Bundesministerium für Bildung und Forschung

Bundesministerium für Wirtschaft und Klimaschutz

KI-Rechnerinfrastruktur in Deutschland: Kapazitäten, Bedarfe, Maßnahmen

Stand: 17.05.2024

Inhalt

1. Vorbemerkungen 2
2. Zusammenfassung der Ergebnisse 3
3. Bedarfsanalyse 5
   - Zusammenfassung 5
   - Im Einzelnen 5
   - Internationaler Vergleich 5
   - Bedarfe im Kontext der KI-Entwicklung 6
4. Mapping der in DEU vorhandenen Kapazitäten 8
   - Zusammenfassung 8
   - Im Einzelnen 8
   - Öffentlich finanzierte Kapazitäten 8
   - Wirtschaft 9
5. Fazit der Kapazitäten- und Bedarfsanalyse 11
6. Maßnahmen 12
7. Zusammenfassung 15

Impressum 16

1. Vorbemerkungen

Neben Fachkräften und Daten sind **Rechenkapazitäten** ein zentraler Treiber für die Entwicklung und Nutzung von Künstlicher Intelligenz (KI). Gerade generative KI-Systeme der neuesten Generationen wie ChatGPT sind auf große Datenmengen und eine hohe Rechenleistung für deren Verarbeitung angewiesen, da sie in erheblichem Maße auf statistischen Methoden operieren. Der Vorsprung der Industrie in den USA bei generativen KI-Modellen speist sich vor allem aus der Tatsache, dass sich derzeit sowohl die größten und wertvollsten Datenbestände als auch die Recheninfrastruktur in amerikanischem Besitz befinden. Durch Netzwerkeffekte wird die marktbeherrschende Stellung weiter gestärkt. Rechenzentren und Datenplattformen werden daher zunehmend zu sogenannten „Essential Facilities"; deren Bereitstellung zu einem Teil der gesellschaftlichen Vorsorge.

Zu beachten ist dabei, dass KI, insb. das Training großer und sehr leistungsstarker KI-Modelle, eine **spezifische Recheninfrastruktur** benötigt. Während Höchstleistungsrechner/Supercomputer in der Vergangenheit vor allem mit klassischen Prozessoren (CPUs) ausgestattet wurden, verschiebt sich dies mit der wachsenden Bedeutung von KI zunehmend in Richtung **Grafikprozessoren (GPUs)**. Diese sind für Matrixrechnungen und parallelisiertes Arbeiten optimiert und damit am besten auf die Funktionsweise aktueller KI-Systeme, insb. neuronaler Netze, abgestimmt. Häufig wird daher zur Beurteilung der Leistungsfähigkeit eines Supercomputers nicht mehr vorrangig auf die FLOPS, d.h. die pro Sekunde durchführbaren Gleitkommazahl-Operationen abgestellt, sondern auf weitere Parameter wie die Anzahl der GPUs. Zu beachten ist dabei jedoch, dass es erhebliche Unterschiede, etwa hinsichtlich der Bauart gibt. Darüber hinaus kommt es für eine effiziente und effektive Nutzung nicht nur auf die Hardware-Spezifikationen, sondern auch auf einem adäquaten Software-Stack, die notwendigen Kompetenzen sowie passende Unterstützungsangebote für den Zugang und die Nutzung an.

Als Reaktion auf die aktuellen Entwicklungen und Debatten zu KI und der dafür verfügbaren und notwendigen Recheninfrastruktur hat die Bundesregierung dieses Whitepaper erarbeitet. Es soll folgende Fragen klären:

- Welche Rechenkapazitäten sind in DEU bereits vorhanden, welche Zugangsbedingungen gibt es dafür und wie ist DEU hier im internationalen Vergleich positioniert?
- Welche Bedarfe an Rechenkapazitäten gibt es mit Blick auf die Erforschung, Entwicklung und Nutzung von KI heute und in Zukunft?
- Welche Maßnahmen hat die Bundesregierung bereits initiiert oder noch geplant, um mögliche Lücken zwischen Kapazitäten und Bedarfen zu schließen?

