Technical University of Munich | Bildungscampus 2 | 74076 Heilbronn

Heilbronn, September 21, 2025

Konsultationsprozess Rechenzentren

Sehr geehrte Damen und Herren,

Bitte finden Sie anbei meinen Input im Rahmen des Konsultationsprozesses zur Rechenzentrumsstrategie.

Mit Freundlichen Grüßen,

Hartwig Anzt
(Chair of Computational Mathematics)

Technical University of Munich
School of Computation, Information and Engineering (CIT)
Chair of Computational Mathematics

Prof. Dr. rer. nat. Hartwig Anzt
Bildungscampus 2
74076 Heilbronn

Tel: +49 (7131) 26418 - 167
Fax: +49 (7131) 26418 - 167

hartwig.anzt@tum.de
hartwiganzt.github.io
www.tum.de

Bayerische Landesbank
IBAN:
DE10 7005 0000 0000 0248 66
BIC: BYLADEMM
Tax ID: 143 / 241 / 80037
VAT ID: DE811193231

Welche Merkmale und Rahmenbedingungen kennzeichnen aus Ihrer Sicht einen „zukunftsfähigen und leistungsstarken“ Rechenzentrumsstandort Deutschland im Jahr 2030?

Die stärkste Komponente eines Rechenzentrumsstandorts ist auf der einen Seite die Kompetenz und Produktivität der Mitarbeiter, insbesondere Hinsichtlich Nutzer-Supports, und auf der anderen Seite die Softwaretechnologie. Die Hardware spielt eine untergeordnete Rolle. Hardware ist deutlich volatiler als Software, und entwickelt sich schneller. “BigTech” Unternehmen wie Google, Amazon und Meta sind letzten Endes Softwareunternehmen. Selbst Nvidia gründet seinen Erfolg nicht auf die GPU Hardware, sondern auf dem Software Ökosystem welches die Nutzung und Akquise von Nvidia GPUs so attraktiv macht. Wie sich die Hardware- und Systemstruktur von Rechenzentren in den nächsten Jahrzehnten verändern werden ist schwer zu sagen: Neuromorphe Systeme? Quanten-Computer? CMOS? Oder analoge Rechnerarchitekturen? Was aber sicher ist: Wir werden auch in den nächsten Jahrzehnten Klima- und Wettersimulationen nutzen, Digital Twins in den Ingenieurwissenschaften, und Strömungssimulationen in der Luft- und Raumfahrttechnologie. Ein zukunftsfähiger Rechenzentrumsstandort in 2030 beschäftigt eine große Anzahl von Experten, die in der Lage sind Algorithmen, Software und Anwendungen zu entwickeln, die auf der dann vorherrschenden Hardware laufen. Deutschland wird aufgrund der hohen Energiekosten nicht mit anderen Nationen in reiner Rechenpower konkurrieren können. Derzeit baut Deutschland Rechenzentren mit 30 bis 40MW, und plant vielleicht 300-400 MW Systeme. In den USA wird derzeit ein Rechenzentrum mit mehr als 3GW gebaut, und es werden noch deutlich größere geplant. Und die USA sind skeptisch, dass sie mit den China konkurrieren werden können. Vor diesem Hintergrund sollte es gar nicht Deutschlands Anspruch sein in reiner Rechenpower zu konkurrieren, sondern in Expertise, Innovation und Effizienz. Deshalb ist die Investition in Kompetenzen und Software sehr viel wichtiger und erfolgsversprechender. Dennoch: wir brauchen Rechenzentren in Deutschland und wir sollten gerade in Anbetracht der Energiekosten in die Forschung und Entwicklung von energieeffizienter Rechenzentren und effizienter Software investieren, welche die verfügbare Hardware maximal ausnutzt.

Welche zentralen Herausforderungen und Chancen sehen Sie für den Rechenzentrumsstandort Deutschland in den kommenden Jahren?

