UNIKATE

2017 | Berichte aus Forschung und Lehre

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

50 Natur-, Ingenieur- und Wirtschaftswissenschaften

High-Performance und Cloud Computing

Thomas Lippert, Dietrich Wolf	**Supercomputer am Forschungszentrum Jülich**	8
Maritta Heisel, Denis Hatebur, Ludger Goeke, Isabelle Côté	**Risiko in der Wolke?**	16
Jure Gujt, Eckhard Spohr	**Molekulardynamik auf Hochleistungsrechnern**	28
Andreas M. Kempf, Fabian Proch, Andreas Rittler, Martin Rieth, Thuong M. Nguyen	**Vorhersage der Turbulenz in Reaktoren und Motoren**	38
Peter Kratzer	**Spinning round**	48
Zoltan Mann, Andreas Metzger, Klaus Pohl	**Ganz privat?**	58
Pedro José Marrón	**SIMON macht mobil**	66
Ludwig Mochty, Michael Wiese	**Virtuelles Risiko oder Segen?**	76
Rossitza Pentcheva, Markus Gruner	**Neue Materialien errechnen**	86
Jörg Schröder, Axel Klawonn, Lisa Scheunemann, Martin Lanser	**Hochleistungsrechnen trifft Hochleistungsstahl**	98
Reinhard Schütte	**Immer mehr Standard**	108
Dominik Brands, Holger Gollan, Olaf Hasemann	**Ein Großgerät für die Forschung**	116

ISSN 1869-3881

I N H A L T

EDITORIAL 6 Ludwig Mochty, Andreas M. Kempf

Thomas Lippert, Dietrich Wolf
Supercomputer am Forschungszentrum Jülich
8 Beschleuniger für computergestützte Wissenschaften an Universitäten

Das Jülich Supercomputing Centre (JSC) versorgt zusammen mit dem Höchstleistungsrechenzentrum in Stuttgart (HLRS) und dem Leibniz-Rechenzentrum (LRZ) in Garching bei München, die gemeinsam das Gauss Centre for Supercomputing. (GCS e.V.) bilden, die deutsche Wissenschaftslandschaft mit Rechenressourcen der höchsten Leistungsklasse.

Maritta Heisel, Denis Hatebur, Ludger Goeke, Isabelle Côté
Risiko in der Wolke?
16 Die Sicherheitsanalyse von Cloud-Anwendungen

Für Cloud-Kund*innen ergeben sich durch die Nutzung von Cloud-Computing-Diensten verschiedene Vorteile. So entfallen beispielsweise die Kosten für den Betrieb einer eigenen IT-Infrastruktur, und der Umfang der genutzten Ressourcen kann je nach Bedarf flexibel angepasst werden. Die Nutzung von Cloud-Computing-Diensten bringt allerdings auch Risiken mit sich.

Jure Gujt, Eckhard Spohr
Molekulardynamik auf Hochleistungsrechnern
28 Selbstorganisation und dynamische Strukturen in Flüssigkeiten

Dieser Artikel erklärt, wie mit Modellen, Hypothesen, Computern und Programmen Forschungsfragen beantwortet und neue gestellt werden (können).

Andreas M. Kempf, Fabian Proch, Andreas Rittler, Martin Rieth, Thuong M. Nguyen
Vorhersage der Turbulenz in Reaktoren und Motoren
38 High-Performance Computing in der Strömungssimulation

Die möglichst akkurate Vorhersage des Strömungsverhaltens von ~~Flüssigkeiten~~ ist nach wie vor eine der größten Herausforderungen der Ingenieurwissenschaften. Hier setzt die Arbeitsgruppe von Andreas Kempf an.

Peter Kratzer
Spinning round
48 High-Performance Computing für maßgeschneiderte Materialien für die Informationstechnologie

Mit Hilfe der Dichtefunktionaltheorie sind wir in der Lage, den Zusammenhang zwischen der atomaren Struktur und den Eigenschaften eines Materials zu erforschen, ohne dass empirisch gewonnene Kenntnisse als Ausgangspunkt benötigt werden. Der Einsatz von High-Performance Computing ermöglicht es uns, diesen parameterfreien wissenschaftlichen Zugang sogar auf komplexe Systeme, zum Beispiel Grenzflächen zwischen zwei Materialien, auszudehnen.

Zoltan Mann, Andreas Metzger, Klaus Pohl
Ganz privat?
58 Der Schutz persönlicher Daten in der Cloud

Persönliche Daten in der Cloud zu speichern und zu verarbeiten, kann ein Risiko sein, wenn die Daten vor Manipulation und unerlaubtem Zugriff nicht sicher sind. Spätestens seit der NSA-Affäre interessiert die Nutzer*innen: Wer besitzt meine Daten? Wer kann darauf zugreifen? Was passiert bei deren Verarbeitung und Auswertung? Wissenschaftler*innen am paluno (The Ruhr Institute for Software Technology) der Universität Duisburg-Essen arbeiten an softwaretechnischen Lösungen zur sicheren Datenverarbeitung beim Cloud Computing.

50

Pedro José Marrón
SIMON macht mobil
66 Die Cloud ermöglicht personalisierte Mobilität

Drei Forscher der Universität Duisburg-Essen haben sich nach mehreren Jahren Forschung mit der Firma Locoslab selbständig gemacht. Locoslab GmbH ist ein 2012 ins Leben gerufene Spin-off der Universität, das cloud-basierte Produkte und Dienstleistungen rund um das Thema Lokalisierung anbietet. Vielfältige Anwendungsmöglichkeiten durch den technischen Einsatz des so genannten RF (Radio Frequencing) Fingerprinting, beispielsweise in den Bereichen Navigation auf Flughäfen, in Einkaufszentren oder auf Messen. So können personalisierte Mobilitätslösungen realisiert werden.

Ludwig Mochty, Michael Wiese
Virtuelles Risiko oder Segen?
76 Cloud Computing in der Wirtschaftsprüfung

Das Aufkommen des Cloud Computing hat gewaltige Auswirkungen auf die IT-Industrie. Unternehmen wie Google, Amazon und Microsoft, bieten immer leistungsfähigere, zuverlässigere und kostengünstigere Cloud Plattformen an. Unternehmen mit Rechenbedarf hinterfragen zunehmend den Nutzen ihrer eigenen IT-Infrastruktur und prüfen, inwieweit sie durch Nutzung von Cloud Computing Fixkosten abbauen und bedarfsgerecht gegen variable Kosten austauschen können.

Rossitza Pentcheva, Markus Ernst Gruner
Neue Materialien errechnen
86 HPC in der Vorhersage neuartiger elektronischer und magnetischer Eigenschaften auf der Nanoskala mittels atomistischer quantenmechanischer Simulationen

Der vorliegende Artikel illustriert den Beitrag der theoretischen Materialphysik im Bereich Hochleistungsrechnen mit dem Fokus auf Materialien für elektronische Bauelemente sowie für die Energiegewinnung (Thermoelektrika), Energiespeicherung (Wasserspaltung) und Wärmetransport (magnetische Kühlung).

Jörg Schröder, Axel Klawonn, Lisa Scheunemann, Martin Lanser
Hochleistungsrechnen trifft Hochleistungsstahl
98 Herausforderungen zweier Welten

In den letzten Jahrzehnten wurden neue, moderne Stahlsorten entwickelt. Diese Stähle haben gegenüber den bisher dagewesenen verbesserte Eigenschaften und stärken die Attraktivität von Stahl gegenüber Alternativen wie Aluminium und Kunststoff.

Reinhard Schütte
Immer mehr Standard
108 Cloud Computing im Kontext von Enterprise Systems

Eine Herausforderung für Unternehmen in Zeiten der Digitalisierung ist es, das Management der gemeinhin als Standardsoftware deklarierten Enterprise Systems in der Cloud zu bewältigen.

Dominik Brands, Holger Gollan, Olaf Hasemann
Ein Großgerät für die Forschung
116 Hochleistungsrechnen auf der magnitUDE

Die technischen Daten und das „Drumherum“ des Wissenschaftlichen Rechnens mit der magnitUDE thematisiert dieser Text des Supportteams des High-Performance Computings an der UDE.

ABONNEMENT 129

HINWEISE, IMPRESSUM 130

EDITORIAL

Verehrte Leser*innen,

die Ihnen vorliegende Ausgabe der UNIKATE ist ein wahres Unikat – konzentrieren sich die meisten Ausgaben doch auf ein einziges Forschungsgebiet. Im vorliegenden Heft liegen Ihnen aber Aufsätze zu wichtigen Werkzeugen der Natur-, der Ingenieur- als auch der Wirtschaftswissenschaften vor – dem Hochleistungsrechnen und dem Cloud Computing. Beiden ist gemein, dass Rechenressourcen bereitgestellt werden, deren Leistungsfähigkeit weit über dem liegt, was ein einzelner Nutzer jemals für sich selber rechtfertigen oder nutzen könnte. Ebensowenig ist für den Nutzer direkt sichtbar, wo sich die Rechenkapazität befindet, wer sie betreibt und wie die Berechnung technisch realisiert wird.

Zwischen Cloud und Hochleistungscomputing gibt es aber auch wesentliche Unterschiede. Anwendungen in der Cloud erfordern oft eine große Speicherkapazität und viel Rechenleistung. Diese kann über eine große Zahl von Recheneinheiten verteilt sein, die nur selten Daten austauschen müssen. Häufig sind auch die über die Recheneinheiten verteilten Aufgaben stark unterschiedlich und die genutzten Prozessoren oft inhomogen. Beim Hochleistungsrechnen dagegen wird eine einzige Anwendung auf viele homogene Rechenressourcen verteilt, die sich aber permanent über die jeweiligen Rechenergebnisse austauschen müssen. Hochleistungsrechnen erfordert daher eine enge Kopplung der laufenden Prozesse, weshalb die verwendeten Prozessoren in räumlicher Nähe zueinander angeordnet und besonders effizient vernetzt sein müssen.

Historisch ist das Hochleistungsrechnen wohl die älteste Form des wissenschaftlichen Rechnens – zunächst zur Militärforschung (Berechnung von Geschossbahnen oder von Kernreaktionen), später zur Wettervorhersage, zur Simulation von Strukturen und deren Verformung (Fahrzeug-Crash-Simulationen), von Strömungen (Verringerung des Kraftstoffverbrauchs), von Stoff- und Materialeigenschaften sowie von chemischen Reaktionen. Dabei sind die physikalisch-mathematischen Probleme von einer Art, die sich nur durch schnelle Vernetzung der einzelnen Rechenwerke effizient lösen lässt.

Dagegen ist das Cloud Computing wohl die neueste Form des wissenschaftlichen Rechnens. Sein technologischer Vorläufer, das Grid Computing, geht auf das Konzept elektrischer Stromnetze zurück: der Nutzer kann über standardisierte Adapter auf das Netz zugreifen, die technische Realisierung bleibt ihm aber verborgen. Das Grid Computing ist gewissermaßen die Übertragung dieses Konzepts auf IT-Ressourcen, wie Rechnerleistung oder Speicherkapazität. Ein Grid besteht aus geografisch verteilten Ressourcen, die administrativ unabhängig von Organisationen betreut werden, standardisierte, offene Protokolle und Schnittstellen verwenden und bestimmte Eigenschaften in Bezug auf Antwortzeitverhalten, Verfügbarkeit und Durchsatz aufweisen. Groß angelegte wissenschaftliche Anwendungen, wie die des Large Hadron Collider des CERN, erforderten mehr Rechenleistung als ein lokales Cluster-System bereitstellen

konnte. Die schnelle Vernetzbarkeit über das Internet beförderte schließlich die Idee einer gemeinsamen, koordinierten Nutzung von IT-Ressourcen über virtuelle Organisationen hinweg. Während wissenschaftliche Grid-Umgebungen ohne komplizierte Preisgestaltung auskommen, haben bei kommerziell betriebenen Netzwerken Geschäfts- und Preisgestaltungsmodelle eine besondere Bedeutung. Cloud Computing berücksichtigt Fragen, die das Geschäftsmodell betreffen, während das dezentrale Management und der Zugriff auf verteilte Ressourcen (entsprechend dem Grid Computing) nicht weiterverfolgt werden. Vielmehr resultieren gerade aus der Zentralisierung von Rechenressourcen entscheidende wirtschaftliche Vorteile (Economies of Scale). Durch die Virtualisierung ermöglichen Clouds die dynamische Bereitstellung und Rückgabe von IT-Ressourcen. Parallel hierzu wird die Ressourcennutzung überwacht, um durch die dynamische Zuteilung von Anwendungen und Ressourcen für einen Belastungsausgleich zu sorgen.

Sowohl Hochleistungsrechnen wie Cloud Computing stellen essentielle Werkzeuge und Forschungsgegenstände an unserer Universität dar – wovon Sie die folgenden Seiten sicher überzeugen werden.

Beim Hochleistungsrechnen ist die UDE mit der neuen Maschine MagnitUDE außergewöhnlich gut aufgestellt – einem der 500 schnellsten Rechner weltweit, und darunter einer der energieeffizientesten! Diese Maschine stellt den Nutzern nicht nur unkompliziert massive Rechenleistung bereit, sondern ermöglicht auch den Schritt zum exklusiven Club der Nutzer der allergrößten deutschen Rechner in Jülich, Stuttgart und München und sogar darüber hinaus.

Cloud Computing wird an der UDE vor allem dazu verwendet, um Studierenden teure kommerzielle Softwarewerkzeuge und praxisrelevante Daten auf abgesicherte Weise zur Verfügung zu stellen. Diese Anwendungen können sowohl vom Hörsaal als auch von zuhause aufgerufen werden. Gegenstand der Forschung ist Cloud Computing an der UDE vor allem hinsichtlich Sicherheit und Prüfbarkeit sowie Lokalisierung und Navigation.

Als Herausgeber dieser Ausgabe bleibt uns, den Autoren aus den verschiedenen Fachbereichen der UDE für ihre Beiträge zu danken. Ihnen allen ist es gelungen, komplizierte Forschungsgegenstände wissenschaftlich und anschaulich darzustellen – und Ihnen eine interessante Lektüre zu ermöglichen.

oben: Ludwig Mochty, unten: Andreas Kempf. Foto: Vladimir Unkovic

Wir wünschen Ihnen viel Spaß beim Lesen!

Prof. Dr. Ludwig Mochty
Lehrstuhl für Wirtschaftsprüfung, Unternehmensrechnung und Controlling

Prof. Dr.-Ing Andreas Kempf
Vorstand des Center for Computational Sciences and Simulation
Lehrstuhl für Fluiddynamik, Institut für Verbrennung Gasdynamik

Das Jülich Supercomputing Centre (JSC) versorgt zusammen mit dem Höchstleistungsrechenzentrum in Stuttgart (HLRS) und dem Leibniz-Rechenzentrum (LRZ) in Garching bei München, die gemeinsam das Gauss Centre for Supercomputing (GCS e.V.) bilden, die deutsche Wissenschaftslandschaft mit Rechenressourcen der höchsten Leistungsklasse.

Supercomputer am Forschungszentrum Jülich

Beschleuniger für computergestützte Wissenschaften an Universitäten

Von Thomas Lippert und Dietrich Wolf

Wenn das Leistungsvermögen universitärer Rechner nicht mehr ausreicht, dann wissen Forscher der Universität Duisburg-Essen, wie Prof. Peter Kratzer aus der Physik oder Prof. Andreas Kempf aus den Ingenieurwissenschaften (siehe deren Beiträge in dieser UNIKATE-Ausgabe), wo sie ihren Bedarf decken können: am Jülich Supercomputing Centre (JSC). Dieses Zentrum versorgt zusammen mit dem Höchstleistungsrechenzentrum in Stuttgart (HLRS) und dem Leibniz-Rechenzentrum (LRZ) in Garching bei München, die gemeinsam das Gauss Centre for Supercomputing (GCS e.V.) bilden, die deutsche Wissenschaftslandschaft mit Rechenressourcen der höchsten Leistungsklasse. Derzeit verfügen sie zusammen über eine Supercomputerkapazität von mehr als 20 Petaflops Leistung. Von diesen Ressourcen liegen die in Jülich für die oben genannten Wissenschaftler besonders nahe: zum einen im wahrsten Sinne des Wortes, denn nur knapp 100 Kilometer trennen die Universität Duisburg-Essen (UDE) von einem der weltweit führenden Supercomputerzentren. So ist im Bedarfsfall lediglich eine kurze Anreise nötig, um schnelle und persönliche Hilfe zu erhalten, eine Veranstaltung zum Höchstleistungsrechnen zu besuchen oder über anstehende, ressourcenintensive Projektvorhaben zu sprechen. Zum anderen ist mit dem Physiker Prof. Dietrich Wolf, selbst seit 30 Jahren Nutzer des Jülicher Supercomputerzentrums, ein UDE-Wissenschaftler in die Verantwortung eingebunden, die in Jülich zur Verfügung stehende Rechenzeit nach DFG-Exzellenzkriterien und gerecht zu verteilen.

Das Jülich Supercomputing Centre (JSC)

Das JSC bietet mehr als nur den Zugang zu leistungsstarken Rechensystemen. Das Angebot reicht von

Dietrich Wolf. Foto: Vladimir Unkovic

allgemeinen Trainingskursen zum parallelen Programmieren über Spezialkurse zu ausgewählten Themen, wie zum Beispiel I/O und dem Umgang mit großen Datenmengen, bis hin zu thematischen Workshops, Studentenprogrammen, Schulen und Konferenzen. Daneben steht eine vielfältige und leistungsstarke Unterstützungsinfrastruktur zur Verfügung. Diese gliedert sich in die Behandlung allgemeiner Fragen auf der einen und die individuelle Beratung auf der anderen Seite. Diese individuelle Beratung wird im Wesentlichen von fachspezifischen Simulationslaboratorien („SimLabs“) und fachgebietsübergreifenden Querschnittsgruppen über Methoden und Algorithmen, Anwendungsoptimierung, Leistungsanalyse von Programmen, Datenwissenschaften oder Visualisierung geleistet. Ein „SimLab” besteht aus einer Handvoll Wissenschaftler*innen einer Fachrichtung, die zusätzlich zu ihrem fachspezifischen Hintergrund über eine ausgeprägte Expertise im Höchstleistungsrechnen (*High-Performance Computing,* HPC) verfügen [1]. Einerseits unterstützen diese Gruppen im Verbund mit angewandten Informatikern ihre Kolleg*innen von den Universitäten bei der Nutzung der Jülicher HPC-Systeme auf vielfältige Weise, betreiben jedoch andererseits auch eigene Forschung und stehen so unter anderem als Projektpartner für das gemeinsame Einwerben von Drittmitteln an der Seite der universitären Forschung. Dies gilt im gleichen Maße auch für die Wissenschaftler*innen der Querschnittsgruppen, deren Forschung und Unterstützung eher anwendungsübergreifend ausgerichtet ist. Besteht Beratungsbedarf mehr im Hinblick auf Hardware-Details, können technisch orientierte Teams („TechLabs“) des JSC mit eingebunden werden, die sich in enger Zusammenarbeit mit namhaften Herstellern (IBM, Intel, NVIDIA) mit den Themen Rechnerentwicklung, -technologien und -architekturen beschäftigen. Insgesamt verfügt das JSC damit über eine umfassende Unterstützungsinfrastruktur für Fragen aus dem Bereich Höchstleistungsrechnen.

Auch die am JSC betriebenen Supercomputer bieten gute Anknüpfungspunkte an die universitäre Rechnerausstattung und kommen damit den Universitätswissenschaftler*innen entgegen. Mit JURECA („Jülich Research on Exascale Cluster Architectures“), einem Intel Haswell Cluster, steht ein leistungsstarkes General Purpose System zur Verfügung [2]. Dorthin lassen sich die auf Universitätsclustern entwickelten Programme leicht übertragen und sind dann bereit für die Bearbeitung zeit- und datenintensiver Fragestellungen. Auf einigen Dutzend der knapp 2.000 Rechenknoten stehen zusätzlich NIVIDIA-Grafikkarten zur Verfügung, die entsprechend programmierte Anwendungen weiter beschleunigen oder für Visualisierungszwecke genutzt werden können. Genügen auch diese Kapazitäten nicht mehr und verlief die Anpassung des Programms unter anderem an eine höhere Parallelität erfolgreich, so steht mit JUQUEEN („Jülich Blue Gene/Q“) ein massiv paralleler Supercomputer der höchsten Leistungsklasse zur Verfügung [3]. Dieses System kann sein Geschwindigkeitspotential allerdings erst voll ausspielen, wenn mehrere zehn- oder sogar hunderttausende Rechenkerne parallel und effizient an einem Programm arbeiten. Um diese hohe Parallelität und Effizienz zu erreichen, ist eine enge Zusammenarbeit mit den oben beschriebenen Unterstützungsstrukturen am JSC höchst angeraten.

Da die Systeme in Jülich teure und nachgefragte Forschungsinfrastrukturen sind, ist der Zugang nicht so frei wie bei universitären Systemen, sondern erfolgt – wie natürlich auch an den beiden anderen Mitgliedszentren des GCS in Stuttgart und München – über ein Begutachtungsverfahren. So wird die Rechenzeit auf JURECA durch eine unabhängige Einrichtung, das John von Neumann-Institut für Computing (NIC), vergeben, nachdem die eingegangenen Anträge einen transparenten, exzellenzorientierten und kompetitiven Begutachtungsprozess – angelehnt an DFG-Begutachtungsverfahren – durchlaufen haben. Für die Rechenzeitvergabe auf JUQUEEN ist der GCS-Lenkungsausschuss zuständig, der sich der lokalen Expertise des NIC bedient. Die Rechenzeit für Projekte von Forschern an Universitäten oder anderen wissenschaftlichen Einrichtungen ist übrigens kostenfrei.

Ein Beispiel für die Zusammenarbeit zwischen JSC und Universitäten

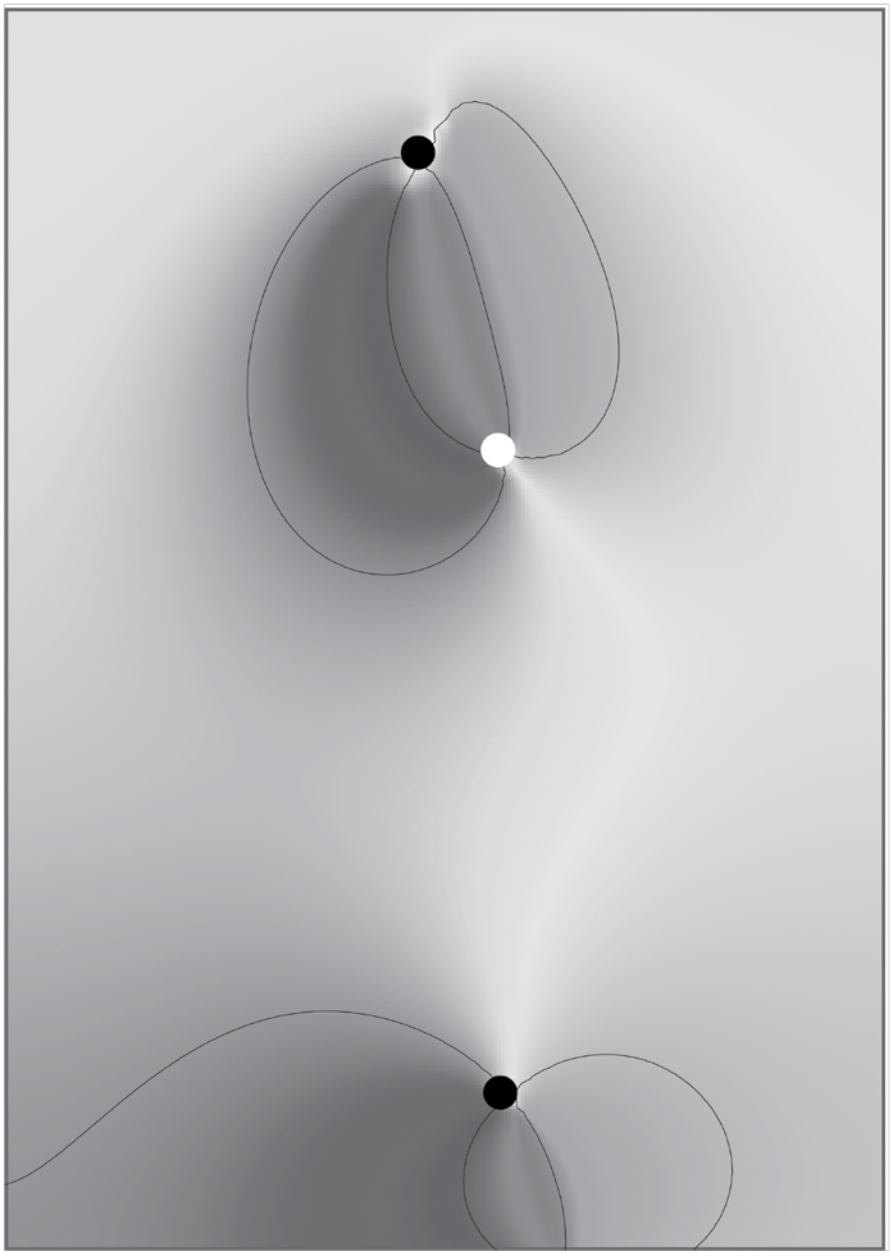
Ein Forschungsthema, über das in den letzten Jahren viel nachgedacht wurde, sind neue, energieeffizientere Möglichkeiten der digitalen Datenspeicherung und -verarbeitung. Dazu gehört vor allem das Feld, für das sich die Bezeichnung „Spintronik“ eingebürgert hat [4]. Informationsträger ist dabei das magnetische Moment der Atome, aus denen die kristallinen Bauelemente bestehen. Diese „Spins“ lassen sich elek-„tronisch“ manipulieren, was die Wortschöpfung erklärt. In der Regel sind die dafür nötigen Ströme viel schwächer als bei der heute verwendeten Computertechnologie. Spintronik hätte das Potential, in einigen Jahrzehnten den enormen Bedarf an elektrischer Energie für die Datenverarbeitung deutlich zu reduzieren.

Mit dem Physik-Nobelpreis 2016 wurden David Thouless, Duncan Haldane und Michael Kosterlitz dafür geehrt, dass sie mit Hilfe des mathematischen Gebiets der Topologie physikalische Phänomene vorhersagen bzw. erklären konnten. Ein gutes Beispiel aus dem Bereich der Spintronik sind magnetische Wirbel, von denen es diverse Arten gibt. Zur Illustration soll hier das sogenannte „magnetische Skyrmion“ beschrieben werden, dessen kürzliche Entdeckung [5] allerdings nicht auf die Nobelpreisträger zurückgeht. In

einer dünnen magnetischen Schicht, bei der die magnetischen Momente bevorzugt parallel und senkrecht zur Schicht stehen, kann man sich ein Skyrmion als winzige, „punktförmige“ magnetische Domäne vorstellen, deren Magnetisierung nach oben aus der Schicht herauszeigt. Von diesem Punkt radial nach außen dreht sich die Richtung der magnetischen Momente allmählich, bis sie in eine Domäne mit nach unten zeigender Magnetisierung münden, welche die punktförmige Domäne in der Mitte rundherum umgibt. Diese magnetische Konfiguration ist topologisch stabil, so dass ihr Vorhandensein als gesetztes Bit interpretiert werden kann.

Um solche durch magnetische Skyrmionen dargestellten Bits zu verschieben, kann man anstelle eines elektrischen Stroms auch ein äußeres Magnetfeld benutzen. Es muss jedoch eine so geringe räumliche Ausdehnung haben, dass ein einziges Skyrmion eine klar erkennbare energetisch günstige Position im Magnetfeld hat. Das lässt sich durch magnetische Rasterkraftmikroskopie realisieren. Dabei wird eine feiner als ein Haar ausgezogene permanentmagnetische Spitze im Abstand von einigen Angström über die magnetische Schicht geführt. Das Skyrmion wird dann im Magnetfeld der Spitze mitgezogen und bremst dabei die Spitze ab. Die Spitze erfährt also eine Reibungskraft, ohne die Schicht, über die hinweg sie sich bewegt, zu berühren. Reibungskraft mal Geschwindigkeit ist die Leistung, die aufgebracht werden muss, um das Skyrmion zu bewegen.

Die magnetische Reibung wurde von der Arbeitsgruppe um Prof. Wolf an der Universität Duisburg-Essen intensiv durch Computersimulationen im Rahmen des Sonderforschungsbereichs 616 „Energiedissipation an Oberflächen“ untersucht [6–8]. Dabei wurden die Bewegungsgleichungen für eine große Zahl wechselwirkender magnetischer Momente in Gegenwart einer bewegten magnetischen Spitze



(1) Magnetische Wirbelstraße hinter der Spitze, die sich über dem oberen schwarzen Punkt befindet und sich nach oben bewegt. Die Grau-Abstufungen kodieren die Richtung des in die Bild-Ebene projezierten magnetischen Moments: Geht man im Uhrzeigersinn um die schwarzen Punkte herum, dreht sich auch die Spinrichtung im Uhrzeigersinn. Oberhalb der schwarzen Punkte zeigt sie nach oben. In den von den dünnen Linien umschlossenen Bereichen zeigt sie nacheinander vorwiegend nach rechts, nach unten und dann nach links. Schwarze Punkte markieren Vortexzentren, weiße Antivortexzentren.
Quelle: M. P. Magiera: Magnetic friction in driven Heisenberg-like systems, Doktorarbeit, Universität Duisburg-Essen (2012), Abb. 7.10.

numerisch gelöst. Es stellte sich bald heraus, dass die Rechenleistung der lokal verfügbaren Clusterrechner nicht ausreichte. Weil das Computerprogramm sich gut parallelisieren ließ, das heißt auf eine große Zahl von Prozessoren verteilt werden konnte und dadurch für den General Purpose Supercomputer in Jülich geeignet war, wurde Rechenzeit am John von Neumann-Institut für Computing (NIC) beantragt. Ein solcher Antrag hat ungefähr dasselbe Format wie ein Forschungsprojekt, das bei der Deutschen Forschungsgemeinschaft beantragt wird, mit

Thomas Lippert. Foto: Vladimir Unkovic

dem Unterschied, dass die benötigten Computerressourcen (Rechenzeit und Speicherplatz) überzeugend abgeschätzt werden müssen. Dazu ist ein Nachweis der Skalierbarkeit des parallelisierten Programms unerlässlich. Damit ist gemeint, dass das Programm im Idealfall nur ein zehntausendstel der Rechenzeit benötigt, wenn es auf zehntausend Prozessoren verteilt wird. Das wird natürlich in der Regel nicht erreicht, weil die Prozessoren, in unserem Beispiel wegen der Wechselwirkung der magnetischen Momente, nicht ganz unabhängig voneinander rechnen können, sondern untereinander Daten austauschen müssen.

Aufgrund der Expertengutachten über den Rechenzeitantrag bewilligt dann die NIC-Rechenzeitkommission Rechenzeit für ein Jahr. Wegen der großen Nachfrage aus den meisten Universitäten Deutschlands ist oft nur eine Teilbewilligung des Antrags möglich.

Dann müssen gegebenenfalls Teile des Arbeitsprogramms zurückgestellt werden. Mit der Bewilligung gehen Verpflichtungen für die Antragsteller einher: Nach einem Jahr ist ein Statusbericht über das Projekt fällig, und selbstverständlich gehört in die mit Hilfe der Supercomputer des JSC entstandenen wissenschaftlichen Publikationen eine entsprechende Danksagung. Herausragende Projekte dürfen sich bei dem zweijährlich stattfindenden NIC-Symposium mit einem Vortrag und einem Buchbeitrag präsentieren (siehe z.B. [4]).

Ein besonders spektakuläres Ergebnis zur magnetischen Reibung ist in Abbildung (1) dargestellt [6]. Wenn die Wechselwirkung der Spitze mit der magnetischen Schicht und ihre Geschwindigkeit Werte in einem bestimmten Bereich haben, lösen sich Wirbel (und Antiwirbel) von der Spitze, so dass diese eine Art Wirbelstraße hinter sich herzieht, ähnlich wie ein schnell durch das Wasser gezogenes Boot. Der Energieeintrag der bewegten Spitze in die magnetische Schicht ist dann groß genug, um Wirbel-Antiwirbel-Paare zu erzeugen. Die Wirbel in Abbildung (1) sind keine Skyrmionen, sondern Vortices, wie sie auch von den Nobelpreisgewinnern Kosterlitz und Thouless betrachtet wurden. Ohne den Supercomputer in Jülich zu benutzen, hätte das Projekt zur magnetischen Reibung nicht so viele neue Erkenntnisse innerhalb von drei Jahren gebracht.

Ausblick

Die unaufhaltsame Zunahme der Parallelität und Leistungsfähigkeit von Rechenknoten in heutigen und zukünftigen Supercomputern, getrieben vom Drang zur Simulation immer realitätsnäherer und komplexerer Modelle, stellt die Wissenschaftler aus den computer-gestützten Anwendungswissenschaften vor schwierige Herausforderungen. Erhielt man vor zwei Jahrzehnten Geschwindigkeitszuwächse von Simulationsprogrammen noch nahezu geschenkt durch die Entwicklung immer schnellerer Prozessoren, so sind diese Zeiten aufgrund physikalischer und technischer Begrenzungen längst vorbei. Heute gilt es, die Parallelität und Effizienz der Anwendungen so weit wie möglich zu steigern und den auf mehreren Ebenen etablierten Parallelismus für die eigene Anwendung nutzbar zu machen. So ist einerseits die Zunahme der Rechenknoten pro System zu beachten, andererseits aber auch die Zunahme der Rechenkerne pro Prozessor (“Multi-Core”) sowie die an die Prozessoren angedockten Beschleunigerelemente, wie Grafikkarten, Intel Xeon Phis oder FPGAs (“Many-Core”). Darüber hinaus sind dann noch die verschiedenen lokalen und entfernten Speicherhierarchien möglichst effizient einzusetzen. Vor diesem Hintergrund kommt der Beherrschung des Werkzeugs Höchstleistungsrechnen eine entscheidende Schlüsselrolle zu im Wettbewerb um genauere, neue oder bisher nicht erzielbare Resultate aus Simulationsrechnungen.

Die Validierung von Simulationsrechnungen stützt sich immer auf Experimente und Beobachtungen. Je detaillierter und genauer diese Messungen sind, desto deutlicher zeigen sich gegebenenfalls Defizite der Simulationsvorhersagen und Hinweise, wie Modelle verbessert werden können. In den letzten Jahren ist die Datenmenge aus Beobachtungen, wie zum Beispiel Sensornetzwerken in der Erdsystemforschung oder Satelliten in der Klimaforschung, durch sowohl räumlich als auch zeitlich immer höher auflösende Sensoren stark angestiegen. Diese Sensoren zeichnen auch Ereignisse auf, die bislang nicht beobachtet wurden. Diese sind natürlich von besonderem Interesse, aber es ist nicht immer leicht, sie aus den Daten herauszufiltern. Altbekannte Verfahren zur Datenanalyse wie die Suche nach Mustern, Ausreißern oder Methoden zur Klassifizierung werden auf große Datenbestände angewandt [9]. Für einige dieser Verfahren werden in kommerziellen Umgebungen spezielle Plattformen, wie Apache Hadoop oder Spark, eingesetzt. Das JSC evaluiert diese Plattformen prototypisch und prüft, ob und wie diese integriert werden können. Gleichzeitig wird an der Skalierung von Datenanalyseverfahren mit OpenMP und MPI auf der bestehenden HPC-Infrastruktur gearbeitet.

Zusammen mit der zunehmenden Bedeutung der Beschleuniger, der Verfügbarkeit schneller Speicher und der oben beschriebenen, immer stärker werdenden Nachfrage nach Datenanalyseverfahren und Deep Learning Methoden durch neu hinzukommende Anwendungsfelder ist eine stete Anpassung der Supercomputing-Strategie des JSC erforderlich. So soll das bisher dual ausgerichtete Architekturmodell, bestehend aus einem General Purpose Supercomputer und einem hochskalierenden System durch ein modulares Modell erweitert werden. Letzteres entwickelt den bisherigen Ansatz konsequent weiter: ein Gene-

ral Purpose Modul wird mit mehreren unterschiedlichen Beschleuniger-Modulen gekoppelt, um so die Anforderungen auch hoch komplexer Anwendungen effizient unterstützen zu können. Die Kopplung der Module erfolgt dabei zum einen über ein gemeinsames hierarchisches Datenspeichersystem, das durch die Kombination von Chip-basiertem Speicher mit Festplatten kosteneffizient höchste Datenraten (im Bereich von TeraByte/s) mit großer Kapazität verbindet. Zusätzlich können die Module direkt durch ein Hochgeschwindigkeitsnetzwerk miteinander verbunden werden. Dieses neuartige Modell ermöglicht sowohl Workflows, die mehrere Module nacheinander nutzen als auch Anwendungen, die bestimmte rechenintensive Teile vom General Purpose Modul auf jeweils geeignete Beschleuniger-Module auslagern. Durch seine inhärente Erweiterbarkeit ermöglicht das modulare Architektur-Modell zudem evolutionären Ausbau und Erneuerung eines Supercomputer-Komplexes in bisher nicht gekannter Weise.

Fazit

Zentren wie das Jülich Supercomputing Centre und die Universitäten ergänzen sich gegenseitig und halten gemeinsam die computergestützten Wissenschaften auf international konkurrenzfähigem Niveau. Der Forschungsbedarf an den Universitäten rechtfertigt im Spitzenbereich die staatlichen Investitionen in teure Supercomputer. Deren effiziente Nutzung erfordert nicht nur einen professionellen technischen Betrieb. Genauso wichtig sind die Service-Infrastruktur am JSC, der intensive Austausch zwischen Nutzer*innen und Betreiber*innen und das transparente Qualitätsmanagement durch die unabhängige Rechenzeitkommission. Das alles ist nach 30 Jahren Höchstleistungsrechnen gut eingespielte und ständig sich erneuernde Routine. Was aber nicht unterschätzt werden sollte, ist der Innovationsmotor, den die Symbiose aus Universitäten und Supercomputerzentren jenseits der gut eingespielten Praxis darstellt. In dieser Symbiose ist der Part des JSC, die neuesten technischen Entwicklungen vorauszudenken und zu bewerten und entsprechende Weichen für zukünftige Investitionen im HPC-Bereich zu stellen. Der Part der Forscher*innen an den Universitäten und Forschungseinrichtungen ist es, mit Kreativität und Fachkompetenz die Chancen zu erkennen, die zukünftige HPC-Infrastrukturen für den Erkenntnisgewinn bieten.

Summary

Supercomputing of the highest performance class is represented in Germany by the Gauss Centre for Supercomputing (GCS), an association of the three national supercomputing centers, the High Performance Computing Center (HLRS) in Stuttgart, the Leibniz Supercomputing Centre (LRZ) in Garching near Munich, and the Jülich Supercomputing Centre (JSC) at Forschungszentrum Jülich, which was the first to be established and will celebrate its 30th anniversary this year. Together they currently offer a performance of more than 20 petaflops (20 times 1015 floating point operations per second) to computational scientists in Germany (and elsewhere in Europe, to some extent). This performance is urgently needed for research projects across many disciplines, such as fluid dynamics and engineering, earth and atmospheric studies, materials science and chemistry, astro- and elementary particle physics, biology and medicine. Most of the resources are granted to research groups at German universities. Scientists gain access to the high-end systems of GCS after successfully passing a well-established, transparent evaluation process that ensures the highest quality. High-performance computing (HPC) at an internationally competitive level is an important driving force for innovation, and requires thinking ahead, both at the operational level, where rational decisions concerning the development of the supercomputer technology have to be made, and at the scientific level, where the potential to create new insights using HPC has to be explored. As an example, this article discusses the energy consumption of novel magnetic structures for spintronic applications.

Danksagung

Die Autoren bedanken sich bei Norbert Attig für die Mitwirkung an den Abschnitten JSC und Ausblick sowie bei Daniel Mallmann für die Mitarbeit am Ausblick.

Anmerkungen/Literatur

[1] N. Attig, P. Gibbon, T. Lippert (2011); Trends in Supercomputing: The European Path to Exascale; Comp. Phys. Comm. 182, (2011) 2041 – 2046 [10.1016/j.cpc.2010.11.011]
[2] D. Krause, Ph. Thörnig (2016); JURECA: General-purpose supercomputer at Jülich Supercomputing Centre. Journal of large-scale research facilities, 2, A62; http://dx.doi.org/10.17815/jlsrf-2-121
[3] M. Stephan, J. Docter (2015); JUQUEEN: IBM Blue Gene/Q Supercomputer System at the Jülich Supercomputing Centre. Journal of large-scale research facilities, 1, A1; http://dx.doi.org/10.17815/jlsrf-1-18
[4] N. Atodiresei, V. Caciuc, P. Lazic, S. Blügel (2012); Understanding molecular electronics and spintronics from first principles simulation; in: NIC Symposium 2012 – Proceedings, eds. K. Binder, G. Münster, M. Kremer, NIC-Series Vol. 45 (Forschungszentrum Jülich, 2012) 177
[5] S. Mühlbauer, B. Binz, F. Jonietz, C.Pfleiderer, A. Rosch, A. Neubauer, R. Georgii, P. Böni; Skyrmion lattice in a chiral magnet; Science 323 (2009) 915
[6] M. P. Magiera, A. Hucht, H. Hinrichsen, S. R. Dahmen, D. E. Wolf; Magnetic vortices induced by a moving tip; EPL 100 (2012) 27004
[7] M. P. Magiera; Energy dissipation of moved magnetic vortices; EPL 103 (2013) 57004
[8] M. P. Magiera, S. Schulz; Magnetic vortices induced by a monopole; IEEE Transactions on Magnetics 50 (2014) 4300604
[9] M.S. Memon, M. Riedel, A. Memon, C. Koeritz, A. Grimshaw, H. Neukirchen; Enabling Scalable Data Processing and Management through Standards-based Job

Execution and the Global Federated File System; Scalable computing 17, (2014) 115–128; DOI:10.12694/scpe.v17i2.1160

Die Autoren

Dietrich Wolf erhielt sein Diplom in Theoretischer Physik 1978 von der Universität zu Köln, wo er 1982 auch promovierte. Nach Auslandsaufenthalten an der Carnegie-Mellon University in Pittsburgh, USA, und am Institut Laue-Langevin in Grenoble, Frankreich, habilitierte er sich 1988 an der Universität zu Köln. Seit 1992 ist Dietrich Wolf Universitätsprofessor für Theoretische Physik an der Fakultät für Physik der Universität Duisburg-Essen. Im Höchstleistungsrechenzentrum des Forschungszentrums Jülich leitete er die Forschungsgruppe Vielteilchenphysik von 1994 bis 1996. 1990 erhielt Dietrich Wolf den Gay-Lussac-Humboldt Preis. Von 2000 bis 2005 war er Vorsitzender des Wissenschaftlichen Rats des John-von-Neumann Instituts für Computing (NIC) in Jülich. 2016 verlieh ihm die Budapest University of Technology and Economics die Ehrendoktorwürde.

Thomas Lippert erhielt sein Diplom in Theoretischer Physik 1987 von der Universität Würzburg. Er promovierte 1993 an der Universität Groningen in Computerwissenschaften und 1998 an der Universität Wuppertal im Bereich der Theoretischen Elementarteilchenphysik. Seine Habilitation schloss er 2001 an der Universität Wuppertal ab. Thomas Lippert ist Direktor des Institutes for Advanced Simulations (IAS) am Forschungszentrum Jülich, Leiter des Jülich Supercomputing Centre (JSC), geschäftsführender Direktor des John von Neumann-Instituts für Computing (NIC) und hat einen Lehrstuhl für Theoretische Physik an der Universität Wuppertal.

*Für Cloud-Kund*innen ergeben sich durch die Nutzung von Cloud-Computing-Diensten verschiedene Vorteile. So entfallen beispielsweise die Kosten für den Betrieb einer eigenen IT-Infrastruktur, und der Umfang der genutzten Ressourcen kann je nach Bedarf flexibel angepasst werden. Die Nutzung von Cloud-Computing-Diensten bringt allerdings auch Risiken mit sich.*

Risiko in der Wolke?

Die Sicherheitsanalyse von Cloud-Anwendungen

Von Maritta Heisel, Denis Hatebur, Ludger Goeke und Isabelle Côté

Der Begriff „Cloud Computing" bezeichnet ein relativ neues Geschäftsmodell und die in diesem Kontext verwendeten Technologien. Im Rahmen des Cloud Computings werden verschiedene Computerressourcen in Form von sogenannten Cloud-Computing-Diensten angeboten. Hierbei bezeichnet „Dienst" unterschiedliche Angebote, wie zum Beispiel Rechenleistung, Server, Speicherplatz und Applikationen. Letztere reichen von einfachen Applikationen, wie beispielsweise Textverarbeitung, bis hin zu komplexen Anwendungen, wie beispielsweise kompletten Lagerverwaltungssystemen. Die Grundidee ist, dass aus einem Ressourcen-Pool (z.B. ein Rechenzentrum, welches virtuelle Maschinen (VMs) nutzt) diese Dienste durch Cloud-Dienstanbieter angeboten werden können.

Hierbei werden die Cloud-Computing-Dienste über ein performantes Netzwerk (Internet) bereitgestellt. Der Zugriff auf die entsprechenden Ressourcen erfolgt mit Hilfe von Standardprotokollen wie beispielsweise TCP/IP. Der Umfang der Nutzung von Ressourcen ist skalierbar und kann sehr schnell an die benötigten Ressourcen eines*einer Kund*in individuell angepasst werden.

Die oben genannten Ressourcen lassen sich drei Ebenen, auch Service-Modelle genannt, zuordnen, die im Folgenden kurz erläutert werden.

Cloud-Ebenen

Cloud-Computing-Dienste werden in Form der Service-Modelle *Software as a Service* (SaaS), Platform as a Service (PaaS) und *Infrastructure as a Service* (IaaS) angeboten. So verwendet ein SaaS-Dienst Funktionalitäten, die durch einen PaaS-Dienst bereitgestellt werden. Ein PaaS-Dienst baut wiederum auf einem IaaS-Dienst auf. Die Service-Modelle auf den verschiedenen Ebenen sind:

- SaaS: Bereitstellung von Applikationen, die innerhalb der Cloud-Infrastruktur ausgeführt werden und durch die Endkund*innen über das Internet genutzt werden. Die Endkund*innen bestellen hier Software, die sie nutzen möchten (z.B. E-Mail, Textverarbeitung, Lagerverwaltung, etc.).
- PaaS: Bereitstellung einer Plattform inklusive einer Programmierschnittstelle und Werkzeugen zur Entwicklung und Ausführung von Cloud-Applikationen durch die Cloud-Kund*innen. Die Kund*innen dieses Dienstes sind häufig Entwickler*innen, welche die Plattform und ihre Schnittstellen als

Maritta Heisel. Foto: Vladimir Unkovic

Grundlage nutzen, um weitere Applikationen zu entwickeln, die dann wiederum als SaaS angeboten werden können.
• IaaS: Bereitstellung von Computerressourcen (Rechenleistung, Speicher, usw.) als Cloud-Infrastruktur. Hierbei ist die unterliegende Cloud-Infrastruktur für Cloud-Kund*innen nicht sichtbar und dementsprechend nicht konfigurierbar. Vielmehr werden den Cloud-Kund*innen beispielsweise Betriebssysteme oder virtuelle Maschinen zur Verfügung gestellt, die diese dann konfigurieren können. Hier können die Kund*innen also Hardware in Anspruch nehmen, ohne dass diese tatsächlich gekauft und in den eigenen Räumen untergebracht werden muss.

Man kann verschiedene Modelle in Bezug auf die Bereitstellung von Cloud-Computing-Diensten unterscheiden[1]:
• Private Cloud (nicht-öffentliche Cloud): Die Cloud-Infrastruktur wird ausschließlich für eine einzige Organisation betrieben. Hierbei kann der Betrieb durch die Organisation selbst oder einen externen Dienstleister erfolgen.
• Community-Cloud: Die Cloud-Infrastruktur ist eine nicht-öffentliche Cloud, die von einer Gemeinschaft von Organisationen mit ähnlichen Anforderungen (z.B. Sicherheitsanforderungen, Verwendungszweck oder Geschäftsfelder) genutzt wird. Hierbei kann eine Community-Cloud durch die Organisationen selbst oder einen externen Dienstleister betrieben werden.
• Public Cloud (öffentliche Cloud): Die Cloud-Infrastruktur kann generell durch die Allgemeinheit oder eine entsprechend große Interessengruppe genutzt werden. Der Betrieb der Cloud-Infrastruktur erfolgt durch eine Organisation, deren Geschäftsszenario in der Bereitstellung von Cloud-Computing-Diensten besteht. Eine solche Organisation wird auch als Cloud-Dienstanbieter bezeichnet. Für die Nutzung der entsprechenden Cloud-Computing-Dienste kann der Cloud-Dienstanbieter je nach Geschäftsmodell ein Entgelt berechnen.
• Hybrid Cloud (hybride Cloud): Die Cloud-Infrastruktur wird durch eine Komposition von zwei oder mehreren Clouds gebildet. Bei den beteiligten Clouds kann es sich um Public Clouds, Private Clouds und/oder Community Clouds handeln. Im Rahmen einer Hybrid Cloud bleiben die einzelnen Clouds eigenständig. Sie werden mit Hilfe entsprechender Technologien verbunden, um den Austausch von Daten zwischen den einzelnen Clouds und die Portabilität von Cloud-Anwendungen zu ermöglichen.

Möglichkeiten des Cloud Computing

Für Cloud-Kund*innen ergeben sich durch die Nutzung von Cloud-Computing-Diensten verschiedene Vorteile. So entfallen beispielsweise die Kosten für den Betrieb einer eigenen IT-Infrastruktur, und der Umfang der genutzten Ressourcen kann je nach Bedarf flexibel angepasst werden. Die Nutzung von Cloud-Computing-Diensten bringt allerdings auch Risiken mit sich. So könnten beispielsweise die Vertraulichkeit und Integrität von Daten der Cloud-Kund*innen durch böswilliges Personal des Cloud-Dienstanbieters oder eine versehentliche Fehlkonfiguration der Cloud-Infrastruktur verletzt werden. Auch die Verfügbarkeit der Kund*innendaten könnte zum Beispiel durch Störungen in Bezug auf die Internetverbindung oder das Rechenzentrum sowie durch die Insolvenz von Cloud-Dienstanbietern gefährdet sein.

Informationssicherheit von Clouds

Ein generelles Hindernis bei der Verbreitung von Cloud-Computing-Diensten besteht darin, dass die Anbieter von Cloud-Computing-Diensten ihre potentiellen Kund*innen nur schwer von der Sicherheit ihrer Dienste überzeugen können. In der Studie „Cloud-Monitor 2016“, die von der Bitkom Research GmbH und KPMG AG erstellt wurde, wird erstmals eine Nutzung von Cloud Computing durch die Mehrheit der befragten Unternehmen beobachtet. Weiterhin sei ein starker Anstieg der Cloud-Nutzung bei Unternehmen von kleiner und mittlerer Größe zu verzeichnen. Es wird allerdings auch angemerkt, dass Sicherheitsbedenken eine noch stärkere Verbreitung von Cloud-Computing-Diensten verhindern würden. In Bezug auf die Public Cloud bezögen sich die größten Befürchtungen der befragten Unternehmen auf unberechtigte Zugriffe auf sensible Unternehmensdaten und den möglichen Verlust von Daten.

Eine stärkere Verbreitung von Cloud Computing könnte erreicht werden, wenn die Cloud-Dienstanbieter die Sicherheitsfunktionalitäten ihrer Dienste transparent für potentielle Kund*innen machen würden. Diese Transparenz kann durch eine Zertifizierung von Cloud-Computing-Diensten durch einen unabhängigen Dritten erreicht werden. Dazu müssen die Sicherheitsanforderungen und die Sicherheitsfunktionalitäten der Cloud-Computing-Dienste genau beschrieben werden.

ISO 27001

Eine Möglichkeit besteht in der Zertifizierung eines sogenannten *Information Security Management Systems* (ISMS). Im Rahmen eines ISMS werden die Prozesse bezüglich der Informationssicherheit einem Management-Rahmen unterstellt. Hierdurch soll erreicht werden, dass ein ausreichendes Niveau an Informationssicherheit innerhalb einer Organisation erreicht und beibehalten wird. Hierbei werden auch Änderungen von Prozessen innerhalb einer Organisation sowie Änderungen in Bezug auf die Gefahrenlage berücksichtigt. Das deutsche Bundesamt für Sicherheit in der

Informationstechnik (BSI) unterstrich 2012 in einer Veröffentlichung[2] ausdrücklich die Notwendigkeit des Betriebes eines ISMS auf Seiten der Cloud-Dienstanbieter.

Bei der *ISO 27001:2013* (ISO/IEC 2013) handelt es sich um einen international anerkannten Standard, der Anforderungen an ein ISMS definiert. Diese Anforderungen beziehen sich auf die Planung, den Betrieb und die kontinuierliche Verbesserung eines ISMS. Ein ISMS nach ISO 27001:2013 wird durch eine akkreditierte Zertifizierungsstelle zertifiziert. Der Managementteil der ISO 27001:2013 definiert die Tätigkeiten, die durchzuführen sind. Hierzu gehören die Definition des Geltungsbereiches des ISMS, die Durchführung einer Risikoanalyse oder die Durchführung interner Audits. Im Anhang A der ISO 27001:2013 werden Maßnahmen definiert, die innerhalb des ISMS umgesetzt werden müssen, wenn sie für die betreffende Organisation relevant sind. Bei diesen Maßnahmen handelt es sich unter anderem um Informationssicherheits-Prozesse oder zu erstellende Informationssicherheits-Richtlinien.

ClouDAT

Ein Problem bei der Umsetzung eines ISMS nach ISO 27001:2013 besteht darin, dass in der ISO-Norm dazu keine konkreten Vorgehensweisen beschrieben werden. Im Rahmen des Forschungsprojektes *ClouDAT*[3] wurde eine musterbasierte Methode für die Planungsphase eines ISMS nach ISO 27001:2013 im Kontext des Cloud Computing entwickelt. Hierbei werden insbesondere die Definition des ISMS-Geltungsbereiches und die Durchführung der Risikoanalyse unterstützt. Zusätzlich wird die entwickelte Methode durch ein Software-Werkzeug unterstützt, das ebenfalls den Namen *ClouDAT* (Cloud Data Analysis Tool) trägt. Das ClouDA-Tool ermöglicht die Erfassung der notwendigen Informationen, sowie die Generierung von notwendigen Dokumentationsartefakten. Dieses Vorgehen unterstützt das Tool durch einen musterbasierten Ansatz. So erfolgt die Definition des ISMS-Geltungsbereichs beispielsweise durch die Instanziierung eines grafischen Musters. Dieses Muster spezifiziert bereits die Elemente, die im Kontext der meisten Cloud-Computing-Dienste relevant sind. Weiterhin werden Kataloge mit textbasierten Mustern in Bezug auf grundlegende Bedrohungen für und Sicherheitsanforderungen an Cloud-Computing-Dienste bereitgestellt. Eine ausführliche Beschreibung der *ClouDAT*-Methode und des ClouDA-Tools geben wir in den Abschnitten „Die ClouDAT-Methode“ und „ClouDA-Tool“. Die ClouDAT-Methode und das ClouDA-Tool wurden im Rahmen einer Fallstudie in Zusammenarbeit mit dem Lanfer Systemhaus aus Dortmund evaluiert. Hierbei wurde ein Cloud-Dienst-Modell in Form von IaaS betrachtet. Ein elementarer Bestandteil von ClouDAT ist die im nächsten Abschnitt beschriebene Risikoanalyse.

Risikoanalyse

Eine Risikoanalyse wird für die Assets einer Organisation durchgeführt. Hierbei bezieht sich der Begriff Asset auf alles, was einen Wert für die Organisation darstellt. So kann es sich bei Assets um Geschäftsprozesse, Informationen oder physische Dinge (z.B. Gebäude, Server, Netzwerk) handeln. Das Ziel der Risikoanalyse besteht darin, einen Wert für das Risiko zu ermitteln, dass die Informationssicherheits-Eigenschaften Vertraulichkeit, Integrität und Verfügbarkeit eines Asset kompromittiert würden. Dazu werden jeweils Werte in Bezug auf die Bedeutung von Assets (engl. Asset Value) bestimmt sowie die Bedrohungen (engl. Threats) für Assets und Schwachstellen (engl. Vulnerabilities) von Assets betrachtet.

Bezüglich der Durchführung der Risikoanalyse definiert die ISO 27001:2013 keine konkrete Methode. Es wird lediglich vorgeschrieben, dass die verwendete Risikoanalysemethode valide Werte produziert, die reproduzierbar und vergleichbar sind. Im Rahmen der ISO 2700x-Normenreihe beschreibt der informative Standard ISO 27005:2011 Hilfestellungen zur Durchführung einer Risikoanalyse.

Die ClouDAT-Methode umfasst ein Vorgehen für die Durchführung der Risikoanalyse, die dementsprechend durch das ClouDA-Tool unterstützt wird. Da die verschiedenen Schritte der ClouDAT-Methode innerhalb des ClouDA-Tools durch separate Plugins umgesetzt werden, könnte die vorhandene Risikoanalysemethode ohne großen Aufwand durch eine andere Methode ersetzt werden. Die Durchführung der ClouDAT-Risikoanalyse wird im Rahmen der Beschreibung der ClouDAT-Methode in Schritt 4 „Risiken bewerten“ ausführlicher diskutiert.

Die ClouDAT-Methode

Mit Hilfe der ClouDAT-Methode können Cloud-Computing-Dienste in Form von SaaS, PaaS und IaaS, sowie die korrespondierenden relevanten Geschäftsprozesse für eine Begutachtung durch Dritte dokumentiert werden. So können für im Aufbau befindliche beziehungsweise bestehende Cloud-Computing-Dienste Sicherheitsrisiken aufgedeckt werden. Sicherheitsrisiken wären beispielsweise das Öffentlichwerden von vertraulich zu haltenden Daten, unbefugte Zugriffe von Mitarbeiter*innen des Cloud-Dienstanbieters auf Daten, Veränderung der Funktionalität von Diensten, oder eine Einschränkung der Verfügbarkeit von Diensten. Die ClouDAT-Methode ist für alle Arten von Clouds einsetzbar.

Bei der Dokumentation der ermittelten Sicherheitsanforderungen an Cloud-Computing-Dienste

müssen das deutsche Recht im Bereich IT-Sicherheit und insbesondere die Einhaltung des Bundesdatenschutzgesetzes durch den Cloud-Dienstanbieter beachtet werden. Ein*e potentielle*r Cloud-Computing-Nutzer*in soll anhand der Dokumentation der realisierten Sicherheitsanforderungen durch den Cloud-Dienstanbieter entscheiden können, ob der erbrachte Dienst seinen*ihren Sicherheitsanforderungen genügt. Wir haben einen Sicherheitsanforderungskatalog erstellt, der für IaaS, PaaS und SaaS eine Zertifizierung nach ISO 27001 ermöglicht. Neben Anforderungen aus Gesetzestexten oder Standards haben individuelle Sicherheitsbedürfnisse der Cloud-Anbieter und der Kund*innen Berücksichtigung gefunden. Zur Dokumentation der Sicherheitsanforderungen werden von ClouDAT Muster bereitgestellt. Konkrete Sicherheitsanforderungen können in diese Muster durch Einsetzen konkreter Elemente eingepasst werden. Die einzusetzenden Elemente und deren Beziehungen untereinander können mit ClouDAT spezifiziert werden.

Cloud-Dienstanbieter wollen oft gegenüber ihren Kund*innen und insbesondere gegenüber Mitbewerbern am Markt ihre Sicherheitsmaßnahmen nicht offenlegen. Auch möchte ein Cloud-Dienstanbieter nicht unbedingt allgemein offenlegen, welche (zertifizierten) Sicherheits-Komponenten verwendet werden. Daher kann nur von einem externen, zur Verschwiegenheit verpflichteten Gutachter geprüft werden, inwieweit die umgesetzten Maßnahmen geeignet sind, die Sicherheitsanforderungen ausreichend zu erfüllen. ClouDAT erlaubt dem Cloud-Dienstanbieter eine erste Validierung bezüglich dieser Fragestellungen selbst durchzuführen.

Für den Zertifizierer werden von ClouDAT Dokumente erzeugt, die die Sicherheitsanforderungen und Maßnahmen beschreiben. Auch werden die Beziehungen der Sicherheitsanforderungen untereinander und auch der Maßnahmen untereinander in nachvollziehbarer Weise dargestellt. Der Zertifizierer kann somit die Validierungen der Anbieter mit geringem Aufwand überprüfen. Die Hauptaufgabe des Zertifizierers ist es, im Rahmen eines Zertifizierungsaudits eines Cloud-Computing-Dienstes zu prüfen, ob die dokumentierten Maßnahmen korrekt realisiert wurden.

ClouDAT hat gegenüber einem informellen Vorgehen folgende Vorteile:

- Es stellt Musterkataloge für Bedrohungen, Sicherheitsanforderungen und eine Liste von möglichen Maßnahmen zur Erfüllung der Sicherheitsanforderungen bereit.
- Es stellt Nachverfolgbarkeits-Verknüpfungen zwischen den verschiedenen Musterkatalogen bereit.
- Es schlägt zum Teil automatisch passende Muster auf Basis der zuvor ausgewählten Muster vor.

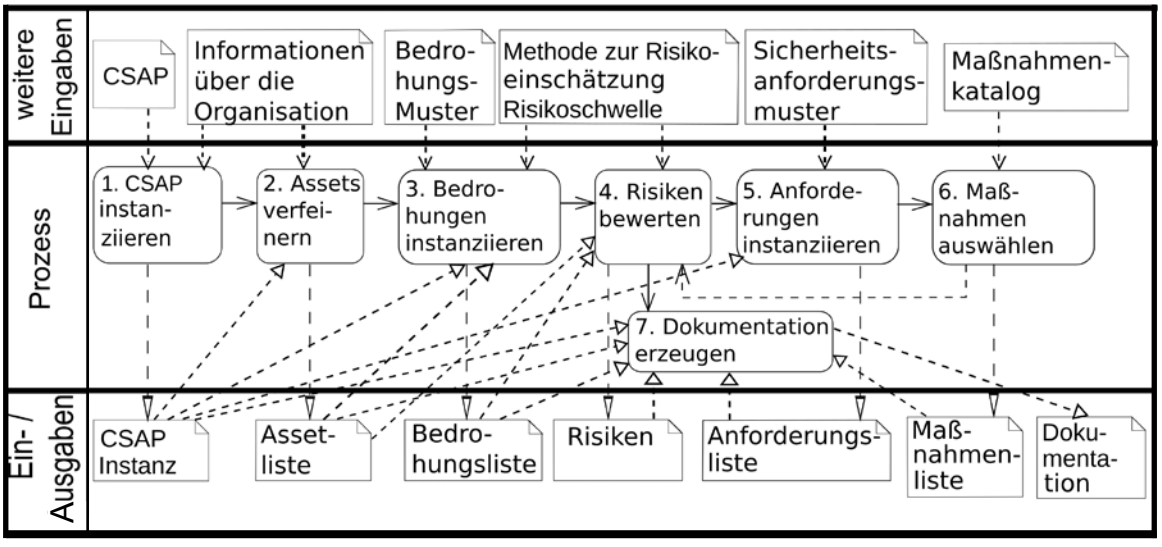
Die von ClouDAT erstellte Dokumentation folgt dem ISO 27001-Standard.

Vorgehensweise zur Erstellung einer geeigneten Dokumentation

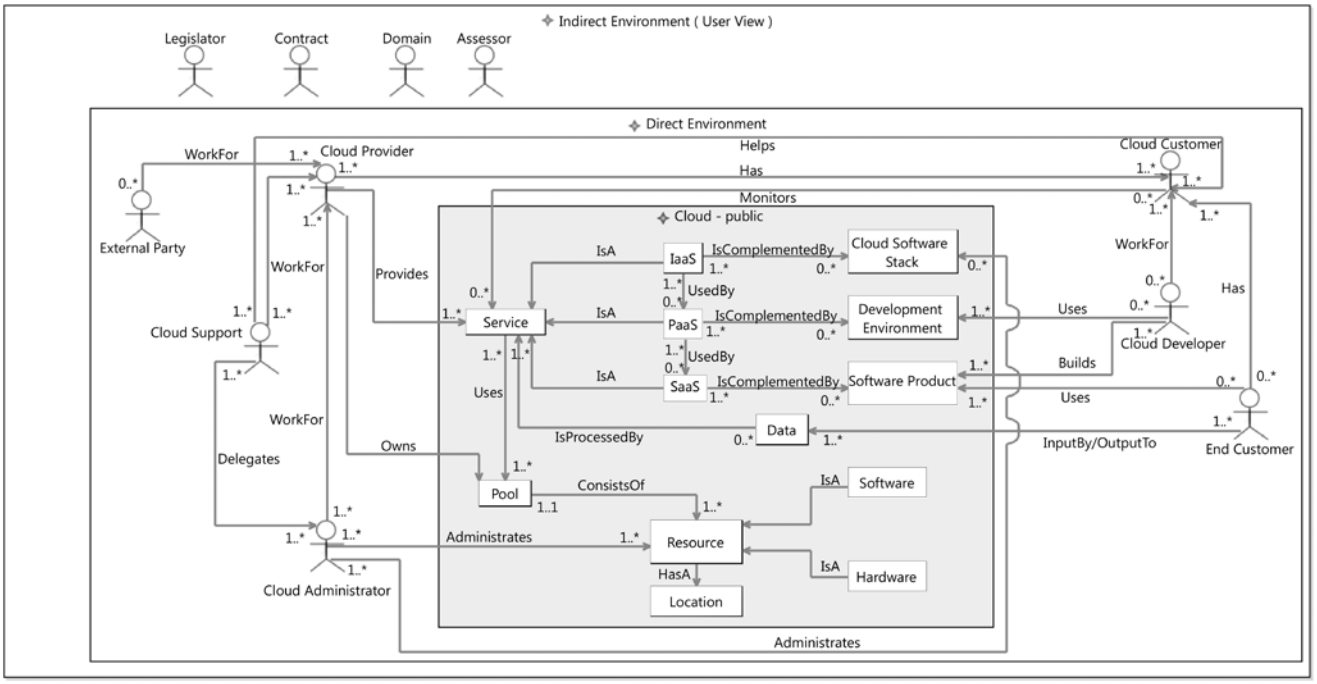
Abbildung (1) gibt einen Überblick über die sieben Schritte der ClouDAT-Methode mit Eingabe- und Ausgabedokumenten, die wir in diesem Abschnitt näher erläutern.

Schritt 1: Instanziierung des CSAP

In diesem Schritt wird der Bereich definiert, für den ein ISMS gelten soll. Dieser Geltungsbereich muss definiert sein, bevor eine Risikoanalyse durchgeführt werden kann. Im Rahmen der ClouDAT-Methode wird dazu das sogenannte Cloud-System-Analysemuster (Cloud System Analysis Pattern, CSAP)[4],



(1) Überblick über die bei ClouDAT angewandte Methode.
Quelle: eigene Darstellung



(2) Das Cloud-System-Analysemuster (CSAP).
Quelle: eigene Darstellung

instanziiert. Im Folgenden diskutieren wir das CSAP, das in Abbildung (2) gezeigt wird.

Das CSAP stellt Elemente zur Verfügung, um einen Cloud-Computing-Dienst zu beschreiben. Hierbei erfolgt eine Beschreibung der direkten Umgebung und indirekten Umgebung des betreffenden Cloud-Computing-Dienstes anhand der Typen von Interessensvertretern (engl. Stakeholders). Zudem wird der Cloud-Computing-Dienst (Cloud) an sich spezifiziert. Hierzu werden verschiedene Typen von Cloud-Elementen verwendet. Wir erklären nun die zuvor genannten CSAP-Elemente und nennen jeweils die Typen der Elemente, die im CSAP definiert sind:

• Indirekte Umgebung (engl. Indirect Environment): Hier erfolgt eine Spezifikation der indirekten Interessensvertreter (engl. Indirect Stakeholder). Diese Interessensvertreter zeichnen sich dadurch aus, dass sie keine direkte Verbindung zum betrachteten Cloud-Computing-Dienst haben. Sie sind allerdings indirekt für den Betrieb des Cloud-Computing-Dienstes relevant. Das CSAP enthält indirekte Interessensvertreter der folgenden Typen

  ◦ Gesetzgeber (engl. Legislator): Gesetzgebungen der Nationen, in der der Cloud-Computing-Dienst betrieben wird.

  ◦ Vertrag (engl. Contract): Vertragliche Vereinbarungen

  ◦ Domäne (engl. Domain): Einschränkungen aus der Anwendungsdomäne, zum Beispiel Basel II/III-Vorschriften für Banken

  ◦ Zertifizierer (engl. Assessor): Unternehmen, das das ISMS des Cloud-Computing-Diensts zertifizieren soll.

• Direkte Umgebung (engl. Direct Environment): Diese Umgebung enthält die Beschreibung der direkten Interessensvertreter, die eine unmittelbare Verbindung zum betreffenden Cloud-Computing-Dienst aufweisen. Die verschiedenen Typen von direkten Interessensvertretern im CSAP spezifizieren Cloud-Dienstanbieter, Cloud-Dienst-Kund*innen (dieser nutzt den Dienst, um Endkunden einen eigenen Dienst anzubieten), Cloud-Dienst-Endkund*innen, Entwickler*innen für Cloud-spezifische Applikationen, Cloud-Administratoren sowie die technische Kund*innenbetreuung (Support) des Cloud-Computing-Dienstes.

• Cloud-Computing-Dienst (Cloud): Hier erfolgt die Definition der verschiedenen Typen von Cloud-Elementen, aus denen sich der betrachtete Cloud-Computing-Dienst zusammensetzt.

Für unseren Beitrag werden wir uns jedoch auf eine Auswahl dieser Typen beschränken: die relevanten Cloud-Dienst-Modelle (IaaS, PaaS, SaaS), den Software-Stack der Cloud-Infrastruktur (Cloud Software-Stack), in der Cloud ausgeführte Softwareprodukte (engl. Software Product), die beteiligten Standorte (Location), die in der Cloud verarbeiteten Daten (Data) sowie die verwendeten Hard- und Software (Hardware, Software) für den Betrieb der Cloud.

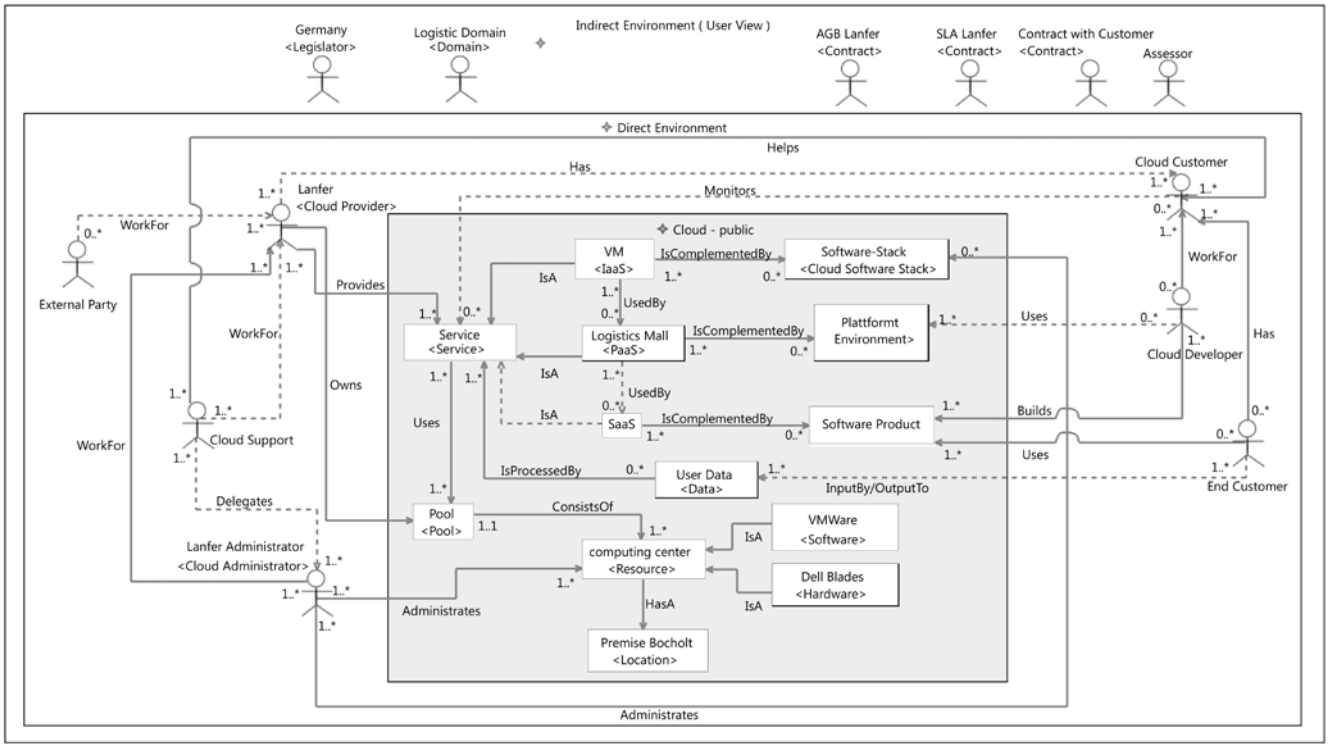
Zusätzlich enthält das CSAP Verbindungen zur Darstellung der logischen Beziehungen zwischen direkten Interessensvertretern und Cloud-Elementen.