2. Zusammenfassung der Ergebnisse

Bei den vorhandenen KI-Rechenkapazitäten ist das Bild gemischt: Während die öffentlich geförderten Supercomputer für die Forschung und den vorwettbewerblichen Bereich bereits gut aufgestellt sind und weiter ausgebaut werden, etwa über das Gauss Centre for Supercomputing (GCS) und die EuroHPC-Initiative, gibt es für Unternehmen im wettbewerblichen Bereich derzeit in Deutschland und Europa noch zu wenig Kapazitäten für die Berechnung großer generativer KI-Modelle. Dies gilt insbesondere für den Mittelstand inklusive KMU und Startups.

Bei der Analyse des **Bedarfs** ist die zentrale Herausforderung, dass die technologischen Entwicklungen im Bereich der KI (etwa generative KI) noch am Anfang stehen und dabei so dynamisch sind, dass die künftige Bedarfsentwicklung nicht verlässlich vorhergesagt werden kann. Eine weitere Steigerung des Bedarfs an Rechenressourcen in Forschung und Wirtschaft ist jedoch technologisch wahrscheinlich, zumindest bei der Entwicklung von großen KI-Modellen. Auch deshalb kündigen einige Staaten derzeit den Ausbau ihrer Infrastruktur an. Ein Training eines mit GPT vergleichbaren KI-Modells ist für Unternehmen, insb. Start-ups, derzeit in Deutschland nur mit Nutzung von US-Hyperscalern machbar. Hyperscaler profitieren damit immer: als Anbieter von KI, aber auch als Anbieter von notwendigen Infrastrukturdienstleistungen für eventuelle Konkurrenzmodelle durch Zugriff auf das zugrundeliegende Know-how. Dies ist nicht nur aus wirtschaftlicher Sicht kritisch, da große KI-Modelle aus außereuropäischen Staaten europäische Werte und Besonderheiten (wie die Sprachvielfalt) nicht ohne weiteres berücksichtigen.

Da es einer hohen Geschwindigkeit bedarf, um den Anschluss insb. gegenüber den USA nicht zu verlieren, sind zunächst **Maßnahmen** sinnvoll, um vorhandene und im Aufbau befindliche Kapazitäten zu nutzen, auszubauen und durch Verbesserung der Zugangsbedingungen sowie erweiterte Serviceangebote verstärkt wissenschaftlichen und wirtschaftlichen Akteuren zur Verfügung zu stellen.

Im Forschungs- bzw. vorwettbewerblichen Bereich gibt es hierfür konkrete Pläne, die sich zum Teil schon in der Umsetzung befinden (insb. Initiativen von GCS und EuroHPC):

Maßnahmen[1]	**in Umsetzung**	**geplant**
Ausbau der KI-Recheninfrastruktur	• Ausbau des GCS-Standorts Jülich mit JUPITER (Inbetriebnahme 2025) • Ausbau des NHR-Standorts Dresden mit ZIH • KI Servicezentren Hardware	• Ausbau Recheninfrastruktur der GCS-Standorte München und Stuttgart • Ausbau der KI-Recheninfrastruktur der NHR-Standorte • EuroHPC-Industrierechner
Verbesserung der Zugangsbedingungen	EuroHPC Special Access	• Weitere Öffnung des Industriezugangs bei GCS
Ausbau von Service-Angeboten	• KI-Servicezentren • GCS-Anwenderunterstützung	• EuroHPC-Projekt European Support Center for AI • Kooperationsmodelle mit Forschungsdienstleistern
Weiterentwicklung des Softwarestack		• EuroHPC-Projekt zur Entwicklung von Methoden, Programmumgebungen und eines Software Stacks

[1] Eine nähere Beschreibung der Maßnahmen ist in den nachfolgenden Kapiteln enthalten.

Diese Maßnahmen sollten im Zusammenwirken dazu führen, dass die öffentlich geförderten Supercomputer für KI-Forschende aus Wissenschaft und Wirtschaft erheblich zugänglicher und die Nutzerzahlen deutlich steigen werden. Ein erheblicher Teil des vermuteten Bedarfs kann somit kurz- und mittelfristig adressiert werden. Ein Vorteil ist, dass bei derzeit offener langfristiger Bedarfslage die Wirksamkeit dieser Maßnahmen überprüft und der tatsächlich bestehende Bedarf an zusätzlichen öffentlichen Computing-Ressourcen evidenzbasiert ermittelt werden kann. Überprüft werden kann auch, ob damit KMU und Startups in der Breite auf ein niedrigschwelliges, planbares und kostengünstiges Compute-Angebot – außerhalb von Cloud-Angeboten - zugreifen können.

