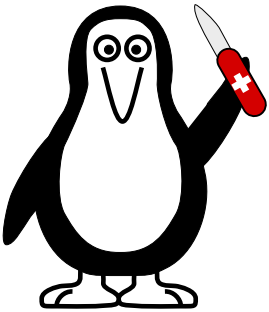
Quartl

2/2025
114. Ausgabe

Inhalt



Das Quartl erhalten Sie online unter `https://www.cs.cit.tum.de/sccs/weiterfuehrende-informationen/quartl/`

Das Quartl ist das offizielle Mitteilungsblatt des *Kompetenznetzwerks für Technisch-Wissenschaftliches Hoch- und Höchstleistungsrechnen in Bayern* (KONWIHR) und der *Bavarian Graduate School of Computational Engineering* (BGCE)

Editorial

Wie gut, dass die Internationalisierung an deutschen Universitäten und insbesondere an der TUM unaufhaltsam voranschreitet. Denn wenn mich nicht eine fürsorgliche eidgenössische Mitarbeiterin auf die absolut unglaubliche Nachricht aufmerksam gemacht hätte, um die es diesmal gehen soll – sie wäre glatt unbemerkt an mir vorbeigezogen.

Denn was da am 27.04.25 auf der Online-Ausgabe der Schweizer Tagesschau zu lesen war, rüttelt an den Urfesten Schweizer Tradition, ja Kultur: „Angela Koller wird erste Frau Landammann in Appenzell Innerrhoden". Sensationell! Ein Landammann ist schließlich nicht irgendwer, sondern in zehn Schweizer Kantonen der – bzw. nun die – Vorsitzende der Kantonsregierung. In den beiden Kanton Glarus und Appenzell Innerrhoden leitet er bzw. sie zudem die so genannte Landsgemeinde, also die verfassungsmäßige, an einem bestimmten Tag unter freiem Himmel und mit feierlichem Zeremoniell abgehaltene Versammlung aller stimmfähigen Einwohner des Kantons.

Beginnen wir mit etwas Landeskunde – das ist doch alles recht verwirrend. Die Geschichte der Schweizer Kantone reicht zurück bis ins 15. Jahrhundert. Es gibt Kantonsverfassungen, Kantonsregierungen und Kantonsparlamente. Somit sind die Kantone also irgendetwas zwischen den deutschen Bundesländern und Regierungsbezirken (abgesehen davon, dass natürlich alles total anders und viel toller ist in der Eidgenossenschaft). Im Laufe der Jahrhunderte kamen Kantone hinzu und verschwanden wieder – sehr kurzlebig waren beispielsweise die Kantone Säntis, Waldstätte, Lugano, Bellinzona und Fricktal. Heute wird ihre Zahl mit 26 oder 23 angegeben. Diese auf den ersten Blick unverständliche Diskrepanz rührt daher, dass es sechs so genannte Halbkantone gibt: Nidwalden und Obwalden, Basel-Stadt und Basel-Landschaft, sowie eben Appenzell-Außerrhoden und Appenzell-Innerrhoden. Appenzell-Innerrhoden (nein, das gängige und von den KFZ-Kennzeichnen bekannte Kürzel AI mag ich ganz und gar nicht benutzen ...) ist der flächenmäßig zweitkleinste und bevölkerungsmäßig kleinste Kanton, mit gerade mal rund

17.000 Einwohnenden. Bereits 1513 trat man dem Bund bei. Dass das kleine Appenzell sich in zwei Kantone gliederte, hat – wie könnte es im Europa des sechzehnten Jahrhunderts anders sein – mit der Religion zu tun: während Außerrhoden sich der Reformation anschloss und evangelisch-reformiert wurde, blieb man in Innerrhoden streng katholisch. Im Gegensatz zu anderswo lief das aber alles recht friedlich ab.

So, und nun nähern wir uns langsam der eingangs zitierten Sensationsnachricht an. Denn noch im April 1990 hatte sich die Landsgemeinde von Appenzell-Innerrhoden gegen die Einführung des Frauenwahlrechts ausgesprochen, aktiv wie passiv. Noch im selben Jahr setzte allerdings das Bundesgericht durch, dass der Kanton – als letzter der 26 – auch auf kantonaler Ebene das Wahlrecht für Frauen einführen musste; was dann auch 1991 umgesetzt wurde.

I.A. am letzten Aprilsonntag kommt die Landsgemeinde zusammen, um große Politik zu machen. Als Nachweis des Stimmrechts gilt dabei ein Papierkärtchen, wobei Männer alternativ auch das „Seitengewehr“ vorzeigen können: ursprünglich wirklich ein Gewehr, heute meist ein Degen, Säbel oder Bajonett. Recht kriegerisch, bzw. wehrhaft, auf jeden Fall. Als eine ihrer Aufgaben wählt die Landsgemeinde den Landammann, ganz grob den Regierungschef. Genauer gesagt, gibt es derer zwei: den regierenden Landammann und den stillstehenden Landammann. Und die Wahlen am 27.4.2025 ergaben nun eben, dass sich Angela Koller fortan mit dem Titel „Frau stillstehender Landammann“ schmücken darf. Dabei setzte sie sich gegen drei männliche Mitbewerber durch. Sie ist erst die vierte Frau in der Standeskommission und die erste als Frau Landammann. Letzteres ist tatsächlich die offizielle weibliche Bezeichnung des Titels – und in der Tat wären weder Landammännin noch Landaffrau bessere Bezeichner. Und nach guter Sitte wird sie damit in zwei Jahren auf den Stuhl des regierendenn Landammanns wechseln und dann als Regierungschefin mächtigste unter den 17,000 Appenzell-Innerrhodlern sein. Sapperlott!

Koller selbst bezeichnete ihre Wahl als Zeichen dafür, dass man die Geschichte hinter sich lassen und in die Zukunft gehen könne. Sie betonte aber auch, nicht nur wegen ihres Geschlechts, sondern auch wegen ihrer langjährigen politischen Erfahrung gewählt worden zu sein. Angesichts der bekannten Gender-Haltung in Appenzell-Innerrhoden sieht der externe Betrachter ohnehin eher den zweiten Punkt und eine Wahl trotz ihres Geschlechts .

Aber wie dem auch sei, was bedeutet das nun für die große Bühne der Weltpolitik? Ist Appenzell-Innerrhoden das gallische Dorf, das sich selbstbewusst gegen das Trumpsche „Kommando zurück!“ in Sachen Gleichstellung stellt, und das damit den Software-Riesen SAP noch erbärmlicher aussehen lässt? Oder ist es einfach ein weiteres Indiz dafür, dass auch eine „Lederhosen-Kultur“ mit Laptops umzugehen vermag? Egal. Erfreulich allemal, und beileibe nicht nur als Input für das Quartl. Insofern wünschen wir der neuen Frau stillstehender Landammann alles Gute für ihr neues Amt sowie ihre weitere politische Karriere.