Die Namen der verschiedenen CSAP-Elemente (Indirect/Direct

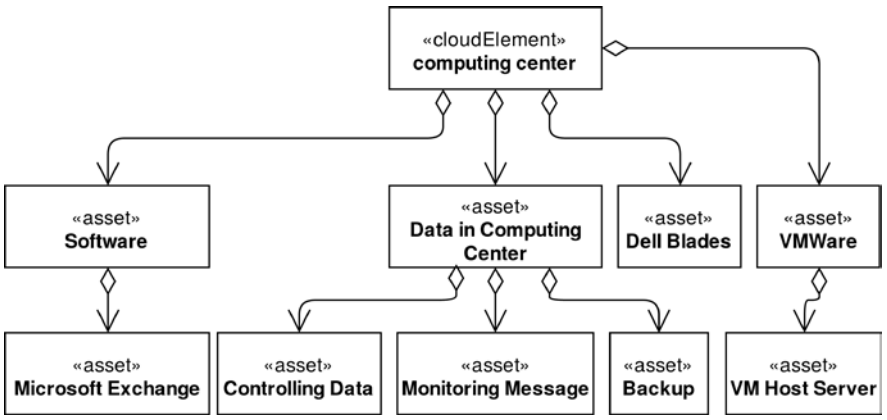
Stakeholder, Cloud-Element) korrespondieren zu ihren sogenannten „Instanztypen“, welche die Art der Elemente beschreiben. In Bezug auf das CSAP enthält dieses beispielsweise einen indirekten Interessensvertreter mit dem Instanztypen „Legislator“ (s. Abb. 2). Bei der Instanziierung des CSAP wird der „Legislator“ durch einen konkreten Namen ersetzt, beispielsweise „Germany“, der den Gesetzgeber der Bundesrepublik Deutschland repräsentiert, siehe Abbildung (3). Die Instanziierung wird für alle relevanten Elemente wiederholt. Sie beginnt zum Beispiel mit der Identifikation der möglichen Kund*innen des Cloud-Computing-Dienstes („Cloud Customer“). Diese nehmen dann die bereitgestellten Cloud-Computing-Dienste in Anspruch, um sie für ihre geschäftlichen Aktivitäten zu nutzen. Der Anbieter der bereitgestellten Cloud-Computing-Dienste muss ebenfalls instanziiert werden. Im Rahmen unserer Fallstudie wird dieser Anbieter durch das Lanfer Systemhaus dargestellt. In dem CSAP werden auch die Gruppen und Regularien betrachtet, die zwar keine direkte Verbindung zu dem Cloud-Computing-Dienst haben, aber trotzdem berücksichtigt werden müssen, wie beispielsweise der Gesetzgeber und andere Regeln (Germany, Dienstleistungsvereinbarungen (engl. Service Level Agreements, SLA), Lanfer, …). Es werden dann auch die Elemente des Cloud-Computing-Dienstes instanziiert (siehe grau hinterlegter Teil in Abb. 3). Diese stellen die vorhandenen Betriebsmittel (z.B. Hardware und Software) und die Dienste, die bereitgestellt werden, dar. Die instanziierten Cloud-Elemente können Assets des Cloud-Dienstanbieters sein. Bezüglich dieser Assets wird die Risikoanalyse ausgeführt, um Sicherheitsrisiken aufzudecken.

Schritt 2: Verfeinerung von Assets

Dieser Schritt bezieht sich auf Assets, die in Form von Cloud-Elementen im CSAP instanziiert wurden. Innerhalb dieses Arbeitsschrittes sind solche Assets relevant, deren Abstraktionsgrad für eine sinnvolle Verwendung innerhalb der Risikoanalyse zu hoch ist. Abbildung (4) zeigt einen Ausschnitt des Verfeinerungsdiagramms aus unserer Fallstudie. Das Rechenzentrum („computing center“) ist ein Asset, dessen Abstraktionsgrad für die Risikoanalyse zu hoch ist. Deshalb ist eine Verfeinerung notwendig: Das Rechenzentrum („computing center“) wird zunächst in die „Software“ und die Daten im Rechenzentrum („Data In Computing Center“) zerlegt. Diese neuen Assets werden weiter zerlegt, da auch hier der Abstraktionsgrad noch zu hoch ist: „Microsoft Exchange“ ist ein Beispiel für „Software“. Daten des Controllings („Controlling Data“), Überwachungsnachrichten („Monitoring Message“) und „Backup“ verfeinern die Daten im Rechenzentrum („Data In Computing Center“). Weitere Verfeinerungen können Abbildung (4) entnommen werden. Wie wir in dieser Abbildung sehen, wurde die „Ist-Bestandteil-von-Beziehung“ benutzt,



(3) Instanz des Cloud-System-Analysemusters für das Lanfer Systemhaus.
Quelle: eigene Darstellung



(4) Beispiel für eine Verfeinerung von Assets.
Quelle: eigene Darstellung

um die Asset-Verfeinerung darzustellen.

Das allgemeine Vorgehen lässt sich wie folgt beschreiben: Betrachte die einzelnen Assets des instanziierten CSAP und nutze gegebenenfalls Prozesse, Richtlinien und Standorte der Organisation, um Assets weiter zu verfeinern.

Für unsere Fallstudie bedeutet das: Um die Assets zu identifizieren, haben wir das instanziierte CSAP als Startpunkt verwendet. Zusätzlich haben wir zur Validierung unseres Ansatzes durch Arbeitsanweisungen, Organigramme, andere Firmendokumente und einen Besuch des Rechenzentrums weitere Assets identifiziert, deren Gesamtanzahl 103 betrug. Unabhängig davon hatten wir dieselben Assets auch durch die Verfeinerungs-Vorgehensweise identifiziert.

Für jedes Asset haben wir (wie auch bei den Cloud-Elementen im CSAP) folgende Informationen dokumentiert:

- Besitzer*in des Assets: Die*Der Besitzer*in ist für das Asset verantwortlich. So muss die*der Besitzer*in beispielsweise die Zugriffsrechte für das Asset kontrollieren.
- Standort des Assets: Hier wird zum Beispiel dokumentiert, dass der „Microsoft Exchange"-Server im „Rechenzentrum 1" steht.
- Art/Typ des Assets: Dokumentation des Typs eines Assets. Dieser Typ entspricht dem Instanztypen von CSAP-Elementen, der bereits in Schritt 1 besprochen wurden. So kann beispielsweise festgehalten werden, ob es sich bei einem Asset um Software oder Daten handelt.
- Relevanz des Assets: Definition, ob ein Asset relevant für die Risikoanalyse ist. Wenn es sich nicht um ein relevantes Asset handelt, wird auch dokumentiert, warum dieses Asset nicht relevant ist. Zum Beispiel wird dokumentiert, dass die Virenschutzsoftware kein Asset an sich ist, sondern zum Schutz der anderen Assets beiträgt.

Die relevanten Assets werden in den folgenden Schritten der Risikoanalyse weiter betrachtet.

Schritt 3: Instanziierung von Bedrohungsmustern mit Schwachstellen

Für alle relevanten Assets aus Schritt 2 wird eine Bedrohungsanalyse durchgeführt. Dabei wird untersucht, ob ein Asset einer Bedrohung ausgesetzt ist, und ob das Asset Schwachstellen besitzt, die durch die bestehenden Bedrohungen ausgenutzt werden könnten. Um diese Analyse zu unterstützen, stellt CloudDAT einen Katalog mit Mustern für Bedrohungen bereit, die auf der Basis des Stands der Technik und der Erfahrung der Autoren erstellt wurden. Der Katalog lässt sich problemlos erweitern. Spezielle Bedrohungen für Cloud-Computing-Dienste sind beispielsweise unsichere Schnittstellen, beziehungsweise Verlust oder ungewolltes Veröffentlichen von Daten.

Die Struktur der Muster für Bedrohungen wird durch ein Modell definiert, bei dem es feste Textpassagen und Platzhalter gibt. Bei den Platzhaltern handelt es sich um die Instanztypen von direkten und indirekten Interessensvertretern und Cloud-Elementen. Die Platzhalter werden durch den Namen des relevanten Elementes aus dem instanziierten CSAP oder der Verfeinerungsdiagramme ersetzt. Der Katalog ist nach den Kategorien Vertraulichkeit, Integrität und Verfügbarkeit sortiert. Für jedes Asset sollten alle Kategorien betrachtet werden. Weitere Kategorien, wie zum Beispiel „Privacy" können ergänzt werden. Ein Beispiel für ein Muster für eine Bedrohung ist:

„Unavailability of [cloud element] for [all end customers and cloud customers]."
(Nicht-Verfügbarkeit eines [Cloud Elementes] für [alle Endbenutzer und Cloudnutzer]).

Um dieses Muster zu instanziieren, muss der Name aus der CSAP-Instanz eingesetzt werden:

„Unavailability of Controlling Data for Lanfer Systemhaus."
(Nicht-Verfügbarkeit der Daten des Controllings für das Lanfer Systemhaus).

Die Verwendung unseres Kataloges reduziert das Risiko, dass wesentliche Bedrohungen in der Analyse nicht betrachtet werden. Ergänzend geben wir Beispiele für Schwachstellen an, die zum Ausnutzen dieser Bedrohung führen können.

Die Identifikation der Bedrohungen und Schwachstellen unterstützt die Einschätzung des Bedrohungspotentials, das wir durch das Interesse des Angreifers und die Größe der Schwachstellen (vulnerability level) beschreiben. Das Inte-

resse des Angreifers klassifizieren wir in LOW, MEDIUM und HIGH als Standardwerte.

Die Schwachstellen beschreiben wir mit L, wenn nahezu alle Schwachstellen mit entsprechenden Maßnahmen adressiert sind, mit M, wenn ein mittlerer Schutz gegeben ist, und mit H, wenn keine geeigneten Maßnahmen existieren, um die Bedrohung abzuwehren.

Aus diesen Informationen werden wir in Schritt 4 das Bedrohungspotential ermitteln. Dieses bestimmt zusammen mit dem Wert des Assets das Risiko, dem das Asset ausgesetzt ist.

Schritt 4a: Risiken bewerten

Wir benutzen eine Risikoschwelle, um zu entscheiden, ob ein Asset im Weiteren behandelt werden muss. Die ISO 27001:2013 verlangt, dass eine angemessene Methode für die Risikoeinschätzung gewählt wird, die vergleichbare und reproduzierbare Ergebnisse liefert, und dass ein Risikowert definiert wird, unterhalb dessen das Risiko akzeptiert wird.

Das Risiko setzt sich aus dem Bedrohungspotential und dem Wert des Asset zusammen. Wir ermitteln es durch Multiplikation zweier numerischer Werte, die diese beiden Faktoren repräsentieren. Dazu definieren wir eine Skala von Werten, anhand derer der Einfluss von Assets auf das Unternehmen untersucht werden kann. Dieser Einfluss bezieht sich darauf, dass die Sicherheitsziele (z.B. Vertraulichkeit, Integrität und Verfügbarkeit) eines Assets im Zuge eines Sicherheitsvorfalls kompromittiert würden. Im Rahmen von CloudAT wird eine Werte-Skala definiert, in der sich der jeweilige Wert für ein Asset aus möglichen Konsequenzen für das Unternehmen ergibt. Der Wert eines Assets ist auf

- 1 zu setzen, wenn keine Konsequenzen aus einem Sicherheitsvorfall erwartet werden müssen,
- 2 zu setzen, wenn mit den Konsequenzen einfach umgegangen werden kann,
- 3 zu setzen, wenn ein moderater Aufwand erforderlich ist, um mit den Konsequenzen umzugehen,
- 4 zu setzen, wenn ein hoher Aufwand erforderlich ist, um mit den Konsequenzen umzugehen und
- 5 zu setzen, wenn das Überleben des Unternehmens durch einen Sicherheitsvorfall gefährdet wäre.

Eine andere Definition der Werte ist einfach möglich.

Das Bedrohungspotential wird folgendermaßen aus den Einschätzungen ermittelt, die wir in Schritt 3 vorgenommen hatten:

- 1 bei L (Interesse des Angreifers), LOW (Schwachstellen),
- 2 bei L, MEDIUM oder M, LOW,
- 3 bei H, LOW oder L, HIGH oder M, MEDIUM,
- 4 bei M, HIGH oder H, MEDIUM
- 5 bei H, HIGH

Nun kann der Risikowert durch eine Multiplikation des Werts des Assets und des Bedrohungspotentials ermittelt werden.

Für unser Asset „Controlling Data“ kamen wir damit zum Beispiel in Bezug auf die Vertraulichkeit auf einen Risikowert von 12 und für unser Asset „VM Host Server“ in Bezug auf die Integrität auf einen Risikowert von 15.

Schritt 4b: Risiko überprüfen

Nachdem das Risiko für alle Assets systematisch abgeschätzt wurde, muss gemäß ISO 27001 geprüft werden, ob der Risikowert ein akzeptables Risiko darstellt. Als Risikoschwelle haben wir den Wert 10 gewählt. Einen Risikowert kleiner oder gleich 10 akzeptieren wir aus den folgenden Gründen:

- Wenn wir keine Konsequenzen erwarten, um mit den Auswirkungen eines erfolgreichen Angriffs umzugehen, oder mit ihnen einfach umgegangen werden kann (Wert des Assets = 1 oder 2), können wir das Risiko akzeptieren, auch wenn Angreifer ein großes Interesse haben und keine Maßnahmen gegen die Schwachstellen vorhanden sind (Bedrohungspotential = 5).
- Wenn wir bei einem erfolgreichen Angriff mittelmäßigen Aufwand erwarten (Wert des Assets = 3), können wir das Risiko akzeptieren, wenn das Bedrohungspotential den Wert 3 hat.
- Wenn durch die Auswirkungen eines erfolgreichen Angriffs das Überleben des Unternehmens gefährdet oder ein hoher Aufwand erforderlich wäre, um damit umzugehen (Wert des Assets = 4 oder 5), können wir das Risiko akzeptieren, wenn nahezu alle Schwachstellen mit entsprechenden Maßnahmen adressiert sind oder bei einem mittleren Schutz von einem Angreifer mit nur mittlerem Interesse ausgegangen werden kann (Bedrohungspotential = 1 oder 2).

Die Vertraulichkeit unseres Assets „Controlling Data“ ist mit dem Risikowert von 12 bewertet und die Integrität unseres Assets „VM Host Server“ ist mit dem Risikowert von 15 bewertet. Für die weiteren hier erwähnten Assets kann bei den gegebenen Maßnahmen das Risiko akzeptiert werden.

Um mit einem Risiko umzugehen, das nicht akzeptabel ist, definiert die ISO 27001:2013 folgende Alternativen:

1. geeignete Maßnahmen anwenden, um das Risiko auf ein akzeptables Maß zu reduzieren,
2. das Risiko explizit akzeptieren,
3. das Risiko vermeiden, oder
4. das Risiko an andere abgeben.

Wir legen unseren Schwerpunkt auf Alternative 1 (Maßnahmen anwenden). Immer, wenn diese Alternative gewählt wird, fahren wir mit Schritt 5 fort. Wenn keine weitere Bewertung einen Risikowert über dem Schwellwert hat beziehungsweise wir eine andere Alternative gewählt haben, können wir zu Schritt 7 gehen.

Schritt 5: Sicherheitsanforderungen instanziieren

In diesem Schritt werden alle Assets betrachtet, bei denen wir ein nicht-akzeptables Risiko identifiziert haben

und für die wir geeignete Maßnahmen durchführen wollen, um das Risiko zu reduzieren. Um geeignete Maßnahmen zu finden, müssen wir Sicherheitsanforderungen definieren. Dazu nutzen wir sogenannte Sicherheitsanforderungsmuster. Diese Muster sind nach dem gleichen Prinzip aufgebaut, wie die Muster für Bedrohungen. Sie enthalten neben fixem Text auch Platzhalter, die den Instanztypen der relevanten CSAP-Elementen entsprechen. Zudem haben wir eine Zuordnung von den Mustern für Bedrohungen zu den Mustern für Sicherheitsanforderungen erstellt. So schlägt unser Werkzeug für jede Instanz eines Bedrohungsmusters die zugeordneten Sicherheitsanforderungsmuster vor, die ebenfalls instanziiert werden. Für die Bedrohung

„Disclosure of stored controlling data of Lanfer Systemhaus by an attacker“,
(Offenlegung von gespeicherten Daten für das Controlling des Lanfer Systemhauses durch einen Angreifer)

ist das entsprechende Muster für eine Sicherheitsanforderung:

„Preserve confidentiality of stored [data] of [cloud customer, end customer] by preventing disclosure by [cloud provider, phone support, cloud administrator, cloud developer, third-party provider, other cloud customer, other end customer, external attacker]“
(Stelle die Vertraulichkeit der gespeicherten [Daten] des [Cloud-Kunden, Endkunden] sicher, indem die Offenlegung durch [Cloud-Anbieter, Hotline, Cloud-Administratoren, Cloud-Entwickler, ausgelagerte Anbieter, andere Cloud-Kunden, andere End-Kunden, externe Angreifer] verhindert wird.)

Dieses Muster kann ausgewählt und wie folgt instanziiert werden (bei der Instanziierung kann die Instanziierung des Bedrohungs-Musters zur Hilfe genommen werden):

SR1 “Preserve confidentiality of stored controlling data of Lanfer Systemhaus by preventing disclosure by an external attacker.“
(Stelle die Vertraulichkeit der gespeicherten Daten für das Controlling des Lanfer Systemhauses sicher, indem die Offenlegung durch externe Angreifer verhindert wird.)

Mit Hilfe der Zuordnung der Muster für Bedrohungen zu den Mustern für Sicherheitsanforderungen haben wir für die Fallstudie Sicherheitsanforderungen für Assets ausgewählt, deren Risiken nicht akzeptiert werden können. Zusätzlich zu SR1 ergaben sich die folgenden Sicherheitsanforderungen:
• Forderung von Integrität in Bezug auf die Kommunikation zwischen den Virtuellen Maschinen und den Mitarbeitern (SR 2),
• Verhinderung von Manipulationen des „VM Host Server“, die zur Nicht-Verfügbarkeit führen würden (SR 3),
• Ausreichender physischer Schutz (keine Fenster im Erdgeschoss, Zutrittskontrolle und beschränkter Zugang durch Besucher*innen, …) (SR 4) und
• Keine Beeinflussung der Verfügbarkeit der bereitgestellten Plattform durch technische Fehler des „VM Host Server“ (SR 5).

Der Vorteil von explizit formulierten Sicherheitsanforderungen besteht darin, dass geprüft werden kann, ob diese erfüllt sind. Wenn alle relevanten Sicherheitsanforderungen für einen Cloud-Computing-Dienst erfüllt sind, ist das erreichte Sicherheitsniveau ausreichend.

Schritt 6: Maßnahmen auswählen

In diesem Schritt werden Maßnahmen ausgewählt, die für die in Schritt 5 identifizierten Sicherheitsanforderungen angemessen sind. Mögliche Maßnahmen haben wir in einem Katalog zusammengefasst und mit den entsprechenden Sicherheitsanforderungen verknüpft, so dass eine Vorauswahl der möglichen Maßnahmen werkzeuggestützt erfolgen kann. Maßnahmen können voneinander abhängig sein. Wenn zum Beispiel die Vertraulichkeit von Daten per Verschlüsselung sichergestellt werden soll, müssen die verwendeten Schlüssel ebenfalls vertraulich sein. Unter Nutzung der Zuordnungen von Sicherheitsanforderungen zu Maßnahmen und der Abhängigkeiten zwischen Maßnahmen haben wir für die genannten Sicherheitsanforderungsmuster die folgenden Maßnahmen aus der ISO 27001:2013 ausgewählt und instanziiert:
• SR 1: zum Beispiel Sicherheit der Geräte, Zugriffskontrolle, und (entsprechend unserer Abhängigkeiten) Sicherheit bezüglich des Personals
• SR 2: zum Beispiel Kryptografische Signaturen und die notwendigen Maßnahmen
• SR 3: Für SR 3 werden die gleichen Maßnahmen, wie für SR 1 gewählt. Zusätzlich werden sichere Bereiche realisiert.
• SR 4: Die Maßnahmen, die SR 4 adressieren, sind bereits zur Behandlung der anderen Sicherheitsanforderungen ausgewählt.
• SR 5: zum Beispiel Maßnahmen für redundante Systeme

Der Katalog wurde auf Basis der ISO 27001, der Cloud Security Alliance Cloud Controls Matrix (Version 3.0.1 von 2014) und eigenen Erfahrungen erstellt. Es ist möglich, ihn zu ergänzen.

Nachdem die Maßnahmen gewählt worden sind, muss geprüft werden, ob mit den Maßnahmen das Risiko für alle Assets auf ein akzeptables Maß reduziert wurde. Dies geschieht durch eine Wiederholung von Schritt 4.

Schritt 7: Dokumentation erzeugen

Zuletzt wird von unserem Werkzeug eine Dokumentation erzeugt. Folgende im Laufe der Anwendung unserer Methode erzeugten Artefakte gehen in die Erzeugung der Dokumentation ein:
Schritt 1: Instanz des CSAP
Schritt 2: Liste der identifizierten Assets
Schritt 3: Liste der Bedrohungen
Schritt 4: Liste der identifizierten Risiken

Schritt 5: Liste der Sicherheitsanforderungen
Schritt 6: Liste der zu realisierenden Maßnahmen

Alle oben genannten Ergebnisse werden in einem Dokument zusammengefasst. Dieses Dokument kann als Grundlage für eine Zertifizierung des ISMS gemäß ISO 27001 dienen.

ClouDA-Tool

ClouDAT ist ein speziell angepasstes Werkzeug für die Analyse und Dokumentation von Sicherheitsanforderungen und Maßnahmen in Bezug auf Cloud-Computing-Dienste. Im Gegensatz zu den sonst oft üblichen Dokumentationen mit Textverarbeitungssystemen und Zeichenprogrammen, können mit Hilfe des ClouDA-Tools logische Zusammenhänge zwischen den erfassten Informationen dargestellt werden. Damit werden interne Konsistenzprüfungen möglich, die zur Verbesserung der Qualität der Analyse beitragen. Durch die standardisierte Darstellung wird eine Prüfung deutlich vereinfacht. Fehler in der internen Konsistenz der Dokumentation sind der größte Kostenfaktor bei einer Zertifizierung. Daher prüft das Werkzeug die interne Konsistenz der Dokumentation, statt auf eine fehleranfällige Prüfung beim Cloud-Dienstanbieter zu setzen. Weiterhin wird durch die Anwendung der ClouDAT-Methode der Arbeitsfluss während der Risikoanalyse verbessert. So werden bei der Instanziierung der Bedrohungs- und Sicherheitsanforderungsmuster für die Ersetzung der Platzhalter potentiell relevante Informationen auf Grundlage der identifizierten Assets automatisch durch das ClouDA-Tool angeboten. Die werkzeuginterne Abbildung der Beziehungen zwischen Assets, Bedrohungsmustern, Sicherheitsanforderungsmustern und Maßnahmen bewirkt, neben der Beschleunigung des Arbeitsflusses, dass diese nicht übersehen werden können. Wir haben verschiedene Designer-Editoren implementiert. Mit diesen können
- das Cloud-System-Analysemuster
- die Kataloge für Bedrohungsmuster, Sicherheitsanforderungsmuster und Maßnahmen sowie
- die Parameter für die Risikoanalyse

modifiziert und/oder erweitert werden. So wird eine hohe Anpassbarkeit und Wiederverwendbarkeit erreicht.

Unser Werkzeug wurde unter der GNU Public License (GPL) veröffentlicht.

Zusammenfassung

Wir haben eine strukturierte Methode zur Risikoanalyse von Cloud-Anwendungen entsprechend dem ISO 27001-Standard vorgestellt. Unsere Methode basiert auf Mustern, um den Kontext und die Struktur eines Cloud-Computing-Dienstes zu beschreiben (CSAP), Bedrohungen zu identifizieren, Sicherheitsanforderungen auszuwählen, sowie entsprechende Maßnahmen zu ermitteln. Unser Werkzeug unterstützt die Anwendung dieser Methode. Sie hat folgende Vorteile:
- Systematische Identifizierung von Bedrohungen durch entsprechende Bedrohungsmuster (engl. Threat Patterns) und deren Beziehung zu Elementen des CSAP, um die Sicherheit von Cloud-Computing-Diensten zu analysieren. Dies erleichtert und beschleunigt die Bedrohungsanalyse.
- Systematische Identifizierung von Sicherheitsanforderungen durch entsprechende Sicherheitsanforderungsmuster und deren Beziehung zu Bedrohungsmustern.
- Systematische Identifizierung von Maßnahmen durch deren Beziehung zu Sicherheitsanforderungsmustern.
- Zusätzliche Effizienz durch das Anwenden der Methode und die Reduzierung des Aufwandes durch die hierarchische Verfeinerung von Assets.

Wir werden in Zukunft unser Werkzeug um weitere Muster erweitern, um damit eine Risikoanalyse für andere Systeme als Clouds durchführen zu können. Zusätzlich planen wir, das Werkzeug so zu erweitern, dass auch die vollständige und kohärente Instanziierung der Muster überprüft werden kann.

Summary

We present a structured method for performing risk analysis for cloud applications according to the ISO 27001 standard. Our method relies on patterns to describe the context and structure of a cloud computing system (using CSAP), to identify threats, to elicit the security requirements, and to select controls. Our ClouDA tool supports the application of this method. Our approach delivers the following main benefits:
- Systematic pattern-based identification of threats using threat patterns and their relationship to CSAP elements, which facilitates and accelerates the threat analysis
- Systematic pattern-based identification of security requirements to be fulfilled by appropriate controls using security requirement patterns and their relationship to threat patterns
- Systematic pattern-based identification of controls using their relationship to security requirement patterns
- Tool support for our approach
- Increased effectiveness of risk analysis by applying the method and reduced documentation effort by hierarchical refinement of assets.

In the future, we want to extend the tool for supporting other types of patterns for performing risk analysis. In addition, we intend to enrich the tool so as to check the complete and coherent instantiation of the patterns.

Danksagung

Wir danken Kristian Beckers für seine Ideen und die Diskussionen mit ihm. Weiterhin danken wir Thomas Santen für seine Kommentare zu diesem Papier.

Anmerkungen

1) entsprechend der vom US National Institute of Standards and Technology (NIST) gegebenen Definition von Cloud Computing 2009, S. 2
2) Eckpunktepapier - Sicherheitsempfehlungen für Cloud Computing Anbieter 2012, S. 25
3)http://www.cloudat.de
4) Beckers et al., 2011

Literatur

– Alebrahim, A., Faßbender, S., Hatebur, D., Goeke, L. & Côté, I. (2015). A Pattern-Based and Tool-Supported Risk Analysis Method Compliant to ISO 27001 for Cloud Systems. In International Journal of Secure Software Engineering (IJSSE)
– K. Beckers, H. Schmidt, J. C. Kuster and S. Faßbender (2011). Pattern-Based Support for Context Establishment and Asset Identification of the ISO 27000 in the Field of Cloud Computing. In Sixth International Conference on Availability, Reliability and Security, pp. 327-333.
– Beckers, K., Côté, I., Goeke, L. & Güler, S., Heisel, M. (2014). A Structured Method for Security Requirements Elicitation concerning the Cloud Computing Domain. In International Journal of Secure Software Engineering (IJSSE)
– ISO/IEC. (2013). International Standard ISO/ICE 27001 Information technology Security techniques Information security management systems Requirements. Second edition.

*Die Autor*innen*

Maritta Heisel ist Professorin für Software Engineering an der Universität Duisburg-Essen und Vorstandsmitglied von paluno – The Ruhr Institute for Software Technology. Sie studierte Informatik an der Universität Karlsruhe, wo sie auch promovierte. Anschließend habilitierte sie sich an der TU Berlin. Es folgten eine Position als Hochschuldozentin an der Universität Madgeburg sowie eine Professur für Praktische Informatik an der Universität Münster. Seit 2004 ist sie Professorin an der Universität Duisburg-Essen. Ihre Forschungsinteressen liegen auf den Gebieten Informations- und Datensicherheit, Anforderungsanalyse, Spezifikation und Entwicklung sicherheitskritischer Software sowie methodischen Aspekten der Softwareentwicklung. Weiterhin ist sie Mitglied in dem seit 2015 von der DFG geförderten Graduiertenkolleg „User-Centred Social Media“.

Isabelle Côté schloss ihr Studium der Informatik 2004 mit Diplom ab und promovierte 2012 an der Universität Duisburg-Essen. Seit 2012 ist sie bei ITESYS Institut für technische Systeme GmbH mit Sitz in Dortmund angestellt. Sie ist für Industrieprojekte im Bereich Automotiv-Safety verantwortlich. Darüber hinaus arbeitet sie auch an Projekten mit dem Schwerpunkt Informationssicherheit mit. Im Rahmen des Forschungsprojektes „ClouDAT“ beschäftigte sie sich mit Informationssicherheit in Cloud-Computing-Systemen. Ihre Forschungsinteressen liegen in der Safety und Security. In diesen Bereichen hat sie verschiedene Konferenz-, Fachzeitschriften- und Workshopbeiträge verfasst und begutachtet.

Denis Hatebur arbeitet an der Universität Duisburg-Essen als wissenschaftlicher Mitarbeiter. Gleichzeitig ist er seit 2003 Geschäftsführer der ITESYS Institut für technische Systeme GmbH in Dortmund. Er arbeitete in verschiedenen Industrieprojekten als Berater. In diesen Safety- und Securityprojekten war er für den Spezifikationen und Testen verantwortlich. Seine Forschungsinteressen liegen in der Entwicklung sicherer und verlässlicher Systeme unter Berücksichtigung des Anforderungs-Engineerings, der Architekturentwicklung und des Testens. In seinem Bereich hat er verschiedene Konferenz-, Fachzeitschriften- und Workshoppapiere verfasst und begutachtet. Er hat ein Diplom der technischen Informatik der Fachhochschule Dortmund, einen Master of Science im Bereich Computer Engineering der Universität Duisburg-Essen und seit 2012 einen Dr.-Ing. der Universität Duisburg-Essen.

Ludger Goeke studierte Allgemeine Informatik an der Fachhochschule Dortmund und schloss 2007 mit Diplom ab. Er ist seit 2007 Angestellter bei der ITESYS Institut für technische Systeme GmbH mit Sitz in Dortmund. In diesem Zusammenhang war er in verschiedenen Industrieprojekten als Berater tätig. Zu Beginn seiner beruflichen Laufbahn bezogen sich die entsprechenden Projekttätigkeiten auf die Themengebiete Änderungs-, Dokumenten- und Prozessmanagement. Seit geraumer Zeit ist er vorrangig an Projekten mit der Thematik Informationssicherheit beteiligt. Hierbei war er für das Testen und die Quellcodeanalyse in Bezug auf Smartcard-Applikationen verantwortlich. Weiterhin befasste er sich im Rahmen des Forschungsprojektes „ClouDAT“ mit der Sicherstellung der Informationssicherheit von Cloud-Computing-Diensten. Seine Forschungsinteressen liegen im Risikomanagement sowie der Systemanalyse zur Sicherstellung von Informationssicherheit. Im Bereich der Informationssicherheit war er an der Erstellung verschiedener Fachzeitschriften- und Workshoppapiere beteiligt.

Dieser Artikel erklärt, wie mit Modellen, Hypothesen, Computern und Programmen Forschungsfragen beantwortet und neue gestellt werden (können).

Molekulardynamik auf Hochleistungsrechnern

Selbstorganisation und dynamische Strukturen in Flüssigkeiten

Von Jure Gujt & Eckhard Spohr

Eine immer wieder gestellte Frage an Theoretiker lautet: „Wozu benötigt man eigentlich so riesige Rechner in der Chemie (oder Biologie, oder Physik, oder ...), wenn doch nur das Experiment die letztendliche Antwort darüber geben kann, wie sich die Natur verhält?“

Die Antwort auf diese Frage kann in vielfältiger Weise gegeben werden, wofür hier eigentlich nicht der Raum ist. Es sollen nur kurz drei Punkte angerissen werden. Erstens ist es so, dass in den Naturwissenschaften, und nicht nur da, trotz oder gerade wegen der heutzutage immer rascheren Generierung von neuen Daten und der Erschließung von neuen Anwendungsfeldern für ‚Big Data‘ das Verständnis und die Interpretation dieser Datenmengen zunehmend wichtiger wird. Dabei fungieren Computer als Datenanalyse-Werkzeuge einerseits und als alternative Datengenerierungswerkzeuge in Computersimulationen andererseits. Zum zweiten gibt es viele Beispiele dafür, dass Daten eben nicht im Labor erzeugbar sind, zum Beispiel solche unter sehr extremen Umgebungsbedingungen. Dann muss man auf Simulationsrechnungen ausweichen, die aber nur wirklich aussagekräftig sind, wenn man vorher in analoger Weise das Verhalten der Substanzen und Systeme unter alltäglicheren Bedingungen studiert und mit dem Experiment verglichen hat. Die Etablierung geeigneter Modelle und Rechenmethoden, die diesen Simulationen zugrunde liegen, ist ein formidables Unterfangen und hat die Naturwissenschaften in den letzten Jahrzehnten von einem Gedankenge-

Jure Gujt. Foto: Vladimir Unkovic

bäude, in dem Theorie und Experiment sich komplementär gegenüberstehen, hin zum Paradigma eines Dreiklangs aus Experiment, analytischer Theorie („mit Papier und Bleistift") und Computersimulation („Computational Science") transformiert. Darin haben Computersimulationen einen festen Platz gefunden, der in seiner Bedeutung ständig zunimmt. Drittens, und nicht zuletzt, machen Computersimulationen Spaß, da man wie in einem Brennglas sehr direkt die Beziehung zwischen Ursache und Wirkung beobachten kann. Manche Wissenschaftler*innen nennen dies auch das „virtuelle Labor".[1]

Was wäre, wenn...?

Chemische Prozesse laufen bevorzugt in Lösung ab, da zum einen im Gegensatz zur festen Phase die Beweglichkeit von Molekülen in Flüssigkeiten sehr hoch ist und chemische Reaktionen rasch ablaufen können, und da zum anderen durch die Wahl des Lösungsmittels auch spezifisch gewünschte Wechselwirkungen zwischen den Reaktanden erreicht werden können. Das Verständnis solcher Steuerungsmechanismen auf der atomaren Ebene ist auch Gegenstand aktueller Forschung, wie sie zum Beispiel im DFG-Exzellenzcluster RESOLV gebündelt ist[2], an dem unsere Arbeitsgruppe beteiligt ist. In einem solchen Verbund ergänzen und stützen theoretische Forschungsprojekte das Experiment, sei es durch Interpretation von Daten, sei es durch Modellrechnungen im Sinne eines „was wäre, wenn..?".

Im Gegensatz zu Feststoffen, deren Eigenschaften sich ganz wesentlich aus ihrer Regelmäßigkeit und Ordnung heraus erschließen, und im Gegensatz zu Gasen, deren Eigenschaften bei relativ geringer Dichte ganz gut aus der Überlegung abgeleitet werden können, dass nur zwei oder drei unterschiedliche Moleküle (in einem ‚Cluster') miteinander wechselwirken und sie ansonsten weitgehend unabhängig voneinander sind, ist die Situation in einer Flüssigkeit viel komplizierter. Aus der Schule ist den meisten noch die Definition der Aggregatzustände fest, flüssig und gasförmig über die Nah- und Fernordnung der Atome und Moleküle geläufig. Die fehlende Fernordnung in einer Flüssigkeit einerseits und die große Zahl von dicht gepackten Molekülen andererseits, die die Nahordnung beeinflussen, legen es nahe, dass man viele unterschiedliche „Konfigurationen" (also Anordnungen von Molekülen in einem chemischen System) separat betrachten und darüber in statistischer Weise mitteln muss, um ihre Eigenschaften zu verstehen. Dieses statistische Verfahren bedeutet einen erheblichen Rechenaufwand, für den große Rechner wie die CRAY-XT6m[3] und die magnitUDE[4] extrem nützlich sind, die vom Zentrum für Informations- und Mediendienste (ZIM)[5] und vom Center for Computational Sciences and Simulation (CCSS)[6] gemeinsam an der Universität Duisburg-Essen betrieben werden.

Statistische Thermodynamik

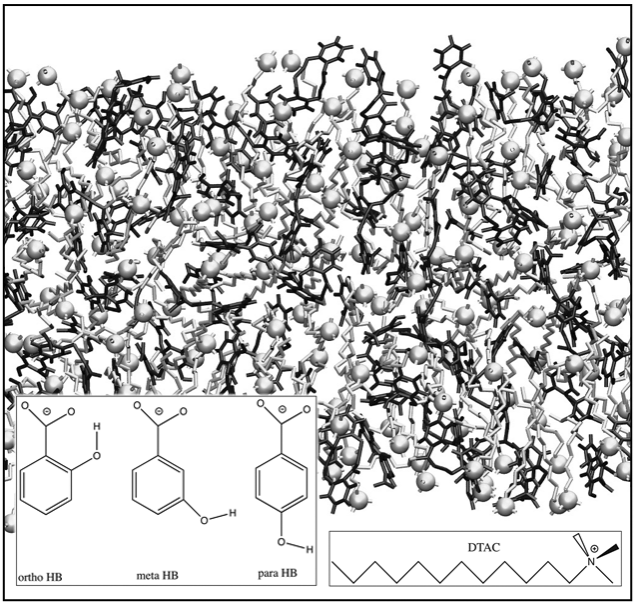
Die wissenschaftliche Disziplin, die sich mit dieser Art der Mittelung über Konfigurationen von Atomen und Molekülen in für die Forschung interessanten chemischen (oder biologischen, oder physikalischen) Systemen befasst, ist die statistische Thermodynamik. Dabei geht es, ganz einfach gesprochen, darum, dass man aus der statistischen Beschreibung von ‚vielen' Atomen und Molekülen im Computer die Eigenschaften eines realen Systems, das noch ‚viel mehr' Atome und Moleküle enthält, vorhersagen beziehungsweise berechnen kann. Offensichtlich ist hier wichtig, was man genau unter ‚viel' und unter ‚viel mehr' versteht. Typische Systeme, die man im Computer realisieren kann, enthalten zwischen einigen Hundert und einigen Millionen Atome, manchmal auch Milliarden. Das ist aber im Vergleich zu realen Systemen immer noch sehr wenig. So sind zum Beispiel in einem Liter Wasser etwa 10^{26} Atome, also 100 Millionen mal eine Milliarde Milliarden Atome enthalten. Das stellt Forscher*innen bei der Computersimulation dann vor die schwierige Aufgabe, wie sie einerseits mit den vorhandenen Computerressourcen ökonomisch umgehen können, und wie andererseits überhaupt abgeschätzt werden kann, ob und, wenn ja, wo eine Nutzung der Ressource Rechenzeit sinnvoll ist, da man vielleicht nie die minimal notwendige Systemgröße oder Beobachtungszeit für die interessanten Prozesse erreichen kann.

Im Verlauf der Erläuterung dieser Problematik wird auch klar werden, wieso *ein einzelner* Rechner nicht ausreicht, sondern wieso man eine ganze Palette unterschiedlich leistungsfähiger Maschinen benötigt, um zum Ziel zu kommen. Da die einzelnen Arbeitsgruppen lediglich kleinere Systeme betreiben können, ist die Zusammenarbeit vieler Forschergruppen zur Bündelung der Anforderungen an einer Universität notwendig. Dies wird in der Kooperation zwischen dem CCSS und dem ZIM seit etlichen Jahren erfolgreich realisiert. Die Kooperation hat seit 2010 zunächst zur Anschaffung des CRAY-Xt6m-Computers und 2016 zur Anschaffung der NEC-MagnitUDE geführt. Beide großen Systeme wurden in der Vergangenheit bis heute (CRAY-Xt6m) beziehungsweise werden in der Gegenwart und Zukunft (NEC-magnitUDE) zur Lösung vielfältiger wissenschaftlicher Probleme an der UDE genutzt. Sie sind aber auch die Grundlage für die Nutzung noch größerer Ressourcen in nationalen und internationalen Rechenzentren, da man für die erfolgreiche Antragstellung an diesen Zentren auch nachweisen muss, wie gut die eigenen Programme und Anwendungen *skalieren*, was, salopp formuliert,

bedeutet, wie gut man Rechnungen auf vielen Rechenkernen oder -prozessoren parallelisieren kann, um schneller zu Resultaten zu gelangen.

Flüssigkeiten und Lösungen

Wir wollen die Vorgehensweise der Nutzung von Rechenanlagen für Computersimulationen am Beispiel der Micellbildung aus amphiphilen Molekülen in wässriger Lösung erläutern. Micellen sind Aggregate von meist kettenförmigen Molekülen, die oft ein geladenes Ende haben, das sich gut in Wasser lösen lässt (man sagt auch, dass die Moleküle mit Wasser „stark wechselwirken"). Chemisch damit verknüpft ist im gleichen Molekül eine mehr oder weniger lange Kette von (meistens) Kohlenstoffatomen, die eher schwach mit Wasser, dafür aber stärker miteinander wechselwirken. Das Resultat ist häufig ein sich schon bei sehr niedrigen Molekülkonzentrationen bildendes mehr oder weniger kugelförmiges Aggregat, bei dem sich außen die gut löslichen geladenen Gruppen und innen die schlechter löslichen Kohlenstoffketten befinden (s. Abb. 1). Dies ist ein Beispiel für das aus der Schule noch bekannte chemische Prinzip, wonach gilt: „Ähnliches löst Ähnliches" (*similis similibus solvuntur*). Die Aggregate sind über relativ lange Zeiträume (auf der Skala der Bewegung der einzelnen Atome) stabil, aber auf der Zeitskala von Sekunden oder Minuten verändern sie ständig ihre Form und zerfallen beziehungsweise bilden sich neu. Ähnliche Prinzipien werden in lebenden Organismen zur räumlichen und zeitlichen Trennung von Funktionalität genutzt. Man kann micellare Systeme prinzipiell auch, da sie ja gut in wässriger Lösung löslich sind, zum Transport von nicht gut wasserlöslichen Molekülen wie zum Beispiel manchen pharmazeutischen Wirkstoffe verwenden, indem man ein solches Molekül im unpolaren Kern der Micelle versteckt.



(1) Schnappschuss einer durch ortho-HB-Ionen stabilisierten zylindrischen Micelle. Etwa 14.000 Wassermoleküle wurden bei der graphischen Darstellung weggelassen.
Quelle: eigene Darstellung

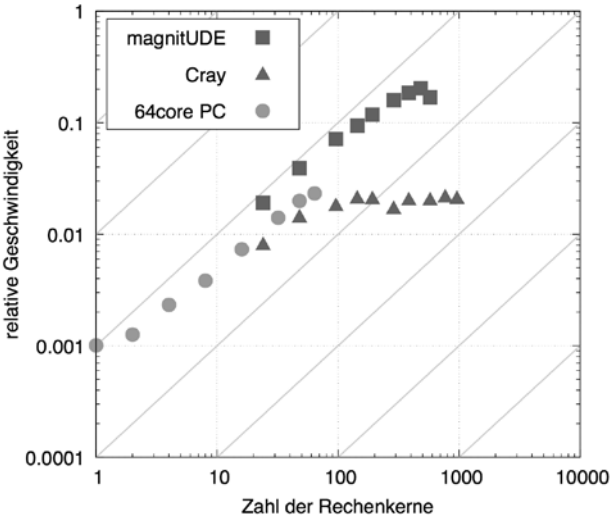
Ausgangspunkt des hier kurz umrissenen Forschungsprojektes war die experimentelle Beobachtung,[7] dass micellbildende Dodecyl-Trimethyl-Ammonium-Moleküle (DTAC-Moleküle) sehr unterschiedlich auf die drei chemisch sehr ähnlichen isomeren Hydroxybenzoat(HB)-Anionen reagieren. Diese Anionen besitzen eine identische Zusammensetzung, jedoch eine unterschiedliche räumliche Anordnung der Atome. Man nennt solche Moleküle *Isomere*; das sogenannte *ortho*-Isomer ist mit dem Wirkstoff Aspirin chemisch verwandt. Es wurde beobachtet, dass schon ein ganz geringer Zusatz dieses *ortho*-Isomeren (im Gegensatz zu den beiden anderen Isomeren) dazu führt, dass sich aus den eher kugelförmigen Micellen lange zylinder- oder fadenförmige Micellen bilden, was sonst meist erst bei sehr viel höherer Konzentration der DTAC-Kettenmoleküle passiert (s. Abb. 1). Da man damit potenziell über einen molekularen Mechanismus zur Steuerung von Form und Beweglichkeit solcher Strukturen verfügt, ist für den Experimentator natürlich die Frage interessant, warum die doch sehr ähnlichen Moleküle solche unterschiedliche Wirkungen auf das Micellaggregat haben. Dieser Frage widmete sich Jure Gujt in seiner Dissertation an der Universität von Ljubljana, die mit zum großen Teil am Lehrstuhl für Theoretische Chemie in Essen durchgeführten Molekulardynamik(MD)-Simulationsrechnungen erarbeitet wurde.

MD-Simulationen

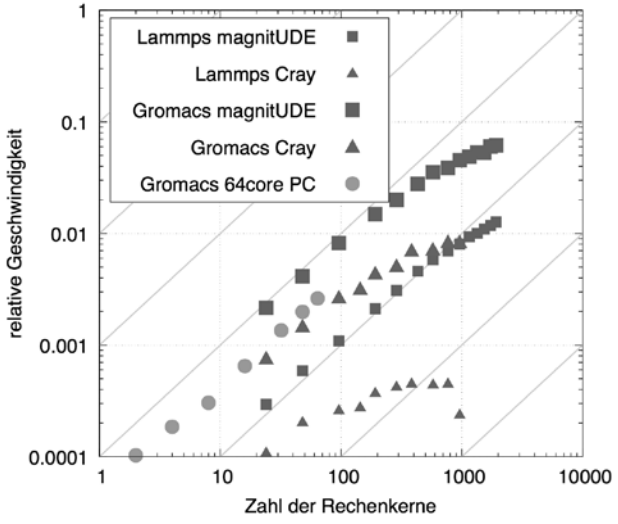
In MD-Simulationen erzeugt man die für die Mittelung über Molekülanordnungen nötigen Konfigurationen dadurch, dass man die zeitliche Entwicklung der Atome und Moleküle unter ihren wechselseitigen Kräften durch numerische Lösung der Newtonschen Bewegungsgleichungen verfolgt, und den Zustand des Systems in regelmäßigen Abständen abspeichert und während beziehungsweise nach der Simulation analysiert. Genauigkeit und Aussagekraft des Verfahrens sind dadurch limitiert, dass entweder bei einer

hochgenauen (und damit rechenzeitintensiven) Berechnung der Wechselwirkungen (z.B. mittels Verfahren der Elektronenstrukturtheorie) die mögliche Systemgröße gering ist, oder in hinreichend großen Systemen die Wechselwirkungen nur in sehr grober Näherung berechnet werden können.

Wir haben mit Hilfe der MD-Simulation zunächst nur die Stabilität einer bereits gebildeten Zylindermicelle untersucht. Um eine hinreichend große Zahl von DTAC-Molekülen (in unserem Fall 180) in Wasser bei niedriger Konzentration zu lösen, sind so viele Wassermoleküle notwendig, dass die Gesamtzahl der Atome notwendigerweise sehr groß ist. Die experimentellen Gegebenheiten kann man gar nicht direkt reproduzieren, stattdessen muss man sowohl bei der Konzentration der Micellbildner, als auch bei den diese Micellen in ihrer Form beeinflussenden HB-Anionen, Abstriche machen und deutlich höher konzentrierte Lösungen als im Experiment simulieren. Dies zeigt auch ein typisches Dilemma: das Experiment kann häufig keine atomistischen Aussagen treffen (da man in der Regel die Atome nicht direkt beobachten kann) und die Computersimulation kann das System in der Regel nicht bei den experimentellen Bedingungen simulieren bzw. die Wechselwirkungen nur sehr approximativ bestimmen. Wir haben uns dann für ein System aus 180 DTAC-Molekülen, 180 HB-Anionen und 180 Natriumchlorid-Ionenpaaren zur Neutralisierung entschieden, die in etwa 14.000 Wassermolekülen gelöst sind. Insgesamt enthalten diese simulierten Systeme mehr als 50.000 Atome (s. Abb. 1).



(2) Skalierungsverhalten auf den drei unterschiedlichen Rechnersystemen für das kleine System aus 51.700 Atomen.
Quelle: J.Gujt, E. Spohr, unveröffentlicht



(3) Skalierungsverhalten für das System aus 549.000 Atomen für die verwendeten Rechnerarchitekturen und zwei verschiedene Molekulardynamik-Codes.
Quelle: J.Gujt, E. Spohr, unveröffentlicht

Zeitintensive Simulationen

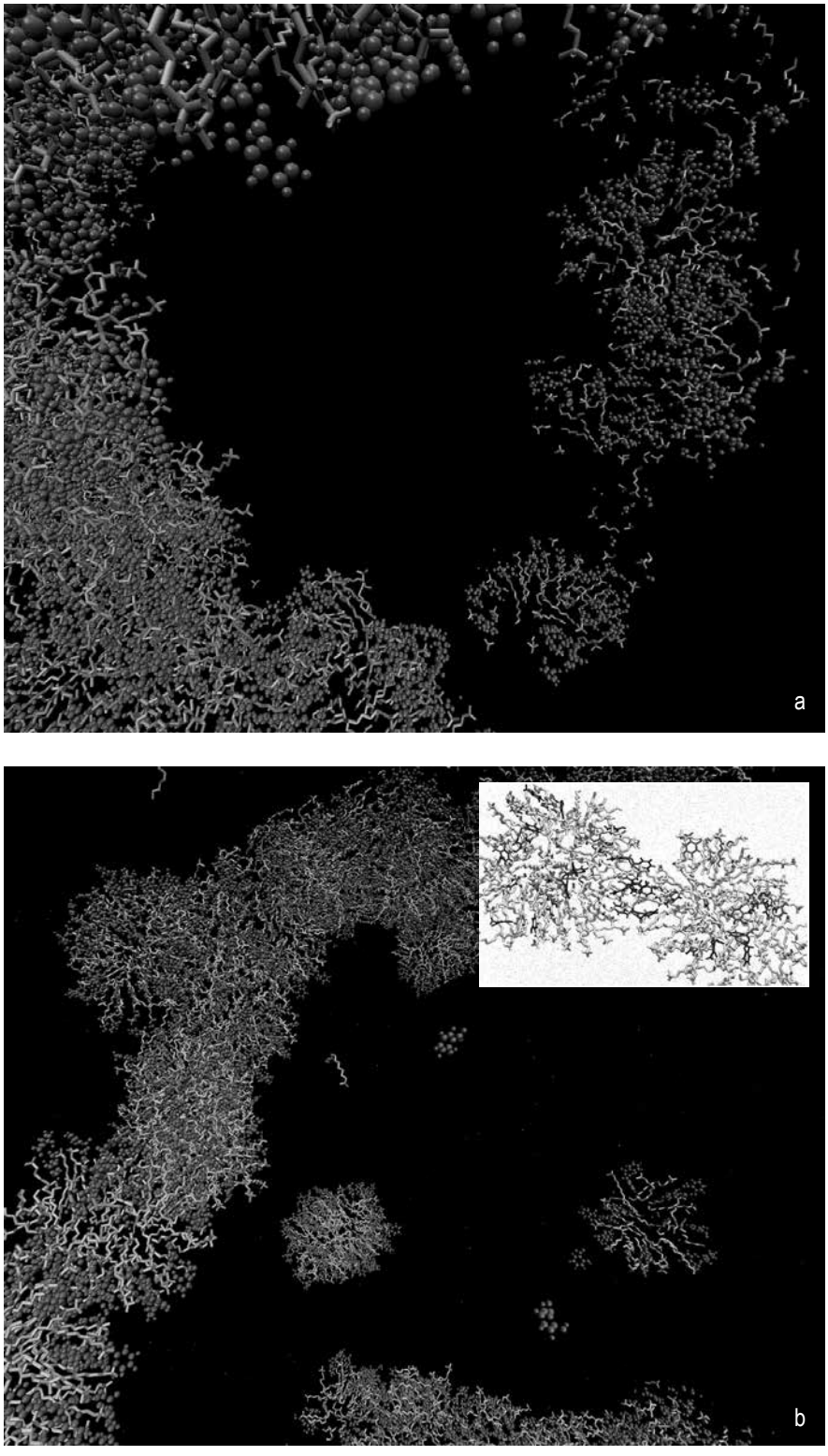
Zentrales Problem einer jeden Computersimulation ist die lange Zeitdauer einer einzelnen Simulationsrechnung, die typischerweise für die genannten Systeme zwischen 5.000 und 10.000 Stunden Prozessorzeit dauern, damit man das System selbst über eine Zeit von größenordnungsmäßig 100 Nanosekunden (also eine zehnmillionstel Sekunde) beobachten kann. Mit einem einzelnen relativ schnellen aktuellen PC-Prozessor würde das also größenordnungsmäßig ein Jahr (vielleicht einige Monate mehr oder weniger) dauern. Angesichts des typischen Zeitraums für eine Dissertation von etwa drei Jahren ist das deutlich zu lange! Man muss also die Rechenzeit verkürzen. Durch weitere Vereinfachungen bei der Beschreibung der Wechselwirkungen zwischen den Atomen ist dies nicht mehr möglich, weil bereits die einfachste gerade noch sinnvolle Approximation bei dieser Rechenzeitabschätzung zugrunde gelegt ist. (Für Expert*innen: die Simulationsrechnungen basieren auf der physikalisch gerade noch akzeptablen, aber häufig gemachten Approximation, dass die Wechselwirkungen zwischen Atomen gruppenweise addiert werden können, dass also der Einfluss anderer Atome auf die Art der Wechselwirkung nur in Form eines effektiven Potentials[8] einbezogen wird.)

Man muss also die Arbeit auf mehr als einen Prozessor verteilen. Kritisch ist dabei, wie gut die Rechnung *skaliert.* Der Idealfall perfekter Skalierung wäre erreicht, wenn bei Verdopplung der Zahl der Prozessoren eine Halbierung der Rechen-

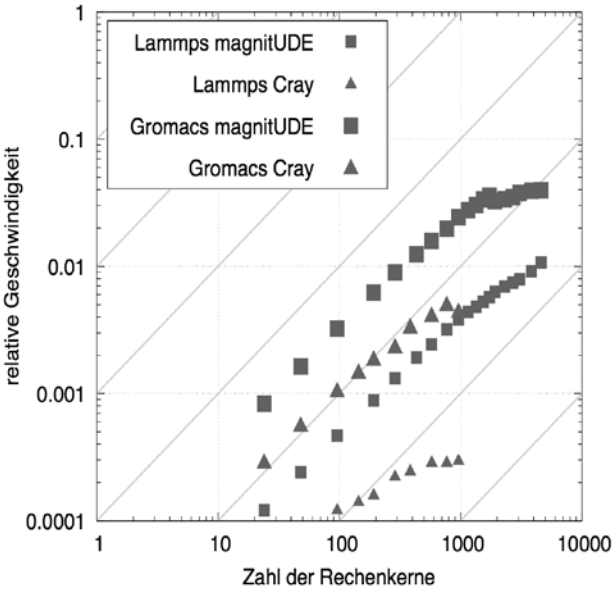
zeit erreicht würde, sich bei erneuter Verdopplung der Prozessorzahl die Rechenzeit noch einmal halbieren würde und so weiter, so dass sich zum Beispiel bei Verwendung von 1.024 Prozessoren die Rechenzeit auf ungefähr ein Tausendstel reduzieren würde. Wie in einer Behörde oder einem Unternehmen kann dies allerdings nicht immer so weiter gehen, weil ja irgendwann der Organisations-‚Overhead' (hier die „Kommunikation" zwischen den Prozessoren) den Zeitgewinn durch zusätzliche Resourcen auffrisst. Wie lange also diese Proportionalität zwischen Resourceneinsatz und Beschleunigung der Rechnung funktioniert, hängt von der Art der Wechselwirkungen zwischen den Atomen, der Organisation des Computercodes und der Architektur des Rechners ab.

Der Idealfall

Im Idealfall müsste also eine lineare Beziehung zwischen der reziproken Rechenzeit *1/T* und der Zahl der Rechenkerne *N* in der Form $1/T = f * N$ bestehen, wobei f ein Faktor ist, der von Problemgröße und Prozessorgeschwindigkeit abhängt. In doppelt logarithmischer Form ist dann $ln(1/T) = ln\, f + ln\, N$. Trägt man also den Logarithmus von *1/T* gegen den Logarithmus der Zahl der Prozessoren auf, dann sollte man eine Gerade mit der Steigung 1 erwarten. Abbildung (2) zeigt diese doppeltlogarithmische Auftragung sowie diverse Hilfslinien mit der idealen Steigung 1. Ein großer Teil der Rechnungen wurde mit einfachen modernen Server-PCs der Arbeitsgruppe durchgeführt. Relativ neue Maschinen mit bis zu 64 Rechenkernen zeigen durchgängig einen Verlauf parallel zu einer der Hilfslinien. Durch Einsatz von 64 Kernen kann man also im Beispiel die Rechnung um einen Faktor von ungefähr 25 beschleunigen, so dass Rechnungen für die Zylindermicellen innerhalb von einigen wenigen Wochen möglich sind. Dass der Faktor nicht nahe bei 64 liegt, ist



(4) Schnappschüsse der simulierten selbstorganisierten Micellstrukturen für die konzentriertere Lösung mit 549.000 Atomen (a) und die verdünntere Lösung mit 1.399.5,00 Atomen (b), wobei die etwa 150.000 beziehungsweise 450.000 Wassermoleküle nicht gezeigt sind. Der Einschub zeigt die Verbrückung zweier Micellen durch die ortho-HB-Moleküle.
Quelle: eigene Darstellung



(5) Skalierungsverhalten für das System aus 1.499.500 Atomen für zwei verschiedene Molekulardynamik-Codes.
Quelle: J.Gujt, E. Spohr, unveröffentlicht

teilweise auch eine Konsequenz der internen Architektur der verwendeten Prozessoren. Der schon etwas betagte CRAY-XT6m-Rechner (Installationsjahr 2010) zeigt bis zur Verwendung von 64 Kernen durchaus ähnliche Leistungen wie der aktuelle Server-PC. Man erkennt aber auch am Abknicken der Kurve für mehr als hundert Kerne, dass die Maschine für die (relativ) kleinen Systeme keine wesentliche Beschleunigung mehr bewirkt. Aufgrund der Vielzahl der verfügbaren Rechenknoten können wir allerdings Parameter systematisch verändern, um die Stabilität unserer Simulationsergebnisse gegen die gemachten Annahmen und physikalischen Approximationen zu überprüfen. Zu guter Letzt zeigt die Kurve für den State-of-the-art-Supercomputer magnitUDE, dass diese moderne Architektur noch einmal eine deutliche Steigerung pro Kern ermöglicht (erkennbar daran, dass die Kurve insgesamt höher liegt). Trotz des eher kleinen zu berechnenden Systems kann durch Einsatz von mehreren Hundert Rechenkernen insgesamt eine weitere Erhöhung der Geschwindigkeit um fast eine Größenordnung, also eine Reduktion der Rechenzeit von einigen Wochen auf zwei bis drei Tage, erreicht werden.

Brückenbildung

Aus diesen Rechnungen lernten wir dann, dass die unterschiedlichen molekularen Zusätze auf existente Micellen in unterschiedlicher Weise wirken. Gerade die Moleküle, die experimentell als besonders stark wirksam aufgefallen waren, lagern sich in sehr charakteristischer Weise in die Außenhülle der Micellen so ein, dass sie sich mit den entsprechenden Molekülen in anderen Micellen verbinden könnten.[9] Daraus entstand die Hypothese, dass die Bildung von Brücken aus zwei HB-Molekülen ein möglicher Mechanismus zur Fusion der kugelförmigen Micellen zu den beobachteten längeren zylindrischen Objekten sein könnte (s. Abb. 4). In Simulationen, in denen wir die Entstehung von Micellen aus den Einzelmolekülen verfolgten, ergaben sich auch Anhaltspunkte für diese Hypothese, jedoch konnten Einzelereignisse des Verschmelzens kleiner Strukturen statistisch nicht hinreichend gesichert werden, und es war auch nicht klar, ob diese langfristig stabil sein würden.

Um viele Verschmelzungs- und Zerfallsereignisse zu beobachten, kann man entweder ein System mit sehr wenigen Micellen über längere Zeit beobachten, oder, was im Rahmen der statistischen Mechanik äquivalent dazu ist, man kann ein System mit vielen Micellen beobachten. Im Idealfall kann man aus beiden Simulationen ähnliche Ergebnisse erhalten (ein Maß für das *ergodische Verhalten*[10] des simulierten Systems). Eine Vergrößerung des Systems ist allerdings aus physikalischen Gründen attraktiver, da man im großen System auch etwas über die dreidimensionale Anordnung der Micell-Strukturen (also die Hierarchie der Strukturen: von Molekülen zu Micellen, von Micellen zu Micellaggregaten) lernen kann. Darüber hinaus hat die Systemvergrößerung den Vorteil, dass man beim Einsatz von Großrechnern mehr Rechenkerne pro Rechenschritt einsetzen kann. Eine Verlängerung (z.B. eine Verzehnfachung) der Beobachtungszeit hätte eben auch eine Verlängerung der Wartezeit bis zur Fertigstellung der Rechnung (Verzehnfachung) zur Folge und wäre damit wenig praktikabel. Wir haben daher sehr viel größere Systeme mit insgesamt 549.000 Atomen simuliert.

Große Systeme

Für die größeren Systeme haben wir wieder Skalierungstests auf den Maschinen durchgeführt, die in Abbildung (3) gezeigt sind. Man erkennt, dass die Server-PCs zwar am besten skalieren, jedoch mit nur 64 Kernen insgesamt zu klein sind, um Ergebnisse in akzeptabler Zeit zu produzieren. CRAY-XT6m und NEC-magnitUDE skalieren bis zum Einsatz einiger Hundert Rechenkerne ganz gut und die Gesamtzeit für die Rechnung nimmt auch bei Einsatz von 1.024 Kernen (CRAY-XT6m) beziehungsweise 2.048 Kernen (NEC-magnitude) noch zu, obwohl dann die Rechnung doch bereits drei- oder viermal so lange dauert als es bei idealem Verhalten zu erwarten wäre, die Rechnung wird dann also ineffizient.

Die Abbildung (4a) zeigt dann ein typisches Arrangement der micellbildenden Moleküle für eines der Systeme, die teilweise auf der CRAY-XT6m, teilweise auf der NEC-magnitUDE berechnet wurden. Man erkennt, dass näherungsweise kugelförmige Gebilde sich aneinander anlagern und, stabilisiert durch die brückenbildenden HB-Moleküle, relative stabile Filamente bilden. Dabei könnte es sich um die Vorläufer, oder, da man die genaue Atomordnung aus dem Experiment auch nicht ermitteln kann, vielleicht sogar um die gesuchten Strukturen selbst handeln. Da bei dieser Simulation die Konzentration der strukturbildenden Moleküle noch zu hoch war, bildeten sich nicht nur perlenkettenartige Filamente, sondern auch Micellhaufen, die im Experiment so nicht beobachtet wurden.

Realistischere Konzentrationen

In einem weiteren Schritt, der den aktuellen Status unserer Forschung widerspiegelt, haben wir daher die Konzentration der Moleküle um einen Faktor 3 reduziert. Dies geschieht dadurch, dass man dreimal so viele Wassermoleküle, die in den Abbildungen nicht gezeigt sind, simuliert (s. Abb. 4b). Dadurch konnte man in der Tat die Filamente besser in der Lösung isolieren, die gerade statistisch analysiert werden. Die Tatsache, dass der bei weitem größte Teil des Rechenaufwands dann in der Simulation der Struktur und Dynamik des Lösungsmittels Wasser steckt, ist eine der zentralen Limitierungen der Computersimulation auf dem Gebiet der wässrigen Lösungen. Diese Begrenzungen können durch größere Rechenkapazität in Kombination mit modernen Simulationsverfahren, die über die oben beschriebene einfache Molekülдynamik in Echtzeit und in kleinen Schritten hinausgehen, teilweise überwunden werden. Dieser Aspekt soll hier aber nicht weiter erörtert werden.

Gromacs und Lammps

Abbildung (5) zeigt das Skalierungsverhalten der beiden Supercomputer für zwei verschiedene Molekulardynamik-Codes, die wir zur Simulation eines Systems aus insgesamt fast 1,5 Millionen Atomen benutzen. Gromacs[11] ist ein vor allem in der molekularen Biologie verwendeter Molekulardynamik-Code, der durch Assemblerprogrammierung zentraler Programmteile extrem schnell für diejenigen Probleme ist, die man damit lösen kann. Lammps[12] ist ein ausnehmend vielseitiger Molekulardynamik-Code, der nicht nur für Flüssigkeiten und Lösungen, sondern auch für Pulver und Festkörper eingesetzt wird. Diese Vielseitigkeit und die damit verbundene interne Struktur führt bei kleinen Prozessorzahlen zu deutlich geringerer Rechenleistung. Allerdings erkennt man auch, dass der Gromacs-Code ab 2.048 Kernen deutlich „einbricht“, während die Steigung der Kurve für den Lammps-Code auf der magnitude ein sehr viel besseres Skalierungsverhalten bis zu ungefähr 5.000 Rechenkernen zeigt. Für noch größere Systeme kann sich also in der Zukunft der Wechsel von Gromacs auf den Lammps-Code als nützlich erweisen, insbesondere, wenn man berücksichtigt, dass Lammps noch nicht so stark „ausoptimiert“ ist wie Gromacs. Abbildung (4b) zeigt eine Struktur der erhaltenen selbstorganisierten Filamente aus Micellen, die ihrerseits wieder jeweils aus 40 bis 100 Molekülen bestehen.

Schluss

Wir wollten mit diesem Artikel zeigen, wie das Wechselspiel aus Computer-Hardware, und -Software, physikalischen Prinzipien und chemischen Modellen dazu verwendet werden kann, Hypothesen zu entwickeln und zu evaluieren, die die Mechanismen der hierarchischen Selbstorganisation von Molekülen hin zu großen Strukturen erhellen. Die Forschung in diesem Feld der molekularen Wissenschaften (molekulare Biologie und Medizin, Chemie, Physik und Materialforschung) ist inhärent interdisziplinär und sehr stark dominiert vom Zugang zu Rechenanlagen verschiedener Größenordnungen. Ohne solche Rechenanlagen ist Forschung auf höchstem Niveau heutzutage in vielen Bereichen fast nicht mehr möglich. Die Universität Duisburg-Essen ist mit den gemeinsam vom CCSS und dem ZIM betriebenen Supercomputern gut aufgestellt. Dies erkennt man u.a. auch daran, dass die NEC-magnitUDE seit ihrer Inbetriebnahme in den “Top500”[13], der Liste der 500 schnellsten Supercomputer, vertreten ist, so wie es die CRAY-XT6m bei Inbetriebnahme im Jahr 2010 war.

Die Supercomputer von heute sind die Desktop-Geräte von morgen. Mit Betrieb und Nutzung eigener Supercomputeranlagen ermöglicht unsere Universität jungen Forscher*innen und Promovierenden die Nutzung von Zukunftstechnologie auf hohem Niveau für ihre wissenschaftliche Karriere.

Summary

Why does chemistry need large computers? Because the interactions between individual atoms and molecules determine the behavior of chemical systems, and because the numerical calculation of these interactions is difficult and numerically expensive. Computer simulations of these interactions form one of the pillars of modern natural sciences. Computers are particularly useful when it comes to disordered systems such as liquids and solutions in which hierarchical processes of self-organization take place on different time and length scales. A current research project of our group is the study of the formation and fusion of so-called micelles (which are, in

turn, self-assembled structures in aqueous solutions composed of many molecules with a hydrophilic (water-loving) and a hydrophobic (water-hating) end) into cylindrical and three-dimensional entities. The dependence of this behavior on small ions with specific activity offers the potential to manipulate the shape and dynamics of the micellar systems themselves. Instrumental and essential for such studies is the use of computer systems of varying computational power, which we demonstrate on the basis of our experience with the performance of server PCs in our group and the supercomputers CRAY-XT6m and NEC-magnitUDE at the Center of Computational Sciences and Simulation at our university.

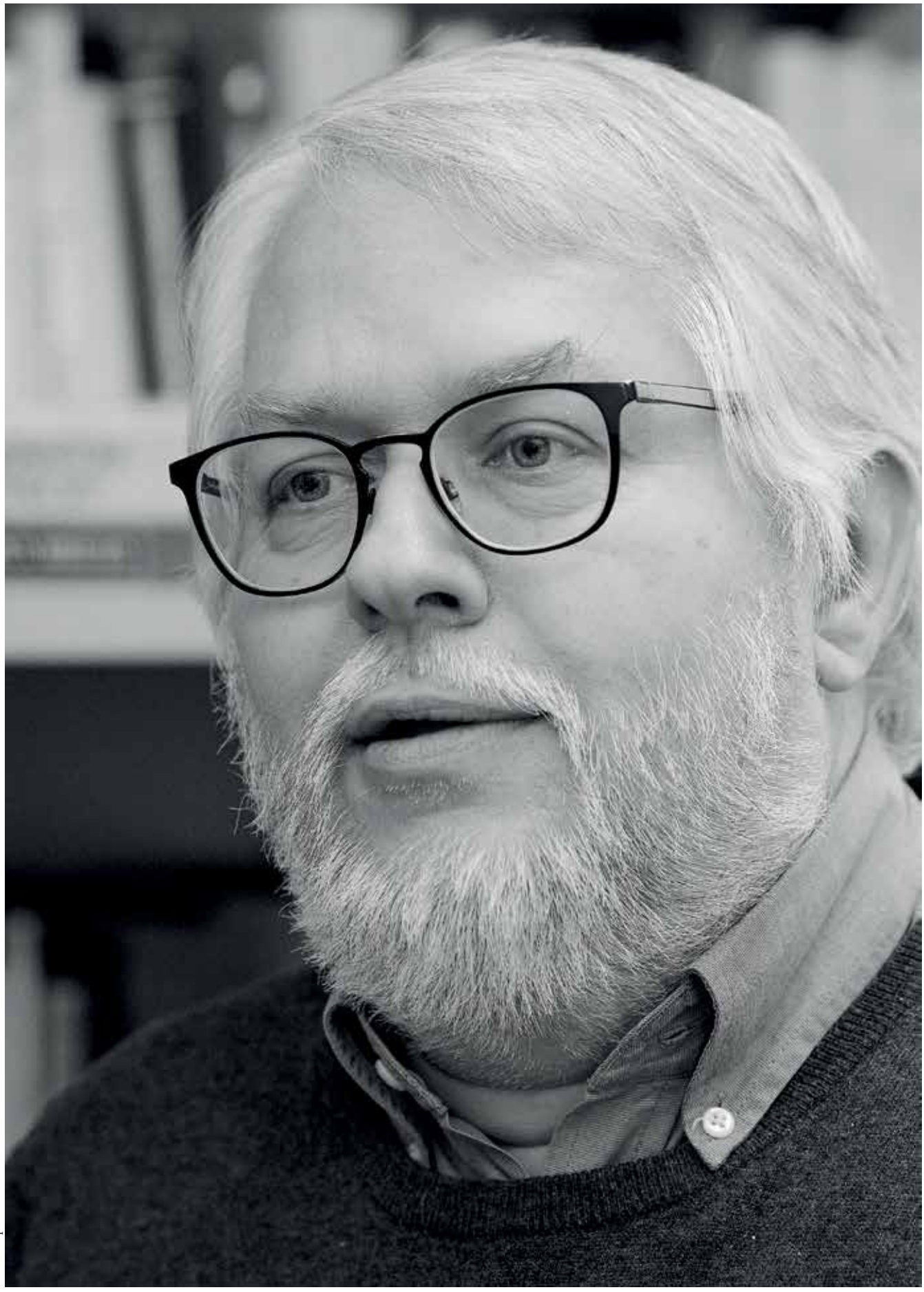
Anmerkungen/Literatur

1) http://www.theochem.rub.de/teaching/lab.de.html
2) https://www.rub.de/solvation
3) https://www.uni-due.de/ccss/cray.shtml
4) https://www.uni-due.de/ccss/magnitude_main.php
5) https://www.uni-due.de/zim
6) https://www.uni-due.de/ccss
7) Šarac, B.; Mériguet, G.; Ancian, B.; Bešter-Rogač, M. (2013): Langmuir 29, 4460–4469.
8) Allen, M.P. and Tildesley, D.J. (1987): Computer Simulation of Liquids, Oxford Science Publications, Clarendon Press, Oxford, 1987.
9) Gujt, J.; Bešter-Rogač, M.; Spohr, E. (2016): Langmuir 32, 8275–8286.
10) Reif, F. (1965): Fundamentals of Statistical and Thermal Physics, McGraw-Hill, Singapore.
11) http://www.gromacs.org
12) http://lammps.sandia.gov
13) http://www.top500.org

Die Autoren

Jure Gujt studierte Chemie in Ljubljana, wo er 2015 auch promovierte. Von 2011 bis 2015 war er Juniorforscher am Lehrstuhl für Physikalische Chemie in Ljubljana; seit 2015 ist er Postdoc in der Arbeitsgruppe von Prof. Eckhard Spohr an der Universität Duisburg-Essen. Für seine Dissertation am Lehrstuhl von Eckhard Spohr erhielt Jure Gujt 2013 und 2014 Stipendien der Slowenischen Regierung. 2015 wurde ihm der Krka Preis für seine Dissertation verliehen.