Eine zentrale Herausforderung für den Rechenzentrumsstandort Deutschland sind die Energiekosten, die Inflexibilität des Stromnetzes die Schwankungen in der Energie Verfügbarkeit abzufedern und die damit verbundene Volatilität von Stromverfügbarkeit und -preisen. Zwar könnte der Strompreis durch die inzwischen massiv vorhandenen erneuerbaren Energien zu Energie-Peak Zeiten gedrückt werden, aber in Zeiten der lokalen Dunkelflaute kann die beschränkte Durchflusskapazität des Stromnetzes anderswo verfügbare Energie nicht effizient genug umverteilen. Ich sehe drei vielversprechende Strategien, diesen Nachteil gezwungenermaßen zu beheben und damit in eine Chance zu verwandeln:

1. Stärkung der Strominfrastruktur durch ein modernisiertes Stromnetz ist zwar eigentlich ein größeres Thema, aber insbesondere für den effizienten Betrieb von Rechenzentren wichtig. Zudem kann auf eine verbesserte Resilienz gegenüber Angriffen auf die Energieinfrastruktur erreicht werden.

2. Verschiebung vorn Computational Workloads. Statt die Energie an Rechenzentren zu bringen, kann man auch umgekehrt die Workloads zu den Rechenzentren verschieben an denen gerade Energie verfügbar ist. Zum Beispiel könnte tagsüber mit Solarstrom in Süddeutschland, nachts und bei Dunkelflaute mit der Energie gespeichert in Speicherseen und bei Verfügbarkeit mit dem Wind der Küste. Workloads zu verschieben erfordert allerdings eine gemeinsames Betriebsmodell der Rechenzentren in Bezug auf Hardware und Software, was eine große Herausforderung ist da derzeit alle Rechenzentren – auch politisch motiviert - ihr eigenes Betriebsmodell fahren und

die Hardwarearchitektur meist zu verschieden ist. Die Idee Workloads zu verschieben war Anfang des Jahrtausends als “Grid-Computing” gescheitert, wird aber durchaus von Tech-Giganten wie Microsoft oder Amazon gefahren die innerhalb des Unternehmens top-down Rechenzentrumsarchitektur und Betriebsmodell homogenisieren.

3. Flexibel auf lokale Energieverfügbarkeit reagieren. Derzeit werden Rechenzentren meist mit einer Energie-Flatrate geplant die eine konstante Auslastung annehmen. Gleichzeitig schwankt die Energieverfügbarkeit und der Spotpreis stark. Als Antwort darauf sollte in adaptive RZ-Infrastruktur investiert werden die Hardware schnell aktivieren aber auch deaktivieren kann um so Schwankungen und Peaks in Energieverfügbarkeit sowie Dunkelflauten auch zum Vorteil der Gesamtbevölkerung abzufedern.

4. Unbürokratische Nutzung von Abwärme der Rechenzentren. Derzeit ist dies aufgrund von Vorteilsnahme schwierig oder unmöglich, das ist aber unökologisch und findet wenig Akzeptanz in der Gesellschaft.

Eine weitere Herausforderung ist die Bürokratie von öffentlichen Bauvorhaben, Erwerb und Installation von Rechenzentrumsinfrastruktur, aber auch die bürokratischen Hürden bei der Nachnutzung von Abwärme.

Eine dritte Herausforderung besteht darin, dass die Vergütung von Mitarbeitern im öffentlichen Dienst gerade im Bereich Hochleistungsrechnen, Künstliche Intelligenz, und Rechenzentrums und Serverbetrieb nicht kompetitiv mit der freien Wirtschaft ist.

Welche Rahmenbedingungen sollten aus Ihrer Sicht wie verändert werden, um Rechenzentrumsinvestitionen zu fördern und Innovation zu ermöglichen?
Derzeit tut sich Deutschland noch schwer Software, Kompetenzen und Support als Infrastruktur zu verstehen. Das behindert Innovation und Erfolg im internationalen Wettbewerb. Rechenpower wird gebraucht, aber Rechenzentren alleine führen nicht zu Forschung und Innovation, dazu braucht es Experten und Software-Stacks die Forschung und Innovation erlauben. Langfristige Förderung von Softwareentwicklung und verstetigte Stellen und eine Karriereperspektive als Research Software Engineer oder Performance Engineer helfen diese Schwächen zu beheben und Talente langfristig zu binden. Rechenzentren sind prädestiniert als Heimat für langfristige Softwareförderung (und das sage ich, obwohl ich an keinem Rechenzentrum arbeite). Des weiteren sollten bürokratischen Hürden bei Bau von Rechenzentrums-Gebäuden und dem Aufbau und dem Betrieb von Rechenzentren abgebaut werden. Die Prozesse müssen beschleunigt werden.