Doch nun wünscht Ihnen die gesamte Quartl-Redaktion einen schönen Sommer und zunächst natürlich viel Spaß mit der neuesten Ausgabe Ihres Quartls!

Hans-Joachim Bungartz.

Iterationsschleife

N=54
06.06.2025

Der Großvenediger ist mit 3657 Metern der höchste Berg Salzburgs, eines Bundeslandes Österreichs, das erst seit dem 1. Mai 1816 zu Österreich gehört[a]. Er wurde am 3. September 1941 erstmals bestiegen. Von 40 Männern, die sich in Neukirchen auf den Weg gemacht hatten, um den Anstieg über das Obersulzbachtal zu führen, erreichten 24 den Gipfel. Unter ihnen war auch der 70-jährige Paul Rohregger. Heute fährt man mit dem Auto bis auf ca. 1100 Meter Höhe und macht sich von dort auf den Weg. Natürlich kann man – wie die Mountainbiker mit den ans Fahrrad gehängten Skiern – der Straße weiter folgen und bis zur Postalm fahren[b].

Aber als Wanderer schlägt man sich links in die Büsche und überquert zuerst einmal den etwa 30 Meter breiten Obersulzbach auf einer Hängebrücke. Auffallend sind die vielen Roten Steine im Bachbett. Die rötliche Färbung kommt aber nicht vom Stein selbst, sondern entsteht durch die Veilchen-Steinalge. Sie blüht in der dauerfeuchten Luft und färbt die Steine leuchtend rot.

[a] Wolfgang Amadeus Mozart (1756 – 1791) ist damit, entgegen anderslautender Behauptungen der österreichischen Fremdenverkehrswerbung, mitnichten ein Österreicher.
[b] Auf dieser Strecke wird einem der Sinn eines SUV erstmals vor Augen geführt.

Nach der Brücke geht es kurz nach links aber dann steil nach rechts. Durch den Wald steigen wir bei steigendem Puls auf bis zur Kampriesenalm auf 1415 Meter Höhe. Der Weg führt uns durch den Nationalpark Hohe Tauern – wenn auch nur durch die Außenzone. Der Wald wird nicht bewirtschaftet, aber auf den Almen weiden Kühe. Die Alm bietet uns ein Mittagessen. Wir genießen die Ruhe die sich im Tosen des Obersulzbach ausdrückt. Russland und die Ukraine, der Streit um Harvard, die Diskussionen über EU, Bundesregierung und Gigafactories sind weit weg.

Es gibt nichts zu sagen und nichts zu schreiben.

Wir beschließen, dass der Großvenediger noch warten kann, denn auch Paul Rohregger musste erst 70 werden, um ihn zu besteigen, und nehmen den linksseitigen Weg zurück entlang des Obersulzbach. Der Tag hat noch genug Zeit um sich mit „The Question Concerning Technology in China“ von Yuk Hui[a] zu beschäftigen. Der Rest muss warten.

M. Resch

[a] Yuk Hui, Chinesischer Philosoph, derzeit Professor an der Erasmus Universität Rotterdam, Promoviert bei Bernard Stiegler (1952 -2020) und beeinflusst von Keiji Nishitani (1900 – 1990).

Programming the Computing Continuum

1 Introduction

The *Internet of Things (IoT)* is extending global communication to devices. Huge amounts of data are collected, stored, and processed for new applications in almost every area of life. Smart house, smart city, autonomous driving, Industry 4.0, and IT-based health care are only a few applications of the Internet of Things.
The IT infrastructure of IoT consists of heterogeneous systems of all sizes, ranging from microcontrollers in sensors/actuators and capability-constrained servers on the edge of the Internet up to cloud data centers with powerful servers even equipped with multiple GPU accelerators for the growing demand of AI workloads. IoT applications combine data collection and processing on resources of the different layers of the *Edge Cloud Continuum* to enable scaling in combination with data privacy and guaranteeing *Service Level Objectives (SLOs)*.

This paper proposes an extension of *serverless computing or Function-as-a-Service (FaaS)* known from today's cloud systems to the entire continuum. Serverless computing enables the use of cloud data centers without having to manage your infrastructure. The application logic is given as event-triggered stateless functions accessing managed backend services. The term serverless indicates, that the deployment of the application does not require any resource management. Function invocations are executed on the servers of the FaaS cloud service and the backend services are also managed by the cloud provider. The startup latency of invocations is a central challenge of FaaS and many techniques for mitigating the *cold start* have been investigated. Instead of microservices, serverless computing supports scale to zero if no events are incoming.

The first FaaS platform offered by a cloud provider was Amazon Lambda[1]. Function invocations are generated by http requests through an application gateway, by other AWS services such as S3 when uploading a new object or when a new MQTT message arrives. Resource management for function invocations is elastic based on the memory allocated to functions. The compute power and network capacity for functions scale with the allocated memory. The accounting is based on memory seconds charging individual function invocations in a pay-per-use model. While all the big cloud providers offer FaaS nowadays, also open source FaaS platforms such as OpenWhisk, OpenFaaS, and Knative are available.
Technical University of Munich extended FaaS to heterogeneous FaaS platforms in the *Function Delivery Network (FDN)* [1, 2]. Function invocations are scheduled to a federated set of FaaS platforms at different cloud providers as well as on the edge so that various cost functions, e.g., energy efficiency, response time, or failure rate, are optimized. The system is based on central scheduling in a dedicated management cluster.

The *Serverless IoT Framework* [3] presented here goes beyond the FDN for IoT applications in various aspects:

1. Serverless programming of IoT devices: The devices will be programmed in an event-triggered approach.

2. Invocation optimization across time and space: SIF allows to migrate invocations in space across the different layers of the continuum to find the best resources and batches computations for energy optimization and reduction of cold start time.

3. Automatic data management: Data will be managed via a distributed key value store which will support invocation migration via prefetching and caching.

[1] aws.amazon.com/lambda

4. Unified application paradigm: Applications on all layers of the continuum will be programmed in an event driven serverless approach.

5. Distributed scheduling: Function invocations can be triggered on all layers from the cloud to the devices and will be managed in a distributed approach surpassing the central concept in FDN.

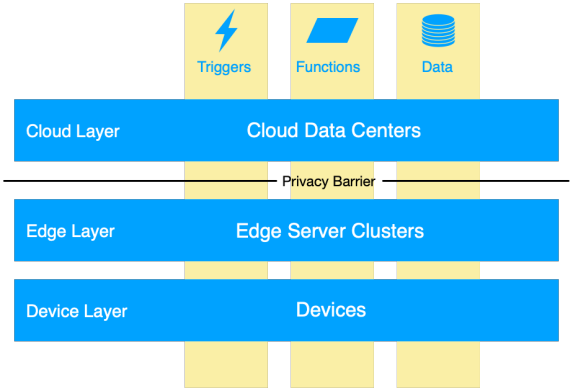


Figure 1: The Serverless IoT Framework spans the whole continuum of IoT devices, edge servers, and the cloud. Applications are implemented in the serverless paradigm based on triggers, functions, and location independent data.