Eckhard Spohr diplomierte 1983 und promovierte 1986 an der Universität Mainz. Dann war er bis 1988 Wissenschaftlicher Mitarbeiter am Max-Planck-Institut für Chemie in Mainz. Es folgte ein Feodor Lynen Fellowship der Alexander von Humboldt Stiftung am Department of Chemistry der University of California in Irvine. Von 1991 bis 2000 war er wissenschaftlicher Mitarbeiter der Abteilung für Theoretische Chemie an der Universität Ulm, wo er sich 1995 habilitierte. Es folgte eine zweijährige Zeit als wissenschaftlicher Mitarbeiter am Institut für Energieforschung am Forschungszentrum Jülich, bevor Eckhard Spohr von 2002 bis 2007 dort seine eigene Arbeitsgruppe zum Thema „Grundlagen der Physikalischen Chemie“ leitete. Seit 2007 bekleidet er den Lehrstuhl für Theoretische Chemie an der Universität Duisburg-Essen. 1988 bekam Spohr die Otto Hahn-Medaille der Max Planck Gesellschaft verliehen und erhielt 1998 den Merckle Forschungspreis der Universität Ulm. Er ist national und international in vielen Gremien wie DFG und DAAD, der Israel Science Foundation oder auch der American Chemical Society als Gutachter gefragt.



Eckhard Spohr. Foto: Vladimir Unkovic

Die möglichst akkurate Vorhersage des Strömungsverhaltens von ~~Flüssigkeiten~~ ist nach wie vor eine der größten Herausforderungen der Ingenieurwissenschaften. Hier setzt die Arbeitsgruppe von Andreas Kempf an.

Vorhersage der Turbulenz in Reaktoren und Motoren

High-Performance Computing in der Strömungssimulation

Von Andreas M. Kempf, Fabian Proch, Andreas Rittler, Martin Rieth & Thuong M. Nguyen

Das Turbulenzproblem

Die Strömung von Fluiden, also von Gasen oder Flüssigkeiten, stellt für Ingenieure noch immer eines der schwierigsten Berechnungsprobleme dar. Die Navier-Stokes-Gleichungen, welche die Strömung typischer Fluide vollständig als Kontinuum beschreiben, sind zwar schon seit langem bekannt, als gekoppelte partielle Differentialgleichungen in drei Raum- und einer Zeitdimension weisen sie aber sehr interessante Eigenschaften auf. Die Lösung im Inneren eines Rechengebietes hängt zwar nur vom Zustand auf dessen Rändern ab (man spricht von den Randbedingungen des Randwertproblems), es wird aber schnell klar, dass die Geometrie der Ränder und der Zustand darauf nicht leicht zu beschreiben sind – man denke hier nur an die vielen Schaufeln einer Turbine, die den bewegten Rand eines durchströmten Gebietes darstellen.

Erschwerend kommt hinzu, dass die Gleichungen im Inneren des Rechengebietes zur Anfachung von Störungen neigen, aus denen schnell ein System scheinbar zufälliger, ineinander eingebetteter Wirbel unterschiedlicher Größe entsteht. Man spricht dann von Turbulenz oder einer turbulenten Strömung – die Pluralform der „Turbulenzen“ bleibt Laien vorbehalten. Turbulenz entsteht in Zonen starker Scherung, also in Schichten zwischen zwei Strömen, die aneinander oder an einer Wand abgleiten. Damit erklärt sich auch, warum in der Luftfahrt Turbulenz meist nur in Bodennähe beobachtet wird und in großen Höhen nur dann, wenn starke Windscherung auftritt, die dort Turbulenz produziert. Die in den großen Wirbeln gespeicherte kinetische Energie wird dann auf kleinere Wirbel übertragen, bis sich eine immer gleiche Verteilung der Energie über die Wirbelgrößen hinweg einstellt – die sogenannte Energiekaskade. Dabei erleben die

Andreas M. Kempf Foto: Vladimir Unkovic

großen Wirbel aufgrund ihrer Trägheit nur sehr wenig Reibung aus der Zähigkeit (Viskosität) des Fluides, wohingegen die kleinen Wirbel geringer Trägheit schnell ausgebremst werden, die kinetische Energie dieser Wirbel wird dann in Wärme umgewandelt, man spricht dabei von Dissipation. Die Größe der kleinsten noch auftretenden Wirbel hängt dann von der Scherrate der Strömung und der Viskosität des Fluides ab. Eine genaue Simulation einer turbulenten Strömung erfordert dann, das gesamte Strömungsgebiet mit Rechenpunkten zu füllen, auf denen die Navier-Stokes Gleichungen gelöst werden, wobei der Abstand der Punkte kleiner sein muss als die kleinsten Wirbel, um diese noch abbilden zu können. Leider würde eine solche „Direkte Numerische Simulation“ (DNS) für fast alle Ingenieursanwendungen zu viele Rechenpunkte und damit auch Rechenleistung erfordern, um praktikabel zu sein.

Die Bildung, Entwicklung und Dissipation turbulenter Wirbel wird als deterministischer Prozess angesehen und ist durch die Navier-Stokes-Gleichungen genau beschrieben, allerdings reagiert der Prozess sehr empfindlich auf kleinste Störungen und Rechenungenauigkeiten, so dass sich einzelne Wirbel bei Wiederholung eines Experiments schon nach kurzer Zeit anders verhalten mögen als zuvor. Diese enorme Sensitivität turbulenter Strömungen auf kleinste Störungen und deren Verstärkung lässt sich leicht an Wettervorhersagen erkennen, die ja kaum mehr als drei Tage zuverlässig vorhersagen können. Statistisch betrachtet sind aber turbulente Strömungen berechenbar, werden also zumindest in Ingenieursanwendungen zu reproduzierbaren Ergebnissen führen. Als Beispiel sei aufgeführt, dass die Anordnung der vielen kleinen turbulenten Wirbel um ein Flugzeug jetzt und einen Augenblick später sicher verschieden ist, die Strömung aber im Wesentlichen gleich bleibt, das Flugzeug fliegt so wie zuvor. Turbulenz lässt sich also fast immer als ein stochastischer (zufälliger) und ergodischer Prozess ansehen, der zu einem konstanten Mittelwert der Strömungsgeschwindigkeit und des Druckes führt – und somit zu einer konstanten Leistung einer Anlage oder eines Luftfahrzeugs.

Für den Mittelwert, aber auch für die Größe der Schwankungen, lassen sich Gleichungen herleiten, die den Navier-Stokes Gleichungen ähneln, allerdings einen „ungeschlossenen Term“ aufweisen, der sich aus dem Gleichungssystem heraus zunächst nicht bestimmen lässt. Ursache für die somit ungeschlossenen Gleichungen ist der unwiederbringliche Informationsverlust, welcher bei der Vereinfachung eines fluktuierenden Strömungsfeldes auf dessen Mittelwerte entsteht. Für die ungeschlossenen Terme lassen sich jedoch Modellannahmen treffen, mit denen eine Lösung der Gleichungen möglich wird. Solche mathematisch formulierten Modellannahmen fasst man allgemein unter der Bezeichnung „Turbulenzmodell“ zusammen.

Klassischerweise wird Turbulenz mit den Reynolds-gemittelten Gleichungen gelöst – also nach Zeitmittelung der Gleichung, die dann auf eine stationäre Lösung führen muss. Diese Lösung ist wegen der zeitlichen Mittelung normalerweise auch sehr glatt, so dass die Lösung mit wenigen Punkten im Raum gefunden werden kann, der Rechenaufwand wird also stark reduziert. Die Dynamik von Strömungsfeldern mit zeitabhängigen Prozessen wie Schallausbreitung, Wirbelablösung oder bewegten Rändern lässt sich so jedoch kaum abbilden. Auch umfassen die ungeschlossenen Terme in diesem Fall sämtliche turbulenten Wirbel aller Größen und Stärken, deren genaue Modellierung kaum möglich ist. Dennoch stellen diese sogenannten Reynolds-Averaged Navier-Stokes (RANS)-Methoden nach wie vor den Stand der Technik dar.

Eine modernere Alternative ist die sogenannte Grobstruktursimulation oder Large-Eddy Simulation (LES).

LES – flexibler Kompromiss aus Genauigkeit und Rechenzeit

Bei der LES wird statt einer Zeitmittelung eine räumliche Filterung der Lösungen und Gleichungen durchgeführt, die Gleichung beschreibt also nur noch die groben Strukturen der Strömung, eine Lösung lässt sich daher mit geringerer räumlicher Auflösung finden. Die wesentliche Dynamik der Strömung bleibt aber erhalten. Auch umfassen die ungeschlossenen Terme nur noch die vielen kleinen Wirbel, die sich wegen ihrer Ähnlichkeit sehr leicht statistisch beschreiben und genau modellieren lassen. Die räumliche Filterung entfernt dabei Strukturen, die kleiner sind als die Filterweite. Durch Einstellung der Filterweite lässt sich steuern, wie stark gefiltert werden soll – oder wie groß die kleinsten verbleibenden Wirbel sein sollen. Somit lässt sich einstellen, wie weit die Punkte des Rechengitters voneinander entfernt und somit wie rechenaufwendig eine Simulation ist – allerdings ist ein Mindestrechenaufwand erforderlich, der deutlich über dem der RANS-Methoden liegt. Natürlich gilt auch hier, dass mit mehr Rechenpunkten eine höhere Genauigkeit erzielt wird, so dass die LES die Möglichkeit bietet, die Zahl der Rechenpunkte als jeweils besten Kompromiss aus Rechenzeitbedarf und Genauigkeitsanforderung zu wählen. Besonders attraktiv an diesem Konzept ist, dass sich die Rechengenauigkeit bei mehr Rechenleistung automatisch erhöht, die Methode eignet sich daher hervorragend für das Hochleistungsrechnen und verbessert sich gemäß des Mooreschen Gesetzes. Die Entwicklung, Verbesserung und Anwendung der LES-Methode für chemisch reagierende Strömungen stellt seit über 15 Jahren den Forschungsschwerpunkt von Professor Kempf dar.

Turbulenz in reagierender Strömung

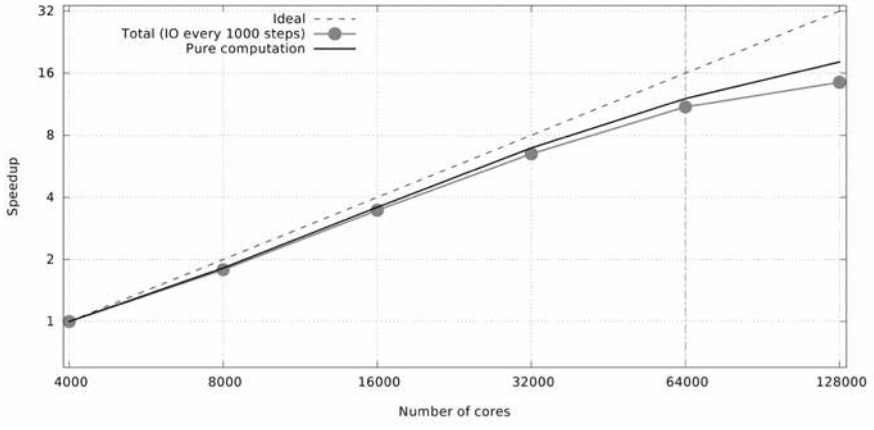
Chemische Reaktoren, Brennkammern in Gasturbinen, Brennräume in Kolbenmotoren und ähnliche Systeme nutzen chemische Reaktionen, die entweder zur Produktion von Stoffen oder zur Wärmefreisetzung und anschließenden Verrichtung von Arbeit dienen. Normalerweise sind solche Systeme mindestens so groß, dass sich eine turbulente Strömung einstellt. Die turbulenten Wirbel fördern die Mischung der Reaktanten oder des Brennstoffs mit der Luft, so dass sich eine schnellere Umsetzung ergibt. Als Resultat lässt sich ein Reaktor oder ein Motor dank turbulenter Strömung noch relativ klein bauen. Da die Vermischungsrate der Reaktanten aber durch die turbulenten Wirbel bestimmt wird, ist auch die chemische Reaktionsrate und die Leistung einer Maschine durch die Turbulenz bestimmt, welche schwierig zu berechnen ist – insbesondere durch die Interaktion von der Turbulenz mit komplizierten chemischen Reaktionen. Bei letzteren ist wiederum zu beachten, dass die Raten der chemischen Umsetzung von vielen Parametern abhängen und nur vollständig berechnet werden können, wenn sämtliche Teilreaktionen bekannt sind. So werden zum Beispiel über dreihundert Einzelreaktionen und deren Reaktionsraten benötigt, um die Verbrennung von Methan (Erdgas) in Luft zu beschreiben. Es ist daher erforderlich, auch die chemischen Reaktionen vereinfacht zu beschreiben, um realistische turbulente Flammen selbst auf Höchstleistungsrechnern simulieren zu können. Solche Methoden, welche die chemische Umsetzung in einer Flamme mathematisch vereinfacht beschreiben, werden als Verbrennungsmodell bezeichnet. Solche Modelle werden von der Gruppe um Andreas Kempf entwickelt, getestet und in Programme implementiert – zum Beispiel in die Eigenentwicklung „PsiPhi“.

Das Simulationsprogramm PsiPhi

Die Entwicklung des Programms PsiPhi[1–3] wurde vor fast zehn Jahren von Andreas Kempf am Imperial College London gestartet, damals noch als kleines Projekt um mehr über Strömungssimulation zu lernen und diese den Doktorand*innen besser zu vermitteln. Daher wurde immer Wert auf eine sehr einfache, kompakte und klare Struktur gelegt; Einfachheit und Verständlichkeit wurde dabei immer als wichtiger angesehen als Recheneffizienz, Mächtigkeit oder Flexibilität. Ebenfalls stand immer die leichte Erweiterbarkeit und Programmierbarkeit im Vordergrund, da die Mitarbeiter*innen der Arbeitsgruppe ja nicht nur bestehende Modelle anwenden, sondern vielmehr neue Methoden und Modelle entwickeln und testen sollen. Dieses Konzept hat sich bis heute bewährt und mit wenigen Entwicklern zu einem enorm leistungsfähigen Programm geführt, das chemisch reagierende Strömungen und Flammen, partikelförmige Brennstoffe, bewegte Geometrien wie Kolben in Motoren, und Wärmeaustausch durch Strahlung genau und effizient simulieren kann – und zwar für laminare und turbulente Strömungen, von fast ruhenden Fluiden bis in den Überschallbereich.

Mittlerweile hat sich das Programm auf verschiedensten Rechnern vom PC bis zum Höchstleistungsrechner bewährt – von einem Prozessorkern bis zu über 128.000 Prozessorkernen auf dem Höchstleistungsrechner JUQUEEN in Jülich (BlueGene/Q). Auch wurden schon Simulationen mit 1,7 Milliarden Rechenpunkten auf SuperMUC in München durchgeführt, für eine turbulente Kohlestaubflamme mit Kohlevergasung, Koksabbrand und Wärmeübergang durch Strahlung.

Die Nutzung von vielen Tausend Prozessoren erfordert ein klares Konzept zur Parallelisierung und erhebliche Optimierung, um zu verhindern, dass sich die Prozessoren nur gegenseitig stören oder aufeinander warten – ähnlich einer aufgeblähten Verwaltung. Die eingesetzte Parallelisierung basiert auf einer Gebietszerlegung, das heißt das gesamte Rechengebiet wird in Blöcke mit der gleichen Zahl von Rechenpunkten zerlegt. Diese Blöcke müssen an ihren Rändern Informationen mit den Nachbarn austauschen, was über das Versenden von „Nachrichten“ geschieht, gemäß dem MPI-Standard (Message Passing Inter-



(1) Test der parallelen Skalierbarkeit des PsiPhi-Programms auf dem Höchstleistungsrechner JUQUEEN in Jülich auf bis zu 128.000 Prozessorkernen. Die gestrichelte Linie gibt die ideale Skalierbarkeit an, die gepunktete Linie die vom Programm mit Schreiben von Ausgabedateien erreichte Skalierbarkeit und die einfache Linie die vom Programm erreichte Skalierbarkeit ohne das Schreiben von Ausgabedateien.
Quelle: Fabian Proch

face). Das andauernde Versenden und Empfangen solcher Nachrichten erfordert aber viel Zeit, weshalb die Nachrichten schon versandt werden, bevor die Information zur Weiterverarbeitung benötigt werden. Trifft die Nachricht dann ein wird sie abgelegt und erst bei Bedarf verarbeitet.

Die hohe Zahl der auszutauschenden Nachrichten ergibt sich bei PsiPhi aus der Möglichkeit, das Fluid nicht nur als kompressibel sondern auch als inkompressibel zu behandeln, was bei niedrigen Geschwindigkeiten einen enormen Effizienzgewinn bedeutet. Leider erfordert die inkompressible Behandlung aber die Lösung eines großen algebraischen Gleichungssystems über alle Blöcke des Rechengebiets hinweg, so dass nach jeder Iteration Nachrichten zwischen den Blöcken versendet werden müssen. Der Aufwand für den Versand dieser Nachrichten ist so hoch, dass typische massiv parallele Programme für Strömungsprobleme auf inkompressible Lösungen verzichten, bei PsiPhi ist aber noch eine effiziente Implementierung gelungen.

Im Bereich des Höchstleistungsrechnens auf vielen tausend Prozessoren ist es wichtig, die wertvollen Prozessoren auch effizient zu nutzen, also die Wartezeiten und den Aufwand für das Versenden, Warten und Empfangen von Nachrichten gering zu halten. Tatsächlich gibt es Programme, die mit zu vielen Prozessoren langsamer laufen als mit wenigen. Bei PsiPhi ist aber eine hohe „parallele Effizienz“ gelungen, das heißt eine Verdopplung der Prozessorzahl führt auch bei sehr vielen Prozessoren noch fast zu einer Halbierung der Rechenzeit. Diese gute parallele Effizienz oder „Skalierbarkeit“ unseres Inhouse-Programms PsiPhi ist im Scaling Plot (Abb. 1) dargestellt.

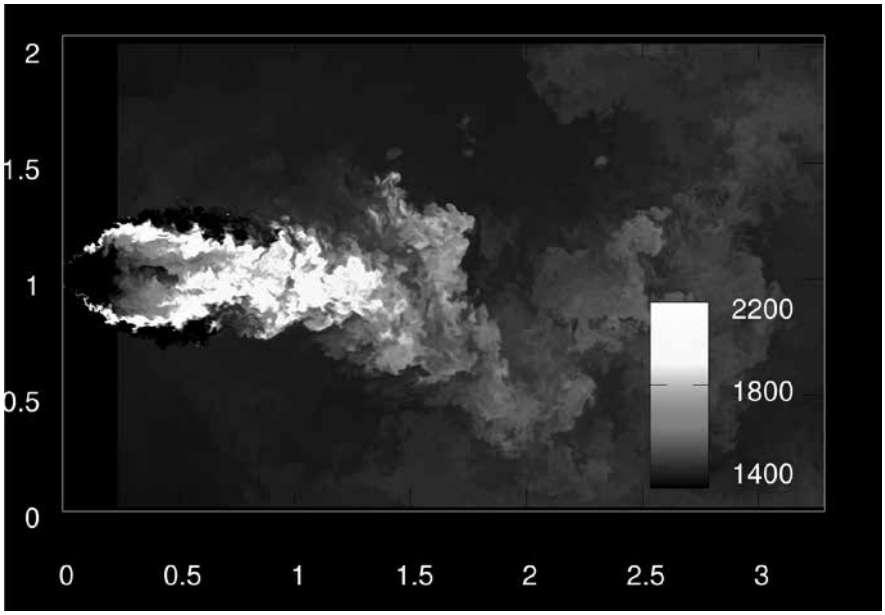
In unserer Forschung ist PsiPhi ein Werkzeug, um neue Modelle und Verfahren zu entwickeln und zu testen, und um komplizierte Strömungsfelder zu untersuchen. Im Folgenden werden Beispiele aus drei interessanten und komplexen Themenfeldern dargestellt.

Simulation von Kohlestaubfeuerung

Kohlestaubfeuerung in Kraftwerken gehört nach wie vor und in absehbarer Zukunft zu den Hauptlieferanten elektrischer Energie, was mit geringen relativen Kosten und der hohen Verfügbarkeit von Kohle zusammenhängt. Allerdings führt die Verbrennung von Kohle zu einem hohen Ausstoß von Kohlenstoffdioxid verglichen mit etwa Erdgas, da Kohle wesentlich aus Kohlenstoff besteht. Um möglichst wenig Kohlendioxid und Schadstoffe auszustoßen, müssen Kohlekessel möglichst effizient ausgelegt und der Verbrennungsprozess möglichst gut verstanden werden. Darüber hinaus wird an modernen Verfahren geforscht um den Ausstoß zu reduzieren, etwa an der Verbrennung mit reinem Sauerstoff und Speicherung des Kohlendioxids aus dem Abgas oder der Nutzung von Biomasse. Simulationen von Kohlestaubfeuerung können wichtige Informationen zur Erforschung und Verbesserung von Prozessen der Feuerung von Kohle liefern, die ansonsten wegen der rauen Verbrennungsumgebung und Größe der Anlagen experimentell nicht zugänglich sind. Mit der Verbesserung der Methoden zur Simulation von Kohlestaubfeuerung befasst sich der Lehrstuhl für Fluiddynamik.

Die Simulation von Kohlestaubfeuerung ist ein sehr komplexes Problem. Zum einen muss, wie bei reinen Gasflammen auch, die Interaktion von Turbulenz und chemischer Reaktion abgebildet beziehungsweise modelliert werden. Zum anderen müssen die Prozesse, die die Kohlepartikel durchlaufen, in der Simulation enthalten sein. Dies beinhaltet die Bewegung der Kohlepartikel in der turbulenten Strömung und den Impulsaustausch zwischen Gas und Partikeln, den konvektiven Wärmeaustausch zwischen Kohlepartikeln und Gas sowie den Wärmeaustausch über thermische Strahlung sowie die Ausgasung und den Koksausbrand der Kohlepartikel. All diese Prozesse werden in PsiPhi[4,5], simuliert, wobei die Kohlepartikel in der Simulation als Punktmassen angenommen werden. Zusätzlich werden bei Simulationen größerer Brennkammern Partikel zu ‚Paketen‘ zusammengefasst, da die Simulation von typischerweise mehreren Milliarden Kohlepartikeln zu aufwendig wäre.

Das Hauptaugenmerk liegt am Lehrstuhl für Fluiddynamik an der Verbesserung der Modelle zur Beschreibung der turbulenten Verbrennung der Gase in einer Kohlestaubfeuerung. Das Modell, das dafür am Lehrstuhl für Fluiddynamik entwickelt wird, beruht auf der Annahme, dass sich die turbulente Flamme als Zusammensetzung vieler einzelner kleiner laminarer, also nicht von Turbulenz gekrümmter und gefalteter Flammen, darstellen lässt. Dies ermöglicht es, die möglichen chemischen Zustände vor der Grobstruktursimulation mittels eindimensionaler Flammensimulationen zu bestimmen und in einer Datenbank abzulegen. Diese Datenbank kann dann auf effiziente Weise während der Grobstruktursimulation abgerufen und der Gaszustand mittels weniger in der Grobstruktursimulation gelösten Parameter bestimmt werden. Das in Abbildung (2) gezeigte Temperaturfeld ist ein Beispiel einer aus der Datenbank ausgelesenen Größe. Die in der Grobstruktursimulation gelösten Parameter beschreiben etwa, aus welchen Bestandteilen sich das Gas ursprünglich, also vor der Reaktion zusammensetzt. Dabei spielt bei der Kohlestaubfeuerung die Tatsache eine besondere Rolle, dass unterschiedliche Gase aus den Kohlepartikeln austreten, je nachdem in welcher Phase der Verbrennung sich das Kohlepartikel befindet. Dabei können im Wesentlichen zwei Phasen unterschieden werden. Zu



(2) Temperaturverteilung in einem Ausschnitt der Grobstruktursimulation einer Kohlebrennkammer semi-industriellen Maßstabs. Das gesamte Simulationsgebiet umfasst 1,7 Milliarden Rechenpunkte und die Simulation wurde auf 16384 Prozesskernen auf dem Höchstleistungsrechner SuperMUC am Leibniz-Rechenzentrum in München durchgeführt. Die Kosten der Simulation belaufen sich auf etwa 3 Millionen Rechenstunden. Achsenbeschriftungen sind in Metern und die Temperatur in Kelvin angegeben.
Quelle: Martin Rieth

Anfang, wenn das Kohlepartikel in den Brennraum eintritt und sich schnell erhitzt, brechen die schwachen Verbindungen in der chemischen Struktur der Kohle, und leichtere Kohlenwasserstoffe sowie schwerere Moleküle wie gasförmiger Teer werden ausgegast. Die Zusammensetzung der Gase ist abhängig davon, wie schnell das Kohlepartikel und bei welcher Temperatur aufgeheizt wird. Nachdem dieser Prozess abgeschlossen ist, liegt das Kohlepartikel als poröse Struktur vor, der noch brennbare Anteil besteht hauptsächlich aus Kohlenstoff. Dieser Kohlenstoff brennt an der Oberfläche der porösen Struktur partiell ab, wenn sich das Kohlepartikel in einem heißen Gemisch aus Abgas und noch verfügbarem Sauerstoff befindet. Die Gase aus diesen beiden Phasen, im ersten Fall ein Gemisch aus unterschiedlichsten Kohlenwasserstoffen und im zweiten Fall im Wesentlichen Kohlenmonoxid, werden in der Gasphase verbrannt und der chemische Zustand des Gases ist dann davon abhängig, aus welcher Phase der Kohleverbrennung das Gas stammt.

Die Methoden, die am Lehrstuhl für Fluiddynamik für die Simulation von Kohlestaubfeuerung entwickelt werden, können in Zukunft auf der einen Seite dazu dienen den Prozess und die zugrundeliegende Physik besser zu verstehen. Auf der anderen Seite können die Methoden in Zukunft von Ingenieur*innen und Wissenschaftler*innen benutzt werden, um möglichst effiziente und saubere Kohlekraftwerke zu entwickeln. Außerdem werden die Grundlagen gelegt, um Simulationen der Feuerung von Biomasse durchzuführen, und so weiter zu einer günstigen CO_2-Bilanz beizutragen.

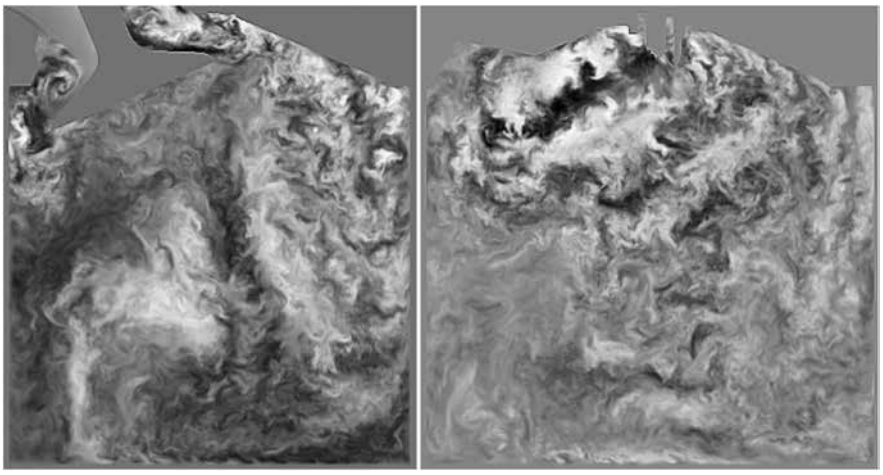
Simulation von Kolbenmotoren

Motorsimulation ist ein großes Thema im Bereich der numerischen Strömungssimulation (CFD). Nicht nur aufgrund der komplizierten Motorgeometrie, sondern auch aufgrund der Dynamik von Fluiden im Motor und dem Verbrennungsverfahren, ist die numerische Modellierung von Verbrennungsmotoren kompliziert und herausfordernd. Der Rechenaufwand ist zudem sehr hoch, da für Statistiken und gemittelte Werte mehrere Motorzyklen simuliert werden müssen.

Durch die komplizierte Motorgeometrie ist die Generierung der Rechengitter äußerst zeitaufwendig und trotzdem oft nur ein Kompromiss in Bezug auf die Qualität. Die Generierung dieser Gitter allerdings ist nicht nur zeitaufwändig, sondern zudem beschwerlich, da Ingenieure ein paar Wochen bis Monaten nur mit der Netzgenerierung beschäftigt sind. Durch die häufige Nutzung einer dynamischen Anpassung des Netzes oder anderer Methoden zu der Behandlung von beweglichen Netzen wird die Komplexität von Motorsimulationen noch zusätzlich erhöht.

Wichtig ist, dass die Turbulenz während des Prozesses vom Einlass über die Kompression bis zur Verbrennung erzeugt wird. Bei der Entwicklung der Turbulenz im Motorzylinder spielt die Interaktion zwischen den beweglichen Kolben, Ventilen und der Trägheitskraft der Fluiden eine bedeutende Rolle. Zur Auflösung dieser Turbulenz im Motorzylinder sind akkurate numerische Verfahren auf feinen Gittern notwendig. Nicht zuletzt, da die Aufrechterhaltung der turbulenten kinetischen Energie eine wichtige Rolle für die Ausbreitung der Flammen im Brennraum spielt. Wie die Abbildungen (3) und (4) der Simulationen mit dem In-house Code PsiPhi zeigen, werden feine turbulente Strukturen in der bildnormalen Geschwindigkeitskomponente der Strömung im Zylinder und die Ausbreitung der Flamme in der Brennkammer aus der Simulation erhalten.

Zur Durchführung von Motorsimulationen wurden Methoden in PsiPhi implementiert, welche die gefilterten Navier-Stokes-Gleichungen kompressibler Gase zusammen mit bewegten Wänden und nicht reflektierenden Randbedin-



(3) Die unmittelbare normale Geschwindigkeit der Strömung im Motorzylinder von Motorsimulation mit 8192 CPUs am Jülich Supercomputer JUQUEEN.
Quelle: Nguyen Thuong

gungen lösen können. Die Verbrennungsmodellierung wird durch ein Flame-Surface-Density Model realisiert, in dem den Wrinkling-Faktor der turbulenten Flame von einem algebraischen Modell berechnet wird. Anstatt auf unstrukturierte Methoden zurückzugreifen, benutzt PsiPhi für die Simulation von Kolbenmotoren äquidistante kartesische Gitter, die aus STL-Daten der Geometrie innerhalb von kurzer Zeit automatisch generiert werden. Zur Behandlung von beweglichen Wänden verwendet PsiPhi einen partikelbasierten Ansatz in Kombination mit einer Immersed Boundary Methode (IBM). Die Bewegung dynamischer Ränder, wie zum Beispiel der Einlass- und Auslassventile oder des Kolbens im Rechengebiet, wird hierbei durch Lagrangepartikel beschrieben. Die Partikelpositionen bestimmen dabei die Grenzen für die IBM, welche die Veränderung des Rechengebietes zur Laufzeit berücksichtigt. Bei dieser Vorgehensweise wird die komplizierte adaptive Netzerzeugung vermieden, wobei die Effizienz der Simulation deutlich verbessert wird. Eine Besonderheit liegt dabei in dem gewählten Ansatz, bei dem sich die Parallelisierung des Partikeltransports als trivial und äußerst günstig erwiesen hat. Die Gebietszerlegung zur Parallelisierung ist dabei unabhängig von der gewählten Motorgeometrie. Bei der Anwendung in massiv-parallelen Simulationen ist dies sehr praktisch und effizient, da das Load Balancing im Laufe der Simulation nicht negativ beeinflusst wird.

Die Leistungsfähigkeit dieser in PsiPhi implementierten Methode, ist anhand von Simulationen aktueller Motorgeometrien, wie zum Beispiel dem Motor an der Universität Duisburg-Essen[6] oder dem an der Technische Universität Darmstadt, erwiesen. Die Skalierbarkeit[7] wurde durch Rechnungen auf hunderten bis mehreren tausend CPUs auf dem Rechencluster Cray-XT6m an der Universität Duisburg-Essen und auf dem Jülicher Supercomputer JUQUEEN (Blue Gene/Q) bewiesen. Bei der Validierung der Methode mit und ohne Verbrennung zeigte sich eine gute Übereinstimmung mit den experimentellen Daten. Die Ausbreitung der Flamme in der Brennkammer sowie der Zylinderdruck und das Geschwindigkeitsfeld wurden in der Simulation gut wiedergegeben.

Simulation der Nanopartikelsynthese

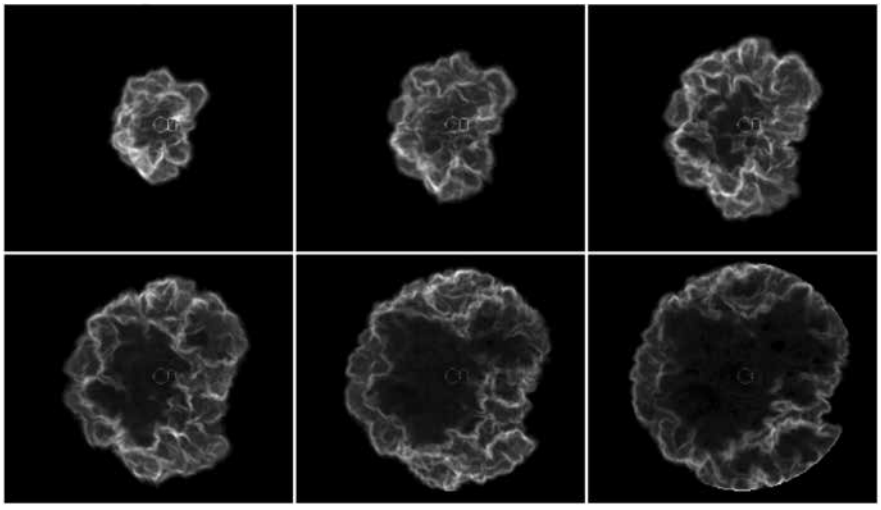
Die chemischen, physikalischen und morphologischen Eigenschaften nanoskaliger Partikel unterscheiden sich teilweise stark von denen ihrer korrespondierenden Festkörper oder größerer Partikel. Zur Kontrolle der Eigenschaften werden daher Partikel mit einer engen Größenverteilung benötigt – diese Nanopartikel werden üblicherweise synthetisch hergestellt. Ein effektives Verfahren zur Herstellung der Partikel ist die Gasphasensynthese. Dabei wird ein Trägerstoff, der sogenannte Präkursor, in flüssiger oder gasförmiger Form in eine heiße Umgebung eingebracht, durch chemische Reaktionen aufgespalten und die Ausgangsmoleküle der Nanopartikel freigesetzt. Darauf folgend entstehen erste Monomer-Partikel durch Nukleation von Molekülen aus der Gasphase. Anschließend erfolgt die Partikeldynamik durch Prozesse wie Konvektion, Diffusion, Thermophorese, Koagulation, Sinterung oder Oberflächenwachstum, die als aerosoldynamische Prozesse zusammengefasst werden. Der dominierende Prozess bei der Partikelsynthese aus der Gasphase ist die Koagulation. Der thermochemische Zustand der Gasphase ist neben der Partikelaufenthaltszeit ein elementarer Prozessparameter zum Einstellen gewünschter Partikeleigenschaften. Neben Experimenten werden Simulationen von Synthesereaktoren genutzt um Strömungs- und Temperaturfelder sowie Gaszusammensetzung zu analysieren und vorherzusagen. Außerdem können durch numerische Methoden experimentell schwer zugängliche Bereiche untersucht werden. Des Weiteren können mit validierten Modellen Parameterstudien durchgeführt werden.

Neben den oben beschriebenen Standardgleichungen für Dichte, Impuls, Spezies und Energie werden weitere Erhaltungsgleichungen benötigt, die die Evolution der Nanopartikel beschreiben. Die aerosoldynamischen Prozesse werden in einer Populations-Bilanz-Gleichung (PBE: „population balance equation“ oder GDE: General Dynamics Equation) zusammengefasst und beschrieben. Die GDE ist eine Inte-

gro-Differentialgleichung und wird in diskrete oder kontinuierliche Form unterteilt. Die GDE beschreibt die Entwicklung einer Partikeleigenschaft, zum Beispiel der Partikelanzahlkonzentration n für Partikel innerhalb eines bestimmten Größenbereichs – zum Beispiel repräsentiert durch das Volumen der Nanopartikel v und v + dv. Um die räumliche (r) und zeitliche (t) Entwicklung einer Partikelgrößenverteilung innerhalb eines Größenspektrums zu beschreiben müssen k GDE gelöst werden: $n_k(r,v,t)$. Hierbei ist die Anzahl der zu lösenden Gleichungen k und die zu berücksichtigenden aerosoldynamischen Prozesse stark vom betrachteten System abhängig. Bei der Synthese von Nanopartikeln aus der Gasphase müssen üblicherweise Partikel kleiner als eins bis hin zu mehreren hundert Nanometern berücksichtigt werden. Allerdings ist die Anzahl der lösbaren Gleichungen durch die numerischen Ressourcen beschränkt. Daher werden diskrete Partikelgrößen in Sektionen zusammengefasst und eine reduzierte Anzahl von GDE für Sektionen anstatt von diskreten Größen gelöst. Eine weitere Vereinfachung ist das Einführen von integralen Größen der GDE, den Momenten der Verteilungsfunktion. Dabei werden Erhaltungsgleichungen beziehungsweise Transportgleichungen für eine stark reduzierte Anzahl an Gleichungen gelöst und anschließend die Größenverteilung aus den Stützgrößen rekonstruiert.

In dem LES und DNS inhouse code PsiPhi wurden ein sektionales und eine modifizierte Version des Kruis Models implementiert und erfolgreich zur Beschreibung der SiO_2-Synthese eingesetzt. In dem untersuchten Synthesereaktor wurde flüssiges Hexamethyldisiloxan (HMDSO) zusammen mit Ethanol durch einen Sauerstoff Dispersionsstrom in die Brennkammer eingebracht. Dabei wurde die Energie einer vorgemischten Methan-Sauerstoff Pilotflamme zur Verdampfung des Sprays sowie zur Zündung und Stabilisierung der Sprayflamme genutzt. In der Simulation wurde für die Spray-, Gas- und Partikel-Phase ein Lagrange-Euler-Euler-Ansatz gewählt[8,9]. Für die Beschreibung der Partikelsynthese wurden neben konvektivem Transport die Nukleation, Koagulation und die Sinterung berücksichtigt. Das eingesetzte modifizierte Kruis-Model berücksichtigte dabei (lokal) die Interaktion von Partikeln gleicher Größe.

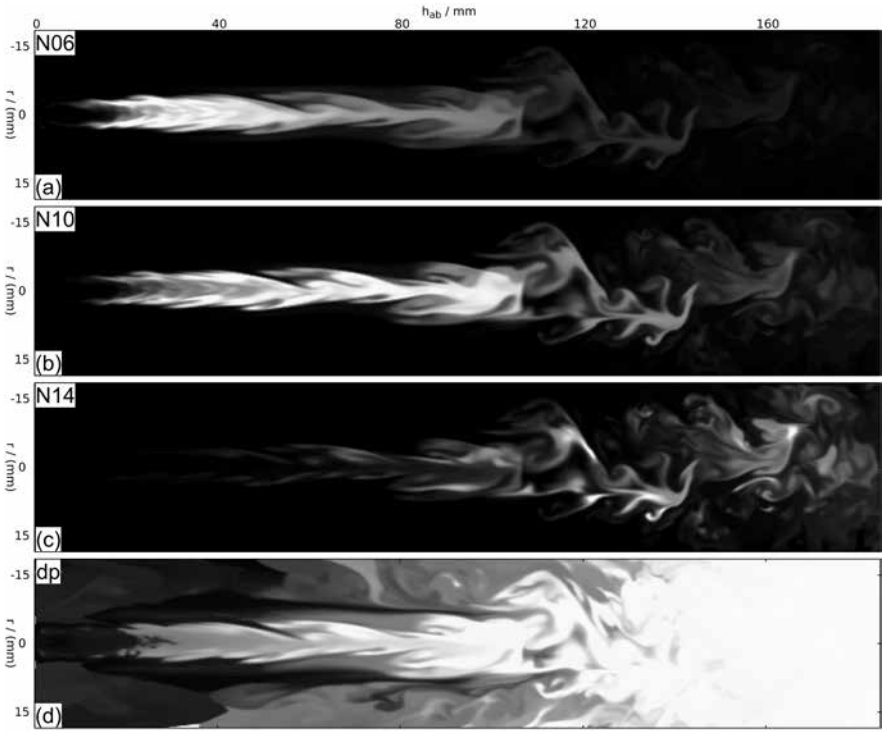
Es konnte gezeigt werden, dass abhängig von der Position in der Brennkammer sowohl größere Aggregate mit kleinen Primärpartikeln, als auch größere gesinterte Primärpartikel zu finden sind. Ein weiteres Ergebnis der Studie war die Erkenntnis, dass die Fluktuationen des Strömungsfeldes und der daraus variierenden Temperatur und Gaszusammensetzung, zu einer breiten Größenverteilung der Partikel führen kann. Eine weitere Studie soll den Einfluss der Turbulenz und gleichzeitig den Einfluss der Polydispersität, das heißt durch Berücksichtigung der Interaktion von Partikeln unterschiedlicher Größe, in Abbildung (5) exemplarisch dargestellt durch drei Anzahlkonzentrationen (a–c) und den resultierenden volumengemittelten Durchmesser (d), untersuchen.



(4) Die Ausbreitung der Flamme in der Brennkammer von Motorsimulation mit 240 CPUs auf dem Rechencluster Cray-XT6m an der Universitat Duisburg-Essen.
Quelle: Nguyen Thuong

Summary

The accurate prediction of fluid flows is still one of the most challenging engineering problems. This is due to the turbulent nature of fluid flows that can only be resolved by numerical simulations for small and simplified problem sizes, even with the massive computational resources that are available today.
When chemical reactions and combustion processes are involved, the computational costs become even larger due to the stiff and non-linear interaction between turbulence and chemistry. Therefore, reduction techniques have been developed in the past, namely the Reynolds averaged Navier-Stokes (RANS) and the large-eddy simulation technique (LES). The latter has the potential to describe the turbulent combustion process more precisely, but still requires models for the unresolved turbulent scales. These models are developed in the group of Prof. Kempf at the University of Duisburg-Essen, where a broad range of applications is covered. Examples are the combustion of coal, the synthesis of nano-particles in flames, and the combustion processes in internal combustion engines. To perform the



(5) Momentaufnahmen der Ergebnisse aus dem sektionalen Modell auf einer 2D-Schnittebene. (a)–(c) zeigen die Evolution der Partikelanzahlkonzentrationen für Partikel mit volumengemittelten Durchmessern von (a) 4nm, (b) 25 nm und (c) 100 nm. Abbildung (d) zeigt den resultierenden volumengemittelten Partikeldurchmesser.
Quelle: Andreas Rittler

required simulations on the largest supercomputers in the world, the group has developed its own inhouse code that is suitable for massively parallel simulations.

Anmerkungen/Literatur

1) A. M. Kempf, B. Geurts, and J. C. Oefelein (2011): Error Analysis of Large-Eddy Simulation of the Turbulent Non-premixed Sydney Bluff-Body Flame, Combustion and Flame 158, 2408–2419.
2) F. Proch, A. M. Kempf (2014): Numerical analysis of the Cambridge stratified flame series using artificial thickened flame LES with tabulated premixed flame chemistry, Combustion and Flame, 161, 2627–2646.
3) F. Proch, A. M. Kempf (2015): Modeling heat loss effects in the large eddy simulation of a model gas turbine combustor with premixed flamelet generated manifolds, Proceedings of the Combustion Institute, 35, 3337–3345.
4) M. Rieth, F. Proch, M. Rabaçal, B. M. Franchetti, F. Cavallo Marincola, A. M. Kempf (2016): Flamelet LES of a semi-industrial pulverized coal furnace, Combustion and Flame, 173, 39–56.
5) M. Rieth, F. Proch, A. G. Clements, M. Rabaçal, A. M. Kempf, Highly resolved flamelet LES of a semi-industrial scale coal furnace, Proceedings of the Combustion Institute (in press).
6) T. M. Nguyen, P. Janas, T. Lucchini, G. D'Errico, S. Kaiser, A. M. Kempf (2014): LES of Flow Processes in an SI Engine Using Two Approaches: OpenFoam and PsiPhi, SAE Technical Paper, 2014-01-1121.
7) T. M. Nguyen, F. Proch, I. Wlokas, A. M. Kempf (2016): Large Eddy Simulation of an Internal Combustion Engine Using an Efficient Immersed Boundary Technique, Flow, Turbulence and Combustion, 97, 191–230.
8) A. Rittler, F. Proch, A. M. Kempf (2015): LES of the Sydney piloted spray flame series with the PFGM/ATF approach and different sub-filter models, Combustion and Flame 162, 1575–1598.
9) A. Rittler, L. Deng, I. Wlokas, A. M. Kempf, Large eddy simulations of nanoparticle synthesis from flame spray pyrolysis, Proceedings of the Combustion Institute (in press)

Die Autoren

Andreas Kempf stammt aus Würzburg und studierte allgemeinen Maschinenbau an der TU-Darmstadt, wo er nach einem Auslandsaufenthalt auch als Wissenschaftlicher Mitarbeiter arbeitete und 2003 im Gebiet der Grobstruktursimulation (LES) turbulenter Flammen promovierte. Im Jahr 2004 erhielt er einen Ruf als Assistenz-Professor an das Imperial College London, wo er neue Simulationsmodelle, Methoden und Programme für die Simulation turbulenter, reagierender Mehrphasenströmungen entwickelte. Im Jahr 2011 folgte Andreas Kempf einem Ruf an die Universität Duisburg-Essen, wo er den Lehrstuhl Fluiddynamik leitet und dem Center for Computational Sciences and Simulation (CCSS) vorsteht. Im Jahr 2016 wurde schließlich der von ihm und Prof. Schröder federführend beantragte Hochleistungsrechner MagnitUDE in Duisburg in Betrieb genommen.

Fabian Proch stammt aus Leichlingen und hat Maschinenbau an der Universität Duisburg-Essen studiert. Seit November 2011 ist er am Lehrstuhl für Fluidmechanik als wissenschaftlicher Mitarbeiter beschäftigt und untersucht die Modellierung von partiell- und vollständig vorgemischten Flammen. Dazu verwendet er hochaufgelöste LES und direkte numerische Simulation (DNS), welche er auf massiv parallelen Hochleistungsrechner durchführt.

Andreas Rittler stammt aus Heidelberg und hat nach seiner Berufsausbildung zum Technischen Zeichner von 2006 bis 2011 Maschinenbau an der Hochschule Mannheim studiert. Seit November 2011 ist er wissenschaftlicher Mitarbeiter am Lehrstuhl für Fluiddynamik und beschäftigt sich mit der Modellierung und numerischen Untersuchung von reaktiven Mehrphasenströhmungen. Der Schwerpunkt seiner Forschung ist die LES der Nanopartikelsynthese aus Sprayflammen.

Martin Rieth stammt aus München und hat an der Universität Duisburg-Essen Wirtschaftsingenieurwesen und Maschinenbau studiert. Seit dem Abschluss des Studiums mit einem Projekt an der University of California, Berkley, ist er wissenschaftlicher Mitarbeiter am Lehrstuhl für Fluiddynamik und beschäftigt sich mit der Modellierung und Simulation von turbulenten Strömungen. Seine Forschungsschwerpunkte liegen auf der LES von Kohlestaubfeuerung sowie der stochastischen Modellierung der Chemie-Turbulenz Interaktion für die Grobstruktursimulation.

Thuong Nguyen hat von 2000 bis 2005 seinen Bachelor in der Informationstechnik für Bauingenieurwissenschaften an der Hanoi University of Civil Engineering Vietnam gemacht. Nach Tätigkeiten im Bauingenieurwesen und in der Entwicklung für Finanzinformationsysteme folgte von 2009 bis 2011 ein Masterstudium an der Ruhr Universität Bochum. Seit November 2011 promoviert er am Lehrstuhl für Fluiddynamik an der Universität Duisburg-Essen. Seine Forschung konzentriert sich auf die numerische Simulationen von Verbrennungsmotoren mittels LES.

Fabian Proch. Foto: Vladimir Unkovic

Mit Hilfe der Dichtefunktionaltheorie sind wir in der Lage, den Zusammenhang zwischen der atomaren Struktur und den Eigenschaften eines Materials zu erforschen, ohne dass empirisch gewonnene Kenntnisse als Ausgangspunkt benötigt werden. Der Einsatz von High-Performance Computing ermöglicht es uns, diesen parameterfreien wissenschaftlichen Zugang sogar auf komplexe Systeme, zum Beispiel Grenzflächen zwischen zwei Materialien, auszudehnen.

Spinning round

High-Performance Computing für maßgeschneiderte Materialien für die Informationstechnologie

Von Peter Kratzer

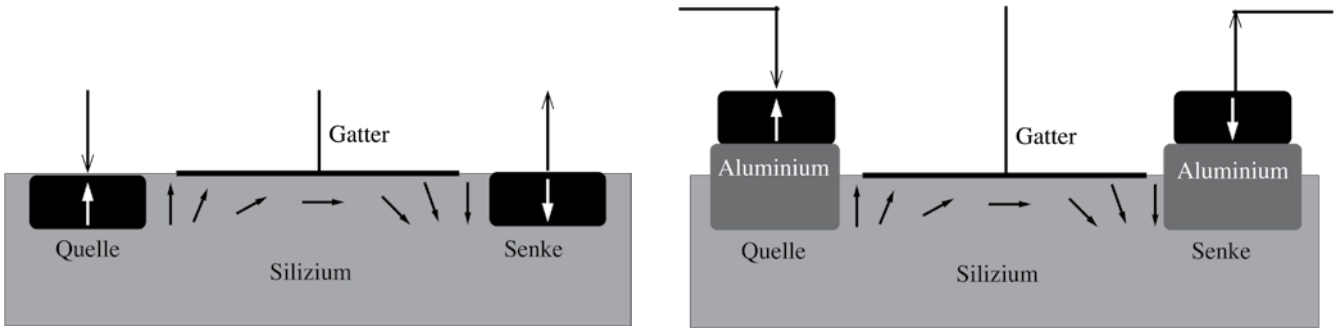
Materialaspekte der Spinelektronik

Die elektronischen Geräte zur Informationsverarbeitung, die wir heute benutzen, arbeiten mit elektrischen Strömen und Spannungen. Das bedeutet, die physikalische Größe, die der Speicherung und Weiterverarbeitung der Information dient, ist die Ladung der Elektronen.

In der digitalen Informationsverarbeitung wird heute die Informationseinheit, das Bit, durch die Ladung einiger hundert bis einiger tausend Elektronen dargestellt. Man unterscheidet zudem zwischen Datenspeichern zur kurzzeitigen Pufferung von anfallenden Daten, die durch elektronische Schaltungen realisiert sind, und Langzeit-Datenspeichern, für die magnetische Speichermedien zum Einsatz kommen.

Die Quantentheorie lehrt uns, dass ein Elektron außer seiner Ladung noch ein weiteres Merkmal, den Eigendrehimpuls (Spin) um eine frei wählbare Achse aufweist. Dieser Spin hat zwei Einstellungsmöglichkeiten, die im Folgenden mit den Begriffen „nach oben“ und „nach unten“ angesprochen werden. Der Spin der Elektronen trägt wesentlich zur Magnetisierung bei, die man in den magnetischen Speichermedien ausnutzt; bei der Verarbeitung der Information macht man aber bisher noch kaum Gebrauch davon. In der materialphysikalischen Grundlagenforschung ist es aktuell eines der Hauptziele, den Spin der Elektronen nicht nur zur Speicherung, sondern darüber hinaus auch zur Verarbeitung von Daten nutzbar zu machen. Dieses Forschungsgebiet trägt den

Peter Kratzer. Foto: Vladimir Unkovic



(a) (b)

(1) Schema eines Spintransistors im geöffneten Zustand (a) wie von Datta und Das vorgeschlagen, und (b) modifiziert durch den Autor. Die schwarzen Pfeile zeigen die Ausrichtung des Spins der Elektronen, die weißen Pfeile zeigen die Magnetisierung von Quelle und Senke.
Quelle: eigene Zeichnung

Namen Spinelektronik oder (auf Englisch) *spintronics*. Man erhofft sich davon eine Reihe von Vorteilen. Zu den wichtigsten zählen ein geringerer Energieverbrauch der Computerchips, sowie die Möglichkeit der Integration von magnetischen und elektrischen Funktionen auf einem einzigen Chip. Durch die letztere Funktion könnten Daten (z.B. durch spezielle magnetische Bereiche) permanent auf dem Chip gespeichert werden und man müsste keinen Datenverlust bei Ausfall der Spannungsversorgung befürchten. Das bedeutet, das elektronische Gerät wird weniger abhängig von Batterien beziehungsweise kann nach einem Ruhezustand schnell wieder starten. Zur Realisierung dieser Technologie muss man in der Lage sein die magnetischen Bereiche auf dem Chip durch ein elektronisches Signal auszulesen.

Das Schlüsselmaterial für die heutige Mikroelektronik ist das Silizium. Silizium ist ein Halbleiter, es ist reichlich vorhanden, sehr beständig gegen Hitze und Umwelteinflüsse und dank jahrzehntelanger Erfahrung von Technologen ist man in der Lage, dieses Material mit ausgefeilten Techniken zu bearbeiten und kleinste Strukturen daraus herzustellen. Aus all diesen Gründen wird Silizium auch in Zukunft das Material der Wahl für elektronische Bauelemente bleiben. Wie können wir nun magnetische Funktionalität in die Silizium-Technologie integrieren? Hier stellen sich zahlreiche Herausforderungen für die Materialforschung. In meiner Arbeitsgruppe berechnen wir Materialeigenschaften auf der Basis grundlegender physikalischer Theorien, und leisten damit einen Beitrag, geeignete Materialklassen für die Spinelektronik zu identifizieren.

Auf welche Eigenschaften der Materialien kommt es uns dabei an? Zur Beantwortung dieser Frage muss man etwas näher auf die beabsichtigten Anwendungen eingehen. Bei der Informationsverarbeitung könnte ein zukünftiges Bauelement zum Beispiel der Spintransistor sein.

Einem Transistor liegt die Idee zu Grunde, dass er durch eine kleine, am Gatter angelegte elektrische Spannung von einem geschlossenen Zustand in einem offenen Zustand (und zurück) umgeschaltet werden kann. Der offene und geschlossene Zustand müssen sich durch ihren jeweiligen elektrischen Widerstand deutlich unterscheiden, und gerade in Hinblick auf Energieersparnis ist man an einer Verringerung der Gatterspannung bei zukünftigen Bauelementen interessiert. Hier kann uns der Spin zu Hilfe kommen: Beim Durchlaufen des Transistors dreht sich der Spin infolge der angelegten Schaltspannung. Der Transistor enthält einen Bereich, der als Quelle, und einen Bereich, der als Senke für die Elektronen fungiert. Die Funktionsweise des Spintransistors beruht auf folgender Idee: Quelle und Senke sollen unterschiedliche Magnetisierung haben. Nur wenn der Spin eines Elektrons sich beim Durchlaufen gerade von oben nach unten gedreht hat, kann das Elektron seinen Weg in die Senke finden, und der Stromfluss ist dann besonders groß. Die erforderliche Drehung des Spins erzeugt einen zusätzlichen Kontrastmechanismus zwischen „offen“ und „geschlossen“.

Deshalb kann eine ziemlich kleine Schaltspannung bereits ausreichend sein, um eine klare Unterscheidbarkeit der Zustände „offen“ und „geschlossen“ zu erreichen.

Die materialwissenschaftliche Herausforderung besteht unter anderem in der Herstellung der Bereiche für Quelle und Senke. Diese sollen magnetisch sein; es erweist sich aber als recht schwierig, magnetische Materialen zu finden, die mit Silizium verträglich sind. Die Materialien müssen hierzu so präzise verarbeitet werden können, dass man Bereiche herstellen kann, die nur einige zehn Nanometer groß sind und scharf gegen benachbarte Bereiche abgegrenzt sind. Das bedeutet, die Materialeigenschaften müssen sich auf einer Strecke von nur wenigen Atomdurchmessern abrupt ändern. Magnetische Materialien, die wir aus unserem Alltag kennen, sind die Metalle Eisen, Kobalt und Nickel sowie davon

abgeleitete Legierungen. In der Fachsprache werden sie als ferromagnetisch bezeichnet. Leider ist Silizium chemisch sehr reaktiv gegenüber den drei genannten Metallen. Wenn man sie zum Beispiel durch Vakuumverdampfung auf einer Siliziumoberfläche aufbringt, kommt es zu einer chemischen Reaktion, und es entstehen nicht-magnetische Silizium-Metall-Verbindungen (Silizide). Als weiterer Nachteil kommt hinzu, dass sich infolge der Heftigkeit der chemischen Reaktion keine atomar scharfe Grenzfläche zwischen Metall und Silizium ausbildet. Die Grenzfläche ist entweder auf atomarer Skala rau, oder, wenn man bei höheren Temperaturen arbeitet, dann lösen sich die Metallatome im Silizium-Festkörper auf. In keinem der beiden Fälle kann man klar definierte Bereiche für die Quelle und Senke der Elektronen durch einfaches Verdampfen herstellen.

Als Ausweg haben wir daher die Möglichkeit verfolgt, das Metall Mangan anstelle von Eisen oder Kobalt zu verwenden. Mangan ist – anders als Eisen – nicht ferromagnetisch, seine chemischen Verbindungen mit Silizium (Mangansilizid, MnSi) sind jedoch in einigen Fällen ferromagnetisch.

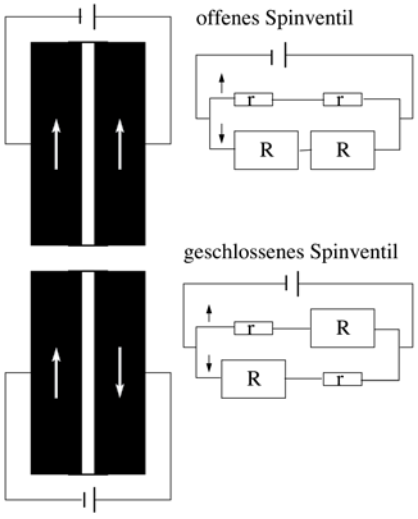
Die chemische Reaktion an der Grenzfläche muss also gar kein Nachteil sein, sondern kann gerade dazu ausgenutzt werden, ein magnetisches Material zu erhalten. Da Mangan außerdem weniger reaktiv ist als Eisen oder Kobalt, ist es etwas einfacher, eine atomar scharfe Grenzfläche herzustellen, wenn man einen geeigneten Temperaturbereich für das Aufdampfen des Mangans findet. Unser Ziel als theoretisch arbeitende Materialforscher war es, die Grenzfläche MnSi-Silizium durch unsere Berechnungen genauer zu charakterisieren und ihre magnetischen und elektronische Eigenschaften herauszufinden.

Etwas andere Aufgabenstellungen für den*die Materialforscher*in ergeben sich, wenn wir uns das elektronische Auslesen eines Magnetspeichers als Ziel setzen. Jedes ferromagnetische Material kann nur unterhalb einer bestimmten, materialspezifischen Temperatur magnetisiert werden. Wenn man das Material über diese sogenannte Curie-Temperatur erwärmt, geht die langreichweitige Ordnung der Spins und damit die messbare Magnetisierung verloren. Ein magnetisches Material, das für die magnetische Datenspeicherung eingesetzt werden soll, muss eine möglichst hohe Curie-Temperatur – weit oberhalb der Zimmertemperatur – aufweisen, damit die Daten vor Verlust geschützt sind. (Dies ist eine Mindestanforderung, eine weitere Anforderung an das Material ist eine hohe magnetische Anisotropie.) Beide Kriterien bedeuten eine erhebliche Einschränkung in der Materialauswahl. Eventuell müsste man, um die Kriterien erfüllen zu können, auf die Verträglichkeit des Speichermaterials mit Silizium verzichten und stattdessen das Silizium durch Aufbringen einer Trennschicht vor dem direkten Kontakt mit dem magnetischen Material schützen.

Für die Spinelektronik möchten wir nun den Magnetspeicher elektronisch auslesen. Dazu müssen wir eine elektrische Spannung an dem magnetischen Speichermaterial anlegen, entweder direkt, zum Beispiel durch Zuleitungen, oder auf indirektem Wege, beispielsweise durch einen Temperaturunterschied verursacht (siehe weiter unten). Der elektrische Widerstand, den die Elektronen beim Durchfließen des magnetischen Materials erfahren, sollte dann vom Magnetisierungszustand abhängen.

Eine solche Abhängigkeit heißt in der Fachsprache Magnetowiderstandseffekt. Meistens ist dieser Effekt recht klein; der Widerstand unterscheidet sich nur um einige Prozent zwischen unterschiedlichen Magnetisierungsrichtungen des Materials. Trotzdem kann dieser kleine Effekt aber schon ausreichend sein, um unterscheiden zu können, welche Information in dem Magnetisierungszustand des Materials gespeichert ist. Es gibt aber auch Materialien und bestimmte geometrische Anordnungen, in denen der Magnetowiderstandseffekt sehr groß werden kann. Das erleichtert dann natürlich das elektronische Auslesen erheblich. Aus Sicht der Materialphysiker*innen kann man dazu folgende Überlegungen anstellen: Man weiß, dass nur Elektronen einer ganz bestimmten Energie zum Stromfluss in einem Metall beitragen.

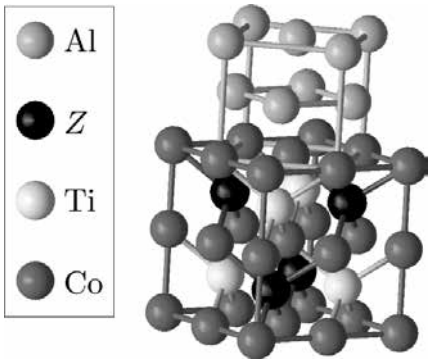
Bei ganz bestimmten Materialien können die Elektronen genau bei dieser Energie nur eine Richtung des Spins aufweisen. Das bedeutet: In einem solchen Material mit vorgegebener Magnetisierung können nur Elektronen, deren Spin in die gleiche Richtung wie diese zeigt, zum Stromtransport beitragen.



(2) Schematischer Aufbau eines magnetischen Speichers (links) aus einer Doppelschicht magnetischer Materialien, getrennt durch eine dünne Barriere. Die Ersatzschaltbilder (rechts) machen deutlich, dass man das Bauelement näherungsweise durch die Parallelschaltung zweier Kanäle darstellen kann, die für die beiden Ausrichtungen des Elektronenspins stehen. Die elektrischen Widerstände *R* und *r* (große oder kleine Rechtecke) hängen davon ab, ob der Spin (schwarze Pfeile) der den Strom tragenden Elektronen parallel oder antiparallel zur Magnetisierungsrichtung (weiße Pfeile) ist. Die unterschiedliche relative Ausrichtung der beiden Magnetisierungen (oben und unten) codiert das gespeicherte Bit.
Quelle: eigene Zeichnung

Betrachten wir nun eine Doppelschicht aus zwei magnetischen Lagen, die in der gleichen oder in entgegengesetzter Richtung magnetisiert sein können (siehe Abb. 2). Im letzteren Fall würden wir erwarten, dass diese Doppelschicht-Anordnung überhaupt keinen Stromfluss zulässt. Denn die Elektronen können die Bedingung, dass ihr Spin gleichgerichtet mit der Magnetisierung sein soll, nur in einer der Schichten, nicht aber in beiden erfüllen. Der elektrische Widerstand wäre also in diesem Fall unendlich. Wenn die beiden Schichten jedoch gleichgerichtet magnetisiert sind, steht nichts dem Stromtransport im Wege, auch wenn wir den Spin der Elektronen berücksichtigen. Die Materialanordnung hat dann nur einen kleinen elektrischen Widerstand *r*. Man kann die beiden Fälle also klar unterscheiden, und darauf kommt es uns an, wenn das elektrische Auslesen der Magnetisierung erfolgreich sein soll. Hier ist anzumerken, dass die Anordnung mit dem unendlich hohen elektrischen Widerstand natürlich eine Idealisierung darstellt. In Wirklichkeit findet man einen hohen, aber endlichen Widerstand *R*, da es, insbesondere bei erhöhter Temperatur, stets Prozesse geben wird, die doch noch einen kleinen Stromfluss ermöglichen. Das Prinzip des elektrischen Auslesens erfordert nur einen klaren Widerstandsunterschied zwischen den Magnetisierungsrichtungen, und dieser ist auch bei erhöhter Temperatur gewährleistet, sofern wir nur genügend weit unter der Curie-Temperatur des Materials bleiben.

Aus materialphysikalischer Sicht müssen wir uns zunächst die Frage stellen: Was sind das für Materialien, bei denen ein solch enger Zusammenhang zwischen der Richtung der Magnetisierung und der Richtung des Spins der stromtragenden Elektronen besteht? Die angesprochenen Materialien werden in der Literatur als ferromagnetische Halbmetalle bezeichnet. Es gibt ganz unterschiedliche Möglichkeiten, wie man



(3) Die Kristallstruktur der Heusler-Legierung (unterer Bildteil) ist kompatibel mit der Kristallstruktur von Aluminium (hellgraue Atome im oberen Bildteil). *Z* kann für Si oder Ge stehen.
Quelle: entspricht Fig. 1a in Benjamin Geisler, Peter Kratzer und Voicu Popescu, Interplay of growth mode and thermally induced spin accumulation in epitaxial Al/Co_2TiSi/Al and Al/Co_2TiGe/Al contacts

ein solches ferromagnetisches Halbmetall herstellen kann. In unseren theoretischen Untersuchungen haben wir uns auf eine ganz bestimmte Materialklasse konzentriert, nämlich intermetallische Verbindungen von Kobalt (Elementsymbol: Co). Unter einer intermetallischen Verbindung versteht man den Spezialfall einer Metalllegierung, bei der ein hohes Maß von atomarer Ordnung herrscht. Die unterschiedlichen Atom-Sorten, aus denen die Legierung besteht, sitzen dabei auf jeweils festen Positionen eines Kristallgitters. Wir untersuchten die intermetallischen Verbindungen Co_2MnSi, Co_2TiSi und Co_2TiGe. Bei diesen Verbindungen sorgen die energetischen Verhältnisse der Kristallbildung dafür, dass die Kobalt-Atome in bestimmten Kristallebenen sitzen, die sich immer mit anderen, schachbrettartig belegten Ebenen abwechseln (s. Abb. 3). Das Schachbrettmuster in den letzteren Ebenen wird von den Silizium- sowie den Mangan- (im Falle von Co_2MnSi) beziehungsweise Titan-Atomen (im Falle von Co_2TiSi) gebildet. Diese spezielle Anordnung der Atome wird nach ihrem Entdecker als Heusler-Legierung bezeichnet. In den genannten Beispielen (Co_2MnSi, Co_2TiSi und Co_2TiGe) sowie in einigen weiteren Heusler-Legierungen hat die spezielle Anordnung der Atome zur Folge, dass ein ferromagnetisches Halbmetall entsteht. Das bedeutet: Wenn wir idealisieren und von einem räumlich perfekt geordneten Atom-Muster ausgehen, dann tragen in diesen Materialen nur die Elektronen mit einer der beiden möglichen Spinrichtungen zum elektrischen Transport bei.

Die besondere Fragestellung, die im Mittelpunkt unserer Berechnungen stand, betrifft die Eigenschaften der Heusler-Legierungen an Grenzflächen. Wir untersuchten einerseits den Kontakt zwischen Co_2TiSi und einem Metall, nämlich Aluminium. Andererseits betrachteten wir Grenzflächen von Co_2MnSi mit einem elektrisch isolierenden Material, Magnesiumoxid. In beiden Fällen konnten wir interessante Auswirkungen der Beschaffenheit der Grenzfläche auf die Materialeigenschaften feststellen. Generell ist der Zusammenhang zwischen Struktur und Eigenschaften ein Thema, bei dem wir die Stärken unserer theoretischen Herangehensweise ausspielen können. In der experimentellen Forschung wird es (jedenfalls zu Beginn) schwierig sein, Proben mit nur einer, genau definierten Grenzflächenbeschaffenheit zu präparieren, während dies in einer Computersimulation einfach möglich ist. Die Ergebnisse unserer Simulationen können deshalb den Experimentatoren Hinweise geben, welche Art von Grenzfläche sich (zunächst einmal unbeabsichtigt) bei der Herstellung vorwiegend ausbildet, um dann später durch Verfeinerung der Präparationsmethoden spezielle Grenzflächen herstellen zu können, die den angestrebten Eigenschaften förderlich sind.