Eine wesentliche **Herausforderung** ist voraussichtlich, die wettbewerbsrelevanten Unternehmensdaten sicher auf eine gemeinschaftlich genutzte und ggf. öffentlich finanzierte Infrastruktur zu transferieren, um die Berechnung großer KI-Modelle zu unterstützen. Hierzu bieten sich zwei Lösungsansätze an:

- Entwicklung von Modellen, bei denen zunächst ein rechenintensives Vortraining auf der öffentlich finanzierten Infrastruktur mit einem reduzierten bzw. öffentlich zugänglichen Datensatz erfolgt und die Verfeinerung mit sensiblen Daten im Anschluss mit kleineren Rechnern durchgeführt wird.
- Beteiligung der Wirtschaft am Ausbau privater Rechenkapazitäten als Alternative zum Angebot von US Hyperscalern (z.B. europäischer Industrierechner).

3. Bedarfsanalyse

Zusammenfassung

Eine präzise Vorhersage der Bedarfe für KI Recheninfrastruktur ist aufgrund der sehr dynamischen technologischen Entwicklung und der Schwierigkeit, Bedarfe anreizkompatibel zu erheben, nicht möglich. Verfügbare Quellen legen jedoch einen aktuell bereits erheblichen und weiter steigenden Bedarf an Rechenkapazitäten in der Forschung und der kommerziellen Nutzung nahe.

Jenseits wirtschaftlicher Erwägungen gibt es gute Gründe, die Entwicklung ressourcenintensiver generativer KI-Modelle auch in Europa anzusiedeln; etwa die in US-Modellen nicht berücksichtigte europäische Sprachenvielfalt und die inhärenten Werte, die in Drittstaaten trainierten Modellen innewohnen und nicht zwingend die europäischen sind, sowie Fragen der digitalen und wirtschaftlichen Souveränität.

Andere Länder investieren erheblich in hochleistungsfähige Recheninfrastruktur.

Im Einzelnen

Eine umfassende und repräsentative Bedarfserhebung ist kurzfristig nicht in anreizkompatibler Form möglich: Es wäre kaum zu verhindern, dass beliebige Bedarfe ohne fachliche Fundierung in der Hoffnung auf staatliche Fördermittel in den Raum gestellt werden. Die Bedarfsanalyse erfolgt daher auf verschiedenen Wegen, insb. einen internationalen Vergleich, die Abschätzung von Rechenbedarfen für aktuelle Entwicklungen im Bereich KI sowie Abfragen bei einzelnen Stakeholdern.

Internationaler Vergleich

Der Stanford AI Index Report 2023 wertet u.a. die Rechnerkapazitäten für die Teilnehmenden am **internationalen KI-Wettbewerb „MLPerf“** aus. Das beste System kommt hier auf **4.216 GPUs**; im Durchschnitt werden 1.859 GPUs genutzt. Alle deutschen KI-Kompetenzzentren kommen derzeit auf ca. 3.150 GPUs (über verschiedene Standorte verteilt), das FZ Jülich verfügt über 3.744 GPUs, die im Rahmen des Exascale-Ausbaus auf ca. 24.000 GPUs anwachsen werden.

Führend im Bereich KI-Rechnerinfrastruktur sind die **USA**, wobei dort die KI-Forschung hauptsächlich von Großunternehmen ohne öffentliche Fördermittel finanziert und durchgeführt wird. Zu den leistungsstärksten Supercomputern in **Wissenschaft & Forschung** zählen u.a. „Frontier“ (Platz 1 @ TOP 500) und „Summit“ (Platz 5 @ TOP 500) am Oak Ridge National Laboratory sowie „Sierra“ (Platz 6 @ TOP 500) am Lawrence Livermore National Laboratory. Hinzu kommen Supercomputer **privater Unternehmen**. Zahlen gibt es zu Systemen von Meta, Tesla, Microsoft Azure, Google und NVIDIA, wobei die Ausbaupläne dynamisch sind.