The approach taken in the *Serverless IoT Framework (SIF)* is illustrated in Figure 1. It introduces three different layers of the continuum, i.e., Cloud, Edge, and Device layer. The *cloud layer* consists of the resources of cloud data centers, the *edge layer* of clusters of edge servers, and the *device layer* of the sensors and actuators.
The vertical columns indicate the global programming paradigm. *IoT applications* consists of a *SIF bundle*. A SIF Bundle is composed of *SIF applications* and the corresponding *SIF application data*.
SIF applications are deployed into the layers of the continuum. The *SIF data*

storage stores and manages the SIF application data.
Each *SIF application* consist of events, triggers generating events and functions performing the processing.
Triggers of an SIF application generate events that lead to invocations of subscribing functions.
Events can carry data and are provided as copies to invocations of subscribing functions. They are created via the application's triggers or functions as well as by backend services or other SIF applications.
Functions subscribe for events and perform computation on data in the SIF data storage. Function invocations can be executed on any layer, where the function is deployed. The distributed scheduling optimizes invocation execution depending on the status and the overall objectives.
Data are handled by the *SIF data storage*, a distributed key value store which allows transparent and optimized access by function invocations on any of the layers. Data placement and replication is handled automatically and in coordination with invocation scheduling.
The implementation of SIF is the *SIF platform*. It consists of the *SIF Device Platform (SIF-D)* and the *SIF Edge-Cloud Platform (SIF-EC)*. The platform implements the distributed, data-aware scheduling of invocations and the SIF data storage. The resources for an IoT application given as a SIF bundle will consist of resources provided by the application owner in Cloud FaaS platforms as well as resources provided by the application owner on the edge and device layer.

2 SIF Platform

The SIF platform orchestrates all the components of SIF applications on the provided resources. It consists of two parts, the SIF Device Platform (SIF-D) for IoT devices and the SIF Edge-Cloud Platform (SIF-EC) for edge servers and the cloud. The conceptual structure of both platforms is the same while the implementation mechanisms are specialized for the device and the edge cloud layer. This article introduces first SIF-D and then SIF-EC.

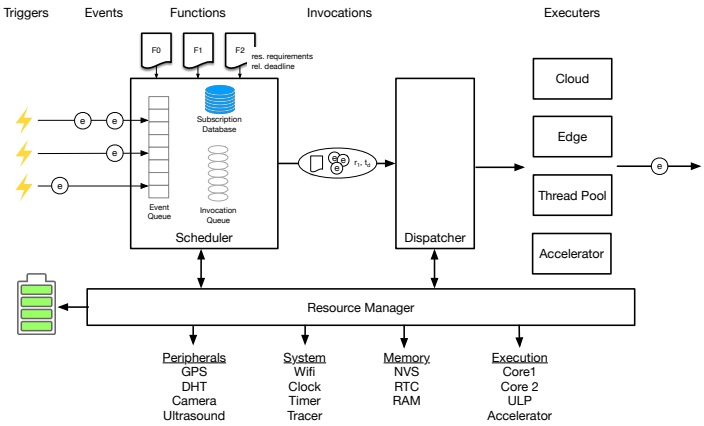


Figure 2: This figure presents the overall architecture of SIF-D. It consists of the scheduler for event and invocation management, the resource manager, and the dispatcher that forwards invocations to executors when all resources are available.

2.1 SIF Device Platform (SIF-D)

SIF-D realizes SIF on microcontrollers used in IoT devices. It was initially implemented for ESP32-based devices based on FreeRTOS and implemented in the ESP IDF in C++ and was recently also ported to Raspberry Pi pico microcontrollers. On the devices, the application components and SIF-D are statically linked as the firmware of the microcontroller. Dynamic deployment of triggers and functions is currently supported by over-the-air updates of the microcontroller's firmware.

Figure 2 outlines the implementation of SIF for IoT devices. The figure shows the main components implemented in the ESP32 SIF framework. All the components are part of the binary that is flashed to the microcontroller. The application only specifies triggers, events, resources, and functions. The SIF framework provides the runtime system for SIF applications and base

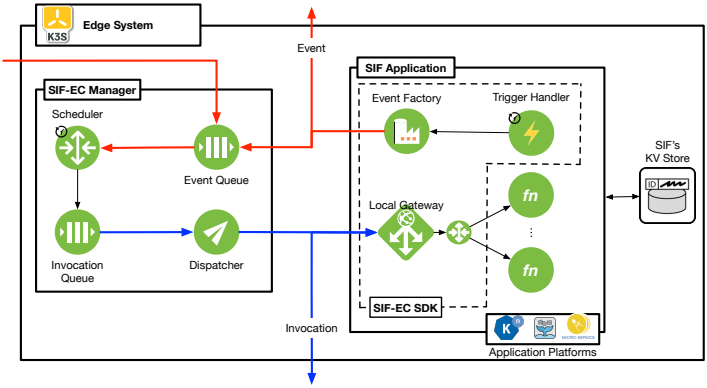


Figure 3: Interaction between remote devices and the SIF-EC for an incoming event or invocation

classes for the application components.
SIF applications use triggers to generate events, that are periodically inserted into the event queue of the scheduler or due to an interrupt. The computation is realized as functions, that subscribe to the events and specify the resource requirements and the deadline for invocations. The scheduler creates invocations from the subscription information, checks the availability of resources and schedules invocations for execution. Once the invocations are scheduled, it is the responsibility of the dispatcher to send the invocations to an executer. On the ESP32, it is a pool of threads that are executed on the two Tensilica cores. Only when the execution is finished, the Dispatcher informs the Resource Manager and the resources are freed again.
Resources can be of type peripherals, system, memory, and execution. The current implementation of SIF on ESP32 has no support for the SIF data storage.

