Karlsruher Institut für Technologie

SCC news

STEINBUCH CENTRE FOR COMPUTING

AUSGABE 1 | 2022

Mit „Future Technologies“ zu den HPC-Systemen von morgen
With “Future Technologies” to the HPC systems of tomorrow

Von gekrönten Buchstaben und metaphorischen Wörtern
Of Crowned Letters and Metaphorical Words

Erster AI-HERO Hackathon zum Thema Energieeffiziente KI
First AI-HERO Hackathon on Energy Efficient AI

Editorial

Liebe Leserin, lieber Leser,

was unsere Vorfahren gemeinsam entwickelt und erarbeitet haben ist auch essentiell für unseren Erfolg, den Erfolg der Nachfolgenden. Die englische Sprache macht dies eindrucksvoll deutlich, denn im „Successor", dem Nachfolger, verbirgt sich auch der „Erfolg". Die Evolution von Hard- und Softwaretechniken aus 50 Jahren zeigt, wie Entwicklungen erfolgreich aufeinander aufbauen und möglicherweise ein Türöffner für weitere technische Revolutionen sind. Davon inspiriert setzt das SCC im Nationalen Hochleistungsrechenzentrum NHR@KIT auf ein mit Zukunftstechnologien ausgestattetes Hard- und Softwaretestbett, dessen innovative und disruptive Komponenten Forschende frühzeitig testen und anwenden können (S. 19).

Im Juni besuchte uns der Informatikprofessor und Turing Award-Gewinner Jack Dongarra und berichtete über seine Forschungsarbeiten zu Soft- und Hardware im High Performance Computing (HPC) der letzten 50 Jahre. Darauf aufbauend forscht Hartwig Anzt am SCC mit seiner Nachwuchsgruppe an Softwarekonzepten für zukünftige Supercomputer und führt so die Arbeiten seines Mentors fort. Nach fünf Jahren nimmt Anzt nun den Ruf an die University of Tennessee an und wird Nachfolger von Dongarra (Titelseite Mitte und S. 28).

Jahrhundertealte religiöse Schriftrollen unserer Vorfahren untersucht ein interdisziplinärer Verbund aus Geisteswissenschaften und Informatik im Projekt Materialisierte Heiligkeit. Das SCC bringt darin seine Expertise im Forschungsdatenmanagement ein. Im Fokus liegt ein Forschungsdatenrepositorium mit modernsten Annotations-, Analyse- und Visualisierungswerkzeugen (S. 23).

Apropos Nachwuchs: Im Februar fand ein Helmholtz-Programmierwettbewerb statt. Angeleitet von unserem KI-Team untersuchten Nachwuchswissenschaftlerinnen und -wissenschaftler die Energieeffizienz von KI-Algorithmen auf HPC-Ressourcen und erarbeiteten besonders energieeffiziente Lösungen in unterschiedlichen Szenarien (S. 32).

Viel Freude beim Lesen

Martin Frank, Martin Nußbaumer, Bernhard Neumair, Achim Streit

Dear reader,

What our ancestors developed and worked on together, is also essential for our success and the success of those who come after us. The English language makes this impressively clear because the word "successor" contains "success". The evolution of hardware and software technologies over 50 years shows how developments successfully built on each other and are possibly opening the door for subsequent technical revolutions. Inspired by this, SCC's National high-performance computing centre, NHR@KIT, offers a hardware and software test-bed equipped with future technologies. Researchers can test and apply the innovative and disruptive components of the NHR@KIT at an early succession stage (p. 19).
In June, computer science professor and Turing Award winner Jack Dongarra visited us to talk about his research on software and hardware in high-performance computing (HPC) over the past 50 years. Building on this work of his mentor, Hartwig Anzt is conducting research with his research group on software concepts for future supercomputers. After five years at SCC, he has accepted the offer to succeed Jack Dongarra at the University of Tennessee (centre cover and p. 28).

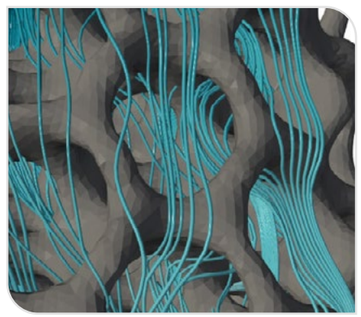
Centuries-old religious scrolls of our ancestors are being studied by an interdisciplinary collaboration of humanities and computer science experts in the Project "Materialisierte Heiligkeit". The SCC is contributing expertise in research data management to this project with focus on building a research data repository, adding state-of-the-art annotation, analysis, and visualization tools (p. 23).

Speaking of young talent. Guided by our AI team, young scientists studied the energy efficiency of AI algorithms on HPC computers and came up with particularly energy-efficient computing solutions at a Helmholtz programming competition in February (p. 32).

Enjoy reading

Martin Frank, Bernhard Neumair, Martin Nußbaumer, Achim Streit

Inhaltsverzeichnis



Papierlose Bewerbung für einen Studiengang am KIT

Zum Sommersemester 2022 war es durch die Einführung der papierlosen Bewerbung am KIT erstmals möglich, den gesamten Bewerbungsprozess für einen Studiengang elektronisch abzuwickeln. Dazu wurden zusammen mit dem Studierendenservice, dem International Students Office sowie Vertreterinnen und Vertretern der KIT-Fakultäten und der Firma CAS Software AG die Prozesse für das am SCC eingesetzte Bewerbungs- und Zulassungsmanagement der CAS Campus Management Software angepasst.

Elisabeth Syrjakow, Thomas Berendonck

Bisher war es für eine Bewerbung für einen Studiengang am KIT notwendig, sich nicht nur am Bewerbungsportal zu registrieren und darüber eine Bewerbung elektronisch abzusenden, sondern im Anschluss daran auch die dazugehörigen Bewerbungsunterlagen auszudrucken und rechtzeitig unterschrieben per Post an das KIT zu schicken. Erst dann konnte die Bewerbung im weiteren Bewerbungsprozess berücksichtigt werden (Abbildung 1).

Durch die Einführung der papierlosen Bewerbung wurde es ermöglicht, dass der gesamte Prozess für einen Studiengang digital abgewickelt werden kann. Für die Realisierung wurde das am KIT eingesetzte Software-Modul BZM (Bewerbungs- und Zulassungsmanagement) der CAS Campus Software entsprechend erweitert und angepasst. Dieser Beitrag stellt die wichtigsten Neuerungen im BZM vor.

Die digitale Sammelakte

Als Voraussetzung für eine dezentrale Bearbeitung der digital eingegangenen Bewerbungsunterlagen war es erforderlich, eine digitale Sammelakte mit Inhaltsverzeichnis und Fußnote einzuführen. Diese enthält die eingereichten Dokumente der Bewerbenden in strukturierter Form und ersetzt die früher verwendete Papierakte.

Die Generierung einer oder mehrerer Sammelakten kann im Bearbeitungsportal jederzeit per Mausklick angestoßen werden. Die Verwaltungsfachkräfte werden in diesem Fall per E-Mail mit Download-Link über die Fertigstellung benachrichtigt. Somit ist es möglich, die Sammelakten zu einem späteren Zeitpunkt aufzurufen, ohne dass die Bearbeitung der Bewerbungen im Bearbeitungsportal unterbrochen werden müsste.

Eine weitere wichtige Eigenschaft der Sammelakte besteht darin, dass mit Hilfe des Rollen- und Rechtekonzepts im BZM unterschiedliche Sichten auf die Inhalte einer Sammelakte konfiguriert werden können. Damit lassen sich beispielsweise mit den verschiedenen Rollen im BZM Sammelakten generieren, welche nur die jeweils für den Rolleninhaber relevanten Dokumente enthalten. Mit Hilfe dieser maßgeschneiderten Sammelakten kann den unterschiedlichen Bedürfnissen der im Prozess involvierten Bearbeitungsgruppen am KIT entsprochen werden.

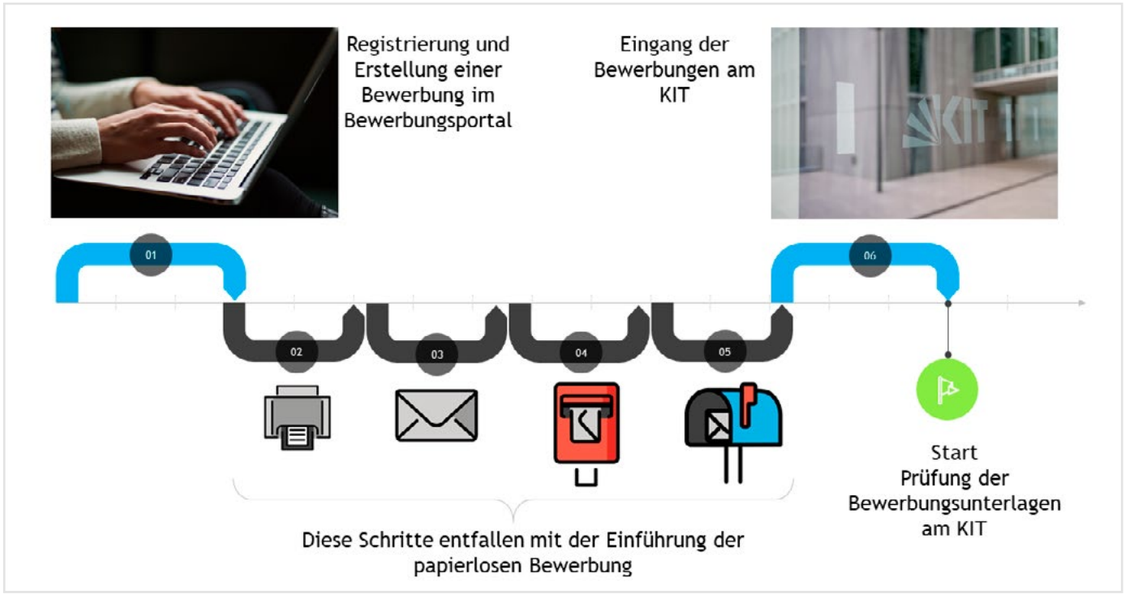


Abbildung 1: Zu optimierender Bewerbungsprozess für einen Studienplatz am KIT

Das Hochladen von Dokumenten

Damit Bewerbende nicht nur ihre Daten eingeben, sondern auch die zugehörigen Dokumente hochladen können, war eine Erweiterung des Bewerbungsportals erforderlich (Abbildung 2). Dazu wurden die jeweiligen Portalbausteine zusätzlich mit einer Upload-Funktionalität ausgestattet und die Prüflogik erweitert, um die Bewerberinnen und Bewerber bei der abschließenden Überprüfung im Bewerbungsportal nicht nur auf fehlende Daten, sondern auch auf fehlende Dokumente hinweisen zu können.

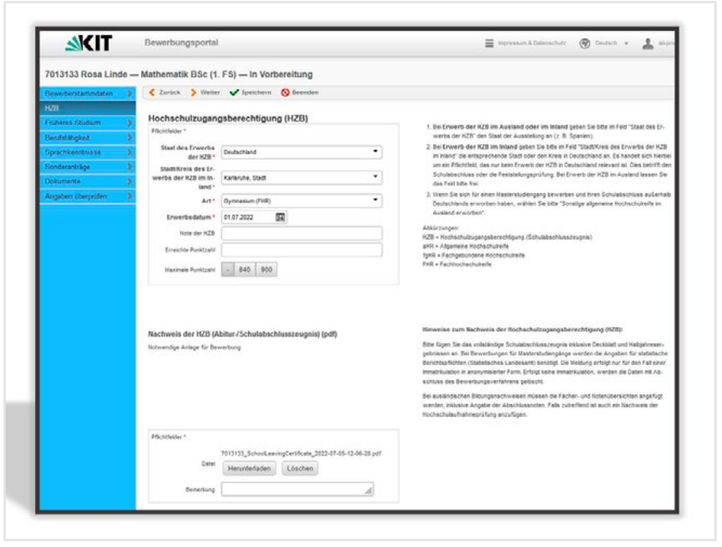


Abbildung 2: Neue Upload-Funktionalität innerhalb der einzelnen Portalbausteine zum Hochladen der zugehörigen Bewerbungsdokumente

Um verwaltungsseitig pro Portalbaustein jede Dateneingabe und jedes Dokument separat bewerten und kommentieren zu können, war ebenfalls eine Erweiterung des Portals notwendig. Des Weiteren mussten alle Studiengänge mit einer Auswahl der zuvor neu spezifizierten elektronischen Dokumenttypen verknüpft werden, so dass Bewerbende die Möglichkeit haben, alle für einen Studiengang erforderlichen Nachweise elektronisch einzureichen. Die Besonderheiten einzelner Studiengänge und unterschiedlicher Gruppen von Bewerbenden erforderten die Definition von Regeln, die steuern, wann ein bestimmtes Dokument innerhalb des Bewerbungsprozesses erforderlich ist.

Das Nachreichen von Dokumenten

Da der gesamte Bewerbungsprozess digitalisiert wurde, musste auch eine Funktionalität eingeführt werden, die das Nachreichen von Dokumenten ermöglicht, falls dies erforderlich ist. Verwaltungsseitig können eingereichte Dokumente entsprechend markiert werden. Bewerbende werden dann sofort per E-Mail informiert, und im Bewerbungsportal darauf aufmerksam gemacht und gezielt aufgefordert, diese zu ergänzen sowie erneut einzureichen.

Das Nachreichen von Dokumenten ist bis zu einer vorher von der Verwaltung festgelegten Frist möglich. Sie kann je nach Dokumentart variiert werden. Dabei findet eine Unterscheidung zwischen zulassungsverhindernden, immatrikulationsverhindernden und optionalen Dokumenten statt. Beispielsweise kann definiert werden, ob ein Nachreichen bis zum Ende der Bewerbungsfrist notwendig ist oder erst bis zum Ende der Frist für die Immatrikulation.

Das Filtern in den Bewerbungslisten

Zusätzlich zu den bereits vorhandenen Filtern in den Bewerbungslisten wurden weitere Filter z.B. nach Studienangebot, Verfahrensart, Dokumentenstatus-Kategorie oder Bildungsausländerstatus eingeführt. In der Verwaltung kann dadurch die Bearbeitung der einzelnen Bewerbungen noch effizienter durchgeführt werden.

Die Filtereinstellungen lassen sich entweder persönlich oder global speichern. Globale Filter haben den Vorteil, dass diese allen in der Verwaltung zur Verfügung gestellt werden können. Dies unterstützt die Anforderung an das Arbeiten in unterschiedlichen Teams. Beispielsweise gibt es Teams, die auf bestimmte Studiengänge spezialisiert sind, während andere ihren Fokus auf den unterschiedlichen Herkunftsländern der Bewerbenden haben.

Workflow für immatrikulationsrelevante Daten und Dokumente

Die bisherigen Statusautomaten für freie, zulassungsbeschränkte und DoSV[1]-Studienangebote wurden erweitert, da die Bewerbenden erst verpflichtende Angaben für die Einschreibung vornehmen und immatrikulationsrelevante Dokumente einreichen müssen, wenn sie den Studienplatz annehmen.

[1] Das Dialogorientierte Serviceverfahren (DoSV) ist ein webbasiertes, hochschulübergreifendes System zur Koordinierung der Bewerbungen und Zulassungen in örtlich zulassungsbeschränkten Studiengängen. DoSV wird von der Stiftung für Hochschulzulassung durchgeführt und ist unter der Webseite dosv.hochschulstart.de erreichbar.

Um seitens der Verwaltung eine erneute Prüfung anstoßen zu können, wurden drei weitere Status („Imma.-Antrag eingereicht“, „Imma.-Antrag unvollständig“ und „Imma.-Antrag vollständig“) sowie entsprechende Statusübergänge eingeführt (Abbildung 3). Mit dieser im Bearbeitungsportal neuen Funktionalität „Imma.-Prüfung abschließen“ können Sachbearbeiterinnen und Sachbearbeiter die Bewerbung nach erneuter Überprüfung automatisch in den korrekten Folgezustand überführen.

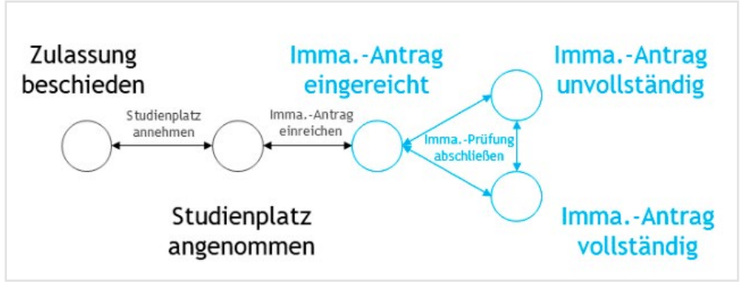


Abbildung 3: Erweiterung der bisherigen Statusautomaten um weitere Status und Statusübergänge für den Immatrikulationsworkflow

Zusammenfassung

Die Einführung dieses rein digitalen Bewerbungsverfahrens hat viele Vorteile mit sich gebracht, von denen im Folgenden die wichtigsten noch einmal zusammengefasst sind:

- Wettbewerbsvorteil
  Ein erhöhter Komfort der Nutzenden steigert den Wettbewerbsvorteil gegenüber den anderen Universitäten und Hochschulen.
- Optimierter Bearbeitungsprozess
  Verwaltungsseitig entfällt eine aufwändige Bearbeitung des Posteingangs, da keine Briefunterlagen mehr entgegengenommen und manuell bearbeitet werden müssen.
- Papierersparnis
  Aufgrund der digitalen Sammelakte entsteht weniger Aufwand im Umgang mit den Akten, da keine Papierakten mehr erstellt und verwaltet werden müssen. Des Weiteren reduziert das Verfahren den natürlichen Ressourcenverbrauch und schont damit die Umwelt.
- Effizientere Zusammenarbeit
  Ein besseres verteiltes Arbeiten innerhalb der Verwaltung und in Zusammenarbeit mit den KIT-Fakultäten wurde ermöglicht, da alle Dokumente nun digital vorliegen.
- Verbesserter Bedienkomfort
  Die Handhabung der Bewerbungen im BZM wurde durch neue unterstützende Funktionalitäten wie z.B. verschiedene Filtermöglichkeiten verbessert.

Fazit

Trotz besonderer Herausforderungen im Rahmen der gesamten Planung und Inbetriebnahme der papierlosen Bewerbung unter Berücksichtigung sich überlappender Bewerbungsphasen, konnte dieser Dienst rechtzeitig zum Sommersemester 2022 am KIT zur Verfügung gestellt werden. Ab dem kommenden Wintersemester wird die Entlastung des Studierendenservice durch den Wegfall der Bewerbungsunterlagen in Papierform deutlich spürbar werden, da ungefähr 15.000 Briefe nicht mehr geöffnet, bearbeitet und abgelegt werden müssen.

Paperless application for a degree program at KIT

The introduction of a paperless application process for the summer term of 2022 makes it possible to complete the entire application process for a degree program at KIT electronically for the first time. For this purpose, the processes for the application and admission management (BZM) of the CAS Campus Management Software used at the SCC were adapted together with the Students Service (SLE), the International Students Office (IStO) as well as representatives of the KIT departments and the CAS Software AG.

The big picture of AAI

Das SCC ist gleich mit zwei Arbeitsgruppen und einer Reihe von Aktivitäten in unterschiedlichen Projekten zum Thema AAI – also Authentication and Authorisation Infrastructures – aktiv. In dieser Ausgabe der SCC-News gibt es daher zwei Artikel, die dieses Thema aus unterschiedlichen Perspektiven beleuchten.
Dieser Artikel widmet sich den Aktivitäten im Kontext europäischer Projekte, die in der ganzen EU sicher nutzbare, föderierte IT-Infrastrukturen für Forschende etablieren. Hier ist das SCC vor allem bei der Entwicklung der Architektur und den Policies, aber auch bei der Entwicklung und Integration von Kommandozeilen-Werkzeugen engagiert.