Methodik der Berechnungen

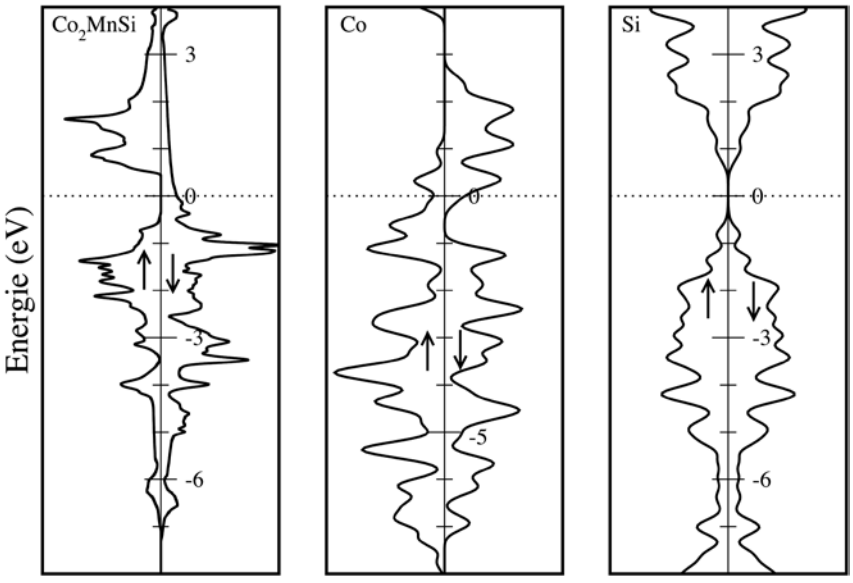
Für ein vertieftes Verständnis magnetischer Materialien und insbesondere für die Eigenschaften von Grenzflächen ist es notwendig, dass wir auf die physikalischen Grundla-

gen zurückgehen und ausgehend von ersten Prinzipien, das heißt ohne Zuhilfenahme empirischer Kenntnisse, diese Eigenschaften berechnen. Die Quantentheorie sagt uns, dass die negativ geladenen Elektronen von den positiv geladenen Atomkernen angezogen werden. Der bevorzugte Aufenthalt einiger Elektronen im Raumbereich zwischen den Atomkernen benachbarter Atome bringt die chemischen Bindungen hervor, die einen Kristall (oder allgemeiner auch ein nicht-kristallines Material) zusammenhalten. Andererseits ist natürlich auch die elektrische Abstoßung zwischen den gleichnamigen Ladungen der Elektronen zu berücksichtigen. Die damit zusammenhängenden Effekte können wir im Rahmen der Dichtefunktionaltheorie behandeln. Dies ist eine Theorie, die es uns gestattet, das äußerst schwierige quantenmechanische Vielteilchenproblem auf eine einfachere Problemstellung für einzelne, unabhängige Teilchen zurückzuführen. Diese Teilchen bewegen sich in einem effektiven Potenzial $V_{\text{eff}}(\vec{r})$, das im Fortgang der Rechnung konsistent zu bestimmen ist. Konkret benötigen wir die Energien $\varepsilon_{n,\sigma}(\vec{k})$ sowie die quantenmechanischen Wellenfunktionen $\psi_{n,\sigma,\vec{k}}(\vec{r})$ einer bestimmten Anzahl *N* von Elektronen. Der erste Index *n* zählt die Energiebänder im Festkörper. Die beiden Möglichkeiten, wie sich der Spin bezüglich einer vorgegebenen Achse einstellen kann, führen zu einem weiteren Index *σ* der Energien und Wellenfunktionen. Dieser Index kann zwei Werte annehmen, das heißt durch die Berücksichtigung des Spins verdoppelt sich die Komplexität der mathematischen Aufgabe. Der dritte Index *k* bezeichnet den Kristall-Impuls, also grob gesprochen die Richtung, entlang der die Elektronen durch den Kristall fliegen. Die Wellenfunktionen $\psi_{n,\sigma,\vec{k}}(\vec{r})$ der (als unabhängig gedachten) Elektronen müssen die sogenannte Kohn-Sham-Gleichung erfüllen,

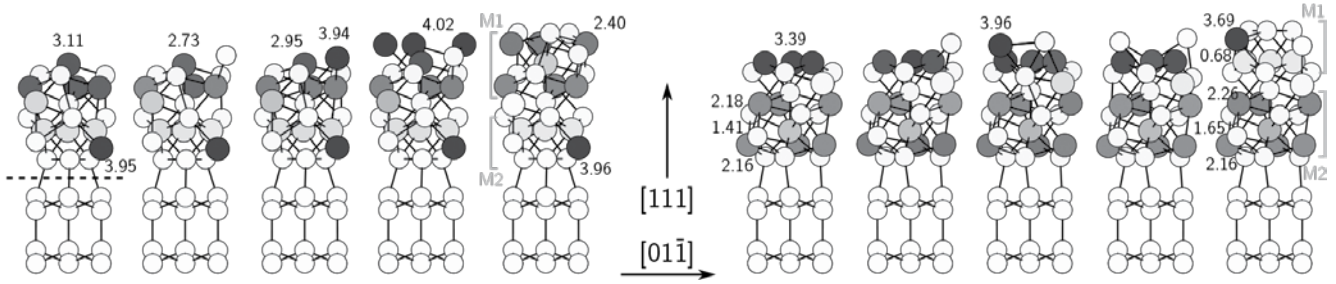
$$\left(\frac{-\hbar^2}{2m}\Delta + V_{\text{eff}}(\vec{r})\right)\psi_{n,\sigma,\vec{k}}(\vec{r}) = \varepsilon_{n,\sigma}\psi_{n,\sigma,\vec{k}}(\vec{r})$$

Aus mathematischer Sicht handelt es sich dabei um ein System von partiellen Differentialgleichungen. Wir benötigen *N* verschiedene Lösungen, wenn das Material oder die betrachtete Schichtstruktur *N* Elektronen enthält. Die numerische Behandlung erfolgt durch Diskretisierung der Differentialgleichung, das heißt, wir berechnen die Werte der Wellenfunktionen auf einem regelmäßigen Punktgitter im Raum. Da es sich um lineare Differentialgleichungen handelt, können besonders effiziente numerische Methoden, zum Beispiel die Methode der schnellen Fouriertransformation, eingesetzt werden. Die Computerprogramme, die zur numerischen Lösung verwendet werden, können parallel auf vielen Prozessoren ausgeführt werden. Da wir *N* Lösungen benötigen, können die Berechnungsaufgaben so verteilt werden, dass jeder Prozessor eine Lösung berechnet. Zugleich kann auch das Punktgitter, das die Wellenfunktionswerte trägt, auf mehrere Prozessoren verteilt werden. Da sich das Potenzial $V_{\text{eff}}(\vec{r})$ in den Kohn-Sham-Gleichungen selbst von der Elektronendichte, und damit indirekt auch von den Wellenfunktionen abhängt, muss das Problem schrittweise iterativ gelöst werden. An Grenzflächen darf man nicht voraussetzen, dass die Atome ihre Positionen auf dem Kristallgitter beibehalten; vielmehr müssen die Positionen der Atome während des Programmlaufs so lange optimiert werden, bis alle Kräfte verschwunden sind. Dies ist ein weiterer Einflussfaktor, der die Anzahl der benötigten Rechenoperationen in die Höhe treibt.

Unsere spezielle Aufgabenstellung, die Suche nach maßgeschneiderten Materialen, bringt noch eine weitere Steigerung der benötigten Rechenzeit mit sich: Wir möchten ja nicht nur eine einzige Anordnung von Atomen betrachten, sondern möchten verschiedene Materialien und Bauelement-Geometrien, das heißt sehr viele verschiedene Anordnungen von Atomen, auf dem Com-



(4) Elektronische Zustandsdichte des ferromagnetischen Halbmetalls Co_2MnSi (links), des magnetischen Metalls Kobalt (Mitte) und des Halbleiters Silizium (rechts). Die nach rechts beziehungsweise nach links aufgetragenen Kurven entsprechen den beiden Spinrichtungen „nach oben“ beziehungsweise „nach unten“. Die Energie derjenigen Elektronen, die zum Stromfluss beitragen, wurde zu Null gesetzt. Man findet, dass der Halbleiter Silizium gar keine Zustandsdichte bei Energie Null aufweist, während das ferromagnetische Halbmetall Co_2MnSi nur für eine Spinorientierung über elektronische Zustände verfügt.
Quelle: eigene Zeichnung



(5) Atomare Struktur unterschiedlich dicker MnSi-Filme auf Silizium. Die Graustufe der Atome und die Zahlenwerte bezeichnen die magnetischen Momente der Mangan-Atome in Einheiten des Bohr-Magnetons.
Quelle: ist Teil von Fig. 7 in Benjamin Geisler und Peter Kratzer, Strain stabilization and thickness dependence of magnetism in epitaxial transition metal monosilicide thin films on Si(111), Phys. Rev. B 88, 115433 (2013)

puter „durchspielen“ und ihre Eigenschaften miteinander vergleichen. Für ein derart ambitioniertes Forschungsprogramm ist deshalb der Zugang zu modernen Hochleistungsrechnern essentiell.

Welche Informationen liefern uns nun die am Computer produzierten Daten? Die N Elektronen werden zunächst die N niedrigsten Energieniveaus $\varepsilon_{n,\sigma}(\vec{k})$ besetzen. Unter den Annahmen der Dichtefunktionaltheorie kann man daraus bereits die Magnetisierung am absoluten Nullpunkt der Temperatur ablesen; wir müssen die Elektronen nur nach ihrem Spin unterscheiden: Die Differenz zwischen der Zahl der „nach oben“ und der „nach unten“ zeigenden Spins ist proportional zur Magnetisierung. Weitere, instruktive Informationen aus den Energien der Elektronen erhält man durch Auswerten der spinaufgelösten elektronischen Zustandsdichte. Dazu teilt man die Energieachse in ein Raster aus Intervallen ein und betrachtet, wie viele der Energien $\varepsilon_{n,\sigma}(\vec{k})$ in ein bestimmtes Intervall fallen. Ein Beispiel ist in Abbildung (4) gezeigt. Die Zustandsdichte, zusammen mit der Information, welche Zustände besetzt sind, ermöglicht uns sofort, das Rechenergebnis mit einer bestimmten Materialklasse zu identifizieren. In einem Halbleiter, wie beispielsweise Silizium, gibt es eine Energielücke (Abb. 4, rechts). Mit anderen Worten: Zwischen den besetzten und den unbesetzten elektronischen Zuständen gibt es einen Energiebereich, in dem keine Lösungen der Kohn-Sham-Gleichung existieren, in dem also die Zustandsdichte gleich Null ist. In einem nicht-magnetischen Material wie Silizium ist die Zustandsdichte der „nach oben“ und der „nach unten“ weisenden Elektronenspins gleich; die Auftragung der Zustandsdichte zeigt also ein symmetrisches Bild. Falls das Material ferromagnetisch ist, wie zum Beispiel Kobalt (Abb. 4, Mitte), so fällt diese Symmetrie weg. Ein ferromagnetisches Halbmetall, wie beispielsweise die oben erwähnte Heusler-Legierung Co_2MnSi, erkennt man daran, dass die Zustandsdichte für den einen Spin eine Energielücke zeigt, während das für die gegenläufige Spinrichtung nicht der Fall ist (Abb. 4, links). Das ferromagnetische Halbmetall kann daher, vereinfachend gesprochen, in seinem elektrischen Verhalten gewisse Merkmale eines Halbleiters und eines Metalls miteinander verbinden.

Die Berechnungsmethode ist auch in der Lage, Aussagen über die elektrische Leitfähigkeit der untersuchten Strukturen zu machen. Diese Information ist im Wesentlichen in den Wellenfunktionen der Elektronen enthalten. Im allgemeinen interessiert man sich für den Durchfluss des elektrischen Stromes durch eine Struktur, die aus einer Quelle, einem Transportbereich, und einer Senke besteht. Für die genaue Bestimmung der Leitfähigkeit sind drei getrennte Berechnungen für die Quelle, die Senke und den Transportbereich, sowie die Anwendung der quantenmechanischen Streutheorie erforderlich. Ferner benötigt man nun auch unbesetzte Lösungen der Kohn-Sham-Gleichungen, und muss sehr viele mögliche Geschwindigkeiten und Flugrichtungen der Elektronen in die Berechnungen einbeziehen. Berechnungen der elektrischen Leitfähigkeit sind deshalb noch einmal um einiges aufwändiger als die Berechnung der weiter oben angeführten magnetischen Eigenschaften.

Anwendungen

Dünne Schichten aus Mangansilizid auf Silizium

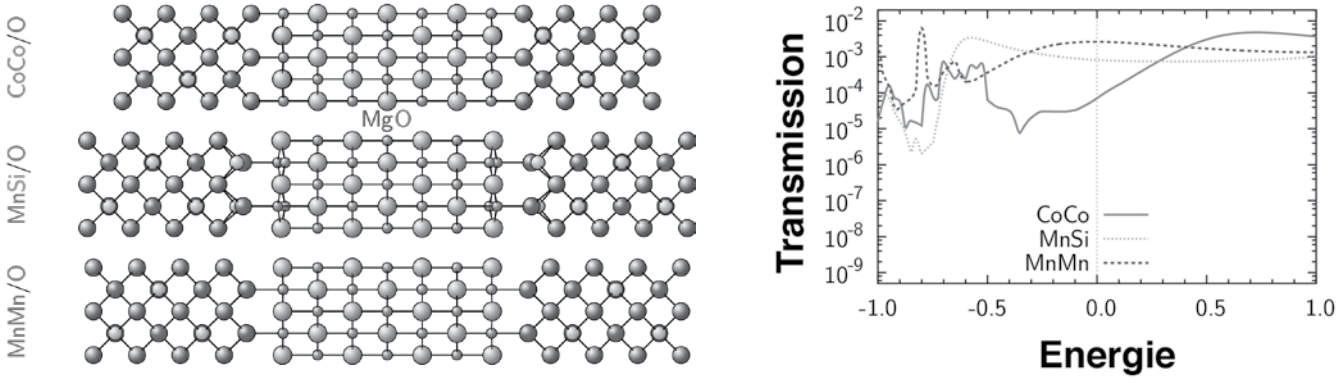
Ein ganz bestimmtes Mangansilizid, das Mangan und Silizium zu gleichen Teilen enthält (chemische Formel MnSi) weist eine Kristallstruktur auf, die sich von der des Siliziums unterscheidet, jedoch entlang einer ausgewählten Wachstumsrichtung mit dem Siliziumkristall verträglich ist. Daher ist es möglich, durch Abscheiden von Mangan und nachfolgende chemische Reaktion dünne MnSi-Filme mit scharf definierter Grenzfläche zum Silizium herzustellen.

Wir haben durch Computersimulationen die Eigenschaften derartiger Filme untersucht. Gewöhnliches MnSi ist unterhalb einer Temperatur von 30 Grad über dem absoluten Nullpunkt ferromagnetisch. Da die MnSi-Filme auf Silizium

einer leichten mechanischen Verspannung ausgesetzt sind, kann sich ihre Curie-Temperatur dadurch bis auf über 40 Grad über dem absoluten Nullpunkt erhöhen, wie unsere Berechnungen nahelegen. Bemerkenswert ist, dass das magnetische Moment, das jedes Mangan-Atom zur Magnetisierung des Films beiträgt, stark davon abhängt, in welcher Atomlage sich das Mangan-Atom befindet (siehe Abb. 5). Die untersuchten MnSi-Filme könnten zum Beispiel in einem Spintransistor Verwendung finden, der bei tiefen Temperaturen arbeitet. Darüber hinaus haben sich kürzlich mehrere andere Forschungsgruppen für MnSi-Filme interessiert, da dieses Material besondere magnetische Wirbel (sog. Skyrmionen) aufweist, von denen man sich interessante neuartige Anwendungen erhofft.

Thermische Spininjektion

Für den Spintransistor (schematische Darstellung in Abbildung 1(a)) ist es eine Grundvoraussetzung, dass die Quelle vorwiegend Elektronen der einen Spinorientierung liefert und die Senke bevorzugt Elektronen der entgegengesetzten Spinorientierung aufnimmt. In dem Halbleitermaterial Silizium ist keine der Spinorientierungen bevorzugt; man muss daher eine ferromagnetische Materialkomponente in die Transistor-Architektur einbauen. Das im obigen Abschnitt untersuchte Mangansilizid hat eine niedrige Curie-Temperatur und ist daher für Anwendungen bei Zimmertemperatur nicht geeignet. Die meisten oberhalb von Zimmertemperatur noch ferromagnetischen Materialien sind Metalle. Man könnte nun einfach ein solches ferromagnetisches Metall mit Silizium zu einem Bauelement kombinieren. Selbst wenn man das schwierige prozesstechnische Problem lösen kann, wie man eine atomare scharfe Grenzfläche der beiden unterschiedlichen Materialien erhalten kann, wird die Effizienz des Spintransistors durch die fehlende Widerstandsanpassung zu wünschen übrig lassen. Damit ist folgendes gemeint: Legen wir an einer Serienschaltung, bestehend aus einem metallischen, einem halbleitenden und wiederum einem metallischen Element von außen eine elektrische Spannung an, so wird der Hauptteil dieser Spannung an dem halbleitenden Element abfallen, da es den größten elektrischen Widerstand aller drei Elemente aufweist. Da die über den ferromagnetischen, metallischen Komponenten abfallende Spannung gering ist, können wir auch nicht erwarten, dass sich die Beiträge der Ströme der Elektronen mit Spin „nach oben“ beziehungsweise „nach unten“ beträchtlich unterscheiden. Genau dies wird aber für einen effizienten Spintransistor benötigt. Es wäre deshalb aussichtsreicher, Materialkomponenten mit ähnlich großen elektrischen Widerständen zu kombinieren, zum Beispiel zwei Metalle. Dies könnte man durch eine Quelle oder Senke erreichen, die aus einer Kombination metallischer Werkstoffe besteht. Ideal wäre die Kombination eines ferromagnetischen Halbmetalls mit einem gewöhnlichen, nichtmagnetischen Metall. Wir haben speziell die Kombination der Heusler-Legierung Co_2TiSi mit Aluminium vorgeschlagen, da die atomaren Kristallgitter beider Materialien gut auf-einander abgestimmt sind (s. Abb. 3) und somit eine atomar glatte Grenzfläche erwarten lassen (wie in Abb. 1b vorausgesetzt). Die Integration von Aluminium-Kontakten in die Silizium-Technologie ist schon lange bekannt und unproblematisch. Die perfekt geordnete Heusler-Legierung würde ausschließlich Elektronen einer Spinrichtung liefern, wäre also die ideale Quelle oder Senke für einen



(6) Atomare Struktur des in Abbildung (2) schematisch gezeigten magnetischen Speichers, bei dem die beiden magnetischen Kontakte (rechts und links) durch eine Magnesiumoxidschicht (Mitte) getrennt sind. Die Transmissionswahrscheinlichkeit, mit der ein Elektron die Oxidbarriere durchdringen kann (ganz rechts) hängt empfindlich davon ab, welche Atomschicht der magnetischen Heusler-Legierung an das Oxid grenzt.
Quelle: entnommen aus Fig. 3 und Fig. 5 in Benjamin Geisler und Peter Kratzer, Spincaloric Transport in Epitaxial $Co_2MnSi/MgO/ Co_2MnSi$ magnetic tunnel junctions, Phys. Rev. B 92, 144418 (2015)

Spintransistor. Da man in realen Materialproben aber die perfekte Ordnung nicht wird erreichen können, schlagen wir auf der Grundlage unserer Berechnungen eine weitere Strategie zur Effizienzsteigerung vor: Anstelle einer äußeren Spannung könnte man eine innere Spannungsquelle verwenden, um den Spintransistor zu betreiben. Dies hätte den Vorteil, dass man das Problem der fehlenden Widerstandsanpassung umgeht. Jede Grenzfläche zwischen zwei Materialien kann als Spannungsquelle dienen, wenn sich die Temperatur über die Grenzfläche hinweg ändert. Dieser Effekt ist generell als thermoelektrischer Effekt bekannt. Durch unsere Berechnungen konnten wir die Größe dieser Thermospannung zwischen Co_2TiSi und Aluminium vorhersagen. Die Spannung für jede der beiden Spinrichtungen der Elektronen ist dabei unterschiedlich. Aufgrund dieses Effekts wird sich in den nichtmagnetischen Aluminiumkontakten, die gewöhnlich keine Spinrichtung bevorzugen, unter dem Einfluss des Temperaturunterschieds doch ein Überschuss von Elektronen einer Spinrichtung einstellen. Diese Beobachtung bezeichnen wir als thermische Spininjektion. Wir kommen durch die Berechnungen zu dem Befund, dass sowohl Größe als auch Richtung der Spininjektion von der atomaren Struktur der Grenzfläche abhängen können.

Elektrisches Auslesen magnetischer Speicher

Eine Möglichkeit, den Inhalt eines magnetischen Speichers durch einen elektrischen Strom auszulesen, bedient sich des Magnetowiderstandseffekts. Dieser Effekt wurde schon in der Einleitung thematisiert und ein passendes Bauelement, das Spinventil, wurde in Abbildung (2) gezeigt. Nun möchte ich genauer auf spezifische Materialen zur Herstellung des Spinventils eingehen. Offensichtlich ist es auch hier von Vorteil, wenn der elektrische Strom vorwiegend von Elektronen nur eines Spins getragen wird. Wenn die beiden Schichten des Spinventils (Abb. 2) in entgegengesetzter Richtung magnetisiert sind, wird ein vollständig spinpolarisierter Strom gar nicht durchgelassen und man findet einen sehr hohen elektrischen Widerstand. Dieser ist klar von dem niedrigen Widerstand im geöffneten Zustand zu unterscheiden. Es empfiehlt sich daher wiederum, ein ferromagnetisches Halbmetall für die beiden Kontakte des Spinventils zu verwenden, um einen möglichst hohen Kontrast zwischen „offen“ und „geschlossen“ zu erhalten. Anders als bei der Spininjektion sind hier die beiden Kontakte durch eine dünne isolierende Schicht, beispielsweise aus Magnesiumoxid, getrennt. Dadurch wird es ermöglicht, nur die Magnetisierung des rechten Kontakts, zum Beispiel durch ein äußeres Magnetfeld, umzudrehen, und auf diese Weise ein Bit in den Speicher zu schreiben. Wenn die isolierende Schicht nur einige Nanometer dick ist, stellt sie kein unüberwindliches Hindernis für den Stromfluss dar, da die Elektronen aufgrund des quantenmechanischen Tunneleffekts durch diese Schicht hindurch gelangen können.

Für den Bau eines Speichers ist es wichtig, dass das verwendete ferromagnetische Halbmetall eine hohe Curie-Temperatur hat. Wir betrachten deshalb nun die Heusler-Legierung Co_2MnSi, die bis über 985 Grad ferromagnetisch bleibt. Besonders wichtig für die Unterscheidbarkeit der beiden Zustände (gleich- oder entgegengerichtete Magnetisierung) und damit für das Auslesen der Information ist der niedrige elektrische Widerstand im „offenen“ Zustand, also bei gleichsinniger Magnetisierung der Kontakte. Deshalb haben wir uns auf diesen Zustand konzentriert. Auch in diesem Fall zeigen unsere Berechnungen den starken Einfluss, den die atomare Struktur der Grenzfläche auf die elektrischen Eigenschaften hat. Wie aus der Auftragung in Abbildung (6) rechts zu sehen, bewirkt eine Anreicherung von Mangan an der Grenzfläche eine besonders hohe Transmissionswahrscheinlichkeit und damit einen niedrigen Widerstand, während bei einer Kobalt-Anreicherung die Durchlässigkeit der Barriere niedrig ist. Diese Kenntnisse kann man bei der Herstellung des magnetischen Speicherelements ausnutzen; eine Grenzschicht mit Mangan-Anreicherung würde bewirken, dass der Kontrast des Stromflusses zwischen offenem und geschlossenem Zustand besonders deutlich ausfällt und somit bequem messbar ist.

Thermisches Auslesen

Die vorgesehene Integration einer großen Anzahl magnetischer Speicherelemente auf einem Chip macht es notwendig, dass jedes einzelne Speicherelement sehr schlank ausgeführt werden muss. Als Folge daraus können sich hohe Stromdichten (Strom pro Querschnittfläche) ergeben, die dem Speicherelement Schaden zufügen können. Wir haben daher über alternative Auslesemöglichkeiten nachgedacht. Anstelle eines Stromes könnte man die elektrische Spannung als Lesesignal verwenden (und dabei den Stromfluss niedrig halten). Die Auslesespannung könnte wiederum in dem Speicherelement selbst durch den thermoelektrischen Effekt erzeugt werden. Dazu haben wir Berechnungen der thermoelektrischen Spannung durchgeführt, die auftritt, wenn sich der linke und rechte Kontakt auf verschiedener Temperatur befinden. Ein Temperaturunterschied von einem Grad führt beispielsweise zu einer Spannung von 65 Mikrovolt, die bereits gut messbar ist. Eine so große Spannung tritt allerdings nur dann auf, wenn die Grenzfläche der Heusler-Legierung zum Magnesiumoxid von einer vollständigen Lage von Kobalt-Atomen gebildet wird. Die Ursache dafür liegt in dem asymmetrischen Verlauf der Transmissions-

kurve um die Energie Null herum, wie in Abbildung (6) rechts zu sehen ist. Diese Vorgabe stellt allerdings hohe Ansprüche an die Präzision bei der Herstellung der Mehrlagenstrukturen.

Zusammenfassung

Mit Hilfe der Dichtefunktionaltheorie sind wir in der Lage, den Zusammenhang zwischen der atomaren Struktur und den Eigenschaften eines Materials zu erforschen, ohne dass empirisch gewonnene Kenntnisse als Ausgangspunkt benötigt werden. Der Einsatz von High-Performance Computing ermöglicht es uns, diesen parameterfreien wissenschaftlichen Zugang sogar auf komplexe Systeme, zum Beispiel Grenzflächen zwischen zwei Materialien, auszudehnen. Die berechneten Eigenschaften umfassen unter anderem die Magnetisierung und den elektrischen Widerstand beim Stromfluss durch die Grenzfläche. Am Computer können wir eine Exploration unterschiedlicher Materialkombinationen durchführen und somit die Eigenschaften im Hinblick auf eine bestimmte Anwendung maßschneidern. Mehrere Anwendungen aus dem Bereich der Spinelektronik wurden vorgestellt: Für den Spintransistor suchen wir nach einem magnetischen Material, das mit Silizium kompatibel ist, und identifizierten das Mangansilizid als einen Kandidaten. Für das Spinventil und die thermische Spininjektion suchten wir nach ferromagnetischen Halbmetallen und konnten Doppellagen-Strukturen aus Heusler-Legierungen identifizieren. In Zukunft wird angestrebt, die Aussagekraft des Verfahrens zu steigern, indem wir auch dynamische Materialeigenschaften einbeziehen, zum Beispiel die Schwingungen der Atome um ihre Gleichgewichtslage oder die Spinwellen, die in magnetischen Materialien auftreten. Beide sind wichtig für die Berechnung der Temperaturabhängigkeit des elektrischen Widerstands.

Summary

Density functional theory puts us in position to investigate the relationship structure-propertty in materials without having to rely on empirical input. The usage of high performance computing enables us to extend this parameter-free approach to complex atomic structures that are encountered for instance at interfaces between two materials. We are thus in position to carry out computer-aided explorative studies of diverse materials combinations, and to tailor the materials properties to specific envisaged applications. Several such applications from the field of spintronics have been presented:
For the spin transistor, a magnetic material compatible with silicon is sought for, and we managed to identify manganese silicide as one candidate. For the spin valve and for thermal spin injection, we explored ferromagnetic half metals and identified double layer structures of Heusler alloys as promising. As a future goal, we aim at enhancing the predictive power of our approach by including dynamical materials properties into our modelling. Examples are oscillations of the atoms around their equilibrium positions as well as spin waves in magnetic materials, both of which have an impact on the temperature dependence of the electrical resistivity.

Literatur

1) Benjamin Geisler und Peter Kratzer, Spin-caloric Transport in Epitaxial Co_2MnSi/MgO/Co_2MnSi magnetic tunnel junctions, Phys. Rev. B 92, 144418 (2015).
2) Benjamin Geisler, Peter Kratzer und Voicu Popescu, Interplay of growth mode and thermally induced spin accumulation in epitaxial Al/Co_2TiSi/Al and Al/Co_2TiGe/Al contacts, Phys. Rev. B 89, 284422 (2014).
3) Benjamin Geisler und Peter Kratzer, Strain stabilization and thickness dependence of magnetism in epitaxial transition metal monosilicide thin films on Si(111), Phys. Rev. B 88, 115433 (2013).
4) Takayuki Suzuki, Theresa Lutz, Benjamin Geisler, Peter Kratzer, Klaus Kern und Giovanni Costantini, Surface morphology of MnSi thin films grown on Si(111), Surf. Sci. 617, 106 (2013).
5) Benjamin Geisler, Peter Kratzer, Takayuki Suzuki,Theresa Lutz, Giovanni Costantini und Klaus Kern, Growth mode and atomic structure of MnSi thin films on Si(111), Phys. Rev. B 86, 115428 (2012).

Der Autor

Peter Kratzer, Jahrgang 1962, promovierte 1993 nach einem Studium der Physik an der Technischen Universität München, wo er sich auch 1998 für das Fach „Theoretische Physik“ habilitierte. Erfahrungen im Ausland sammelte Peter Kratzer von 1993 bis 1995 als Post-Doc am „Center for Atomic-Scale Materials Physics“ in Lyngby, Dänemark, sowie bei einem Gastaufenthalt am Semiconductor Materials Interdisciplinary Research Centre am Imperial College London im Jahre 1998. Mit Materialien für die Informationstechnologie beschäftigte sich Peter Kratzer schon während seine Tätigkeit von 1997 bis 2006 als Arbeitsgruppenleiter in der Abteilung „Theorie“ des Fritz-Haber-Instituts der Max-Planck-Gesellschaft in Berlin. Seit 2006 hat er eine Professur für Theoretische Physik an der Fakultät für Physik der Universität Duisburg-Essen inne und ist Mitglied von CENIDE.

Klaus Pohl. Foto: Vladimir Unkovic

*Persönliche Daten in der Cloud zu speichern und zu verarbeiten, kann ein Risiko sein, wenn die Daten vor Manipulation und unerlaubtem Zugriff nicht sicher sind. Spätestens seit der NSA-Affäre interessiert die Nutzer*innen: Wer besitzt meine Daten? Wer kann darauf zugreifen? Was passiert bei deren Verarbeitung und Auswertung? Wissenschaftler*innen am paluno (The Ruhr Institute for Software Technology) der Universität Duisburg-Essen arbeiten an softwaretechnischen Lösungen zur sicheren Datenverarbeitung beim Cloud Computing.*

Ganz privat?

Der Schutz persönlicher Daten in der Cloud

Von Zoltan Mann, Andreas Metzger & Klaus Pohl

Die Cloud-Dienste im Internet werden immer zahlreicher, komfortabler und sind leicht zu nutzen. Beispiele sind die Apple iCloud, GoogleDocs, Amazon Web Services oder DropBox. Für die Nutzer*innen wird es jedoch zunehmend schwerer nachzuvollziehen, was solche Cloud-Dienste mit den persönlichen Daten machen und ob persönliche Daten in unerwünschte Hände gelangen. Dies liegt daran, dass Clouds und die Anbieter von Cloud-Diensten sich untereinander mehr und mehr vernetzen und somit Daten untereinander austauschen. Um Lastspitzen abzufangen, können Daten und Software-Komponenten auch in andere Cloud-Rechenzentren verschoben werden (sog. Cloud-Migration). Ein Beispiel: Wenn der Andrang auf einer Online-Einkaufsplattform kurz vor Weihnachten sehr hoch ist, erlaubt die Cloud, Daten in andere Rechenzentren auszulagern. Genau diese Flexibilität – im Cloud-Kontext oft auch Elastizität genannt – ist einer der wesentlichen Vorteile der Cloud. Aus Sicht des Datenschutzes birgt sie jedoch zusätzliche Gefahren.

Die Datenschutz-Grundverordnung der EU sieht vor, dass persönliche Daten die EU nicht verlassen dürfen. In dem Beispiel der Online-Einkaufsplattform dürfen die Daten der EU-Bürger*innen nicht außerhalb der Europäischen Union verschoben werden, weil andere Gesetze gelten. Allerdings kann dies dennoch passieren, etwa wenn Cloud-Dienste fehlerhaft entwickelt wurden und somit persönliche Daten zum Beispiel in Cloud-Rechenzentren in den USA verschoben werden.

Cloud-Systeme sind in hohem Maße dynamisch, so dass man zum Schutz persönlicher Daten in der Cloud die sich ständig ändernden Ziele und Rahmenbedingungen berücksichtigen muss. Die Dynamik von Cloud-Systemen hat viele Gründe: Die Rechenlast der Cloud-Dienste ändert sich kontinuierlich je nach Nutzer*innen-Verhalten, Rechner fallen manchmal aus, Software-Komponenten werden zwischen Rechnern migriert und so weiter. Beim Thema Datensicherheit kommen zusätzliche Dimensionen der Dynamik hinzu. Beispielsweise kann ein*e Nutzer*in die Sicherheitseinstufung seiner*ihrer Daten ändern, so dass Datensätze, die bisher nicht geschützt werden mussten, plötzlich als schützenswert eingestuft sind oder umgekehrt. Auch Änderungen von Software-Komponenten können ähnliche Effekte erzeugen, zum Beispiel wenn eine Software-Komponente, die bisher anonymisierte Daten ausgegeben hat, nach einer Programmänderung persönliche Daten ausgibt.

Weitere, subtilere Gefahrenquellen für den Datenschutz gehen von anderen Cloud-Nutzer*innen aus. Um die Kosten niedrig zu halten, ist

es im Cloud Computing üblich, dass Software-Komponenten für unterschiedliche Nutzer*innen auf demselben Rechner im Rechenzentrum laufen, gegebenenfalls gekapselt in sogenannte virtuelle Maschinen, aber immerhin auf derselben physischen Maschine. Dies ermöglicht prinzipiell, dass eine bösartige Software-Komponente eines Nutzers die Daten eines anderen Nutzers, die auf demselben Rechner beheimatet sind, ausspioniert oder verfälscht. Gegen solche Angriffe hilft auch nicht, dass das Rechenzentrum sich in einem erlaubten Land befindet; ein effektiver Schutz kann nur durch geeignete Sicherheitsmechanismen gewährleistet werden.

Lösungsansatz

Um mit dieser oben motivierten vielfältigen Dynamik und den Herausforderungen beim Schutz persönlicher Daten in der Cloud umgehen zu können, ist Adaptionsfähigkeit der Cloud-Systeme gefordert. Ein gutes Fundament dazu bildet das Wissen, das Softwareingenieure in den letzten 10 bis 15 Jahren über so genannte Selbst-adaptierende (Software-)Systeme gesammelt haben. Diese sind Systeme, die in der Lage sind, Änderungen der Umgebung zu beobachten und sich selbst – in vordefiniertem Rahmen – an die Änderungen der Umgebung anzupassen, wie in Abbildung (1) schematisch dargestellt.

Die Grundidee der paluno-Lösungen ist, dieses allgemeine Muster auf das Problem der Datensicherheit in der Cloud anzuwenden. Dazu werden die Konfiguration des Cloud-Systems (inklusive physischer und virtueller Maschinen, Anwendungskomponenten, Datensätze und Datenflüsse) sowie die Datensicherheitsanforderungen der Nutzer*innen kontinuierlich beobachtet. Falls die Beobachtungen potentielle Verletzungen feststellen, passen sich die Cloud-Dienste an, um solche Verletzungen zu verhindern oder zu mitigieren.

Dementsprechend bestehen die paluno-Lösungen aus zwei wesentlichen Pfeilern. Diese sind (1) die Beobachtung der Cloud-Systeme

(1) Adaptive Software-Systeme: Das Bild zeigt die beiden wesentlichen Phasen bei der Selbst-Adaption: Laufzeit-Beobachtung und dynamische Anpassung.
Quelle: paluno

während der Laufzeit und des Betriebs, um mögliche Datenschutzgefährdungen zu erkennen, (2) die dynamische Anpassung der Cloud-Systeme, um auf Datenschutzgefährdungen zu reagieren und diese möglichst zu vermeiden beziehungsweise zu mitigieren.

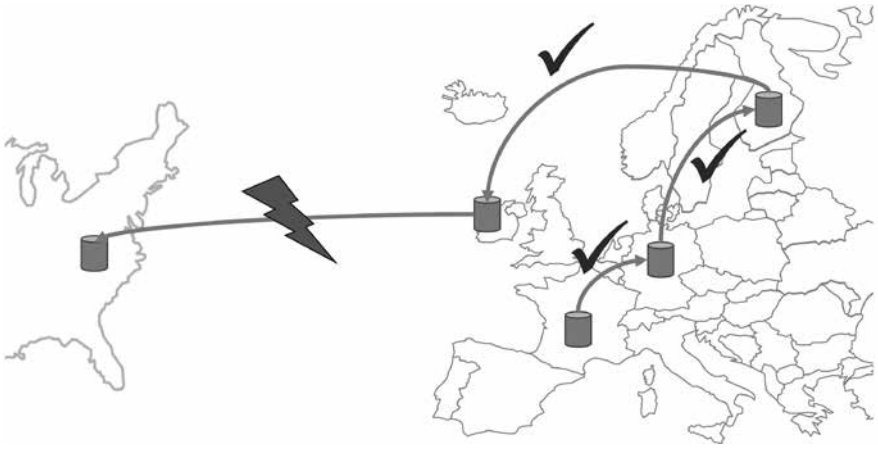
Pfeiler 1: Laufzeit-Beobachtung von Cloud-Systemen

Abbildung (2) zeigt ein plakatives Beispiel, wie es zu einer Verletzung des Datenschutzes beim Betrieb von Cloud-Diensten kommen kann. Persönliche Daten, die in der Cloud gespeichert sind, können in ein entfernt liegendes Datencenter zur Laufzeit verschoben werden. Durch eine solche Verschiebung können somit Daten in eine ausgeschlossene Geo-Lokation gelangen. In dem Beispiel würden etwa unerlaubt persönliche Daten in ein Datencenter in den USA gelangen. Die gezeigten Verschiebungen nach Deutschland, Finnland oder Irland wären in diesem Beispiel jedoch ohne Probleme.

Die Forschungsarbeiten von paluno zur Datenschutz-Beobachtung zielen darauf ab, solche Datenschutzverletzungen während des Betriebs von Cloud-Systemen zu erkennen. Im Rahmen des Gemeinschaftsprojekts iObserve erarbeiten paluno-Wissenschaftler*innen der Arbeitsgruppe von Prof. Klaus Pohl mit Forschungskolleg*innen der Christian-Albrechts-Universität zu Kiel und dem Karlsruher Institut für Technologie neue Techniken zur Cloud-Beobachtung. Das iObserve-Projekt gehört seit 2012 zum Schwerpunktprogramm 1593 „Design For Future – Managed Software Evolution“, das von der Deutschen Forschungsgemeinschaft (DFG) gefördert wird.

Zur Cloud-Beobachtung werden in iObserve neuartige Software-Programme entwickelt, die – sozusagen als Kontrollinstanzen – die Cloud-Systeme automatisiert beobachten, selbst wenn sich ein Cloud-Anbieter nicht beliebig in die Karten schauen lässt. Aus diesen Beobachtungen werden so genannte Laufzeitmodelle abgeleitet. Diese erlauben komplexe Analysen von Daten und Datenflüssen und liefern Hinweise, ob Datenschutzvorgaben eingehalten oder verletzt werden. Ausgehend von diesen Analysen können Handlungsempfehlungen generiert werden, um Datenschutzverletzungen zu verhindern. Insbesondere erlauben die iObserve-Lösungen Verletzungen von Geo-Lokations-Richtlinien zu erkennen. Diese treten auf, wenn Daten in unzulässige Regionen oder Länder verschoben werden (wie im Beispiel in Abb. 2 gezeigt).

Eine solche Cloud-Beobachtung während des Betriebs ist notwendig, da – wie oben erläutert – Software-Komponenten und Daten dynamisch – auch zwischen Rechenzentren – verschoben werden können. Diese Verschiebungen sind zur Entwurfszeit, sprich während der Entwicklung der Cloud-Dienste, nicht bekannt, da sie erst zur Laufzeit zur Optimierung der Performanz, der Verfügbarkeit und der Kosten ausgelöst werden. Durch die dynamische Verschiebung von Software-Kompo-



(2) Verschiebung von Daten zwischen Cloud-Rechenzentren kann zu Datenschutzverletzungen führen. Verschiebung innerhalb der EU ist laut EU-Datenschutz-Grundverordnung zulässig, jedoch nicht eine Verschiebung von Daten von EU-Bürger*innen in die USA.
Quelle: paluno

nenten, wie zum Beispiel von Datenbanken oder Komponenten zur Datenanalyse, könnten Daten in unerlaubte Länder gelangen.

Der in iObserve entwickelte Ansatz zur Datenschutz-Beobachtung von Cloud-Systemen würde eine Verletzung wie die in Abbildung (2) erkennen. In solchen Fällen erkennt der iObserve-Ansatz, dass eine potentielle Verletzung der Datenschutzrichtlinien aufgetreten ist, und kann somit die dynamische Anpassung der Cloud-Systeme auslösen.

Pfeiler 2: Dynamische Anpassung von Cloud-Systemen

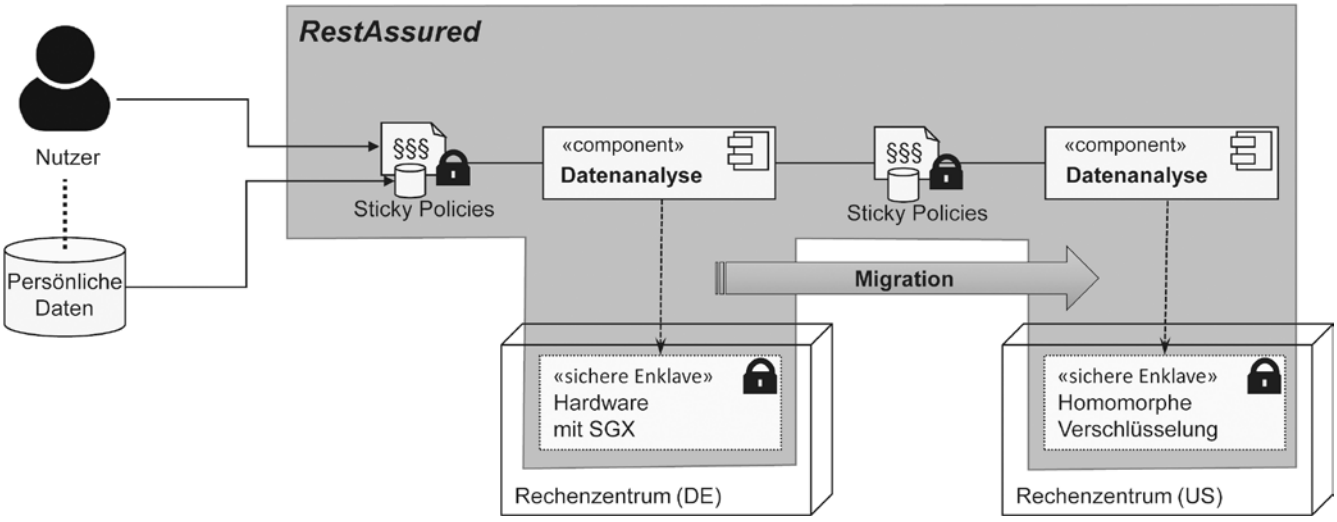
In den letzten Jahren gab es in den Bereichen der Systemsicherheit und der Kryptographie große Fortschritte und es wurden Sicherheitsmechanismen erarbeitet, die prinzipiell das Potenzial haben, die speziellen Sicherheitsbedenken in der Cloud zu adressieren. Diese Sicherheitsmechanismen bieten zusätzliche Anpassungsmöglichkeiten, die bei der dynamischen Anpassung von Cloud-Systemen genutzt werden können.

Das von der Europäischen Union geförderte Projekt „RestAssured – Secure Data Processing in the Cloud“ hat zum Ziel, mit Hilfe neuartiger Sicherheitsmechanismen Datensicherheit in der Cloud zu gewährleisten. In RestAssured arbeiten paluno-Wissenschaftler*innen der Arbeitsgruppen von Prof. Maritta Heisel und Prof. Klaus Pohl mit namhaften Partnern aus Forschung und Industrie zusammen, unter anderem mit IBM, Thales und der Universität Southampton. Gemeinsam werden Lösungen erarbeitet, die auf den neuesten wissenschaftlichen Erkenntnissen aufsetzen und gleichzeitig für den praktischen Einsatz tauglich sind.

Als konkrete Maßnahmen zur Anpassung des Systems mit dem Ziel des Datenschutzes werden in RestAssured drei innovative Sicherheitstechniken eingesetzt.

Hardware-Enklaven sind virtuelle Bereiche eines Hardware-Speichers, in denen Software-Komponenten ihre Daten und ihren Code sicher speichern können. Dadurch ist es möglich, sowohl die Daten als den Code kritischer Cloud-Dienste vor unbefugtem Zugriff zu schützen. Der Schutz ist wirksam nicht nur gegen Angriffe anderer Nutzer*innen, sondern sogar gegen potenzielle Angriffe seitens des Rechenzentrumbetreibers. Das Konzept von Hardware-Enklaven ist schon seit einigen Jahren bekannt. 2016 hat der Chiphersteller Intel mit dem SGX-Befehlssatz (Software Guard Extensions) erste Prozessoren auf den Markt gebracht, die dieses Konzept tatsächlich auch umsetzen. Diese Chips nehmen langsam Einzug in die Cloud-Rechenzentren. Allerdings kann man nicht davon ausgehen, dass auch in absehbarer Zeit alle Rechner diese Sicherheitsmechanismen unterstützen werden. Aus diesem Grund kann man SGX auch nicht per se einsetzen, sondern nur für die konkreten Fälle, in denen die Vermeidung von Datenschutzverletzungen angezeigt ist, und insoweit diese Mechanismen im konkreten Cloud-Rechenzentrum verfügbar sind.

Vollhomomorphe Verschlüsselung ist eine revolutionäre kryptographische Methode, die ein direktes Arbeiten auf dem Chiffretext ermöglicht. Traditionell werden verschlüsselte Daten zuerst entschlüsselt um anschließend verarbeitet werden zu können. Dies hat zur Folge, dass der Cloud-Dienst, der die Verarbeitung durchführt, zwangsläufig Zugriff auf die Daten bekommt und somit eine Sicherheitslücke darstellt. Im Gegensatz dazu brauchen die Daten bei homomorpher Verschlüsselung nicht entschlüsselt zu werden: Die nötigen Verarbeitungsschritte können direkt an den verschlüsselten Daten durchgeführt werden. Das bedeutet, dass die verarbeitende Anwendung die eigentlichen Daten gar nicht kennen muss, was die Risiken bezüglich Datensicherheit deutlich reduziert. Dass ein Verschlüsselungsverfahren mit den nötigen, sogenannten Homomorphie-Eigenschaften überhaupt existiert, wurde erst vor wenigen Jahren von Forscher*innen bei IBM nachgewiesen. Die heute bekannten homomorphen Verschlüsselungsverfahren sind entweder im Umfang der unterstützten Operationen eingeschränkt oder sind zu langsam für die praktische Anwendung, aber die großen Fortschritte der letzten Jahre lassen hoffen, dass vollhomomorphe Verschlüsselung in



(3) Beispiel für die Verwendung innovativer Sicherheitstechniken im Projekt RestAssured. Nutzer*in legen Datensicherheitsanforderungen mittels Sticky Policies fest. Bei einer Verschiebung der Daten aus Deutschland in die USA wird automatisch homomorphe Verschlüsselung eingesetzt, so dass keiner außer dem*der Nutzer*in Zugriff auf die Daten in unverschlüsselter Form erlangt.
Quelle: paluno

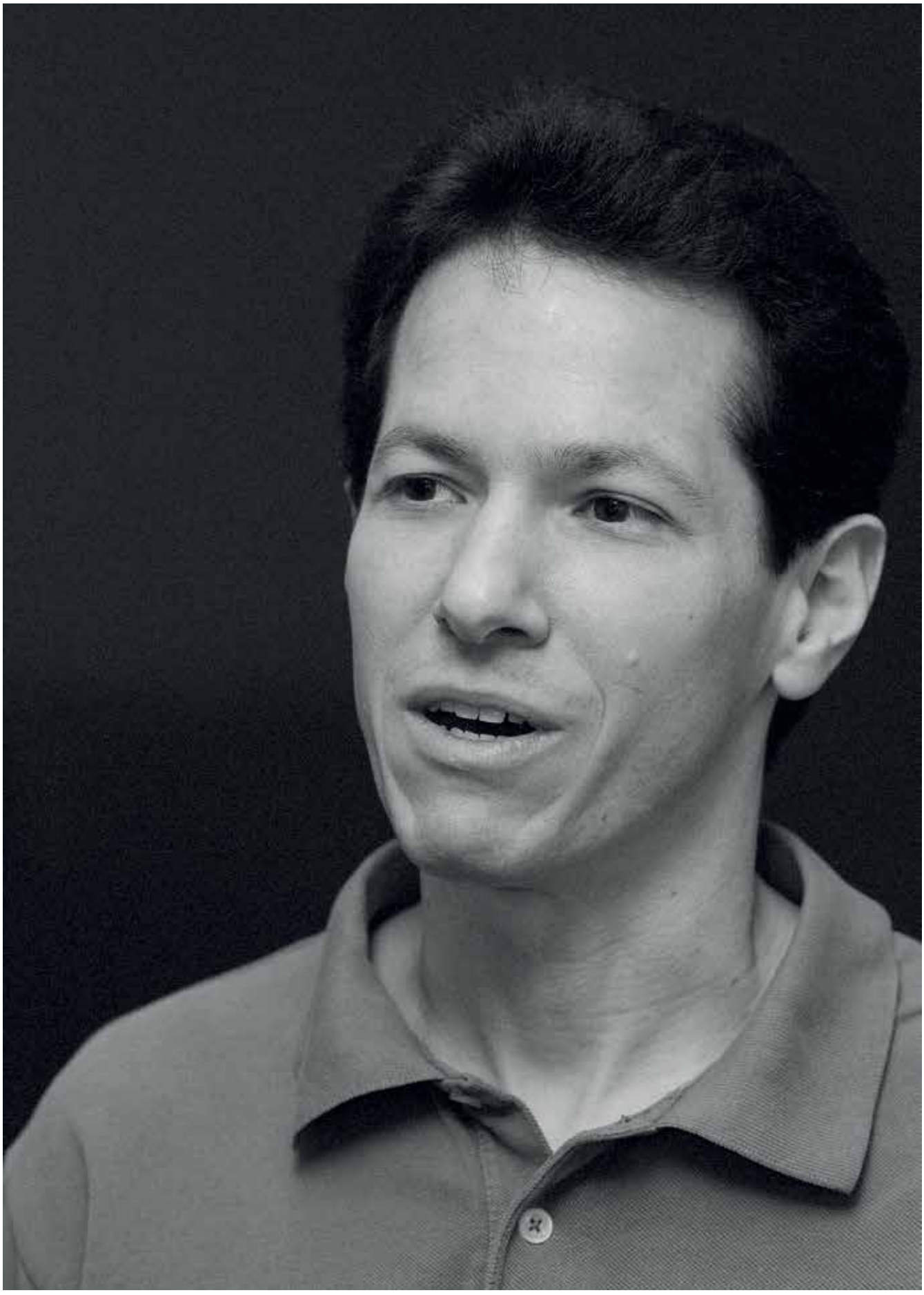
absehbarer Zukunft zu einer marktreifen Technologie wird. Wegen der Performanzeinbußen sollte vollhomomorphe Verschlüsselung auf jeden Fall nur dann eingesetzt werden, wenn die Vermeidung von Datenschutzverletzungen angezeigt ist.

Sticky Policies ermöglichen die Kennzeichnung der Kritikalität einzelner Datensätze. Dadurch kann gesteuert werden, wer welche Operationen auf diesen Datensätzen durchführen darf. Durch entsprechende kryptographische Verfahren wird sichergestellt, dass bei einer Bewegung des Datensatzes (z.B. Migration) auch die zugehörige Sticky Policy mitgeht (daher die Benennung „sticky“, da sie an dem Datensatz haftet) und auch eingehalten wird. Im Projekt RestAssured sorgt ein verteiltes Management der Sticky Policies dafür, dass der gesamte Lebenszyklus der Daten kontrolliert wird und der Schutzbedarf der Daten stets korrekt bestimmt ist.

Diese Sicherheitsmechanismen, zusammen mit den üblichen Anpassungsmechanismen von Cloud-Systemen (wie z.B. die Migration von virtuellen Maschinen zwischen physischen Maschinen) stellen die Grundlagen für eine integrierte und adaptive Datensicherheitslösung dar. Insbesondere können Sticky Policies genutzt werden, um die Datensicherheitsanforderungen zu spezifizieren und später auch zu ändern. Solche Änderungen können beobachtet werden und somit als Auslöser für Anpassungen dienen. Mögliche Anpassungen umfassen die Migration zwischen physischen Maschinen mit und ohne Hardware-Enklaven oder das Ein- und Ausschalten von homomorpher Verschlüsselung, wie auch in Abbildung (3) schematisch dargestellt.

Wichtig bei der Wahl einer geeigneten Anpassung ist, dass eine konkrete Anpassungsmöglichkeit neben einer Auswirkung auf die Sicherheitseigenschaften auch Nebenwirkungen auf andere Qualitätsattribute wie Performanz oder Kosten umfasst. Beispielsweise hat, wie oben erwähnt, die vollhomomorphe Verschlüsselung einen sehr hohen Rechenbedarf und damit eine Auswirkung auf Qualitätsattribute wie „Durchsatz“ oder „Antwortzeit“. Aufgrund der unterschiedlichen Auswirkungen der jeweiligen Anpassungsmöglichkeiten auf Datensicherheit und andere Qualitätsattribute ergeben sich daher verschiedene mögliche Anpassungsalternativen, aus welchen sich verschiedene Optima auswählen lassen basierend auf der relativen Wichtigkeit der einzelnen Optimierungskriterien.

Die im Projekt RestAssured erarbeiteten Lösungen werden anhand praktischer Anwendungsfälle validiert. Ein Anwendungsfall bezieht sich auf eine Online-Plattform für Pflegedienste, auf der die Zuordnung zwischen Pfleger*innen und Pflegebedürftigen verwaltet wird. Dabei kommt dem Schutz der Daten der Pflegebedürftigen eine besondere Bedeutung zu, da insbesondere auch medizinische Daten der Pflegebedürftigen gespeichert sind. Bei einem weiteren Anwendungsfall handelt es sich um eine fahrverhaltensbasierte Kfz-Versicherung, in der Sensoren im Automobil Daten über das Fahrverhalten der Autofahrer*innen erfassen und an ein Cloud-System senden. In der Cloud werden die gesammelten Daten ausgewertet und daraus die Versicherungsgebühren berechnet. Auch hier spielt der Schutz persönlicher Daten eine große Rolle.



Zoltan Mann. Foto: Vladimir Unkovic

Andreas Metzger. Foto: Vladimir Unkovic

Zusammenfassung

Die sichere Speicherung und Verarbeitung persönlicher Daten in der Cloud ist eine große Herausforderung. Solange dieses Problem nicht gelöst ist, kann man von den Vorteilen des Cloud Computing in Bereichen mit kritischen Daten nur sehr beschränkt profitieren.

Die kontinuierliche Datenschutz-Beobachtung stellt eine wichtige Grundlage für zukünftige Sicherheitslösungen dar. Insbesondere die Verletzung von gesetzlich vorgegebenen Geo-Lokations-Richtlinien wird im Projekt iObserve von Wissenschaftler*innen der UDE behandelt. Im Projekt RestAssured geht man einen Schritt weiter und nutzt innovative Sicherheitsmechanismen um die Cloud-Dienste bei einer möglichen Datenschutzverletzung anzupassen. Durch eine solche Anpassung an veränderte Rahmenbedingungen, beziehungsweise Sicherheitsanforderungen, zielt man darauf ab, Sicherheitsanforderungen stets in optimaler Weise zu erfüllen.

Summary

Cloud computing has many benefits for service providers and service users alike. However, by storing and processing data in the cloud, users lose control over their data. The data security breaches that became publicly known in the last couple of years have shown that this is a real threat that can hinder the adoption of cloud computing, especially in domains with strong data security requirements.

Researchers at paluno (The Ruhr Institute for Software Technology) are working to address this challenge as part of two ongoing research projects. The developed solutions consist of two main pillars: (1) monitoring the adherence to data protection requirements, which is developed in the iObserve project, and (2) adapting the cloud system to react to – actual or imminent – requirements violations, which is developed in the RestAssured project.

In the iObserve project, funded by the DFG, the focus is on monitoring data security in a cloud system. Through appropriate mechanisms, the violation of security requirements can be detected at runtime, making it possible to avoid or mitigate security breaches. In particular, violations of geo-location policies can be detected and appropriate counter-measures may be taken.

In the RestAssured project, funded by the European Union, the focus is on adaptation by leveraging innovative security mechanisms such as hardware enclaves and homomorphic encryption to guarantee security in an adaptive and integrated way. By also considering the impact of the used security mechanisms on other quality attributes like costs and performance, security requirements can be fulfilled in a globally optimal way.

Danksagung

Die vorgestellten Forschungsarbeiten werden gefördert im Rahmen des Schwerpunktprogramms SPP1593: „Design For Future – Managed Software Evolution“ der Deutschen Forschungsgemeinschaft (DFG) unter Förderkennzeichen PO 607/3 (iObserve), sowie im Rahmen des EU Horizon 2020 Forschungs- und Innovationsprogramms unter Förderkennzeichen 731678 (RestAssured).

Literatur

– E. Schmieders, A. Metzger, K. Pohl (2017): Modellbasierte Datenschutzprüfung datenintensiver Cloud Dienste, In: Jan Jürjens, Kurt Schneider (Hrsg.): Software Engineering 2017, Hannover, Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn.
– Z. Mann, A. Metzger (2017): Optimized Cloud Deployment of Multi-tenant Software Considering Data Protection Concerns. In: J. Carretero, M. Parashar, J. Garcia-Blas, G.C. Fox, F. Cappello (Hrsg.): 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2017), Madrid, IEEE.
– E. Schmieders, A. Metzger, K. Pohl (2015): Runtime Model-Based Privacy Checks of Big Data Cloud Services. In: A. Barros, D. Grigori, N.C. Narendra, H.K. Dam (Hrsg.): Service-Oriented Computing: Proceedings of the 13th International Conference on Service Oriented Computing (ICSOC 2015), Paris, LNCS 9435, Springer, Berlin, Heidelberg.
– E. Schmieders, A. Metzger, K. Pohl (2014): A Runtime Model Approach for Data Geo-Location Checks of Cloud Services. In: X. Franch, A.K. Ghose, G.A. Lewis, S.Bhiri (Hrsg.): Service-Oriented Computing: Proceedings of the 12th International Conference on Service Oriented Computing (ICSOC 2014), Ghoa, LNCS 8831, Springer, Berlin, Heidelberg.

Die Autoren

Zoltan Mann ist wissenschaftlicher Mitarbeiter am paluno (The Ruhr Institute for Software Technology) an der Universität Duisburg-Essen. Er studierte Informatik und Mathematik in Budapest (Ungarn), wo er 2005 in Informatik promovierte. Zwischen 2004 und 2014 arbeitete er bei verschiedenen Unternehmen in den Bereichen Softwareentwicklung, IT-Beratung und Unternehmensberatung in Deutschland und Ungarn. Zwischen 2007 und 2016 war er an der Technischen und Wirtschaftswissenschaftlichen Universität Budapest beschäftigt, anfangs als „Assistant Professor“, später als „Associate Professor“.

Andreas Metzger ist Akademischer Oberrat an der Universität Duisburg-Essen und Leiter der Bereiche Adaptive Systeme und Big Data Anwendungen am paluno (The Ruhr Institute for Software Technology). Er ist stellvertretender Generalsekretär der Europäischen Big Data Value Association (BDVA) und Vize-Vorstand der Europäischen Technologieplattform NESSI (Networked European Software and Services Initiative). Er studierte Informatik an der TU Kaiserslautern, wo er 2004 in Informatik zum Dr.-Ing. promovierte.

Klaus Pohl ist Professor für Software Systems Engineering an der Universität Duisburg-Essen und Direktor von paluno (The Ruhr Institute for Software Technology). Von 2005 bis 2007 war er wissenschaftlicher Gründungsdirektor des Irish Software Engineering Research Centre (lero). Er ist Vorstandsmitglied der Europäischen Technologieplattform NESSI (Networked European Software and Services Initiative) und Mitglied des Lenkungsausschusses der deutschen Innovations-Allianz SPES (Software-Plattform for Embedded Systems).

Drei Forscher der Universität Duisburg-Essen haben sich nach mehreren Jahren Forschung mit der Firma Locoslab selbständig gemacht. Locoslab GmbH ist ein 2012 ins Leben gerufene Spin-off der Universität, das cloud-basierte Produkte und Dienstleistungen rund um das Thema Lokalisierung anbietet. Vielfältige Anwendungsmöglichkeiten durch den technischen Einsatz des so genannten RF (Radio Frequencing) Fingerprinting, beispielsweise in den Bereichen Navigation auf Flughäfen, in Einkaufszentren oder auf Messen. So können personalisierte Mobilitätslösungen realisiert werden.

SIMON macht mobil

Die Cloud ermöglicht personalisierte Mobilität

Von Pedro José Marrón

Als Teil der Aktivitäten im Bereich personalisierte Mobilität, haben sich Mitarbeiter*innen der Firma Locoslab gemeinsam mit Forscher*innen der Universität Duisburg-Essen der Herausforderung gestellt, Navigationsanwendungen zu implementieren, die für Personen mit eingeschränkter Mobilität besonders relevant sind. Im Rahmen des SIMON EU-Projekts wurden unter anderem Lösungen für multimodales Routing erforscht, die in der Lage sind, Outdoor- und Indoor-Routing nahtlos zu kombinieren. Die Integration dieser Lösungen ist in einer App namens SIMON Mobile erfolgt, die für Android und iOS verfügbar ist.

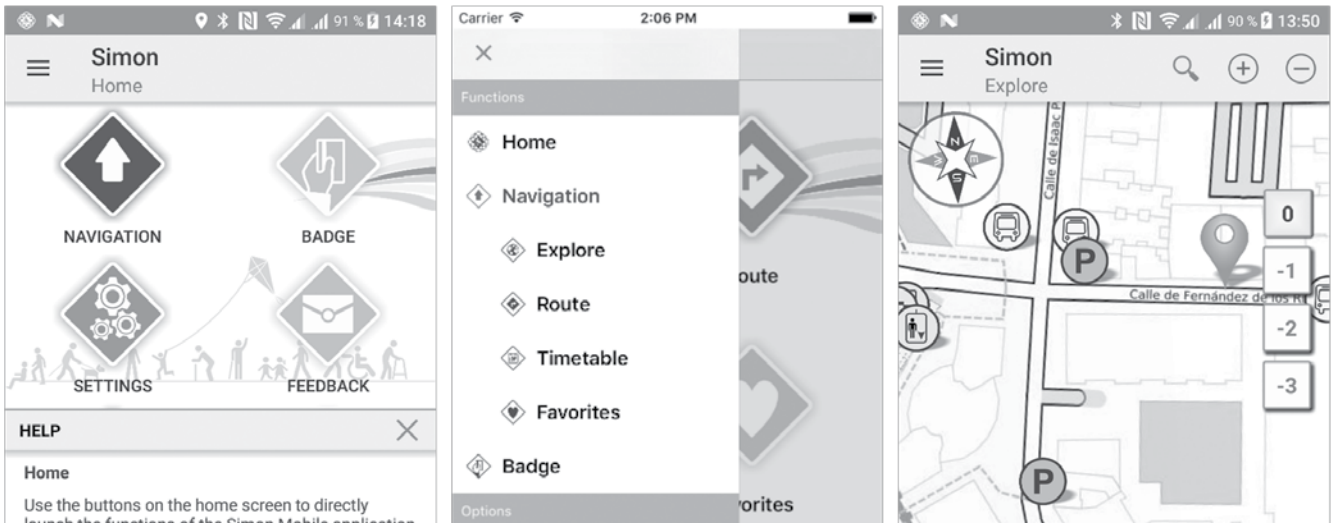
Erfahrungen mit Indoor Navigation in SIMON Mobile

Mit der Ausbreitung GPS-fähiger Geräte haben sich mobile Navigationsanwendungen zu einem weit verbreiteten Werkzeug entwickelt, die Menschen während ihrer täglichen Fahrten unterstützen. Da GPS in Gebäuden nicht verfügbar ist, werden solche Anwendungen meist nur im Außenbereich genutzt. Infolgedessen kann eine nahtlose End-to-End Nutzung nicht gewährleistet werden. Dies ist besonders dann problematisch, wenn man schnell durch ein Gebäude laufen muss, um zum Beispiel eine bestimmte U-Bahn Haltestelle zu erreichen.

Das Ziel des Europäischen Forschungsprojektes SIMON ist die Entwicklung von Tools und technischen Lösungen, die eine Verbesserung der Mobilität von Menschen mit Einschränkungen ermöglichen. Eine dieser Lösungen ist eine mobile Navigationsanwendung, die speziell für Nutzer*innen mit Geh- oder Sehbehinderungen konzipiert wurde. Die Anwendung wird zur Zeit in Lissabon, Parma, Reading und Madrid getestet.

Ein wichtiger Teil der Anwendung ist die Bereitstellung von nahtlosen und multimodalen Navigationsrouten, die sowohl im Außen- wie im Innenbereich funktionieren und verschiedene Verkehrsmittel wie

Pedro José Marrón. Foto: Vladimir Unkovic



(1) SIMON Mobile.
Quelle: eigene Darstellung

Bus, U-Bahn, Auto oder zu Fuß unterstützen. Die Arbeit baut auf Ergebnissen eines früheren EU-Projekts namens GAMBAS auf [2], das an der UDE von unserer Forschungsgruppe koordiniert wurde. Um die neue resultierende Anwendung testen zu können, haben wir ein auf Bluetooth basierendes Indoor-Lokalisierungssystem in Moncloa, eine große Transit-Station in der Stadt Madrid installiert. Darüber hinaus haben wir die Anwendung für Android und iOS über den Google Play Store und den Apple App Store zur Verfügung gestellt. Zum Zeitpunkt des Schreibens hat die Anwendung etwa 1.000 Downloads mit ungefähr 400 aktiven Installationen.

Simon Mobile

Um das Risiko zu minimieren, eine Navigationsanwendung zu bauen, die nicht den Bedürfnissen unserer primären Nutzer*innengruppe entsprechen würde, folgten wir einem partizipativen Designprozess. In einer Reihe von Workshops mit Behörden in den Zielstädten sowie Mitgliedern von Gruppen und Organisationen mit Mobilitätseinschränkungen wurden die Anforderungen und Informationsquellen definiert und zu einem späteren Zeitpunkt validiert.

Im Laufe der Arbeit haben wir schnell erkannt, dass die notwendigen Informationen oft nicht leicht zugänglich waren oder über verschiedene Informationsanbieter und -systeme hinweg verteilt waren. Zum Beispiel musste eine Excel-Tabelle mit den Orten der Aufzüge händisch ausgewertet und mit den Fahrplaninformationen des U-Bahnnetzbetreibers kombiniert werden. Um Routen für die Intercity-Busse berechnen zu können, mussten wir ein Tool entwickeln, das die Buspläne aus etwa 600 PDF-Dateien heraus extrahiert, konsolidiert und in die Cloud-Infrastruktur integriert, die als Basis für die Ausführung der Client-Apps benutzt werden.

Basierend auf den in den Workshops gesammelten Endbenutzer*inneneingaben haben wir die Anwendung entwickelt, die in Abbildung (1) zu sehen ist. Mit einem zweistufigen hierarchischen Menü (1) kann ein*e Benutzer*in auf die Funktionen der Hauptanwendung zugreifen. Jeder Screen ist als Einzelansicht implementiert (Fragmente auf Android und ViewController auf iOS). Für jede Ansicht kann ein*e Benutzer*in eine kontextsensitive Hilfe aufrufen, die entweder auf der Unterseite (im Hochformat) oder auf der rechten Seite (im Querformat) angezeigt wird. Die Hilfe beschreibt den Zweck der Ansicht sowie deren Funktionalität. Wir haben auch ein Navigationselement integriert (2), das persistent in der oberen linken Ecke angezeigt wird und von jeder Ansicht angeklickt werden kann, indem man eine Swipe-Geste von links nach rechts durchführt. Das Menü bietet einen direkten Zugriff auf wichtigen Funktionen und auf die Hilfe-Ansicht.

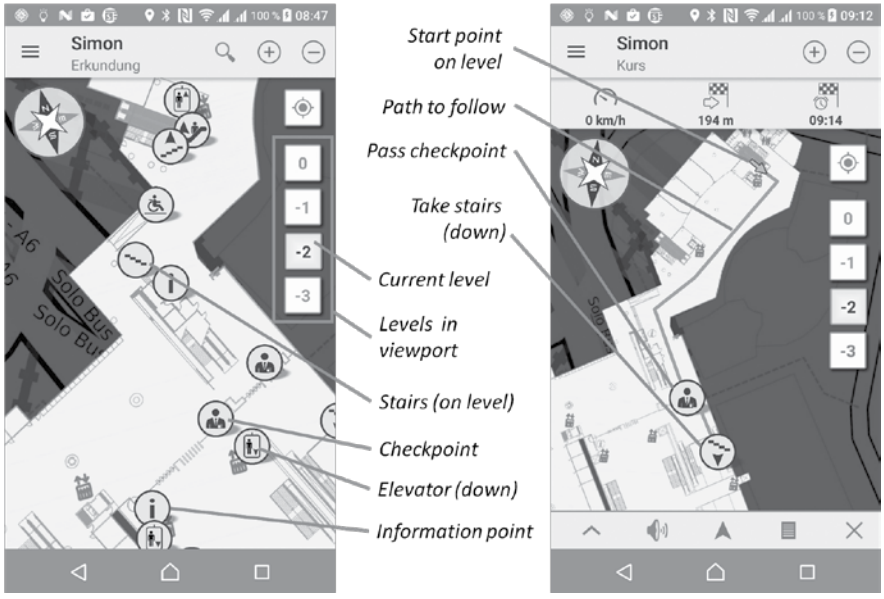
Ähnlich wie bei anderen Navigationsanwendungen, ermöglicht eine Explorationsansicht (Abb. 3) dem*der Benutzer*in die Karte unter Verwendung von Panning, Pinching oder Rotieren zu manipulieren. Wenn man den Zoom betätigt, wird die Karte automatisch geladen und relevante Orte werden als Overlay markiert und visualisiert. Dabei werden in der Visualisierung verschiedene Symbole zur Differenzierung der Kategorien verwendet. Außerdem werden Farben benutzt, um dynamische Informationen anzuzeigen (z.B. besetzte Parkplätze mit einem roten „P“ Symbol und verfügbare Parkplätze mit einem grünen „P“ Symbol). Mit einem kurzen Tap auf den Screen kann ein*e Benutzer*in einen Marker setzen, mit einer langen Berührung des Bildschirms wird ein anderer Ort ausgewählt, um Details darstellen zu können. Diese Detailansicht enthält

eine Gruppe von Typ-spezifischen Schaltflächen, die es dem*der Benutzer*in ermöglichen, einen Ort als Eingabe für weitere Aktionen zu verwenden (z.B. um einen Zeitplan für eine Bushaltestelle abzurufen oder um den Ort als Start- bzw. Zielort für eine Route auszuwählen).

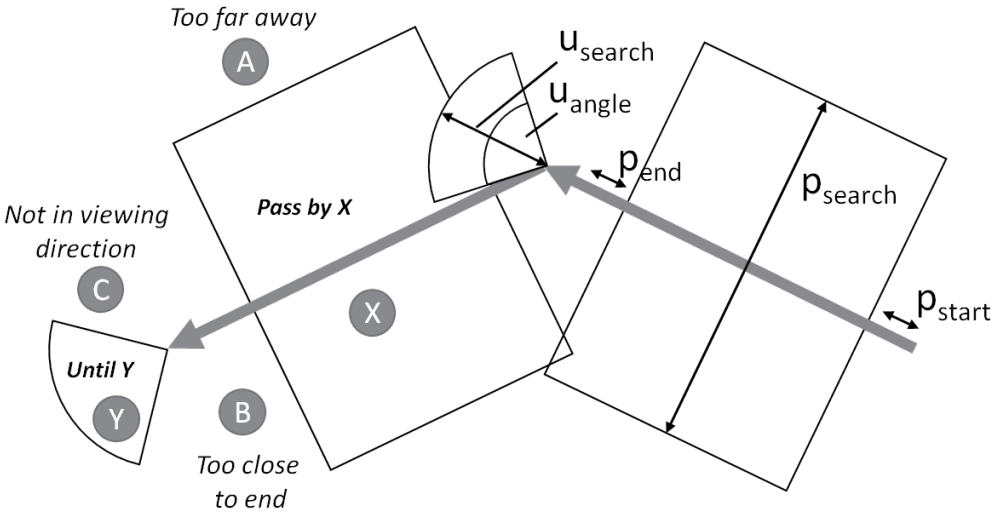
Bei der Suchansicht (Abb. 4) kann der Benutzer Orte nach Namen oder Adressen suchen. Ob eine Suche lokal oder global durchgeführt wird, hängt davon ab, ob die Informationen mit dem aktuellen Kontext (z.B. Standort) verbunden sind oder nicht. Die globale Suche erfordert von dem*der Benutzer*in Teile eines Stadtnamens einzugeben und diese wird dann mit den gespeicherten Daten in der Cloud verglichen. Die lokale Suche verwendet den aktuellen Benutzer*innenkontext, um die Ergebnisse entsprechend der Distanz zum Standpunkt zu ordnen und einzuschränken. Der*Die Benutzer*in kann auch eine Reihe von Schaltflächen nutzen, um die Ergebnisse auf einen bestimmten Ort zu begrenzen. Wenn ein Ergebnis angezeigt wird, werden die Suchbegriffe und der Suchverlauf gespeichert. Somit ist es dann möglich, zu einem späteren Zeitpunkt vorherige Anfragen schnell wieder durchführen zu können.