2.2 SIF Edge-Cloud Platform (SIF-EC)

SIF-EC implements the SIF model on the edge and cloud layer. Currently, it is written in Python using OOP and is composed of:

1. SIF-EC Software Development Kit (SDK) and
2. Manager (SIF-EC Manager)

The SDK provides two types of triggers: one-shot and periodic. Furthermore, it includes a base event factory class, which developers extend to accommodate custom event types. Considering the OOP model, the triggers rely on an instance of the event factory. Once the trigger fires, the event fabric creates an event and sends it to the Manager.
Each application is containerized and consists of triggers, events, and functions. The functions can be part of a container and/or deployed as FaaS functions. An application is deployed to an edge server or the cloud by deploying its containers as pods into Kubernetes and FaaS functions into the serverless platforms.
Due to the distributed nature of SIF-EC, functions are executed upon HTTP requests. If the function is part of the container, the container provides a `LocalGateway`, offered by the SDK, that allows the SIF-EC Manager to execute function invocations in the container. If it is a FaaS function, the SIF-EC Manager executes the function on the FaaS platform via the gateway of the FaaS platform. The `LocalGateway` is also used to inform the SIF-EC Manager about available deployments of functions. For example, in Listing 1, Line 14 deploys `fn-name` into the container and informs the Manager about it. Line 20 deploys the same function into the available FaaS platform.
Figure 3 shows an edge cluster with an SIF-EC application. Within the SIF-EC Manager, the scheduler creates and manages function invocations. It also manages registered functions and their event subscriptions, available endpoints, and incoming events. Upon event arrival, an event is placed in an event queue. From this queue, the scheduler creates an invocation for subscribing functions. Invocations can be in a waiting state iff a function

subscribes to more than one event. Once all necessary events arrived, the invocation becomes ready for execution.

```
# Function example to be invoked by an event
def evt_handler(data):
    # Do something with
    # incoming data

# Instantiates the application router
app = LocalGateway(
   sch="http://sch.sif.svc.cluster.local:9000"
   )

# Instantiates the FaaS Platform router
# It enables deploying serverless functions dynamically
knative = KnativeGateway()

# Registers a function as microservice by creating
# a local route in the form /api/evt_handler
app.deploy(evt_handler, "fn-name", "TriggerEvent")

# Register a serverless platform handler
app.registerHandler(knative)

# Registers a function as FaaS using the available
# platform handlers
app.deployFn("docker.io/<user>/fn_handler:latest",
    "fn-name", "TriggerEvent"
    )

# Instantiates a custom event factory
evt_fb = ExampleEventFactory()

# Instantiates a 30 sec periodic trigger
trg = PeriodicTrigger(evt_fb, "30s", "1m")
```

Listing 1: SIF-EC application using the available SDK.

If a function is deployed in multiple platforms, e.g., a container or FaaS, the scheduler selects the best deployment and forwards it to the dispatcher for execution. Currently, it uses weighted round-robin scheduling where the weights are automatically adapted to the capacity of the different deployments. The dispatcher forwards the invocation to the target, monitors the execution, and restarts it in case of a failure.

Although the same container produces and implements the function that subscribes to `TriggerEvent`, the event is sent to the Manager to allow for resource-aware scheduling. In the spirit of SIF, events can be used to synchronize applications across clusters. Therefore, events cannot only be sent to the local Manager but also remote ones. Invocations can be forwarded to any container and/or FaaS platform in the Continuum where the function is deployed. Such behavior also enables collaboration between IoT devices and edge-cloud resources.

3 Senior Homecare Application

Many countries do have an aging population. In Germany the very few places in nursing homes with very high monthly costs do not meet the demand for seniors requiring support. Furthermore, seniors prefer to stay in their private homes as long as possible to have an independent life. TUM is developing a digital twin for seniors staying at home to support caregivers - frequently relatives - in their caring for their elders.
The digital twin is realized on an edge computer located at the senior's home, currently a Raspberry Pi is used, which needs not be connected to the internet. It is implemented as a SIF bundle and deployed via a set of containers on the edge system. Figure 4 introduces the main components of the digital twin. The components are organized in three layers: resource, data, and operational layer. The resource layer provides resources such as databases (non-SQL and SQL), the FaaS platform Knative and the underlying Kubernetes implementation. The data layers provides the raw data and the trained models describing the senior's behavior as well as data about the status of the digital twin and the underlying hard- and software. The operational layer provides four components, monitoring, modeling, actuation, and visualization. The modeling components provides functions to create behavioural models from the raw data. These functions are triggered by the monitoring component which is the overall control component of the digital twin. Triggers are implemented here that periodically start modeling

functions or analyses to detect anomalies. The result of these analyses might trigger actuation functions that forward information to the visualization component or directly inform the caregiver. The visualization component offers a todo list and an information list which provide information to the caregiver.

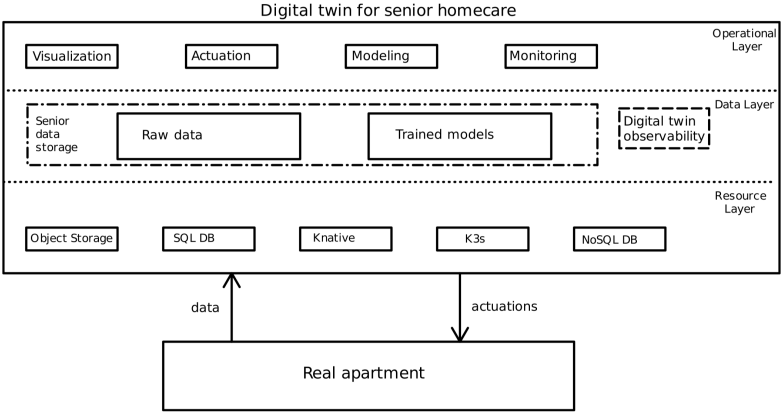


Figure 4: This figure presents the overall architecture of the Senior Homecare Digital Twin.

As sensors, we currently provide PIR-based room sensors capturing room occupancy. Furthermore, we developed a smart bottle capturing drinking behaviour and are working on smart slippers for gait analysis together with the Technical University of Denmark.

Michael Gerndt, Isaac Nunez

References

[1] Mohak Chadha, Anshul Jindal, and Michael Gerndt. Towards federated learning using faas fabric. In Proceedings of the 2020 Sixth International Workshop on Serverless Computing, WoSC'20, page 49–54, New York, NY, USA, 2020. Association for Computing Machinery, Association for Computing Machinery.

[2] Anshul Jindal, Michael Gerndt, Mohak Chadha, Vladimir Podolskiy, and Pengfei Chen. Function delivery network: Extending serverless computing for heterogeneous platforms. CoRR, abs/2102.02330, 2021.

[3] Isaac Nunez, Michael Gerndt, and Shajulin Benedict. Serverless iot framework. In 30th International Workshop on High-Level Programming Models and Supportive Environments (HIPS 2025), 2025.

BGCE 2025: Eröffnungswochenende am Sylvensteinspeicher

Elitenetzwerk Bayern

Wie jedes Jahr traf sich vom 25. bis 27. April 2025 die ganze Familie der Bavarian Graduate School of Computational Engineering (BGCE) zum Auftaktwochenende im Seminarzentrum Jäger von Fall am Sylvensteinspeicher.

Abbildung 1: Die ganze BGCE-Familie auf einem Bild: Juniors, Seniors, Koordinatoren, und Professoren. © Michael Wiedemann

Los ging es kurz nach Mittag am Freitag, als die 16 Auserwählten für den BGCE Jahrgang 2025 am Sylvensteinsee eintrafen. Es folgte eine kurze Einführung in die Geschichte, das BGCE-Programm und dessen Ausgestaltung durch Prof. Dr. Hans-Joachim Bungartz, gefolgt von der Präsentation der organisatorischen Details des Programms durch die Koordinatoren.