Diana Gudu, Marcus Hardt, Gabriel Zachmann

Architektur

Die Grundlage der Aktivitäten vieler Datenmanagementprojekte am SCC bildet die im Rahmen der AARC- und AARC2-Projekte[1] identifizierte Blueprint Architecture for an interoperable AAI[2].

[1] aarc-project.eu
[2] aarc-community.org/guidelines/aarc-g045

Diese Blueprint Architecture (oder auch BPA) wurde vom AARC-Architecture-Team aus den AAI-Architekturen mehrerer internationaler Forschungskollaborationen zusammengesetzt. Hierbei wurden die jeweils unterschiedlichen Aspekte in einem Schaubild erfasst und mit einer einheitlichen Nomenklatur beschrieben (Abbildung 1).

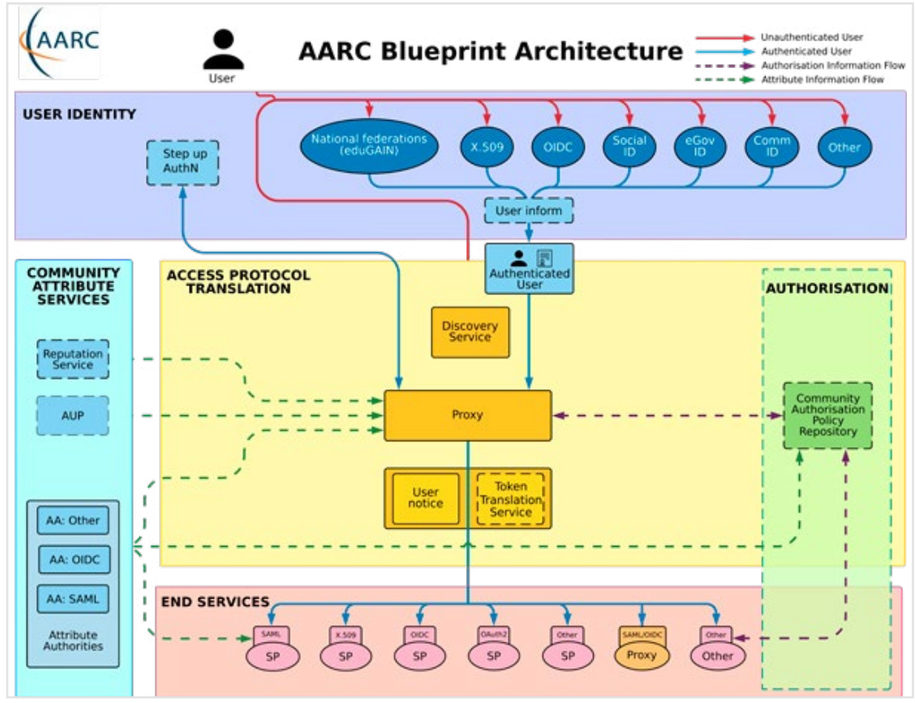


Abbildung 1: Die AARC Blueprint Architecture (BPA) zeigt die unterschiedlichen funktionalen Blöcke, die aus den verschiedenen föderierten AAIs zusammengetragen wurden. Die Pfeile repräsentieren hierbei nicht die Bewegungen der Nutzenden, sondern vielmehr den Fluss der Informationen in der Architektur.

Bei dieser Architektur wurde streng darauf geachtet, dass mindestens die Protokolle SAML, OpenID Connect oder X.509 eingesetzt werden können.

Die zentrale Ebene *Access Protocol Translation* enthält hierbei den sogenannten SP-IdP-Proxy (SP: Service Provider, IdP: Identity Provider). Diese Komponente wurde von fast allen untersuchten Infrastrukturen eingeführt, weil die direkte Herausgabe von Attributen häufig zu Problemen führte. Dies liegt in der Regel daran, dass IdP-Betreiber bevor sie Attribute an jeden SP herausgeben oftmals eine Einzelfallprüfung vornehmen. Dies skaliert aber bei gut 3.500 SPs und fast 5.000 IdPs ohne den Proxy-Ansatz nicht. Der europäischen Datenschutzgrundverordnung (DSGVO) wird hierbei entsprochen, weil es sich bei den verarbeiteten Daten nicht um kritische persönliche Daten (wie z.B. Geburtsdatum oder Religionszugehörigkeit) handelt, und weil nur solche Daten

transferiert werden, die zur Autorisierung auf verteilten Systemen benötigt werden. Mehrere Datenschutzbeauftragte haben den Ansatz unabhängig voneinander untersucht und dabei festgestellt, dass es ausreicht, die Nutzenden über die herausgegebenen Daten zu informieren, und die zuständigen Stellen zu benennen.

Die Ebene *Community Attribute Services* in Abbildung 1 beinhaltet Informationen über ein User-Objekt, um dieses mit weiteren – zumeist community-spezifischen – Attributen anzureichern. Dies sind oftmals Informationen über Gruppenmitgliedschaften. Die Ebene *Authorisation* überlappt mit Diensten und Proxy, was andeutet, dass Autorisierungsentscheidungen – je nach Ausprägung der Architektur – an mehreren Stellen getroffen werden können.

Skalierbarkeit

Im Laufe des AARC-Projekts wurde die Blueprint Architecture weiterentwickelt und eine weitere Proxy-Komponente eingeführt. Hierbei wird der SP-IdP-Proxy in eine Community AAI und einen Infrastructure Proxy aufgeteilt. Eine Community AAI wird üblicherweise für jeweils eine wissenschaftliche Community aufgebaut. Die Größe der Community ist dabei variabel. Beispielsweise haben sich verschiedene ESFRI-Projekte[3] aus dem Life Science-Bereich entschlossen, ihre Community AAIs zu einer gemeinsamen zusammenzufassen. Kleinere Community AAIs, wie z.B. die von eduTEAMS betriebene AAI für das Latin American Giant Observatory (LAGO), haben aber ebenso ihre Berechtigung. Oft geben das Vertrauen der Dienste in die Benutzer- und Gruppenverwaltung einerseits und die organisatorischen Strukturen innerhalb von Communities und Infrastrukturen andererseits den Ausschlag für die Entscheidung, wie groß eine Community definiert ist oder wird.

[3] European Strategy Forum on Research Infrastructures (www.esfri.eu)

Der Infrastructure Proxy versammelt ganze Infrastrukturen hinter sich. Dies können beispielsweise einzelne HPC-Geräte sein. Hier würde der Infra-Proxy z.B. die föderierten Identitäten in ein lokales LDAP-Verzeichnis schreiben, so dass Nutzende sich via SSH anmelden können. Die im Projekt bwIDM vom SCC entwickelte AAI-Software RegApp stellt ebenfalls eine derartige Teilfunktionalität als Infrastruktur-Proxy zur Verfügung. Sie wird unter anderem für Landesdienste eingesetzt, die Hochschulen und Universitäten in Baden-Württemberg nutzen (Seite 11).

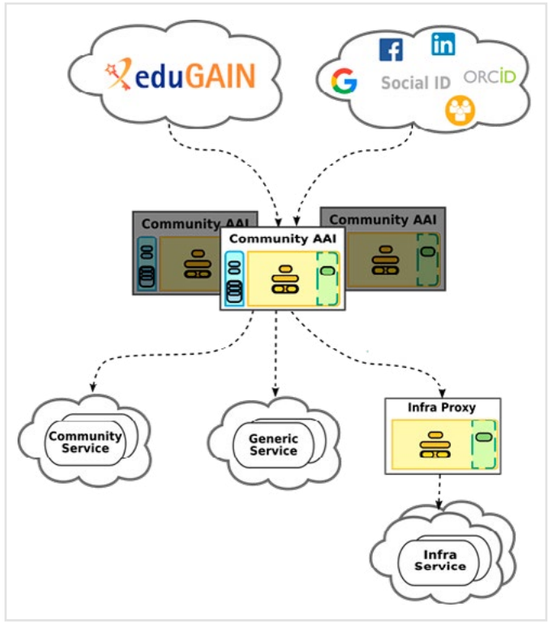


Abbildung 2: Infrastructure Proxy-Dienste (laut AARC BPA) können an eine spezielle Community angebunden sein (Community Service), allgemein mehreren Communities zur Verfügung stehen (Generic Service) oder über einen Infrastruktur-Proxy angeschlossen werden (Infrastructure Service)

Attribute

Großer Beliebtheit erfreut sich die Proxy-Lösung nicht nur aufgrund der praktikableren Attributfreigabe, sondern auch weil dieser Proxy die Möglichkeit erlaubt, eine Community zu verwalten. Das bedeutet, dass Gruppen definiert werden können, die beispielsweise die Zugehörigkeit zu einem wissenschaftlichen Experiment spezifizieren. Oft können einzelnen Nutzenden spezifische Rollen innerhalb einer Gruppe zugeordnet werden. Manche Communities verwalten am Proxy sogar die Zugriffsrechte für einzelne Datensätze.

Angeschlossene Dienste können, basierend auf diesen Gruppenmitgliedschaften, Zugriffsentscheidungen fällen. Natürlich werden auch relevante Attribute aus der Heimateinrichtung vom Proxy an die Dienste weitergereicht. Hierdurch ist es Diensten möglich, so wie z.B. in der Landesföderation bwIDM praktiziert, Zugriffsentscheidungen zusätzlich auf der Grundlage dieser Attribute zu treffen. Ein Beispiel hierfür wären die in der jeweiligen Heimateinrichtung gepflegten sogenannten entitlement-Attribute, mit denen festgelegt wird, welche ihrer User einen Dienst wie z.B. bwSync&Share nutzen dürfen. Manche Infrastrukturen erlauben als dritte Möglichkeit auch eine „Assurance“ (also Zusicherungen über die Qualität einer Identität) mit in die Zugriffsentscheidung einfließen zu lassen.

All diese Möglichkeiten, Nutzende, die an einer Föderation teilnehmen, am eigenen Dienst korrekt einordnen zu können, und die richtigen Zugriffsrechte zu erteilen, erfordert eine genaue Definition der Attribute und der Verfahrensweisen, wie diese erzeugt bzw. weitergeleitet werden.

Ein wesentlicher Teil der Arbeit aus den EU-Projekten AARC und AARC2, der aktuell im Projekt GEANT4-3 und künftig auch in GEANT5-1 fortgeführt wird, entfällt daher auf die Standardisierung und Harmonisierung solcher Attribute. Die AEGIS-Gruppe[4] beauftragt technische Teams, Richtlinien – die sogenannten AARC Guidelines[5] – zu entwickeln, und definiert manche hiervon für die in AEGIS zusammenarbeitenden Infrastrukturen als bindend[6]. AEGIS bringt viele internationale Forschungsinfrastrukturen zusammen, unter anderem EGI, WLCG, ELIXIR/Lifesciences, Internet2, HIFIS, FENIX, PaNOSC, PRACE, DARIAH und XSEDE.

Die definierten Attribute und die Blueprint Architecture sind zwar prinzipiell unabhängig voneinander, werden aber im gegenseitigen Kontext entwickelt und zumeist auch gemeinsam eingesetzt.

Policies

Die Zusammenarbeit der internationalen Identitätsprovider via eduGAIN mit den großen internationalen Forschungsinfrastrukturen erfordert selbstverständlich Regelungen. Die Herausforderungen liegen darin, die angeschlossenen Dienste einerseits und die Identitäts- und Attribut-Provider andererseits zusammenzubringen. Typische Fragestellungen eines jeden Dienstes sind: „Wie gut kann man den Daten vertrauen?“ und „Was tut man, wenn jemand einen Dienst für illegale Aktivitäten missbraucht?“. Identitäts- und Attribut-Provider hingegen müssen sichergestellt wissen, dass Daten, die sie über Nutzende herausgeben, vom Empfänger sachgemäß behandelt werden. Die zusätzliche Schwierigkeit ist hierbei, dass die Anzahl der Teilnehmenden zu groß ist, als dass jeder mit jedem einen Vertrag unterzeichnen könnte.

Um das notwendige Vertrauen einerseits, aber auch eine gewisse Flexibilität andererseits zu gewährleisten, hat das AARC Policy-Team einen Satz an Rahmenregelungen (Policy Frameworks) zusammengestellt. Ebenso wie bei der Blueprint Architecture, führte man existierende Regelwerke zusammen, anstatt neue zu verfassen. Hierbei konzentrierte man sich darauf, die Kompatibilität zwischen den unterschiedlichen Regelungen sicherzustellen. Eine besondere Berücksichtigung fand die DSGVO, die im Laufe des AARC-Projekts in Kraft trat. Konkret wurden die Policy Frameworks auf Vereinbarkeit mit der DSGVO geprüft und (wo sinnvoll) Vorlagen für DSGVO-konforme Policies (z.B. für die dienstbezogene Privacy Policy) bereitgestellt.

Das Konzept der Rahmenregelungen wurde gewählt, um die nötige Flexibilität bei der Umsetzung in unterschiedlichen Infrastrukturen und Ländern zu gewährleisten. Bei diesem Ansatz wird vorgegeben, zu welchen Punkten eine Regelung tatsächlich benötigt wird. Oftmals wird ein konkretes Regelwerk nur als Beispiel gegeben.

Insgesamt neun Dokumentvorlagen für Regelungen samt Begleitmaterial wie Moodle-Kurse, Youtube-Videos und weiterführende Links werden als Policy Development Kit fortlaufend gepflegt. Sie sind auf den Webseiten der AARC-Community[7] zu finden.

Die drei Aspekte Policies, Attribute und Architektur sind immer gemeinsam gemeint, wenn es um eine Implementierung der AARC Blueprint Architecture geht. An allen Bestandteilen der Architektur war das SCC wesentlich beteiligt.

Implementierung in Projekten

Diese Blueprint Architecture wurde bereits in einer Reihe von neueren Projekten implementiert. So ist beispielsweise die Helmholtz AAI im Rahmen von HIFIS (Helmholtz Federated IT Services) auf Basis von Vorarbeiten aus der Helmholtz Data Federation (HDF) entstanden. In HIFIS wurden die Policies vom SCC dahingehend angepasst, dass diese die Anforderungen der deutschen Helmholtz-Zentren erfüllen. Das SCC brachte Erfahrungen aus dem bwIDM-Projekt ein, welche sich in der Struktur des Policy Development Kit ausdrücken und so die internationale Zusammenarbeit vereinfachen.

HIFIS setzt zudem stark auf die Definition von Virtuellen Organisationen, die als Gruppen an der HIFIS Community-AAI gepflegt werden. HIFIS favorisiert vorwiegend das Token-basierte Protokoll OpenID Connect. Es erlaubt, im Vergleich zu SAML, zusätzlich zur Browser-basierten Authentifizierung, auch die AAI für Delegation, REST- oder Kommandozeilen-Schnittstellen zu verwenden. Am SCC werden Werkzeuge entwickelt, die dies unterstützen, sowie OpenID Connect-Identitätsprovider betrieben.

Im Landesprojekt bwIDM2 wird maßgeblich vom SCC die bereits genannte Software RegApp so erweitert, dass diese neben SP-IdP- und Infrastruktur-Proxy auch als Community AAI und mit delegierbarer Community-Verwaltung eingesetzt werden kann. Dazu kommt die Erweiterung um eine Zwei- bzw. Multifaktor-Authentifizierung, die für alle angeschlossenen Dienste aktiviert werden kann. Diese Funktionalität erlaubt einen vielfältigen Einsatz in den unterschiedlichsten Projekten und Anwendungsszenarien.

Beim Aufbau der vielfältigen Nationalen Forschungsdateninfrastrukturen (NFDI) bringt das SCC das Architektur-Know-how für AAIs mit ein. Hier werden eine ganze Reihe von Community AAIs benötigt, die auf eine teilweise gemeinsam genutzte Infrastruktur zugreifen. Dabei kommt der Zusammenarbeit von Mitgliedern aus den unterschiedlichsten Forschungsgemeinschaften ein besonderes Augenmerk zu.

[4] wiki.geant.org/display/AARC/AEGIS
[5] aarc-community.org/guidelines
[6] wiki.geant.org/display/AARC/AARC+Documents+Approved+by+AEGIS

[7] aarc-community.org/policies/policy-development-kit

Einführung einer europaweiten ID

Forschende, die im Laufe ihrer Karriere mehrfach die Heimateinrichtung wechseln, haben aktuell keinen eindeutigen persönlichen Bezeichner (Identifier). Die Einführung einer nationalen EDU-ID ist ein Ansatz, Wissenschaftlerinnen und Wissenschaftlern diesen eindeutigen unveränderlichen Identifier zu geben. Die Vernetzung mit internationalen Partnern soll für eine – wenigstens europaweit – harmonische Lösung sorgen.

Viele EU-Projekte im Umfeld der European Open Science Cloud (EOSC) verwenden entweder die EGI-Lösung RCIAM[8], die EUDAT-Lösung[9] oder den von GEANT entwickelten Dienst eduTEAMS[10].

Der allgemein zu beobachtende Trend zeigt klar, dass bei modernen Diensten die Integration über OpenID Connect gegenüber SAML bevorzugt wird. X.509-Zertifikate für die Identifikation von Nutzenden werden an vielen Stellen (wie z.B. im WLCG) durch Token-basierte OpenID Connect-Lösungen ersetzt. Aufgrund der Möglichkeit, am SP-IdP-Proxy zwischen den verschiedenen Technologien zu vermitteln, gibt es theoretisch keine Web-Dienste, die nicht an eine moderne AAI angebunden werden können. Praktisch zeigt sich allerdings, dass Dienste allzu oft Probleme damit haben, die föderierten Attribute korrekt in die interne Logik zu übersetzen. Und selbst einfache Autorisierungsfunktionen, um den Zugriff zumindest rudimentär zu regeln, fehlen ebenfalls bei den meisten Diensten. Gerade im Praxisbetrieb zeigt sich, wie wichtig harmonisierte und standardisierte Attributdatensätze sowie Autorisierungsfunktionen am SP-IdP-Proxy sind.

Werkzeugentwicklung am SCC

Am SCC liegt der Fokus neben der Weiterentwicklung verschiedener Aspekte der Blueprint Achitecture auf der Entwicklung von sogenannten non-Web-Szenarien und der Unterstützung von Nutzendengruppen, die vorwiegend auf der Kommandozeile arbeiten. Ebenso werden Werkzeuge und Bibliotheken entwickelt, die serverseitige Anwendungen in die Lage versetzen, Token-basierte Protokolle zu verarbeiten.

Die Basis hierfür bietet der am SCC entwickelte oidc-agent[11], der mittlerweile Bestandteil zahlreicher Linux-Distributionen ist und auch für macOS und Windows bereitgestellt wird. Der oidc-agent implementiert eine zum SSH-Agenten analog konzipierte Methode, um jederzeit – und wenn möglich ohne weitere Nutzerinteraktion – aktuelle Access-Tokens zu erhalten. Dies macht die Nutzung von OpenID Connect für viele nachgelagerte Tools möglich.

Um auch aus Computing Jobs heraus oder auf länger laufenden virtuellen Maschinen dauerhaft Access-Tokens zu erhalten, wird bereits eine Erweiterung entwickelt. Das hierfür aufgesetzte mytoken-Projekt[12] ist aus der Erfahrung mit dem oidc-agent in einer Masterarbeit am SCC entstanden.