Unter Verwendung einer Zeittabellenansicht (Abb. 5) kann ein*e Benutzer*in an den Haltestellen auf die Abfahrt- und Ankunftszeiten von öffentlichen Verkehrsmitteln zugreifen. Dafür kann der*die Benutzer*in entweder die Explore-Ansicht oder die Suchansicht betätigen. Wenn die Suchansicht verwendet wird, werden die Filtertasten ausgeblendet und die Suchansicht automatisch so eingestellt, dass sie nur Stopps und Stationen anzeigen. Wenn die Abfahrtsinformationen für eine bestimmte Haltestelle angefragt werden, werden die Ergebnisse Zeile für Zeile gruppiert. Für jede Zeile werden nicht nur die Abfahrtsinformationen pro Stunde angezeigt sondern auch Echt-Zeit Informationen und Verhinderungen, die einen Einfluss auf den Verlauf der Route haben können. Bei Bedarf werden diese Informationen in verschiedenen Sprachen automatisch in der Cloud übersetzt. Dafür haben wir unser Backend mit den APIs von Microsoft integriert, die eine automatische Übersetzung zur Verfügung stellen (https://azure.microsoft.com/en-us/services/cognitive-services/translator-text-api/).

In der Routenansicht (Abb. 5) kann ein*e Benutzer*in Routen von einem Start zu einem Ziel unter Berücksichtigung verschiedener Bewegungsmodalitäten (zu Fuß, mit dem Auto oder mit öffentlichen Verkehrsmitteln) berechnen. Bei der Eingabe des Start- beziehungsweise Endpunktes der Reise können verschiedene Ansichten benutzt werden. Falls einer der Punkte nicht spezifiziert wird, benutzt das System die jetzige Position des*der Nutzer*in. Um Reisen planen zu können, kann auch eine Uhrzeit angegeben werden, an der die Reise stattfinden soll. Darüber hinaus erstellt SIMON Mobile weitere Optionen für die Berechnung der Routen, die mit dem Profil des*der Nutzer*in in Verbindung gebracht werden können. Zum Beispiel kann der*die Nutzer*in die Routenberechnung auf „nur Bus“ oder „nur Bahn“ einschränken oder angeben, dass er Mobilitätseinschränkungen hat und nur Aufzüge und rollstuhlgerechte Zugänge auszugeben sind. Die Berechnung der Route findet in der Cloud statt. Die Ergebnisse wiederum werden in der App dargestellt und je nach Einschränkung entweder lediglich visuell angezeigt oder zusätzlich automatisch vorgelesen. Dies ermöglicht, dass Reisende mit Seheinschränkungen unsere App



(2) Indoor Visualisierung.
Quelle: eigene Darstellung



(3) Berechnung der Landmarks.
Quelle: eigene Darstellung

auch benutzen können. Die Liste der möglichen Routen wird nach Relevanz sortiert und dem*der Benutzer*in präsentiert. Um alternative Ergebnisse schnell vergleichen zu können, verwenden wir verschiedene Icons, die die einzelnen Teile der Route darstellen. So wird zum Beispiel ein Bus oder eine U-Bahn für das Segment benutzt, wo der*die Nutzer*in diese Verkehrsmittel nutzen muss oder ein*e Fußgänger*in, wenn er*sie zu Fuß die angezeigten Anzahl der Meter laufen muss. Darüber hinaus werden auch zusammenfassende Informationen wie Abfahrtszeit, Ankunftszeit, Umstiege oder die Gesamtdauer der Reise berechnet und gemeinsam mit jeder Route dargestellt. Diese Informationen werden entweder in Text-Form oder auf einer Karte angezeigt, damit der*die Nutzer*in auch die Möglichkeit hat, sich auf der Karte zu orientieren und visuell zu entscheiden, welche Route sich besser eignet.

Wenn die Schritt-für-Schritt-Navigation für eine ausgewählte Route gestartet wird, schaltet die Anwendung zu einer Tracking-Ansicht (8), die die verbleibende Route des Benutzers als Polylinie auf einer Karte visualisiert. Außerdem zeigt die Ansicht die jetzige Position des*der Nutzer*in sowie wichtige Points of Interest (POIs). Die Ansicht unterstützt zusätzlich einen „Follow-me"-Modus, der den*die Benutzer*in immer in der Mitte der Karte platziert und die Karte schiebt und rotiert, wenn der Nutzer in die eine oder andere Richtung läuft. Um den Drehwinkel zu berechnen, verwenden wir die Ausrichtung des Mobiltelefons, die entweder von GPS oder vom eingebauten Beschleunigungssensor/Magnetometer lokal erfasst wird. Die Tracking-Ansicht zeigt auch die Geschwindigkeit des*der Benutzer*in, die verbleibende Entfernung, sowie die erwartete Ankunftszeit. Der „Follow-me"-Modus wird automatisch deaktiviert, wenn der*die Benutzer*in beginnt, selber die Karte zu manipulieren (z.B. durch Pinching, Panning, usw.), und es kann wieder aktiviert werden durch Drücken des „Locate me"-Buttons in der oberen rechten Ecke der Anwendung.

Während der Schritt-für-Schritt-Navigation kann ein* Benutzer*in weiterhin alle Funktionen der Anwendung benutzen. Wenn die Navigation aktiv ist, zeigt die Anwendung kontinuierlich eine Leiste am unteren Rand, die die nächsten eine bis zwei Anweisungen mit Icons und Text visualisiert. Die Leiste kann aber auch minimiert werden, um mehr Platz auf dem Screen zu schaffen, ohne die Möglichkeit zur Navigation zu verlieren.

Indoor-Navigation

Um Indoor-Navigation in das System zu integrieren, mussten wir die zugrundeliegenden Visualisierungen, Dienste und Algorithmen anpassen. Insbesondere mussten wir Erweiterungen in der Modellierung und Visualisierung der einzelnen Elemente durchführen.

Modellierung

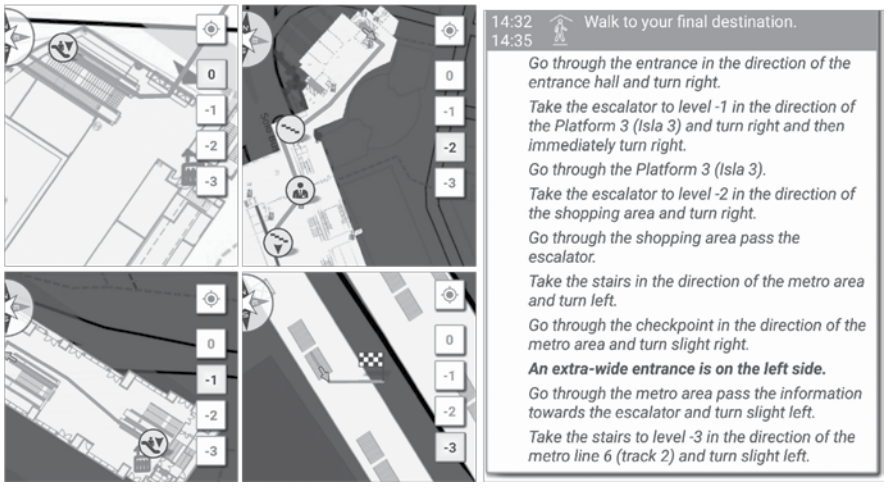
Für die Außennavigation setzen wir auf Daten des OpenStreetMap-Projekts. Das Datenmodell besteht aus Knoten, Wegen und Beziehungen. Ein Knoten repräsentiert einen bestimmten Punkt auf der Karte, der als WGS84 Koordinate definiert ist. Ein Weg stellt eine Reihe von Punkten dar, die eine Polylinie sein können (z.B. um eine Straße zu repräsentieren) oder ein Polygon (z.B. ein See oder ein Wald, etc.). Unter Verwendung von Relationen ist es möglich, Gruppen von Objekten (z. B. eine Menge von Straßen, die eine Autobahn repräsentieren) zu definieren, die bestimmte Eigenschaften haben (z.B. Beschränkung an einer

bestimmten Straßenkreuzung). Um die Semantik dieser Elemente zu definieren, werden Attribute-Value Pairs benutzt, deren Bedeutung und Verwendung in den Richtlinien des OpenStreetMaps Projekts detailliert beschrieben werden.

Allerdings ist es nicht möglich, diese Daten direkt aus dem OpenStreetMap Projekt zu verwenden. Dafür ist es notwendig, einen Renderer zu schreiben, der diese Daten für die Außennavigation vorbereitet, synthetisiert und basierend auf der semantischen Information, die richtigen geometrischen Elemente extrahiert.

Darüber hinaus benötigen wir eine Routing-Engine, die auf den Graphen arbeitet, die durch die topologische Struktur der Wege definiert werden. Dadurch können Straßen erkannt und zusammen gefasst werden. Für die Indoor-Navigation benutzen wir das gleiche Modell, um die Wege in einem Gebäude zu definieren und als Input für die Routenberechnung bereit zu halten. Um mehrere Etagen zu unterstützen, fügen wir dem Knoten des Modells einen Integer hinzu, der die Etage darstellt, wo sie zu finden sind. So wird eine 0 für das Erdgeschoss benutzt, positive Zahlen für höhere Etagen und negative Werte, um unterirdische Etagen wie bei einer U-Bahn-Station zu repräsentieren. Wir generieren dann Links zwischen zwei Knoten auf zwei verschiedenen Ebenen, um zwei Etagen miteinander zu verbinden, womit wir die Möglichkeit haben, Treppen, Rolltreppen, Rampen oder Aufzüge zu modellieren. Die Unterschiede zwischen einer Treppe und einem Aufzug wird durch die Nutzung von semantischen Informationen an den jeweiligen Knoten realisiert. Orte, die für den*die Nutzer*in von Interesse sind, werden als Knoten (z.B. ein Geschäft), Wege (z.B. einen Gang) oder Polygone (z.B. eine Einkaufszone) in dem Modell dargestellt und mit besonderen semantischen Informationen versehen.

Um den Berechnungsaufwand zu minimieren, wird dieses Modell als Input der Routing-Engine in der Cloud verwendet. Die Darstellung der einzelnen Grundrisse, Etagen und so weiter werden vorberechnet und als Bild auf die Client-Anwendung übertragen. Durch die Definition von drei Ankerpunkten auf der Basis der OpenStreetMap-Daten, können wir die vorberechneten Bilder auf den WGS84-Koordinaten abbilden, die von den Ankerpunkten dargestellt werden. Durch eine affine Transformation werden dann lokal von der Client-Anwendung die Bilder auf der Karte dargestellt und als Overlay an der korrekten Stelle angezeigt.



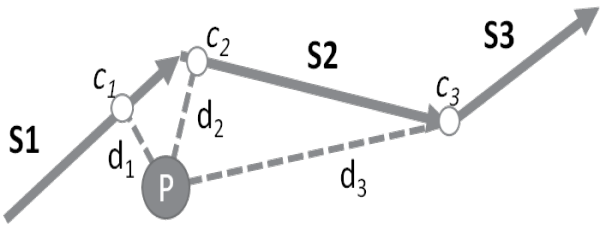
(4) Beispielroute mit Textanweisungen.
Quelle: eigene Darstellung

Visualisierung

Für die Visualisierung der Innenräume stellen wir zwei Dienste zur Verfügung. Der erste Dienst wird benutzt, um die Overlay-Bilder zu empfangen. Da die Auflösung der vorberechneten Bilder hoch sein kann, werden die Bilder in Teile gesplittet, die individuell mit dem slippy Addressierungsmechanismus abgerufen werden können [4], die von OpenStreetMap und anderen Kartenanbieter verwendet wird. Der zweite Dienst ermöglicht die Berechnung der Etagen, die in einer View der Karte zu finden sind. Dazu werden Bounding-Boxes mit Etagen-Informationen aus den Modell-Elementen herausgefiltert und zurückgegeben.

Beim Testen der Visualisierung fanden wir schnell heraus, dass eine zweidimensionale Darstellung der verbindenden Elemente (siehe Abb. 2) auf verschiedenen Etagen (z.B. Treppen, Rolltreppen, usw.) nicht immer von allen Nutzer*innen korrekt interpretiert wurde. Insbesondere kam die Frage „Welche Rolltreppe fährt nach oben und welche nach unten? Welche soll ich nehmen? Manchmal war es so, dass es nicht ersichtlich war, welche zwei Etagen durch eine Rolltreppe verbunden sind und das hat in solchen Fällen für Verwirrung gesorgt. Ein ähnliches Problem hing mit der Darstellung der 3D-Routen zusammen. SIMON Mobile unterstützt zwar die Berechnung von Routen, die über mehrere Etagen laufen, aber die Darstellung von solchen Ergebnissen war für viele Nutzer nicht intuitiv. In manchen Fällen war es nicht klar, welche Pfade mit welcher Etage gekoppelt waren, womit unsere Testgruppe Schwierigkeiten hatte, den Anfang und das Ende der Route auf den jeweiligen Etagen zu finden.

Um diese Probleme zu lösen, fügten wir Informationen über die dritte Dimension hinzu anhand von Annotationen auf der Karte, die nicht nur die Art der Verbindung



(5) Positionsbestimmung auf einer Linie mit drei Segmenten.
Quelle: eigene Darstellung

zwischen zwei Etagen darstellen, sondern auch die Richtung die für die Route relevant ist. In der Karte sind diese zusätzlichen Informationen für alle sichtbaren Elemente zu finden aber in der Routen-Darstellung werden nur die Markierungen gezeigt, die für die Route relevant sind. Dadurch kann sich der*die Benutzer*in auf die wesentlichen Routeninformationen konzentrieren und wird weniger durch Extra-Informationen abgelenkt. Für den Anfang der Route wurde einen Pfeil hinzugefügt, der die Richtung anzeigt, der der*die Nutzer*in zu folgen hat. Die daraus resultierende Visualisierung ist in Abbildung (2) dargestellt.

Zusammenfassung der dargestellten Informationen

Da mehrere unserer Tester*innen Sehbehinderungen haben, war eine Zusammenfassung der Indoor-Route in Textform notwendig, um die Anweisungen nicht nur visuell sondern auch als Sprachausgabe darstellen zu können. Darüber hinaus sind Anweisungen in Bezug auf Meter oder Winkel nicht so nützlich wie andere, die auf Landmarks basieren und diese als Ankerpunkte gebrauchen [3]. Die richtige Mischung aus Hinweisen und Textbeschreibungen haben wir in der Zusammenarbeit mit den Nutzergruppen des Konsortiums über mehrere Tests herausgearbeitet und in der Application integriert. Folgende Elemente des Modells haben wir mit den folgenden Eigenschaften erweitert:

- *Links:* Die Beschreibungen der Route beinhalten explizite Informationen über die Verknüpfung der folgenden Etagen wie zum Beispiel „Nehmen Sie den Aufzug runter bis Ebene -2.“. Somit ist es klar, welche zwei Etagen ein Link verbindet und diese kann auch bei den Routing-Anweisungen verwendet werden.
- *Landmarks*: Wie in [3] beschrieben, wurden Landmarks als primäres Element benutzt, um Wendepunkte zu identifizieren. Personen, die das Gebäude gut kennen, orientieren sich eher über bekannte Orte, zum Beispiel „Laufen Sie an Geschäft X vorbei und anschließend biegen Sie rechts ab“. Aber auch andere Nutzer*innen, die sich nicht im Gebäude auskennen, orientieren sich lieber im Innenbereich über Landmarks als über relative Distanzangaben, wie es in Autonavigationsanwendungen üblich ist.
- *Kompression*: Für Anweisungen, die schnell nacheinander durchgeführt werden sollten, weil sie in einer relativ kurzen Distanz auftreten, ist es wichtig, beide Anweisungen in einem Satz auszugeben. Das gibt dem*der Nutzer*in die Möglichkeit, sich vorzubereiten und nicht über die nächste Anweisung überrascht zu werden. Aus diesem Grund generiert unsere Anwendung komprimierte Hinweise, die zum Beispiel wie folgt aussehen: „Gehen Sie die Treppe hinunter und dann biegen Sie rechts ab“.
- *Zonen*: Die Beschreibungen des Verkehrsbetreibers beinhalten häufig Anweisungen über Zonen, die von den Nutzer*innen eindeutig identifiziert werden können. Zum Beispiel „Gehen Sie in Richtung Metro“ oder „Gehen Sie in Richtung Einkaufszone“. Diese semantischen Informationen müssen in das Modell eingepflegt werden und als Landmark benutzt werden, um die Anweisungen verständlicher zu machen.
- *Anmerkungen*: Interessanterweise wurden auch solche Aussagen und Hinweise von den Nutzer*innen erwartet wie „Bitte beachten Sie die Haltestellenregeln.“ oder „Bitte Vorsicht bei der Einfahrt des Zuges“, sowie Hilfestellungen wie „Ein rollstuhlgerechter Eingang befindet sich auf der linken Seite der Barriere.“. Diese Anmerkungen und Hinweise wurden nicht nur modelliert, sondern auch an der entsprechenden Stelle wie zum Beispiel beim Betreten des Gebäudes oder der U-Bahn Zone ausgegeben.

Basierend auf diesen Beobachtungen und Anforderungen der Nutzer*innen, haben wir Notizen und Zonen zu unserem Modell hinzugefügt, so dass es möglich ist, Sprachausgaben an den jeweiligen Stellen durchzuführen, die für die Route, Zone oder Wendepunkte notwendig sind. Um diese Informationen in unser Modell zu integrieren, haben wir einen Nachbearbeitungsschritt eingefügt, der auf der räumlichen Analyse der Routen basiert und in fünf Schritten durchgeführt wird.

Zuerst gruppieren wir die Pfade, die auf einer gerade Linie auf einer bestimmten Ebene sind. Dafür berechnen wir die Winkel zwischen den Segmenten der Polylinie der Route und teilen sie an den Stellen, an denen der Winkel einen voreingestellten Schwellenwert überschreitet oder Übergänge repräsentiert werden (z.B. Treppen, Aufzüge, Kontrollpunkte, usw.). Das Ergebnis ist eine Reihe von Gruppen, die entweder gerade Linien oder Transitionen repräsentieren. Zwischen den Gruppen wird der Winkel berechnet,

um Abbiege-Anweisungen zu generieren.

In dem zweiten Schritt werden die semantischen Informationen identifiziert, die zu den Zonen gehören, die sich mit der Route schneiden. Dafür wird eine topologische Analyse durchgeführt, die die entsprechenden Elemente aus dem Modell extrahiert, die für die Route relevant sind.

Als dritter Schritt wird für jede Gruppe, die keinen Übergang darstellt, eine räumliche Analyse durchgeführt, um zu bestimmen, welcher Landmark benutzt werden soll, um diesen Teil der Route zu identifizieren. Dafür wird ein Verfahren benutzt, was grafisch in Abbildung (3) zu sehen ist. Dadurch wird auch erkannt, welche Landmarks erreicht werden und an welchen vorbei gelaufen wird.

Bei der Identifizierung der Landmarks, an denen vorbei gelaufen wird, muss gewährleistet werden, dass diese Landmarks nicht zu nah an den Start- und Endpunkten der einzelnen Gruppen sind. Bei der Identifizierung der erreichten Landmarks müssen wir feststellen, ob die Distanz zwischen der Route und des Landmarks bei einem optimalen Winkel kleiner ist, als ein vordefinierter Schwellenwert.

Im vierten Schritt versuchen wir, Gruppen zu identifizieren, die in einer komprimierten Einzelanweisung auftauchen sollen. Dafür identifizieren wir zuerst Gruppen, die eine vordefinierte Länge unterschreiten und keinen assoziierten Landmark oder Zone haben. Um sicher zu stellen, dass wir nicht die komplette Route in einer langen Anweisung komprimieren, limitieren wir die Kompression auf zwei Segment-Längen.

Schließlich verwenden wir die resultierenden komprimierten Gruppen mit Landmarken, Zonen, Notizen und Abbiege-Informationen, um die Anweisungen zu erzeugen. Eine einfache Grammatik wird für die automatische Generierung der Anweisungen benutzt, so dass mehrere Sprachen leichter unterstützt werden können. Eine einfache Version der Templates, die wir dafür verwenden ist unten angegeben. Die eckigen Klammern zeigen optionale Teile an. Eine Beispielroute mit den resultierenden Textzusammenfassungen ist in Abbildung (4) zu sehen.

A) Nehmen Sie den <LinkType> [auf Ebene <targetLevel>] [in Richtung <NextZone>] [dann wiederum <TurnType> [und dann sofort <nextTurnType>]]

B) Gehen Sie [durch <current-Zone>] [vorbei an <passingPlace>] [bis <ErreichtPlace>] [dann wiederum <TurnType> [und dann sofort <nextTurnType>]] [in Richtung <nextZone>].

Tracking

Schritt-für-Schritt-Navigation erfordert die Lokalisierung des*der Nutzer*in auf der Route. Basierend auf der verbleibenden Route kann die Anwendung die geeignete Beschreibung für die nächste Aktion generieren, auf dem Bildschirm darstellen und gegebenenfalls sprachlich ausgeben.

Die Bestimmung der verbleibenden Strecke wird anhand der Strategie gemacht, die in Abbildung 5 zu sehen ist. Diese Strategie funktioniert gleichermaßen für den Außen- und den Innenbereich und basiert auf der Berechnung des nächsten Punktes (c1–3) zu jedem Liniensegment (S1–3), die die Route definiert. Dadurch werden die Linien berechnet, die senkrecht zu jedem Segment liegen und die durch die erfasste Position des*der Nutzer*in laufen. Das Segment, das die kleinste Distanz aufweist, wird dann als die Position des*der Nutzer*in auf der Route betrachtet.

Ein wesentlicher Unterschied zwischen Indoor- und Outdoor-Routen besteht darin, die Segmente auf verschiedenen Etagen voneinander zu unterscheiden und nur die Segmente zu nehmen, die auf der gleichen Ebene wie die erfasste Position des*der Nutzer*in sind.

Falls die minimale Distanz nicht berechnet werden kann, zum Beispiel weil der*die Benutzer*in sich auf einer Etage befindet, die nicht direkt zu der Route gehört oder die berechnete Distanz zu der zur Zeit aktiven Route zu groß ist, geht das System davon aus, dass der*die Benutzer*in sich nicht auf der Route befindet, vermutlich in eine andere Richtung gelaufen ist und eine neue Berechnung der Route erfolgen muss, damit die richtigen Anweisungen dargestellt werden können.

Im Laufe des SIMON-Projekts haben wir mit Anweisungen auf verschiedenen Granularitätsniveaus experimentiert. Diese gingen von den üblichen Anweisungen wie bei der Outdoor-Navigation (z.B. „in X Meter bitte nach links abbiegen“) bis hin zu allgemeinen Anweisungen wie „Gehen Sie in Richtung Norden“. Bei den Tests mit den Endnutzer*innen haben wir aber festgestellt, dass die Anweisungen, die am besten verstanden und als korrekt eingestuft werden, diejenigen sind, die mit Landmarks arbeiten und nicht unbedingt davon ausgehen, dass die Positionierung eine maximale Ungenauigkeit aufweist. Nicht nur die Positionierung, sondern auch die Bestimmung der Bewegungsrichtung anhand des Magnetometer-Sensors ist nicht in jedem Fall fehlerfrei zu schaffen und darf auch nicht uneingeschränkt benutzt werden. Stromleitungen im Gebäude oder Stahlkonstruktionen spielen hierbei eine besonders wichtige Rolle und führen oft zu Fehlern, die sich dann in falsche Anweisungen übersetzen lassen.

Erfahrungen

Im Laufe des Projektes haben wir 18 Monate gebraucht, bis wir die Backend-Services in der Cloud entwickelt haben, die Datenquellen integriert haben und die erste Version der Client-Anwendungen für Android mit Hilfe unserer Nutzer*innengruppen frei verfügbar gemacht haben. Danach haben wir

kleine Nutzertests gemacht, die uns erlaubt haben, weitere Entwicklungen und Iterationen mit den Nutzern zu machen. Dadurch haben wir 33 fortlaufende Versionen der Applikation innerhalb von fünf Monaten entwickelt, die dann als Beta-Version zu einer kleinen Gruppe von Tester*innen verschickt worden ist. Nachdem diese Tests erfolgreich gewesen sind, haben wir die öffentliche Version der Android Applikation gelauncht und nebenher die Portierung zu iOS initiiert. Sechs Monate später haben wir die iOS Version veröffentlicht. Seitdem haben wir weitere Updates und Bugfixes zur Verfügung gestellt, die die Anwendungen immer besser und einfacher nutzbarer gemacht haben.

Wir haben unsere Anwendung zusätzlich instrumentiert, um verstehen zu können, wie die Endnutzer*innen mit unserer App interagieren. Dafür haben wir ein Logging-Framework integriert, das UI-Interaktionen sowie Interaktionen zwischen der App und des Cloud-Backends protokolliert. Wir haben selbstverständlich Vorkehrungen getroffen, um sicherzustellen, dass die erfassten Daten nicht die Identität des*der Nutzer*in offenlegen und nur von uns sichtbar sind. Wir verwenden auch Pseudonyme, die bei der Installation der App neu generiert werden und sich auch manuell ändern lassen können. Darüber hinaus haben wir die Nutzer*innen beim ersten Start der Anwendung darüber informiert, dass Nutzungsdaten erfasst und für Optimierungs- und Verbesserungszwecke anonymisiert in der Cloud evaluiert werden.

Bis heute wurde die Anwendung 1.065 Mal installiert (Android: 803, iOS: 262), von denen schätzungsweise 415 Installationen derzeit noch aktiv sind. Gemeinsam haben die Benutzer*innen nach 13.831 Standorten gesucht, 1.748 Anfragen für Routen gestellt und 292 Fahrpläne ausgesucht. Interessanterweise wurde die Karten-basierte Explorationsansicht 2.806 mal benutzt, was ähnlich ist wie die Ansicht, die für die Berechnung der Routen benutzt wird (2.212 mal). Es sieht deswegen so aus, dass die Nutzer*innen die Routen-Berechnungs-Ansicht sechsmal häufiger nutzen als die Fahrplaninformationsansicht. Dieses Verhalten unterstreicht die Wichtigkeit, solche Fahrplaninformationen direkt in der Routen-Ansicht darzustellen.

Positiv zu bewerten ist auch die Tatsache, dass alle komplexere Berechnungen in der Cloud durchgeführt werden und die Client-Anwendungen „lediglich“ als Darstellungsoberfläche für die Cloud-Services dienen. Durch die Nutzung einer Cloud-basierten Infrastruktur können verschiedene mobile Applikationstechnologien schneller unterstützt werden. Auch die Verbesserung der Routing-Algorithmen oder Bugfixes in der Software können einfacher durchgeführt werden und in den meisten Fällen benötigen sie nicht mal die Installation einer neuen App. Ausgeschlossen sind natürlich Änderungen, die mit der Benutzer*innenoberfläche der einzelnen Apps zu tun haben.

Zusammenfassung

SIMON Mobile ist eine der wenigen Anwendungen, die in der Lage ist, multimodale Indoor- und Outdoor-Navigation nahtlos zu ermöglichen. Darüber hinaus unterstützt das System nativ auch Nutzer*innen mit Mobilitäts- sowie Seheinschränkungen und passt die Routenberechnung und die Darstellung der Ergebnisse der entsprechenden Zielgruppe an. Die Cloud-basierte Infrastruktur und die Tatsache, dass die Berechnungen in der Cloud durchgeführt werden, hat uns ermöglicht, die Portierung unserer Lösung für verschiedene Plattformen sehr schnell durchzuführen. Das Feedback der Nutzer*innen in Madrid, Parma, Lissabon und Reading ist hervorragend und mit dem ausdrücklichen Wunsch gekoppelt, das System auch nach Ende des Projekts weiter laufen zu lassen und mit zusätzlicher Funktionalität zu erweitern.

Summary

Three researchers at the University of Duisburg-Essen opened a company called Locoslab after several years of research in the area of localization. This article discusses the experiences of the research team in the design, development and real-life testing of SIMON Mobile, a mobile application that uses a cloud infrastructure to provide support for mobility impaired users for different cities like Madrid, Parma, Lisbon and Reading. The application is the main product of three years of collaborative research as part of the SIMON European project, that involves researchers from Spain, Italy, Portugal and Germany. The main asset of the application is the seamless integration of indoor and outdoor navigation routes and the fact that the computed routes take into account the characteristics of the user in terms of capabilities and preferences. At the time of this writing, SIMON Mobile has been available for several months at the Google and Apple App Stores and has been downloaded over 1000 times with very good user feedback. Additionally, all of the operations from the computation and retrieval of the map tiles to the algorithm that computes the optimal route for the user, are performed in the cloud. The client application simply acts as the local device where the information from the cloud is presented in the appropriate form to the user. This has the advantages that routing algorithms, for example, can be improved on the cloud backend and do not require the update of the application on the client side in order to use the new and improved services.

Danksagung

Wir danken allen Mitgliedern unserer Benutzer*innengruppen, die wertvolles Feedback für die Entwicklung der SIMON Mobile Apps geliefert haben. Darüber hinaus danken wir den Vertretern der Behörden von Lissabon, Parma, Reading und Madrid für den Zugang zu den Datenquellen, die für die Applikation benutzt wurden sowie ihr Feedback und ihre Hilfe während der Entwicklung und Installation der Cloud Dienste, die für die Applikationen notwendig waren.

Anmerkungen/Literatur

1) 2017. Google Maps. (2017). https://www.google.com/maps Besucht am: 2017-02-02.
2) M. Handte, S. Foell, S. Wagner, G. Kortuem, and P. J. Marron. 2016. An Internet-of-Things Enabled Connected Navigation System for Urban Bus Riders. IEEE Internet of Things Journal 3, 5 (Oct 2016), 735–744. DOI: http://dx.doi.org/10.1109/JIOT.2016.2554146
3) Andrew J. May, Tracy Ross, Steven H. Bayer, and Mikko J. Tarkiainen. 2003. Pedestrian navigation aids: information requirements and design implications. Personal and Ubiquitous Computing 7, 6 (2003), 331–338. DOI: http://dx.doi.org/10.1007/s00779-003-0248-5
4) OpenStreetMap Project. 2017. OpenStreetMap Wiki. http://wiki.openstreetmap.org/wiki/Slippy_map_tilenames. (2017).

Der Autor

In Spanien geboren, studierte **Pedro José Marrón** das Fach „Computer Engineering“ an der University of Michigan in Ann Arbor bis 1998. Ab 1999 arbeitete er an der Universität Freiburg an seiner Doktorarbeit, die er 2001 mit Auszeichnung abschloss. Von 2003 bis 2007 leitete er die Arbeitsgruppe „Mobile Data Management and Sensor Network“ an der Universität Stuttgart. Seine erste Professur führte Pedro Marrón 2007 an die Universität nach Bonn. Dort leitete er in der Computerwissenschaft die Arbeitsgruppe „Sensor Networks and Pervasive Computing“. Seit 2009 ist er Professor an der Universität Duisburg-Essen und war bis 2015 gleichzeitig führender Wissenschaftler am Fraunhofer Forschungsinstitut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE) in Wachtenberg. Pedro Marrón ist einer der Gründer der Locoslab GmbH, einem Unternehmen, das kostengünstige Lösungen zur Lokalisierung in geschlossenen Räumen anbietet. Ferner ist er Vorsitzender der Steuerungskommittees von UBICITEC, kurz für „European Center for Ubiquitous Technologies and Smart Cities“. An diesem virtuellen Zentrum sind etwa 20 europäische Partner aus Wissenschaft und Industrie beteiligt. Das Ziel von UBICITEC ist die Koordination der Forschungsvorhaben zum Thema Smart Cities.

Das Aufkommen des Cloud Computing hat gewaltige Auswirkungen auf die IT-Industrie. Unternehmen wie Google, Amazon und Microsoft, bieten immer leistungsfähigere, zuverlässigere und kostengünstigere Cloud Plattformen an.Unternehmen mit Rechenbedarf hinterfragen zunehmend den Nutzen ihrer eigenen IT-Infrastruktur und prüfen, inwieweit sie durch Nutzung von Cloud Computing Fixkosten abbauen und bedarfsgerecht gegen variable Kosten austauschen können.

Virtuelles Risiko oder Segen?

Cloud Computing in der Wirtschaftsprüfung

Von Ludwig Mochty und Michael Wiese

Cloud Computing gehört zweifelsohne zu den großen IT-Trends der letzten Jahre. Mit Hilfe des Cloud Computing[1] können IT-Ressourcen wie Rechnerleistung und Speicherplatz, über das Internet bedarfsgerecht angemietet und wieder an den IT-Dienstleister zurückgegeben werden. Dadurch wird die traditionelle Rolle des IT-Dienstleisters zweigeteilt: In einen Infrastrukturanbieter, der Rechnerplattformen betreibt und IT-Infrastruktur gegen nutzungsabhängige Preise vermietet, und einen IT-Dienstleister, der von einem oder mehreren Infrastrukturanbietern IT-Ressourcen anmietet, um IT-Dienstleistungen für einen End-User zu erbringen.

Das Aufkommen des Cloud Computing hat gewaltige Auswirkungen auf die IT-Industrie. Unternehmen wie Google, Amazon und Microsoft, bieten immer leistungsfähigere, zuverlässigere und kostengünstigere Cloud Plattformen an. Unternehmen mit Rechenbedarf hinterfragen zunehmend den Nutzen ihrer eigenen IT-Infrastruktur und prüfen, inwieweit sie durch Nutzung von Cloud Computing Fixkosten abbauen und bedarfsgerecht gegen variable Kosten austauschen können.

Tatsächlich kann eine solche Reorganisation der firmeneigenen IT-Landschaft eine Reihe von attraktiven Vorteilen mit sich bringen: (1) Es bedarf keiner investiven Vorleistungen. Der*Die Nutzer*in mietet je nach Bedarf IT-Ressourcen von der Cloud und zahlt für deren Nutzung. (2) Es lassen sich IT-Verarbeitungskosten senken, weil die in Anspruch genommene Kapazität nicht am Spitzenbedarf ausgerichtet sein muss. Angemietete IT-Ressourcen können wieder freigegeben werden, sobald der Bedarf sinkt. (3) IT-Dienstleistung wird auf diese Weise flexibel skalierbar: Mehrere IT-Infrastruktur-Anbieter können ihre Ressourcen zusammenlegen und erforderlichenfalls wieder entflechten. IT-Dienstleistungsanbieter sind durch den kurzfristigen Zugriff auf einen solchen Ressourcenpool in der Lage, sich geradezu blitzartig an Schwankungen in der Nachfrage nach IT-Dienstleistungen anzupassen. (4) Der Zugriff auf in der Cloud angebotene IT-Dienstleistungen ist üblicherweise sehr leicht realisierbar, weil er über das Internet erfolgt. (5) Ein IT-Dienstleistungsanbieter kann sein Hardware-Risiko an IT-Infra-

Ludwig Mochty. Foto: Vladimir Unkovic

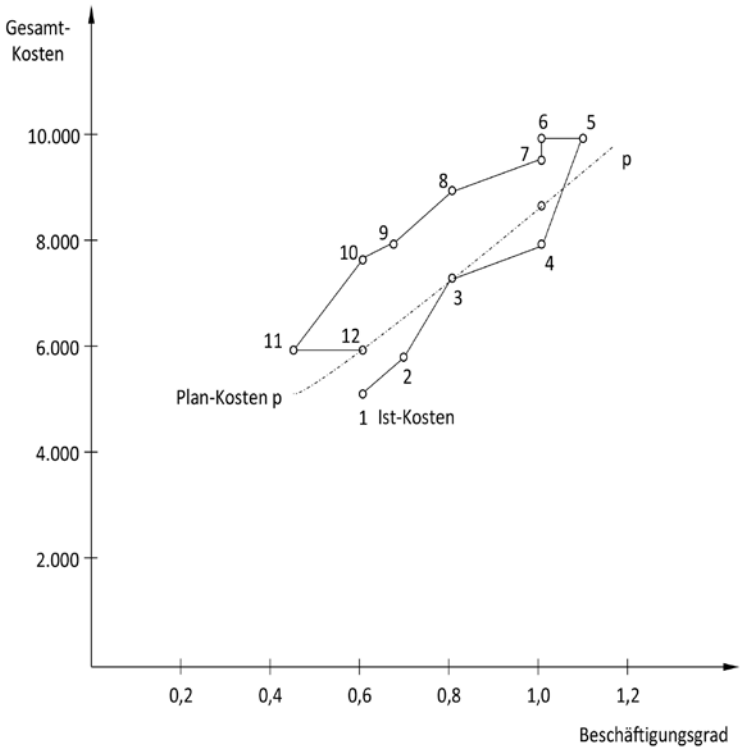
strukturanbieter outsourcen, die oftmals dafür besser ausgerüstet sind, und kann damit die hardwareabhängigen Instandhaltungs- und Personalkosten einsparen. Natürlich hat Cloud Computing für den End-User nicht nur Vorteile – aber davon später.

Betriebswirtschaftliche Fragestellungen

Für das Cloud Computing gibt es bisher zwar eine Reihe von Verbrauchserfassungssystemen, allerdings sind diese lediglich auf die mitlaufende Erfassung der Leistungs- und Ressourceninanspruchnahme anhand technischer Metriken (CPU-Zeiten, GigaByte, etc.) ausgerichtet, aber mangels betriebswirtschaftlicher Fundierung nicht geeignet, die kaufmännische Entscheidungsfindung zu unterstützen. Dies hat negative Auswirkungen auf die verursachungsgerechte Preis- und Produktpolitik, auf die betriebswirtschaftliche Optimierung des IT-Produktionsbetriebs sowie auf die Investitionspolitik von Cloud Computing-Dienstleistungsanbietern. Schließlich fehlen damit auch IT-Prozesskostensätze, die es einem Unternehmen erlauben würden, die Entscheidung zwischen Eigenfertigung und Fremdbezug (Eigene IT versus Cloud Computing) kaufmännisch fundiert zu fällen.

Durch das Cloud Computing wird deshalb nicht nur die Computerwissenschaft, sondern auch die Betriebswirtschaftslehre (BWL) herausgefordert, denn durch seine dynamische Virtualität stellt Cloud Computing klassische betriebswirtschaftliche Fragestellungen in ein völlig neues Licht. Einige dieser Fragestellungen seien im Folgenden explizit angesprochen.

Das kaufmännische Management von Cloud Computing-Systemen verlangt nach einer Buchführung und insbesondere nach einer Kosten- und Leistungsrechnung, die geeignet ist, Prozesse in virtuellen Organisationen verursachungsgerecht abzubilden. Die traditionelle Kostenrechnung wurde ursprünglich für die arbeitsintensive Massenfertigung relativ homogener Erzeugnisse konzipiert. Sie ist nach betrieblichen Funktionen (Materialbeschaffung und -lagerung; Fertigung; Verwaltung und Vertrieb) und nicht nach Geschäftsprozessen gegliedert. Deshalb kommt es häufig zu einer nicht verursachungsgerechten Proportionalisierung großer Kostenanteile, wenn diese Art der Kostenrechnung auf flexible industrielle Fertigungsprozesse angewendet wird. Und zwar umso mehr, je heterogener, komplexer und variantenreicher die Erzeugnisstruktur ist und je stärker die Gemeinkosten von anderen Kostentreibern als den menschlichen Leistungsstunden abhängig sind. Von dieser Problematik ist die Kalkulation von Cloud Computing-Leistungen in extrem verschärfter Form betroffen. Da also die traditionelle Kostenrechnung nicht geeignet ist, Cloud Computing-Leistungen verursachungsgerecht abzubilden, bedarf es eines Kostenrechnungssystems, das speziell für Prozessstrukturen konzipiert ist. Ein solches wurde von der Betriebswirtschaftslehre in Form der Prozesskostenrechnung[2] (Activity Based Costing, ABC) entwickelt. Es konnte sich aber wegen des großen Arbeitsaufwands, der mit der Erhebung der Teilprozesse und Aktivitäten sowie mit der Erfassung der Prozessgrößen verbunden ist, nicht flächendeckend durchsetzen. Hier bringt das Cloud Computing ganz entscheidende Vorteile mit sich. Denn anders als bei konventionellen Geschäftsprozessen steht beim Cloud Computing ein mitlaufendes prozessorientiertes Verbrauchserfassungssystem zur Verfügung, auf das betriebswirtschaftliche Abrechnungs- und Entscheidungsunterstützungssysteme in Echtzeit zurückgreifen können. In dieser Hinsicht verfügt Cloud Com-



(1) Darstellung des asymmetrischen Kostenverlaufs (Hysteresis).
Quelle: in Anlehnung an Heinen, E.: Zum Problem der Kostenremanenz. In: Zeitschrift für Betriebswirtschaft (ZfB), 36. Jahrgang 1966, S. 1–18, hier: S. 3

puting gewissermaßen über eine Art von „Fahrtenschreiber“ (Meta-Daten, die wie Kuppelprodukte anfallen), sodass es einer aufwendigen und fehleranfälligen manuellen Erfassung der Prozessgrößen nicht bedarf. Es wird noch zu prüfen sein, welche Verfeinerungen der Prozesskostenrechnung durch die virtuelle Natur des Cloud Computing erforderlich werden. Sollte es aber gelingen, am Beispiel von Cloud Computing ein funktionsfähiges, betriebswirtschaftlich fundiertes Kosten- und Leistungssystem zu entwickeln, müsste dieses auch auf andere virtuelle Organisationen und Unternehmen übertragbar sein.

Mit dem Cloud Computing kann aus betriebswirtschaftlicher Sicht ein interessantes weiteres Phänomen verbunden sein, das erstmals 1927 beschrieben worden ist. Es handelt sich um das Phänomen der Kostenremanenz[3], das im Zusammenhang mit Beschäftigungsschwankungen zu beobachten ist: Der Zuwachs der Kosten nimmt bei einer Beschäftigungszunahme (hier: bei der Steigerung der nachgefragten Rechenleistung) einen anderen Verlauf als bei der Beschäftigungsabnahme (hier: Absenkung der nachgefragten Rechenleistung). Dieser asymmetrische Kostenverlauf in Form einer Hysteresis wurde ursprünglich in Analogie zur magnetischen Induktion beim Auf- und Abbau eines elektromagnetischen Felds beschrieben und vermutlich deshalb von der BWL bisher kaum beachtet. Sollte dieses Phänomen beim Cloud Computing auftreten, stünden erstmals Massendaten zur Verfügung, um umfangreiche und detailreiche Kostenanalysen durchführen zu können.

Allgemein ist die Kostenrechnung nur geeignet, um kurzfristige Entscheidungen zu fundieren. Für die Unterstützung langfristiger Investitions- und Desinvestitionsentscheidungen dient die Investitionsrechnung. Im Zusammenhang mit dem Cloud Computing interessiert dabei insbesondere die Fragestellung: In welchem Ausmaß soll im Kalkulationszeitpunkt bei der Wahl der Größe einer Anlage die in der Zukunft erwartete Steigerung der Nachfrage berücksichtigt werden? Die Wirtschaftlichkeit von erweiterbaren Anlagen unter Berücksichtigung einer zukünftigen Bedarfszunahme wurde bereits 1939 im Zusammenhang mit der Erweiterung eines Telefonnetzwerks analysiert[4]. Auch diese in der betriebswirtschaftlichen Ausbildung wenig beachtete Fragestellung erfährt im Zusammenhang mit Investitionen in dynamische virtuelle Organisationen ein Redivivum.

Besonders schwierig zu beantworten ist für die BWL zurzeit noch die Frage nach der zweckmäßigsten Gestaltung nutzungsabhängiger Preise für die „Rechenleistung aus der Steckdose“. Eine betriebswirtschaftlich fundierte Lösung dieser Aufgabenstellung setzt indes voraus, dass die Kosten von Cloud Computing-Serviceleistungen zuverlässig und verursachungsgerecht kalkuliert werden können. Und hier steht die BWL erst am Anfang.

Rechtliche Einordnung

Voraussetzung für den Einsatz der genannten Entscheidungsunterstützungskalküle ist die korrekte Erfassung der verschiedenen Aspekte des Cloud Computing in der Buchführung. Denn diese bildet die Datenbasis für alle weiteren Auswertungen. Deshalb soll zuerst eine grobe rechtliche Einordnung erfolgen, bis wir uns mit der Frage beschäftigen werden, welche Anforderungen an die Sicherheit und Ordnungsmäßigkeit des Rechnungswesens im Zusammenhang mit Cloud Computing zu erfüllen sind und wie die Einhaltung dieser Anforderungen durch den freien Beruf der Wirtschaftsprüfer überwacht wird.

Die Grenzen zwischen Cloud Computing und IT-Outsourcing verschwimmen zusehends, da klassische IT-Outsourcing-Dienstleistungsunternehmen zunehmend Techniken des Cloud Computing einsetzen.[5] Man kann Cloud Computing als eine besondere Form des IT-Outsourcings ansehen. Ein besonderer Vertragstyp für Cloud Computing-Leistungen existiert im deutschen Recht nicht. Relevant sind insbesondere der Mietvertrag (§§ 535ff. BGB), der Werkvertrag (§§ 631ff. BGB) und der Dienstvertrag (§§ 611ff. BGB). Für die Beurteilung von Cloud Computing-Verträgen ist auf die enthaltenen Teilleistungen abzustellen, denn in der Regel handelt es sich um gemischte Vertragstypen. Cloud Computing-Leistungen umfassen im Wesentlichen „Software as a Service“ (SaaS), „Infrastructure as a Service“ (IaaS) und „Platform as a Service“ (PaaS)[6].

- Bei SaaS besteht die Dienstleistung in der Bereitstellung einer IT-Anwendung auf dem Server des Anbieters. Das auslagernde Unternehmen hat keinen Einfluss auf die der genutzten IT-Anwendung zugrunde liegenden IT-Infrastruktur. Die Benutzeroberfläche der IT-Anwendung wird in der Regel innerhalb eines Browserfensters dargestellt, während die Anwendung selbst in Echtzeit auf dem Server des Anbieters ausgeführt wird. Die Installation der Software auf den Endgeräten des auslagernden Unternehmens ist nicht vorgesehen. SaaS-Verträge sind Dauerschuldverhältnisse[7]. Das Leistungsentgelt (Software-Miete) richtet sich regelmäßig nach der Anzahl der vereinbarten Softwarenutzern. Viele Unternehmen entscheiden sich für den Bezug von SaaS (z.B. ERP-Systeme oder Office-Dienste wie E-Mail-Programme), weil ihre eigene Software-Architektur den Anforderungen volatiler Märkte und der gestiegenen Innovationsgeschwindigkeit nicht mehr entspricht.
- IaaS-Servicemodelle bestehen in der flexiblen Anmietung von IT-Ressourcen (z.B. Rechenleistung und Speicherkapazität), die das auslagernde Unternehmen für seine Unternehmenszwecke nutzen kann. Diese Infrastruktur besteht nicht aus

realen, sondern aus virtuellen Servern, die physisch im Serverzentrum des Cloud Computing-Anbieters untergebracht sind. Die Mitarbeiter des auslagernden Unternehmens greifen über einen Internetbrowser oder mobile Endgeräte auf den IaaS-Service zu und nutzen diese Infrastruktur zum Aufbau eigener Services zum internen oder externen Gebrauch[8]. Das IaaS-Servicemodell wird in der Regel durch einen gemischten Vertragstyp abgebildet, der miet-, dienst- und werkvertragliche Elemente beinhaltet[9]. Das Leistungsentgelt richtet sich in der Regel nach dem Umfang der während der Abrechnungsperiode in Anspruch genommenen Ressourcen (z.B. Rechenleistung oder Speicherkapazität).

• PaaS bezeichnet die bedarfsorientierte Bereitstellung von IT-Infrastruktur (z.B. Rechenleistung, Speicherkapazität und Internetanbindung, Laufzeitumgebungen) und Entwicklerwerkzeugen zur Entwicklung und Ausführung von Cloud-basierten Softwarelösungen. Die primäre Zielgruppe von PaaS-Angeboten sind daher Software-Entwickler[10].

Cloud Computing-Anbieter bieten ihren Kunden in der Regel standardisierte Leistungen an. Daher werden insbesondere für den Online-Bezug dieser Leistungen anstelle von Individualvereinbarungen vorformulierte Allgemeine Geschäftsbedingungen (AGB) eingesetzt.

Service Level Agreements

Das Service Level Agreement stellt üblicherweise einen gesonderten Teil oder eine Anlage des Cloud Computing-Vertrags dar. Es bezieht sich auf die Leistungsbeschreibung und regelt, welche Service Levels gelten (d.h. welche Funktionalitäten und welche Qualität geschuldet sind) und welche Rechtsfolgen eine Vertragsverletzung auslöst. Standardbestandteile des Service Level Agreements sind Regelungen zur Verfügbarkeit der Cloud Computing-Leistung (üblicherweise ausgedrückt durch einen Prozentwert in Bezug auf einen bestimmten Betrachtungszeitraum), zu Entstörzeiten im Falle einer Störung der Cloud Computing-Leistungen, zu den Rechtsfolgen einer Schlechtleistung, zu Vertragsstrafen oder pauschaliertem Schadensersatz sowie zum Kündigungsrecht im Falle einer Nichteinhaltung von Service Levels[11]. Im Service Level Agreement sollten ebenfalls Regelungen zum Gerichtsstandort, dem anwendbaren Recht sowie die Vertragssprache enthalten sein. Darüber hinaus sind Eigentums- und Urheberrechte an Daten, Systemen, Software und Schnittstellen festzulegen[12].

Datenschutzrecht und Compliance

Charakteristisch für Cloud Computing ist die Weitergabe von Daten vom auslagernden Unternehmen an den Dienstleister. Das deutsche Datenschutzrecht kommt zur Anwendung, sobald personenbezogene Daten im Inland erhoben oder verwendet werden. Sofern die Datenerhebung oder -verarbeitung durch ein Unternehmen mit Sitz in der EU oder im EWR erfolgt, gilt nach § 1 Abs. 5 BDSG das Datenschutzrecht im Ansässigkeitsstaat dieses Unternehmens. Nach §§ 4b, 4c BDSG werden Datenverarbeitungen innerhalb der EU beziehungsweise des EWR wie Datenverarbeitungen im Inland behandelt. Wenn Cloud Computing-Dienstleistungen durch Unternehmen oder durch physische Rechenzentren in Drittstaaten erbracht werden, ist unter Datenschutzgesichtspunkten die Zulässigkeit des Auslandstransfers zu klären[13]. Das auslagernde Unternehmen hat sich als Auftraggeber nach § 11 BDSG bereits vor Beginn der Datenverarbeitung und im Anschluss regelmäßig von der Einhaltung der Bestimmungen des Datenschutzes zu überzeugen.

Zusätzlich zu den datenschutzrechtlichen Anforderungen ist sicherzustellen, dass sich der Dienstleister an die sonstigen rechtlichen Bestimmungen (Compliance) des auslagernden Unternehmens hält. Hierbei liegt es in der Verantwortung des auslagernden Unternehmens, die konkreten rechtlichen Anforderungen zu benennen. Anforderungen können sich zum Beispiel ergeben aus dem Telekommunikationsgesetz (TKG), der Abgabenordnung (AO) bei der Verarbeitung steuerrechtlicher Daten, dem Handelsgesetzbuch (HGB) bei der Verarbeitung buchführungsrelevanter Daten und dem Strafgesetzbuch (StGB).

Anforderungen an die Sicherheit und Ordnungsmäßigkeit

Die gesetzlichen Vertreter müssen die Einhaltung der gesetzlichen Anforderungen an die Sicherheit und Ordnungsmäßigkeit der IT-gestützten Rechnungslegung sicherstellen. Die Sicherheitsanforderungen umfassen Vertraulichkeit, Integrität, Verfügbarkeit, Autorisierung, Authentizität und Verbindlichkeit[14]. Zu den Ordnungsmäßigkeitskriterien zählen Vollständigkeit (§ 239 Abs. 2 HGB), Richtigkeit (§ 239 Abs. 2 HGB), Zeitgerechtheit (§ 239 Abs. 2 HGB), Ordnung (§ 239 Abs. 2 HGB), Nachvollziehbarkeit (§ 238 Abs. 1 Satz 2 HGB), Unveränderlichkeit (§ 239 Abs. 3 HGB)[15]. Das Management implementiert zu diesem Zweck Grundsätze, Verfahren und Maßnahmen (Regelungen), die gerichtet sind auf: die Sicherung der Wirksamkeit und Wirtschaftlichkeit der Geschäftstätigkeit (hierzu gehört auch der Schutz des Vermögens, einschließlich der Verhinderung und Aufdeckung von Vermögensschädigungen), zur Ordnungsmäßigkeit und Verlässlichkeit der internen und externen Rechnungslegung sowie zur Einhaltung der für das Unternehmen maßgeblichen rechtlichen Vorschriften. Die Gesamtheit dieser Regelungen wird als internes Kontrollsystem bezeichnet. Das interne Kontrollsystem

umfasst die Komponenten Risikobeurteilung, Kontrollumfeld, Risikobeurteilungen, Kontrollaktivitäten, Information und Kommunikation und Überwachung des internen Kontrollsystems durch das Management des Unternehmens[16]. Relevante Standards, die für eine Implementierung herangezogen werden können, sind COBIT©, COSO©, ISO EC 27001 und IDW RS FAIT 1.

Die gesetzlichen Ordnungsmäßigkeitsanforderungen einschließlich der Beachtung der Grundsätze ordnungsmäßiger Buchführung (§ 239 Abs. 4 HGB) und die damit verbundenen Anforderungen an die Sicherheit der IT-gestützten Rechnungslegung gelten uneingeschränkt auch für den Fall des IT-Outsourcings. Mit der Auslagerung von betrieblichen Funktionen im Rahmen des IT-Outsourcings müssen die im IDW RS FAIT 1 formulierten Sicherheits- und Ordnungsmäßigkeitsanforderungen auch im Zusammenspiel mit den Dienstleistungsunternehmen – einschließlich eventueller Subdienstleistungsunternehmen – erfüllt werden. Aufgrund der zunehmenden Auslagerung von Prozessen und Daten auf Subdienstleister wurde der IDW RS FAIT 5 in Zusammenarbeit mit dem Fachausschuss IT entwickelt, um den Risiken aus der Auslagerung der Daten und Prozesse nach den Sicherheits- und Ordnungsmäßigkeitsanforderungen des HGB zu genügen. IDW FAIT 5 ergänzt und konkretisiert die aus den §§238, 239 und 257 HGB sowie der IDW RS FAIT 1 (Grundsätze ordnungsmäßiger Buchführung bei Einsatz von Informationstechnologie) resultierenden Anforderungen an die Führung der Handelsbücher mittels IT-gestützter Systeme. FAIT 5 ist ab dem 1. Januar 2016 anzuwenden.

Die gesetzlichen Vertreter haben die Aufgabe, das Cloud Computing von der Entscheidung zur Auslagerung bis zur Beendigung zu steuern. Das interne Kontrollsystem muss auch im Hinblick auf die ausgelagerten Funktionen angemessen ausgestaltet und wirksam sein. Ausgehend von der Risikobeurteilung muss das auslagernde Unternehmen den sich durch das IT-Outsourcing ergebenden Risiken durch eine entsprechende Ausgestaltung des internen Kontrollsystems begegnen. Alle involvierten Subdienstleister sind ausnahmslos Bestandteil des internen Kontrollsystems.

Service Level Agreements

Die mit dem Dienstleister getroffenen Vereinbarungen müssen die Sicherheit und Ordnungsmäßigkeit der ausgelagerten Daten und Prozesse sicherstellen. Risiken aus nicht getroffenen Vereinbarungen müssen durch kompensierende Kontrollen adressiert werden.

Administration

Die gesetzlichen Vertreter des auslagernden Unternehmens sind für die Änderungen an den Programmen und Zugriffsberechtigungen des Dienstleisters verantwortlich. Das auslagernde Unternehmen muss sicherstellen, dass nur solche Änderungen durchgeführt und Berechtigungen vergeben werden, die die Sicherheit und Ordnungsmäßigkeit der gespeicherten, transportierten oder verarbeiteten Daten nicht gefährden.

Überwachung

Eine regelmäßige Berichterstattung und Auswertung des Dienstleisters hinsichtlich der Vereinbarungen in den Service Level Agreements ist erforderlich. Die unternehmenseigenen Risiken aus der Auslagerung und die dafür entwickelten Kontrollen müssen regelmäßig überwacht werden. Das auslagernde Unternehmen sollte ein umfassendes Bild vom Dienstleistungsunternehmen haben, zum Beispiel durch Einsichtnahme in aktuelle Service Organization Control (SOC) Berichte (IDW PS 951 n.F., ISAE 3402) oder in Prüfungsberichte der internen Revision.

Netzwerke

Die Infrastruktur des Dienstleisters muss separiert sein in vertrauenswürdige Netzwerke (z.B. Lokales Administrationsnetzwerk) und nicht sichere Netzwerke (z.B. End-User-Netzwerk). Deren Einstellungen sind durch das auslagernde Unternehmen regelmäßig zu überprüfen. Die für die Übermittlung der Daten relevanten Komponenten und die relevante Infrastruktur unterliegen den gleichen Integritäts- und Verfügbarkeitsanforderungen wie die Orte der Speicherung der Daten.

Nationale Besonderheiten

Die Behandlung von rechnungslegungsrelevanten Daten in Drittländern muss den Sicherheits- und Ordnungsmäßigkeitsanforderungen des HGB genügen. Besonderheiten im Steuer-, Straf- und Zivilrecht sowie des Datenschutzes dürfen zu keinem Wiederspruch mit der deutschen Gesetzgebung führen. Die gesetzlichen Vertreter des auslagernden Unternehmens sind auch nach Beendigung der vertraglichen Beziehungen mit dem Dienstleister für die Einhaltung der Ordnungsmäßigkeits- und Sicherheitsanforderungen verantwortlich. Darunter fallen beispielsweise Anforderungen an die Archivierung rechnungslegungsrelevanter Daten oder an die Datenlöschung auf Seiten des Dienstleistungsunternehmens.

Risikoorientierter Prüfungsansatz

Gegenstand der Abschlussprüfung ist neben dem Jahresabschluss die Ordnungsmäßigkeit der Buchführung. Hierbei untersucht der Abschlussprüfer in der Regel das eingerichtete interne Kontrollsystem. Da eine lückenlose Prüfung von Buchhaltung und Rechnungslegung wirtschaftlich nicht möglich ist und vom Gesetzgeber in § 317 HGB auch nicht verlangt wird, erfolgt die Abschlussprüfung auf Basis einer risikoorientierten Auswahl von

Geschäftsvorfällen und Abschlussposten. Durch Kombination von Art, Zeitpunkt und Umfang der Prüfungshandlungen sollen hinreichend viele (ausreichend) beweiskräftige (angemessene) Prüfungsnachweise erhalten werden, um das Prüfungsurteil im Bestätigungsvermerk zu begründen.

Ausgelagerte Dienstleistungen sind für die Prüfung des internen Kontrollsystems relevant, wenn sie dazu dienen, Daten über Geschäftsvorfälle oder betriebliche Aktivitäten zu speichern oder zu verarbeiten, die entweder direkt in die Rechnungslegung einfließen oder dem Rechnungslegungssystem als Grundlage für Buchungen zur Verfügung gestellt werden. Typische ausgelagerte Unternehmensfunktionen umfassen Beschaffung, Einkauf, Materialwirtschaft/Logistik, Produktion, Entwicklung, Buchhaltung, Vertrieb oder Personalabrechnung. Die Auslagerung dieser Funktionen hat zum Teil erhebliche Auswirkungen auf Aufbau- und Ablauforganisation, IT-Infrastruktur, IT-Anwendungen und IT-gestützte Geschäftsprozesse. In jüngerer Vergangenheit finden sich auch Auslagerungen von Analysekapazitäten im Bereich Controlling und Management („Big Data in der Cloud“) mit teilweise erheblichen unmittelbaren Risiken für die Steuerung von Unternehmen und die Finanzberichterstattung.

Beim Cloud Computing ergeben sich neben IT-Sicherheitsrisiken, Ordnungsmäßigkeitsrisiken besondere rechtliche Risiken in Bezug auf Organisation und Aufgabenteilung, Schnittstellen und genutzte Übertragungswege, eingesetzte Technologien, Datenspeicherung und Speicherort. Hierzu gehören auch Maßnahmen für den Not- und Katastrophenfall sowie Kontrollen, die eine Funktionsfähigkeit dieser Maßnahmen gewährleisten. Als Teil der Prüfung des internen Kontrollsystems hat der Prüfer die Wirksamkeit der dafür eingerichteten Kontrollen zu beurteilen.

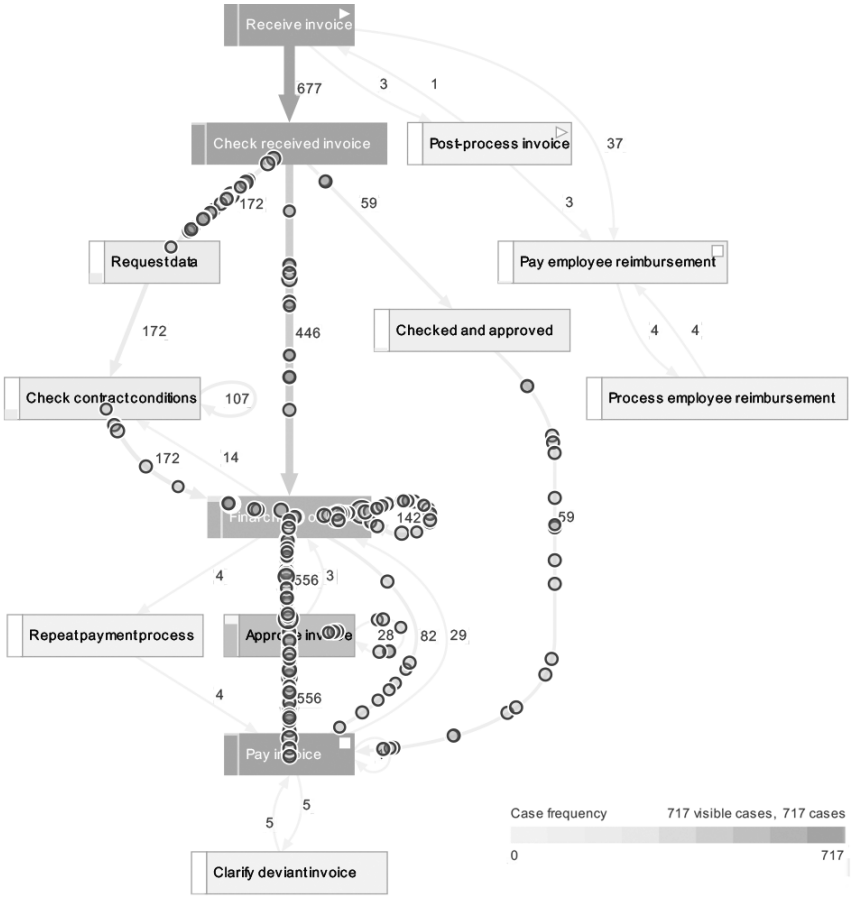
Verlust der Transparenz

Mit der IT-gestützten Automatisierung, dem Outsourcing und dem Schritt in die Cloud („Digitalisierung“) versuchen Unternehmen vor allem, die Kosten interner Abläufe zu reduzieren und interne Abläufe zu beschleunigen. Das kann dann gelingen, wenn Aktivitäten innerhalb der Prozesse häufig und gleichartig sind und daher standardisiert werden können. Durch Standardisierung lässt sich die Anzahl von Prozessvarianten reduzieren. Die IT-Integration zieht jedoch auch eine erhöhte Komplexität der Abläufe und damit einhergehend eine erhöhte Intransparenz nach sich. Für den Nicht-Standardfall sind Mitarbeiter häufig gar nicht ausgebildet oder es existiert systemseitig keine Möglichkeit, den Fall abzubilden. Für Prüfer IT-gestützter Systeme und Prozesse stellt sich das Prüfproblem deutlich schärfer als das Ausführungsproblem dar. Denn der Prüfer verfügt in der Regel nicht über die Detailkenntnisse der internen Prozesse wie beispielsweise eine interne Revision. Durch Beobachtungen von Prozessabläufen und Befragungen der beteiligten Mitarbeitern lässt sich diese Wissenslücke in komplexen IT-gestützten Systemen nicht schließen. Denn beobachten lässt sich nur der Standardfall. Sonderfälle sind entweder nicht vorgesehen, sind noch nicht aufgetreten oder werden nicht offengelegt, zum Beispiel im Fall vorsätzlicher Eingriffe. Aufgabe der Prüfer ist es aber, die tatsächliche Umsetzung der definierten Abläufe und die Zuverlässigkeit des Systems auch im Sonderfall zu beurteilen. Das gelingt zunehmend schlechter und hat negative Auswirkungen auf das Risikomanagement. Klarerweise sind moderne IT-Prüfer darin geschult, Risiken auf Ebene der IT-Infrastruktur und innerhalb der IT-Anwendung zu erkennen. Doch gerät der Bezug zur individuellen Ausführung der Prozessaktivitäten und zum Jahresabschluss als Prüfungsgegenstand oftmals aus dem Blick. Mitunter unterminiert die IT die Fähigkeit der Prüfer, die wesentlichen Risiken zu erkennen.

Process Mining

Digitale Systeme und Prozesse haben aber auch eine Eigenschaft, die sich zur Lösung des Problems der Komplexität und Intransparenz nutzen lässt. Um den gesetzlichen Ordnungsmäßigkeits- und Sicherheitsanforderungen gerecht zu werden, speichern moderne Enterprise Ressource Systeme (ERP) nahezu alle relevanten digitalen Verarbeitungsschritte von Geschäftsvorfällen, Systemaktivitäten und durchgeführten Kontrollen. Die Datentabellen mit den individuellen Verarbeitungsschritten (sogenannte Event Logs) für jeden Geschäftsvorfall können für Prüfungszwecke extrahiert werden. Unter Einsatz entsprechender Algorithmen werden diese Events zeitlich sortiert und zu gleichartigen Aktivitäten zusammengefasst. So gelingt es, die tatsächlichen Prozessabläufe zu rekonstruieren, zu visualisieren und für prüferische Zwecke zu analysieren. In Anlehnung an den Begriff „Data Mining“, dem „Schürfen“ nach Informationen aus Massendaten, wird der Ansatz als „Process Mining“ bezeichnet.

Erkennbar ist, dass die individuellen Rechnungen (dargestellt als kleine Kreise) unterschiedliche Wege durch das Verarbeitungssystem nehmen. In der dynamischen Anwendung werden außerdem die unterschiedlichen Durchlaufzeiten erkennbar. Bislang war der Prüfer auf Schilderungen der prozessbeteiligten Mitarbeiter nur eines typischen Falles, zum Beispiel einer Rechnungsprüfung, beschränkt. Durch die Komplexität sind bei der konventionellen Prüfung des internen Kontrollsystems Unvollständigkeiten und Irrtümer kaum zu vermeiden.



(2) Ablauf eines Kontrollprozesses für Eingangsrechnungen, beginnend mit dem Rechnungseingang bis zum Zahlungsausgang und zur Differenzenklärung.
Quelle: eigene Darstellung

Unser Forschungsschwerpunkt ist die methodische Integration des Process Mining in die Ablaufmodelle betriebswirtschaftlicher Prüfungen mit besonderem Schwerpunkt auf die Abschlussprüfung durch Wirtschaftsprüfer. Für die praktische Anwendbarkeit ist hierbei besonders relevant, inwieweit das Process Mining geeignet ist, die Ziele der internationalen Prüfungsstandards des IAASB für die Prüfung des internen Kontrollsystems zu erreichen. Die Prüfungsstandards geben ein zweistufiges Vorgehen vor. ISA 315 verlangt ein grundlegendes Verständnis für die Komponenten des internen Kontrollsystems: Kontrollumfeld, Risikobeurteilung, Kontrollaktivitäten, Information und Kommunikation sowie Überwachung des internen Kontrollsystems durch das Management.

Kontrollumfeld

Das Kontrollumfeld stellt den Rahmen dar, innerhalb dessen die Grundsätze, Verfahren und Maßnahmen der internen Kontrolle eingeführt und angewendet werden[17]. Es wird unter anderem bestimmt durch die Unternehmenskultur, Fachkompetenz der Mitarbeiter und die Aufbau- und Ablauforganisation. Unter Einsatz von Process Mining ist es möglich, viele der relevanten Eigenschaften des Kontrollumfelds, die besonders bei der Auslagerung von Funktionen nur schwer beobachtbar sind, indirekt prüfbar zu machen. Weil ERP-Systeme neben den reinen Events auch den Systemnutzer speichern, ist es möglich, große Teile der Aufbauorganisation zu rekonstruieren. Ein anschließender Vergleich mit der in Organigrammen und Arbeitsplatzbeschreibungen festgelegten Organisation zeigt oftmals erstaunliche Unterschiede zwischen dem, was Mitarbeiter tun, und dem, was sie eigentlich tun sollten. Identifiziert werden kritische Aktivitäten und Mitarbeiter, bei deren Ausfall Schaden droht. Gegebenenfalls sind Vertretungsregeln zu überarbeiten. Anhand von überdurchschnittlichen Bearbeitungszeiten werden Engpässe sichtbar, die auf Mängel in der Personalplanung schließen lassen. Häufige Wiederholungen von Aktivitäten und Rückkopplungen zeigen Nachlässigkeit oder mangelnde fachliche Eignung auf. Für den Prüfer wird erkennbar, ob das Management seiner Verantwortung für die Überwachung der Aufbau- und Ablauforganisation gerecht wird. Gerade in Krisenzeiten einer Unternehmung zeigt sich, dass interne Abläufe zunehmend degenerieren. Das Kontrollumfeld kann durch Process Mining wie ein Film beobachtet werden.

Risikobeurteilungsprozess

Unternehmen sind einer Vielzahl von finanziellen, rechtlichen, leistungswirtschaftlichen oder strategischen Risiken ausgesetzt. Diese Risiken müssen vom Unternehmen erkannt, analysiert und beurteilt werden[18]. Process Mining unterstützt hier in zweifacher Hinsicht. Zum einen können Risiken aus der Ablauforganisation identifiziert werden, die bisher regelmäßig erst zu spät erkennbar werden. Beispielsweise können Mängel im Prozess des Forderungsmanagements das Unternehmen in die Zahlungsunfähigkeit treiben. Zum anderen kann der Risikobeurteilungsprozess selbst mit Process Mining analysiert werden. Besonders relevant ist diese Anwendung für kapitalmarktorientierte Unternehmen, die verpflichtet sind, ein Risikofrüherkennungssystem einzurichten und permanent zu überwachen. Der Risikobeurteilungsprozess ist sicherlich für Outsourcing wenig geeignet. Diejenigen Risiken

Michael Wiese. Foto: Vladimir Unkovic

aber, die sich aus Outsourcing und Cloud Computing ergeben, müssen zwingend Bestandteil permanenter Überwachung auf Basis des Process Mining sein.

Kontrollaktivitäten

Als eigenständige Kontrollaktivität, beispielsweise im Rahmen der internen Revision, und bei der Aufbau- und Ablaufprüfung von Kontrollaktivitäten spielt Process Mining seine Stärke aus. Kontrollaktivitäten sind Grundsätze und Verfahren, die sicherstellen sollen, dass die Entscheidungen des Managements beachtet werden. Sie tragen dazu bei, dass notwendige Maßnahmen getroffen werden, um den Unternehmensrisiken zu begegnen[19]. Beispielsweise dienen Kreditwürdigkeitsprüfungen von Kunden dazu, Zahlungsausfällen und Bewertungsrisiken im Jahresabschluss vorzubeugen. Mit Process Mining kann beobachtet werden, wann und durch wen für welche Geschäftsvorfälle welche Verarbeitungsschritte durchgeführt wurden. Daraus abgeleitet wird offensichtlich, welche unterschiedlichen Varianten es gibt. Basierend auf dem vorgegebenen Ablauf können die zulässigen von den potentiell schädlichen Varianten getrennt werden. Es wird nicht nur bestimmt, dass Schwächen in der internen Kontrolle existieren, sondern auch, mit welcher Eintrittswahrscheinlichkeit, welche Schäden zu erwarten sind. Auf diese Weise kann das interne Kontrollsystem einem Stresstest unterzogen werden, wie er im Bankensektor bereits üblich ist.

Unternehmenskommunikation

Information und Kommunikation dienen dazu, dass die für die unternehmerischen Entscheidungen des Managements erforderlichen Informationen in geeigneter und zeitgerechter Form eingeholt, aufbereitet und an die zuständigen Stellen im Unternehmen weitergeleitet werden[20]. Mit Process Mining kann der Informationsaustausch anhand von Log-Files der Netze und Server analysiert werden. Durch Kombination mit Methoden der sozialen Netzwerkanalyse kann der Informationsfluss zwischen Systemnutzern visualisiert und analysiert werden. Entwicklungen hierzu stehen noch an ihrem Anfang, haben aber großes praktisches Potential.