Der administrative Teil des Programms war damit abgeschlossen und die neuen *Juniors* von FAU und TUM konnten sich im Soft Skill Seminar *When Teamwork works* besser kennenlernen, angeleitet durch professionelle Coaches. Damit verging der Nachmittag wie im Flug, und das Programm belohnte alle Teilnehmer mit einem üppigen Buffet zum Abendessen. Hier stießen schließlich auch die *Seniors*, der BGCE Jahrgang 2024, zu uns dazu. Der erfolgreiche erste Tag fand seinen Ausklang mit einem Kaminvortrag zum Thema Chip-Design, der von BGCE-Alumni Michael Rippl (Infineon) gehalten wurde. Das Thema stoß auf großes Interesse und es folgten lange Unterhaltungen am Kaminfeuer.

Der nächste Tag begann für den Senior Jahrgang mit dem ganztägigen *Step Out* Programm. Im Angesicht der majestätischen bayerischen Alpen lehrt die Outdoor-Experience den Studenten elementare Werte wie Zusammenarbeit und Vertrauen.

Abbildung 2: Auf dem Gipfel ist es doch am Schönsten - Die BGCE Seniors ganz oben. © Michael Wiedemann

Für die Junioren begann der Tag etwas ruhiger, aber nicht minder spannend mit einem Kennenlernen zwischen ihnen, den assoziierten Professoren und Koordinatoren des BGCE-Programms. Im Vordergrund stand dabei der Austausch von Erfahrungen und das Angleichen der Erwartungen. Am Nachmittag schlossen sich dem Dialog dann wieder Soft Skill Kurse durch die erfahrene Coaching-Truppe um Michael Wiedemann an.

Der letzte Tag beinhaltete ein gemeinsames Programm für Seniors und Juniors namens *Group Challenge* und *Consulting Circle*. Bei ersterem mussten die beiden Jahrgänge gemeinschaftlich ein komplexes Problem in kürzestmöglicher Zeit lösen. Bei der zweiten Aktivität stand dagegen der Erfahrungsaustausch im Vordergrund. Die Seniors konnten so den Junioren erfolgreich ihr Wissen um das umfassende Angebot des BGCE-Programms vermachen.

Abbildung 3: Beide Jahrgänge bei einer der Gruppenaktivitäten am Sonntag. © Michael Wiedemann

In diesem Sinne war 2025 das letzte BGCE-Eröffnungswochenende seiner Art, da der Jahrgang 2025 das Programm nun komplettiert. Und obwohl es nächstes Jahr damit anders sein wird, kann man doch schon festhalten, dass der neue Jahrgang für die kommenden zwei sicherlich aufregenden letzten Jahre im BGCE-Programm bestens gerüstet ist.

Jonas Schuhmacher, Tobias Neckel

Wissenschaftsminister Blume zu Gast am Zentrum für Nationales Hochleistungsrechnen Erlangen

Am Zentrum für Nationales Hochleistungsrechnen Erlangen an der Friedrich-Alexander-Universität Erlangen-Nürnberg (NHR@FAU) ist der Aufbau des neuesten Supercomputers mit modernsten NVIDIA-GPUs abgeschlossen. Zu den bereits 2024 installierten 384 H100-GPUs (wir berichteten im Quartl 4/2024) gesellten sich im April 2025 weitere 384 H200-GPUs. Bayerns Wissenschaftsminister Markus Blume gab im Mai den Startschuss für das aktuell leistungsfähigste KI-Cluster an Hochschulen in Deutschland. Der Name: „Helma“ – benannt nach Wilhelmine von Brandenburg-Bayreuth, Gattin des FAU-Gründers Friedrich III. von Brandenburg-Bayreuth.

Der Vorhang lüftet sich langsam – Helma als Wegbereiterin des Bayerischen Basismodells?!

Wissenschaftsminister Markus Blume: „In Erlangen investieren wir nicht nur in Beton, sondern auch in Künstliche Intelligenz: Am NHR@FAU sind die ersten 384 GPUs des Typ NVIDIA H200 für die Entwicklung des Bayerischen KI-Basismodells installiert – ein weiterer bedeutender Beitrag beim Aufbau einer leistungsstarken KI-Rechnerinfrastruktur für unsere Hochschulen. Das ist erst der Anfang, denn weitere GPUs folgen. Mit herausragend besetzten KI-Professuren verfügen unsere bayerischen Hochschulen über die erforderliche Expertise. Nun stellen wir ihnen eine für die Entwicklung und das Training von KI-Modellen optimierte Technik zur Verfügung. Das ist eine bundesweit einmalige Initiative. Denn für uns ist klar: Bei Schlüsseltechnologien vorne mit dabei zu sein, ist die Grundlage unserer Souveränität und unseres Wohlstands in der Zukunft.“

Abbildung 1: Helma ist benannt nach Markgräfin Wilhelmine von Bayreuth, eine der wichtigsten Gründungsfiguren der FAU (Bild: G. Iannicelli, FAU)

Abbildung 2: Der Bayerische Wissenschaftsminister Markus Blume und Gerhard Wellein werfen einen Blick ins Innere des Superrechners Helma (Bild: G. Iannicelli, FAU)

Dass in Bayern – und natürlich auch in Franken – bundesweit Einmaliges passiert, daran sind wir ja fast schon gewöhnt. Dass es in Zeiten knapper Kassen und gekürzter universitärer Budgets geschieht, freut uns und die KI-Forschenden. Gleichzeitig müssen wir uns in besonderer Weise am Mehrwert dieser Investitionen messen lassen. Im Rahmen der BayernKI, für die ja schon 320 H100-GPUs am NHR@FAU und am LRZ beschafft wurden (siehe (Quartl 1/2025) sehen wir bereits den Mehrwert in einer breiten Nutzung über zahlreiche bayerische Hochschulen und Universitäten. Nun geht der Freistaat den ersten Schritt hin auch zu den Spitzenanwendungen im KI-Bereich – hoffentlich weiterhin Hand in Hand mit der Breite der Nutzerschaft. In beiden Fällen soll es nur ein erster Schritt sein, „denn weitere GPUs folgen“. Stay tuned!

Weitere Informationen zum neuen Meilenstein der KI-Infrastruktur für bayerische Forschende: https://go.fau.de/1bidb

Gerhard Wellein

KONWIHR Project: Optimizing LuKARS

In a collaborative effort of German and French universities, the LuKARS model has been developed since 2018 [1]. The LuKARS model is a semi-distributed hydrological model which simulates the water flows through karst aquifier systems making use of buckets and transfer functions. The LuKARS model was developed to improve the simulation of complex karst groundwater systems, which are difficult to model due to their heterogeneous and dynamic nature. By incorporating multiple hydrotopes—landscape units with similar land use and soil characteristics—LuKARS enables a more accurate representation of spatial variability within karst catchments.