Welche Rolle sollte der Staat bei der Entwicklung einer souveränen und resilienten Recheninfrastruktur einnehmen?
Aufbau und Ausbau von Stromnetz und Kommunikationsnetz und die Rahmenbedingungen für langfristige Förderung von Software und Talent.

Gibt es konkrete Maßnahmen oder Best Practices aus Ihrer Praxis/Erfahrung, die in die Strategie aufgenommen werden sollten?
Von 2017-2024 habe ich maßgeblich im US Exascale Computing Projekt (ECP) mitgewirkt, zunächst als Projektmitarbeiter, später als Principal Investigator und Leiter von Schwerpunktprogrammen. Das

US Exascale Computing Projekt war unter Präsident Barack Obama mit einem Volumen von 4Mrd US Dollar ins Leben gerufen um den Wettlauf zum Exascale zu gewinnen. Wie alle Projekte hatte auch ECP Schwächen, aber einige Stärken die das Projekt zu einem Erfolg gemacht haben:

- Das Projekt hatte von Anfang an ein klar definiertes Ziel und eine top-down Struktur in der jeder Arbeitsgruppe, jedem Betriebssystem, jedem Industriepartner, letztendlich jedem Mitarbeiter klar war wie er oder sie zum Erreichen dieses Ziels beiträgt.

- Die Schwerpunkte des Programms – Themen von nationaler Relevanz - wurden vorgegeben, so dass zwar ein Wettbewerb der Ansätze, aber nicht der Anwendungen entstand.

- Etwa die Hälfte des Budgets wurde in Softwareentwicklung und Research Software Engineering investiert und nur die andere Hälfte in die Entwicklung und Akquise von Hardware (RZ Infrastruktur). Teil der Hardwareinvestition waren Fördergelder für die Industrie um “non-recurring development” zu finanzieren, das die Hardware speziell für die Anforderungen des Projektes zuschneiden sollte. Als Hardware-Plattform wurden GPU-Supercomputer von AMD und Intel entwickelt – obwohl bereits zu Projektbeginn NVIDIA Supercomputer verbreiteter waren und mehr Software-Unterstützung hatten. Letztendlich war es die Investition des ECP Projektes in Softwaretechnologie, die das Projekt trotz fraglich optimaler Hardwareinvestition zu einem Erfolg hat werden lassen.

Ein wichtiger Aspekt zum Erfolg des ECP Projektes war, dass es als top-down Projekt implementiert wurde. Zwar unterstehen in den USA die Open Science National Labs dem Department of Energy, was die Implementation einer top-down Strategie vereinfacht. Aber auch dort gibt es Konkurrenz zwischen den National Labs, die Situation ist der deutschen Rechenzentrumslandschaft also nicht unähnlich.

Die massive Investition in Softwaretechnologie war der Schlüssel zum Erfolg. ECP hat ein Software-Ökosystem geschaffen, das wissenschaftliches Arbeiten auf verschiedenen Computerarchitekturen erlaubt, und damit auch ein Verschieben von Workloads von einem System auf ein anderes. Zudem wurde als Beiprodukt des Projektes eine große Anzahl an Softwareentwicklern ausgebildet und in der Teamarbeit geschult, dieses Personal stellt jetzt einen signifikanten Anteil der Softwareentwickler bei AMD, Google, HPE, Intel, Nvidia und Meta. Als dritte Lektion hat sich die ECP Strategie eines “Contingency Funds” bewährt. Mit diesen Mitteln wurde während des Projektes auf ein sich änderndes Lagebild reagiert. Zum Beispiel wuchs der AI Fokus über die Laufzeit des Projektes. Auf der Hardware-Seite wurden Tensor-Cores/Matrix-Cores für AI Anwendungen in die Hardware integriert, und der Contingency Fund konnte Projekte fördern, die die Nutzung dieser Spezialhardware für andere Anwendungen untersucht haben.