Erhebliche KI-Rechenkapazitäten gibt es auch in Großbritannien (Cloud des Unternehmens „XTX Markets“), Italien (u.a. Supercomputer „HPC5“ des ENI-Konzerns), Japan (AI Bridging Cloud Infrastructure des NAIST) und Frankreich („Jean Zay” Supercomputer des GENCI). Über Aktivitäten in CHN sind kaum belastbare Daten bekannt, da sich CHN u.a. nicht mehr am TOP500-Ranking beteiligt.

Viele Staaten haben einen weiteren Ausbau ihrer Rechenkapazitäten angekündigt: Großbritannien plant Investitionen in Höhe von 900 Mio. £ in Rechnerkapazitäten, die zu einem großen Teil für KI-Sprachmodelle genutzt werden sollen („BritGPT“). Auch Japan und Südkorea wollen jeweils über 200 Mio. Dollar in KI-Rechnerinfrastruktur investieren. In den USA wird derzeit am Argonne National Laboratory der Exascale-Supercomputer „Aurora“ mit 63.744 GPUs aufgebaut.

Hochleistungsrechner	Standort	GPU
Aurora	Argonne, USA (im Aufbau)	63.744
Frontier	Oak Ridge, USA	37.632
Summit	Oak Ridge, USA	27.648
Sierra	Livermore, USA	17.280
AI Research Super Cluster (Meta)	Menlo Park, USA	16.000
Private Cloud (Tesla)	USA	7.360
EOS NVIDIA DGX SUPERPOD (NVIDIA)	USA	4.608
Eagle (Microsoft)	USA	14.400
A3 (Google)	USA	26.000
Private Cloud (XTX Markets)	Großbritannien	10.000
Isambard-AI	Bristol Großbritannien (geplant)	5.448
HPC5 (ENI-Konzern)	Ferrera Erbogogne, ITA	7.280
Jean Zay (GENCI)	Orsay, FRA	3.136
AI Bridging Cloud Infrastructure	Tokio, JPN	4.352

Bedarfe im Kontext der KI-Entwicklung

In den letzten Jahren sind die **Parameterzahlen bedeutender KI-Modelle stark angestiegen**. Dieser Trend legt einen weiterhin steigenden Bedarf an Rechenkapazitäten nahe. So rechnet die LEAM-Machbarkeitsstudie für einen Trainingsdurchlauf eines „Foundation Modells“ mit ca. 390.000 Knotenstunden, d.h. Stunden auf einem GPU-Knoten mit 8 GPUs. Für das Training von ChatGPT (175 Mrd. Parameter) wird von einer Rechnerinfrastruktur mit ca. 10.000 GPUs ausgegangen. Darüber hinaus werden gerade großen generativen KI-Modellen **erhebliche ökonomische Potenziale** zugeschrieben: Einer aktuellen McKinsey-Studie zufolge könnte generative KI weltweit eine zusätzliche Wertschöpfung von 2,6 bis 4,4 Bio. US-Dollar pro Jahr generieren; für Deutschland wird ein Wachstum des Marktwertes von generativer KI von aktuell 1,0 auf 8,3 Mrd. US-Dollar bis 2030 prognostiziert.

Die Wirtschaft hat das Potenzial erkannt: In der o.g. BDI-Abfrage sieht eine Mehrheit der Mitglieder ein großes bis sehr großes Potential großer Sprachmodelle. Die **LEAM-Initiative** fordert in ihrer Machbarkeitsstudie eine KI-Rechnerinfrastruktur mit **4.480 GPUs** und hält hierfür Investitionen zwischen 350 und 400 Mio. Euro für den Aufbau und Betrieb für vier Jahre erforderlich.

Grundsätzlich kann davon ausgegangen werden, dass der Entwicklungspfad basierend auf großen KI-Modellen, die vor allem statistische Verfahren nutzen und zunehmend mehr Daten und Rechenleistung erfordern, sich weiter fortsetzen wird. Der Nutzung dieser Modelle stehen jedoch die **hohen Kosten**

für Hardware, Energie und Datenakquisition gegenüber. Bei der wirtschaftlichen Nutzung werden daher Modellen mit geringerem Ressourcenbedarf bedeutende Chancen eingeräumt. Der Bedarf an Rechenressourcen wird sich dabei je nach Branche, Unternehmenstypus und -strategie unterscheiden; Start-Ups und große Unternehmen, die eigene Modelle erstellen, benötigen mehr Rechenressourcen als ggf. KMU, die kleinere Open Source-basierte Spezialmodelle nutzen und lokal betreiben möchten. Gleichzeitig wurden zuletzt auch bedeutende KI-Modelle veröffentlicht, die geringere Parameterzahlen ausweisen und dennoch leistungsfähig sind (z.B. LLaMa; Luminous). Auch die Erweiterung Statistik-basierter Modelle um gut verstandene genuine KI-Methoden bietet erhebliche neue Potentiale, die bisher nicht ausgeschöpft sind.