LuKARS has been successfully applied in a range of studies. Applications include assessing land-use change impacts in the Kerschbaum catchment [1], comparing land cover and land use effects in lumped parameter models for forested karst areas [2], and linking hydrological and hydrochemical processes using hydrochemical data [3]. The model has also supported uncertainty quantification through active subspace methods [4] and has been used to evaluate prediction uncertainties tied to meteorological forcing and the soil-vegetation-atmosphere continuum [1].

LuKARS faces scalability issues with long time series, large datasets, and advanced analysis, which require many model realizations. In semi-distributed setups with multiple hydrotopes, the high parameter count leads to increased computational demand, limiting its effectiveness for extensive sensitivity and uncertainty assessments. For this reason, in 2024 the Applied Geology and Environmental Systems Modeling group of FAU applied for a KONWIHR project to optimize the performance of the original model [1]. At first, the Numba just-in-time compiler for Python was integrated. By compiling the Python code to close-to-machine code, the performance was increased by a large factor (1.14 seconds to 0.03 seconds) with a 420-fold reduction in runti-

me. With recommendations from NHR@FAU through KONWIHR about the data handling in the innermost loops, the runtime was further reduced to 0.01 seconds. However, Numba compiles the functions every time the application is called; this code generation takes multiple seconds and Numba's feature to cache generated kernels did not work. Morris sampling [6], a method used to explore the parameter space and identify those parameters with the greatest impact on model performance, was applied by evaluating the agreement between simulated and observed spring discharge. With Morris sampling, a large number of input variants could be covered in one program run, mitigating the JIT compilation issue.

The benefit of LuKARS 3.0's fast computation time was evident during this analysis: Approximately 220,000 simulation runs for the Morris analysis were completed in just 30 minutes. In comparison, a previous study [4] required 4.3 hours for 43,000 simulations. This computational efficiency enables extensive parameter space exploration, supporting more robust sensitivity analysis and uncertainty quantification. Furthermore, it facilitates the application and comparison of multiple sensitivity analysis approaches within practical timeframes.

In 2025, the group applied again for a KONWIHR project to couple LuKARS 3.0 with the library PHREEQC [7] to incorporate aqueous geochemical calculations, mixing and dilution processes. The coupled flow-transport model was applied for the test case of the Kerschbaum spring in Austria considering a complete mixing approach. The coupling was performed sequentially, meaning that first the flow equations are solved for each bucket and time step, and then the reactive transport is computed in each bucket. The reactive transport is solved at each time step based on the mixing percentage of water contributions from the other buckets of the model. Each water contribution is represented in PHREEQC by the individual keyword SOLUTION, for which PHREEQC computes chemical speciation. The model was further extended to an arbitrary number of model buckets and flow components. Finally, the user friendliness of the code was improved by allowing the modification of model structure, parameters, and other inputs by means of a configuration

file without the need to further modify the model script.

[1] Bittner, D., Narany, T. S., Kohl, B., Disse, M., Chiogna, G. (2018). Modeling the hydrological impact of land use change in a dolomite-dominated karst system. Journal of Hydrology, 567, 267-279.

[2] Sivelle, V., Jourde, H., Bittner, D., Richieri, B., Labat, D., Hartmann, A., and Chiogna, G. (2022). Considering land cover and land use (LCLU) in lumped parameter modeling in forest dominated karst catchments, Journal of Hydrology, 612(C), 128264. https://doi.org/10.1016/j.jhydrol.2022.128264

[3] Richieri, B., Bittner, D., Sivelle, V., Hartmann, A., Labat, D., and Chiogna, G. (2024). On the value of hydrochemical data for the interpretation of flow and transport processes in the Baget karst system, France. Hydrogeology Journal. https://doi.org/10.1007/s10040-024-02801-2

[4] Teixeira Parente, M., Bittner, D., Mattis, S. A., Chiogna, G., Wohlmuth, B. (2019). Bayesian calibration and sensitivity analysis for a karst aquifer model using active subspaces. Water Resources Research, 55(8), 7086-7107.

[5] Morris, M.D., 1991. Factorial Sampling Plans for Preliminary Computational Experiments. Technometrics 33, 161-174. https://doi.org/10.1080/00401706.1991.10484804

[6] https://www.usgs.gov/software/phreeqc-version-3

Thomas Gruber, NHR@FAU
Beatrice Richieri & Gabriele Chiogna, Geozentrum Nordbayern

Graduate School Week 2025
Kurse - Kommunikation - Kultur

Die NHR Graduate School ist ein strukturiertes Stipendienprogramm für talentierte Doktorandinnen und Doktoranden in den Bereichen HPC und Informatik. Es wird von den neun NHR-Zentren organisiert und bietet seinen „Fellows“ eine Finanzierung für drei Jahre, professionelle Betreuung, einen sechsmonatigen Aufenthalt an einem anderen Zentrum und außerdem jedes Jahr eine spannende Kurswoche an einem der NHR-Zentren. Dieses Mal hatte das NHR@FAU die Ehre, die Veranstaltung auszurichten, und darf auf eine erfolgreiche und sehr intensive „NHR Graduate School Course Week“ zurückblicken, in der aber auch der Spaß nicht zu kurz kam.

Vom 19. bis 23. Mai gab es für 21 Fellows aus ganz Deutschland neben einem vollgepackten Programm zu HPC- und KI-Themen quer durch verschiedenste Anwendungsgebiete, Networking pur und intensives Mentoring durch erfahrene HPC-Profis. Drei der NHR-Zentren (NHR@TUDa, NHR@ZIB und NHR@FAU) führten in zwei parallelen Tracks Kurse mit Vorträgen und praxisorientierten Arbeitsphasen durch: GPU-Programmierung, Codegenerierung für HPC, moderne KI-Frameworks und CI/CD-Workflows. Zum besseren Kennenlernen stellten die Stipendiatinnen und Stipendiaten gleich zu Beginn der Kurswoche ihre eigenen Forschungsprojekte im Rahmen einer Postersession vor und tauschten sich lebhaft untereinander aus.

Am Dienstag und Donnerstag ließ auch das unterhaltsame Rahmenprogramm keinen Raum für Langeweile: Die Promovierenden lauschten den Rechenvorgängen der einzigen noch funktionsfähigen Zuse Z23 aus dem Jahr 1962, die in der Informatik-Sammlung Erlangen (ISER) ihre Heimat gefunden hat, ließen sich durch die Nürnberger Altstadt führen und erkundeten die Geschichte des Brauwesens in den alten Bierkellern am „Berch“, dem kultigen (und historischen) Wahrzeichen der Bierproduktion in Erlangen.