Überwachung

Durch die Unternehmensleitung ist permanent zu beurteilen, ob das interne Kontrollsystem sowohl angemessen ist als auch kontinuierlich funktioniert. Darüber hinaus hat das Management dafür Sorge zu tragen, dass festgestellte Schwächen im internen Kontrollsystem in geeigneter Weise abgestellt werden[21]. Die dargestellten Aktivitäten und Merkmale des internen Kontrollsystems können durch das Process Mining im Zeitverlauf dargestellt werden. Durch Kombination mit den Hauptbuch- und Nebenbuchjournalen können Kennzahlen zur Vermögens-, Finanz- und Ertragsentwicklung mit prozessbezogenen Kennzahlen ergänzt werden. Durch

Definition von Grenzwerten für die Ausprägung bestimmter Kennzahlen erhält der Prüfer einen guten Überblick, ob das Management seiner Überwachungsfunktion gerecht wird und wie es auf problematische Entwicklungen reagiert hat.

Ausblick

Das Verständnis für die internen Prozesse und Kontrollen geht bei zunehmender Digitalisierung verloren. Durch Einsatz des Cloud Computing wird diese Problematik sogar noch verschärft. Dies geschieht durch die Virtualität des Cloud Computing, wodurch der Prüfer die Zuordnung zwischen den virtuellen und realen Ressourcen und Dienstleistungen nicht mehr direkt beobachten kann. Diese Zuordnung ändert sich beim Einsatz des Cloud Computing ständig.

Durch Process Mining ist es möglich, die Übersicht wieder zu erlangen und die Zuordnung zwischen realer und virtueller Organisation im Zeitverlauf zu rekonstruieren und damit prüfbar zu machen. Es besteht jedoch Bedarf, das Process Mining zu ergänzen und zu verfeinern, insbesondere durch die Integration von Events aus der Cloud. Denn für Unternehmen und Prüfer gelten die Ordnungsmäßigkeits- und Sicherheitsanforderungen unabhängig davon, ob es sich um eine reale oder eine virtuelle Organisation handelt. Wir kooperieren mit einer Reihe internationaler Experten, um die notwendige betriebswirtschaftliche Forschung voranzutreiben und Anwendungsbarrieren aus Sicht der Praxis zu reduzieren.

Summary

To date, Cloud Computing has primarily been subject to research in the area field of Computer Science. We describe the managerial challenges with an emphasis focus on cost accounting, capital budgeting, and auditing when outsourcing internal functions to the cloud. In a cloud environment, activities of the real organization and the virtual organisationorganization are interrelated and not directly observable. Further, the interrelationship between real and virtual components changes quickly. Both, management and auditors are challenged by an increasing complexity and lack in of transparency of in processes outsourced to the cloud. Complexity and non-transparency cause risks in regard to security and compliance requirements. We propose to use Process Mining to handle complexity and non-transparency for to mitigateing the risks. Process Mining facilitates an enhanced understanding of processes through visualization and detailed analysis. Process Mining can be used by auditors to assess the operating effectiveness of internal controls. However, additional event data are necessary in order to apply Process Mining in the general form of Event Mining for auditing cloud computing systems, additional event data are necessary.

Anmerkungen/Literatur

1) Zur Abgrenzung von Cloud Computing gegenüber den verwandten Technologien Grid Computing, Utility Computing, Virtualisierung und Autonomic Computing siehe Zhang, Qi; Cheng, Lu; Boutaba, Raouf (2010): Cloud computing: state-of-the-art and research challenges. In: Journal of Internet Services and Applications 1, S. 7–18.
2) Glaser, H. (1992): Prozesskostenrechnung – Darstellung und Kritik. In: Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung (ZfbF) 44. (3), S. 275–288.
3) Brasch, H.D. (1927): Zur Praxis der Unkostenschwankungen und ihrer Erfassung. In: Betriebswirtschaftliche Rundschau, 4 (4–5), 65–72; Günther, Th. W.; Riehl, A.; Rößler, R. (2014): Cost stickiness: state of the art of research and implications. In: Journal of Managerial Control 24, S. 301–308.
4) Rapp, Y. (1939): Economic Stages of Extension of a Telephone Network. Ericsson Technics, Stockholm 1939, Nr. 5, zitiert nach: Schneider, E. (1973): Wirtschaftlichkeitsrechnung, 8. Aufl., Tübingen u. Zürich 1973, S. 121.
5) Eine Abgrenzung wird vom BSI vorgenommen, vgl. Bundesamt für Sicherheit in der Informationstechnik, Eckpunktepapier (2012): Sicherheitsempfehlungen für Cloud Computing Anbieter, S. 18f.
6) Vgl. BSI, S. 17f.
7) BGH-Urteil 15.11.2006, XII ZR 120/04.
8) Lissen/Brünger/Damhorst (2014): IT-Services in der Cloud und ISAE 3402 – Ein praxisorientierter Leitfacen für eine erfolgreiche Auditierung, Heidelberg, Springer Gabler, S. 15f.
9) BGH-Urteil vom 4.3.2010, III ZR 79/09.
10) Lissen/Brünger/Damhorst, S. 16f.
11) Bundesministerium für Wirtschaft und Energie, Kompetenzzentrum Trusted Cloud (2014): Leitfaden – Haftungsrisiken beim Cloud Computing 10, S. 18f.
12) BSI, S. 69f.
13) Für Einzelheiten zum Datenschutz siehe BSI, S. 73ff.
14) Institut der Wirtschaftsprüfer in Deutschland e.V. (IDW): Stellungnahme zur Rechnungslegung: Grundsätze ordnungsmäßiger Buchführung bei Einsatz von Informationstechnologie (IDW RS FAIT 1), Tz. 23.
15) IDW RS FAIT 1 Tz. 25.
16) Institut der Wirtschaftsprüfer in Deutschland e.V. (IDW): Prüfungsstandard: Feststellung und Beurteilung von Fehlerrisiken und Reaktionen des Abschlussprüfers auf die beurteilten Fehlerrisiken (IDW PS 261 n.F.) Tz. 29.
17) IDW PS 261.30.
18) IDW PS 261.31.
19) IDW PS 261.32.
20) IDW PS 261.33.
21) IDW PS 261.34.

Die Autoren

Ludwig Mochty studierte Technische Mathematik an der TU Wien, wo er auch promovierte, Betriebswirtschaftslehre und Operations Research am Institut für Höhere Studien in Wien und Wirtschaftsinformatik an der Universität Wien. Seit 1994 hat er den Lehrstuhl für Wirtschaftsprüfung, Unternehmensrechnung und Controlling an der Universität Duisburg-Essen inne. Neben seiner akademischen Tätigkeit ist er seit 1982 in verschiedenen Funktionen in der Wirtschaftsprüfunspraxis tätig, wo er unter anderem zeitgemäße computergestützte Prüfungsverfahren entwickelt und anwendungsbezogen testet.

Michael Wiese studierte Wirtschaftswissenschaften an der Universität Duisburg-Essen und promovierte am Lehrstuhl für Wirtschaftsprüfung, Unternehmensrechnung und Controlling. Er ist Wirtschaftsprüfer bei der Ernst & Young GmbH Wirtschaftsprüfungsgesellschaft. Dort leitet er die Forschung und Entwicklung für die Serviceline Assurance in Deutschland, Österreich und der Schweiz. Sein Entwicklungsfokus liegt auf dem Einsatz computergestützter Verfahren in der Prüfungspraxis, zum Beispiel dem Process Mining.

Der vorliegende Artikel illustriert die Bedeutung des Hochleistungsrechnens für die Entwicklung neuartiger energieeffizienter Materialien für elektronische Bauelemente sowie für die Energiegewinnung (Thermoelektrika), Energiespeicherung (Wasserspaltung) und Wärmetransport (magnetische Kühlung).

Neue Materialien errechnen

HPC in der Vorhersage neuartiger elektronischer und magnetischer Eigenschaften auf der Nanoskala mittels atomistischer quantenmechanischer Simulationen

Von Rossitza Pentcheva & Markus Ernst Gruner

Bei der Suche nach neuen Wegen, die Herausforderungen unserer hochtechnisierten Gesellschaft auf ökonomisch vertretbare Weise zu lösen, sind moderne Höchstleistungs-Computersysteme nicht mehr wegzudenken. Dies umfasst die Wertschöpfungskette in ihrer gesamten Länge, angefangen von der Optimierung von Prozessen und Abläufen auf der makroskopischen Skala bis hin zur Entdeckung und Ergründung neuer physikalischer Phänomene auf der Nanoskala, die komplett neue Lösungsansätze für etablierte Technologien bieten. Ein wichtiger Schwerpunkt liegt hier in der Entwicklung neuartiger Materialsysteme (für einen umfassenden Überblick über die gemeinsamen Aktivitäten der Universitätsallianz Ruhr im Bereich der Materialforschung sei hier auf den kürzlich erschienenen Band 48 der Unikate „Materials Chain Ruhr“ verwiesen).

Der vorliegende Artikel soll nun den Beitrag der theoretischen Materialphysik im Bereich Hochleistungsrechnen zu diesem Aufgabenspektrum anhand konkreter Beispiele illustrieren – mit dem Fokus auf Materialien für elektronische Bauelemente sowie für die Energiegewinnung (Thermoelektrika), Energiespeicherung (Wasserspaltung) und Wärmetransport (magnetische Kühlung).

Parameterfreie Berechnung von Materialeigenschaften

Die Grundlagen der modernen Verfahren zur theoretischen Bestimmung der physikalischen und chemischen Eigenschaften von Festkörpern und Molekülen gehen zurück auf den 1926 von Erwin Schrödinger eingeführten mathematischen Formalismus, der die exakte Beschreibung von atomaren und elektronischen Systemen in Form einer quantenmechanischen Wellengleichung erlaubt. Vorausgesetzt werden hierbei im Prinzip lediglich Informationen über beteiligte Atome und der zugehörigen Elektronen (bekannt aus dem Periodensystem der Elemente) und ihrer geometrischen Anordnung im Kristallgitter beziehungsweise im Molekül.

Die fundamentale Bedeutung dieser Entwicklung wurde bereits früh vorausgesehen, zum Beispiel von Paul Dirac, der zusammen mit Schrödinger 1933 den Nobelpreis für Physik für die Weiterentwicklung

Rositza Pentcheva. Foto: Vladimir Unkovic

der Quantenmechanik erhielt. Dirac schrieb dazu 1929:

„Die grundlegenden Physikalischen Gesetze, die für eine mathematische Theorie weiter Teile der Physik und der gesamten Chemie notwendig sind, sind damit vollständig bekannt und die einzige Schwierigkeit ist es daher, dass die exakte Anwendung dieser Gesetze zu Gleichungen führt, die zu kompliziert sind, um lösbar zu sein".

Im Hinblick auf die Materialphysik konnte diese Herausforderung erst mit Hilfe der modernen Computertechnologie ernsthaft angenommen werden. Seit Beginn der 1990er Jahre hat sich die kumulierte Rechenleistung der 500 weltweit größten Supercomputer alle vier Jahre beinahe verzehnfacht (Quelle: www.top500.org). Demgegenüber steht ein immenser und mit der Genauigkeit der Beschreibung ansteigender numerischer Aufwand. Für den Siegeszug in der Festkörperphysik war deshalb eine Vereinfachung entscheidend, die bereits in den 60er Jahren des vergangenen Jahrhunderts von Walter Kohn und seinen Mitstreiter*innen vorangetrieben und konsequenterweise 1998 ebenfalls mit dem Nobelpreis für Chemie geehrt wurde. Im Zuge der Entwicklung der Dichtefunktionaltheorie wies Kohn nach, dass zur exakten Bestimmung der Grundzustandseigenschaften Schrödingers komplexe Wellenfunktion, die von den Positionen aller Elektronen im System abhängt, durch die viel einfachere Elektronendichte ersetzt werden kann und schlug ebenfalls eine sehr erfolgreiche Näherung für die mathematisch schwer zu fassenden Austausch- und Korrelationseffekte zwischen den Elektronen vor, die immer noch in vielen Arbeiten genutzt wird.

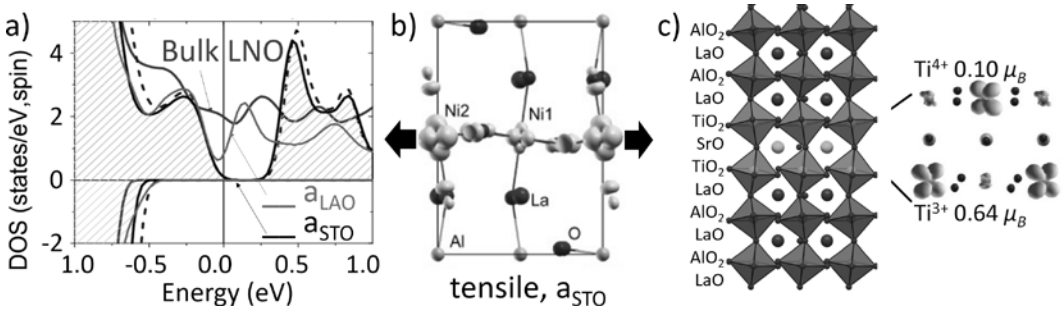
Auch heute sind diese Bemühungen noch lange nicht beendet. Die Notwendigkeit, immer größere Strukturen mit hoher Genauigkeit zu simulieren, erfordert stete Weiterentwicklung, sowohl im Bereich neuer Methoden als auch der Computersysteme. Supercomputer, wie die im letzten Jahr an der UDE installierte „magnitUDE" (Platz 279 der Top 500 Liste der schnellsten Supercomputer im Juni 2016), bieten Wissenschaftler*innen die Möglichkeit, an dieser Entwicklung an vorderster Front zu partizipieren.

Oxidgrenzflächen als neue Materialien

Die fundamentalen Grenzen der Miniaturisierung elektronischer Bauelemente erfordern die Entwicklung neuartiger Materialien mit maßgeschneiderter Funktionalität, die das Potential haben, konventionelle Halbleiter zu ersetzen. Eine vielversprechende Materialklasse sind die Übergangsmetalloxide. Zum einen sind sie weitverbreitet und damit kostengünstig, zum anderen sind sie umweltfreundlich und chemisch stabil. Des weiteren zeigen sie selbst im Volumenmaterial eine Reihe spannender Eigenschaften wie Magnetismus, Ferroelektrizität, Supraleitung. Letztere sind das Ergebnis des komplexen Zusammenspiels von Gitter-, Spin-, Orbital- und Ladungs-Freiheitsgraden[1].

In verschiedenen Anwendungen wie die Katalyse oder in elektronischen Bauelementen spielen Grenzflächen und Oberflächen eine zentrale Rolle. Oder, wie Herbert Kroemer in seiner Nobelpreis-Rede 2000 formulierte: „Die Grenzfläche ist das Bauelement". Hier waren allerdings „nur" herkömmliche Halbleitergrenzflächen gemeint. Die Grenzflächen von Übergangsmetalloxiden erscheinen noch interessanter, da sie neuartige elektronische und magnetische Eigenschaften aufweisen können, die nicht im Volumenmaterial vorhanden sind. Ein prototypisches Beispiel sind die Grenzflächen zwischen den einfachen nichtmagnetischen Bandisolatoren $SrTiO_3$ (STO) und $LaAlO_3$ (LAO), an denen sich ein zweidimensionales Elektronengas ausbildet sowie magnetische und supraleitende Effekte einstellen. Die intensive Erforschung von diesem und anderen Systemen hat sich zu einem florierenden Forschungsgebiet entwickelt. In 2007 hat die Zeitschrift „Science" Oxidgrenzflächen als einen von zehn wissenschaftlichen Durchbrüchen identifiziert. Hierzu tragen zum einen die Entwicklung von Wachstumstechniken wie die Molekularstrahlepitaxie oder gepulste Laserdeposition bei, mit denen es heutzutage möglich ist, Oxidheterostrukturen mit atomarer Präzision aufzuwachsen. Die neue Funktionalität zusammen mit der Fähigkeit, diese Systeme experimentell zu realisieren, eröffnet eine vielversprechende Perspektive im Design funktionaler Materialien. Allerdings erfordern die Anwendungen ein fundamentales Verständnis der Phänomene und Mechanismen. Hierzu leisten Dichtefunktionaltheorie-Rechnungen einen wesentlichen Beitrag. Die korrelierte Wechselwirkung zwischen den Elektronen der *d* und/oder *f*-Schalen und die komplexe Struktur stellen eine Herausforderung für die theoretische Beschreibung dar. Da die Rechenzeit quadratisch bis kubisch mit der Systemgröße ansteigt, ist für Heterostrukturen mit bis zu Hunderten von Atomen die Nutzung von Hochleistungsrechnern erforderlich. Für die Beschreibung korrelierter Elektronen sind Methoden über die Standard-Näherung für das Austausch-Korrelationsfunktional hinaus notwendig. Einige von diesen erhöhen den numerischen Aufwand um mindestens einer Größenordnung. Hinzu kommen gestiegene Anforderungen an Arbeitsspeicher und Laufzeiten, die auch nur mit Hilfe des High-Performance Computing erfüllt werden können.

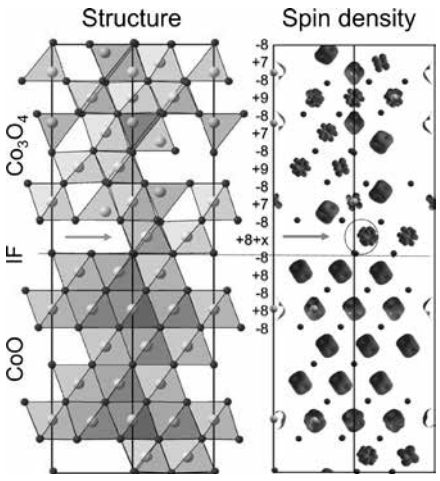
Um ein fundamentales Verständnis für die Phänomene an Oxidgrenzflächen zu erzielen, müssen auch die Kontrollparameter für dieses Verhalten identifiziert und systematisch untersucht werden. Damit befasst sich ein von der Deutschen Forschungsgemeinschaft gefördertes Projekt im Sonderforschungsbereich TR80 „Von elektro-



(1) a) Die Zustandsdichte (DOS) zeigt die Anzahl der Zustände pro Energieintervall an: Während im Volumenmaterial LNO ein Metall ist (nicht-verschwindende Zustandsdichte an der Fermi-Kante, dargestellt als dunkelgraue Linie, die besetzte von unbesetzten Zuständen trennt), öffnet sich bei dem $(LNO)_1/(LAO)_1(001)$ Übergitter unter Zugspannung bei der lateralen Gitterkonstante von STO eine Bandlücke (schwarze Linie).
b) Die Elektronendichteverteilung der unbesetzten Zustände zeigt, dass dabei zwei inäquivalente Ni-Plätze (Ni1 und Ni2) mit größeren/kleineren magnetischen Momenten und Ni-O Bindungslängen entstehen.
c) Heterostruktur aus 4 LAO und 2 STO Schichten mit zwei polaren Grenzflächen von n-Typ.
Der Elektronenüberschuss von 0.5 e/Ti Platz resultiert in einem Ladungs- und Orbital-geordneten Zustand mit Ti^{3+} und Ti^{4+} Plätzen.
Quelle: a) und b) modifiziert aus A. Blanca-Romero and R. Pentcheva, Confinement induced metal-to-insulator transition in strained $LaNiO_3$/$LaAlO_3$ superlattices, Phys. Rev. B 84, 195450 (2011). © American Physical Society, c) aus D. Doennig and R. Pentcheva, Control of orbital reconstruction in $(LaAlO3)_M/(SrTiO3)_N(001)$ quantum wells by strain and confinement, Sci. Rep. 5, 7909 (2015). © Nature Publishing Group

nischen Korrelationen zur Funktionalität". Einige der identifizierten und untersuchten Parameter werden im Folgenden dargelegt. Die Ausbildung von Grenzflächen ist mit einer Brechung von Symmetrien verbunden. Diese kann zu einer weiteren Aufspaltung von Energieniveaus führen und damit zur selektiven Besetzung bestimmter Orbitale (Orbitalpolarisation) und folglich zu elektronischen Übergängen zum Beispiel von einem Metall zu einem Isolator (Metall-Isolator-Übergang: MIT). Diese Phänomene koppeln stark auch an die sogenannte epitaktische Verspannung, ausgeübt durch das Substrat, auf dem die Heterostrukturen aufwachsen (s. Abb. 1b und c). Durch die Variation des Substrats und damit des lateralen Gitterparameters lassen sich strukturelle Verzerrungen wie Verdrehungen und Verkippungen der Oktaeder (s. Abb. 1c) gezielt verändern. Diese wiederum steuern den Orbitalüberlapp und können damit auch zum MIT führen. Ein weiterer wichtiger Parameter ist die Reduktion der Dimensionen (engl: „confinement", „finite size effects", „dimensional crossover"). Dabei wird die Schichtdicke des aktiven Materials auf nur wenige Lagen reduziert. Ein Beispiel für das Zusammenspiel der reduzierten Dimensionen, der Verspannung und der elektronischen Korrelationen sind die Übergitter, die aus je einer Schicht des Bandisolators $LaAlO_3$ und des korrelierten Metalls $LaNiO_3$ aufgebaut sind. Hier zeigen Dichtefunktionaltheorie-Rechnungen mit einem zusätzlichen Hubbard-Term (DFT+*U*), der den hier starken elektronischen Korrelationen Rechnung trägt, dass sich bei Zugspannung eine Bandlücke in der Bandstruktur öffnet und das System isolierend wird, obwohl $LaNiO_3$ im Volumen bei allen Temperaturen metallisch ist (s. Abb. 1a). Als Ursache dafür wurde die Entstehung zweier inäquivalenter Ni-Ionen identifiziert (Ni1 und Ni2 in Abb. 1b). Diese unterscheiden sich stark in ihrem magnetischen Moment, sowie in den Bindungslängen zum umgebenden Sauerstoff[2].

Eine weitere Möglichkeit, neuartiges Verhalten einzustellen, ist die Ausnutzung der Polarität von Grenzflächen. Zum Beispiel wechseln sich in $Sr^{2+}Ti^{4+}O^{2-}_3$ in der (001) Richtung SrO- und TiO_2-Lagen, die formell neutral sind. Im Gegensatz dazu sind die $(LaO)^+$ und $(AlO_2)^-$ Schichten in $La^{3+}Al^{3+}O^{2-}_3$ jeweils positiv und negativ geladen. Damit ergibt sich an den Grenzflächen, die entweder aus einer LaO- und TiO_2 oder SrO- und AlO_2 Abfolge bestehen, eine Elektronen- (*n*-Typ) oder Loch- (*p*-Typ) dotierte Grenzfläche. An der *n*-Typ Grenzfläche zwischen LAO und STO, wie oben bereits erwähnt, führt die elektrostatische Dotierung zu einer Teilbesetzung des ursprünglich leeren *3d*-Bandes von Ti und damit zur Entstehung eines zweidimensionalen Elektronengases. Dabei hängt das Verhalten stark von der Probengeometrie (Heterostruktur oder dünner Film) ab. Wenn man in einer Heterostruktur mit *n*-Typ Grenzflächen die Schichtdicke bis auf zwei Monolagen reduziert, so lokalisiert die überschüssige Ladung auf jedem zweiten Ti-Ion in der Grenzflächenschicht und es kommt zu Ladungsordnung von Ti^{3+} und Ti^{4+} mit einer abwechselnden Besetzung des d_{xz} und d_{yz} Orbitals an den Ti^{3+} Plätzen (s. dazu die Elektronendichteverteilung um die Fermi-Kante in Abb. 1c)[3].



(2) Seitenansicht und Spindichte der Heterostruktur bestehend aus den Antiferromagneten Cobalt-Monoxid (CoO) und Cobalt-Spinell (Co_3O_4). Insbesondere ergibt sich an der Grenzfläche eine Ladungsdiskontinuität, die zur Reduktion eines oktaedrischen Co^{3+} auf der Spinellseite zu Co^{2+} mit einem nichtkompensierten magnetischen Moment führt. Quelle: modifiziert aus Zi-An Li, N. Fontaíña-Troitiño, A. Kovács, S. Liébana-Viñas, M. Spasova, R. E. Dunin-Borkowski, M. Müller, D. Doennig, R. Pentcheva, M. Farle and V. Salgueiriño, Electrostatic doping as a source for robust ferromagnetism at the interface between antiferromagnetic cobalt oxides, Sci. Rep. 5, 7997 (2015), © Nature Publishing Group

Polare Diskontinuitäten treten nicht nur in Perowskiten auf, sondern auch in vielen anderen Heterostrukturen, zum Beispiel führen sie an den Grenzflächen zwischen Hematit (α-O_3) und Ilmenit ($FeTiO_3$)[4] oder CoO (Kochsalzstruktur) und Spinell (Co_3O_4) zur Entstehung von Ferromagnetismus zwischen antiferromagnetischen Isolatoren (s. Abb. 2). Die theoretische Untersuchung des letzteren Systems hat zur Aufklärung des robusten ferromagnetischen Signals oberhalb der Raumtemperatur in CoO/ Co_3O_4 Kern-Hülle-Nanopartikeln in der AG Farle (UDE) beigetragen[5].

Zurzeit verschiebt sich das Interesse von den viel untersuchten (001)-ausgerichteten Perowskit-Heterostrukturen zu Systemen mit (111)-kristallographischer Orientierung wie in Abbildung (3d) gezeigt. Dabei ändert sich zum einen die Abfolge der Schichten in der ABO_3 Struktur von AO-BO_2-AO in der (001)-Richtung zu AO_3-B-AO_3 in der (111)-Richtung (s. Abb. 3d). Das Besondere daran ist, dass je zwei B-Schichten ein nicht-planares Honigwabengitter ausbilden (s. Draufsicht in Abb. 3d). Dieses Strukturmuster zeigt nicht nur strukturelle Analogien zu Graphen und anderen topologischen Isolatoren, sondern verspricht durch die Präsenz von *d*-Elektronen anstelle der sonst üblichen *s*- und *p*-Bindungen zusätzliche Funktionalität. Des Weiteren hat Haldane – einer der Nobelpreisträger für Physik 2016 – bereits 1988 in Modellrechnungen vorhergesagt, dass auf einem Honigwabengitter ein sogenannter Chern-Isolator stabilisiert werden kann. Die Chern-Isolatoren unterscheiden sich von den topologischen Isolatoren darin, dass bei ihnen die Zeitinversionssymmetrie gebrochen ist (also ferromagnetisch sind) und damit kein externes Magnetfeld notwendig ist für ihre Funktion zum Beispiel in zukünftigen Bauelementen. Die Herausforderung liegt darin, geeignete Materialien zu identifizieren, die dieses Verhalten zeigen und zudem Bandlücken aufweisen mit einer Größe, die Raumtemperatur-Anwendungen erlaubt.

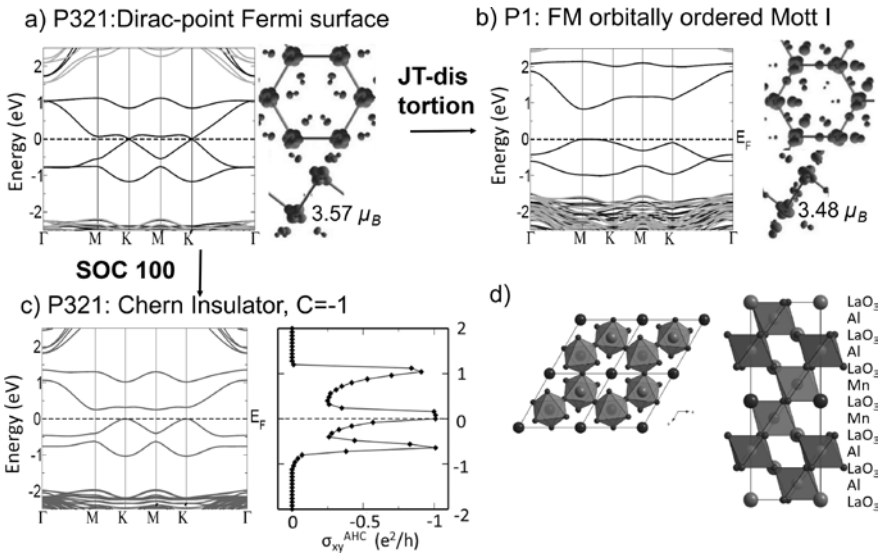
Durch systematische DFT+U-Rechnungen mit einem zusätzlichen Hubbard-Term konnten mehrere Systeme mit unterschiedlicher *d*-Bandbesetzung identifiziert werden, die eine Bandstruktur wie bei Graphen aufweisen mit zwei sich linear kreuzenden Bändern am K-Punkt (s. Abb. 3a). Der Kreuzungspunkt – der sogenannte Dirac-Punkt – liegt dabei an der Fermi-Kante als höchst besetztes Niveau. Das Mangan-basierte Honigwabengitter $(LaMnO_3)_2$ eingebettet in $LaAlO_3$(111) wurde als ein vielversprechender Kandidat identifiziert, da die Spin-Bahn-Kopplung, die typischerweise für *3d*-Materialien klein ist, hier einen großen Effekt zeigt (Abb. 3c). Dabei öffnet sich eine Bandlücke von 150 meV, die ausreichend für Raumtemperaturanwendungen ist. Die weitere Analyse der anomalen Hall-Leitfähigkeit (Abb. 3c) zeigt, dass dieses System ein Chern-Isolator ist, wenn strukturelle Verzerrungen, die es zu einem Mott-Isolator machen (Abb. 3b), unterdrückt werden[6]. Dieser metastabile Zustand kann zum Beispiel durch gezielte Anregungen erreicht werden. Mit den Erkenntnissen aus dieser Untersuchung, insbesondere was die Rolle der *d*-Band-Besetzung betrifft, wurden weitere Kandidaten in der *4d*- und *5d*-Reihe ermittelt[7].

Maßgeschneiderte Thermoelektrika mit Oxidischen Multilagensystemen

Neben der Erforschung dieser und ähnlicher exotischer elektronischer Systeme befasst sich ein weiteres Projekt im SFB/TR80 mit der Optimierung von oxidischen Thermoelektrika. Die Rückgewinnung von Energie ist ein wichtiger Schritt zur effizienteren Nutzung unserer Ressourcen. Eine interessante Möglichkeit bieten hier thermoelektrische Generatoren, die Temperaturunterschiede direkt in elektrische Arbeit umsetzen können. Ihr Wirkungsgrad ist durch die verfügbaren Materialien limitiert, die durch einen dimensionslosen Gütefaktor

$$ZT = \frac{S^2 \sigma T}{\varkappa_E + \varkappa_G}$$

quantifiziert werden. Hierbei ist im Zähler T die Arbeitstemperatur, S der Seebeck-Koeffizient, der den Zusammenhang zwischen Temperaturdifferenz und elektrischer Spannung beschreibt, σ die elektrische Leitfähigkeit. Im Nenner finden wir die Wärmeleitfähigkeit des Materials, sie setzt sich zusammen aus der durch die Elektronen vermittelten Wärmeleitfähigkeit $\varkappa_E$, die eng mit der elektrischen Leitfähigkeit zusammenhängt und der Wärmeleitfähigkeit des Kristallgitters $\varkappa_G$, die auf den Schwingungen der Atome beruht und damit von $\varkappa_E$ und σ weitgehend unabhängig ist. Die besten derzeit bekannten Thermoelektrika wie zum



(3) Bandstruktur und Elektronendichteverteilung von $(LaMnO_3)_2/(LaAlO_3)_4(111)$ a) mit zwei äquivalenten Mn-Untergittern (P321 Symmetrie); b) nach einer strukturellen Verzerrung (Jahn-Teller-Effekt) und c) ausgehend von a) nach Berücksichtigung der Spin-Bahn-Kopplung (SOC). Die Bandstruktur in a) zeigt eine lineare Kreuzung der Bänder am K und K´ Punkt exakt am Fermi-Niveau. Durch die SOC öffnet sich eine Bandlücke und die Berechnung der anomalen Hall-Leitfähigkeit ergibt, dass das System ein Chern-Isolator ist. Andererseits führt die Jahn-Teller Verzerrung in b), die dem Grundzustand entspricht, zu einem trivialen Mott-Isolator aufgrund der Orbitalordnung auf den zwei jetzt inäquivalenten Mn-Untergittern. d) Drauf- und Seitenansicht der Simulationszelle. Die Draufsicht zeigt das nichtplanare Honigwabengitter.
Quelle: modifiziert aus D. Doennig, S. Baidya, W.E. Pickett and R. Pentcheva, Design of Chern and Mott insulators in buckled *3d* oxide honeycomb lattices, Phys. Rev. B 93, 165145 (2016) © American Physical Society

Beispiel Bi_2Te_3 besitzen einen *ZT* im Bereich von eins. Verbesserungen können beispielsweise durch Nanostrukturierung erzielt werden, die den Beitrag des Gitters zur Wärmeleitung erheblich verringern kann[8,9].

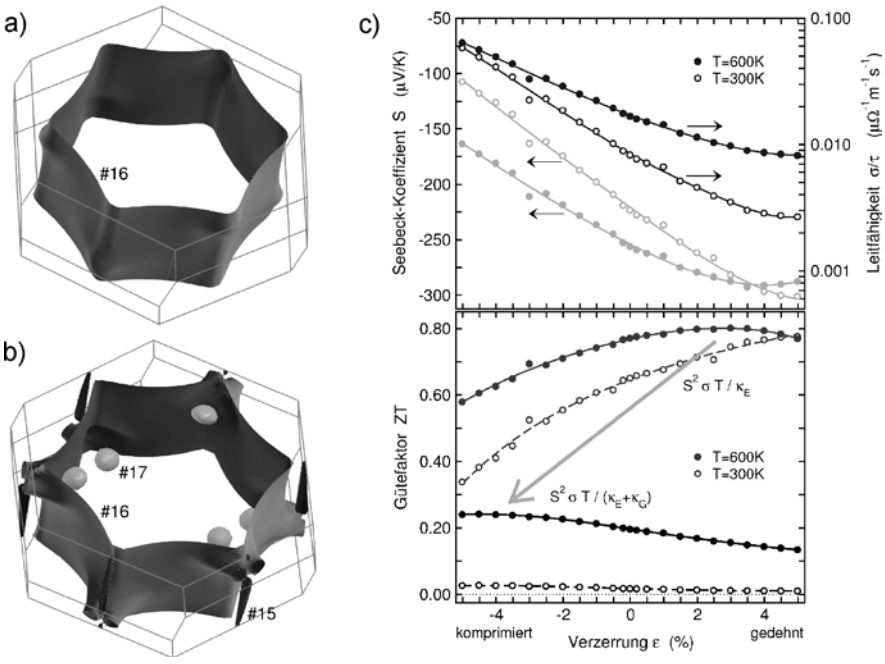
Zwar erreichen Übergangsmetalloxid-Thermoelektrika zurzeit noch nicht die *ZT*-Werte von Halbleiter-basierten Thermoelektrika, sind aber durch ihre thermische Stabilität interessant für Hochtemperatur-Anwendungen. Weiteres Potential für Verbesserung bergen die Orbital-Freiheitsgrade, die Reduktion der Dimensionen oder die Rolle der vom Substrat ausgeübten Verspannung. Diese, insbesondere die gezielte Optimierung der elektronischen Eigenschaften, sind Gegenstand der Untersuchungen in einem weiteren Forschungsprojekt im SFB/TR80 gemeinsam mit Wissenschaftler*innen der Universität Augsburg und des MPI in Stuttgart.

Bis auf $\varkappa_G$ lassen sich alle Größen, die in die Gleichung für *ZT* eingehen, mit Input aus den DFT-Rechnungen im Rahmen der semiempirischen Boltzmann-Transport-Theorie bestimmen. Diese besitzt als einzigen Parameter eine Zeitkonstante τ, die die Relaxation des Systems ins Gleichgewicht beschreibt, während alle anderen relevanten Größen aus Elektronenstrukturrechnungen gewonnen werden können. Relaxationszeit und $\varkappa_G$ erhält man wiederum durch den Vergleich mit dem Experiment oder weiterführenden Berechnungen.

Für ein optimales *ZT* ist es wichtig, dass die unterschiedlichen Beiträge zur Leitfähigkeit gut ausbalanciert sind. Einen besonders großen Seebeck-Koeffizienten S erhält man in der Regel, wenn die elektrische Leitfähigkeit σ sehr klein wird. Dann wird zwar auch $\varkappa_E$ klein, aber $\varkappa_G$ begrenzt weiterhin die Performance. Umgekehrt, wenn σ groß ist, ist auch $\varkappa_E$ groß während *S* klein wird. Bei Halbleitern stellt man den besten Kompromiss für σ meist durch gezielte Verunreinigungen, das heißt das Dotieren mit Fremdatomen, ein. In Materialien mit stark anisotroper, d.h. richtungsabhängiger Leitfähigkeit lässt sich ein ähnlicher Effekt auch durch mechanische Verspannung erreichen.

Am besten lässt sich der Zusammenhang zwischen elektronischer Struktur, epitaktischer Verspannung und thermoelektrischen Eigenschaften an einem einfachen System wie den „natürlichen Multilagensystemen“ $PdCoO_2$ und $PtCoO_2$ verdeutlichen[10]. Diese bestehen abwechselnd aus einer isolierenden Schicht aus CoO_6-Oktaedern und einer metallischen Lage von PdO_2- beziehungsweise PtO_2-„Hanteln“, die sich jeweils ein Sauerstoffatom teilen. Das führt zu sehr guten Leifähigkeitseigenschaften in der Ebene, während senkrecht dazu, in Richtung der Stapelfolge, das Material nahezu isolierenden Charakter hat. Für die Transporteigenschaften sind hier wiederum die höchsten besetzten Zustände an der Fermi-Energie E_F verantwortlich, die sich mit geringem Energieaufwand anregen lassen.

Aufschlussreich ist daher die dreidimensionale, räumliche Auftragung der elektronischen Zustände an der Fermi-Energie, die sogenannte Fermi-Fläche (Abb. 4a). Die Auftragung erfolgt im reziproken Raum, der Fourier-Transformation des Ortsraums. Die Koordinaten entsprechen dann dem Impuls der Elektronen im Kristallgitter. Die Anisotropie in der Leitfähigkeit zeigt sich nun in dem quasi-zweidimensionalen Charakter der Fermi-Fläche, die eine hexagonale Struktur in der Ebene aufweist, sich senkrecht dazu aber fast unverändert fortsetzt. Das bedeutet, dass in senkrechter Richtung die Energie der Elektronen nahezu unabhängig von ihrem Impuls ist, was im Sinne der Newtonschen Mechanik einer unendlichen beziehungsweise sehr großen Masse entspricht, die die Beweglich-



(4) (a) Die Fermi-Fläche von $PtCoO_2$ des unverzerrten Kristallgitters ergibt ein quasi-zweidimensionales Sechseck, das nahezu unverändert in der dritten Dimension fortgesetzt wird. Nur das 16. Band trägt zur Leitfähigkeit bei. (b) Bei einer kompressiven Verspannung des Gitters in der Ebene der CoO_6- und PtO_2-Schichten tragen zwei weitere Elektronenbänder bei und der quasi-zweidimensionale Charakter geht verloren. (c) Thermoelektrische Eigenschaften von $PtCoO_2$ senkrecht zu den Schichtebenen aufgetragen als Funktion der epitaktischen Verzerrung aus unseren Rechnungen für zwei verschiedene charakteristische Temperaturen T=300K und T=600K. Sowohl der Seebeck-Koeffizent S als auch die elektrische Leitfähigkeit lassen sich durch Verspannung des Gitters stark variieren. Im unteren Teil ist die daraus berechnete Gütezahl *ZT* aufgetragen. Vernachlässigt man die Wärmeleitfähigkeit des Gitters erhält man einen sehr guten Wert von 0.8 für eine drei Prozent-Dehnung. Trägt man jedoch auch $\varkappa_G$ Rechnung, verringert sich *ZT* auf maximal 0.25, was dann jedoch bei einer kompressiven Verzerrung von vier Prozent erreicht wird. Entlang der Schichtebenen bleiben *S* und *ZT* hingegen klein.
Quelle: Fermi-Flächen und Simulationsdaten entnommen aus M. E. Gruner, U. Eckern and R. Pentcheva, Impact of strain-induced electronic topological transition on the thermoelectric properties of $PtCoO_2$ and $PdCoO_2$, Phys. Rev. B 92, 235140 (2015). © American Physical Society

keit der Ladungsträger in dieser Richtung stark einschränkt.

Die elektronischen Zustände an der Fermi-Energie gehören bislang alle zu einem einzigen Band, das vornehmlich von Pt-Elektronen besetzt wird. Setzt man das Material einer mechanischen Verspannung aus, zum Beispiel indem man das Oxid epitaktisch auf einem Substrat aufwächst, das eine kleinere Gitterkonstante hat, durchstoßen weitere Elektronenbänder die Fermi-Kante und tragen zur Leitfähigkeit bei. In Abbildung (4b) äußert sich das darin, dass weitere Strukturen hinzukommen und die Fermi-Fläche nun auch in senkrechter Richtung räumlich variiert, was die Leitfähigkeit in dieser Richtung stark verbessert. Solche charakteristischen Änderungen der Fermi-Fläche werden auch als elektronischer topologischer Übergang bezeichnet. Abbildung (4c) zeigt nun für $PtCoO_2$ die Auswirkungen einer epitaktischen Verspannung auf die wichtigsten thermoelektrischen Eigenschaften und *ZT* für elektronischen Transport senkrecht zu den Schichtebenen. Man erkennt, dass auf diese Weise S um dem Faktor drei und σ sogar um eine Größenordnung variiert werden kann. Das hat deutliche Auswirkungen auf den Gütefaktor *ZT*. Vernachlässigt man zunächst die Wärmeleitfähigkeit des Gitters, so würde bei einer Temperatur von 600K man einen sehr guten Wert von 0.8 bei einem gedehnten System erwarten, bei dem die Fermi-Fläche noch glatter und zweidimensionaler ist als im nicht verspannten System. Rechnet man jedoch $\varkappa_G$ mit hinzu, sinkt *ZT* auf einen immer noch passablen Wert von 0.25. Nun wird *ZT* jedoch maximal bei einer kompressiven Verspannung, bei der sich der oben beschriebene elektronische topologische Übergang zeigt. Dies muss bei der experimentellen Realisierung berücksichtigt werden und demonstriert, wie wichtig quantenmechanische Simulationen für das konkrete Design thermoelektrischer Systeme sind.

Während die oben gezeigten anisotropen Transporteigenschaften in einem Volumensystem auftreten, beschäftigen sich unsere gegenwärtigen Untersuchungen mit komplexeren Multilagensystemen, die weitaus mehr Rechenleistung benötigen. Die Kombination verschiedener Oxide in Heterostrukturen kann nicht nur die Leitfähigkeit verbessern sondern auch $\varkappa_G$ minimieren. Ein konkretes Beispiel sind Heterostrukturen, die aus $LaNiO_3$ und $SrTiO_3$ bestehen[11], bei denen die Polarität der Grenzflächen ausgenutzt wird, um kompatible *n*- und *p*-Typ Thermoelektrika zu erzeugen.

In einem weiteren Projekt innerhalb des von der DFG geförderten SPP1613 „Regenerativ erzeugte Brennstoffe durch lichtgetriebene Wasserspaltung: Aufklärung der Elementarprozesse und Umsetzungsperspektiven auf technologische Konzepte“, werden die Eigenschaften oxidischer Anodenmaterialien für die Wasserspaltung optimiert. Zum Beispiel wurde kürzlich eine verbesserte Leistung durch Variation der Zusammensetzung von Kobalt/Nickel Ferrit erzielt.[12]

Materialien für die Kühlsysteme von morgen

In unserer Gesellschaft macht die Erzeugung von Kälte einen signifi-

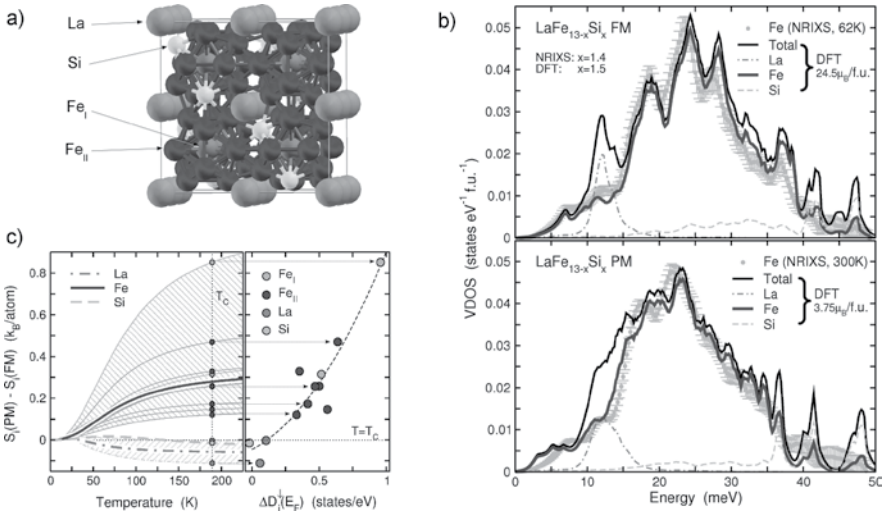
kanten Anteil am industriellen und privaten Energiebedarf aus. Heutige Kühlschränke und Klimaanlagen arbeiten meist noch nach dem Gas-Kompressor-Prinzip. Insbesondere energieeffiziente Anlagen werden dabei oft noch mit klimaschädlichen Gasen betrieben, wie zum Beispiel den Fluorkohlenwasserstoffen FKW. Alternative Kühlmittel sind oft entweder weniger effizient, giftig (Ammoniak), brennbar (HFO-1234yf) oder können nur bei hohen Drücken eingesetzt werden (Kohlendioxid) – was den Einsatz und die Wartung der Geräte einschränkt. Eine vielversprechende Alternative bieten Festkörper-basierte Kühlsysteme, die magnetische, ferroelektrische oder strukturelle Phasenübergänge für den Wärmetransport nutzen.

Setzt man ein Gas nahe der Kondensationstemperatur einem hydrostatischen Druck aus, verflüssigt es sich und gibt die Umwandlungswärme an die Umgebung ab. Nimmt man den Druck adiabatisch weg, das heißt während kein Wärmeaustausch mit der Umgebung stattfinden kann, kühlt das Medium ab und kann dann beim Verdampfen Wärme aus dem Inneren des Kühlgerätes aufnehmen. Ein magnetisches Kühlsystem funktioniert analog. An Stelle des Kompressors tritt ein hinreichend starkes, variables Magnetfeld das beispielsweise durch rotierende Permanentmagnete erzeugt werden kann. Dieses ordnet ein ferromagnetisches Material nahe der Curie-Temperatur, das sich dabei erwärmt. Nachdem die Wärme an die Umgebung abgegeben worden ist, nimmt man das Feld adiabatisch weg und das Material kühlt wieder ab, während es in den magnetisch ungeordneten (paramagnetischen) Zustand zurückkehrt, und kann somit überschüssige Wärme von der kalten Seite aufnehmen.

Das Prinzip der magnetischen Kühlung ist schon sehr lange bekannt und wurde bereits 1933 von William Giauque genutzt, um für seine im Jahr 1949 mit dem Nobelpreis für Chemie geehrte Arbeit Temperaturen nahe des absoluten Nullpunkts zu erreichen. Dennoch rückte seine Anwendung bei Raumtemperatur in Kühlschränken und Klimaanlagen erst in den 1990er Jahren in den Fokus der Forschung und kommerzielle Prototypen wurden erst in den letzten Jahren vorgestellt. Im Gegensatz zu herkömmlichen Kühlgeräten, die auf eine über 100-jährige Entwicklungsgeschichte zurückblicken dürfen, müssen alternative magnetokalorische Kühlsysteme aus dem Stand kommerziell erfolgreich sein. Dies erfordert einen hohen Entwicklungsstand von Beginn an, insbesondere auch was die verwendeten Materialien betrifft. Für die Massenanwendung müssen die Materialien umweltfreundlich und ungiftig sein, effizient arbeiten und dürfen natürlich auch keine exorbitant teuren oder potentiell knappen Rohstoffe enthalten. Sie sollten eine hohe adiabatische Temperaturänderung in moderaten Magnetfeldern von 1 bis 2 Tesla zeigen, die von herkömmlichen Permanentmagneten erzeugt werden können, in Verbindung mit einer großen isothermen Entropieänderung am Phasenübergang. Für eine maximale Performance sollte der Phasenübergang erster Ordnung, das heißt diskontinuierlich, sein und darf zudem nur eine minimale Hysterese aufweisen, da diese mit Energiedissipation, das heißt Verlusten verbunden ist. Gleichwohl stehen bereits heute eine Reihe vielversprechender Kandidaten für diese Aufgabe bereit. Die nach obigen Gesichtspunkten interessantesten Materialklassen sind $La(Fe_{1-x}Si_x)_{13}$, $MnFe(P_{1-x}Si_x)$, and Ni-Mn-basierte Heusler-Legierungen[13].

Im Rahmen des von der Deutschen Forschungsgemeinschaft geförderten Schwerpunktprogramms SPP1599 „Kalorische Effekte in ferroischen Materialien: Neue Konzepte der Kühlung“, nehmen an der Universität Duisburg-Essen theoretische und experimentelle Gruppen an der Suche nach verbesserten Materialien teil. Ziel ist es, zunächst ein grundlegendes Verständnis über die physikalischen Mechanismen zu gewinnen, die zu einer guten Performance oder auch Hysterese führen, um diese dann gezielt zu beeinflussen. Mit Hilfe dieser Information können neue Strategien entwickelt werden, wie die magnetokalorischen Eigenschaften der Materialsysteme gezielt verbessert werden können. Während zum Beispiel der Gesamtbetrag der isothermen Entropieänderung experimentell durch Kalorimetrie gut zu bestimmen ist, bereitet es im Experiment deutlich größere Probleme, die entsprechenden Teilbeiträge der verschiedenen mikroskopischen Freiheitsgrade, repräsentiert durch magnetische Ordnung, Gitterdynamik und elektronische Struktur, zu identifizieren und damit deren Relevanz für die magnetokalorischen Eigenschaften des Materials aufzuklären. Abhilfe können hier moderne, parameterfreie Elektronenstrukturrechnungen im Rahmen der Dichtefunktionaltheorie (DFT) schaffen. Diese sind in der Lage, ohne empirische Parameter das Zusammenspiel von elektronischen, magnetischen und Gitterfreiheitsgraden auf der elektronischen Skala zu quantifizieren. Durch die Kombination von großskaligen DFT-Rechnungen und inelastischer Röntgenstreuung an resonanten Atomkernen (nuclear resonant inelastic X-ray scattering, NRIXS), die am Argonne National Laboratorium durchgeführt wurden, konnten wir so zusammen mit den experimentellen Arbeitsgruppen von Heiko Wende aus Duisburg-Essen, von Oliver Gutfleisch aus Darmstadt und Beatriz Roldán Cuenya aus Bochum für das technologisch wichtige System $La(Fe_{11.5}Si_{1.5})_{13}$ demonstrieren, dass die Gitterschwingungen einen wesentlichen Anteil zur isothermen Entropieänderung beitragen[14].

Die Schwierigkeit für die theoretische Beschreibung liegt vor allem in der Größe des zu berechnenden Systems. Die kubische Einheitszelle von $La(Fe_{1-x}Si_x)_{13}$ beherbergt 112 Atome (Abb. 5a). Für einen Si-Anteil



(5) (a) Einheitszelle von $La(Fe_{11.5}Si_{1.5})_{13}$ mit insgesamt 112 Atomen (8xLa, 12xSi, 92xFe). (b) Vergleich der mit NRIXS gemessenen Phononenszustandsdichten (VDOS) für ^{57}Fe als Funktion der Phononen-Energie (helle Kreise mit Fehlerbalken) mit den mittels DFT berechneten Daten (Linien), sowohl in der ferromagnetischen Phase (oben) als auch in der paramagnetischen (unten). Die dickeren Linien beschreiben die partielle Fe-VDOS, die sehr gut mit dem Experiment übereinstimmt. Die dünnere schwarze Linie markiert die gesamt-VDOS, während die punktierten und gestrichelten Linien den Anteil der anderen beiden Elemente wiedergeben. (c) Linke Seite: Aus der VDOS berechnete Entropiedifferenz zwischen der ferromagnetischen (FM) und der paramagnetischen (PM) Phasen. Die dicken Linien beschreiben den Mittelwert für die jeweilige Atomsorte. Die dünnen Linien die Entropiedifferenz ΔS_i für jeden Gitterplatz i. Rechte Seite: Die lokalen ΔS_i (jeweils bestimmt an der Curie-Temperatur T_C) aller Gitterplätze korrelieren direkt mit der Änderung der lokalen elektronischen Zustandsdichte an der Fermi-Energie ΔD_i, was eine enge Kopplung zwischen magnetischer Ordnung, elektronischer Struktur und Gitterentropie beweist.
Quelle: Grafiken entnommen aus M. E. Gruner, W. Keune, B. Roldan Cuenya, C. Weis, J. Landers, S. I. Makarov, D. Klar, M. Y. Hu, E. E. Alp, J. Zhao, M. Krautz, O. Gutfleisch und H. Wende, Element-Resolved Thermodynamics of Magnetocaloric $LaFe_{13-x}Si_x$, Phys. Rev. Lett. 114, 057202 (2015). (c) American Physical Society

von x=1.5 können wir zwar auf eine vereinfachte primitive Zelle mit 28 Atomen zurückgreifen, die Bestimmung der Vibrationsfreiheitgrade erfordert jedoch deren Vervielfachung auf 224 Atome. Damit müssen insgesamt 1744 Valenz-Elektronen in unsere Rechnung mit einbezogen werden, was nur auf massiv-parallelen Supercomputersystemen wie der Cray XT/6m oder magnitUDE an der Universität Duisburg-Essen möglich ist.

Um die Schwingungsmoden des Kristallgitters zu bestimmen, werden einzelne Atome aus ihren Gleichgewichtspositionen ausgelenkt und mittels DFT die rücktreibenden Kräfte berechnet, mit denen die deformierten Elektronenwolken die Ionenrümpfe wieder in ihre ursprünglichen Positionen zurückziehen. In einer einfachen Anschauung kann man sich das Kristallgitter als periodisches System von Federn denken, die zwischen den Atomen in der Gleichgewichtslage gespannt sind (Born-von-Kármán-Modell). Die entsprechenden Federkonstanten ergeben sich aus den oben beschriebenen rücktreibenden Kräften, deren Kenntnis die Berechnung der Schwingungsmoden und Eigenfrequenzen in dem jeweiligen Kristallgitter erlaubt. Insgesamt waren für die vollständige Bestimmung der Kraftkonstantenmatrix beider Phasen 112 solcher Rechnungen nötig. Relevant für die thermodynamischen Eigenschaften wie die spezifische Wärmekapazität oder die Gitterentropie ist die Verteilung (Anzahl) der Schwingungsmoden, die mit einer vorgegebenen Energie angeregt werden können, die sogenannte Phonon-Zustandsdichte (vibrational density of states, VDOS). Diese kann experimentell mit NRIXS bestimmt werden. Dabei werden mit harter Röntgenstrahlung die Atomkerne selbst angeregt – in diesem Fall das Isotop ^{57}Fe – und die Energie der gleichzeitig angeregten Gitterschwingungen gemessen werden. Im Gegensatz zum theoretischen Ansatz erhält man damit jedoch nur die Beiträge von ^{57}Fe. Die Abbildungen (2b) und (2c) zeigen, dass die theoretischen Vorhersagen für die Beiträge von Fe nahezu perfekt mit den gemessenen partiellen Phononen-Zustandsdichten übereinstimmen, sowohl in der ferro- als auch in der antiferromagnetischen Phase. Das erlaubt uns, die Rechnungen zu nutzen, um die Entropiedifferenz auch für die anderen Elemente oder sogar gitterplatzabhängig zu bestimmen (Abb. 5c).

Der Vergleich der beiden Teilgrafiken in Abbildung (5b) zeigt aber auch, dass sich die Form der Kurven abhängig vom magnetischen Zustand signifikant unterscheiden, die Verteilung der Gitterschwingungen also stark von der magnetischen Ordnung abhängt. Auch dies lässt sich mit Hilfe der elektronischen Struktur aufklären. Die Zahl der Elektronen im höchsten besetzten Zustand, der Fermi-Energie E_F spielen dabei eine besondere Rolle, da diese mit dem geringsten Aufwand bei einer Verformung des Kristallgitters umverteilt werden können. Die magnetische Unordnung resultiert in einer starken Erhöhung der Zahl der elektronischen Zustände an der Fermi-Energie E_F. Dadurch wird das Gitter weicher, was wiederum die Gitterentropie vergrößert. Dieser Zusammenhang gilt lokal für jedes einzelne Atom, was durch die starke Korrelation an jedem Gitterplatz i zwischen der Änderung der elektronischen Zustandsdichte $\Delta D_i(E_F)$ und der Entropieänderung ΔS_i in Abbildung (5c) aufgezeigt wird. Letztlich konnten wir also zeigen, dass die magnetischen, elektronischen und

Gitterfreiheitsgrade in diesem System eng miteinander verwoben sind und gemeinsam zu den guten magnetokalorischen Eigenschaften dieses Materials beitragen. Das genaue Verständnis dieser Mechanismen ermöglicht uns nun die gezielte Veränderung des Materials, beispielsweise durch Hinzulegieren anderer Elemente, um unerwünschte Eigenschaften wie die Umwandlungshysterese zu minimieren, ohne die positiven Eigenschaften nachhaltig einzuschränken.

Summary

State-of-the-art supercomputers have meanwhile become an indispensable tool in the quest for new materials addressing the technological challenges of our society. One important strategy is a knowledge-based search relying on the understanding of materials properties on the electronic scale with parameter-free, material-specific calculations in the framework of density functional theory. This method yields an accurate, quantum-mechanical access to materials properties, which is independent from experiment but in turn also very computationally expensive, in particular, when the size of investigated systems becomes large. Further challenges concern the treatment of electron-electron interactions. Within this contribution, we illustrate how high performance computing contributes to this task using three paradigmatic examples related to interface-induced electronic and magnetic phases, materials for energy harvesting and magnetic refrigeration. Transition metal oxide heterostructures not only allow us to combine functional properties in one composite material, but, most importantly, novel electronic states can be stabilized at their interfaces that are not available in the bulk parent compounds. In this contribution we show how the functionality at the interface can be tailored exploring systematically a set of parameters such as polar discontinuities that can give rise to 2D conductivity and/or magnetic phases between insulators. Moreover, strain and confinement can tune the electronic behavior between a metal and an insulator. Last but not least, buckled honeycomb patterns in oxide superlattices can lead to exotic topologically nontrivial states that bear analogies to graphene with additional functionality due to the correlated nature of *d* and *f*-electrons. Furthermore, we demonstrate how strain can be used to enhance the thermoelectric performance of highly anisotropic cobalt delafossites. In future, one may thus expect to exploit spin- and orbital degrees of freedom in the design of electronic devices or the development of highly efficient thermoelectric generators. The third example aims at the intimate coupling between magnetic, lattice and electronic degrees of freedom in modern magnetocaloric materials. Their understanding is of primary importance for the development of energy-saving refrigeration and air-conditioning concepts.

Danksagung

Die Autoren bedanken sich bei W.E.Pickett, U. Eckern, P. Entel, A. Blanca-Romero, D. Doennig, B. Geisler, H. Hajiyani, O. Köksal und zahlreichen weiteren Kollaborationspartnern im In- und Ausland für die gute Zusammenarbeit, wichtige Anregungen und Diskussionen. Bei M. Farle, W. Keune, H. Wende, O. Gutfleisch und den zahlreichen Mitgliedern ihrer Arbeitsgruppen bedanken wir uns insbesondere für die enge experimentelle Zusammenarbeit. Der Deutschen Forschungsgemeinschaft gebührt unser Dank für die finanzielle Unterstützung unserer Projekte im Rahmen des SFB/TR80, des SPP1599 und SPP1613.
Ebenso wären die oben vorgestellten Ergebnisse nicht möglich gewesen ohne die Bereitstellung von HPC Ressourcen und den Support durch das Leibniz Rechenzentrum in München, das NIC am Forschungszentrum Jülich und das Center for Computational Sciences and Simulation (CCSS) der Universität Duisburg-Essen.

Anmerkungen und Literatur

1) Y. Tokura, N. Nagaosa, Orbital Physics in Transition Metal Oxides, Science 288, 463 (2000)
2) A. Blanca-Romero and R. Pentcheva, Confinement induced metal-to-insulator transition in strained $LaNiO_3/LaAlO_3$ superlattices, Phys. Rev. B 84, 195450 (2011).
3) D. Doennig and R. Pentcheva, Control of orbital reconstruction in $(LaAlO_3)_M/(SrTiO_3)_N(001)$ quantum wells by strain and confinement, Sci. Rep. 5, 7909 (2015).
4) R. Pentcheva and H.Sadat Nabi, Interface magnetism in $Fe_2O_3/FeTiO_3$ Heterostructures, Phys. Rev. B 77, 172445 (2008).
5) Zi-An Li, N. Fontaíña-Troitiño, A. Kovács, S. Liébana-Viñas, M. Spasova,R. E. Dunin-Borkowski, M. Müller, D. Doennig, R. Pentcheva, M. Farle and V. Salgueiriño, Electrostatic doping as a source for robust ferromagnetism at the interface between antiferromagnetic cobalt oxides, Sci. Rep. 5, 7997 (2015).
6) D. Doennig, S. Baidya, W.E. Pickett and R. Pentcheva, Design of Chern and Mott insulators in buckled 3d oxide honeycomb lattices, Phys. Rev. B 93, 165145 (2016).
7) H. Guo, S. Gangopadhyay, O. Köksal, R. Pentcheva, and W. E. Pickett, Wide gap Chern Mott insulating phases achieved by design, npj Quantum Materials 2, 4 (2017).
8) für einen weiterführenden Überblick siehe auch den Beitrag von Gabi Schierning, Roland Schmechel und Dietrich Wolf im Unikate-Heft Nr. 43
9) S. Hébert und A. Maignan, Thermoelectric Oxides, in: Functional Oxides, D. W. Bruce, D. O'Hare und R. I. Walton (Ed.), Wiley (2010), S. 203
10) M. E. Gruner, U. Eckern and R. Pentcheva, Impact of strain-induced electronic topological transition on the thermoelectric properties of $PtCoO_2$ and $PdCoO_2$, Phys. Rev. B 92, 235140 (2015)
11) B. Geisler, A. Blanca-Romero and R. Pentcheva, Design of n- and p-type oxide thermoelectrics in $LaNiO_3/SrTiO_3(001)$ superlattices exploiting interface polarity Phys. Rev. B. 95, 125301 (2017).
12) K. Chakrapani, G. Bendt, H. Hajiyani, I. Schwarzrock, T. Lunkenbein, S. Salamon, J. Landers, H. Wende, R. Schlögl, R. Pentcheva, M. Behrens, S. Schulz, Role of composition and size of cobalt ferrite nanocrystals in the oxygen evolution reaction, ChemCatChem (2017), DOI:101002/ctcc.201700376
13) für einen Überblick s. Liu, T. Gottschall, K. P. Skokov, J. D. Moore und O. Gutfleisch, Giant magnetocaloric effect driven by structural transitions, Nature Materials 11, 620 (2012)
14) M. E. Gruner, W. Keune, B. Roldan Cuenya, C. Weis, J. Landers, S. I. Makarov, D. Klar, M. Y. Hu, E. E. Alp, J. Zhao, M. Krautz, O. Gutfleisch und H. Wende, Element-Resolved Thermodynamics of Magnetocaloric $LaFe_{13-x}Si_x$, Phys. Rev. Lett. 114, 057202 (2015)

*Die Autor*innen*

Markus Gruner studierte von 1989 bis 1995 Physik an der damaligen Universität-Gesamthochschule Duisburg. Seine Diplomarbeit schrieb er zu dem Thema „Simulation von statistischen Modellen zur Beschreibung von Magnetovolumen-Instabilitäten". Er promovierte 2003 an der Universität Duisburg-Essen in theoretischer Physik über das Thema „Monte-Carlo-Simulationen von Magnetovolumeneffekten in Festkörpern und Nanopartikeln". Seine Habilitation in theoretischer Physik folgte 2013 ebenfalls an der Universität Duisburg-Essen mit dem Titel „Understanding the functional properties of nanostructured magnetic materials from first-principles."Markus Gruner war von 1996 bis 2013 wissenschaftlicher Mitarbeiter in der Arbeitsgruppe von Prof. Peter Entel, darunter von 2008 bis 2012 als Projektleiter in einem von ihm selbst eingeworbenen Forschungsprojekt. Er stellte in dieser Zeit ebenfalls mehrere erfolgreiche Anträge auf Rechenzeit an den Großrechnern JUBL, JUROPA und JUQUEEN am Forschungszentrum Jülich. Zwischen Ende 2013 und Anfang 2016 war Markus zunächst am IFW Dresden (Arbeitsgruppe von PD Dr. Sebastian Fähler) und später am Forschungsreaktor Heinz-Meier-Leibnitz FRM-II der TU München beschäftigt und arbeitete im Projekt von Prof. Rossitza Pentcheva an den thermoelektrischen Eigenschaften von oxidischen Heterostrukturen. Seit dem Frühjahr 2016 leitet Markus Gruner wieder ein eigenes Projekt an der Universität Duisburg-Essen in der AG Pentcheva über die Simulation magnetokalorischer Materialien aus ersten Prinzipien.
Markus Gruner erhielt den Preis der Universität-Gesamthochschule Duisburg für seine Diplomarbeit und promovierte mit Auszeichnung. Er erhielt dreimal den CENIDE-Best Paper Award. Er ist Mitglied in CENIDE und im Vorstand des CCSS.

Rossitza Pentcheva studierte Physik an der Universität Sofia und Köln und schloss ihr Studium mit einer Diplomarbeit am Forschungszentrum Jülich über den Zusammenhang zwischen Struktur und Eigenschaften von ultradünnen Metallschichten. Ihr Abschluss wurde mit einem Preis für ausländische Studierende der DAAD ausgezeichnet. Anschließend promovierte sie am Fritz-Haber-Institut der MPG und der Freien Universität Berlin über die elektronischen und magnetischen Eigenschaften und das Wachstum metallischer Heterostrukturen und Filme. Nach der Promotion in 2000 wechselte sie an die Ludwig Maximilians-Universität, wo sie zunächst ein DFG-Projekt „eigene Stelle" leitete und später als wissenschaftliche Mitarbeiterin, akademische Rätin auf Zeit und akademische Oberrätin auf Zeit eine eigene Gruppe mit selbst eingeworbenen Drittmitteln aufbaute. Nach einem Auslandsaufenthalt an der University of California, Davis in 2005 habilitierte sie im Jahr 2008 im Fach Materialwissenschaften über die theoretische Beschreibung oxidischer Grenzflächen. Seit 2014 forscht sie als Professorin für Theoretische Physik an der Universität Duisburg-Essen. Sie ist Vorstandsmitglied im Center for Computational Science and Simulation (CCSS), Mitglied im Center for Nanointegration (CENIDE) und im wissenschaftlichen Beirat der Universitätsallianz Ruhr. Sie leitete und leitet zahlreiche Projekte finanziert durch die Deutsche Forschungsgemeinschaft sowie European Science Foundation, unter anderem in den Schwerpunktprogrammen 1236 und 1613. Seit 2010 ist sie Gründungsmitglied, Projektleiterin und Mitglied des Vorstands des SFB/TR80 „Von elektronischen Korrelationen zu Funktionalität" sowie seit 2016 Gründungsmitglied und Projektleiterin im SFB 1242 „Nichtgleichgewichtsphänomene in kondensierter Materie in der Zeitdomäne". Für die Realisierung dieser Projekte ist die erfolgreiche Beantragung und Nutzung von Ressourcen unter anderem an den Höchstleistungsrechnern am Leibniz Rechenzentrum sowie eine Erweiterung des neuinstallierten Rechners magnitUDE an der UDE maßgeblich. Im Fokus ihrer Forschung steht die theoretische Modellierung von elektronischen und magnetischen Phänomenen an Oxidgrenzflächen mit Anwendungspotential in der Elektronik, Spintronik und der Energiekonversion (Thermoelektrika, Wasserspaltung).

Markus Ernst Gruner. Foto: Vladimir Unkovic

Jörg Schröder. Foto: Vladimir Unkovic

In den letzten Jahrzehnten wurden neue, moderne Stahlsorten entwickelt. Diese Stähle haben gegenüber den bisher dagewesenen verbesserte Eigenschaften und stärken die Attraktivität von Stahl gegenüber Alternativen wie Aluminium und Kunststoff.

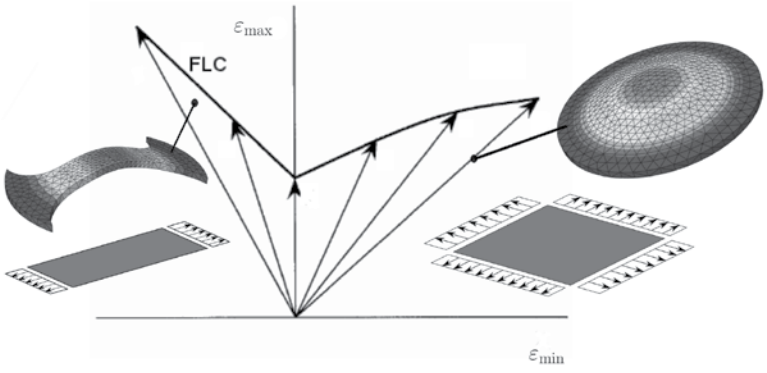
Hochleistungsrechnen trifft Hochleistungsstahl:

Herausforderungen zweier Welten

Von Jörg Schröder,
Axel Klawonn, Lisa Scheunemann
& Martin Lanser

Simulation moderner Hochleistungsstähle im virtuellen Labor

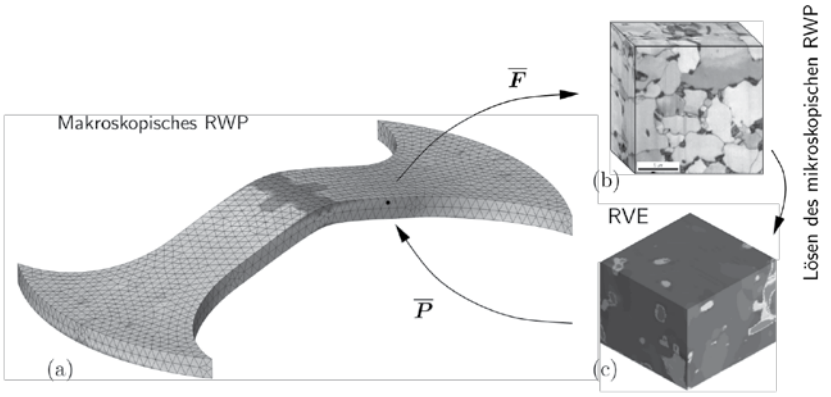
Um den heutigen Anforderungen im Automobilbau nach besserem Schutz der Insassen sowie leichteren Konstruktionen zur Emissionsverringerung zu genügen, wurden in den letzten Jahrzehnten neue, moderne Stahlsorten entwickelt. Diese Stähle haben gegenüber den bisher dagewesenen verbesserte Eigenschaften und stärken die Attraktivität von Stahl gegenüber Alternativen wie Aluminium und Kunststoff. Durch Zugabe unterschiedlicher Legierungselemente wie Mangan, Nickel oder Chrom lassen sich die Eigenschaften des Endprodukts je nach Anwen-



(1) FLC mit Simulationen zweier Beispielgeometrien des Nakajima-Tests, die zur Bestimmung der Kurve herangezogen werden. Rechts: Approximation des biaxialen Zugversuchs, links: Approximation des einachsigen Zugversuchs in Probenmitte.
Quelle: Scheunemann und Schröder 2017

dung anpassen. Dual-Phasen (DP) Stähle, deren Struktur aus martensitischen Einschlüssen in einer ferritischen Matrix besteht, haben diese gewünschten Eigenschaften. Sie kombinieren eine niedrige Streckgrenze mit einer hohen Zugfestigkeit. Als Streckgrenze bezeichnet man die Spannung (Kraft pro Fläche) im Material, die überwunden werden muss um bleibende Verformungen zu erhalten, während die Zugfestigkeit die maximal ertragbare Spannung angibt, bevor das Bauteil versagt.