Die Tools mccli[13] und motley-cue[14] implementieren die Möglichkeit, weitere Protokolle mit OpenID Connect zu verwenden. Als erstes Beispiel wurde hier SSH gewählt. Das besondere an der Lösung ist, dass weder der SSH-Client noch der SSH-Server modifiziert werden müssen. Hierzu wurde in Zusammenarbeit mit dem Poznan Supercomputing and Networking Center (PSNC) in Polen ein PAM-Modul entwickelt, das Access-Tokens verwenden kann. Motley-cue ist ein weiterer serverseitiger Teil, der für das mapping von föderiertem Nutzerobjekt und lokaler Identität sorgt. Selbstverständlich können hierbei lokale Identitäten (unter anderem) mit der RegApp (alternativ: lokale Accounts oder LDAP) abgeglichen werden. Clientseitig kommt mccli zum Einsatz. Dieser wrapper klärt zunächst die serverseitige Autorisierung mit motley-cue ab, und ruft dann den lokal installierten SSH-Client mit den passenden Parametern auf[15]. Diese Lösung funktioniert unter Linux und macOS. Für Windows wird aktuell an einer Erweiterung des beliebten SSH-Clients PuTTY gearbeitet, und natürlich auch an der Integration des nativen Open-SSH-Clients unter Windows.

Ausblick

In Zukunft wird das Thema Self Sovereign Identities (SSI) auf der Agenda stehen. Hierbei stehen die Nutzenden im Mittelpunkt. Diese sammeln unterschiedliche Aussagen über ihre Identität (sogenannte verifiable credentials) und können dann selbst entscheiden, welche davon sie einem Dienst übermitteln wollen. Hiermit soll den Nutzenden die vollständige Kontrolle über die Herausgabe ihrer Daten gegeben werden, unter anderem, um den Anforderungen der DSGVO noch besser gerecht zu werden. Das SCC wird sich dabei darauf konzentrieren, dass der Schritt in Richtung SSI möglichst in Einklang mit den bereits aufgebauten Infrastrukturen erfolgen kann.

[8] aai.egi.eu
[9] b2access.eudat.eu
[10] eduteams.org
[11] github.com/indigo-dc/oidc-agent
[12] mytoken.data.kit.edu
[13] pypi.org/project/mccli
[14] pypi.org/project/motley-cue
[15] ssh-oidc-demo.data.kit.edu

The big picture of AAI

SCC is actively involved in the area of Authentication and Authorization Infrastructures (AAI) with two working groups and a number of activities in various projects. In this issue of SCC News, there are two articles that look at this topic from different perspectives. This article focuses on activities in the context of European projects that are establishing IT infrastructures for researchers throughout the EU. Here, the SCC is particularly involved in the development of AAI architectures and policies, but also in the development and integration of command line tools.

bwIDM2 – Security & Communities

Im zweiten Artikel zum großen Thema AAI berichten wir über das am 1. August gestartete Nachfolgeprojekt zu bwIDM – Föderiertes Identitätsmanagement der baden-württembergischen Hochschulen. Das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg fördert darin Universitäten und Hochschulen für eine Laufzeit von zwei Jahren. In den vergangenen Jahren sind die Anforderungen an sichere IT-Föderationen in der Wissenschaftslandschaft enorm gestiegen. Unter Berücksichtigung der aktuellen technischen Entwicklungen steht in diesem Projekt neben der hochschulartenübergreifenden Einbindung von Diensten unter anderem eine delegierbare Gruppen-/Rollenverwaltung für überregionale und nationale Communities im Fokus.

Matthias Bonn, Michael Simon, Ulrich Weiß

Die Hochschulen des Landes stellen eine Vielzahl verteilter Dienste und Ressourcen zur Verfügung, welche den Mitgliedern über die im Projekt bwIDM[1] erarbeiteten Zugangsmechanismen angeboten werden. Dieses in Baden-Württemberg erfolgreich betriebene föderative Identitätsmanagement hat auch über Landesgrenzen hinaus Vorbildcharakter für andere Bundesländer wie z.B. Rheinland-Pfalz und Nordrhein-Westfalen. Das Angebot heute existierender Landesdienste umspannt dabei ganz unterschiedliche Bereiche: von High Performance Computing (HPC) über Daten- und Ressourcenmanagement bis hin zu Online-Kursen. Die Dienste haben dabei sukzessive Anforderungen an die zentrale Softwareplattform definiert, die seit Projektende um viele Funktionalitäten erweitert worden ist.

Ausgangsbasis bwIDM und daran anschließende Entwicklungen

Bei der zentralen Komponente in bwIDM handelt es sich um die federführend am SCC entwickelte RegApp[2], ein Identitätsmanagementsystem, das die Voraussetzungen schafft, Personen aus unterschiedlichen Account-Quellen den sicheren Zugang zu Diensten bzw. Anwendungen jeglicher Art an jedwedem Standort zu ermöglichen. Die Kernfunktionalität der RegApp besteht darin, IT-Dienste und Anwendungen mit einer Authentifizierungs- und Autorisierungsinfrastruktur (AAI) zu verbinden. Dazu stehen vielfältige Integrationstechnologien zur Verfügung: Ein typischer Anwendungsfall wäre zum Beispiel, wenn ein Dienst nicht in der Lage ist, SAML-föderierte Identity Provider direkt anzusprechen. Hier kann das System Personenidentitäten vermitteln und als sogenannter Service-Provider-Identity-Provider-Proxy (SP-IdP-Proxy) das föderierte Authentifizierungsprotokoll verbergen. Weiterhin kann die Komplexität einer Föderation so leichter gehandhabt werden, da das System dem anzubindenden Dienst einen geeigneten Zugangspunkt anbietet sowie die Benutzenden zur Authentifizierung an deren Heimat-IdPs weiterleitet (Abbildung 1). Damit hat das Projekt bwIDM mit der IDM-Plattform RegApp bereits 2013 die wesentlichen Aspekte der auf Seite 7 detailliert beschriebenen AARC Blueprint Architecture in einem produktiv eingesetzten, hoch skalierenden System umgesetzt.

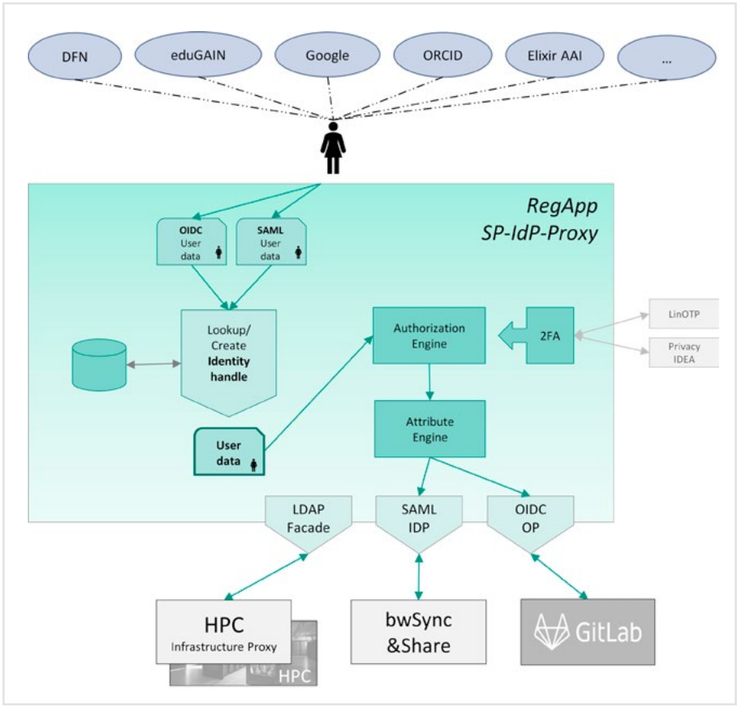


Abbildung 1: RegApp in bwIDM als SP-IdP-Proxy

[1] bwidm.de
[2] www.scc.kit.edu/dienste/regapp

Um die Möglichkeiten zur Anbindung von Diensten zu erweitern, bietet die RegApp Endpunkte sowohl für SAML als auch – als Neuentwicklung der letzten Monate – für das inzwischen vermehrt eingesetzte OpenID Connect (OIDC) an, das von vielen Diensten, freien sowie kommerziellen Produkten unterstützt wird. Die Anbindung von Diensten ist mit OIDC im Vergleich zu SAML einfacher zu handhaben, zudem ist das Protokoll in einigen Punkten flexibler bei der Integration von Dritt-Diensten. Die RegApp bildet dabei als Protokollübersetzungsdienst SAML Assertions auf OpenID Connect Claims ab. Zur Unterstützung lokaler Konten mit Benutzernamen/Passwort-Authentifizierung wird auch der LDAP-basierte Zugriff unterstützt. Diese sogenannte Infrastruktur-Proxy-Funktionalität kann mit der RegApp mit sehr geringem Aufwand realisiert werden, indem das System einfach zwischen Anwendung und föderierte Authentifizierung geschaltet wird. Somit werden alle typischen Single-Sign-On-Szenarien (SSO) für beliebige webbasierte Anwendungen abgedeckt.

Der Zugriff auf nicht-webbasierte Systeme mit Kommandozeilenschnittstellen mit erhöhten Sicherheitsanforderungen ist ein weiterer Anwendungsfall. Dazu können authentifizierte Personen über das Registrierungsportal die Liste aller verfügbaren Dienste einsehen und sich für diese registrieren. Je nach Dienst können spezialisierte Passworte gesetzt werden; etwa, wenn dies durch Policies erforderlich ist. Gleiches gilt für die üblicherweise verwendeten SSH-Schlüssel, die mit unterschiedlichen Profilen (Login oder Remote-Befehlsausführung) eingestellt werden können. Gerade diese vereinfachte und zentrale SSH-Key-Verwaltung bringt eine signifikante Erhöhung des Sicherheitslevels (siehe auch SCC News 01/2021: Innovative Multifaktor-Authentifizierung für sicheren HPC-Zugang).

Desweiteren kann ein Multifaktor-Authentifizierungs-Dienst (auf Basis von LinOTP) verwendet werden, um Transaktionsnummern zu prüfen. Das Portal bietet

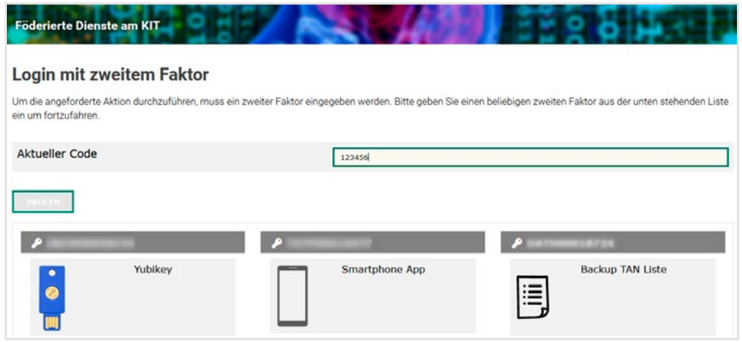


Abbildung 2: Multifaktor-Login auf der RegApp

dafür eine Self-Service-Registrierung von TOTP-basierten (Time-based One-time Password) Smartphone-Authenticator-Apps (Abbildung 2). So kann der Zugang zu allen an der RegApp angebundenen Diensten mit einem zweiten Faktor abgesichert werden, selbst wenn dies die jeweilige Heimateinrichtung nicht anbietet. Die in der RegApp zusätzlich realisierten Autorisierungsfunktionen auf

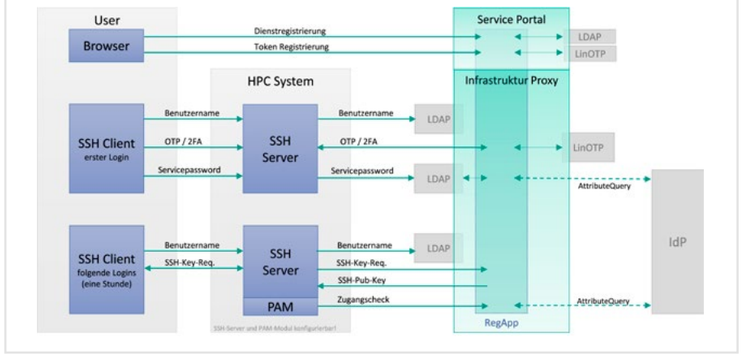


Abbildung 3: Ablauf eines HPC-Logins: Registrierung → Benutzername + Dienstpasswort + TOTP → SSH-Key

Basis von beliebigen Kontomerkmalen eröffnen Möglichkeiten, Dienstzugänge auf bestimmte Personenkreise einzuschränken. Das ist ein deutlicher Gewinn an Sicherheit, gerade wenn Dienste selbst derartiges nicht vorsehen.

Dies kann auch für den sicheren Login an Nicht-Web-Systemen genutzt werden: Beim initialen (Kommandozeilen-)Login auf beispielsweise einem HPC-System wird neben dem Dienstpasswort auch ein Zweifaktor-Code abgefragt und von der RegApp geprüft. Ist das erfolgreich, wird dort für eine begrenzte Zeitdauer der hinterlegte öffentliche SSH-Schlüssel zum Download freigeschaltet, so dass die folgenden Logins mit dem üblichen SSH-Mechanismus erfolgen können (Abbildung 3). Hierfür stellt die RegApp entsprechende REST-Schnittstellen bereit. Um diese beim Login anzusprechen, muss auf Seiten des HPC-Systems keine bestehende Software verändert oder ausgetauscht werden. Die dort üblichen Login-Module erlauben es, den oben beschriebenen sicheren Login durch reine Rekonfiguration auszuführen.

Das aktuell für die Landesdienste am KIT aufgebaute System besteht aus 13 Apache SSL-Proxies im Clustermodus und vollvermaschten Anwendungsservern, die hinter zwei Netzwerk-Lastverteilern

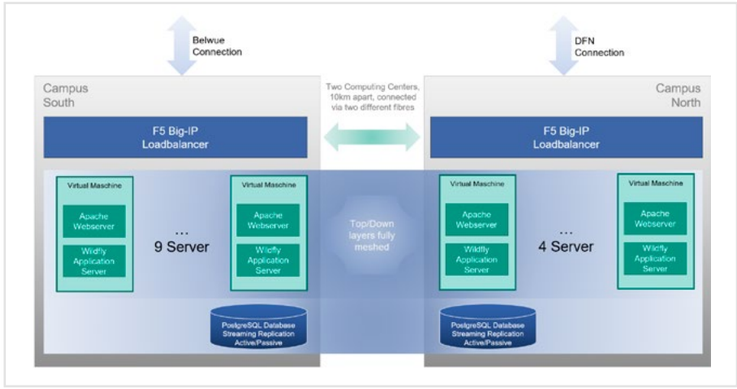


Abbildung 4: Hochverfügbares und skalierbares Betriebskonzept der RegApp

in beiden KIT-Standorten virtualisiert betrieben werden (Abbildung 4). Die verwendete PostgreSQL-Datenbank wird ebenfalls verteilt betrieben, so dass mehrere Redundanzstufen einen hochverfügbaren und entsprechend der anfallenden Last skalierbaren Betrieb ermöglichen.

bwIDM2 mit mehr Sicherheit und Communities Features

Das Projekt bwIDM2[3] widmet sich den gestiegenen Anforderungen an die IT-Sicherheit und berücksichtigt aktuelle technische Entwicklungen. Es schafft die Voraussetzungen zur hochschularten-übergreifenden Einbindung von Diensten

[3] www.bwidm.de/bwidm2

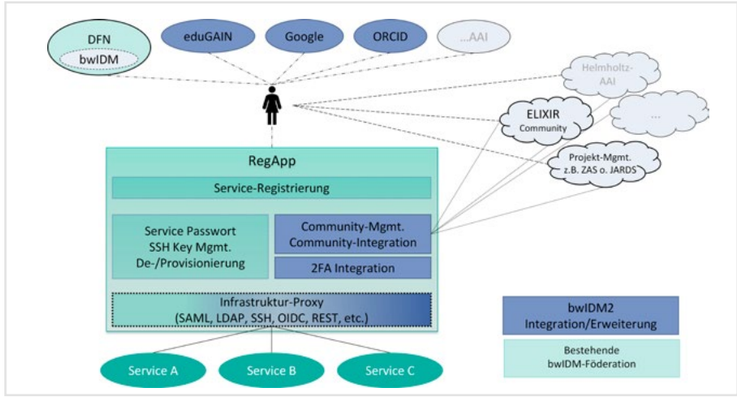


Abbildung 5: bwIDM2 – Architektur der RegApp im Kontext von Identitätsprovidern (IdP), Service Providern (SP), Authentifizierungs- und Autorisierungsinfrastrukturen (AAI) sowie der im Rahmen von bwIDM2 geplanten Erweiterungen zur Integration weiterer Föderationen und AAIs

Der Fokus des föderierten Dienstes bwIDM liegt auf dem verlässlichen Transport der für den sicheren Betrieb notwendigen Informationen und auf der Ertüchtigung der Softwareinfrastruktur. Die dienstspezifische Datenpflege, Zuständigkeiten, rechtliche Fragen und die Harmonisierung lokaler Ordnungen klären die lokalen Einrichtungen bzw. Dienstbetreiber. Dasselbe gilt für Aspekte und etabliert eine Gruppen-/Rollenverwaltung für überregionale und nationale Communities mit Delegationsmechanismen. Darüber hinaus werden Fachkonzepte zur Integration eines langfristigen, für das Forschungsdatenmanagement notwendigen Personenidentifikators erarbeitet.

Projektziele

Eine einfache und gesicherte Nutzung von Diensten, die an anderen Standorten angeboten werden, ist für den Betrieb von Wissenschaft und Lehre essentiell. Dabei soll ein gesicherter Dienstzugang mit Hilfe einer Zwei- oder Multifaktor-Authentifizierung (2FA/MFA) nach einer Evaluation bereits eingesetzter Technologien ermöglicht werden. Im Weiteren ist geplant, OpenID Connect als zusätzliche, über das etablierte SAML hinausgehende zweite Technologie für föderative Authentifizierung und Autorisierung zu evaluieren und zu unterstützen.

Eine zentrale Rolle spielt dabei die Verwaltung von Projekten/Communities, die dienstübergreifend genutzt werden kann und auch Personen ohne Account einer baden-württembergischen Landeseinrichtung beinhalten. Dabei stellt sich die Frage, wie externe Wissenschafts-Communities mit eigener AAI (vgl. S. 7) des Datenschutzes, die jeweils für die angeschlossenen Dienste im Sinne des jeweiligen Zwecks zu betrachten sind und dabei auch die bwIDM-Infrastruktur einschließen.

Die genannten Anforderungen sollen auch für kleinere Hochschulen, die bislang noch nicht an bwIDM teilnehmen, umsetzbar sein. Dies soll dem Konzept „Hilfe zur Selbsthilfe" folgend eine Umsetzung in allen Einrichtungen ermöglichen. Die enge Zusammenarbeit mit der Digitalen Hochschule NRW stellt sicher, dass sich Baden-Württemberg und Nordrhein-Westfalen in ihren jeweiligen IDM-Projekten bwIDM2 und IDM.NRW[4] abstimmen und alle Einrichtungen an den Erfahrungen anderer partizipieren. Des Weiteren sind diese Erfahrungen auch für das Nationale Hochleistungsrechnen (NHR) von Bedeutung, wo gemeinsam evaluiert wird inwieweit in die IDM-Lösung auch NHR-Ressourcen eingebunden werden können.

[4] idm.dh.nrw

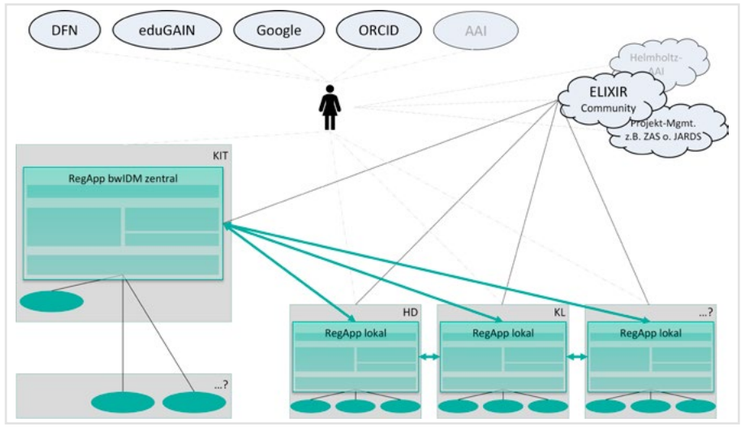


Abbildung 6: Föderierter Betrieb der bwIDM-Software und Interoperabilität mit anderen AAIs

unter Vermeidung von Dopplung der Identitäten und des Arbeitsaufwands eingebunden werden können. Ein weiteres Hauptziel besteht in der Konzeption und der Realisierung einer Projektgruppen- bzw. Communityverwaltung, mit der delegierbare Autorisierungsentscheidungen über die Nutzung von Diensten getroffen werden können. Abbildung 5 zeigt die geplanten Erweiterungen bzw. den geplanten Ausbau der vorhandenen Komponenten und Funktionen.