Die Forschung an KI benötigt große Rechenressourcen. Die nächste Generation der großen KI Modelle wird durch Forschung vorangetrieben und auch häufiges Wechseln der Trainingsparameter, wie sie in der Forschung naturgeben vorkommen, benötigen ausreichend Kapazitäten.

Des Weiteren ist es ein Bedarf der Forschung ebenso wie der Unternehmen einen unkomplizierten Zugang zu Rechenkapazität zu erhalten.

4. Mapping der in DEU vorhandenen Kapazitäten

Zusammenfassung

Die öffentlich finanzierten Rechenkapazitäten in der Forschung sind in Deutschland gut aufgestellt und werden aktuell weiter ausgebaut. Informationen über private Kapazitäten sind deutlich schwerer zu erfassen; es zeigt sich jedoch, dass die Kapazitäten, zumindest für das Training großer KI-Modelle, für Unternehmen nicht ausreichend verfügbar bzw. zugänglich sind, und dass der Bedarf steigt.

Im Einzelnen

Deutschland hat seit Jahren im Forschungsbereich eine sehr gut aufgestellte Rechnerlandschaft. In der aktuellen **TOP 500-Liste** (Stand: Juni 2023; siehe www.top500.org) liegt Deutschland bei der Anzahl der Supercomputer mit 36 auf **Rang 3**, hinter USA (150) und CHN (134). Bei der aufsummierten Leistung erreichen zusätzlich JPN, ITA und FIN höhere Werte, wobei dies bei ITA und FIN an den bereits installierten Systemen aus der gemeinsamen europäischen Initiative, EuroHPC (siehe unten) liegt, an der Deutschland beteiligt ist.

Öffentlich finanzierte Kapazitäten

Das deutsche, auch an die internationale Wissenschaft gerichtete HPC-Angebot ist in drei Ebenen strukturiert:

- Die erste Ebene bilden die drei leistungsfähigsten Supercomputer Deutschlands unter dem Dach des **Gauss Centre for Supercomputing (GCS)**. Dieses wird vom BMBF und den Ländern Nordrhein-Westfalen, Baden-Württemberg und Bayern gemeinsam finanziert. Das GCS umfasst das Jülich Supercomputing Center, das Leibniz-Rechenzentrum München und das Höchstleistungs-rechenzentrum Stuttgart. Insgesamt umfasst das GCS sechs Supercomputer in den TOP 500. Hervorzuheben ist das **JUWELS-Booster-Modul in Jülich**, das **mit 3.744 GPUs** auch für KI-Anwendungen international wettbewerbsfähig ist und u.a. vom Projekt OpenGPT-X und von der LAION-Initiative für große Open-Source-KI-Modelle genutzt wird.

  Rechenzeit am GCS wird derzeit **vorrangig für Forschungsvorhaben** in einem Peer-Review-Verfahren nach wissenschaftlichen Kriterien vergeben. Bewerbungen sind in der Regel jederzeit möglich, Rechenzeit wird für Großprojekte und für die Nutzung des JUWELS-Booster-Moduls jedoch nur zweimal pro Jahr vergeben. Die von Universitäten oder Forschungseinrichtungen beantragten Forschungsprojekte werden häufig von Konsortien durchgeführt, auch unter Beteiligung der Industrie. Aufgrund des hohen Bedarfs an Rechenzeit sind die Rechner am GCS aktuell **vollständig ausgelastet**.
- Die zweite Ebene umfasst **überregionale HPC-Zentren mit Hochleistungsrechnern** an Forschungseinrichtungen und Hochschulen, die das BMBF gemeinsam mit den Ländern im **Verbund Nationales Hochleistungsrechnen (NHR)** fördert oder im Rahmen der Förderung der **außeruniversitären Forschungseinrichtungen**, der **KI-Kompetenzzentren** sowie des Deutschen Netzwerks für Bioinformatik-Infrastruktur (**de.NBI**) bereitstellt. Auch hier gibt es HPC-Systeme mit ca. 1.000 GPUs, etwa an der Universität Tübingen (1.013), am Max-Planck-Institut für Intelligente Systeme (960) und an der Universität Frankfurt (864).
- Die dritte Ebene bilden vor allem regionale Rechenzentren, die eine Vielzahl von Anwendungen mit geringeren Leistungsanforderungen für den Bedarf vor Ort bedienen.