Abbildung 1: Die Teilnehmenden der NHR Graduate School Kurswoche im Innenhof des RRZE, Bild: M. Drees, NHR-Verein

Abbildung 2: Die Fellows präsentieren ihre Forschungsprojekte in einer Postersession, Bild: J. Carl, NHR@FAU

Wir bedanken uns herzlich bei allen, die dabei waren – bei den engagierten Lehrkräften, den vielen helfenden Händen, dem NHR-Verein und natürlich ganz besonders bei unseren Stipendiatinnen und Stipendiaten für das tolle Engagement im Kursprogramm. Ein begeisterter Teilnehmer beschrieb die Woche als „unglaublich gut organisiert, intellektuell bereichernd und sehr einladend“. Man kann sich also kaum einen besseren Ort vorstellen, um gemeinsam zu lernen, Gespräche zu führen und Spaß zu haben, als die NHR Graduate School Course Week.

Mehr Informationen zur NHR Graduate School: `https://www.nhr-verein.de/die-graduiertenschule`

Judith Carl

LRZ nimmt HPC-System CoolMUC in Betrieb

Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften

Am LRZ ist ein neues HPC-Cluster in Betrieb gegangen: CoolMUC vereint die Leistung diverser Prozessorarchitekturen und steht Forschenden an bayerischen Hochschulen ab sofort zur Verfügung. Insgesamt 119 Rechenknoten bieten geballte Performance, darunter 106 mit Intel Sapphire Rapids-Chips, die mit 512 Gigabyte Arbeitsspeicher und 112 Rechenkernen besonders für parallele Anwendungen mit geringem Speicherbedarf geeignet sind. Für datenintensivere Jobs kommen die 12 Ice-Lake-Knoten zum Einsatz, mit jeweils 80 Rechenkernen sowie einem Terabyte Kurzzeitspeicher. Ergänzt wird das System durch "Teramem", einem Knoten mit Cooper-Lake-Prozessor, 96 Rechenkernen und rund 6 Terabyte Arbeitsspeicher. Der aktuelle CoolMUC wird demnächst mit GPUs ergänzt und bis 2026 durch weitere CPUs ausgebaut. Neben den klassischen HPC-Disziplinen nutzen immer weitere Fachbereiche wie Biologie, Umwelt- und Lebenswissenschaften oder Sozial- und Wirtschaftswissenschaften das Rechencluster für ihre Forschung. Weitere Informationen: `https://www.lrz.de/news/detail/coolmuc-system`.

Susanne Vieser

Digitale Zwillinge programmieren

Das menschliche Herz, die Lunge oder das Hirn berechnen und digital modellieren: Für solche Aufgaben empfiehlt sich das Open Source-Programmpaket deal.II, das auf Basis von C++ rund 600.000 Zeilen Code umfasst und viele mathematische Werkzeuge zur Entwicklung innovativer Solver für partielle Differenzialgleichungen bietet. Auf deren Basis lassen sich Strömungen von Flüssigkeiten und Gasen berechnen oder Deformationen von Festkörpern modellieren: die Grundlagen für den Aufbau digitaler Zwillinge von Organen. Martin Kronbichler hat diese Programmbibliothek mitentwickelt, der Mathematiker lehrt an der Ruhr-Universität in Bochum und leitet ein interdisziplinäres Projekt: Mit 12 Hochschulen und Forschungseinrichtungen, darunter das Leibniz-Rechenzentrum (LRZ), werden deal.II, damit entwickelte Codes sowie Anwendungen an die nächste Generation heterogener Supercomputer auf Exascale-Niveau angepasst. So soll „dealii-X“ entstehen – ein Exascale-Framework zur Erschaffung digitaler Zwillinge des menschlichen Körpers. Mehr dazu `https://www.lrz.de/news/detail/digitale-zwillinge-programmieren`.

Gerhard Wellein

KI-Modelle: Rechnen statt trainieren

Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften

Künstliche Neuronale Netze werden immer größer und komplexer – und brauchen daher für das Training mit Daten immer mehr Energie: Ein Forschungsteam um Felix Dietrich, Professor für Physics-Enhanced Machine Learning an der Technischen Universität München (TUM), arbeitet daran, Trainingsschritte durch mathematische Methoden zu ersetzen. Das senkt den Strombedarf und hilft, die Funktionsweise von Künstlicher Intelligenz besser zu verstehen.

„Wir ersetzen iterative Trainingsschritte durch probabilistische Berechnung und suchen dazu gezielt Werte in den Datensätzen, die sich durch Änderung von Parametern besonders stark und schnell ändern“, erklärt Dietrich das Verfahren in einem Interview mit dem Leibniz-Rechenzentrum (LRZ). Durch die Berechnungen wachsen zwar die Künstlich Neuronalen Netzwerke, dafür werden aber zig-tausende Trainingsläufe unnötig. Zurzeit funktioniert die Strategie bei einfachen Feedforward und Recurrent Networks, die beim maschinellen Lernen für Zeitreihen- und Tabellendaten eingesetzt werden, sowie bei Graphmodellen für die Bild- und Mustererkennung.
Zurzeit arbeitet das Team um Dietrich an mathematischen Lösungen für Convolutional sowie Attention Layers, die beim Entwickeln von generativer

Abbildung 1: Prof. Dr. Felix Dietrich, Bildnachweis: Andreas Heddergott, TUM

KI von Bedeutung sind. Forschenden, die mit KI Daten auswerten wollen, rät der Mathematiker „das Rad nicht stetig neu zu erfinden, sondern auf bereits existierende Modelle zu setzen. Auch für eigene, neue Datensätze können solche bestehenden Modelle angepasst und austariert werden. Das ganze Interview finden Sie hier auch auf Englisch: `https://www.lrz.de/news/detail/ki-modelle-trainingsschritte-berechnen`.

Susanne Vieser

KONWIHR: New projects from round 2025-1

The competence network for scientific high-performance computing in Bavaria welcomes the new projects that succeeded in the application round 2025-1. In every round, we accept proposals for "normal" (up to 12 months) and "small" (up to 3 months) projects, as well as "basis" projects to establish contact points. In this round, the following projects were funded:

- *Auto-calibration of physics-based fully-distributed model*
  by Dr. Xinyang Fan,
  Applied Geology and Modeling of Environmental Systems, FAU Erlangen-Nürnberg

- *Efficient Training Principles for Small Neural Networks on High-Performance Computing(HPC) Systems*
  by Prof. Dr. Andreas Kist,
  Artificial Intelligence in Communication Disorders, FAU Erlangen-Nürnberg

- *High-performance parallel I/O for SIMSON - a pseudo-spectral solver for incompressible boundary layer flows*
  by Prof. Dr. Philipp Schlatter,
  Chair of Fluid Mechanics, FAU Erlangen-Nürnberg

- *Exa-scale simulations of turbulent pipe flow on trillions of grid points using thousands of GPU nodes*
  by Prof. Dr. Philipp Schlatter,
  Chair of Fluid Mechanics, FAU Erlangen-Nürnberg

- *Performance insights into Julia-based codes at HPC scale*
  by Dr. Valentin Churavy
  High-Performance Scientific Computing, University of Augsburg
- *Bringing MGLET's particle module to GPUs*
  by Prof. Dr.-Ing. Michael Manhart
  Associate Professorship of Hydromechanics, TUM

You can find more details about these projects at

`https://www.konwihr.de/konwihr-projects/`

From Kyushu to Hokkaido!