Auch kleinere Einrichtungen sollen in die Lage versetzt werden, Landesdienste zu nutzen bzw. eigene Dienste anbieten zu können, wobei hier primär der Bedarf nach einfachen und nachvollziehbaren Anleitungen zum Anschluss der eigenen Einrichtung an die bwIDM-Infrastruktur im Vordergrund steht. Zu diesem Zweck wird ein Betriebskonzept entwickelt, nach dem die dafür entscheidenden Komponenten sowohl bei den Einrichtungen selbst als auch zentral als Dienst betrieben werden können (Abbildung 6). Alle Entwicklungen und Ergebnisse werden zur Nutzung durch die Hochschulen Baden-Württembergs ausführlich dokumentiert, um so die Eintrittsbarrieren für kleinere Einrichtungen zu senken.

Nutzen und Mehrwert

Der Hauptnutzen der föderierten IDM-Lösung besteht klar in einem einheitlichen Zugang zu einem erweiterten Diensteportfolio. Ohne die Föderation würden die meisten Dienste nur einer eingeschränkten Personengruppe – nämlich der der betreibenden Heimatorganisation – zur Verfügung stehen. Die Erweiterung der Single-Sign-On-Protokolle sowie die MFA-Integration erweitern das mögliche Dienstspektrum und steigern gleichzeitig die IT-Sicherheit für alle Beteiligten erheblich. Für die Hochschulen wird insofern ein Mehrwert generiert, weil nicht jede Einrichtung Standarddienste betreiben muss, sondern diese – in Kombination mit der erweiterten AAI-Infrastruktur – kooperativ vorgehalten werden können. Ein immenser Vorteil für die Hochschulen ist die über Projektgruppen steuerbare flexible Autorisierung.

Weiterhin fordern allgemeine Entwicklungen kommerzieller Cloud-Lösungen für den Geschäfts- und Wissenschaftsbetrieb – wie Microsoft 365 bzw. Azure AD – solche ausgereiften AAI-Konzepte. bwIDM2 kann hier, gerade für kleinere Hochschulen, einen wichtigen Beitrag liefern. Die Allianz mit IDM.NRW bietet die Chance auf eine stärkere Verbreitung und Weiterentwicklung der Plattform auch außerhalb von Baden-Württemberg. Die Mitwirkung Nordrhein-Westfalens führt zudem zu einer Stärkung des Entwicklungsteams.

Für Endanwenderinnen und -anwender verbessert sich die Sicherheit bei der Nutzung von Diensten bei nur geringem Mehraufwand durch die einmalige Vorab-Registrierung von Hardware-Token oder Smartphone-Multifaktor-Apps. Zusätzlich werden immer mehr Dienste durch Single-Sign-On ohne Eingabe von persönlichen Zugangsdaten wie Nutzername und Passwort nutzbar sein. Für die Hochschulen erleichtern sich Arbeitsabläufe insofern, als dass für viele Dienste nicht selbst ein Installations-, Betriebs- und Wartungskonzept erstellt werden muss. Der Aufbau von lokalem Wissen darüber, welcher Standort für welchen Dienst verantwortlich ist und Unterstützung – insbesondere in Fehlerfällen – bieten kann, ist jedoch unabdingbar. Die Schaffung einer dienstübergreifenden Projekt-/Communityverwaltung in bwIDM macht die Entwicklung und Pflege einer lokalen Autorisierungssteuerung in vielen Fällen hinfällig.

bwIDM2 – Security & Communities

On August 1st 2021, the successor project to bwIDM started with the objective to further develop federated identity management for universities and institutes of higher education in the state of Baden-Württemberg. The two-year project is funded by the Ministry of Science, Research and the Arts of the state of Baden-Württemberg. In recent years, the requirements for secure IT federations in the academic landscape have increased enormously. Taking into account current technical developments, the focus of the project is on the integration of IT services across universities and the delegation of group and role management for communities that work across regional and at national level.

Einfach sicher: Zwei-Faktor-Authentifizierung

Die Haustür abschließen, vorausschauend fahren, in Corona-Zeiten Maske tragen: wir sind gerne bereit, aktiv für unserer Sicherheit zu sorgen. Der Schutz von Daten im digitalen Raum ist aber ebenso wichtig wie der Schutz von Haus und Hof. Es lohnt sich, hier etwas mehr Sorgfalt zu investieren und die Sicherheit der eigenen digitalen Identität durch aktuelle Mechanismen zu erhöhen. Aus diesem Grund sind die Zugänge zu wichtigen Anwendungen und Diensten am KIT, wie SAP oder Campus Management, mit einem „zweiten Faktor" abgesichert.

Karin Schäufele

Sicherheit und Komfort scheinen sich zu widersprechen, wird doch gesteigerte Sicherheit oft nur mit erhöhtem Aufwand erreicht. Dieser wird umso mehr akzeptiert, wenn die Anwendung zur Gewohnheit geworden ist oder sich an eine bestehende Gewohnheit „anhängt". Über viele Schutzmechanismen wie z.B. das Anlegen eines Sicherheitsgurts im Auto denken wir schon lange nicht mehr nach. Die Sicherheit der digitalen Identität wird jedoch, obwohl ebenso wichtig, oft nicht genug beachtet.

Für den Schutz der digitalen Identität am KIT (KIT-Account) stellt das SCC die „Zwei-Faktor-Authentifizierung am KIT" bereit. Zur Anmeldung an SAP, campus.kit.edu, bestimmten Logins am VPN und an den HPC-Rechnern sowie am Job-Portal des Personalservice muss zusätzlich zu Konto und Passwort ein zweiter Faktor – eine sechsstellige Ziffernfolge – eingegeben werden. Kontobezeichnung und Passwort sind den Nutzenden bekannt, die benötigte Ziffernfolge wird dagegen auf einem separaten Gerät jede Minute neu erzeugt. Um sich erfolgreich an einem Portal anmelden zu können, muss man also nicht nur das Passwort wissen, sondern auch das zusätzliche Gerät besitzen. Mit der Abfrage einer solchen Ziffernfolge wird das Erschleichen einer Identität erschwert und die Zugriffsicherheit auf die Daten erheblich erhöht. Ein weiteres Plus: über die genannten Dienste hinaus können weitere Anwendungen am KIT an das zentrale Single-Sign-On-Verfahren (Shibboleth) angebunden und dabei die Zwei-Faktor-Authentifizierung eingerichtet werden.

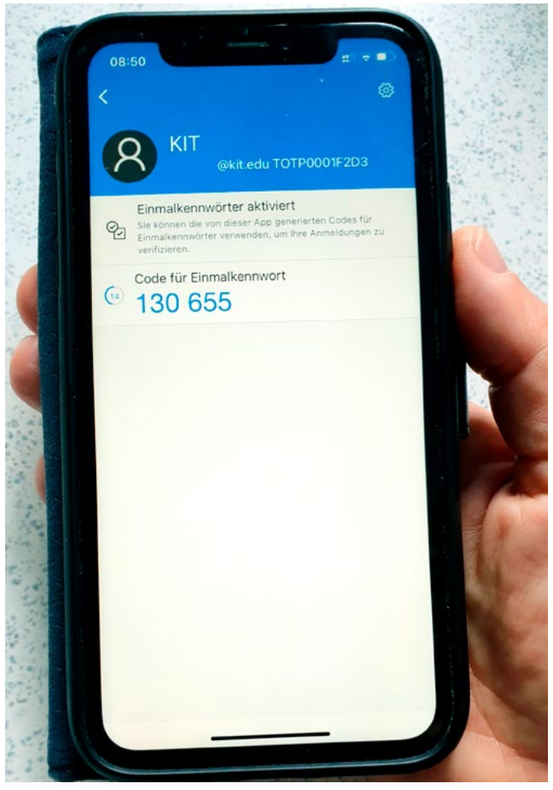


Abbildung 1: Denkbar einfach – die Generierung eines Einmalkennworts. Hier mit der Smartphone-App Microsoft Authenticator.

Ende 2017, als das KIT die Zwei-Faktor-Authentifizierung einführt (SCC-News 2/2017), wurden für die Mitarbeiterinnen und Mitarbeiter sogenannte KIT-Token mit Display ausgegeben. Heute sind davon knapp 8.500 im Einsatz. Diese Hardware-Token der ersten Generation kommen nun ans Ende ihres Lebenszyklus und müssten in großer Stückzahl ersetzt werden. Dieser Zeitpunkt ist auch eine gute Gelegenheit, auf einen Smartphone-Token umzusteigen, denn dieses bietet eine Reihe von Vorteilen. Zum einen kann ein vorhandenes Gerät genutzt werden, Herstellung und Beschaffung weiterer Hardware sind nicht nötig. Zum anderen geraten Hardware-Token leicht aus dem Blick, werden in Schreibtischschubladen vergessen oder gehen doch mal verloren. Im Gegensatz dazu stellen viele Smartphone-Nutzende bereits unbewusst sicher, es nicht zu verlieren. Für wen das Smartphone als zentraler Speicher für Kontakte und Korrespondenz sowie als Gedächtnisstütze in vielen Bereichen dient, der achtet auch besonders darauf, dass das Smartphone nicht abhanden kommt.

Dazu ist es weit verbreitet: In der Altersgruppe von 20–59 Jahren nutzen mehr als 92 % der Deutschen ein Smartphone[1]. Sie sind es gewohnt, Apps auf dem Mobiltelefon anzuwenden. Deshalb stellt

[1] de.statista.com/statistik/daten/studie/459963/umfrage/anteil-der-smartphone-nutzer-in-deutschland-nach-altersgruppe

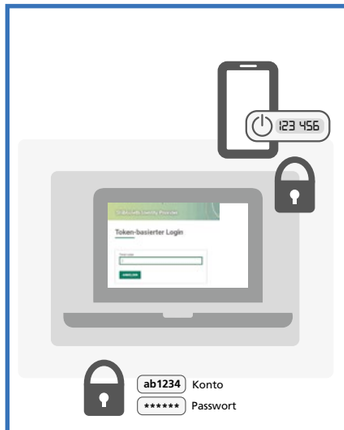
die Installation und Nutzung einer App gemäß RFC 6238 (z.B. Google Authenticator, Microsoft Authenticator, FreeOTP, Authy oder Sophos Authenticator) aus dem entsprechenden App-Store kaum eine Herausforderung dar (Abbildung 1).

Eine zusätzliche KIT-Software benötigt das Smartphone nicht, es muss lediglich in der Tokenverwaltung[2] im Self-Service-Portal registriert werden.

Heute sind am KIT bereits mehr als 10.300 Smartphone-Token registriert. Ein Großteil wird von Studierenden genutzt, die freiwillig auch den Zugang zur KIT-Lernplattform ILIAS mit dem zweiten Faktor schützen können. Aber gerade im Zusammenhang mit mobiler Arbeit und Telearbeit ist der Smartphone-Token auch für Mitarbeitende von Vorteil. Er kann entweder zusätzlich zum Hardware-Token eingerichtet werden, wenn dieser an einem zugriffssicheren Ort am Arbeitsplatz bleiben soll, oder ihn komplett ersetzen. Flexibles und sicheres Arbeiten ist damit einfach möglich. Der Smartphone-Token ist immer dabei. Wenn kein Smartphone genutzt werden kann oder dies einmal doch nicht zu Hand ist, hilft als Backup-Lösung, eine zuvor über my.scc.kit.edu generierte TAN-Liste (Abbildung 2) mit fünf Token-Codes, die beispielsweise im Portemonnaie steckt. Der SCC-Servicedesk gibt bei Fragen oder Problemen Hilfestellung und tauscht, wenn nötig, auch defekte Token aus.

Weitere Infos zur Zwei-Faktor-Authentifizierung:
www.scc.kit.edu/dienste/2fa.php

Abbildung 2: Immer mit dabei – eine Liste mit Token-Codes als Backup-Lösung



Für diejenigen, die einen KIT-Token mit Display nicht benutzen können, liegen USB-Token bereit. Dieser Token wird einfach in einen USB-Port gesteckt und anschließend von allen modernen Betriebssystemen als USB-Tastatur erkannt. Wird anschließend die runde Sensorfläche kurz berührt, trägt der Token automatisch den nötigen Token-Wert in das aktuelle Eingabefeld ein. Während USB-Token z.B. Sehgeschädigten eine echte technische Hilfestellung geben können, sind in vielen Fällen ‚normale' Hardware-Token nicht nötig und können durch Smartphone-Token ersetzt werden.

Simply secure: two-factor authentication
Locking the front door, driving with foresight, wearing a mask in Corona times: we're happy to take active steps to keep ourselves safe. But protecting data in the digital space is just as important as protecting your home. It's worth investing a little more care here and enhancing the security of your own digital identity with up-to-date mechanisms. For this reason, access to important applications and services at KIT, such as SAP or Campus Management, is secured with a "second factor".

[2] my.scc.kit.edu/token

Moderne DNS-Protokolle am SCC

Das Domain Name System, kurz DNS, ist ein weltweit verbreiteter, hierarchischer, dezentraler Verzeichnisdienst. DNS wird in nahezu allen Anwendungen verwendet, die mit anderen Systemen im Internet mittels Hostnamen kommunizieren – überwiegend zur Auflösung von Hostnamen auf die dazugehörigen IP-Adressen, aber auch für erweiterte Funktionen bei der Zustellung von E-Mails oder der Service Discovery. Das ursprüngliche DNS-Protokoll wurde bereits 1983 spezifiziert und im Laufe der Zeit um technische Aspekte für die Verbesserung der Sicherheit und den Schutz der Privatsphäre erweitert. Dieser Artikel beschreibt ausgewählte Erweiterungen des DNS und wie sie am SCC eingesetzt werden.

Julian Schuh

Das Domain Name System (DNS) ist ein weltweit verbreiteter, hierarchischer, dezentraler Verzeichnisdienst. Dieser wird überwiegend zur Auflösung von Hostnamen (`z.B. host.example.org`) auf die dazugehörigen IP-Adressen (z.B. `2001:db8:5ef:9c81:6247:ff7:ab6a:5894`) genutzt, spielt aber auch eine zentrale Rolle bei der E-Mail-Zustellung und bildet die Basis für weitergehende Funktionen wie Service Discovery und automatische Konfiguration von Anwendungen. DNS wird in nahezu allen Anwendungen verwendet, die mit anderen Systemen im Internet mittels Hostnamen kommunizieren. An der Abwicklung von DNS-Anfragen (z.B. die Auflösung eines Hostnamens zu einer IP-Adresse) sind heutzutage meistens zwei Typen von DNS-Servern beteiligt: autoritative DNS-Server, welche den Inhalt einzelner Domains kennen und bereitstellen, sowie DNS-Resolver (kurz Resolver), welche die hierarchische Namensauflösung für Endsysteme durchführen.

Authentizität von DNS-Daten

Das DNS-Protokoll, welches 1983 in Form von RFC 882 und RFC 883 spezifiziert wurde, wird dabei den heutigen Anforderungen an Sicherheit und Schutz der Privatsphäre nicht mehr gerecht. Aus diesem Grund wurden im Laufe der Zeit Erweiterungen des DNS-Protokolls spezifiziert, welche ausgewählte Sicherheitsaspekte verbessern sollen: So kann die Authentizität der DNS-Daten mittels Domain Name System Security Extensions (DNSSEC, RFC 4033, RFC 4034, RFC 4035) sichergestellt werden. Dies verhindert die Manipulation von DNS-Daten auf dem Weg zwischen autoritativem DNS-Server und den Resolvern. Die Voraussetzung hierfür ist, dass der Eigentümer einer Domain diese Domain mittels DNSSEC signiert, und dass der Resolver diese Signatur vor Auslieferung an den anfragenden Nutzenden validiert. Während der Einsatz von DNSSEC im Jahr 2011 für erste Top-Level-Domains ermöglicht wurde und 2022 für 1365 der insgesamt 1487 Top-Level-Domains möglich war, wird DNSSEC noch nicht flächendeckend eingesetzt: So waren im Juni 2022 nur ca. 1,5% der .de-Domains mittels DNSSEC gesichert[1].

Defizite bei Sicherheit und Schutz der Privatsphäre

Resolver werden üblicherweise vom lokalen Netzdienstleister betrieben und bereitgestellt, es gibt jedoch auch zentral im Internet betriebene, z. B. von Google[2] und Quad9[3]. Generell profitieren Nutzende davon, wenn sie einen geteilten Resolver benutzen, da dieser die Antworten im Cache zwischenspeichert, und damit die Last auf den autoritativen DNS-Servern sowie die Antwortzeiten zum Endgerät reduziert. Der Betreiber eines Resolvers kann alle Anfragen der Nutzenden einsehen und für seine Zwecke nutzen, daher sollte bei der Auswahl eines Resolvers neben Aspekten wie Geschwindigkeit und Zuverlässigkeit stets auch der Datenschutzaspekt berücksichtigt werden.

Da der Resolver DNS-Anfragen im Namen des Endsystems durchführt, muss ein Vertrauensverhältnis zwischen dem Nutzenden und dem Resolver bestehen. Dies ist auch dann relevant, wenn der Resolver die Validität der autoritativen Daten mittels DNSSEC überprüft und das Ergebnis der Prüfung an den Nutzenden übermittelt. Da das DNS-Protokoll an dieser Stelle keine Authentifizierung der Gegenstelle (also des Resolvers) vorsieht, und auch keine Maßnahmen zum Schutz der Authentizität der Daten anwendet, besteht hier unter anderem das Potential für einen Man-in-the-Middle-Angriff: Angreifende können sich in die Kommunikation zwischen Endsystem und Resolver einklinken und die DNS-Anfragen bzw. -Antworten an das Endsystem zu ihren Gunsten manipulieren. Damit können Nutzende z.B. auf einen von Angreifenden kontrollierten Server umgeleitet werden.

Standards zur Verbesserung der Sicherheit und zum Schutz der Privatsphäre

Um hier Abhilfe zu schaffen, wurden in den letzten Jahren Erweiterungen des DNS-Protokolls spezifiziert, welche die Authentizität eines Resolvers gegenüber einem Endsystem sicherstellen und die DNS-Anfragen und -Antworten verschlüsseln, sodass Netzbetreiber oder Angreifer mit den zuvor skizzierten Fähigkeiten keine Informationen mehr über die Anfragen der Endnutzenden an

[1] www.denic.de/en/know-how/statistics/monthly-statistics-of-de/
[2] dns.google
[3] www.quad9.net

den Verzeichnisdienst erhalten können: Hierbei handelt es sich um die Standards DNS over TLS (DoT, RFC 7858) und DNS over HTTPS (DoH, RFC 8484). Während die Nutzung von DoT und DoH primär zur Sicherung der DNS-Kommunikation zwischen Endsystem und Resolver eingesetzt wird, ist diese nicht darauf beschränkt: In einigen aktuellen Implementierungen von DNS-Server-Software kann DoT auch zur Absicherung anderer Transaktionen, z.B. zwischen zwei autoritativen Servern, verwendet werden. Beide Protokolle setzen eine Transportverschlüsselung basierend auf Transport Layer Security (TLS) ein – die gleiche Technologie, die auch beim verschlüsseltem Zugriff auf Webseiten mit HTTPS zum Einsatz kommt. Im Falle von DoT werden DNS-Anfragen direkt via TLS verschlüsselt; bei DoH werden DNS-Anfragen und -Antworten in einer verschlüsselten HTTPS-Transaktion übermittelt. Während DoT auf Protokollebene einfacher zu implementieren ist, hat DoH den Vorteil, dass ein außenstehender Angreifer oder Netzbetreiber nicht mehr ohne Weiteres zwischen DNS-Anfragen des Endsystems und dem Aufruf von Webseiten unterscheiden kann, da es sich in beiden Fällen aus Sicht des Netzwerks um einen mittels HTTPS verschlüsselten Datenverkehr handelt.