Das Ziel ist, jederzeit ein bestmögliches System zur Verfügung zu stellen, das in vielfältigen Anwendungsfeldern flexibel nach den Bedarfen der Wissenschaft eingesetzt werden kann.[2]

Das BMBF beteiligt sich darüber hinaus am **Gemeinsamen Unternehmen EuroHPC**. Mit dieser europäischen Partnerschaft fördern die Europäische Union und die Mitgliedstaaten den Aufbau europäischer Rechenkapazitäten und ihre Vernetzung sowie die Entwicklung neuer Technologien und Anwendungen für die europäische Wissenschaft und Wirtschaft. Über EuroHPC hat die deutsche Wissenschaft & Forschung Zugriff auf sehr leistungsstarke Rechner (vgl. Tabelle). Jeweils die Hälfte der Rechenzeit der genannten EuroHPC ko-finanzierten Systeme wird europaweit ausgeschrieben[3]. Neben der Recheninfrastruktur erfolgt in EuroHPC voraussichtlich in Q2/2024 der Aufbau eines Europäischen KI-Supportzentrums, das das Angebot der vorgenannten nationalen KI-Servicezentren ergänzen und die Anwenderunterstützung weiter stärken wird.

Hochleistungsrechner	**Standort**	**GPU**
JUPITER	Jülich, Deutschland (ab 2025)	Rd. 24.000
LUMI	Kajaani, Finnland	11.912
LEONARDO	Bologna, Italien	13.824
MARENOSTRUM 5	Barcelona, Spanien	4.480
VEGA	Maribor, Slowenien	240
MELUXINA	Bissen, Luxemburg	800
KAROLINA	Ostrava, Tschechische Republik	576
DISCOVERER	Sofia, Bulgarien	--
DEUCALION	Guimarães, Portugal	132
JULE VERNES	FRA, in Planung	Noch unklar

Wirtschaft

Ein umfassendes Bild über vorhandene Rechenkapazitäten der Wirtschaft zu erhalten, ist schwer möglich. Es gibt nur wenige öffentliche Informationen; viele Unternehmen halten sich auch auf Nachfrage sehr bedeckt. Das BMWK plant deshalb die Durchführung einer Studie zu privatwirtschaftlichen Rechenzentren, die ein klareres Bild über die in Deutschland vorhandene und geplante Rechenkapazität, etwaige Engpässe für Unternehmen und die Eigentümerstruktur geben soll.

Den Zugriffsmöglichkeiten durch die Wirtschaft / private Akteure auf die o.g. Rechnerkapazitäten für Wissenschaft & Forschung sind ordnungspolitisch und **durch das Beihilferecht Grenzen gesetzt**.

[2] Bspw. wurden ab Mitte März 2020 Rechnerkapazitäten prioritär für die Erforschung des Virus SARS-CoV-2 und der Pandemieausbreitung zur Verfügung gestellt

[3] Der 50:50 Schlüssel gilt bei den großen EuroHPC-Systemen („High-End“), zu denen das JUPITER-System zählen wird. Bei den kleineren Systemen („Mid-Range“) beträgt das europäisch vergebene Rechenzeitkontingent 35 %.

Insgesamt ist v.a. eine Bereitstellung im vorwettbewerblichen Bereich möglich, etwa für Forschungsprojekte an denen Unternehmen im Verbund mit der Wissenschaft beteiligt sind. Zudem stellt das Training eines großen KI-Sprachmodells besondere Anforderungen an Rechenkapazitäten.