日本からの温かい歓迎。 A warm welcome from Japan! Even this simple sentence eclipsed my knowledge of Japanese when I started my journey to my research stay on the 25th of January. Ten weeks, eight for work, two for vacation, later, it still does. This is the story about my research stay at the RIKEN Centers for Computation Science (R-CCS), where I was hosted by Miwako Tsuji in the Quantum-HPC Hybrid Software Division (a mouthful, I know). Speaking of which, officially RIKEN is called.

国立研究開発法人理化学研究所 (Kokuritsu Kenkyū Kaihatsu Hōjin Rikagaku Kenkyūsho), which thankfully is shortened to the first syllables of the last two words 理研。 RIKEN is located in Kobe, a port town near the bay of Osaka, which is well known for its rich culture, sake (rice wine) districts, and Kitano, a district where houses mimic European architectural styles. My travels were unsurprisingly uneventful. Planes – depart on time; airport busses – depart on time; trains (local, rapid, high-speed) – depart on time. Google Maps even recommends which train compartments align best with stairs leading to the exits you need to take.

とても便利. The only difficulty is navigating multi-story train stations; Multiple train operating companies co-exist but are separated by gates for which you need a public transport card. Once you figure out the system, it is a breeze, and connections are available every two minutes at peak hours. After 20 hours of traveling, including layovers, I arrived at my target destination: 7-chōme-1-21 Kitanagasadōri, Chuo Ward, Kobe, Hyogo 650-0012, Japan. My luxurious apartment of $20m^2$ is pictured in Fig. 1. To my advantage, it had a mattress and bed and not a traditional futon, allowing me to fully rest for my internship starting the next day.

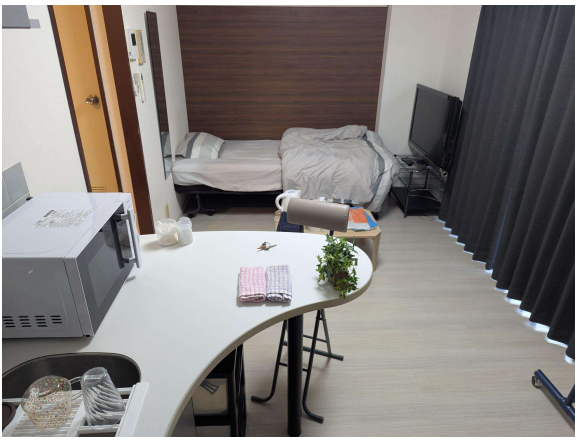


Figure 1: My entire apartment.

R-CCS is located on an island at the edge of the city, where it hosts one of the largest supercomputers FUGAKU. It is a world-leading facility in HPC, and as the name of my division suggests, starting to investigate quantum solutions. As part of their quantum initiative, they have procured multiple quantum systems and run large-scale simulations. Experiments[2] have shown the utility of using hybrid systems. Quantum devices pose unique challenges, that the software stack has to account for. One of the challenges is the variety and intricacies of different quantum technologies. The QDMI[3]., spearheaded by the LRZ, proposes a solution to this exact problem. My goal for the research stay was to integrate and adapt QDMI to the system at RIKEN. Work in Japan follows strict social norms; fortunately, RIKEN is very progressive. When arriving at the office everyone will be greeted (including facility management).

[2] arXiv:2405.05068
[3] QDMI Github

おはようございます！ . Lunch is set from 12:00 to 12:50 (as stipulated in the contract), and the day ends with saying goodbye to everyone. Interaction outside of work also follows social norms. My welcome dinner was set at an Okonomiyaki (see Fig. 2) restaurant, aimed at overcoming the initial awkwardness by "sampling" the local spirits and their varieties. After that, work continued mostly smoothly, although communication worked best by email as written text was translated significantly faster. I successfully integrated QDMI into the Super Quantum Software stack, laying a foundation for future collaborations between RIKEN and the LRZ.

Figure 2: Hiroshima Okonomiyaki including soba and scallops.

I would be lying if I said I solely traveled to Japan for research purposes. Japanese culture exhibits an interesting duality; on one side, upholding century-old traditions and mastering skills for years. Mastery covers a vast area of life: tee ceremonies, calligraphy and penmanship, cooking and baking, tailoring, etc. On the other side, they aspire to include as many Western ideas (especially French) as possible. I dare any true-blooded Bavarian to visit the New München beerhall in Kobe. Also, for anyone considering real bread a

staple food, Japanese baking is largely underwhelming (the pastries are great though). Only by sheer coincidence did I find a decent sourdough on a day trip to Hiroshima. The infamous Shinkansen, pictured in Fig. 3 sporting a Hello Kitty paint job,

かわいい！ , allows a thorough exploration of major cities all over Japan.

Figure 3: The rare Hello Kitty Shinkansen (bullet train) at Hiroshima Station.

The trip to Hiroshima, for example, only takes about 1.5 hours from Shin-Kobe. Concluding my successful adventure, I enjoyed the diversity Japan has to offer. Swimming in the 10°water on a beach visit in Fukuoka; skiing in the famous powder in Sapporo; seeing the cherry blossoms at Ueno Park.

次回まで、日本。

Philipp Seitz

* Notiz * Notiz * Notiz *

Termine 2025

- **Upcoming SIAM Conferences & Deadlines**
  `https://www.siam.org/conferences/calendar`
- **Supercomputing 2025:**
  The International Conference for High Performance Computing, Networking, Storage and Analysis – SC 25 in St. Louis, Missouri, USA: 16.11.-21.11.2025
  `https://sc25.supercomputing.org`
- **KONWIHR News** `https://www.konwihr.de`
- **Durham HPC Days** `https://edin.ac/3DbxZlv`

.

.

.

Quartl* - Impressum

Herausgeber:
Prof. Dr. A. Bode, Prof. Dr. H.-J. Bungartz, Prof. Dr. U. Rüde
Redaktion:
S. Herrmann, Dr. S. Zimmer
Technische Universität München
School of Computation, Information and Technology
Boltzmannstr. 3, 85748 Garching b. München
Tel./Fax: ++49-89-289 18611 / 18607
e-mail: herrmasa@in.tum.de,
https://www.cs.cit.tum.de/sccs
Redaktionsschluss für die nächste Ausgabe: **01.09.2025**

* Quartel: früheres bayerisches Flüssigkeitsmaß,
→ das Quart: 1/4 Kanne = 0.27 l
(Brockhaus Enzyklopädie 1972)