Resolver-Dienste am KIT

Das SCC betreibt die klassischen Resolver des KIT in einem hochverfügbaren, Anycast-basierten Aufbau. Im Zuge der Grunderneuerung des Resolver-Dienstes stellt das SCC zusätzlich DoT- und DoH-Resolver in einem Probebetrieb zur Verfügung. Der DoT- und DoH-Probebetrieb richtet sich dabei in einem ersten Schritt an einen technisch versierteren Kreis von Anwendenden, welche besonderen Wert auf die Absicherung ihrer Kommunikation und den Schutz der Privatsphäre legen. Durch das Bereitstellen eines abgesicherten Resolver-Dienstes durch das SCC wird für Anwendende mit entsprechenden Anforderungen die Möglichkeit geschaffen, DoT und DoH ohne Datenabfluss an die zuvor genannten externe Dienstleister zu nutzen; ein Ausweichen auf zuvor genannte externe Dienste ist nicht mehr notwendig. Interessierte des KIT können am Probebetrieb teilnehmen, indem sie bei einer kompatiblen Anwendung oder einem kompatiblen Endgerät den DoT-Resolver `dot.scc.kit.edu` eintragen bzw. den Endpunkt `https://doh.scc.kit.edu/dns-query` für DoH-kompatible Anwendungen nutzen. Jedoch ist das Risiko eines Man-in-the-Middle-Angriffs auch für Endgeräte, welche den herkömmlichen Resolver-Dienst nutzen, sehr gering, da die DNS-Anfragen von Endsystemen innerhalb des KITnet verbleiben. Unabhängig von der Art des Zugriffs – DoT, DoH oder herkömmliches DNS – wird bei Nutzung des Resolver-Dienstes des SCC die Privatsphäre der Nutzenden gewahrt, da Dritte keine Einsicht in die DNS-Anfragen der Endsysteme erhalten.

Android unterstützt die manuelle Konfiguration eines DoT-Resolvers, auf unterstützten Linux-Distributionen kann DoT beispielsweise mittels des Systemdienstes systemd-resolved genutzt werden. Das Protokoll DoH wird von Windows 10 und 11 nach manueller Konfiguration und verschiedenen Webbrowsern (Mozilla Firefox, Google Chrome, Microsoft Edge) nativ unterstützt. Die Konfiguration der Endsysteme unterscheidet sich dabei je nach eingesetzter Anwendung oder eingesetztem Betriebssystem[4].

Wie auch bei dem bereits bestehenden Resolver-Dienst werden die DoT- und DoH-Resolver im KITnet per Anycast angebunden: Der DoT- und DoH-Dienst wird jeweils am Campus Nord und Campus Süd bereitgestellt; für den Zugriff können jedoch immer die gleichen Hostnamen bzw. IP-Adressen verwendet werden. Der DoT- und DoH-Dienst wird nur via IPv6 bereitgestellt, was durch den vollständigen Rollout von IPv6 am KIT in 2021 möglich ist. Systeme, welche ausschließlich via IPv4 angebunden sind, können den Dienst daher nicht nutzen. Weiterhin steht der Dienst aktuell nur für die Nutzung aus dem KITnet heraus bereit.

[4] www.scc.kit.edu/sl/dot-doh

Ausblick

Nach einem erfolgreichen Probebetrieb sollen die DoT- und DoH-Resolver in den Produktivbetrieb überführt werden. Darüber hinaus wird die Entwicklung des im Mai 2022 verabschiedeten Standards DNS over QUIC (DoQ, RFC 9250) beobachtet und soll – vorausgesetzt der eingesetzte Software Stack unterstützt das – ebenfalls als Dienst verfügbar gemacht werden. Neben der Bereitstellung des gesicherten Resolver-Dienstes an sich werden weiterhin Möglichkeiten evaluiert, eine automatische Nutzung der DoT- oder DoH-Resolver durch Endsysteme zu ermöglichen. Hierfür wird der Status und die Implementierung des sich in Entwicklung befindlichen IETF-Standards „Discovery of Designated Resolvers“ beobachtet. Auf Seite der autoritativen DNS-Dienste plant das SCC das Ausrollen von DNSSEC, um damit die DNS-Daten der eigenen Domains abzusichern.

Modern DNS protocols at SCC

The Domain Name System (DNS) is a worldwide, hierarchical, decentralized naming system. DNS is used in almost all applications that communicate with other systems on the Internet by means of hostnames - predominantly for resolving hostnames to the associated IP addresses, but also for advanced use-cases like the delivery of e-mail or service discovery. The original DNS protocol was specified as early as 1983 and has been extended by new standards that aim to enhance the security of DNS and protect the privacy of users. SCC operates the classical resolvers at KIT and, in the course of the overhaul of the resolver service, provides DNS over TLS (DoT) and DNS over HTTP (DoH) resolvers in a trial operation.

DoT: dot.scc.kit.edu
2a00:1398::53:853:1
2a00:1398::53:853:2
DoH: https://doh.scc.kit.edu/dns-query

Mit „Future Technologies“ zu den HPC-Systemen von morgen

Die Welt des High Performance Computing wurde in den letzten Jahren von zwei Hardware-Architekturen dominiert: x86-Prozessoren von Intel/AMD und GPU-Beschleuniger von NVIDIA. Die Potentiale alternativer Architekturen wurden hingegen kaum genutzt. Mit der „Future Technologies Partition“ soll dies im Rahmen von NHR@KIT konsequent und umfassend geändert werden.

Simon Raffeiner, René Caspart

Über drei Jahrzehnte hinweg bot sich beim Blick auf die jeweils aktuellen Supercomputer ein äußerst buntes Bild. Allein unter den Systemen der Vorgängereinrichtungen des KIT, der Universität Karlsruhe und des Forschungszentrums Karlsruhe, finden sich in den 1980er und 1990er Jahren die unterschiedlichsten Prozessorarchitekturen wie beispielsweise MasPar, IBM POWER, Fujitsu Siemens S600, NEC VPP und weitere. Fast jeder der großen Hardwarehersteller produzierte bis in die frühen 2000er Jahre hinein eigene Prozessoren. Teilweise wurden ganze Produktlinien ausschließlich für den Einsatz in Supercomputern entwickelt. Das ist nicht verwunderlich – am obersten Ende der Leistungsskala fanden und finden sich immer schon sehr komplexe und kostspielige Sonderanfertigungen. Die Hersteller operieren hier per Definition am Rand des technisch Machbaren.

Von der Vielfalt zur „Monokultur“

Mit dem zunehmenden technischen Fortschritt und der starken Konkurrenz zwischen den Herstellern stiegen Aufwand und damit Kosten für die Weiterentwicklung der Hard- und Software sprunghaft an. Statt teurer Sonderanfertigungen setzten sich zunehmend sogenannte „Commodity-Cluster“ durch, die aus kostengünstigen Standardkomponenten zusammengesetzt wurden. Hersteller wie Intel und AMD, die mit ihren Prozessoren auf Basis der x86-Architektur schon lange mehr als nur ein Marktsegment bedienten, konnten ihre Entwicklungskosten besser verteilen und durch Massenproduktion andere Anbieter verdrängen. Statt proprietärer Netzwerktechnologien kommen in Commodity-Clustern offene Standards wie Ethernet oder InfiniBand zum Einsatz, statt spezialisierter Betriebssysteme fast immer das quelloffene Linux. Viele Hersteller wandten sich in der Folge von ihren Eigenentwicklungen ab, und es entstand eine nunmehr fast zehn Jahre andauernde Monopolsituation. Die halbjährlich erscheinende Liste der 500 schnellsten Rechner weltweit (Top500[1]) weist bereits im Juni 2010 nur noch 49 Systeme aus, die nicht auf x86-Prozessoren von Intel oder AMD basieren. Bis Juni 2022 war diese Zahl sogar auf 18 gesunken. Ähnlich sieht es bei Systemen mit zusätzlich installierten Beschleuniger-Chips aus: Im Juni 2012 waren 53 von 58 mit Chips von NVIDIA ausgestattet, im Juni 2022 154 von 168. Auch der im Juni 2021 eingeweihte „Hochleistungsrechner Karlsruhe“ (HoreKa) ist ein Intel/NVIDIA-System.

Die „kambrische Explosion“ der Architekturen

Monokulturen haben durchaus nicht nur Nachteile. Anwendungen müssen nicht mehr für jedes System neu angepasst, optimiert und validiert werden. Die Pflege unterschiedlichster Versionen für verschiedene Architekturen entfällt. Wie in allen anderen Bereichen auch, belebt Konkurrenz aber das Geschäft. Maßgeschneiderte und alternative Architekturen versprechen nicht nur ökonomische Vorteile, sondern können je nach Anwendungsfall auch eine höhere Leistung, eine bessere Energieeffizienz oder andere Vorteile bieten. Insbesondere im weiterhin wachsenden Marktsegment der Künstlichen Intelligenz (KI) und des Machine Learning (ML) wurde in den letzten Jahren eine ganze Reihe von Startups gegründet, die speziell auf diese Anwendungsfälle optimierte Chips auf den Markt gebracht haben. Gleichzeitig drängt die früher hauptsächlich in embedded systems und Mobilgeräten eingesetzte Arm-Architektur in den Markt für Standardserver. Mittlerweile haben alle größeren Cloud- und Hardware-Anbieter Arm-Server im Programm, auch NVIDIA plant den Einstieg in die Produktion eigener Arm-Prozessoren. Mit RISC-V steht zusätzlich eine weitere, offene Architektur in den Startlöchern, die unter anderem von der European Processor Initiative[2] für zukünftige Exascale-Systeme auf Basis von in Europa entwickelten und hergestellten Prozessoren vorangetrieben

Eines der Racks der Future Technologies Partition (FTP). (Foto: Simon Raffeiner)

[1] www.top500.org

[2] www.european-processor-initiative.eu

wird. Man kann also – was das zeitgleiche Auftreten neuer Prozessorarchitekturen angeht – derzeit fast schon von einer „kambrischen Artenexplosion“ sprechen.

NHR@KIT bietet Technologien der Zukunft

Die Potentiale alternativer Architekturen können nur dann ausgeschöpft werden, wenn Nutzende und Betreiber diese auf einfache Art und Weise unter realen Bedingungen evaluieren können. Das SCC hatte daher bereits in den letzten Jahren immer wieder auch einzelne solcher Systeme beschafft. Unter der Bezeichnung „Future Technologies Partition“ (FTP) wurde das Konzept als Teil des Nationalen Hochleistungsrechenzentrums erweitert und verstetigt.

Die FTP ist ein Hardware- und Softwaretestbett, in dem Forschende und Betreiber die Potentiale neuer, innovativer und disruptiver Hardware frühzeitig testen und ihre Anwendungen darauf vorbereiten können. Im Gegensatz zu den großen Produktivsystemen wird die FTP nicht nur alle fünf Jahre erweitert, sondern meist mehrfach pro Jahr. NHR@KIT ist aktuell der einzige Standort im NHR-Verbund, der ein solches Konzept konsequent umsetzt.

Derzeit stehen in der FTP unter anderem Systeme mit MI100-Beschleunigern von AMD, Arm-Prozessoren (A64FX, Ampere Altra), KI-Beschleunigern von Graphcore oder auch neuartigen All-Flash-Datenspeichersystemen zur Verfügung. Weitere Systeme sind bereits in Beschaffung. Im Rahmen von Partnerschaften erhält NHR@KIT von den Herstellern auch frühzeitig Zugriff auf kommende Technologien. Beispielsweise konnte das SCC als einzige Institution weltweit ein Cluster aus NVIDIA HPC Development Kits beschaffen, und das KIT erhielt als erste akademische Einrichtung in Europa ein Graphcore POD16 System (s. SCC-News 2/2021).

Die Übertragbarkeit der in der Future Technologies Partition gewonnenen Erkenntnisse auf die in den Produktionssystemen gewohnte Umgebung ist von großer Wichtigkeit.

Die Softwareumgebung ist daher soweit wie möglich identisch zu der auf den Produktivsystemen, auch der Zugriff auf die gewohnten Dateisysteme von HoreKa und die Large Scale Data Facility (LSDF) ist möglich. Zudem werden die Systeme so weit wie möglich symmetrisch ausgelegt. Wenn es etwa ein System mit einer x86-CPU und NVIDIA-GPUs gibt, dann gibt es auch eines mit einer Arm-CPU und NVIDIA-GPUs. Um die aktuelle Matrix der Systeme zu vervollständigen, wurde daher beispielsweise auch ein Arm-System mit AMD-GPUs beschafft – eine Kombination, die so sonst noch nirgends im Einsatz ist.

Ein Gewinn für alle Seiten

Im Idealfall sollen die Nutzenden sich rein auf die Portierung, Optimierung und Vermessung ihrer Anwendungen konzentrieren können. Bis dahin ist es aber oft ein weiter Weg, der manchmal schon auf den untersten Ebenen des Systems beginnt.

Beispielsweise kann derzeit von den Arm-Systemen nur über Umwege, mittels NFS, auf das parallele Dateisystem von HoreKa zugegriffen werden, da IBM noch keine Version des Spectrum Scale-Clients für Arm-Systeme anbietet. Nicht alle von HoreKa gewohnten Dateisystem-Kommandos funktionieren daher wie gewohnt, was die Nutzenden entsprechend berücksichtigen müssen. Nach Gesprächen mit dem Hersteller ist eine Arm-Version von Spectrum Scale mittlerweile offiziell geplant. Von der Vorarbeit des SCC profitieren damit später auch andere Betreiber weltweit.

Für die Unterstützung der Nutzenden bei der Portierung, Evaluierung und Vermessung ihrer Anwendungen stehen die im Rahmen der Helmholtz-Programmatik und NHR@KIT am SCC etablierten Support-Strukturen zur Verfügung.

Diese bestehen unter anderem aus den „Simulation and Data Life Cycle Labs“ (SDL) (Seite 26) und dem „Software Sustainability and Performance Engineering Team“ (SSPE). Die Future Technologies Partition ist auch in weitere Dienste wie etwa die Continuous Integration-Infrastruktur integriert. Auch bei größeren Beschaffungen, wie der für 2023 geplanten zweiten Phase von HoreKa, fließen die Erfahrungen aus der FTP ein. Man darf gespannt sein, welche der neuen, disruptiven Technologien dann ihren Auftritt in größerem Umfang feiert.

Mehr Informationen zur Future Technologies Partition sind zu finden unter nhr.kit.edu.

With "Future Technologies" to the HPC systems of tomorrow

With the so-called “Future Technologies Partition”, a second operating environment in addition to the HoreKa supercomputer was established. Through this hardware and software testbed, users gain access to promising, disruptive technologies that are on their way to market maturity or have not yet achieved sufficient market penetration, and are therefore not represented in larger production systems.

The Future Technologies Partition enables users to test their applications on as many platforms and architectures as possible, to optimize them for these systems and thus make their software fit for the future.

The testbed includes a wide range of platforms and architectures, among them processors and GPU-based accelerators from all well-known manufacturers and architectures (Intel CPUs, AMD CPUs/GPUs, Fujitsu/Ampere Armv8 CPUs, NVIDIA GPUs). This offering is expanded continuously.

More information about the Future Technologies Partition can be found at nhr.kit.edu

Beschleunigung von numerischen Simulationen mithilfe moderner KI-Methoden

Numerische Simulation ist ein wichtiger Baustein in zahlreichen Industrie- und Forschungsanwendungen. Obwohl die verfügbare Leistung moderner Hochleistungsrechner immer weiter steigt, sind manche hochkomplexe Systeme jedoch nach wie vor außer Reichweite herkömmlicher numerischer Simulationen. Hier können KI-basierte Assistenzsysteme die erforderlichen Berechnungen enorm beschleunigen. So kann mit der am SCC entwickelten Software Suite Kinetic Transport Solver for Radiation Therapy (KiT-RT) die Strahlentherapieplanung in der Onkologie optimiert werden.

Steffen Schotthöfer

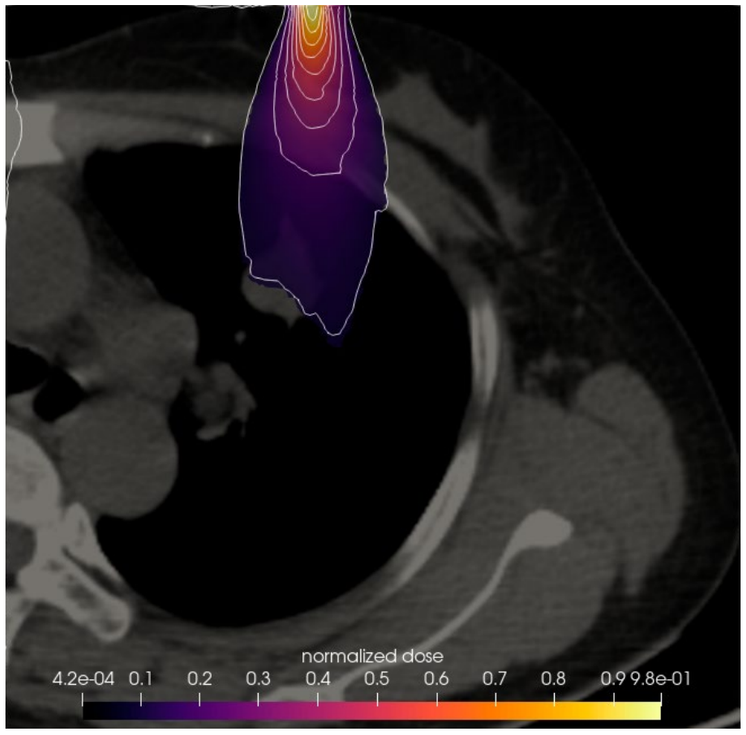


Abbildung 1: Simulation der Strahlendosis bei der Bestrahlung einer Lunge

Numerische Simulation hochkomplexer Systeme ist ein wichtiger Baustein einer digitalisierten Industrie und Wissenschaft, z.B. in Form von Simulationen von Strömungen, Strahlungstransport oder Wetter und Klima. Generell gilt, je feiner aufgelöst die Simulation sein muss und je größer die räumliche Ausdehnung des simulierten Phänomens ist, desto höher ist der Rechenaufwand.

Strahlentransport in der Medizin

Bestimmte Anwendungen, wie die Strahlentherapieplanung in der Onkologie, erfordern zum einen eine Simulation der schwierig zu berechnenden Interaktion der Strahlung und dem Körper des Patienten. Um die dafür notwendigen Patienteninformationen zu gewinnen, werden Computertomographiebilder der zu bestrahlenden Region zugrunde gelegt. Eine numerische Simulation auf Basis dieser Bilder erfordert jedoch hoch aufgelöste Rechengitter. In jeder Gitterzelle wird die lokale Strahlungsdosis und der Strahlentransport zu Nachbarzellen berechnet (Abbildung 1). Niedrige Fehlertoleranzen und schnelle Berechnungszeiten sind erforderlich, um dem ärztlichen Fachpersonal zeitnah bei der Therapieplanung zu helfen.
Dies stellt numerische Verfahren vor eine Herausforderung.