Aus öffentlich verfügbaren Quellen und nicht repräsentativen Befragungen der Wirtschaft liegen zu **Rechenkapazitäten von privaten Akteuren/Unternehmen** folgende Informationen vor:

- Der BDI hat im Rahmen einer anonymen Umfrage unter seinen Mitgliedern die Kapazitäten und den Bedarf an Rechenkapazitäten erhoben. Während die große Mehrheit der Unternehmen den potentiellen Wert von großen KI-Modellen erkennt, sind das notwendige Know-How und die Rechenkapazitäten nicht vorhanden. Große Unternehmen bemühen sich hier, eigenständige Infrastrukturen aufzubauen.
- Einer aktuellen **Bitkom-Studie** zufolge wuchsen die Kapazitäten der Rechenzentren in DEU (gemessen an der IT-Anschlussleistung) von 2010 bis 2022 um über 90%. Der Trend geht dabei zu größeren Zentren sowie Cloud- und Colocation-Modellen. Gesonderte Aussagen zu KI oder Forschung enthält die Studie nicht.
- **Aleph Alpha** hat im Sommer 2022 ein Rechenzentrum mit mindestens 512 GPUs eröffnet (stateof.ai geht von 1.044 GPUs aus). Daneben sind nur noch zwei weitere Rechenzentren aus der Industrie in den Top 500 gelistet.
- Auch SAP, die Robert Bosch GmbH, Siemens, die Deutsche Telekom, BASF und Continental verfügen (nach eigenen Angaben) über leistungsstarke Rechenzentren u.a. auch für KI-Anwendungen. Konkrete Daten zur Hardware sind (auch aus Geschäftsgeheimnisgründen) nicht öffentlich bekannt, die Unternehmen werden aber von KI-Hardwareanbietern als Referenzkunden angegeben.
- Eine aktuelle Umfrage legt nahe, dass **74% der deutschen Unternehmen auf externe Rechnerkapazitäten zurückgreifen**, also zumeist Ressourcen großer, auch/vor allem internationaler Anbieter nutzen. Konkrete Kooperationen sind bspw. von Carl Zeiss (mit Microsoft), VW (mit IBM), Daimler (mit Infosys, NOR), Bayer (mit Google) und der Deutschen Bank (mit NVIDIA) bekannt.[4]

[4] https://www.idc.com/getdoc.jsp?containerId=prEUR149762022

5. Fazit der Kapazitäten- und Bedarfsanalyse

Allgemein lässt sich festhalten, dass sich die Anforderungen an die Hardware durch große KI-Modelle qualitativ und quantitativ stark verändert haben. Diese benötigen eine im Vergleich zu „herkömmlicher KI“ riesige Menge an KI-spezialisierter Rechenleistung. Diese ist *für kommerzielle Zwecke* momentan in DEU und EU so nicht gegeben. Zudem wird der Bedarf aller Voraussicht nach weiter anwachsen.

Für **die Wissenschaft bzw. den vorwettbewerblichen Bereich** verfügt Deutschland mit dem JUWELS-Rechner über eine international wettbewerbsfähige KI-Recheninfrastruktur für Wissenschaft & Forschung. Daneben gibt es für Wissenschaft & Forschung zahlreiche weitere nationale und europäische Recheninfrastrukturen. Diese Kapazitäten sind aktuell ausgelastet und es wird von einem steigenden Bedarf ausgegangen, weshalb der Ausbau der KI-Rechenkapazitäten für die Forschung bereits vorangetrieben wird (siehe „6. Maßnahmen“).

Über **private Rechnerkapazitäten für die Wirtschaft** in Deutschland ist deutlich weniger bekannt. Hier besteht nach aktuellen Erkenntnissen, zumindest in Hinblick auf das Training großer KI-Modelle, jedoch der größte Nachholbedarf. Der Bedarf an Rechenkapazitäten für „herkömmliche“ KI steigt zwar gleichermaßen und bedarf ebenfalls des Weiteren Ausbaus. Allerdings besteht hier unserer Kenntnis nach kein grundsätzliches Kapazitätsproblem.

Andere Staaten investieren aktuell massiv in Rechnerinfrastruktur. Die USA hat sich hier bereits einen erheblichen Vorsprung erarbeitet. DEU darf in diesem Wettbewerb nicht zurückfallen.