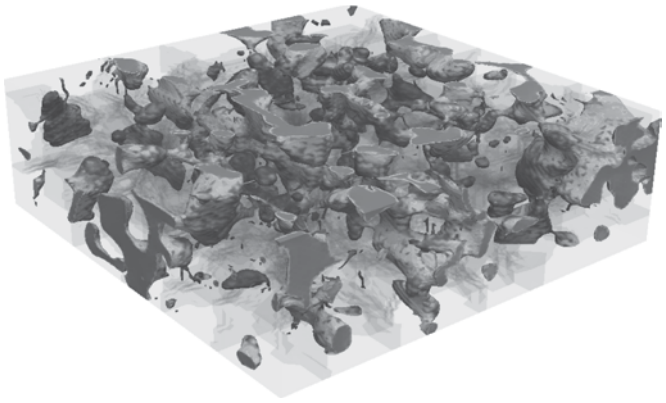
DP-Stähle finden im Automobilbau seit Jahren Verwendung. Dabei spielt das Verformungsverhalten des Materials eine große Rolle, nicht nur für die spätere Sicherheit im Fahrzeug, sondern schon bei der Herstellung der Bauteile. Der Stahl eignet sich besonders für Bauteile, die in Tiefzieh-Prozessen aus Blechen hergestellt werden, bei denen an bestimmten Stellen sehr starke Dehnungen auftreten. Dieser Herstellungsprozess muss so ausgelegt sein, dass das Stahlblech nach dem Umformen keine Risse aufweist. Das in der DIN EN ISO 12004 geregelte Grenzformänderungsdiagramm („Forming Limit Curve“ (FLC), siehe Abb. 1) hilft Konstrukteur*innen, zwischen dem sicheren Bereich kleinerer Formänderung und den Bereichen größerer Formänderung, bei denen das Bauteil versagt, zu unterscheiden. Diese sicheren und unsicheren Bereiche der Verformung werden durch die FLC beschrieben. Der Bereich unterhalb der Kurve ist sicher, oberhalb der Kurve ist ein Versagen des Bauteils zu erwarten. Um ein solches Diagramm für ein Material zu erstellen, sind viele Versuche im Rahmen des Nakajima-Tests notwendig. Jeder dieser Versuche, in denen verschiedene Probegeometrien benutzt werden, beschreibt genau einen Punkt der FLC. Im Versuch werden die Probekörper in einer Prüfmaschine mit einem Stempel verformt, bis es zum Materialversagen kommt. Anhand eines vorher auf die Probe



(2) Erläuterung des Konzepts der FE²-Methode anhand des Nakajima Tests, (a) makroskopisches RWP mit Spannungsverteilung auf der Makroskala und (b) mikroskopisches RWP. Das RVE besteht hier aus martensitischen Einschlüssen (schwarz) und einem ferritischen Polykristall (versch. Graustufen), (c) Approximation der Spannungsverteilung nach Lösen des mikroskopischen RWPs.
Quelle: Scheunemann und Schröder 2017

gezeichneten Rasters oder mit Hilfe von Bildaufnahmen während des Tests, kann nun an der maximal beanspruchten Stelle des Blechs die Dehnung abgelesen und ein Wert im Grenzformänderungsdiagramm bestimmt werden. Simulationsergebnisse basierend auf zwei Probegeometrien des Nakajima-Tests sind in Abbildung (1) dargestellt, welche mit Hilfe der Finite-Elemente-Methode berechnet wurden.

Die Simulation solch aufwendiger Versuche mit Hilfe computergestützter Methoden bietet die Möglichkeit, neue Stahlzusammensetzungen vorab im virtual lab, also in einem virtuellen Labor zu testen und den Einfluss veränderter Parameter auf die Umformbarkeit und das Materialverhalten zu untersuchen. Dazu hat sich im Rahmen der Finite-Elemente-Methode die FE²-Methode, ein Multiskalenansatz, als zielführend erwiesen [1]. Schematisch ist das Prinzip der FE²-Methode in Abbildung (2) dargestellt: Der makroskopische Deformationsgradient $\overline{\boldsymbol{F}}$, der die lokale Verformung an einem makroskopischen materiellen Punkt beschreibt, definiert die Randbedingungen des mikroskopischen Randwertproblems (RWP) und übermittelt so die makroskopische Verformung an die zugrundeliegende Mikrostruktur. Die Lösung des RWP beschreibt die Deformationen und die Spannungen auf der Mikroskala, die ihrerseits den Zustand der Spannungen und Verformungen auf der Makroskala beeinflussen. Dieser Lösungsprozess erfolgt iterativ, und liefert den tatsächlichen Spannungs- und Verformungszustand auf beiden Skalen. Statt eines phänomenologischen Materialgesetzes, das versucht, die mikroskopischen Effekte mit Hilfe eines makroskopischen Materialmodells zu beschreiben, wird eine Mikrostruktur in der Simulation direkt berücksichtigt. An jedem Berechnungspunkt des globalen RWPs wird so ein Ausschnitt der Mikrostruktur des Materials, ein sogenanntes repräsentatives Volumenelement (RVE), verwendet und die daraus entstehenden Effekte berechnet. Da das Zusammenspiel der einzelnen Komponenten (Ferrit und Martensit) auf der Mikroskale großen Einfluss auf das makroskopische Materialverhalten hat und oftmals durch rein makroskopische phänomenologische Gesetze nur unzureichend beschrieben werden kann, birgt die Methode die Möglichkeit der genauen Erfassung der tatsächlich vorkommenden mikroskopischen Effekte. Um mit der FE²-Methode Simulationen in einer realistischen Größenordnung auf



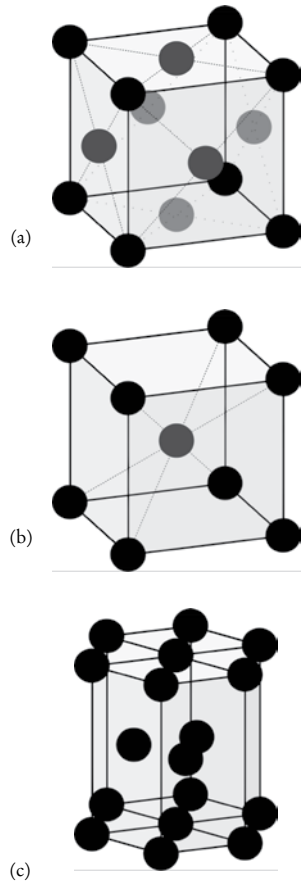
(3) 3D-Mikrostruktur eines Dual-Phasenstahls mit Abbildung der von Mises-Spannungen auf den Einschlussoberflächen. (3D EBSD/FIB-Messungen wurden am Max-Planck Institut für Eisenforschung, Düsseldorf, Prof. D. Raabe, durchgeführt. Für Details sei auf Brands et al. 2011 verwiesen.
Quelle: Scheunemann und Schröder 2017

modernen Supercomputern durchführen zu können, ist zudem eine effiziente und parallel skalierbare Implementierung notwendig. Mit der in Kapitel „FE2TI: Eine extrem skalierbare Implementierung der FE²-Methode“ beschriebenen FE2TI-Software ist eine solche Umsetzung erstmals in Kombination mit Gebietszerlegungsverfahren gelungen.

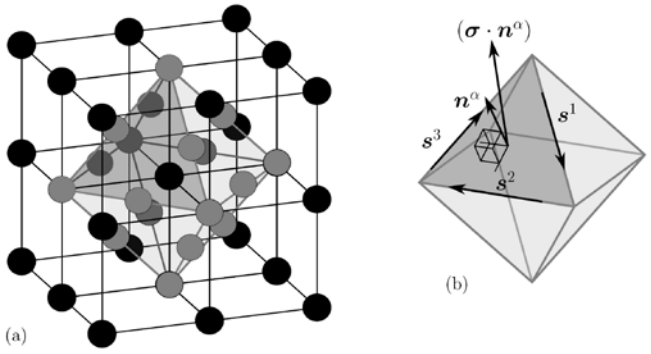
Vom Bauteil zur Mikrostruktur

Zur Berücksichtigung der realen Mikrostruktur in Computersimulationen sind Messungen notwendig, die eine geometrische Darstellung der inneren Struktur im Mikrometerbereich ermöglichen. Ein Beispiel eines Ausschnitts aus der Mikrostruktur eines DP-Stahls der Größe 15.9µm × 16.4µm × 5µm ist in Abbildung (3) dargestellt. Hier sieht man zusätzlich zur komplexen Geometrie die berechnete Spannungsverteilung auf der Oberfläche der Martensiteinschlüsse während eines Zugversuchs. Bereits hier wird der enorme Aufwand deutlich, der zur Simulation einer solchen Struktur notwendig ist, es wurde eine FE-Diskretisierung mit 8.500.000 Freiheitsgraden verwendet. In diesem Beispiel werden die beiden Bestandteile des DP-Stahls, Martensit und Ferrit, als homogene Phasen angenommen. Tatsächlich weisen beide Materialien eine polykristalline Struktur auf. Wie in vielen anderen Werkstoffen, sind im Stahl die Atome in einer Kristallstruktur angeordnet, die das Verhalten des Werkstoffs prägen. Die kristalline Gitterstruktur kann dabei durch eine Einheitszelle beschrieben werden, die je nach Material zum Beispiel kubisch flächenzentrierte (kfz), kubisch raumzentrierte (krz) Formen oder hexagonal dichteste Kugelpackungsform (hdP) aufweisen kann. Bekannte Vertreter der kubisch flächenzentrierten Kristallgitter sind neben Kupfer und Aluminium auch bestimmte Mischkristalle im Stahl, beispielsweise Austenit. Andere Zustandsformen des Stahls weisen kubisch raumzentrierte Einheitszellen auf, wie zum Beispiel Ferrit und Martensit. Verschiedene Typen von kristallinen Einheitszellen sind in Abbildung (4) dargestellt.

Wenn ein Kristallgitter verformt wird, geht man davon aus, dass reversible Verformungen durch eine Verzerrung des gesamten Gitters entstehen, die Anordnung der Atome zueinander aber gleich bleibt. Bei irreversiblen Verformungen gibt es Versetzungen im Gitter, die durch die relative Verschiebung der Atome zueinander charakterisiert sind. Um diese ursächlichen Mechanismen mit Hilfe geeigneter Modelle zu beschreiben, wird im Rahmen der Kontinuumsmechanik das Modell der Kristallplastizität verwendet.[2] Ähnlichkeiten im Deformationsverhalten kristalliner Strukturen wurde bereits in der ersten Hälfte des letzten Jahrhunderts untersucht und dienten dazu, Gesetzmäßigkeiten zu formulieren. Dabei wurde festgestellt, dass eine Scherung im Material parallel zu bestimmten Ebenen des Kristallgitters auftritt und Spannungen senkrecht zu diesen Gleitebenen einen untergeordneten Einfluss auf das Verformen des Kristalls haben, was im Schmidschen Schubspannungsgesetz mündete. Abbildung (5) zeigt dies in der Projektion des Spannungstensors σ auf eine Gleitebene mit dem Normalenvektor n^{α}, auf welcher die Gleitrichtungen



(4) Einheitszellen verschiedener Kristallgitterstrukturen: (a) kubisch flächenzentriert (kfz), (b) kubisch raumzentriert (krz) und (c) hexagonal dichteste Kugelpackung (hdP).
Quelle: Scheunemann und Schröder 2017



(5) (a) Gitterstruktur eines kfz-Kristalls mit den im Oktaeder zueinander angeordneten Gleitebenen (nur die Atome auf den Flächendiagonalen, die zum Oktaeder der Gleitflächen gehören, sind dargestellt) und (b) Projektion des Spannungstensors auf eine Gleitebene.
Quelle: Scheunemann und Schröder 2017

durch die Richtungsvektoren s^α definiert sind. Die Scherspannung τ^α auf den betrachteten Gleitebenen ergeben sich damit durch

$$\tau^\alpha = (\sigma \cdot n^\alpha) \cdot s^\alpha \qquad (1)$$

Zusammen definieren die Vektoren s^α und n^α Gleitsysteme, von denen je nach Kristallstruktur unterschiedlich viele existieren. Abhängig vom zugrundeliegenden Gittertyp können weitere Spannungskomponenten als die oben genannten Einfluss auf die Deformation haben, wie beispielsweise in kubisch raumzentrierten Kristallgittern, was hier allerdings nicht weiter vertieft wird. Im Allgemeinen bilden die Ebenen der dichtesten Anordnung der Atome im kfz-Kristallgitter die Gleitebene aus, die darauf liegenden Gleitrichtungen sind durch den dichtesten Atomabstand definiert. Durch unterschiedliche Abstände der Atome zueinander muss unterschiedlich viel Energie aufgebracht werden, um Versetzungen im Kristallgitter zu erzeugen beziehungsweise voranzutreiben. In der Richtung der dichtesten atomaren Packung ist dieser Abstand und damit auch die aufzubringende Energie am geringsten. Im kfz-Kristallgitter lassen sich acht Gleitebenen identifizieren, auf denen je drei Gleitrichtungen liegen. Daraus ergeben sich 24 (± 12) Gleitsysteme, auf denen jeweils eine Scherspannung nach Gleichung (1) definiert ist. In Abbildung (5) sind die Gleitebenen im kubisch flächenzentrierten Kristallgitter dargestellt. Dabei sind für eine bessere Übersichtlichkeit die flächenzentrierten Atome nur im Bereich des Oktaeders dargestellt, der die Lage der Gleitebenen zueinander verdeutlicht. Selbstverständlich enden die Gleitebenen nicht an der Kante des Oktaeders, sondern setzen sich durch das Kristallgitter entsprechend der Raumstruktur fort. Das Ende einer Gleitebene ist durch die Korngrenzen, die einen Kristall von seinem Nachbarn trennen, definiert.

Im Allgemeinen ist das Fließen, also die plastische Deformation, ein ratenabhängiger Prozess. Es gibt also einen Unterschied in der Spannungsantwort des Materials abhängig von der Geschwindigkeit, mit der eine Belastung aufgebracht wird. Zeitunabhängige Formulierungen werden jedoch häufig bei der Beschreibung quasistatischer Prozesse favorisiert. In diesem Fall unterscheidet man im Gegensatz zur ratenabhängigen Formulierung zwischen aktiven und nicht aktiven Gleitsystemen, wobei nur die aktiven Gleitsysteme zur Entstehung der plastischen Verzerrung beitragen. Die Schwierigkeit der ratenunabhängigen Formulierung besteht in der Nichteindeutigkeit der aktiven Gleitsysteme. Beispielhaft sei die Problematik der Nichteindeutigkeit der Lösung an einem akademischen Beispiel gezeigt, in dem sich im Material zwei senkrecht aufeinander stehende Gleitsysteme befinden, siehe Abbildung (6). Die axiale Belastung des Stabes führt zu identischen Schmid-Spannungen auf beiden Gleitsystemen und einer singulären Jakobimatrix $\boldsymbol{A}$, sodass das zu lösende Gleichungssystem die Form

$$\begin{bmatrix} a & a \\ a & a \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} b \\ b \end{bmatrix} \rightarrow \boldsymbol{A}\boldsymbol{\gamma} = \boldsymbol{b} \quad (2)$$

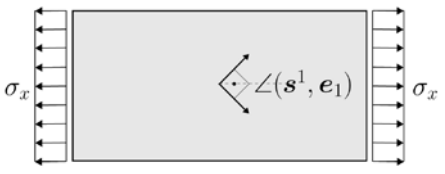
annimmt. In diesem Fall ist die Jakobimatrix singulär und eine Lösung des Gleichungssystems nur mit Hilfe spezieller Regularisierungstechniken möglich, die unter anderem von Miehe und Schröder untersucht wurden. Ein ratenabhängiger Ansatz regularisiert die Jakobimatrix und bestimmte Annahmen der Parameter liefern den ratenunabhängigen Spezialfall als Grenzwert des ratenabhängigen Modells. Aus diesen Gründen wird hier eine ratenabhängige Formulierung verfolgt.

Im kontinuumsmechanischen Materialmodell bestimmen Evolutionsgleichungen die Entwicklung der plastischen Verzerrungen $\dot{\boldsymbol{\varepsilon}}^p$, wobei gilt

$$\dot{\boldsymbol{\varepsilon}}^p = \sum_\alpha \dot{\gamma}^\alpha \, \mathrm{sym}(\boldsymbol{s}^\alpha \otimes \boldsymbol{n}^\alpha) \qquad (3)$$

mit $\dot{\gamma}^\alpha$ als der Rate der Gleitung auf dem System α . Eine Zeitintegration dieser Gleichung ist notwendig zur Lösung der Problemstellung, welche die zu bestimmenden Größen abhängig von der plastischen Gleitung definiert. Dabei ist die Rate der Gleitungen definiert als

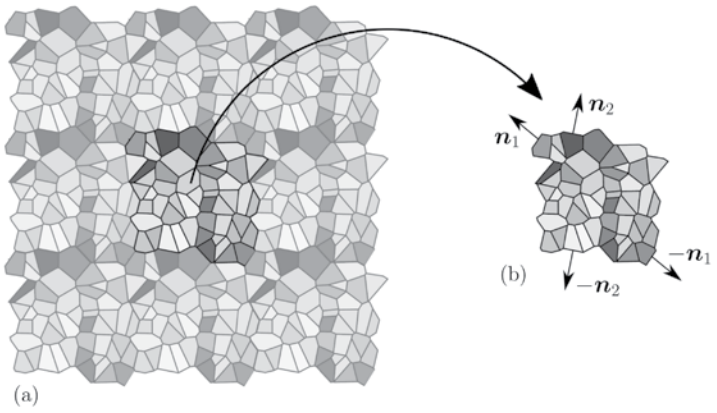
$$\dot{\gamma}^\alpha = \dot{\gamma}_0 \left| \frac{\tau^\alpha}{g^\alpha} \right|^{p-1} \frac{\tau^\alpha}{g^\alpha} \qquad (4)$$



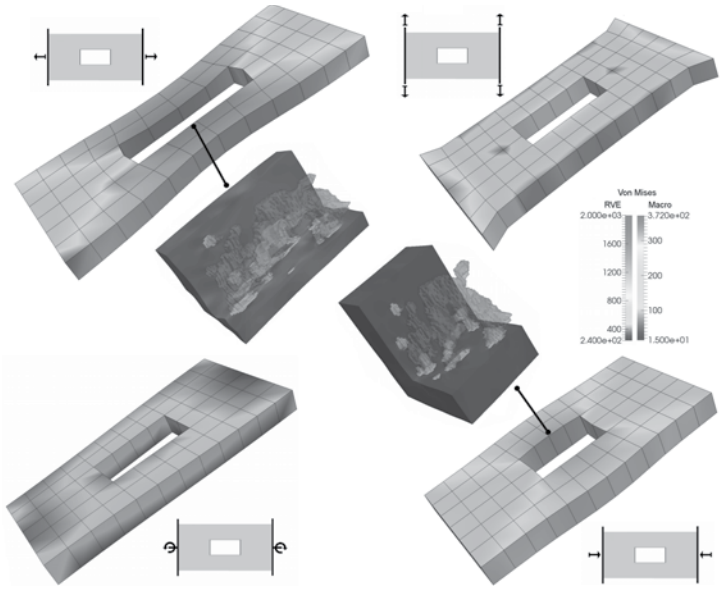
(6) Zugstab mit zwei senkrecht aufeinander stehenden Gleitsystemen.
Quelle: Miehe und Schröder 2001

mit den Viskositätsparametern p und $\dot{\gamma}_0$ und dem Gleitwiderstand g^α auf Gleitsystem α. Durch die Abhängigkeit der Scherspannung τ^α von der plastischen Gleitung auf allen Gleitsystemen muss hier ein Gleichungssystem gelöst werden, um die plastische Gleitung auf allen Systemen zu bestimmen. Dies wird iterativ mit Hilfe des Newton-Verfahrens umgesetzt. Verfestigungseffekte, also ein Ansteigen des Gleitwiderstandes während des Deformationsprozesses, werden durch zusätzliche Evolutionsgesetze definiert.

Die Betrachtung einer polykristallinen Mikrostruktur im Rahmen der FE²-Methode erfordert geeignete Beschreibungen ihrer Geometrie. Ein synthetisches Modell eines Polykristalls ist in Abbildung (7) dargestellt, wobei eine große Mikrostruktur durch Nebeneinandersetzen polykristalliner Einheitszellen erzeugt werden kann. In der Simulation muss nun nur die periodische Einheitszelle betrachtet werden, wobei periodische Randbedingungen zum Einsatz kommen, welche die Konformität der Verschiebungen und Spannungen auf den Randflächen sichern. Die Normalenvektoren assoziierter Randflächen zeigen dabei in entgegengesetzte Richtungen. Jedes der dargestellten Körner des Polykristalls wird als Einkristall mit einer bestimmten Orientierung betrachtet. Ihr Zusammenspiel im Polykristall ermöglicht eine realitätsnahe Simulation des oftmals komplexen Materialverhaltens, das auf der Makroskala zu beobachten und dessen Quantifizierung ohne die Betrachtung der mikrostrukturellen Effekte unmöglich ist. Zudem liefern Computersimulationen Einblicke in Effekte auf der Mikroskala des Materials, die ansonsten nur schwer oder unmöglich abbildbar sind, aber maßgeblich zum makroskopischen Materialverhalten beitragen. Die Realisierung eines virtuellen Labors zur Analyse komplexer dreidimensionaler Strukturen, wie der hier betrachteten Polykristalle, ist dabei nur unter Verwendung von Supercomputern und dem Einsatz effizienter Algorithmen möglich.



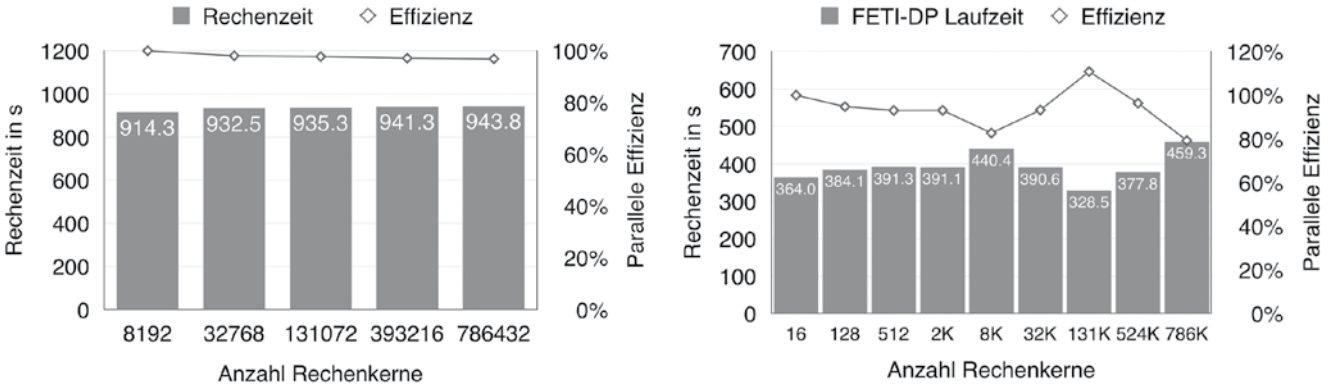
(7) (a) Polykristalline Mikrostruktur und (b) periodische polykristalline Einheitszelle. Die Randflächen stehen hierbei zueinander in Beziehung und weisen entgegengesetzte Normalenvektoren n auf.
Quelle: Scheunemann und Schröder 2017



(8) Vier verschiedene makroskopische Deformationsprozesse simuliert mit der FE2TI Software. Zwei der insgesamt 2048 mikroskopischen Randwertprobleme sind ebenfalls exemplarisch abgebildet.
Quelle: Klawonn, Lanser, Rheinbach 2017

FE2TI: Eine extrem skalierbare Implementierung der FE²-Methode

Die FE²-Methode punktet nicht ausschließlich aus der Perspektive der Mechaniker*innen durch die präzise Vorhersage der Spannungen auf mikroskopischer und makroskopischer Skala, sondern auch durch die effiziente und natürliche Umsetzbarkeit auf modernen Supercomputern. Dass deren hohe Rechenkapazität absolut notwendig für die Simulation von Deformationsprozessen von DP-Stählen ist,



(9) Links: Schwache parallele Skalierbarkeit der FE2TI-Software für ein hyperelastisches Problem.
Rechts: Schwache parallele Skalierbarkeit von FETI-DP für ein heterogenes hyperelastisches Randwertproblem.
Quelle: Weitere Details finden sich in Klawonn et al. 2015a; 2016

kann man anhand einer einfachen Rechnung veranschaulichen. Um die feinen mikroskopischen Strukturen aufzulösen, ist es notwendig, Finite Elemente in der gleichen und damit kleinen Größenordnung zu verwenden. Beispielsweise werden zwischen 10^3 und 10^9 Finite Elemente zur Diskretisierung eines Würfels mit einer Seitenlänge von einem Mikrometer benötigt. Rechnet man dies hoch, so erhält man eine Diskretisierung mit $10^{18} - 10^{24}$ Finiten Elementen für ein, immer noch kleines, $1m^2$ großes Stahlblech von lediglich einem Millimeter Dicke. Die unmittelbare Lösung solch gigantischer FE-Probleme ist tatsächlich nicht einmal mit modernsten Supercomputern zu erreichen - sie ist bereits zu aufwendig und erfordert den Einsatz effizienter Algorithmen und vor allem von Homogenisierungsansätzen wie der FE^2-Methode. Die FE^2-Methode verringert zunächst die Problemgröße um einen Faktor von bis zu 10^6 und ist im Hinblick auf die algorithmische Struktur exzellent für den Einsatz auf Supercomputern geeignet. Dies soll im Folgenden etwas näher erläutert werden.

Generell ist die Berücksichtigung der Architektur moderner Supercomputer essentiell für die Konstruktion effizienter numerischer Verfahren. Die Zeiten, in denen sich die Rechengeschwindigkeit einzelner Prozessoren alle 12 bis 18 Monate verdoppelte und somit die gleiche Simulation in der Hälfte der Zeit durchgeführt werden konnte, sind lange vorbei. Moderne Computerarchitekturen setzen stattdessen zur Leistungssteigerung auf den parallelen beziehungsweise gleichzeitigen Einsatz vieler Rechenkerne. So hat der derzeit schnellste Supercomputer der Welt, der *Sunway TaihuLight* (China), mehr als 10 Millionen Rechenkerne und Europas größter Rechner *Piz Daint* (Schweiz) immerhin derer 200.000. Geeignet für den Einsatz auf diesen Computern sind somit nur Verfahren, deren Berechnungen sich zum größten Teil gleichzeitig durchführen lassen. Die FE^2-Methode erfüllt diese Bedingung in besonderem Maße, da die Simulation der vielen Randwertprobleme auf der mikroskopischen Skala nur durch gelegentliche makroskopische Lösungen gekoppelt werden und somit nahezu unabhängig und parallel berechnet werden können.

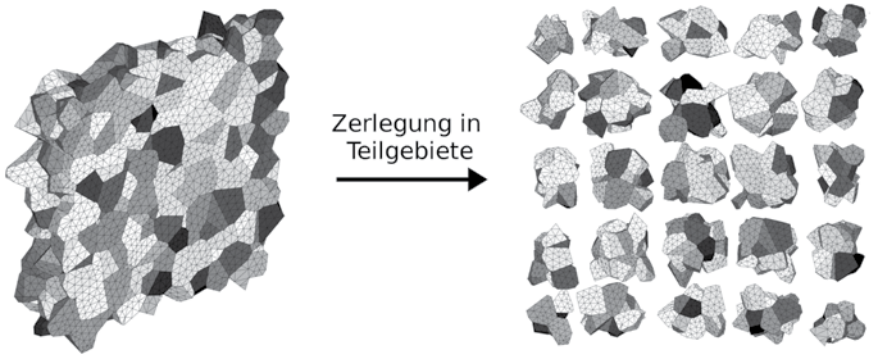
Die erste konsequente Implementierung (FE2TI) der FE^2-Methode für den Einsatz auf den weltweit größten Supercomputern erfolgte im Rahmen des Projekts *EXASTEEL* innerhalb des DFG-Schwerpunktprogramms 1648 *Software for Exasxale Computing*. Mit Hilfe dieses Softwarepaketes wurden bereits erfolgreich große Simulationen auf der JUQUEEN im Forschungszentrum Jülich, einem Supercomputer vom Typ IBM BlueGene/Q, durchgeführt. Die Ergebnisse dieser Simulationen sind in Abbildung (8) dargestellt. Dabei wurde ein Lochblech verschiedenen Deformationen unterworfen und in jeder Simulation wurden 512 mikroskopische Randwertprobleme auf der mikroskopischen Skala, mehrere hundert Mal, parallel auf über 250.000 Rechenkernen gelöst. Bislang wurde ein elasto-plastisches Materialgesetz verwendet, das sowohl die Heterogenität als auch die plastische Verformung des DP-Stahls berücksichtigt, jedoch noch nicht die explizite Kristallstruktur auf der mikroskopischen Skala betrachtet.

Die FE2TI-Software bewies bereits im Jahr 2015 ihre hohe Effizienz und parallele Skalierbarkeit. Simulationen unter Ausnutzung aller verfügbaren Rechenkerne liefen bereits auf Mira (Argonne National Laboratory, USA) [4], dem derzeit neuntschnellsten Supercomputer der Welt, sowie JUQUEEN. Letzteres führte wiederum zur Aufnahme von FE2TI in den sogenannten High-Q Club [5], einer Sammlung von Software, die effizient auf der ganzen Maschine liefen. Eine Studie der schwachen parallelen Skalierbarkeit auf Mira ist in Abbildung (9) (links) zu finden. Die Anzahl der zu lösenden mikroskopischen RWP steigt hier proportional zur Anzahl der verwendeten Rechenkerne und die benötigte Rechenzeit bleibt nahezu konstant, was im Sinne der

schwachen Skalierbarkeit ein optimales Ergebnis darstellt. Die Parallelisierung in FE2TI wurde mit Hilfe von MPI (Message Passing Interface) umgesetzt, einem Protokoll, das den Datenaustausch zwischen einzelnen Prozessen beziehungsweise Rechenkernen verwaltet. Verschiedene Prozessoren sind in sogenannten Kommunikatoren organisiert und Daten werden zwischen einzelnen Prozessoren oder innerhalb des gesamten Kommunikators ausgetauscht, gesendet oder empfangen. In FE2TI werden zunächst alle Prozessoren einem globalen Kommunikator zugeordnet, der für die wenigen makroskopischen Berechnungen zuständig ist. Auf einer zweiten Ebene werden die Prozessoren gleichmäßig auf mehrere gleichwertige Kommunikatoren beziehungsweise Subkommunikatoren aufgeteilt, wobei jeder dieser Kommunikatoren genau einem mikroskopischen Randwertproblem zugeordnet wird. Diese Struktur erlaubt die effiziente und parallele Durchführung aller mikroskopischen Berechnungen, ohne unnötigen Datenaustausch oder Synchronisation der Subkommunikatoren untereinander. Daten werden ausschließlich über den globalen Kommunikator in den Phasen der makroskopischen Berechnungen zwischen den Subkommunikatoren ausgetauscht.

Zur Steigerung der parallelen Effizienz kommen robuste und ebenfalls für parallele Berechnungen ausgelegte Gebietszerlegungsverfahren bei der Lösung aller mikroskopischen Randwertprobleme zum Einsatz. In FE2TI fällt die Wahl auf das FETI-DP (Finite Element Tearing and Interconnecting –Dual-Primal) Gebietszerlegungsverfahren – diese Wahl ist auch ursächlich für das TI im Namen FE2TI. Gebietszerlegungsverfahren setzen, wie auch das FE²-Verfahren, auf das *Teile und Herrsche*-Prinzip. Wie in Abbildung (10) schematisch dargestellt, wird das gesamte FE-Problem (Abb. 10, links), hier eine Finite Elemente-Diskretisierung einer polykristallinen Mikrostruktur, in kleine Teilprobleme zerlegt, indem die zugrunde liegende Geometrie in Teilgebiete zerlegt wird (Abb. 10, rechts). Die Lösung der einzelnen Teilprobleme kann im Wesentlichen gleichzeitig erfolgen, was Gebietszerlegungsverfahren ebenfalls für den Einsatz auf modernen Computerarchitekturen qualifiziert. In der Regel werden die Teillösungen in einem iterativen Prozess wieder zu einer Gesamtlösung zusammengefügt, weshalb Gebietszerlegungsverfahren auch zur Klasse der iterativen Verfahren gezählt werden. Obgleich FETI-DP innerhalb der FE2TI-Software lediglich als Löser für die mikroskopischen Randwertprobleme eingesetzt wird und damit nur auf den vergleichsweise kleinen Subkommunikatoren operiert, sind viele Gebietszerlegungsverfahren absolut in der Lage auch als alleinstehende Löser große FE-Probleme zu bewältigen. Eine Studie zur schwachen Skalierbarkeit von FETI-DP auf Mira findet sich in Abbildung (9) (rechts).

Zusammengefasst verfügt die FE2TI-Software somit über zwei Level paralleler Berechnungen. Zunächst erzeugt die FE²-Methode viele voneinander unabhängige, beziehungsweise durch das makroskopische Problem leicht gekoppelte Randwertprobleme, die wiederum durch das FETI-DP-Verfahren in viele Teilgebiete zerlegt werden. Durch diese Struktur ist das Potential für den Einsatz auf den weltgrößten Supercomputern sehr groß und die Skalierbarkeit noch längst nicht ausgereizt.



(10) Zerlegung einer polykristallinen Mikrostruktur in Teilgebiete.
Quelle: Klawonn, Lanser, Scheunemann, Schröder 2017

Summary

Modern high-tech steels offer many possibilities for improvements in engineering applications, while virtual labs can help to taylor the materials to be best suited for an underlying application. A realistic modeling of the inherent phenomena acting on the microstructural level are inevitable for reliable computations. In the case of dual-phase steel, the interaction of the two phases in the microstructure together with the polycrystalline structure of each phase contribute to the complex macroscopic material response. Homogenization approaches are necessary to account for the effects on both scales, such as the FE²-method. Due to the enormous size of the simulations, the computational power of modern supercomputers is necessary. State-of-the-art computer architectures provide up to ten million parallel processing units and to exploit them efficiently, it is mandatory that the mayor part of the underlying computations of a numerical method can be performed in parallel. This is possible in the FE²-method as well as in domain decomposition methods as FETI-

DP (Finite Element Tearing and Interconnecting – Dual-Primal). The combination of both methods, named FE2TI, managed to run efficiently on some of the largest supercomputers.

Anmerkungen

1) Schröder 2014
2) Hierzu sind in Havner 1982, Taylor 1934, Peirce et al. 1982, Asaro 1983 die grundlegenden Ansätze beschrieben.
3) Miehe und Schröder 2001
4) siehe Klawonn et al. 2015a; 2016
5) Klawonn et al. 2015b

Literatur

– R. J. Asaro. Crystal plasticity. Transactions of the ASME, 50:921–934, 1983.
– D. Brands, D. Balzani, J. Schröder, and D. Raabe. Simulation of DP-steels based on statistically similar representative volume elements and 3D EBSD data. In Computational Plasticity XI – Fundamentals and Applications, pages 1552–1563, Barcelona, Spain, September 7–9 2011.
– K.S. Havner. Mechanics of Solids. the Rodney Hill 60th Anniversary Volume, chapter The theory of finite plastic deformation of cryscrystal solids, pages 265–302. Pergamon Press: Oxford, 1982.
– A. Klawonn, M. Lanser, and O. Rheinbach. Toward extremely scalable nonlinear domain decomposition methods for elliptic partial differential equations. SIAM Journal on Scientific Computing, 37(6):C667–C696, 2015a. doi: 10.1137/140997907. URL http://dx.doi.org/10.1137/140997907.
– A. Klawonn, O. Rheinbach, and M. Lanser. FE2TI. EXASTEEL – bridging scales for multiphase steels, 2015b. http://www.fz-juelich.de/ias/jsc/EN/Expertise/High-Q Club/FE2TI/node.html.
– A. Klawonn, M. Lanser, and O. Rheinbach. FE2TI: Computational scale bridging for dual-phase steels. In IOS Series Advances in Parallel Computing, Volume 27, Parallel Computing: On the Road to Exascale; Proceedings of ParCo2015, pages 797–806, 2016. URL http://dx.doi.org/10.3233/978-1-61499-621-7-797. Also TUBAF Preprint: 2015-12, http://tu-freiberg.de/fakult1/forschung/preprints.
– C. Miehe, J. Schröder. A comparative study of stress update algorithms for rate independent and rate dependent crystal plasticity. International Journal for Numerical Methods in Engineering, 50:273–298, 2001.
– D. Peirce, R. J. Asaro, and A. Needleman. An analysis of nonlinear and localized deformation in ductile single crystals. Acta Metall., 30:1087–1119, 1982.
– L. Scheunemann, J. Schröder. Preliminary studies for the prediction of yield surfaces with crystal plasticity models. Technical Report 141, Institut für Mechanik, Bauwissenschaften, Universität Duisburg-Essen, 2017.
– J. Schröder. Plasticity and Beyond - Microstructures, Crystal-Plasticity and Phase Transitions (CISM Lecture Notes 550, Eds. J. Schröder, K. Hackl), chapter A numerical two-scale homogenization scheme: the FE^2 -method. Springer, 2014.
– G. I. Taylor. The mechanism of plastic deformation of crystal. Part I: Theoretical. Proc. Roy. Soc. London A., 145:362–387, 1934.

Lisa Scheunemann. Foto: Vladimir Unkovic

*Die Autor*innen*

Axel Klawonn ist Professor für Angewandte Mathematik an der Universität zu Köln und derzeit Prodekan für Forschung der Mathematisch-Naturwissenschaftlichen Fakultät. Seit 2012 ist er als Professor an der Fakultät für Mathematik der Universität Duisburg-Essen kooptiert. Sein Arbeitsgebiet liegt in der Numerischen Mathematik, insbesondere in der Numerik partieller Differentialgleichungen, der Numerischen Analysis und Linearen Algebra sowie dem Wissenschaftlichen Rechnen mit den Schwerpunkten Gebietszerlegungsverfahren, paralleles Wissenschaftliches Rechnen und Simulation in Festkörper- und Biomechanik. Zusammen mit seiner Arbeitsgruppe sowie nationalen und internationalen Kooperationspartnern arbeitet er an der Entwicklung effizienter, numerischer Methoden zur Simulation von Problemen aus den Ingenieur-, Natur-, und Lebenswissenschaften. Dies umfasst die Entwicklung effizienter Algorithmen, ihre theoretische Analyse, sowie deren Implementierung auf sehr großen Parallelcomputern mit bis zu mehreren hunderttausend Prozessorkernen.
Axel Klawonn studierte Mathematik und Informatik an der Universität Münster, wo er auch promovierte. An seiner Dissertation arbeitete er in Münster und den USA, wo er

am Courant Institute of Mathematical Sciences in New York als Visiting Graduate Student studierte und dort insbesondere mit Olof Widlund arbeitete. Nach seiner Promotion war er als Wissenschaftlicher Assistent an der Universität Münster und als Nachwuchswissenschaftler am Institut für Algorithmen und Wissenschaftliches Rechnen (SCAI) der GMD (später Fraunhofer-Gesellschaft) in Sankt Augustin tätig. Er habilitierte sich 2001 an der Universität zu Köln und erhielt 2002 einen Ruf an die Universität Essen (später Universität Duisburg-Essen) auf eine Professur für Numerische Mathematik. Nach einem Ruf an die TU Darmstadt im Jahr 2006 hatte er von 2007 bis 2012 den Lehrstuhl für Numerische Mathematik an der Universität Duisburg-Essen inne. Im Jahr 2012 wurde er an die Universität zu Köln berufen. Er hielt sich zu längeren Forschungsaufenthalten im Ausland auf, unter anderem am Courant Institute of Mathematical Sciences in New York, USA, an der University of Colorado at Boulder, USA, am Argonne National Laboratory, USA und an den Sandia National Laboratories, Albuquerque, USA.

Martin Lanser studierte Mathematik mit Nebenfach Informatik an der Universität Duisburg-Essen, wo er 2011 seinen Abschluss als Diplom-Mathematiker erlangte. Anschließend war er als Wissenschaftlicher Mitarbeiter tätig – zunächst an der Universität Duisburg-Essen und ab dem Jahr 2012 an der Universität zu Köln. Dort promovierte er auch im Jahr 2015 im Fach Mathematik. Die Themen der Promotion waren nichtlineare Varianten der Gebietszerlegungsverfahren FETI-DP und BDDC, sowie die parallele und hoch skalierbare Implementierung von Multiskalen-Methoden. Während seiner Promotion führte ein längerer Forschungsaufenthalt Martin Lanser an das Lawrence Livermore National Laboratory, USA. Zudem arbeitete er aktiv im Projekt EXASTEEL mit, welches Teil des DFG Schwerpunktprogramm SPPEXA ist. Seit dem Jahr 2016 ist er als Akademischer Rat an der Universität zu Köln tätig.

Lisa Scheunemann studierte von 2006 bis 2009 Bauingenieurwesen (B.Sc.) an der Universität Duisburg-Essen und erlangte dort 2011 ihren Abschluss im Masterstudiengang „Computational Mechanics“. Seit 2011 ist sie am Institut für Mechanik der Universität Duisburg-Essen als wissenschaftliche Mitarbeiterin tätig. Im Rahmen ihrer Promotion beschäftigt sie sich mit gekoppelten Mikro-Makro-Übergangsmodellen für DP Stähle und den zugrundeliegenden Materialmodellen sowie der Konstruktion statistisch ähnlicher RVEs.

Jörg Schröder studierte zunächst Bauingenieurwesen an der FH Hildesheim/Holzminden; dieses Studium schloss er 1988 mit der Diplomprüfung ab. Anschließend nahm er an der Universität Hannover das Studium des Bauingenieurwesens auf, wo er 1989 Stipendiat der Studienstiftung des Deutschen Volkes wurde. Ende 1992 schloss er sein Studium ab und wurde Assistent am Institut für Baumechanik und Numerische Mechanik im Fachbereich Bauingenieurwesen der Leibniz Universität Hannover. Mit einer Dissertation über die theoretische und algorithmische Beschreibung anisotropen Materialverhaltens wurde er Ende 1995 zum Dr.-Ing. promoviert. Ab 1996 war Jörg Schröder Akademischer Rat am Institut für Mechanik (Bauwesen) am Lehrstuhl I bei Christian Miehe an der Universität Stuttgart, wo er 2000 in dem Fach Mechanik mit einem Thema im Bereich der Homogenisierungsmethoden in der nichtlinearen Kontinuumsmechanik habilitierte. Im Herbst 2000 folgte er einem Ruf auf eine Professur für Mechanik an die Technische Universität Darmstadt. Bereits im Jahre 2001 folgte ein Ruf auf eine Professur für das Fachgebiet Mechanik der Fakultät Bauwesen an die Universität Duisburg-Essen, den er zum Herbst 2001 annahm. Von 2011 bis Anfang 2015 war er Prorektor für Forschung, wissenschaftlichen Nachwuchs und Wissenstransfer, seit März 2016 ist er es erneut. Ferner ist er Mitglied der Akademie der Wissenschaften und der Literatur Mainz und im Auswahlausschuss für Forschungsstipendien der Alexander von Humboldt-Stiftung. Die Forschungsschwerpunkte von Jörg Schröder liegen im Bereich der Formulierung und Konstruktion anisotroper Materialgleichungen sowie derer algorithmischer Durchdringung und Umsetzung und der Weiterentwicklung der Finite-Elemente-Methode. Im Rahmen der theoretischen und computerorientierten Materialtheorie werden hierbei die folgenden Themenkomplexe behandelt: Anisotrope Elasto-Plastizität bei finiten Deformationen, anisotrope Hyperelastizität, insbesondere unter Berücksichtigung spezieller mathematischer Anforderungen, mit Anwendungen in der Biomechanik und magneto-elektromechanisch gekoppelte Funktionskeramiken. Als spezielle Zugänge dienen hierbei überwiegend Elemente der Invariantentheorie und geeignete Tensor-Repräsentations-Theoreme. Er ist Sprecher der DFG-Forschergruppe 1509 „Ferroische Funktionsmaterialien – Mehrskalige Modellierung und experimentelle Charakterisierung“ und des DFG-Schwerpunktprogramms 1748 „Zuverlässige Simulationstechniken in der Festkörpermechanik – Entwicklung nichtkonventioneller Diskretisierungsverfahren, mechanische und mathematische Analyse“.

Axel Klawonn und Martin Lanser. Foto: Vladimir Unkovic

Reinhard Schütte. Foto: Vladimir Unkovic

Eine Herausforderung für Unternehmen in Zeiten der Digitalisierung ist es, das Management der gemeinhin als Standardsoftware deklarierten Enterprise Systems in der Cloud zu bewältigen.

Immer mehr Standard

Cloud Computing im Kontext von Enterprise Systems

Von Reinhard Schütte

Anwendungssysteme sind die automatisierten Teile der Informationssysteme von Unternehmen. Die Digitalisierung führt dazu, dass der Ausdehnungsbereich der Anwendungssysteme zunimmt und die ausschließlich dem Organisationssystem zugänglichen Informationen, das heißt die nicht automatisierten Informationen, abnehmen.[1] Der Erfolg von Anwendungssystemen in Unternehmen ist nur durch die situative Einbettung in das Unternehmen möglich. Die integrative Berücksichtigung dieser beiden Systemarten ist der zentrale Gegenstandsbereich der Wirtschaftsinformatik, die sich nicht nur mit dem technischen System auseinandersetzt.

Die Anwendungssysteme werden aus einer betriebswirtschaftlichen Perspektive von den Softwareherstellern dominiert, die immer umfassendere „Standardsysteme" – sogenannte Enterprise Systems (ES) – anbieten.

Enterprise Systems – keine Standardsoftware im engeren Sinne

Eine besondere Bedeutung bei der Einführung, dem Betrieb und auch der strategischen Würdigung von ES kommt dem Verständnis zu, was unter Standardsoftware zu verstehen ist. Dabei dürfte bereits die Wahl des Wortes Standardsoftware zu vielen Missverständnissen in der Wissenschaft und auch in der Praxis geführt haben. Möglicherweise angeregt durch die Nutzung von Standardsystemen wie den Microsoft-Office-Produkten wurde unter Standardsoftware der universelle Einsatz des gleichen Softwaresystems in unterschiedlichen Kontexten verstanden. Während dieses Verständnis für die Textverarbeitung, das Tabellenkalkulations-, das Präsentations-, das E-Mail-Programm oder auch die vielen anderen „Apps" auf dem Smartphone zutreffend sein mag, so gilt dies nicht für Enterprise Systems.

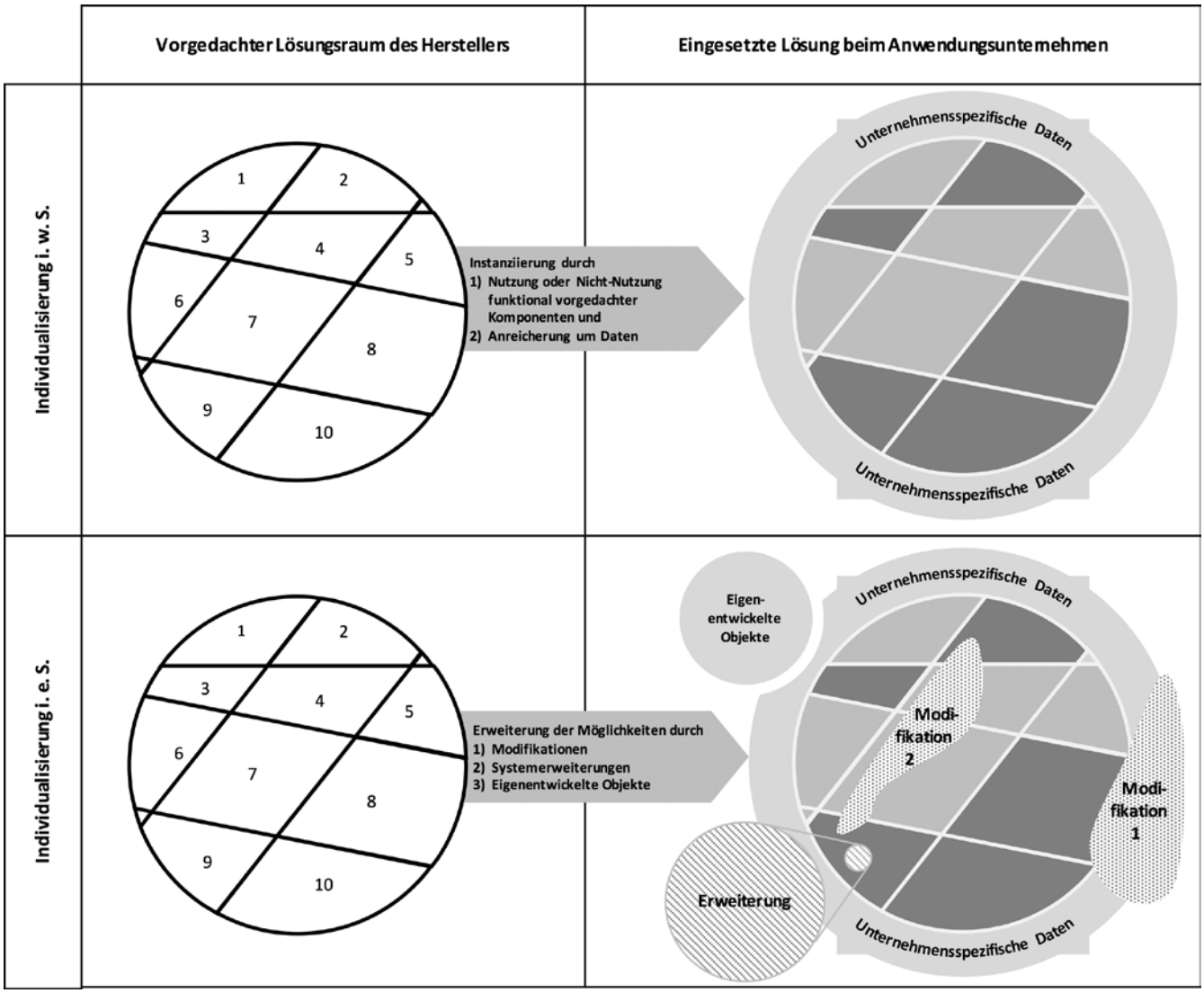
Unter ES werden die zentralen Applikationen in Unternehmen subsumiert, die sich in den letzten Jahrzehnten in Unternehmen global etabliert haben. ES sind integrierte

Anwendungssysteme, die die Wertschöpfungskette innerhalb eines Unternehmens und auch zu den Lieferanten und Kunden abdecken sollen. Je nach Ausdehnungsgrad umfassen sie die klassischen Enterprise-Resource-Planning-Funktionen, Customer-Relationship-Management- sowie Supplier-Relationship-Management-Funktionalitäten.[2] Aus einer historischen Betrachtung sind ES aus MRP-, MRP II- und ERP-Systemen hervorgegangen.[3] Die Historie offenbart auch den Ausdehnungsgrad von Anwendungssystemen innerhalb von Organisationen: ein stetiger Trend zu einer weitergehenden Digitalisierung von Unternehmen, der seit Jahrzehnten fortgeführt wird und von Softwareherstellern wie SAP, Oracle oder Microsoft vorangetrieben wurde. Insbesondere auch durch den globalen Trend der Prozessorientierung, der durch ERP-Systeme hochgradig gefördert wurde, wurden diese Systeme in sehr vielen Unternehmen dominant.

Die überragende Bedeutung von ES für Unternehmen, die keine Individualsoftware einsetzen, eröffnet nun die Frage, was für eine Art von Software diese darstellen. Einige Soziologen haben aus ihrer Forschungsperspektive noch Anfang des Jahrtausends den potentiellen Erfolg von Standardsoftwareherstellern für unmöglich erachtet:

„For many social scientists, especially those informed by sociology and anthropology, the large software suppliers like SAP should not be successful. Sociological/anthropological theory tells them that organisations are too diverse to deploy these highly generic systems. Many studies therefore end up suggesting – on the basis of difficulties and complications witnessed during fieldwork – that ERP systems and the like have no more than limited potential for extension.“[4]

Der Ausgangspunkt der Missinterpretationen liegt in dem Verständnis dessen, was standardisierte ERP-Systeme sind. Es gibt kein ERP-System, welches gleichermaßen in



(1) Die Individualisierungsstufen bei der Einführung von Enterprise Systems.
Quelle: eigene Darstellung

vielen Unternehmen im Einsatz ist. Dies soll als Ausgangspunkt der weitergehenden Ausführungen dienen, wenn die Rede von Standardisierung ist. Es gibt bei ES keine Software, die von dem Lieferanten lauffähig ausgeliefert wird. Es ist der fundamentale Unterschied zwischen „echter" Standardsoftware wie Microsoft Word und einem ES, dass ein ES erst durch den Prozess der „Entstandardisierung" einsatzfähig wird.[5] Ein ES bildet eben nicht eine denkmögliche Situation ab, sondern versucht Software für eine Klasse von Unternehmenssituationen darzustellen. Somit ist es aber zwingend, dass von einer abstrakteren Perspektive (Klasse von Unternehmen) auf eine konkrete, anwendungsbezogene Ebene gewechselt wird (Instanz, ein Unternehmen). Während bei der Instanziierung von Software von „echter" Standardsoftware keine Veränderung der funktionalen Fähigkeiten der Software stattfindet, wird bei ES in der Regel ein zweiphasiger Vorgang erforderlich. Erstens wird der vordefinierte funktionale Lösungsraum des ES-Herstellers reduziert (siehe Abb. 1, oberer Teil). Zweitens werden mit Hilfe technischer Lösungsmöglichkeiten des Enterprise Systems (der Entwicklungsumgebung) neue Lösungen definiert, die von dem jeweiligen Software-Hersteller nicht vorgesehen oder umgesetzt wurden (siehe Abb. 1, unterer Teil).

Individualisierung von Enterprise Systems

Anwendungsunternehmen haben aus der Notwendigkeit heraus, diese Form der Reduktion und der Erweiterung der Lösung bei der Einführung gleichermaßen umzusetzen, in der Regel die Notwendigkeit, auf Beratungsinstitutionen im Netzwerk der jeweiligen ES-Hersteller zurückzugreifen. Dabei wird zwar eine standardnahe Einführung häufig als Postulat formuliert; dessen Rationalität erscheint aber keinesfalls zwingend. Die Frage, was noch als standardkonform oder aus wirtschaftlichen Gründen für angemessen gelten kann und was für die weitere Nutzung des Enterprise Systems als nicht mehr adäquat gelten sollte, darüber gibt es in der wissenschaftlichen Literatur kaum eine Antwort.

ES sind nach der Installation nicht lauffähig, das heißt das Anwendungsunternehmen kann sich noch nicht einmal davon überzeugen, was das ES kann. Empirische Untersuchungen deuten an, dass wenige Unternehmen Systeme mit nur geringen Customizing-Aktivitäten oder Veränderungen implementieren.[6] Dabei wird hier unter Customizing die Einrichtung der Daten und Funktionen verstanden, die ein Unternehmen bei der Einführung eines ES gemäß der vom Softwarehersteller implementierten Vorgehensweise vorzunehmen hat. Allerdings bleibt unklar, was unter geringen Customizing-Aktivitäten zu verstehen ist. Die Anreicherung der Systeme um kundenspezifische Daten ist für die Funktionsfähigkeit ohnehin zwingend, sodass ohne ein im Sinne des Herstellers vorgenommenes Customizing die Systeme in keinem Fall ihre betriebswirtschaftlichen Aufgaben erfüllen können. Dabei gibt es aktuell kein wissenschaftlich akzeptiertes Verfahren, was ein Customizing-Grad ist, wann ein System stärker oder weniger stark „customized" ist und vor allem wie es eine situative, die Ziele des Unternehmens beachtende, hinreichende Lösung gibt.

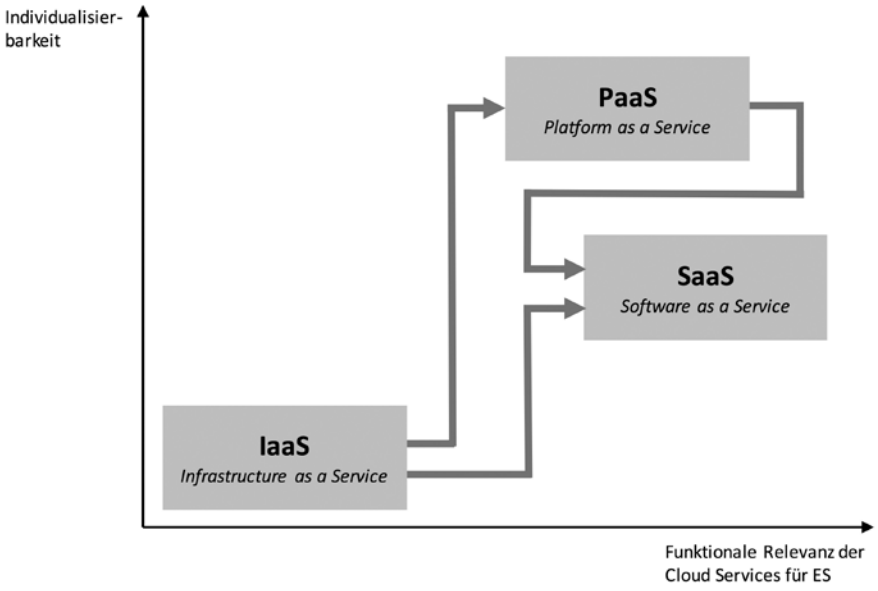
Die im Zuge des Customizings vorzunehmende Reduktion des funktionalen Möglichkeitsraumes des Enterprise Systems kommt strategisch eine nur geringe Bedeutung zu, denn sie verändert nicht den Softwarecode des Systems. Daher wird das Customizing hier auch nur als Individualisierung im weiteren Sinne verstanden. Die Individualisierung der Software im engeren Sinne ist immer dann gegeben, wenn ein Anwendungsunternehmen die Software entweder mit Hilfe der vom Hersteller bereitgestellten Erweiterungsmöglichkeiten verändert oder gar mit der bereitgestellten Entwicklungsumgebung gänzlich eigenständige Lösungen entwickelt (vgl. hierzu Abb. 1, unterer Teil).

Bei den standardmäßig vorgesehenen technischen Individualisierungsoptionen können unter anderem unterschieden werden:

- auf der Benutzer*innenebene: Nutzung von standardmäßig vorgesehenen Exits (GUI-Exits) oder Definition von eigenen graphischen Benutzeroberflächen, die der benutzerspezifischen Interaktion Rechnung tragen und sich nicht oder nur wenig an der Struktur des Mensch-Maschine-Dialogs des Enterprise Systems ausrichten.
- auf der Applikationsebene: Veränderung funktionaler oder prozessualer Gegebenheiten der Standardsoftware, indem Funktionen in einer veränderten Ablauflogik verwendet werden, neue Funktionen oder neue Prozesse definiert und anschließend mit den Entwicklungswerkzeugen des Standardsoftwareherstellers umgesetzt werden.

Die Nutzung der Entwicklungsumgebung der ES-Hersteller bedeutet, dass das Anwendungsunternehmen von dem vorgedachten Weg des Herstellers in seiner betriebswirtschaftlichen Deutung der Anwendungssituation abweichen kann. Die damit in der Regel einhergehende längerfristige Modifikation gegenüber der Applikationslogik des ES-Herstellers ist vor allem vor dem Hintergrund der strategischen Differenzierung zu Wettbewerbern zu bewerten, die auch das Produkt des Standardsoftwareherstellers nutzen, sofern den digitalen Fähigkeiten von Unternehmen eine strategische Relevanz zugesprochen wird, die dieses von Wettbewerbern abgrenzt.

Empirische Erkenntnisse über die Individualisierung von ES im engeren Sinne existieren kaum. Es ist nach den Einschätzungen des Verfassers insbesondere bei den großen Konzernen ein erheblicher Umfang an Programmen entstanden, die die Einsetzbarkeit der ES erst ermögli-



(2) Cloud-Lösungen und strategische Bedeutung aus Sicht von Enterprise Systems.
Quelle: eigene Darstellung

chen und dadurch auch einen Wettbewerbsvorteil gegenüber Konkurrenten eröffnen. Damit wird die ehemals geführte Diskussion von Standard- versus Individualsoftware aus einer strategischen Perspektive obsolet. Die Unternehmen entwickeln längst auf Basis der Release-Stände von Softwareherstellern eigenständige Release-Versionen der bei ihnen eingeführten Systeme, sodass die entstandenen Systeme einen strategischen Vorteil bieten können (aber nicht müssen). Sie sind ein eigenständiger Softwarelieferant geworden, der neben den Entwicklungen des ES-Herstellers eine eigene Strategie zur Weiterentwicklung der Enterprise-Software verfolgt. Die weiterhin enormen Programmieraufwendungen, die in diesen Unternehmen zu beobachten sind, deuten jedenfalls an, dass noch immer Software individuell entwickelt wird. Dabei ist lediglich der Gegenstandsbereich der Individualisierung verändert worden. An Stelle der großen eigenentwickelten Systeme, die komplett in den Unternehmen auf Basis eigener Datenstrukturen, eigener Begrifflichkeiten und eines eigenen Verständnisses von den Prozessen des Unternehmens entstanden sind, werden heute auf Basis von Strukturen, Prozessen und Begrifflichkeiten des Softwareherstellers und auch der zur Verfügung stehenden Technologien die unternehmenseigenen Prozesse „in der Welt der Enterprise Systems“ konstruiert.

Enterprise Systems in der Cloud

Cloud Computing, die Bereitstellung von mehr oder weniger umfangreichen Soft- oder Hardware-Services im Internet, ist seit geraumer Zeit ein nachhaltiger Trend, mit dem sich die Literatur wie auch die betriebliche Praxis gleichermaßen beschäftigt. Zunächst ist beim Angebot von Services in der Cloud zu konstatieren, dass die Technologiebranche eine Branche ist, in der wenige Anbieter auf der Welt den Markt dominieren; es liegt ein Angebotsoligopol vor. Dieser wettbewerbspolitische Hintergrund ist bei den Anwendungsunternehmen immer zu bedenken, denn es gibt ausgeprägte Prinzipal-Agenten-Probleme, die in der IT-Branche nur ungenügend erörtert werden, obgleich sie insbesondere wegen diverser informationsasymmetrischer Gegebenheiten zu erheblichen Problemen führen können.

Die grundsätzliche Differenzierung von Cloud Services folgt in der Regel der Dreiteilung von Infrastructure as a Service, Platform as a Service und Software as a Service.[7] Im Kontext von Enterprise Systems sind die drei unterschiedlichen Cloud-Angebote für Unternehmen von differierender strategischer Implikation:

Bei einer Diskussion von ES unter der Problemstellung, ob *Infrastructure as a Service* ein relevantes, neues Entscheidungsproblem darstellt, ist festzustellen, dass die Verlagerung des Rechenzentrum-Betriebs ein bekanntes Problem darstellt. Dieses betriebswirtschaftliche Entscheidungsproblem ist von seiner Struktur nicht neu, denn es wird unter dem Rubrum „Outsourcing“ seit Langem diskutiert. Den Entscheidungskriterien der Vergangenheit sind auch keine grundsätzlich neuen hinzuzufügen. Ein stärker zu gewichtendes Kriterium dürfte aufgrund der Risiken dabei das Thema der Datensicherheit sein, denn das Sicherheitsproblem ist in der Cloud ausgeprägter. Welches führende Unternehmen einer Branche ist daran interessiert, dass die besonders sensiblen Einkaufsdaten oder Kundenbeziehungsdaten in der Cloud potentiellen Angriffen ausgesetzt sind?

Bei den unter *Platform as a Service* offerierten Angeboten der ES-Hersteller, wie beispielsweise der SAP HANA Cloud Platform, wird eine Entwicklungsplattform in der Cloud für Unternehmen bereitgestellt. Die Zielsetzung der ES-Hersteller besteht darin, über das traditionelle Angebot hinausgehende Lösungen für Kunden verfügbar zu machen, die über die Cloud dann auch – analog dem Apple Store – als Software bezogen werden können. Außerdem besteht für Unternehmen die Möglichkeit, die Individualisierung ihres eigenen ES auf diese Art umzusetzen. Somit werden erste Entwicklungsrichtungen für eine Koexistenz von cloudbasierten und nicht cloudbasierten Lösungen auf-

gezeigt, indem die Cloud als Entwicklungsplattform für das Unternehmen genutzt werden kann.

Die im Kontext von Enterprise Systems und Cloud fundamentale Fragestellung lautet aber, ob ein Enterprise System gänzlich oder in Teilen als *Software as a Service* aus dem Netz bezogen werden soll. In diesem Fall wird das ES als Software-Service von den Herstellern bezogen.[8] Es kommt zu einer veränderten Nutzungsart und einer anderen vertraglichen Basis zwischen dem Anwendungsunternehmen und dem Softwarehersteller. Aus einer strategischen Perspektive ist es wichtig zu beachten, dass die Nutzung von Enterprise Systems as a Service nur dann vom eigenen Betrieb der Software abweicht, wenn die Lösung ohne größere Individualisierungen in der Cloud betrieben wird. Nur unter dieser Rahmenbedingung kann ernsthaft von Software as a Service gesprochen werden. Sofern eine hohe Individualisierung des Enterprise Systems angestrebt wurde oder wird, ist eine private Cloud-Lösung für den gesamten Betrieb erforderlich, sodass aus einer ehemals angestrebten Software-as-a-Service Lösung eine Infrastructure-as-a-Service-Lösung wird.

Enterprise Systems als Software as a Service – Strategische Diskussion

Die Verbreitung von Software as a Service über die Cloud hat vor allem zwei Ursachen, die sowohl von den Anbietern als auch von den Anwendungsunternehmen immer vorgetragen werden: Kosteneinsparungen und Innovationsfähigkeit. Bereits die durchaus antinom wirkenden Zielsetzungen sollten Entscheidungsträger vorsichtig werden lassen. Kosteneinsparungen bei dem Betrieb oder der Erstellung von Services können – zunächst auf der Ebene desjenigen, der etwas neu erstellt oder eine bestehende Lösung betreibt – nur erreicht werden, wenn es zu einem Degressionseffekt kommt. Dieser wiederum ist nur realisierbar, wenn aus technischer Sicht betrachtet eine Standardisierung der Leistungserbringung erreicht wird. Nur für den Fall, dass mehrere Anwendungsunternehmen auf einer Maschine, einem System usw. mit einer ähnlichen Funktionalität betrieben werden können, werden Degressionseffekte realisiert. Diese anbieterseitigen Degressionseffekte bilden die Voraussetzung, dass die niedrigeren Kosten der Leistungserbringung auch den Anwendungsunternehmen zugutekommen. Ein zweiter wichtiger Aspekt ist, dass die Leistungsinanspruchnahme der Unternehmen keine statische Fragestellung, sondern ein dynamisches Problem darstellt. Handelsunternehmen haben beispielsweise eine hohe Last auf ihren Systemen zu verzeichnen, wenn die Inventur am Jahresende durchgeführt wird; eine geringere Last ist nachts vorhanden, währenddessen Handelsunternehmen in anderen Zeitzonen eine zeitinverse Leistungsanforderung haben. Nicht nur innerhalb einer Branche und über Zeitzonen verteilt, sondern auch zwischen unterschiedlichen Branchen und Leistungsanforderungsprofilen eröffnet die Zusammenführung vieler Unternehmen in der Cloud die Möglichkeit zur Effizienzsteigerung. Die Steigerung der Nachfrage nach Cloud-Leistungen aus Kostengründen setzt voraus, dass der Anbieter von Cloud-Diensten Degressionseffekte gegenüber den einzelnen Unternehmen realisieren kann. Diese wiederum dürften vor allem dann entstehen, wenn Leerkosten bei den Unternehmen angenommen werden. Diese sind immer dann gegeben, wenn die einzelnen Unternehmen ihre Rechnerkapazitäten an der Kapazitätsnachfrage ausrichten (aus Sicht der Produktionsplanung würde von der Synchronisation von Absatz und Produktion gesprochen werden, während bei einer Emanzipation von Absatz und Produktion das Problem so nicht entstehen würde). Beispielsweise ist in sehr vielen Unternehmen vormittags die Nutzung der Systeme und damit die Belastung der Rechenkapazität besonders ausgeprägt, während diese nachmittags signifikant abnehmen. Beim Cloud Computing, in diesem Fall vor allem dem Infrastructure as a Service, wird unter Zuhilfenahme von Virtualisierungs- und Verteiltechniken ein Ausgleich zwischen den Anforderungsprofilen von vielen Anwendungsunternehmen möglich, sodass Kostenreduzierungen beim Betrieb von Software entstehen.

Die Innovationsfähigkeit soll durch Cloud Services insbesondere dadurch gesteigert werden, dass die Release-Zyklen der ES-Hersteller verkürzt werden. Die Umsetzung weiterer Releases des ES-Herstellers durch das Anwendungsunternehmen setzt dabei voraus, dass diese auch genutzt werden können, das heißt die Release-Änderungen mit den Individualisierungen „verträglich" sind. Dabei handelt es sich um eine potentielle Innovation, denn die Nutzbarmachung der Innovation im Unternehmen setzt voraus, dass die allgemeine Innovationshypothese auch für den konkreten Anwendungsfall gilt.

Die Realisierung von Kosteneinsparungen und auch die Möglichkeiten zur schnellen Etablierung von Neuerungen im Unternehmen setzen voraus, dass das vom Hersteller vorgedachte Enterprise System und die konkrete Einführung in einem Anwendungskontext strukturell und funktional nur geringe Unterschiede aufweisen. An dieser Stelle knüpfen Software as a Service an; sie implizieren einen besonders ausgeprägten Standardisierungsgrad, der den bisherigen Einführungsprojekten von Enterprise Systems nicht zu eigen war. Es sei an dieser Stelle auf das Zitat von Williams und Pollock verwiesen, denn Enterprise Systems sind gerade nicht Standardsoftware. Die Lösungen in der Cloud hingegen erheben genau diesen Anspruch. Dies führt zu einer weit größeren strategischen Problematik als dies bisher bei Enterprise Systems der Fall gewesen ist. Während in der Vergangenheit die Individualisie-

rung, die Entstandardisierung, in vielen Fällen der Erfolgsfaktor war, der die Wettbewerbsvorteile trotz der Einführung von Enterprise Systems ermöglichte, würden Cloud-Angebote genau das Gegenteil bewirken: Es kommt zu einem Standardisierungszwang, der zu einer Gefahr für die Wettbewerbsfähigkeit werden kann.

Dies ist das strategische Grundproblem bei der Diskussion um Enterprise Systems in der Cloud. Die Services in der Cloud setzen, um wirtschaftlich vorteilhaft zu sein, eine Standardisierung voraus, die aktuell nicht gegeben ist. Es ist daher geradezu widersinnig, dass beispielsweise der führende Hersteller von Enterprise Systems, die SAP SE in Walldorf, sich als Protagonist von Cloud-Lösungen darstellt. Die großen Kunden der SAP haben sämtlich individualisierte Systeme im Einsatz, die sich nicht für eine Übertragung in die Cloud anbieten, es sei denn, es geht um den reinen Betrieb der Lösung. Für diesen Fall ist, wie bereits ausgeführt, die Cloud nur eine weitere Outsourcing-Alternative, keinesfalls aber strategisch so bedeutsam. Die Fragestellung, ob allerdings Software as a Service bezogen werden soll, ist eine Fundamentalfragestellung einer jeden IT-Strategie, sofern damit wettbewerbsrelevante Prozesse betroffen sind. Die Nutzung von Software-as-a-Service-Diensten für jene Funktionen oder Funktionsbereiche eines Unternehmens, in denen es vor allem um die effiziente Abwicklung geht (Reisekostenabrechnung, Personalabrechnung, Buchhaltung, etc.), ist strategisch nicht kritisch, bei den strategisch wichtigen Bereichen hingegen ist die Nutzung von Software as a Service problematisch.