Auf Deep Learning basierte Hilfsmodelle

Die Simulation von Strahlentransport erfolgt durch die Berechnung der Boltzmanngleichung, einer Differenzialgleichung mit einer großen Zahl zu berechnender Variablen. Numerische Methoden zur Berechnung dieser Gleichung bestehen aus einem Ensemble von Berechnungstechniken, von denen manche sehr effizient sind, andere jedoch bis zu 90% der gesamten Berechnungszeit der Simulation benötigen.

Hier können Deep Learning-Methoden, eine Unterdisziplin von künstlicher Intelligenz (KI), helfen, die Simulationszeit um bis zu 85% zu senken. Somit können nicht nur Zeit, sondern auch Ressourcen gespart werden.

Die im Deep Learning eingesetzten neuronalen Netze werden speziell konstruiert, um physikalische Gesetze wie Masseer-

haltung und Positivität der Lösung einzuhalten. Somit kann eine akkurate und schnelle Simulation geliefert werden.

Hybride Software

Neuronale Netze können besonders effizient auf Grafikprozessoren (GPUs) implementiert werden, während Differenzialgleichungslöser häufig für herkömmliche Prozessoren (CPUs) programmiert werden. Hybride numerische Methoden, die sowohl traditionelle Techniken als auch neuronale Netze einsetzen, stellen neue Herausforderungen an die Softwarearchitektur.

Der Kinetic Transport Solver for Radiation Therapy, kurz KiT-RT (Abbildung 2), ist ein Open Source Softwarepaket und implementiert eine hybride Architektur, die CPU-Differenzialgleichungslöser und GPU-basierte neuronale Netze verbindet.

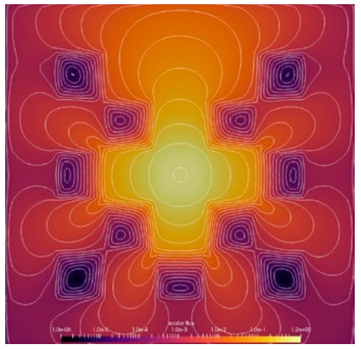


Abbildung 2: Strahlentransportsimulation mit KiT-RT. Eine Strahlungsquelle (hell, Mitte) ist umgeben von Bremsstäben (dunkel)

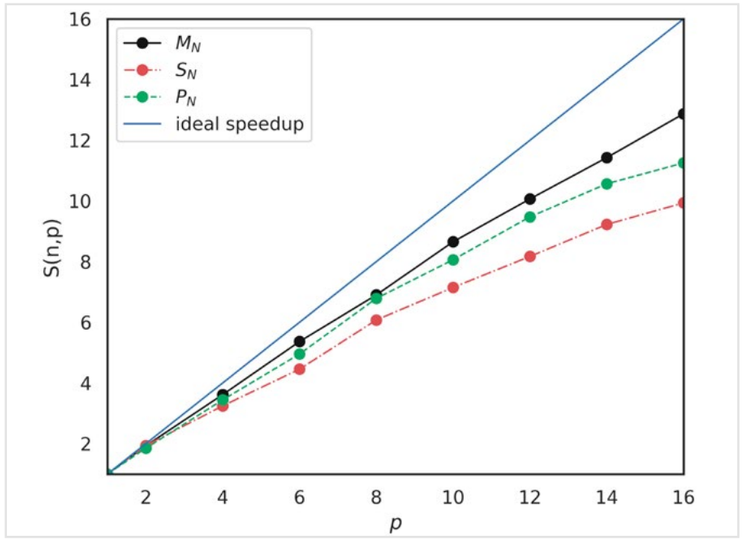


Abbildung 3: Skalierbarkeit verschiedener Löser der Software KiT-RT. Die X-Achse beschreibt die Anzahl der Prozessorkerne und die Y-Achse die resultierende Beschleunigung. Lineare Skalierbarkeit ist das theoretische Optimum.

Das Softwarepaket ist nach dem Steckkastenprinzip gebaut. Es besteht aus einer Vielzahl einfach zu handhabender Module, mit deren Hilfe numerische Löser für Strahlentherapie und Strahlentransport gebaut werden können. Der Simulator KiT-RT ist mit modernen Programmiertechniken parallelisiert (Abbildung 3).

Die Software Suite KiT-RT (Kinetic Transport Solver for Radiation Therapy) unterstützt hybride, Deep Learning-basierte Methoden und steht auf der Plattform Github unter »»» **github.com/CSMMLab/KiT-RT** zum Download bereit.

Acceleration of numerical Simulation using modern AI methods

In recent years, numerical simulation of complex systems has enjoyed tremendous traction due to eased access to high performance computing resources. Nevertheless, some large-scale or fine-grained physical systems are still prohibitively expensive to compute. Examples reach from ultra-large-scale weather simulation, fine-grained fluid dynamics and radiation transport. The latter is applied in medical radiotherapy planning, where accuracy as well as planning speed is crucial.

Radiation transport equations, consist of multiple parts, of which very few take the majority of the simulation time, consuming up to 90% of all necessary computational resources. Specialized neural networks can be constructed to bypass these computation heavy modules, while obeying all physical laws required to obtain an accurate simulation.
The open-source software KiT-RT implements this approach and provides an easy to use, dynamically extendable simulation suite for radiotherapy planning.

Von gekrönten Buchstaben und metaphorischen Wörtern

Altägyptische Pyramiden, aristotelische Handschriften oder frühneuzeitliche Sprachlehrwerke – die Forschungsgegenstände, die das SCC gemeinsam mit Kolleginnen und Kollegen aus den Geisteswissenschaften untersucht, waren seit jeher unglaublich vielfältig und spannend. Seit Anfang des Jahres 2022 versprechen nun das neue BMBF-Projekt, Materialisierte Heiligkeit' sowie der neue Sonderforschungsbereich 1475, Metaphern der Religion' erstmalige und interdisziplinäre Einblicke in die faszinierenden Objekte und Fragestellungen der Judaistik und Religionswissenschaft.

Danah Tonne, Laura Frank und Vandana Jha

„Gott würfelt nicht!"[1] – diesen oder einen ähnlich markigen Ausspruch hat bestimmt schon fast jeder gesehen, sei es auf T-Shirts, Tassen oder im neuesten Meme. Was dabei aber den wenigsten bewusst sein dürfte: vielfach wird dabei das Stilmittel der Metapher eingesetzt, um große, geradezu unfassbare Weltzusammenhänge auf den Punkt zu bringen. Wie auch sonst sollte man religiöse Erfahrungen beschreiben oder gar erklären, wenn ihre Grundlage – das Transzendente – niemals wörtlich artikuliert werden kann? Es ist also nicht verwunderlich, dass Religion und Metaphorik über Kulturen und Zeiten hinweg sehr häufig Hand in Hand gehen.

Forscherinnen und Forscher vermuten, dass Religion seit Anbeginn der Menschheit existiert, und versuchen schon lange, ihren Geheimnissen auf die Spur zu kommen. Seit Anfang des Jahres 2022 ergänzen nun zwei neue Projekte mit Beteiligung des SCC das vielfältige Forschungsspektrum und verfolgen dabei einen digitalen Ansatz zur interdisziplinären Untersuchung von mittelalterlichen Torarollen und Metaphern in der religiösen Sinnbildung. Doch was haben Krönchen auf hebräischen Buchstaben oder Lichtmetaphern in Forumsbeiträgen eigentlich mit Forschungsdatenmanagement zu tun?

Mittelalterliche Torarollen im Blick

In dem vom Bundesministerium für Bildung und Forschung (BMBF) geförderten neuen Forschungsprojekt ‚Materialisierte Heiligkeit' steht die jüdische, rituell reine Torarolle als zentrales Forschungsobjekt im Fokus. Aus unterschiedlichen Perspektiven sollen mittelalterliche Torarollen des 11.–15. Jahrhunderts untersucht und analysiert werden – mit dem Ziel, einen umfassenden digitalen Wissensspeicher dieses spannenden Forschungsgegenstandes zu erstellen. Gemeinsam mit Judaistinnen und Soziologinnen der Freien Universität Berlin (FU Berlin) und Materialforscherinnen der Bundesanstalt für Materialforschung und -prüfung (BAM) bilden Informationswissenschaftlerinnen des SCC das interdisziplinäre Team des auf vier Jahre ausgelegten Projekts.

Die rituell reine Torarolle ist ein bis heute handschriftlich verfasstes Werk, das die ersten fünf Bücher der hebräischen Bibel enthält (Abbildung 1). Dabei verkörpert die bis zu 90 Meter lange Torarolle das Wort Gottes und stellt als ‚materialisierte Heiligkeit' bei der traditionellen Lesung den zentralen

Abbildung 1: Magdeburger Torarolle aus dem 14. Jahrhundert[2]

[1] Basierend auf einem Satz aus einem der überlieferten Briefe Albert Einsteins: „Jedenfalls bin ich überzeugt, daß der nicht würfelt."
[2] Cod. Guelf. 148 Noviss. 2°, Herzog August Bibliothek Wolfenbüttel, diglib.hab.de/?db=mss&list=ms&id=148-noviss-2f

Gegenstand des synagogalen Gottesdienstes dar. Diese Heiligkeit spiegelt sich nicht nur in der klar festgelegten Beschaffenheit des Pergamentes und der verwendeten Tinte wider, sondern auch der Prozess des Schreibens unterliegt strengen Regulierungen, um die rituelle Reinheit zu gewährleisten.

Ein Gegenstand – vier Perspektiven

In den vier Modulen des Projektes wird das handschriftliche Wunderwerk Torarolle aus unterschiedlichen Perspektiven beleuchtet.

Modul I: Schreiberliteratur
Ein weites Spektrum an Regelwerken diskutiert die ‚ideale' Gestalt und Beschaffenheit der Torarolle. Dabei spielen sowohl rabbinische Werte als auch individuelle und persönliche Einstellungen eine Rolle.

Modul II: Materialwissenschaft
Durch verschiedene Techniken und Analysemethoden der Materialwissenschaft können Torarollen zeitlich und örtlich bestimmt werden. Zusätzlich können unterschiedliche Schichten des Pergamentes und variierende Tinten auf Retuschen und nachträgliche Bearbeitungen hinweisen.

Modul III: Schriftanalyse
Das Schriftbild unterliegt ebenfalls strengen Regeln, Verzierungen der Buchstaben sind nur vereinzelt erlaubt und klar definiert. Dennoch finden sich Ausnahmen in Form von verschnörkelten und ‚gekrönten' Buchstaben (Abbildung 2), deren Vorkommen und Bedeutung erforscht werden soll.

Modul IV: Schreiberinnen und Schreiber der Gegenwart
Die Bedeutung und Auslegung dieses besonderen Berufs im Wandel der Zeit soll durch Interview- und Film-Material erfasst und dokumentiert werden.

Abbildung 2: Zwei lamed mit ungewöhnlichen Fähnchen am oberen Ende (oben) und auf ungewöhnliche Weise gekrönte Buchstaben schin, pe und tet (unten) aus Annett Martini: Torarollen, in: Mathias Kluge (Hg.), Mittelalterliche Geschichte. Eine digitale Einführung (2021). URL: mittelalterliche-geschichte.de/martini-annett-01

Das SCC steht dabei mit umfangreicher Erfahrung im Forschungsdatenmanagement an der Seite der Forscherinnen. Im Fokus liegt vor allem das modulübergreifende digitale Repositorium zur zentralen Speicherung aller gesammelten Daten und Metadaten. Verschiedene Werkzeuge – etwa ein Annotationstool zur Auszeichnung schriftlicher Besonderheiten oder ein Vokabulareditor zur Vereinheitlichung der Nomenklatur – erlauben den Zugang und die weitere Analyse der Daten. Miteinher geht die Konzeption und Implementierung einer ‚Virtuellen Torarolle', welche erstmals die Erfassung, Analyse und Verknüpfung unterschiedlicher Torarollen ermöglicht.

Ein neuer Sonderforschungsbereich erblickt das Licht der Welt

Auf den ersten Blick erscheinen die Forschungsgegenstände gänzlich anders, doch der infrastrukturelle Kern wird auch im neuen Sonderforschungsbereich (SFB) 1475 ‚Metaphern der Religion: Religiöse Sinnbildung in sprachlichen Prozessen' zur Anwendung gebracht. Der seit dem 1.1.2022 erstmals von der Deutschen Forschungsgemeinschaft (DFG) geförderte SFB stellt Metaphern in den Mittelpunkt eines interdisziplinären Forschungsverbundes.

Angesiedelt an der Ruhr-Universität Bochum (RUB), soll der religiöse Gebrauch von Metaphern über Zeiten und Kulturen hinweg verstanden und methodisch erfasst werden. Die beteiligten Teilprojekte sind in drei Sektionen organisiert, die sich an den grundlegenden sprachlichen Domänen des Physischen, des Psychischen und des Sozialen orientieren. Inhaltlich werden eine Vielzahl von Schriften aus Christentum, Islam, Judentum, Zoroastrismus, Jainismus, Buddhismus und Daoismus untersucht, die aus Europa, dem Nahen und Mittleren Osten sowie Süd-, Zentral- und Ostasien stammen und die Zeitspanne von 3000 v. Chr. bis heute umfassen. Erstmals werden auf diese Weise vergleichende Studien in einzigartigem Umfang ermöglicht.

Viele Perspektiven – ein gemeinsamer Ansatz

Der Sonderforschungsbereich setzt ganz bewusst den Schwerpunkt auf digitale Methoden, um den Austausch zwischen den Teilprojekten zu fördern und innovative Ansätze zur Identifizierung, Analyse und Visualisierung der Metaphernverwendung in religiösen Texten zu ermöglichen. Die dafür notwendige Infrastruktur unterstützt also nicht nur die einzelnen Forschenden, sondern fördert die konzeptionelle Integration des SFB. Dazu gehört beispielsweise eine gemeinsame methodische Basis – wie etwa dieselbe Definition von Metapher in allen Teilprojekten – oder ein gemeinsamer sprachunabhängiger Thesaurus, der die vergleichende Forschung über Sprachen und religiöse Traditionen hinweg erlaubt (Abbildung 3).

Im so genannten Informationsinfrastrukturprojekt ‚Metaphern-Basislager' entwickeln Forschende aus den Religionswissenschaften, der Computerlinguistik und den Informationswissenschaften gemeinsam die digitale Forschungsinfrastruktur für alle Teilprojekte des SFB. Im Einklang mit der Methodologie des Verbundes werden drei methodologische Ebenen ausgestaltet:

(1) Basisinfrastruktur
Die Daten aller Teilprojekte werden als strukturierte Datenobjekte mit allen relevanten Metadaten in einem zentralen Repositorium abgelegt und so für die weitere Anreicherung und Analyse zugänglich gemacht. Ein Annotationsdienst erleichtert die Metaphernannotation gemäß gemeinsamem Schema. Als zentrales, sprachunabhängiges Referenzsystem für die Annotation wird ein konzeptueller Thesaurus geschaffen, der durch die Beiträge aus den Teilprojekten erweitert wird. Die Verknüpfung konkreter metaphorischer Ausdrücke mit einer zentralen Ressource ermöglicht das Auffinden konzeptionell verwandter Metaphern aus verschiedenen Korpora und die Untersuchung der in Metaphern verwendeten semantischen Domänen.

(2) Korpusanalyse-Toolbox
Einfach zu bedienende Korpuswerkzeuge wie Suchen, Konkordanzen oder statistische Kollokationsanalysen sollen für die Forschenden bereitgestellt werden, um ihnen die Aufdeckung von Mustern zu ermöglichen.

(3) Fortschrittliche Analysemethoden
Zwei der Teilprojekte mit computerlinguistischer Expertise arbeiten an Werkzeugen zur semi-automatischen Metapherndetektion und -analyse, die

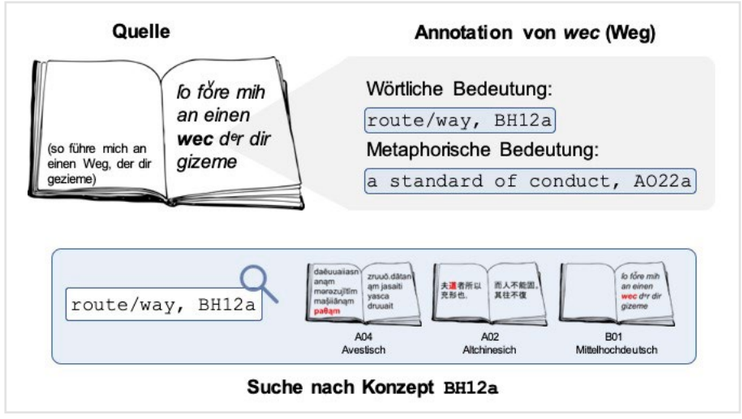


Abbildung 3: Der metaphorisch gebrauchte Ausdruck ‚wec' aus einem mittelhochdeutschen Text (links oben) wird annotiert und sowohl mit passenden Konzepten für die wörtliche als auch die metaphorische Bedeutung verknüpft (rechts oben). Wird anschließend nach dem Konzept ‚BH12a' gesucht, können Verwendungen in anderen Korpora beispielsweise aus dem Avestischen oder Altchinesischen (unten) aufgedeckt werden.

für die anderen Teilprojekte in der Infrastruktur integriert werden sollen.
Das SCC bringt einmal mehr seine Expertise im Forschungsdatenmanagement in den Verbund ein. Im Fokus liegen daher die Basiskomponenten der digitalen Infrastruktur – ein Forschungsdatenrepositorium mit modernsten Annotations-, Analyse- und Visualisierungswerkzeugen für die geisteswissenschaftlichen Daten. Gemeinsam mit den Wissenschaftlerinnen und Wissenschaftlern der Teilprojekte sollen innovative Wege interdisziplinärer Metaphernforschung beschritten werden.

Setzt die Segel!

Spannende Fragestellungen aus zahlreichen Disziplinen wie der Handschriftenkunde, Religionswissenschaft, Soziologie, Computerlinguistik, Material- und Informationswissenschaft warten in den nächsten Jahren auf die Forscherinnen und Forscher der beiden Projekte. Sie begeben sich auf die Spuren der mittelalterlichen Torarollen in Europa und des religiösen Gebrauchs von Metaphern in der Menschheitsgeschichte und freuen sich – metaphorisch gesprochen – auf die gemeinsame Reise!

Of Crowned Letters and Metaphorical Words

Since the beginning of 2022, two new projects with SCC participation complement the diverse range of religion research, taking a digital approach to the interdisciplinary study of medieval Torah scrolls and metaphors in religious meaning-making. The project ‘Materialized Holiness’, funded by the German Federal Ministry of Education and Research, analyses Torah scrolls from a multitude of perspectives to develop the first comprehensive repository for these fascinating research objects. The new Collaborative Research Center ‘Metaphors of Religion’, funded by the German Research Foundation, aims to understand and methodologically grasp the religious use of metaphors across times and cultures. With its extensive experience in research data management, SCC is at the side of the researchers in both collaborations and is developing a research data repository with state-of-the-art annotation, analysis and visualization tools for the heterogeneous humanities research data.

Energie und Mobilität – Herausforderungen für Hochleistungsrechnen und Maschinelles Lernen

Das Simulation and Data Life Cycle Lab (SDL) 'Engineering in Energy and Mobility' am SCC wurde im Rahmen des Nationalen Hochleistungsrechenzentrums (NHR@KIT) und der Plattform Helmholtz AI gegründet. Damit sind die Aufgaben dieses SDLs sehr breit aufgestellt – es werden Forschende aus ganz Deutschland bei deren Supercomputing-Ingenieursaufgaben mit Schwerpunkt Energie und Mobilität unterstützt und auch Teams mit Energieprojekten auf dem Gebiet des Maschinellen Lernens beraten. Gleichzeitig arbeiten die Wissenschaftlerinnen und Wissenschaftler des SDLs an ihren eigenen Kooperationsprojekten. *Jordan Denev, Charlotte Debus*

Sowohl Simulationen in den Bereichen Energieumwandlung und Mobilität mit Supercomputern, als auch die Methoden des Maschinellen Lernens (ML) stellen komplexe und sehr dynamische Forschungsthemen dar.[1] Um diese vielfältigen Forschungsrichtungen mit modernsten Methoden der Künstlichen Intelligenz (KI) auszurüsten, wurde am SCC das Simulation and Data Life Cycle Lab (SDL) „Engineering in Energy and Mobility" ins Leben gerufen. Hier arbeiten mehr als zehn Wissenschaftlerinnen und Wissenschaftler des SCC zusammen, um die langjährigen Erfahrungen der Strömungs- und Energiesimulation mit der methodischen Kompetenz des Maschinellen Lernens zu verknüpfen. Somit ergeben sich neue Kooperationsmöglichkeiten – wie sie z.B. bei einem Projekt zur Schadstoffreduzierung bei städtischen Transportmitteln zum Tragen kommen.

Projekte und Kooperationen

Thorsten Zirwes, ein Wissenschaftler des SDLs, hat nach seiner erfolgreichen Promotion am KIT im letzten Jahr ein 18-monatiges Stipendium vom Deutschen Akademischen Austauschdienst (DAAD) erhalten und forscht gegenwärtig als visiting PostDoc an der Stanford Universität in der Gruppe von Prof. Matthias Ihme. Dort beschäftigt er sich mit der Energieumwandlung in porösen Medien. Bei diesem Ansatz werden metallische oder keramische Bauteile, bestehend aus vielen kleinen Poren, von einem chemisch reagierenden Fluid durchströmt. Die durch die Reaktionen freigesetzte Wärme kann durch die große Kontaktoberfläche zwischen Gas und Feststoff vom porösen Medium aufgenommen und gezielt zum Stabilisieren der chemischen Reaktionsfront genutzt werden. Dies ermöglicht den Einsatz neuer kohlenstofffreier Brennstoffe, wie zum Beispiel Ammoniak. Die detaillierte Simulation der Strömung und deren chemischer Prozesse innerhalb des porösen Mediums ist sehr rechenintensiv, da die vielen kleinen Poren des Feststoffs aufgelöst werden müssten. Eine solche Simulation erfordert daher den Einsatz modernster Supercomputer wie beispielsweise den Hochleistungsrechner HoreKa am KIT.

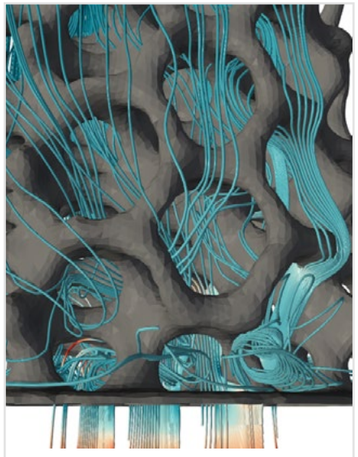


Abbildung 1: Stromlinienverlauf eines reagierenden Gases im porösen Medium (© Thorsten Zirwes)

Ein weiteres Projekt in Kooperation mit dem Engler-Bunte-Institut und dem Institut für Kolbenmaschinen am KIT analysiert große Messdatenmengen und mikroskopische Aufnahmen anhand von ML-Methoden. Das Ziel hier ist es, die Ausbreitung von Schadstoffen, die als Gase oder Feinpartikelstaub emittiert werden, zu minimieren. Dabei liegt der Fokus auf dem Kaltstart und den ersten zehn Kilometern, die von PKWs in der Stadt zurückgelegt werden – hier soll eine Gewichtung der zahlreichen Einflussparameter vorgenommen und letztendlich neue Steuerungsalgorithmen zur Unterstützung der Hersteller erarbeitet werden.

Das Grundlagenprojekt „Entwicklung und Validierung eines hybriden Gitter-/Partikelverfahrens für turbulente Strömungen, unterstützt durch Hochleistungsrechnungen mit OpenFOAM" forscht auf dem Gebiet der Turbulenzmodellierung[1]. Hier werden neue Turbulenzmodelle für die Ingenieuranwendungen in den Bereichen Energie und Mobilität entwickelt, die auf Lagrange'schen Partikelbewegungen basieren. Das Projekt, eine Kooperation mit der Universität Rostock, wird sukzessiv komplexere Strömungsansätze behandeln. Die numerische Herausforderung liegt in der Parallelisierung einer Großzahl an Partikeln und deren effizienter Portierung auf Grafikprozessoren (GPUs).

[1] www.scc.kit.edu/forschung/14972.php

Skalierbare Methoden der künstlichen Intelligenz

Auch die Methoden des Maschinellen Lernens (ML) und der Künstlichen Intelligenz (KI) benötigen heutzutage besonders leistungsfähige Rechensysteme und Ansätze des High Performance Computings (HPC). Im Rahmen von Beratungstätigkeiten unterstützt das Helmholtz AI-Team Forschende aus dem Bereich Energie darin, HPC- und ML-Methoden in ihren Forschungsprojekten zur Anwendung zu bringen. So werden beispielsweise auf Basis von Überflugbildern von Drohnen die mangelnde Wärmedämmung und ineffiziente Wärmeisolation von Gebäuden, sogenannte „Thermal Bridges", mittels Computer Vision identifiziert. Da die hierfür verwendeten tiefen neuronalen Netze eine Vielzahl von freien, trainierbaren Parametern aufweisen und darüber hinaus die auszuwertenden Videodaten sehr groß sind, ist eine entsprechende Analyse nur durch den Einsatz von Supercomputern möglich.

Open Source Software für datenintensive KI

Für die Unterstützung der Nutzenden stellt das AI-Team auch Softwarepakete für Maschinelles Lernen auf HPC-Systemen bereit. Das Open Source Toolkit HeAT[2] beispielsweise bietet verteilte Tensoroperationen und Algorithmen des Maschinellen Lernens sowohl auf CPUs als auch auf GPUs. Da-durch lassen sich sehr große Datenmen-gen in kürzester Zeit auswerten, wie dies beispielsweise in einem Projekt gemein-sam mit dem DLR bei der Auswertung von Verbrennungsdaten von Raketen erfolgte[3].

Die jüngste Entwicklung der KI-Gruppe ist das Softwarepaket HyDe[4], das im Rahmen eines gemeinsamen Projektes mit dem Helmholtz-Zentrum Dresden-Rossendorf entwickelt wurde. Ziel des Projektes war die Weiterentwicklung und Bereitstellung von Algorithmen für das hyperspektrale Entrauschen (Hyperspectral Denoising), welches für die Prozessierung von Remote Sensing-Daten zum Einsatz kommt. Mit HyDe stellen die Wissenschaftlerinnen und Wissenschaftler des SCC das erste Python-Paket für das GPU-beschleunigte Hyperspectral Denoising bereit, welches der Remote Sensing Community eine schnelle, zuverlässige und nutzerfreundliche Software für ihre Datenauswertung bietet.

KI für Energie und Energie für KI

Ein weiterer Fokus der Forschungsanstrengungen der KI-Gruppe liegt auf der Energieeffizienz von modernen KI-Algorithmen. Während Modelle und Algorithmen des Maschinellen Lernens sich kontinuierlich weiterentwickeln, sowohl hinsichtlich Speicher- als auch Rechenbedarf, konzentriert sich die KI-Community hauptsächlich auf die Verbesserung der Vorhersagemetriken. Um diesem nicht nachhaltigen Trend in der KI-Forschung entgegenzuwirken, arbeitet die Gruppe auch an der Frage, wie sich KI-Algorithmen zukünftig energieeffizienter und ressourcenschonender einsetzten lassen. Einen signifikanten Teil macht hierbei die Messung des Stromverbrauchs von KI-Workloads aus. HoreKa, der Karlsruher Supercomputer, bietet als erstes System über den XClarity Controller von Lenovo und ein eigens vom SCC hierfür entwickeltes Slurm-Plugin die Möglichkeit, den Stromverbrauch einzelner Knoten sehr genau zu erfassen. Dadurch konnten die Teams KI-Forschung und HPC-Betrieb die Energieeffizienz verschiedener KI-Modelle auf heterogenen Hardware-Knoten messen. Um die Gemeinschaft der KI-Forschenden weiter für das Thema Energieeffizienz zu sensibilisieren, hat Helmholtz AI Anfang des Jahres eigens dafür einen Hackathon veranstaltet (Seite 32).

Vielseitige Nutzendenunterstützung für das Hochleistungsrechnen

Die Nutzenden des Nationalen Hochleistungsrechenzentrums NHR@KIT werden durch die Bereitstellung von Simulationssoftware, diverse Kurse und ein Voucher-System zur Projektbeantragung unterstützt. Auch die neuen Mitarbeitenden am NHR@KIT engagieren sich aktiv an der Unterstützung von Nutzenden – sie beantworten Fragen, helfen beim Erstellen von Kursmaterial und nehmen selbst an den Kursen teil. Des Weiteren gibt es die Möglichkeit über ein extra dafür vorgesehenes Voucher-System eine umfangreiche Projektunterstützung bei den AI Consultants zu beantragen (siehe SCC-News 2/2020, S. 20). Hier gehen die KI-Fachleute tiefer auf die Aufgaben und Bedürfnisse der einzelnen Anwender-Teams ein, und nicht selten entstehen so längerfristige Kooperationen.

[2] github.com/helmholtz-analytics/heat

[3] Debus, C. et al. "High-performance data analytics of hybrid rocket fuel combustion data using different machine learning approaches" AIAA Scitech 2020 Forum. 2020.

[4] github.com/Helmholtz-AI-Energy/HyDe

Energy and Mobility – National Level of Supercomputing boosted by Machine Learning

Supercomputing simulations of energy conversion and vehicle mobility, as well as machine learning methods, represent complex and very dynamic research areas. The SCC SimDataLab "Engineering in Energy and Mobility" was established to better manage and link these diverse research areas. More than ten SCC scientists have come together to combine their many years of experience in flow and energy simulation with the methodological expertise of machine learning (ML for short). This opens up new opportunities for cooperation – for example in a new project that aims to reduce pollutants from urban means of transport when the vehicles start at cold engine conditions.

Anzt tritt in die Fußstapfen von Turing Award Winner Jack Dongarra

Im Jahr 2017 wurde die Helmholtz-Nachwuchsgruppe "Fixed-Point Methods for Numerics at Exascale" (FiNE) von Hartwig Anzt am SCC gegründet. Zunächst hatte die Gruppe nicht viel Manpower, aber eine klare Vision, inspiriert von Anzts Zeit als PostDoc im Innovative Computing Laboratory (ICL) an der University of Tennessee. Die Idee, Algorithmen und Software für das Hochleistungsrechnen als freies Gut für die wissenschaftliche Community zu entwickeln, ist sehr vom Wunsch geprägt, damit zum technologischen Fortschritt beizutragen. Jedoch braucht es einen langen Atem, die Stakeholder von der Schlagkraft dieser Idee zu überzeugen. Auch die Gruppe FiNE hatte eine längere Anlaufphase, in der Erfolge und Anerkennung rar blieben. Sicherlich hat die Partnerschaft mit Jack Dongarras ICL geholfen, diese Zeit zu überstehen; essentiell aber war ein motiviertes Team, das die Überzeugung teilte, mit diesem Ansatz das Richtige zu tun. Hinzu kommt, dass das SCC mit seinem Aufgabenspektrum, das vom Betrieb von Hochleistungsrechnern und Services bis hin zur Forschung an neuen Algorithmen reicht, ein idealer Ort ist, um Algorithmen und Software als wissenschaftliche Infrastruktur zu entwickeln und zu betreiben.

FiNE behauptet sich im internationalen Wettbewerb

Über die Jahre stellte sich der gewünschte Erfolg ein und ließ die Forschungsgruppe FiNE weit über die Erwartungen hinaus wachsen. In den fünf Jahren der Helmholtz-Förderung hat die Gruppe 80 wissenschaftliche Artikel in Zeitschriften und auf Konferenzen publiziert und gleichzeitig Drittmittelprojekte in Höhe von insgesamt 5,5 Mio. Euro akquiriert. Darunter waren Projekte mit hoher Sichtbarkeit aus den Forschungsprogrammen Horizon 2020, EuroHPC und dem US Exascale Computing Project. Für das Team viel wichtiger aber war und ist, dass die gemeinsam entwickelte Softwarebibliothek Ginkgo (SCC-News 2/2018, 1/2019 und 1/2020) von der Community angenommen wird, und sowohl der Entwicklungsprozess als auch das Softwaredesign ein Vorbild für andere Projekte wurde. Die entwickelten Algorithmen und Implementierungen tragen somit zur Forschung unter anderem in der Elektrophysiologie, Physik von Mehrphasenströmungen, Fusionstechnologie und Windkraftsimulationen bei. Bei allem in klassischen Metriken erfassbarem Erfolg betont Hartwig Anzt aber besonders den Teamgeist seiner Gruppe: „Ich bin unglaublich dankbar für diese fünf Jahre und die Erfahrungen, die ich in dieser Zeit machen durfte.

Bei seiner Verabschiedung übergibt Professor Jack Dongarra „den Staffelstab" an Hartwig Anzt, den nun neuen Direktor des dortigen Innovative Computing Laboratory

Gemeinsam eine solche Gruppe aufbauen und mit so talentierten Wissenschaftlerinnen und Wissenschaftlern arbeiten zu dürfen, ist schon ein besonderes Privileg."

Ein weiterer Sprung auf der Karriereleiter

Im November 2021 wurde Hartwig Anzt auf eine befristete Juniorprofessur in der KIT-Fakultät für Informatik berufen – ein für die akademische Laufbahn wichtiges Sprungbrett.

Im April 2022 erhielt Professor Jack Dongarra die höchste Auszeichnung, die es in der Informatik gibt: den Turing Award, auch „Nobelpreis der Informatik" genannt. Diese Auszeichnung ist auch eine Anerkennung für das Thema Software im wissenschaftlichen Hochleistungsrechnen allgemein und unterstreicht die Relevanz des Forschungsfeldes für den wissenschaftlichen Fortschritt. Vor allem ist sie aber der krönende Abschluss Dongarras wissenschaftlicher Karriere, der bereits ein Jahr zuvor angekündigt hatte, Ende Juni 2022 in den Ruhestand zu gehen.

Jack Dongarra besucht das SCC

Bereits kurz nach der Verleihung des Turing Awards im Mai besuchte Jack Dongarra das SCC – eine einmalige Gelegenheit für das KIT, einen Turing Award-Gewinner persönlich zu treffen. Neben dem Besuch des Supercomputers HoreKa und des Neutrino-Experiments KATRIN stand ein gemeinsames Mittagessen auf dem Programm. Am Nachmittag hielt Prof. Dongarra einen Vortrag an der KIT-Fakultät für Informatik. Anschließend nutzten viele derzeitige und ehemalige Professorinnen und Professoren des KIT die Gelegenheit zum Austausch in lockerer Atmosphäre.

Anzt wird Nachfolger von Dongarra als Direktor des ICL

Im Frühjahr 2022 erhielt Hartwig Anzt den Ruf auf eine Professur an die University of Tennessee, verbunden mit der Direktorenstelle im Innovative Computing Lab als Jack Dongarras Nachfolger. Diesen Karrieresprung verdankt er nach eigenen Worten auch dem SCC: „Dass ich mich im offenen Verfahren gegen renommierte und sehr etablierte Wissenschaftlerinnen und Wissenschaftler durchgesetzt habe, liegt auch daran, dass ich am SCC die Möglichkeit hatte, eine Forschungsgruppe erfolgreich aufzubauen und zu führen." Die Geschichte der Forschungsgruppe FiNE ist damit jedoch noch nicht zu Ende: Anzt bleibt in einem Anbindungsmodell am SCC und leitet die Gruppe weiter. So hatte er es bei seinem Doktorvater Dongarra erlebt, der gleichzeitig mit der University of Manchester verbunden war. „Ich sehe ganz großes Potential für Synergien durch die enge Verknüpfung von Forschungsgruppen auf zwei Kontinenten, die weit über den Austausch von Promovierenden und PostDocs hinausgehen", ist Anzt überzeugt und lässt das SCC wissen: „Ich habe die Flugtickets zu einem nächsten Aufenthalt am KIT bereits gebucht."

Anzt follows in the footsteps of Turing Award Winner Jack Dongarra

In 2017, the Helmholtz Young Investigator Group "Fixed-Point Methods for Numerics at Exascale" (FiNE) was founded by Hartwig Anzt at the SCC – initially without much manpower, but with a clear vision inspired by Anzt's postdoctoral time in Jack Dongarra's Innovative Computing Laboratory at the University of Tennessee. The idea of developing algorithms and software for high-performance computing as a free good for the scientific community has driven the success of the FiNE group over the years, and now, after the 5 years of initial Helmholtz funding are completed, the group has established the Ginkgo numerical software in the scientific community, published over 80 scientific articles in journals and conferences, and acquired third-party funded projects totaling €5.5 million, including high-visibility projects from Horizon 2020, EuroHPC and the US Exascale Computing Project. The success comes not unnoticed by the community, and Hartwig Anzt received a call to a professorship at the University of Tennessee, combined with the director position in the Innovative Computing Lab as Jack Dongarra's successor. In his own words, Anzt acknowledges the importance of the SCC to enable this career leap: "The fact that I prevailed in the open procedure against renowned and well-established scientists is certainly also due to the fact that I had the opportunity at the SCC to prove that I can successfully build and lead a group." However, the story of the FiNE research group does not end: Hartwig Anzt will remain at the SCC under an affiliation model and continues the group leadership. Anzt: "I see quite a lot of potential for synergies through the close linking of research groups on two continents, which goes far beyond the exchange of doctoral researchers and postdocs. I have already booked the plane tickets for my next stay at KIT."

Internationale Praktikanten am SCC

Rached Chaaben aus Tunesien, der sein Masterstudium mit einem Auslandspraktikum abschließt, und Stefano Maurogiovanni, ein ERASMUS+-Praktikant der EuroHPC-Partneruniversität Pavia in Italien, berichten im Interview von ihren ersten Eindrücken am SCC.

Im Frühjahr 2022 kamen zwei internationale Praktikanten ans SCC: Rached Chaaben aus Tunesien, der sein Masterstudium mit einem Auslandspraktikum abschließt, und Stefano Maurogiovanni, ein ERASMUS+-Praktikant von der EuroHPC-Partneruniversität Pavia in Italien. Da sie einen sehr unterschiedlichen wissenschaftlichen und kulturellen Hintergrund haben, wollten wir mehr darüber erfahren, wie Stefano und Rached zum SCC gekommen sind, welche ersten Eindrücke sie von ihrer neuen Heimat gewonnen haben und was ihre größten Herausforderungen waren, um hierher zu kommen.

Die beiden internationalen Praktikanten Rached Chaaben (links) und Stefano Maurogiovanni (rechts) vor dem SCC-Gebäude am Campus Nord (Foto:Yu-Hisang Mike Tsai)

Rached und Stefano, könnt ihr euren Bildungshintergrund kurz zusammenfassen?

Rached: *Ich besuchte das Vorbereitungscurriculum für das Ingenieurstudium (Äquivalent zum französischen CPGE: Classe Préparatoire pour les Grandes Ecoles), das ich nach zwei Jahren mit Schwerpunkt auf Mathematik und Physik absolvierte. Anschließend bestand ich die Aufnahmeprüfung für die nationalen Ingenieurschulen und belegte Platz 13 von mehr als 1.000 Bewerbenden. Dies ermöglichte mir die Aufnahme an der Ecole Polytechnique de Tunisie, wo ich jetzt allgemeines Ingenieurwesen mit dem Schwerpunkt Signale und Systeme studiere (gleichbedeutend mit einem Masterabschluss).*

Stefano: *Ich habe als Bachelor-Student für Bioingenieurwesen an der Universität Pavia (UNIPV) begonnen. Während dieser Zeit arbeitete ich als studentischer Tutor und konzentrierte mich auf Deep Learning-Anwendungen in der Bioinformatik. Danach habe ich mich für einen Master in Computertechnik – insbesondere Data Science – eingeschrieben, ebenfalls an der UNIPV, wo ich als Laborassistent tätig war. Derzeit befinde ich mich im letzten Semester meines Masterstudiums.*

Wie seid ihr auf das SCC aufmerksam geworden und schließlich hier gelandet?