Enterprise Systems in einer koexistenten Zukunft von Cloud- und On-Premise-Lösungen

Die Ausführungen haben gezeigt, dass Enterprise Systems nicht einfach als Software as a Service betrieben werden sollten, ohne dass die unterschiedlichen Prozessbereiche der Systeme auf ihre wettbewerbsstrategische Bedeutung hin untersucht werden. Aufgrund der Tatsache, dass Enterprise Systems gerade keine Standardsoftware darstellen, sondern erst durch den Akt der Individualisierung einsatzfähig werden, scheint mit dem Ansatz „Software as a Service“ eine besondere strategische Situation für Anwendungsunternehmen zu existieren. *Erstens* werden die Anbieter ihre Leistungsangebote nachahmen, ein häufig beobachtetes Marktverhalten der oligopolistischen Technologieanbieter. Es wird somit zu immer größeren Cloud-Lösungsangeboten kommen. *Zweitens* wird in Zeiten zunehmender Digitalisierung der Zusammenhang zwischen Wettbewerbsvorteil und informationstechnologischen Gegebenheiten immer enger werden. *Drittens* bedarf es konzeptioneller Ansätze und Informationssystem-Architekturen, welche die Integration zwischen Cloud Services und On-Premise-Lösungen sicherstellen.

Die aktuelle Situation ist strategisch durch ein hohes Maß an Unsicherheit geprägt. Zunächst ist es wichtig, dass im Rahmen des IT-Strategie-Managements die aktuellen Standardisierungszwänge von Enterprise Systems erkannt werden. Ob diese Standardisierung in einem konkreten Anwendungsfall durchaus sinnvoll ist, kann hier nicht allgemeingültig beantwortet werden. Sofern die These vertreten wird, dass die Individualisierung die Voraussetzung für den erfolgreichen Einsatz eines Enterprise Systems in einem Unternehmen gewesen ist, so wäre es gefährlich, ohne eine strategische Analyse Software-as-a-Service-Lösungen einzusetzen.

Die Situation in den Unternehmen ist auch nicht so dichotom wie dargestellt, denn in vielen Bereichen lassen sich heute Services miteinander verbinden. Die diskutierten Probleme treten in jenen Bereichen auf, in denen die Standardisierbarkeit so nicht gegeben ist. In diesen Fällen ist auch eine Nutzbarkeit von Cloud Services aktuell noch nicht vorhanden. Es bedarf konzeptioneller und technischer Lösungen, um eine Koexistenz von Cloud-basierten und individualisierten Komponenten von Enterprise Systems zu ermöglichen. Dabei gilt es im Rahmen des strategischen IT-Managements die skizzierte Problemstellung weiter zu entfalten und für die betriebliche Praxis handhabbar zu machen, eine Integration der Deployment-Formen auf einer konzeptionellen Ebene, beispielsweise in Enterprise Architectures vorzunehmen sowie die technische Integration sicherzustellen,[9] damit die Interoperabilität von Services möglich wird.

Summary

Enterprise Systems are the core application systems of today’s companies. Over the last decades, their scope has broadened more and more, today encompassing more business functions than ever. They are developed by producers of standard software and are adopted by companies in extensive implementation projects. At the same time, by using cloud computing various forms of hardware or software services can be provided. Considering today’s practice of making standard software systems available through individualization, a strategic perspective is taken in order to identify potential applications that cloud computing may offer for Enterprise Systems.

Anmerkungen

1) Zur Digitalisierung und deren Konsequenzen, diskutiert am Beispiel von Handelsunternehmen, vgl. Schütte, Vetter, 2017.
2) Vgl. Olson, Kesharwani, 2010, S. 35.
3) Ein MRP-System ist ein Materials-Resource-Planning-System, ein MRP II-System wird als Manufacturing Resource Planning und ein ERP als Enterprise Resource Planning System verstanden. Die Ausdehnung

des Ressourcenbegriffs vom Material über die Fertigung bis hin zum Unternehmen charakterisiert die Ausdehnung der Ressourcenperspektive bis hin zum ganzen Unternehmen. Aktuell wird mit dem Begriff Enterprise System ein noch ganzheitlicher Ansatz vertreten, da eine Reduzierung der Systeme auf Ressourcen sprachlich vermieden werden sollte.
4) Pollock, Williams, 2009, S. 8
5) Die Problematik der Individualisierung von Standard-Software und der dahinterliegenden idealisierenden Organisationsmodelle wird von Moormann thematisiert, von dem auch der Begriff der Entstandardisierung übernommen wurde, Moormann, 2016, S. 16.
6) Vgl. hierzu u.a. die Ausführungen von Mijač Picek, Stapič 2013; Zhao, Kirche, 2013. Vgl. auch Olson, Kesharwani, 2010, S. 56.
7) Diese Unterscheidung soll hier für die weitere Diskussion angenommen werden, auch wenn dies eine Vereinfachung gegenüber realen Lösungsangeboten darstellt. Vgl. exemplarisch für die Vielzahl an Autoren, die diese Einteilung von Cloud Services verwenden, Hentschel, Ley, 2016; Schubert, Adisa, 2011.
8) Es wird hier davon ausgegangen, dass es sich um eine Public-Cloud-Lösung handelt.
9) Unter Enterprise Architectures werden in der Literatur solche verstanden, die die Strukturen und Elemente der beiden unterschiedlichen Systemarten (Anwendungssystem und Organisationssystem) nach unterschiedlichen Methoden holistisch abbilden.

Literatur

– Hentschel, R.; Ley, C. (2016): Cloud Computing: Gestern, heute, morgen. In HMD, 53 (2016), 563–579.
– Mijac, M.; Picek, R.; Stapic, Z. (2013): Cloud ERP System Customization Challenges. In Proceedings of the Central European Conference on Information and Intelligent Systems (CECIIS), 132–140. (http://www.ceciis.foi.hr/app/public/conferences/1/papers2013/637.pdf, 10.1.017)
– Olson, D.L.; Kesharwani, S. (2010): Enterprise Information Systems. Contemporary Trends and Issues. World Scientific, New Jersey et al. 2010.
– Pollock, N.; Williams, R. (2009): Software and Organisations, The Biography of the Enterprise-Wide System or How SAP Conquered the World, Routledge, London, New York.
– Moormann, H. (2016): Das Projekt SAP. Zur Organisationssoziologie betriebswirtschaftlicher Standardsoftware, transcript Verlag, Bielefeld 2016.
– Schubert, P.; Adisa, F. (2011): Cloud Computing for Standard ERP Systems: Reference Framework and Research Agenda, Arbeitsbericht aus dem Fachbereich Informatik der Universität Koblenz, Nr. 16/2011.
– Schütte, R.; Vetter, T. (2016): Analyse des Digitalisierungspotenzials von Handelsunternehmen. In: R. Gläß, B. Leukert (Hrsg): Handel 4.0, Strategien, Technologien, Transformation, Springer, Heidelberg 2016, 75–113.
– Zhao, F.; Kirche, E. (2013): Continuing On-Premise or Adopt On-Demand? An Empirical Study of ERP-Adoptions in SMEs. In: Kurosu M. (eds), Human-Computer Interaction, Users and Contexts of Use, HCI 2013, Lecture Notes in Computer Science, Vol 8006, Springer, Berlin, Heidelberg, 492–500.

Der Autor

Reinhard Schütte, geboren 1967, studierte Betriebswirtschaftslehre an der Westfälischen Wilhelms-Universität in Münster. Er war Wissenschaftlicher Mitarbeiter am Institut für Wirtschaftsinformatik der Universität in Münster und promovierte dort in Wirtschaftsinformatik zum Thema „Grundsätze ordnungsmäßiger Referenzmodellierung“ mit Auszeichnung. Er wechselte als Wissenschaftlicher Assistent an das Institut für Produktion und Industrielles Informationsmanagement von Prof. Dr. Stephan Zelewski an die Universität Essen. Hier hat er sich intensiv mit Fragen der Informationsmodellierung auseinandergesetzt, vor allem Fragen zur Güte der Qualität von Modellen und grundsätzlichen Aspekten der Konstruktion und Anwendung von allgemeingültigen Modellen. Nach seiner Zeit an der Universität Essen vertrat er den Lehrstuhl für Industrielles Informationsmanagement an der Universität Koblenz und habilitierte sich an der Westfälischen Wilhelms-Universität in Münster, die ihm die venia legendi für Wirtschaftsinformatik verlieh. Von 2013 bis 2015 war er an der Zeppelin Universität in Friedrichshafen als Gastprofessor für Informationsmanagement und Handelsinformationssysteme aktiv, bevor er den Ruf auf die Professur für Wirtschaftsinformatik und integrierte Informationssysteme an der Universität Duisburg-Essen annahm.

Die technischen Daten und das „Drumherum“ des Wissenschaftlichen Rechnens mit der magnitUDE thematisiert dieser Text des Supportteams des High-Performance Computings an der UDE.

Ein Großgerät für die Forschung

Hochleistungsrechnen auf der magnitUDE

Von Dominik Brands, Holger Gollan & Olaf Hasemann

Vorgeschichte und Antrag

Im Sommer 2010 wurde an der Universität Duisburg-Essen der Hochleistungsrechner Cray XT6m installiert, der mit einer realisierten Rechenleistung von mehr als 26 TFlop/s (26 Billionen Berechnungen pro Sekunde) damals auf Platz 447 in der Liste der 500 schnellsten Rechner der Welt (www.top500.org) stand. Gleichzeitig wurde an der UDE das Center for Computational Sciences and Simulation (CCSS) gegründet, in dem die Aktivitäten im Bereich des Wissenschaftlichen Rechnens gebündelt werden sollten. Im CCSS sind Arbeitsgruppen aus den Fakultäten Biologie, Chemie, Ingenieurwissenschaften, Mathematik, Medizin, Physik und Wirtschaftswissenschaften aktiv, was die breite Verankerung des Wissenschaftlichen Rechnens in vielen Bereichen der Forschung dokumentiert.

Wurde die Cray XT6m zunächst vor allem von Chemiker*innen, Physiker*innen und Mathematiker*innen genutzt, so zeigte sich in der Folgezeit – insbesondere auch durch Neuberufungen – ein immer stärker werdender Bedarf an Rechenzeit, vor allem für die Ingenieurwissenschaften. Zusammen mit dem Alter der Cray XT6m führte diese Situation zu Überlegungen für einen neuen und leistungsstärkeren Hochleistungsrechner als Nachfolger der Maschine aus dem Jahr 2010. Die Anforderungen der beteiligten beziehungsweise interessierten Wissenschaftler *innen aus dem CCSS wurden gesammelt und über das Rektorat wurde im Februar 2014 ein Großgeräteantrag bei der Deutschen Forschungsgemeinschaft (DFG) nach Art. 91 b GG gestellt. Dieses

Dominik Brands. Foto: Vladimir Unkovic

Verfahren ermöglicht die Beschaffung von Großgeräten für die Forschung bis zu einem Gesamtvolumen von fünf Millionen Euro. Dabei muss bei einem erfolgreichen Antrag die Universität nur zehn Prozent der Finanzierung tragen, 40 Prozent werden vom Land NRW, 50 Prozent vom Bund über die DFG als Förderung bewilligt. Federführender Antragsteller war Prof. Andreas Kempf, der zusammen mit Prof. Jörg Schröder auch den Anteil der UDE an der Finanzierung beisteuerte. Wissenschaftlich ergänzt und begründet wurde der Antrag durch die Beteiligung weiterer Wissenschaftler*innen aus dem CCSS, so dass schließlich insgesamt zwölfArbeitsgruppen ihre Bedarfe angemeldet und damit die Anforderungen an den neuen Hochleistungsrechner definiert hatten. Der Großgeräteantrag an die DFG hatte letztlich ein Volumen von 3,3 Millionen Euro und wurde im März 2015 bewilligt.

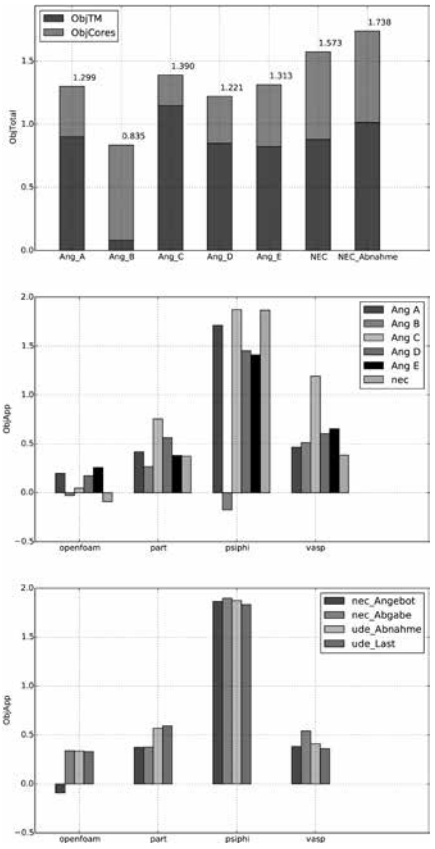
Da heutzutage die Leistung einzelner Prozessoren nicht mehr kontinuierlich ansteigt (Moores Law: Verdoppelung der Rechenleistung alle 18 Monate) gehen die Hersteller schon seit Jahren dazu über, immer mehr Rechenkerne in einem Rechenknoten (Server) zusammenzufassen. Das gleiche Prinzip wird auch im Consumerbereich angewendet; so verfügen moderne Smartphones über vier bis acht Rechenkerne. Ein typisches High-Performance Computing-System (HPC-System) besteht aus einer größeren Anzahl von Rechenknoten, die über ein schnelles Netzwerk verbunden sind und auf ein gemeinsames Speichersystem zugreifen.

Zur Unterstützung und zur besseren Vernetzung mit den Anwendern wurde vom Zentrum für Informations- und Mediendienste (ZIM) und dem CCSS das Support-Team für das Wissenschaftliche Rechnen, kurz HPC-Support-Team, an der UDE gegründet. Der Schwerpunkt seiner Arbeit liegt in der Nutzerbetreuung auf den HPC-Systemen der Universität Duisburg-Essen. Neben einer verbesserten Anwenderunterstützung soll durch diese zentralen Ansprechpartner der direkte und indirekte Wissenstransfer zwischen den Anwender*innen intensiviert werden. Darüber hinaus besteht ein Beratungsangebot bei zukünftigen Beschaffungen im Bereich des wissenschaftlichen Rechnens. Das Team wurde im Wintersemester 2015/16 initial mit zwei Mitarbeitern aus CCSS-Arbeitsgruppen im Rahmen von Teilumsetzungen besetzt. Dieses Konzept verfolgt unter anderem das Ziel, dass die Mitglieder des Support-Teams weiterhin eine aktive Rolle in der Wissenschaft einnehmen und so aktuelle Entwicklungen „hautnah erleben".

Auswahl und Beschaffung

Aufgrund von Beschaffungsrichtlinien wurde die Ausschreibung für das neue HPC-System europaweit durchgeführt. Dabei wurde das zu beschaffende System im Wesentlichen durch drei Aspekte charakterisiert: maximaler Anschaffungspreis, grundsätzliche Ausschlusskriterien und Berechnungsbenchmarks. Der erste Punkt ergab sich trivialerweise durch die Fördersumme von 3,3 Millionen Euro, die in der Bewilligung des Großgeräteantrags zur Verfügung gestellt wurde.

Wesentlich schwieriger gestaltete sich die Gestaltung der übrigen Eckdaten. Bei den Ausschlusskriterien spielten zum einen grundsätzliche Ausstattungsmerkmale des Systems eine Rolle (z.B. Speicher- und Softwareausstattung sowie Dokumentationspflichten), die aufgrund von Erfahrungswerten und Anwenderhinweisen aufgestellt wurden. Darüber hinaus mussten aber auch Kriterien zusammen gestellt werden, die sich direkt (z.B. Stellfläche, max. Stromaufnahme) oder indirekt (z.B. Art der Kühlung) aus dem zukünftigen Standort des Systems ergeben sollten. Jedoch stand zum Zeitpunkt der Ausschreibungsplanung der finale Standort nicht fest, da bis dato



(1) Vergleich der ausgewerteten Zielfunktion für die angebotenen Systeme insgesamt (oben) und bezogen auf die Anwendungen aus der Testmatrix (Mitte) sowie bezogen auf das ausgelieferte System (unten).
Quelle: Olaf Hasemann

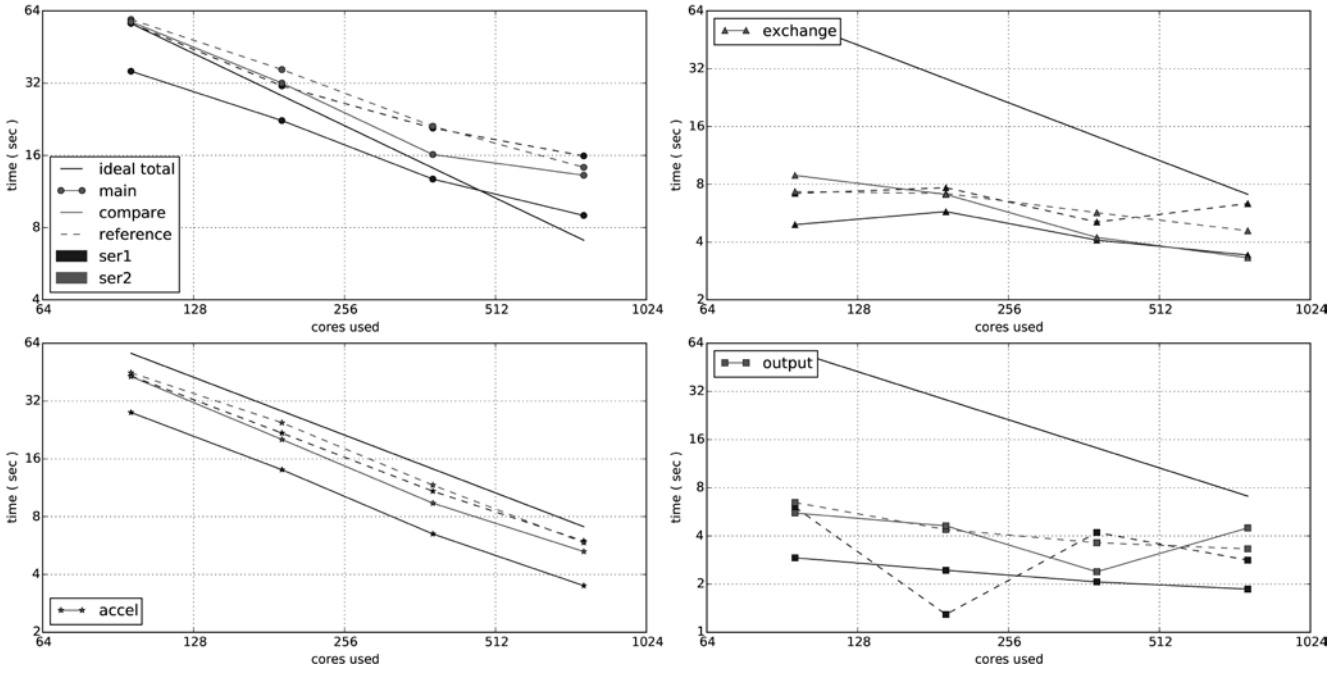
kein geeigneter Standort an den beiden Campi in Duisburg und Essen gefunden werden konnte. Bereits die Ertüchtigung der Rechnerräume für das bestehende HPC-System Cray XT6m war mit aufwändigen Baumaßnahmen ermöglicht worden. So mussten diese Kriterien mittels Schätzungen und Herstellerinformationen mühsam gesammelt werden. Jedoch zeichnete sich zeitgleich zum Beginn der Ausschreibung ab, dass ein geeigneter externer Standort in Aussicht stehe. Die DU-IT, eine Tochtergesellschaft der Duisburger Stadtwerke, plante die Errichtung eines eigenen, neuen Rechenzentrums im Logport I in Duisburg-Rheinhausen und hatte in dieser, als Einhousing-Lösung geplanten Baumaßnahme, noch Raumkapazitäten frei. Dabei konnten im Synergieeffekt noch weitere,

dringend benötigte Rechenzentrumsflächen für IT-Infrastruktur und Server des ZIM der UDE angemietet werden, so dass eine Gesamtfläche von etwa 215 Quadratmetern für das neue DataCenter SZ zur Verfügung stand.

Mit diesen Kriterien wurden nun die Rahmenbedingungen der Ausschreibung definiert. Den eigentlichen Kern bildeten die Anwendungsbenchmarks, die sogenannte Testmatrix, die die Bewertung der Leistungsfähigkeit des angebotenen Systems ermöglichen und damit eine effiziente Nutzung der Fördermittel sicherstellen sollten. Die Testmatrix umfasste einen Mix aus vier Anwendungen: PsiPhi (ein InHouse-Code aus dem Bereich der Fluiddynamik), OpenFOAM (ein Open Source-Code mit konträren Programmier-Paradigmen, ebenfalls aus dem Bereich der Fluiddynamik), VASP5 (ein semi-kommerzielles Softwarepaket aus der Physik) und Particles (ein Beispiel für ein weiteres Parallelisierungsparadigma aus der Lehre), die einem ‚kleinen' Strong Scaling Test unterzogen werden sollten. Zielsetzung der Testmatrix ist, die benötigte Rechenzeit für typische Anwendungen der Hauptnutzergruppen zu minimieren. Die Testmatrix umfasste Problemstellungen für maximal 1.024 Rechenkerne, sowie einen optionalen Fall für etwa 2.000 Kerne. Im Vergleich zur zu erwartenden Größe (ca. 10.000 Rechenkerne) der angebotenen Systeme wurden bewusst „kleine Fälle" gewählt, um den Anbietern die Durchführung auf kleineren Testinstallationen zu ermöglichen und somit den Kreis der potentiellen Anbieter nicht auf die wenigen „großen" einzuschränken. Gleichzeitig mussten die Testfälle so konstruiert werden, dass mögliche Skalierungseinbrüche auf den angebotenen Systemen bereits bei dieser kleinen Problemstellung sichtbar werden konnten. Die Anbieter mussten bei Angebotseinreichung die Ergebnisse der von ihnen durchgeführten Testmatrix mitliefern. Diese Daten wurden seitens des HPC-Support-Teams ausgewertet und ein Ranking der vorliegenden Angebote erstellt, siehe Abbildung (1). Zur Durchführung des Rankings wurde eine Zielfunktion definiert, mit der alle Ergebnisse der Testmatrix in einer Kennzahl zusammengefasst wurden, die es zu maximieren galt. In der Definition dieser Zielfunktion ist die Beschreibung und Bewertung der zum Teil konträren Anforderungen an das HPC-System seitens der primären Anwendungen enthalten. Neben der Auswahl der geeigneten Hardware für die Aspekte Berechnung, Speicher (RAM u. File IO) sowie Interconnect (Vernetzung der Rechenknoten), ist die Auswahl eines geeigneten Softwarestacks und das Know-how des Anbieters bezüglich wissenschaftlicher HPC-Anwendungen entscheidend für einen guten Wert der Zielfunktion. Auf technische Vorgaben wie zum Beispiel die Art des Interconnects oder welches parallele Speichersystem eingesetzt werden sollte, wurde vollständig verzichtet – den Anbietern wurde diesbezüglich freie Hand gelassen.

Der entscheidende Begriff im High-Performance-Computing ist die Skalierbarkeit. Bei zunehmend steigenden Problemgrößen, die in den Wissenschaften behandelt werden und dem gleichzeitigen Tod von Moore's Law, verbleibt beim

(2) Strong Scaling Test der Anwendung Particles aus der Testmatrix, durchgeführt auf dem ausgelieferten System. Die Serien ser1 und ser2 sind charakterisiert durch unterschiedliche Parallelisierungskonzepte. Quelle: Olaf Hasemann



aktuellen Stand der Technik als Lösung lediglich das massiv parallele Rechnen. Die Skalierbarkeit sagt dabei, wie gut eine Anwendung auf einem HPC-System von der Hinzunahme weiterer Rechenressourcen profitiert, ein idealer Fall würde bei gleicher Problemstellung und bei einer Verdoppelung der genutzten Kerne die Hälfte der Rechenzeit benötigen. Abbildung (2) zeigt ein Beispiel eines solchen Strong Scaling Tests für die Anwendung Particles aus der Testmatrix und verdeutlicht, dass in diesem Fall der Interconnet (exchange) und das Filesystem (output) die für die Skalierbarkeit limitierenden Faktoren sind.

Ein gut skalierendes System zur Verfügung zu stellen, war bei der Definition der Testmatrix ein wesentlicher Aspekt, nicht zuletzt um den Wissenschaftler*innen der UDE mit ihren Anwendungen den Sprung auf die größten europäischen Höchstleistungsrechner wie SuperMUC oder JUQUEEN zu ermöglichen.

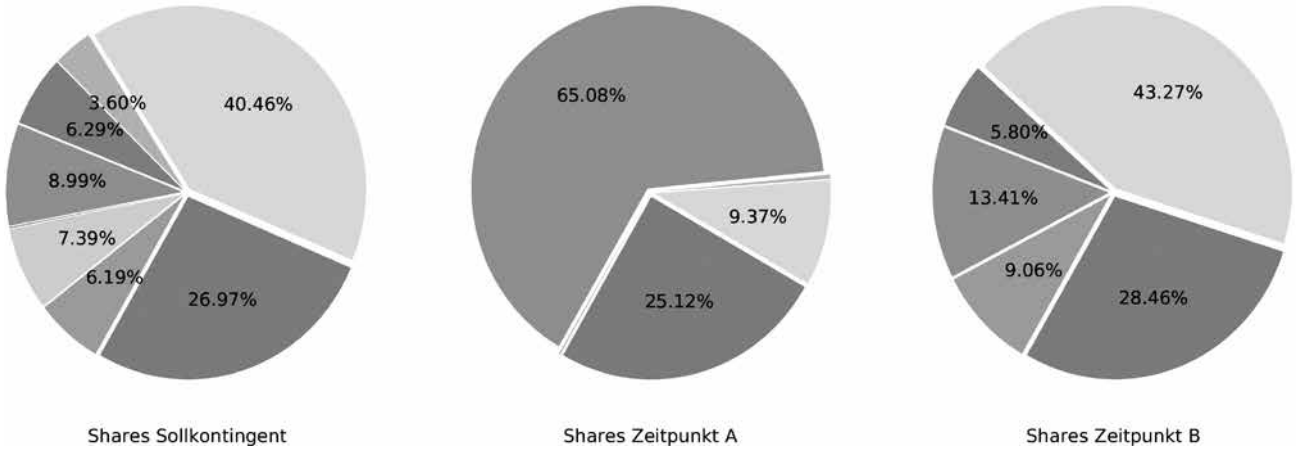
Am Ende der Ausschreibung war die Firma NEC Deutschland GmbH erfolgreich und bestellt wurde ein System mit 564 Rechenknoten, die insgesamt 13.536 Rechenkerne enthielten. Knapp 1/8 der Knoten (72) wurden mit 128 GB Hauptspeicher bestückt, die restlichen Knoten erhielten jeweils 64 GB, so dass das System insgesamt über 40.704 GB Hauptspeicher verfügt. Die Rechenknoten sind mit Intel OmniPath Interconnect als schnellem Netzwerk vernetzt; an das System sind 480 TB Festplattenplatz über ein paralleles Speichersystem angebunden, so dass viele Knoten gleichzeitig auf den gemeinsamen Speicher zugreifen können. Das System wird indirekt durch wassergekühlte Rücktüren klimatisiert und hat eine maximale Leistungsaufnahme von 200 kW Strom. Es ist in zwei Reihen à fünf Serverschränke untergebracht.

Administration und Betrieb

Im Februar 2016 begann die Auslieferung und der Aufbau der magnitUDE im neu gebauten Rechenzentrum am Logport und Anfang Mai wurde die Betriebsbereitschaft seitens des Herstellers erklärt. In der folgenden Phase des Stabilitätstests wurden zum einen die HPL-Benchmarks für die Einträge in den TOP-500 Listen durchgeführt, zum anderen wurden die Benchmarks der Testmatrix unter den Betriebsbedingungen „Leeres System“ sowie „80 Prozent Last“, wie in der EU-Ausschreibung gefordert durch das HPC-Support-Team durchgeführt und ausgewertet. Bei den TOP 500-Läufen wurde eine Rechenleistung von über 437 TFlop/s realisiert, was mehr als 90 Prozent der theoretisch maximal erreichbaren Leistung von rund 476 TFlop/s entspricht und womit das System im Juni 2016 Platz 279 in der TOP 500-Liste belegte. In dieser Phase wurden auch Skalierungstests über das gesamte HPC-System durchgeführt, für die zwei Anwendungen aus der Testmatrix und eine zusätzliche Anwendung aus dem Bereich der massiv parallelen Algorithmen für Mehrskalen-Simulationen verwendet wurden. Die Ergebnisse zeigten, dass das ausgelieferte System besser als zuvor angenommen skaliert. Dies wurde mit großer Erleichterung aufgenommen, da für die Hauptanwendergruppen die Skalierbarkeit über das gesamte HPC-System eine wesentliche gewünschte Eigenschaft darstellt und diese Eigenschaft in der Testmatrix mit ihren „kleinen“ Testfällen nur indirekt über die Definition der Zielfunktion abgebildet werden konnte.

Der Nachweis der Eignung eines HPC-Systems für die geplanten Anwendungen auf einem wohldefinierten, zumeist exklusiv zur Verfügung stehenden System, ist die eine Sache; den reibungslosen Betrieb für eine große Anzahl von Nutzer*innen aus unterschiedlichen Arbeitsgruppen mit ihren jeweiligen speziellen Anforderungen zur Zufriedenheit aller zu gewährleisten, ist eine vollständig andere Aufgabe. Auf der Basis der mit dem Betrieb der CRAY XT6m gesammelten Erfahrungen begann das HPC-Support-Team zusammen mit dem CCSS noch vor Beginn der EU-Ausschreibung mit der Ausarbeitung der Nutzerrichtlinien für das neue HPC-System. Durch diese Richtlinien sollte gewährleistet werden, dass zum einen die Vorgaben der Fördergeber (Land, Bund, DFG) eingehalten werden, zum Beispiel Verbot zur Nutzung bei Auftragsforschung, als

(3) Die magnitUDE in ihrem Zuhause im neuen DataCenter SZ.
Quelle: Tanja Behrendt



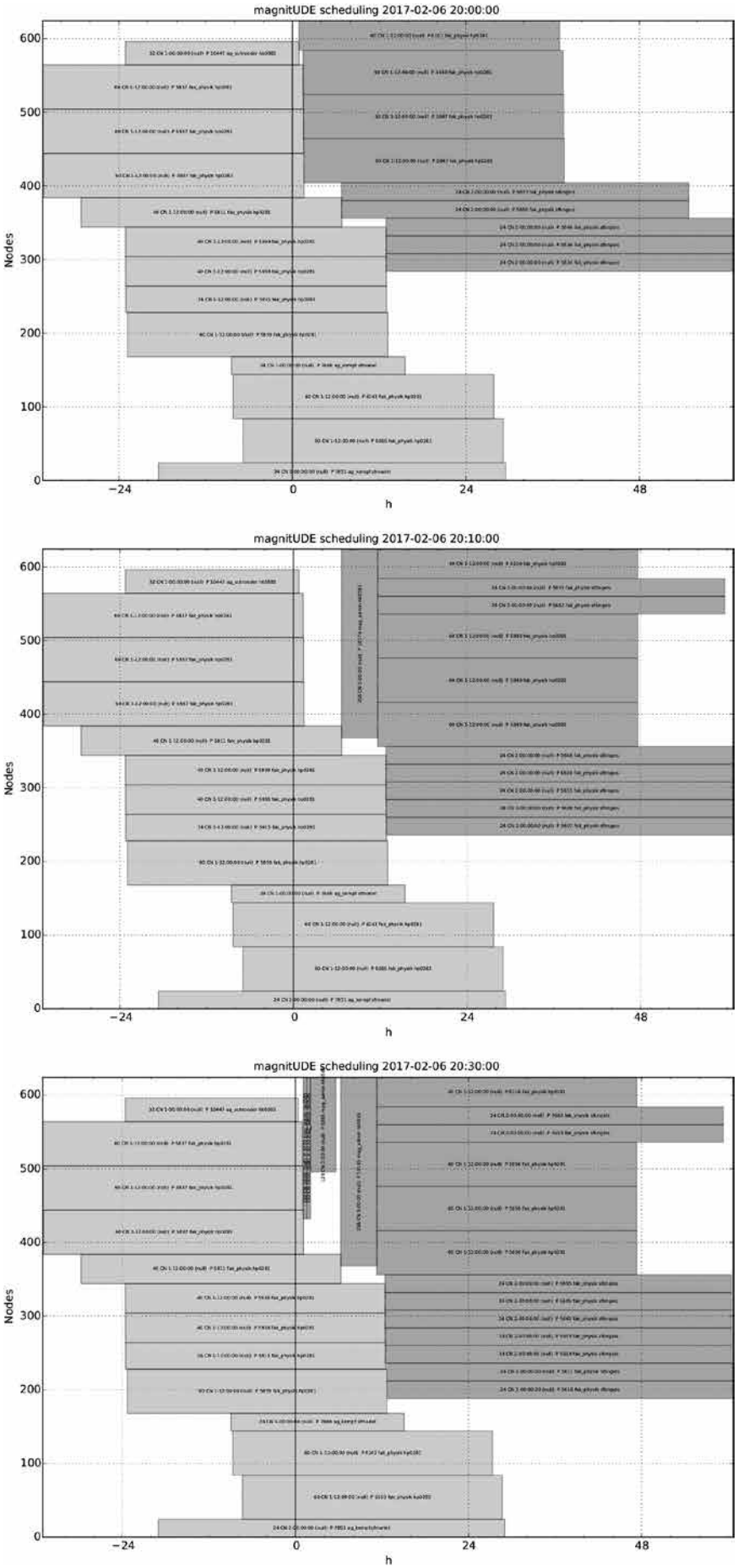
(4) Fairshare-Kontingente: Soll-Werte (links) und beispielhafte Verteilungen im Betrieb (mitte und rechts).
Quelle: Olaf Hasemann

auch ein langfristig geregelter und stabiler Betrieb sichergestellt wird. Die Abstimmungen und Diskussionen mit den Koordinatoren der Arbeitsgruppen konnten kurz vor Inbetriebnahme abgeschlossen werden, so dass die technischen Vorgaben der Nutzerrichtlinien durch die Konfiguration des eingesetzten Queueing Systems SLURM umgesetzt werden konnten.

Scheduling

In diesem Zusammenhang ist die Entscheidung bezüglich der eingesetzten Scheduling-Policy (Jobplanungs-Strategie) sicherlich die für die Anwender*innen folgenreichste. Sie bestimmt, wann eine Rechnung eines Anwenders gestartet wird beziehungsweise wie lang die Wartezeit bis zum Start ist. Gewählt wurde eine Tree-Fairshare Strategie in Kombination mit Backfilling. Fairshare geht dabei davon aus, dass es kein festes Kontingent an Ressourcen pro Zeiteinheit gibt, üblicherweise Kernstunden/Monat, nach deren Verbrauch ein*e Anwender*in keine weitere Rechnung mehr durchführen kann, sondern lässt die von einem*einer Anwender*in verbrauchten Ressourcen in die Berechnung eines Priorisierungsfaktors einfließen, der dann über den Startzeitpunkt einer Rechnung entscheidet. Oft werden dabei Anwender*innen, zum Beispiel entsprechend ihrer Arbeitsgruppenzugehörigkeit, zu sogenannten Kontingentgruppen zusammengefasst. Das Fairshare-Prinzip ermöglicht den Anwender*innen kurzfristig weitaus mehr Ressourcen zu benutzen, als es ihnen entsprechend ihrer Kontingentzuordnung zusteht. Erst wenn weitere Anwender*innen aus anderen Kontingentgruppen mit bis dahin geringerem Ressourcenverbrauch ebenfalls Rechenzeit anfordern, werden diese bevorzugt. Der für die Berechnung der Priorisierung herangezogene Ressourcenverbrauch wird einem Halbwertverfahren unterzogen, dessen Länge auf die typische Laufzeit der Rechnungen abgestimmt wird. Bei der Konfiguration auf der magnitUDE wurde als Halbwertzeit 14 Tage bei einer maximalen Laufzeit pro Rechnung von vier Tagen festgelegt. Dieses bedeutet, dass der Einfluss eines Ressourcenverbrauchs auf die Prioritätsberechnung nach 14 Tagen halbiert wird und nach etwa zwei Monaten kaum noch spürbar ist. Dies ermöglicht eine flexible Handhabung des in den Kontingentgruppen anfallenden, stark schwankenden Rechenzeitbedarfs, der zum Beispiel kurz vor einem Abgabetermin für einen wissenschaftlichen Beitrag entsteht. Abbildung (4) zeigt eine solche Situation, bei der die Mitglieder einer Kontingentgruppe kurzfristig ein Vielfaches ihres Kontingentes verbraucht haben, eine andere Gruppe ist in etwa im ‚Soll' und eine weitere Gruppe hat deutlich weniger als langfristig benötigt verbraucht. Der Zusatz „Tree" zu „Fairshare" versucht, unerwartete „Ungerechtigkeiten" bei der Berechnung der Faktoren zu vermeiden, die entstehen, wenn die Kontingentgruppen eine stark unterschiedliche Anzahl von Nutzer*innen aufweisen.

Fairshare basiert lediglich auf dem vergangenen Ressourcenverbrauch, zum Beispiel in Kernstunden, und sagt noch nichts über die Handhabung von Typen von Rechnungen, insbesondere bei Ressourcenknappheit, aus. Ein Beispiel einer einfachen Strategie hierfür ist, dass freiwerdende Ressourcen unmittelbar mit Rechnungen einer möglichst passenden Größe unabhängig von der Laufzeit belegt werden. Diese Strategie führt zur Minimierung des Leerstandes, hat jedoch den Nachteil, dass massiv parallele und gut skalierende Anwendungen mit kurzen Laufzeiten „verhungern", da die große Anzahl an benötigten Ressourcen nie gleichzeitig verfügbar ist. Gerade diesen Typ von Rechnungen durchführen zu können, also kurze Laufzeiten bei massiv parallelem Ressourceneinsatz, war jedoch ein



(5) Zeitliche Abfolge (von oben nach unten) eines Backfilling-Vorgangs.
Quelle: Olaf Hasemann

erklärtes Ziel bei der Beschaffung des HPC-Systems. Hier fiel die Entscheidung auf den Einsatz von Backfilling, welches dem umgekehrten Prinzip folgt: je besser eine Rechnung skaliert, um so höher wird sie priorisiert. Also führt die Anforderung von mehr Ressourcen bei kürzerer Laufzeit zu einem höheren Priorisierungsfaktor. Die Rechnung mit dem höchsten Faktor wird allgemein auch als „Top-dog" bezeichnet – davon ausgehend wird das weitere Scheduling geplant. Abbildung (5) zeigt einen solchen Scheduling-Vorgang, wobei die aktuell laufenden Rechnungen links der Nulllinie und die auf ihre Ausführung wartenden rechts der Nulllinie beginnen. Die Ausgangssituation ist vergleichsweise homogen: das HPC-System ist mit Rechnungen belegt, die bei einer Laufzeit von ein bis zwei Tagen 24 bis 60 Rechenknoten angefordert haben, was als mittelgut skalierbar eingestuft werden kann. Nun kommt ein anderer Rechnungstyp hinzu, der bei 256 Rechenknoten und einer Laufzeit von fünf Stunden als Top-dog eingestuft wird. Als solcher wird er an die frühestmögliche Position gestellt, an der genügend Ressourcen zur Verfügung stehen. Alle übrigen Rechnungen werden um ihn herum geplant, in diesem Fall hinter ihn gestellt, da keiner von ihnen bis zum Beginn des Top-dogs beendet sein kann. Diese Scheduling-Policy erzeugt naturgemäß größere Lücken, da das System für den Top-dog leer gefahren wird. Damit der dadurch entstehende Leerstand nicht zu groß wird, können die Lücken mit kleineren/kürzeren Rechnungen gefüllt werden. Diese werden vor dem Top-dog gestartet, sofern sie bis zu dessen geplanter Startzeit beendet sein können. Diese Vorgehensweise stellt das eigentliche Backfilling dar. Das Beispiel zeigt weiterhin, dass eine solche Scheduling-Policy in hohem Maße von der Mitwirkung der

Anwender abhängt – sie müssen zumindest einen Teil ihrer Rechnungen so parametrieren, dass sie vom Scheduler zum Backfilling verwendet werden können. Der Scheduler bietet hierfür Hilfen an, indem zum Beispiel eine gewünschte Laufzeitspanne angegeben werden kann. Dies lässt beispielsweise folgende Interpretation zu: Für eine Rechnung liegt nach einer Minimallaufzeit von sechs Stunden ein erstes Ergebnis vor, wenn möglich, soll die Rechnung aber auch länger durchgeführt werden.

*Nutzer*innenrichtlinien*

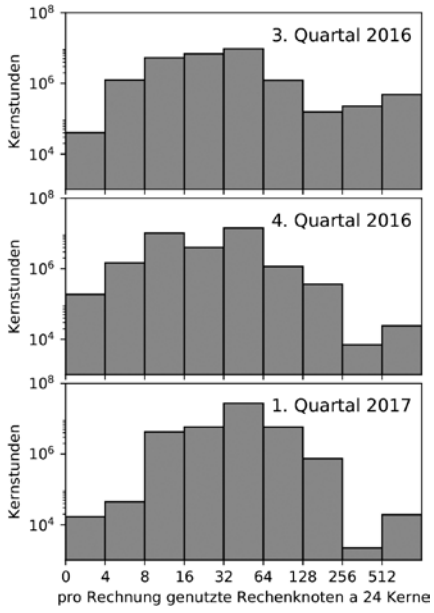
Das Hauptaugenmerk der oben genannten Nutzer*innenrichtlinien zielt auf einen geregelten, fairen und langfristig stabilen Betrieb der magnitUDE. Neben den technischen Anforderungen an das Queueing-System spielen dabei zwei weitere Aspekte eine maßgebliche Rolle: die Art der Nutzer*innenzugänge inklusive deren Beantragung und die Verwendung des lokalen Festplattenspeichers, im Folgendem „SCRATCH“ genannt. Der Grundgedanke ist es, den Zugang zur magnitUDE allen Mitgliedern der UDE zu ermöglichen, soweit sie über einen begründeten Bedarf an HPC-Ressourcen verfügen. Bei den Zugängen wird zwischen einem normalen Zugang, einem Testzugang, einem Zugang für Studierende und Zugang für Externe im Rahmen einer wissenschaftlichen Kooperation unterschieden. Das wesentliche Unterscheidungsmerkmal liegt in der Gültigkeitsdauer der Zugänge, die bei sechs Wochen (Testzugang) bis zu einem Jahr (normaler Zugang) liegen. Durch die begrenzte Gültigkeitsdauer wird dem „Verwaisen“ von Zugängen aufgrund der hohen Personalfluktuation im Wissenschaftsbereich vorgebeugt. Die entsprechenden Anträge werden durch das HPC-Support-Team verwaltet und an die Koordinatoren der betreffenden Kontingentgruppe zur Begutachtung weitergeleitet. Gleichzeitig gibt das HPC-Support-Team eine Empfehlung hinsichtlich der Eignung der geplanten Anwendung ab. Nach einer positiven Begutachtung führt das Support-Team die notwendigen Schritte durch: Freischaltung in der ZIM-Benutzerverwaltung und Einrichtung der lokalen Profile (SLURM, Disk-Quota, SCRATCH-Verzeichnis).

Ein weiterer wichtiger Baustein für einen reibungslosen Betrieb ist die ausreichende Verfügbarkeit lokaler Speicherressourcen, denn gerade bei massiv parallelen Rechnungen entstehen große Datenmengen; teils temporär (z.B. Restart-Daten) und teils dauerhaft (z.B. für das Post-Processing). Hierzu steht den Nutzern das SCRATCH zur Verfügung, welches auf den Rechenknoten hochverfügbar und massiv parallel eingebunden ist und aktuell über ein Gesamtspeichervolumen von etwa 480 TeraByte verfügt. Um dessen Speicherplatznutzung effizient zu gestalten, wurden neben benutzerspezifischen Festplattenkontingenten (Disk-Quota) verschiedene zeitliche Löschfristen definiert, die das Löschen von Daten bezüglich ihres Erstellungszeitpunkts erlauben. Dadurch wird dem Fehlgebrauch der lokalen Festplattenspeicher, die durch ihre massiv parallele Anbindung sehr exklusiv sind, als Langzeitdatenspeicher vorgebeugt. Gleichzeitig wird das System nach einer gewissen Zeit von Daten abgelaufener Nutzer*innenzugänge bereinigt. Diese zwar aus erster Sicht sehr restriktive Vorgehensweise beruht auf den Erfahrungen, welche beim Betrieb des lokalen Speichersystem der Cray XT6M gemacht worden sind. Auf dessen Speichersystem sind keine Disk-Quota und (lange Zeit) keine Löschfristen definiert worden und gleichzeitig war eine ständige Speicherauslastung über 90 Prozent zu beobachten.

Nutzung

Inzwischen ist die magnitUDE mehr als ein halbes Jahr im normalen Anwenderbetrieb und seit Anfang 2017 um etwa zehn Prozent vergrößert worden. Aktuell ist der Rechner voll ausgelastet, das heißt es steht immer eine größere Anzahl an Rechnungen in der Warteschlange. Dieser Zustand hat sich jedoch erst vergleichsweise spät (nach ca. 5 Monaten) eingestellt; ein Grund hierfür ist sicherlich darin zu sehen, dass die Hauptanwender*innengruppe erst zu diesem Zeitpunkt ihre wichtigste Anwendung lauffähig bekommen hat. In welchem Umfang das Potential zum massiv-parallelen Rechnen auf der magnitUDE tatsächlich genutzt wird, kann der Abbildung (6) entnommen werden. Sie zeigt für die ersten drei Quartale der Nutzung die verwendete Rechenleistung in Kernstunden gruppiert nach der Anzahl pro Rechnung verwendeter paralleler Ressourcen, angegeben in Rechenknoten à 24 Rechenkerne. In der ersten Nutzungsphase (3. Quartal 2016) fällt eine in etwa gleichmäßige Verteilung der Rechnungen mit mittlerer Parallelisierung auf sowie eine große Anzahl von Rechnungen, die die magnitUDE annähernd vollständig ausnutzen. Es ist anzunehmen, dass diese Phase von Tests hinsichtlich der Skalierbarkeit dominiert ist. In der zweiten Nutzungsphase (4. Quartal 2016) erfolgt eine Verschiebung zu einer geringeren Parallelisierung, bei der die Anwender*innen für ihre Produktivläufe scheinbar eine erste Abschätzung der Skalierbarkeit ihrer Anwendungen gefunden haben. In der dritten Nutzungsphase (1. Quartal 2017) erfolgt eine leichte Verschiebung zu höher parallelisierenden Rechnungen. Dies zeigt zum einen den Bedarf an diesem Rechnungstyp; zum anderen verdeutlicht es, dass es keine triviale Aufgabe ist, die gewünschte Parallelisierung „im Alltag“ auch umzusetzen. Dieses erschwert insbesondere den HPC-Neueinsteigern die effektive Nutzung eines solchen HPC-Systems. Um die Zugangshürde etwas zu reduzieren, ist ein erhebliches Kontingent für diese Anwender reserviert.

Holger Gollan. Foto: Vladimir Unkovic



(6) Größenverteilung der Rechnungen für die ersten drei Nutzungsquartale.
Quelle: Olaf Hasemann

Seit Inbetriebnahme der magnitUDE ist eine stetig wachsende Anzahl von Anfragen an das HPC-Support-Team hinsichtlich der Nutzung des System zu verzeichnen. Dabei kommen die Anfragen sowohl von „HPC-Neueinsteiger*innen“ als auch von erfahrenen Anwender*innen. Die HPC-Neueinsteiger benötigen besonders Unterstützung im Bereich der grundsätzlichen Bedienung und der Umsetzbarkeit beziehungsweise Portierung ihrer Anwendungen auf einem HPC-System. Die erfahreneren Nutzer*innen benötigen hingegen zumeist Ratschläge für die optimale Konfiguration der SLURM-Parameter für ihre Rechnungen oder für die Optimierung ihrer Anwendungsinstallationen beziehungsweise Kompilierung ihrer Programmcodes. Bei beiden Anwendergruppen versucht das Support-Team die Bereitstellung von benötigten numerischen Bibliotheken oder Softwarepaketen zentral zu realisieren und über sogenannte module-Umgebungen bereitzustellen. Damit sollen dem*der Nutzer*in gleichzeitig eine komfortable Verwaltung und optimale Konfiguration dieser Softwarepakete und -umgebungen ermöglicht werden. Ebenfalls trägt dieses Konzept zur effizienten Speichernutzung bei, da zum Beispiel nicht jede*r Nutzer*in gängige numerische Bibliotheken und Compiler in seinen*ihren nutzer*innenspezifischen Datenverzeichnissen ablegen muss.

Ausblick

Im Jahre 2016 wurde im Rahmen des Sonderforschungsbereichs 1242 von Prof. Peter Kratzer und Prof. Rossitza Pentcheva aus der Fakultät für Physik der UDE ein weiterer Großgeräteantrag bei der DFG zur Finanzierung von Rechenleistung gestellt, die im Rahmen des SFB anfallen werden. Nach erfolgreicher Bewilligung wurden die eingeworbenen Mittel genutzt, um eine Erweiterung der magnitUDE von etwa zehn Prozent zu ermöglichen. Auch hier hat das HPC-Support-Team bei der Planung und Beratung unterstützt, so dass Anfang 2017 nun 60 weitere Rechenknoten in das System eingebaut werden konnten, jeweils mit 256 GB Hauptspeicher. Damit besitzt das System in der jetzigen Ausbaustufe nun 624 Rechenknoten mit insgesamt 14.976 Rechenkernen und in der Summe 56.064 GB Hauptspeicher. In Testläufen wurden auf dem Gesamtsystem ca. 489 Tlop/s erreicht, was mehr als 92 Prozent der theoretisch maximal erreichbaren Rechenleistung von 528 Tflop/s entspricht. Es bleibt abzuwarten, welche Position die erweiterte magnitUDE damit in der neuen TOP 500-Liste erreicht, die im Juni 2017 veröffentlicht wird.

Derzeit gibt es Pläne, das System eventuell noch einmal zu erweitern. Die Randbedingungen am Aufstellungsort lassen dabei bis zu 120 weitere Rechenknoten zu, was etwa 20 Prozent der ursprünglichen Maschine bedeuten würde. Will man auch bei der zweiten Erweiterung noch eine weitgehend homogene Maschine erhalten (bisher unterscheiden sich die einzelnen Rechenknoten nur durch die Größe des jeweiligen Hautspeichers), so müsste diese sicherlich noch im Jahr 2017 erfolgen, bevor das Preis-Leistungs-Verhältnis der Prozessoren nicht mehr akzeptabel ist.

Abschließend bleibt zu sagen, dass die magnitUDE, in welcher Ausbaustufe auch immer, den Wissenschaftler*innen der Universität Duisburg-Essen Möglichkeiten zum Hochleistungsrechnen bietet, die es in dieser Form an der UDE bisher noch nicht gegeben hat. Insbesondere die Größe der Maschine ermöglicht den Nachweis, solche Systeme effektiv nutzen zu können, was letztlich eine Voraussetzung ist, zukünftig noch stärker dann auch Zugang auf die großen Höchstleistungsrechner in Jülich, München und Stuttgart zu erhalten.

Summary

In 2010 the first supercomputer CRAY XT6m at the UDE was brought into service. Four years later the demand on HPC-resources at the UDE has been increased eminently. Thus, in 2014 the UDE applied successfully under the supervision of the CCSS for the financial support of a new, more powerful supercomputer. With the grant of 3.3 million € from the German Research Council in 2015, the supercomputer magnitUDE was purchased from NEC and installed in the first half of 2016; it reached rank 279 in the TOP 500 of HPC ranking. Now, nine months into operation, we can state that the process to setup a well-working HPC-system satisfying the demands of all researchers was not as simple as it might look.

The first crucial part was the definition of the technical criteria for the tender, since they constrain the framework of possible configurations for the potential vendors. In this context the measurement of the performance was defined and a suitable testmatrix of typical applications of prospective usergroups was composed. Concurrently, the search for a

suitable location for magnitUDE had to be started, since the resources at both UDE-campi were not valid or exhausted. Finally, the UDE rented a new DataCenter in the logport Duisburg-Rheinhausen, which will host further IT-infrastructure of the ZIM as well. Obviously, the parallel establishment of the technical criteria and the location authored a few dilemmas, e.g. expectable power consumption. After the installation another two essential aspects had to be decided: the scheduling policy for the computations and the usage policy. A combination of fairshare and backfilling strategy was selected to weight the resource usage by the usergroup and take into account the scaling ability of their applications. Until now this decision has shown a good exploitation of magnitUDE. To ensure a robust operation the definition of user-specific disk quota and respites of file age were established in the usage policy as well.

Right at the beginning of the installation the first extension of magnitUDE was planned and, thus, early in 2017 additional 60 compute nodes were added. With this extension magnitUDE will hopefully still be part of the TOP 500 rankings in 2017/18.

In conclusion, the magnitUDE enables the researchers of the UDE to improve and demonstrate the performance of their computational applications. This opportunity has not been locally available to the researchers before and so is a new resource in the field of HPC at the UDE.

Die Autoren

Dominik Brands studierte von 2000 bis 2006 Maschinen- und Anlagenbau an der Gerhard-Mercator-Universität Duisburg, später Universität Duisburg-Essen (UDE). Im Jahr 2006 begann er als wissenschaftlicher Mitarbeiter am Institut für Mechanik der UDE, promovierte dort 2012 über die numerische Simulation von heterogenen Materialien und ist dort seit 2015 akademischer Oberrat. Der Schwerpunkt seiner Forschungsthemen sind aktuell im Bereich der mehrskaligen Festkörpermechanik und der massiv parallelen Finite-Elemente Simulation. 2013 wurde er in den Vorstand des Center for Computational Sciences and Simulation (CCSS) der UDE gewählt. Seit 2015 ist er gleichzeitig im Zentrum für Informations- und Mediendienste (ZIM) der UDE beschäftigt mit Aufgaben beim „Support des Wissenschaftlichen Rechnens" – kurz HPC-Supportteam. Im Rahmen dessen war er maßgeblich in Beantragung und Beschaffung der magnitUDE eingebunden und ist aktuell für die Administration mitverantwortlich.

Holger Gollan studierte von 1980 bis 1986 Mathematik an der damaligen Universität GHS Essen und promovierte dort im Jahr 1988 über modulare Gruppen- und Darstellungstheorie. Bis 1999 war er am Fachbereich Mathematik und im Institut für Experimentelle Mathematik als wissenschaftlicher Mitarbeiter beschäftigt, 1990/91 verbrachte er, unterstützt durch eine DFG-Förderung, ein Jahr an der Carleton University in Ottawa. 1999 wechselte er an das damalige Hochschulrechenzentrum der Universität Essen, das später in das ZIM der UDE überging. Hier betreut er unter anderem den Bereich des wissenschaftlichen Rechnens und ist Leiter des Geschäftsbereichs Lerntechnologien. Außerdem ist er seit 2011 Geschäftsführer des CCSS der UDE.

Olaf Hasemann studierte von 1983 bis 1993 Informatik an der Universität Dortmund mit dem Vertiefungsgebiet Physikalische Chemie. Seine Schwerpunkte waren Artificial Intelligence, Intelligente Systeme und Robotik. Zwischen 1990 und 1999 war er an der Universität Witten/Herdecke zunächst als studentischer, später als wissenschaftlicher Mitarbeiter tätig und befasste sich in zahlreichen Projekten mit der Modellierung und Optimierung biologischer sowie biochemischer Systeme. Ab 1999 wechselte er zur AQ.ON Water Technologie GmbH, die den Einsatz intelligenter Systeme zur Steuerung biotechnischer Prozesse in Form von Fernwirksystemen betrieb. 2005 folgten Projekte am Institut für Aerospace-Technologie der Hochschule Bremen, die in Kooperation mit der ansässigen Luft- und Raumfahrtindustrie Techniken der Virtual Reality und des Entwurfs Mechatronischer Systeme in den Produktentstehungszyklus einbanden. 2012 wechselte er zum Lehrstuhl für Fluiddynamik der UDE, wo er die Konzeption und Administration der HPC-Ressourcen einschließlich der Beratung der Anwender übernahm. Seit 2014 hat er im Rahmen des HPC-Supportteams diese Dienstleistung auf die gesamte Universität ausgeweitet; so war er maßgeblich an Beschaffung und Betrieb des HPC-Systems magnitUDE beteiligt. Generierung und Optimierung chemischer Reaktionsmechanismen mittels Strategien des Deep Learning sind ein weitere Schwerpunkt seiner Arbeiten.

Olaf Hasemann. Foto: Vladimir Unkovic

ERSCHIENEN

UNIKATE-Ausgaben 1992 bis 2017

1 Medizin
Krebsforschung (1992). Federführung: Manfred F. Rajewsky. 69 Seiten (vergr.).
ISBN 3–934359–01–9

2/3 Kommunikation
Design (1993). Federführung: Vilim Vasata. 121 Seiten (vergr.).
ISBN 3–934359–02–7

4/5 Naturwissenschaft
Umweltforschung: Globale Risiken (1994). Federführung: Günter Schmid. 117 Seiten (vergr.).
ISBN 3–934359–04–3

6/7 Geisteswissenschaft
Fremdsein: Historische Erfahrungen (1995). Federführung: Paul Münch. 173 Seiten.
ISBN 3–934359–06–X

8 Geisteswissenschaft
20 Jahre "poet in residence" (1996). Federführung: Jürgen Manthey. 90 Seiten.
ISBN 3–934359–08–6

9 Bildung durch Wissenschaft?
Ein Kolloquium (1997). Federführung: Justus Cobet, Klaus Klemm. 137 Seiten (vergr.).
ISBN 3–934359–09–4

10 Medizin
25 Jahre Transplantationsmedizin in Essen (1998). Federführung: Friedrich W. Eigler. 113 Seiten.
ISBN 3–934359–10–8

11 Naturwissenschaften
Physik: Unordnung und Selbstähnlichkeit (1999). Federführung: Fritz Haake. 108 Seiten.
ISBN 3–934359–11–6

12 Wirtschaft
Die Europäische Union im Zeichen des Euro (1999). Federführung: Dieter Schmitt. 120 Seiten.
ISBN 3–934359–12–4

13 Materialwissenschaft
Grundlagen für die Technik der Zukunft (2000). Federführung: Günter Schmid. 130 Seiten.
ISBN 3–934359–13–2

14 Europäische Gesellschaft
Annäherung an einen Begriff (2000). Federführung: Wilfried Loth. 96 Seiten.
ISBN 3–934359–14–0

15 Klinische Onkologie
Fortschritte in der Tumorbekämpfung (2001). Federführung: Herbert Rübben. 56 Seiten.
ISBN 3–934359–15–9

16 Erfahrung
Über den wissenschaftlichen Umgang mit einem Begriff (2001). Federführung: Paul Münch. 104 Seiten.
ISBN 3–934359–16–7

17 Design & Neue Medien
Kommunikationsgestaltung für eine global vernetzte Gesellschaft (2002). Federführung: Norbert Bolz. 100 Seiten.
ISBN 3–934359–17–5

18 Wirtschaftsinformatik
Wissensmanagement und E-Services (2002). Federführung: Heimo H. Adelsberger. 88 Seiten.
ISBN 3–934359–18–3

19 Umwelt Ruhr
Vitalität einer Region I (2002). Federführung: Wilfried Loth. 142 Seiten.
ISBN 3–934359–19–1

20 Herz-Kreislaufmedizin
Experimentelle und klinische Kardiologie (2003). Federführung: Gerd Heusch. 70 Seiten.
ISBN 3–934359–20–5

21 Geisteswissenschaften
Religion und Gewalt (2003). Federführung: Hubertus Lutterbach. 98 Seiten.
ISBN 3–934359–21–3

22 Medizin
Neurowissenschaften (2003). Federführung: Dietmar Stolke. 120 Seiten.
ISBN 3–934359–22–1

23 Ingenieurwissenschaft
Bauwesen – Ein Leistungsspektrum (2004). Federführung: Jörg Schröder. 120 Seiten.
ISBN 3–934359–23–X

24 Bildungswissenschaften
Bildungsforschung nach PISA (2004). Federführung: Klaus Klemm. 132 Seiten.
ISBN 3–934359–24–8

25 Medizin
Unsere Hormone (2005). Federführung: Klaus Mann. 110 Seiten.
ISBN 3–934359–25–6

26 Germanistik
Arbeit an/in der Kultur (2005). Federführung: Rüdiger Brandt. 110 Seiten.
ISBN 3–934359–26–4

27 Medizin: Immunologie
Das Immunsystem – Freund oder Feind? (2006). Federführung: C. Hardt, H. Grosse-Wilde. 110 Seiten.
ISBN 3–934359–27–2

28 Neue Medien
Interaktivität und Ubiquität (2006). Federführung: Edgar Heineken. 98 Seiten.
ISBN 3–934359–28–0

29 Wirtschaftswissenschaften
Empirische Wirtschaftsforschung (2007). Federführung: Walter Assenmacher. 98 Seiten.
ISBN 978-3–934359–29–1

30 Essen im Blick
Ein interdisziplinärer Streifzug (2007). Federführung: Heiko Schulz. 124 Seiten.
ISBN 978-3–934359–30–7

31 Ingenieurwissenschaften
Computersimulationen: Von Nano bis Giga (2007). Federführung: Dieter Hänel, Andrés Kecskeméthy. 124 Seiten.
ISBN 978-3–934359–31–4

32 Naturwissenschaften
Physik: Energieumwandlungen an Oberflächen. Federführung: Dietrich von der Linde. 128 Seiten.
ISBN 978-3–934359–32–1

33 Mathematik
Eine lebendige Wissenschaft (2008). Federführung: Axel Klawonn. 128 Seiten.
ISBN 978-3–934359–33–8

34 Geschichtswissenschaft
Europa: Geschichte und Kultur (2009). Federführung: Wilfried Loth. 144 Seiten.
ISBN 978-3–934359–34–5

35 Natur-/Geisteswissenschaften
Sterben, Tod – und dann? (2009). Federführung: Dieter Bingmann, Hubertus Lutterbach. 112 Seiten (vergr.).
ISBN 978-3–934359–35–2

36 Naturwissenschaften/Medizin
Kernmagnetische Resonanz (2009). Federführung: Christian Mayer. 70 Seiten.
ISBN 978-3-934359-36-9

37 Naturwissenschaften/Medizin
Nano meets Bio (2010). Federführung: Günter Schmid. 90 Seiten.
ISBN 978-3-934359-37-6

38 Geistes-/Gesellschaftwissenschaften
Beiträge zur „Ruhr 2010" (2010). Federführung: Jörg Engelbrecht. 106 Seiten.
ISBN 978-3-934359-38-3

39 Ingenieurwissenschaften
Herausforderung Elektromobilität (2011). Federführung: Ferdinand Dudenhöffer. 114 Seiten.
ISBN 978-3-934359-39-0

40 Social Sciences
Comparative, International, Transnational (2011). Federführung: Karen Shire. 140 Seiten.
ISBN 978-3-934359-40-6

41 Geschlechterforschung
Blick hinter die Kulissen (2012). Federführung: Maren A. Jochimsen. 132 Seiten.
ISBN 978-3-934359-41-3

42 Translationale Krebsforschung
Auf dem Weg zu neuen Therapien (2012). Federführung: Angelika Eggert. 150 Seiten.
ISBN 978-3-934359-42-0

43 NanoEnergie
Materialentwicklung für eine nachhhaltige Energieversorgung (2013). Federführung: Christof Schulz, Marion Franke. 150 Seiten.
ISBN 978-3-934359-43-7

44 Medizin
50 Jahre Universitätsklinikum: Highlights aus der Forschung (2013). Federführung: Gerd Heusch, Raimund Erbel. 146 Seiten (vergr.).
ISBN 978-3-934359-44–4

45 Fusionen
Was sagen die Fakultäten? (2014). Federführung: Ursula Renner-Henke, Patrick Eiden-Offe. 150 Seiten.
ISBN 978-3-934359-45–1

46 Patente
Motivation fur die Wissenschaft? (2014). Federführung: Stefanie Peschel, Oliver Locker-Grütjen. 82 Seiten.
ISBN 978-3-934359-46–8

47 Globale Kooperationsforschung
Transdisziplinäre und transkulturelle Perspektiven (2015). Federführung: Tobias Debiel. 98 Seiten.
ISBN 978-3-934359-47–5

48 Materials Chain
Forschung aus dem UA Ruhr-Profilschwerpunkt (2016). Federführung: Jörg Schröder, Ralf Drautz, Wolfgang Tillmann. 130 Seiten.
ISBN 978-3-934359-48–2

49 Mehrsprachigkeit im Ruhrgebiet
Vielfältig und doch individuell (2016). Federführung: Katja Cantone, Anastasia Moraitis. 112 Seiten.
ISBN 978-3-934359-49–9

in Vorbereitung:

51 Wasser

ABONNEMENT

Die UNIKATE im Abonnement

Ob im Bereich der Kommunikation, der Wirtschaft, der Medizin, der Ökonomie oder der Ökologie – im Zeitalter globaler Vernetzung sind wissenschaftliche Erkenntnisse zur entscheidenden Zukunftsinvestition geworden. Gleichzeitig führte der weltweite Zuwachs an Wissen zu immer feiner differenzierten Fächern, Disziplinen und Fachsprachen: die Wissenschaften erscheinen zunehmend unübersichtlicher und hermetischer.
Mit den UNIKATEN hat sich die Universität Duisburg-Essen die Aufgabe gestellt, Sie in verständlicher Form regelmäßig und unmittelbar aus der Forschungspraxis heraus über die aktuellen Entwicklungen in Wissenschaft und Forschung zu informieren – mit Ergebnissen und Erklärungen aus erster Hand.

ABONNEMENT

Vorname, Name _____
Straße / Postfach _____
PLZ, Ort _____

Hiermit abonniere ich die UNIKATE für 1 Jahr (2 Ausgaben) zum Preis von 12,50 € inkl. Versandkosten.
Das Abonnement verlängert sich automatisch um 1 Jahr (2 Ausgaben), wenn es nicht im laufenden Abonnementzeitraum gekündigt wird.

Bitte senden Sie mir die bereits erschienene Ausgabe
der UNIKATE zum Einzelbezugspreis von 7,50 € einschl. Porto (für Abonnenten 5,- €).

☐ Ich bin Abonnent/Abonnentin.

Datum _____ Unterschrift _____

☐ Die Bestellung kann durch Mitteilung an die Heinrich-Heine Buchhandlung, Viehofer Platz 8, 45127 Essen innerhalb von 7 Tagen widerrufen werden. Es gilt das Datum des Poststempels.

☐ Datum _____ Unterschrift _____

Ich zahle per Bankeinzug zu Lasten meines Kontos Nr.

bei _____ BLZ _____

Wenn mein Konto die erforderliche Deckung nicht aufweist, besteht seitens des kontoführenden Instituts keine Verpflichtung zur Einlösung.

☐ Datum _____ Unterschrift _____

Bitte schicken Sie mir eine Rechnung. Ich überweise den Betrag. _____

Per Post oder Fax an:

Heinrich-Heine Buchhandlung · Viehofer Platz 8 · 45127 Essen · Tel.: (0201) 820700 · Fax: (0201) 8207016

☐ _____

HINWEISE

Die UNIKATE

Schon 1992 verfolgte die Universität Essen das Konzept, die Öffentlichkeit mit der Herausgabe des damals noch ESSENER UNIKATE genannten Magazins für Wissenstransfer tiefergehend über die an der Hochschule erzielten Ergebnisse in Forschung und Lehre zu informieren. In einer Zeit, in der sich Wissenschaft wie auch akademisch interessierte Leserschaft in hohem Maße ausdifferenziert haben, soll dieses Magazin für Wissenstransfer die an der Universität erarbeiteten Informationen in differenzierter Weise widerspiegeln und – klassisch aufbereitet – der Öffentlichkeit themenzentriert transparent machen.

Im Mittelpunkt jeder Ausgabe stehen die Wissenschaftler*innen der Universität Duisburg-Essen mit ihren Originalbeiträgen und -berichten. Die Ausgaben orientieren sich dabei an den Herausforderungen, vor denen einzelne Fächer gegenwärtig stehen, wie auch an aktuellen wissenschaftlichen Zeitfragen, zu denen eine Universität insgesamt und nicht zuletzt auf Grund ihres öffentlichen Auftrags Stellung beziehen sollte.

Die UNIKATE erscheinen ausschließlich in Form von Themenheften; bisher hat sich die Reihe unter anderem mit der Krebsbehandlung, der Herz-Kreislaufmedizin, den Entwicklungen im Kommunikations- und Industrial Design, den globalen ökologischen Risiken, dem Kräfteverhältnis zwischen Bildung und Wissenschaft, der Chaosphysik, den Materialwissenschaften, dem Lebensraum Ruhrgebiet, der sich herausbildenden europäischen Gesellschaft und der Bildungsforschung nach PISA beschäftigt.

Durch die Konzentration auf jeweils ein Fachgebiet oder ein interdisziplinär ausgeleuchtetes Thema können wissenschaftliche Sachverhalte breiter dargestellt und komplexe Zusammenhänge fächerübergreifend verständlich erläutert werden.

Die UNIKATE werden vom Rektorat der Universität Duisburg-Essen in einer Auflage von derzeit 4.500 Exemplaren herausgegeben. Ansprechpartner für alle redaktionellen Belange sowie für Vertrieb und Anzeigenverwaltung ist das SSC – Science Support Centre an der Universität Duisburg-Essen. Das Magazin ist zum Preis von 7,50 € im Buchhandel erhältlich. Im Abonnement (zwei Ausgaben pro Jahr, 12,50 €) sind die UNIKATE über die Heinrich-Heine-Buchhandlung, Viehofer Platz 8, 45127 Essen zu beziehen.

UNIKATE

IMPRESSUM

Herausgegeben mindestens zweimal jährlich vom Rektorat der Universität Duisburg-Essen, 45117 Essen.

Auflage: 4.500

Redaktions- und Verlagsanschrift:
UNIKATE
Universität Duisburg-Essen/
Wissenschaftsverlag SSC
Science Support Centre
45117 Essen
Tel.: 02 01/1 83-32 54
Fax: 02 01/1 83-46 94
E-Mail: unikate@uni-duisburg-essen.de

Wissenschaftlicher Beirat:
Prof. Dr. Eckart, Hasselbrink, Prof. Dr.-Ing. Wojciech Kowalczyk, Prof. Dr. Renate Martinsen, Prof. Dr. Christoph Marx, Prof. Dr. Ludwig Mochty, Prof. Dr.-Ing. Jörg Schröder, Prof. Dr. Astrid Westendorf

Federführung der Ausgabe 50:
Prof. Dr.-Ing. Andreas M. Kempf,
Prof Dr. Ludwig Mochty
Redaktion: Dr. Barbara Bigge (verantw.)
Korrektorat: Esther Hasan
Layout/Grafik: Paran Pour,
mail@paranpour.com
Gestaltungskonzept: Prof. Vilim Vasata
Fotografie und Bildbearbeitung:
Vladimir Unkovic
Druck: Asterion Germany GmbH, Viernheim
Buchhandels- und Abonnementvertrieb:
Heinrich-Heine-Buchhandlung, Viehofer Platz 8, 45127 Essen; Tel.: 02 01/820 70-0; Fax: 820 70-16; E-Mail: heine.buchhandlung@t-online.de

Einzelverkaufspreis (Buchhandel): 7,50 €
Abonnement (2 Hefte/Jahr, inkl. Versand): 12,50 €
Die UNIKATE finden Sie im Internet unter:

www.uni-duisburg-essen.de/unikate

Gedruckt auf chlorfreiem Papier.

Nachdruck und Reproduktion von Text, Fotos und Grafiken nur nach Abstimmung mit der Redaktion. Die Redaktion bemüht sich regelmäßig, die Rechteinhaber von veröffentlichten, jedoch nicht selbst erstellten Bild- und Grafikbeiträgen zu ermitteln und die Rechte abzugelten. Bei nicht zu ermittelnden oder inkorrekt angegebenen Nachweisen bitten wir um Nachsicht. Alle Rechte vorbehalten.

ISBN: 978–3–934359–50–5
ISSN: 1869–3881

© Universität Duisburg-Essen

Gerichtsstand: Essen

DuEPublico

Duisburg-Essen Publications online

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

ub | universitäts bibliothek

Dieser Text wird über DuEPublico, dem Dokumenten- und Publikationsserver der Universität Duisburg-Essen, zur Verfügung gestellt. Die hier veröffentlichte Version der E-Publikation kann von einer eventuell ebenfalls veröffentlichten Verlagsversion abweichen.

DOI: 10.17185/duepublico/70369
URN: urn:nbn:de:hbz:464-20201015-150919-5

Alle Rechte vorbehalten.