Rached: *Zu Beginn dieses akademischen Jahres habe ich mich auf die Suche nach einem Abschlusspraktikum gemacht, das für mein Studium obligatorisch ist. Ich war auf der Suche nach einem Forschungsthema, das HPC und Mathematik kombiniert. Ich hatte das Glück, auf Hartwig Anzt zu stoßen, und war begeistert von der Arbeit, die er in seiner Gruppe leistet. Eine andere Person, die ich während meiner Suche kennenlernte, half mir, mit Hartwig in Kontakt zu treten. Dieser war offen für ein Treffen und nahm mich glücklicherweise als Praktikant in seine Gruppe auf.*

Stefano: *Sowohl das KIT als auch die UNIPV gehören zu den Partnern des EuroHPC MICROCARD-Projekts, zu dem ich hoffentlich mit meiner Masterarbeit beitragen kann. Daher haben Hartwig Anzt und meine Betreuer vor Ort vereinbart, dass ich das Sommersemester 2022 als Forschungsstudent am SCC verbringen werde.*

Woran arbeitet ihr?

Rached: *Ich arbeite an der Implementierung eines Sparse Matrix-Matrix-Multiplikationsalgorithmus, der für die verschiedensten High Performance Computing-Anwendungen wichtig ist und häufig genutzt wird. Wir planen in diesem Praktikum, diesen Algorithmus für Multicore-CPUs und GPUs zu implementieren. Außerdem wollen wir mit dieser Arbeit einen Beitrag zum Ginkgo-Projekt leisten, einer Hochleistungsbibliothek für lineare Algebra, die derzeit in unserer Forschungsgruppe entwickelt wird.*

Stefano: *Ich arbeite an der Implementierung eines verteilten Lösers, der auf algebraischen Mehrgittermethoden (AMG) für kardiologische Elektrophysiologie-Simulationen basiert. Der Code wird vollständig in C++ geschrieben und stützt sich stark auf MPI und die Bibliothek Ginkgo.*

Was sind eure Freizeitinteressen?

Rached: *Ich treibe Sport und verbringe viel Zeit mit Tischtennis und Volleyball. Und seit ich in die Gruppe von Hartwig gekommen bin, habe ich mich für das Wandern begeistert. Außerdem koche ich sehr gerne und schneide auch gerne Haare.*

Stefano: *Ich habe eine Vorliebe für Science Fiction-Romane und -Filme. Außerdem mag ich Basketballsport – den ich seit mehr als zehn Jahren betreibe –, außerdem mag ich Wandern, Basteln und für Freunde kochen.*

Welches war die größte Herausforderung, um hierher zu kommen?

Rached: *Das Visa-Verfahren, eine Unterkunft zu finden, was mich immer noch beschäftigt, und der ganze Papierkram drumherum waren aufwändige Prozesse, die Hartwig und ich durchlaufen mussten. Ich kann sagen, dass Hartwig sich sehr viel Mühe gegeben hat, und dafür bin ich ihm sehr dankbar.*

Stefano: *Die Wohnungssuche in Karlsruhe und die Immatrikulation am KIT verliefen für mich recht reibungslos. Die zeitaufwändigste Aufgabe bestand wohl darin, meinen Wohnsitz bei der Stadt anzumelden.*

Auf Wanderungen erlebt Rached zum ersten mal Schnee

Und euer schönstes Erlebnis bisher?

Rached: *Nach nun mehr als zwei Monaten in der Gruppe kann ich sagen, dass ich kein angenehmeres Umfeld und keine angenehmeren Teamkollegen haben könnte. Es ist auch eine lohnende Erfahrung hier zu sein.*

Stefano: *Die Arbeit in einer Gruppe von kompetenten, motivierten und leidenschaftlichen Menschen, von denen ich viel lernen kann, ist für mich eine großartige Gelegenheit. Außerdem machen wir am Wochenende gemeinsam verschiedene Unternehmungen – meistens sind das Wanderungen –, die ich sehr schätze, weil sie uns zusammenschweißen und wir etwas außerhalb unseres Arbeitsplatzes teilen.*

Stefano und Hartwig besichtigen den HoreKa

Was sind eure Pläne für die Zukunft? Was werdet ihr nach diesem Praktikum machen?

Rached*: Ich möchte in der Forschung weitermachen, daher ist die Bewerbung um eine Doktorandenstelle mein Plan A. Das wird von der Qualität meiner Arbeit während dieses Praktikums abhängen. Als Plan B kann ich mir auch die Suche nach einem Job vorstellen.*

Stefano: *Ich überlege noch, ob ich mich für eine Arbeitsstelle oder eine Doktorandenstelle bewerben soll, je nachdem, welche Ziele ich mit meinem Dissertationsprojekt erreichen kann und welches Feedback ich von den anderen Gruppenmitgliedern bekomme.*

International Interns at SCC

In spring 2022, Rached Chaaben from Tunesia, who completes his master's studies with an international internship, and Stefano Maurogiovanni, an ERASMUS+ intern from the EuroHPC partner university of Pavia in Italy did their internship at SCC. Coming from very different scientific and cultural backgrounds, we wanted to know more about why Stefano and Rached came to SCC, their first impressions of their new home, and their biggest challenges getting here.

Erster AI-HERO Hackathon zum Thema Energieeffiziente KI

Anfang Februar fand der erste AI-HERO Hackathon zu Energieeffizienter Künstlicher Intelligenz statt. Über drei Tage hatten die sieben teilnehmenden Teams Zeit, KI-Modelle für die Lösung zweier Anwendungsfälle aus den Gebieten Energie und Gesundheit zu entwickeln und dabei so wenig Strom wie möglich für Rechenzeit zu verwenden. Gerechnet wurde auf der HAICORE-Partition des Hochleistungsrechners HoreKa am KIT, wo ein neues SLURM-Plugin eine sehr genaue Messung des Energieverbrauchs pro Job zulässt. Organisiert wurde das Event von Wissenschaftlerinnen und Wissenschaftlern des DKFZ und KIT im Rahmen ihrer Helmholtz-Inkubator-Plattformen.

Charlotte Debus, Markus Götz

In den letzten zehn Jahren hat Künstliche Intelligenz (KI) in vielen Bereichen der Wissenschaft und Technik große Fortschritte ermöglicht. Mit dem zunehmenden Einsatz von KI in immer komplexeren Fragestellungen wächst jedoch auch der dafür erforderliche Bedarf an Rechenressourcen. Moderne Beschleunigerhardware und große Compute Cluster ermöglichen zwar die Implementierung immer größer werdender KI-Modelle, die Bereitstellung bzw. die Reduktion des enormen Strombedarfs dieser IT-Infrastrukturen stellen jedoch zunehmend eine Herausforderung dar, insbesondere im Hinblick auf den Klimawandel und den Übergang zu erneuerbaren Energien.

Die „Helmholtz Energy- and Resource-awareness Operation for Artificial Intelligence" (AI-HERO) ist ein Zusammenschluss von Wissenschaftlerinnen und Wissenschaftlern aus vier Helmholtz-Inkubator-Plattformen an zwei verschiedenen Helmholtz-Zentren. Gemeinsam wollen die Mitglieder von Helmholtz AI, Helmholtz Imaging, Helmholtz Metadata Collaboration (HMC) und die Graduiertenschule HIDSS4Health am DKFZ und KIT das Bewusstsein für den Energieverbrauch moderner KI-Forschung und -Entwicklung schärfen und zur Verbreitung von „Green IT"-Ansätzen beitragen (Abbildung 1).

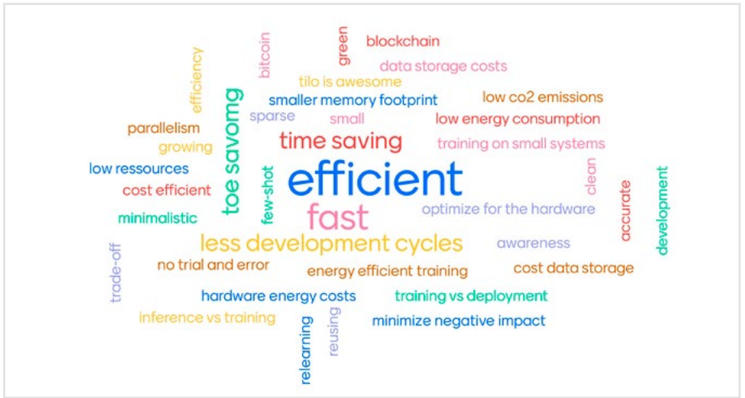


Abbildung 1: Ergebnis des Brainstormings zur Green IT

Anfang Februar 2022 hat ein Team aus KI-Forscherinnen und -Forschern des SCC und des Instituts für Automation und angewandte Informatik (IAI) am KIT sowie des Deutschen Krebsforschungszentrums (DKFZ) deshalb einen dreitägigen virtuellen Hackathon veranstaltet. Im Rahmen des Programmier-Events wurden energieeffiziente KI-Modelle entwickelt, d.h. Modelle, die bei Modellentwicklung, -training und -vorhersage möglichst wenig Strom verbrauchen.

19 Teilnehmende aus zehn verschiedenen Helmholtz-Zentren und Partneruniversitäten bzw. Forschungseinrichtungen nahmen an dem Hackathon teil, in dessen Verlauf jeweils ein Anwendungsfall aus dem Bereich Energie und aus dem Bereich Gesundheit in Teams von je drei Personen bearbeitet wurde. Im Use Case Gesundheit sollten die Teilnehmenden ein Modell entwickeln, das eine COVID-19-Infektion anhand von Röntgenbildern möglichst zuverlässig vorhersagt (Abbildung 2). Die für das Training verwendeten Daten wurden in verschiedenen Krankenhäusern aufgenommen und im Rahmen der COVID-Net Open Initiative[1] gesammelt.

Im Use Case Energie war die Aufgabe, die elektrische Last, d.h. den Verbrauch, vorherzusagen: Basierend auf Daten des stündlichen Stromverbrauchs verschiedener Städte, gemessen über drei Jahre, sollte der zukünftige Verbrauch für einen Zeitraum von einer Woche vorhergesagt werden (Abbildung 3).

Für die Entwicklung, das Training und die Inferenz der Modelle wurde die Helmholtz AI Computing Resources (HAICORE)-Partition des Supercomputers HoreKa am KIT verwendet. Durch den Einsatz der Server Management Software „XClarity Controller" von Lenovo und eines neuartigen SLURM-Plugins ermöglicht HoreKa eine hochaufgelöste Messung des Stromverbrauchs pro Knoten. Während des gesamten Hackathons wurde damit der Stromverbrauch aller auf dem Hochleistungsrechner durchgeführten Berechnungen überwacht und so der Gesamtverbrauch für die Modellentwicklung gemessen.

[1] alexswong.github.io/COVID-Net

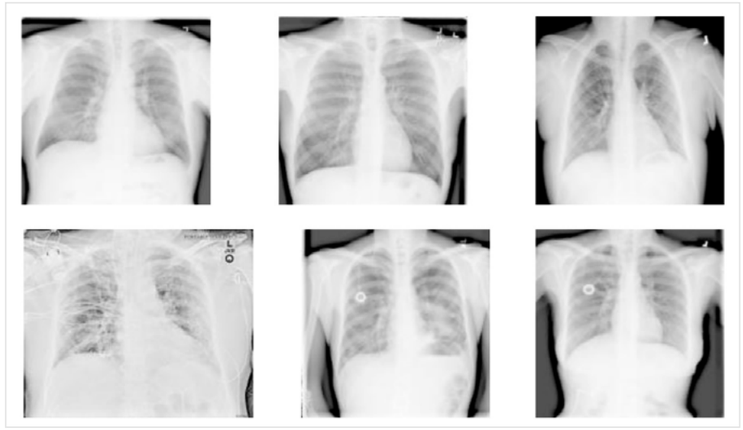


Abbildung 2: Beispiele von Röntgenbildern aus dem Anwendungsfall Gesundheit

Am Ende des Hackathons wurden die von den Teams entwickelten finalen Modelle zur Evaluierung eingereicht. Um die Grundsätze von Open Science aktiv voranzutreiben, wurde hierbei explizit bei allen Schritten der Modellentwicklung auf die FAIR-Prinzipien (Findable, Accessible, Interoperable, Reproducible) Wert gelegt: Die fertigen, trainierten Modelle wurden von den Teilnehmenden auf den Speicherdienst Zenodo hochgeladen und der zugehörige Code in einem GitHub-Repositorium veröffentlicht. Diese Beiträge wurden dann vom AI-HERO-Team verwendet, um die Modelle auf den Testdatensatz anzuwenden. Dadurch konnten die Modelle unabhängig sowohl in Bezug auf ihre Leistung und ihren Energieverbrauch, als auch auf ihre Reproduzierbarkeit bewertet werden.

Den Anwendungsfall Energie gewann das Team „Dynamic Ants" vom Helmholtz-Zentrum Hereon. Sie verwendeten einen DeepAR-Ansatz und konnten damit das Modell mit der höchsten Vorhersagegenauigkeit entwickeln, während ihr gesamter Entwicklungsprozess nur 1,74 Megajoule (MJ) verbrauchte. Dies war rund ein Drittel des Energieverbrauchs des zweitplatzierten Teams „Red Warriors". Auch in der Inferenz konnte das Modell der „Dynamic Ants" mit einem Energieverbrauch von 36,39 Kilojoule (kJ) überzeugen. Im Anwendungsfall Gesundheit machte das Team „Skeleton Suns" vom DKFZ das Rennen um den ersten Platz. Ihr auf einem ResNet-18 basierendes Modell war zwar nur das zweitbeste im Ranking des Entwicklungsenergieverbrauchs (22,10 MJ), resultierte aber in der höchsten Genauigkeit (66,43 %) und einem geringen Stromverbrauch in der Inferenz (106,64 kJ).

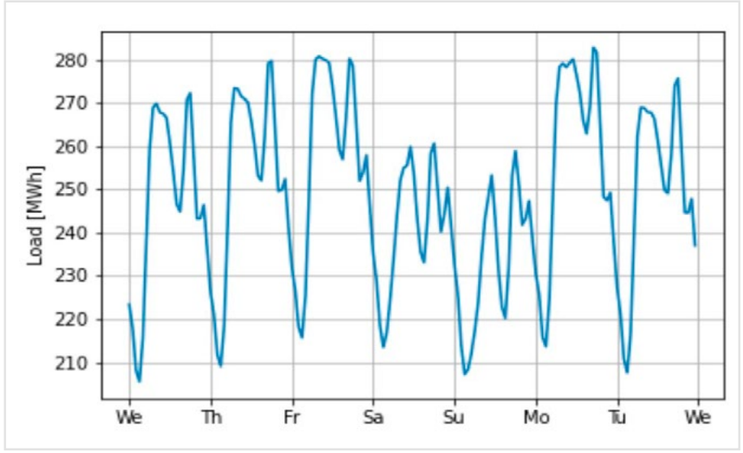


Abbildung 3: Eine 7-Tage-Lastkurve aus dem Anwendungsfall Energie

Alle Modelle und Ergebnisse können auf der Webseite des AI-HERO Hackathon[2] eingesehen werden. Neben der reinen Modellentwicklung fand auch ein Rahmenprogramm mit Keynote-Vorträgen zum Thema Green AI und zum Industrieansatz von Lenovo für energieeffiziente IT- und Computerhardware statt. Die Veranstaltung wurde durch die Helmholtz Information & Data Science Academy (HIDA) und das NHR-Zentrum NHR@KIT unterstützt.

First AI-HERO Hackathon on Energy Efficient AI

The first AI-HERO Hackathon on Energy Efficient Artificial Intelligence took place in early February. Over three days, the seven participating teams had time to develop AI models for solving two use cases from the fields of energy and health while using as little power for computing time as possible. The computation was performed on the HAICORE partition of the high-performance computer HoreKa at KIT, where a new SLURM plugin allows a very precise measurement of the energy consumption per job. The event was organized by KIT and DKFZ scientists in the context of their Helmholtz Incubator platforms.

[2] ai-hero-hackathon.de/

20 Jahre Grid Computing Centre Karlsruhe – Fotoimpressionen der Jubiläumsfeier

Weitere Informationen
www.scc.kit.edu/ueberuns/16298

Fotos Kira Heid (KIT)

Neues aus den SCC-Abteilungen

Neue Leitung der Abteilung Scientific Computing and Mathematics

Seit 1. März 2022 ist Jasmin Hörter Leiterin der Abteilung Scientific Computing and Mathematics (SCM) am SCC. Sie übernahm diese Aufgabe von Ivan Kondov, der die Abteilung zuvor kommissarisch leitete und weiterhin als stellvertretender Abteilungsleiter fungieren wird.

Nach ihrem Mathematikstudium in Karlsruhe und London war Frau Hörter von 2016 bis 2021 Doktorandin am Institut für Analysis des KIT und schloss ihre Promotion im Bereich der Geometrischen Analysis im Frühjahr 2021 ab.

Im Anschluss war sie für das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg tätig. Dort betreute sie im Referat für Digitalisierung und Informationsinfrastruktur die Hochschulrechenzentren des Landes und begleitete neben SCC-Projekten wie bwHPC-S5, Simulierte Welten oder bwCampusnetz auch die Gründungsphase des Verbunds Nationales Hochleistungsrechnen (NHR).

In der Abteilung SCM treffen drei Bereiche aufeinander. In der Forschungsgruppe Computational Science and Mathematical Methods (CSMM) dreht sich alles um Modellierung und numerische Methodenforschung. Für die HPC-Nutzendenunterstützung stehen Expertinnen und Experten der Simulation and Data Lifecycle Laboratories (SDLs) zur Verfügung, und die HPC Outreach-Projekte Simulierte Welten und CAMMP vermitteln Schülerinnen und Schülern in Workshops und Projekttagen mathematische Lösungsansätze für reale Modellierungsprobleme.

IMPRESSUM

SCC news
Magazin des Steinbuch Centre for Computing

Herausgeber
Präsident Professor Dr.-Ing. Holger Hanselka
Karlsruher Institut für Technologie (KIT)
Kaiserstraße 12
76131 Karlsruhe

Anschrift
Steinbuch Centre for Computing (SCC)
Karlsruher Institut für Technologie (KIT)
Redaktion SCC-News
Zirkel 2
76131 Karlsruhe
oder:
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen
Fax: +49 721 608-24972

Redaktion
Achim Grindler (verantwortlich), Karin Schäufele, Andreas Ley
E-Mail: redaktion@scc.kit.edu

Gestaltung, Satz und Layout
Nicole Gross, Elias Kobel, Heike Gerstner
AServ – Crossmedia – Grafik (CroM)
Karlsruher Institut für Technologie (KIT)
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen

Titelfoto
Jack Dongarra (m.) neben Martin Frank (l.) und Hartwig Anzt (r.) im Kaltgang des Hochleistungsrechners Karlsruhe am KIT Campus Nord. (Foto: Achim Grindler, KIT, SCC)

Fotos
SCC, KIT

Druck
Systemedia GmbH, 75449 Wurmberg

Erscheinungstermin dieser Ausgabe
August 2022

www.scc.kit.edu/publikationen/scc-news

Karlsruher Institut für Technologie (KIT)
Steinbuch Centre for Computing (SCC)

ISSN: 1866-4954

www.scc.kit.edu
www.scc.kit.edu/twitter
contact@scc.kit.edu