6. Maßnahmen

Ziel der Bundesregierung ist es, die HPC-Kapazitäten in Forschung und Wirtschaft so auszubauen und zugänglich zu machen, dass Deutschland im KI-Bereich international wettbewerbsfähig bleibt. Dabei geht es nicht allein um die Nutzung von KI-Technologien, sondern um die Fähigkeit, selbst KI-Lösungen zu erforschen, weiterzuentwickeln und anzubieten. Unternehmen wie Start-ups brauchen dafür flexiblen Zugriff auf vorhandene HPC-Ressourcen. Zudem geht es angesichts der dynamischen Entwicklung vor allem auch um Geschwindigkeit.

Vor diesem Hintergrund bauen alle Maßnahmen **auf vorhandenen und sich bereits im Aufbau befindlichen Kapazitäten in Forschung und Wirtschaft auf**. Ein gänzlich neues KI-Rechenzentrum stünde vor erheblichen Herausforderungen hinsichtlich Personalgewinnung, Standort, Zugangsmanagement und Auslastung. Der Zeitaufwand für den Aufbau eines neuen Rechenzentrums wäre zudem erheblich höher, bspw. auch aufgrund der notwendigen Zertifizierungen.

Folgende Maßnahmen befinden sich bereits **in der Umsetzung**:

- Der **Ausbau der Rechenkapazitäten am Gauss Center for Supercomputing** (GCS) ist bereits angelegt und wird wie geplant umgesetzt. Das gilt sowohl für die durch Bund und Länder finanzierten GCS-Kapazitäten als auch für die in EuroHPC europäisch-national kofinanzierten Rechnereinheiten. Bis 2025 wird mit Mitteln von EuroHPC das **erste europäische Exascale System (JUPITER) in Jülich** aufgebaut. Nach aktuellen Angaben der Firma NVIDIA, die die GPUs für das System liefern wird, wird JUPITER über rd. 24.000 GPUs[5] verfügen und damit zu den leistungsstärksten Supercomputern für KI-Berechnungen der Welt gehören.[6] Auch ein sukzessiver Ausbau der Rechenkapazitäten in München und Stuttgart ist bereits geplant und wird zu einer deutlichen Erhöhung der KI-Kapazitäten an diesen Standorten führen.
- Parallel zu den GCS-Zentren wird auch die Recheninfrastruktur an den neun Standorten der universitären Rechenzentren Deutschlands, die sich im **Verbund Nationales Hochleistungsrechnen (NHR)** zusammengeschlossen haben, im Rahmen der zwischen Bund und Ländern vereinbarten, gemeinsamen Förderung des nationalen Hochleistungsrechnens ausgebaut.[7]
- Seit November 2022 fördert das BMBF den Aufbau von **vier KI-Servicezentren** mit dem Ziel, den Zugang zu KI-Recheninfrastruktur für Wissenschaft und Wirtschaft, insb. kleine und mittlere Unternehmen zu erleichtern, passende Serviceangebote zu entwickeln und bereitzustellen und damit den Transfer von KI in die Praxis voranzutreiben. Die KI-Servicezentren greifen dabei einerseits auf vorhandene Hardware der Partnereinrichtungen zurück, etwa das Jülich Supercomputing Centre und haben andererseits zusätzliche Rechencluster beschafft. Die Hardware-Neubeschaffung an den KI-Servicezentren liegt bei 700 GPU. Durch passende Rahmenverträge konnten die Beschaffungen größtenteils in den Jahren 2022 und 2023 abgeschlossen werden. Ende des Jahres 2023 waren bereits 468 GPUs in Betrieb genommen. Die KI-Servicezentren gehen jetzt sukzessive in den aktiven Betrieb über und bieten bereits erste Services an.

[5] Es handelt sich dabei um NVIDIA GH200 Superchips der neuesten Generation, die gegenüber den aktuellen H100 Chips deutliche Performancesteigerungen aufweisen.

[6] Nach Aussage der Firma NVIDIA werden in dem JUPITER-System (Inbetriebnahme Ende 2024 geplant) rd. 24,000 NVIDIA GH200 Superchips eingebaut (Zum Vergleich: Der neue KI-Supercomputer von Großbritannien (Isambard AI in Bristol) wird über 5.000 dieser Chips verfügen). Dies führt zu folgender Aussage: „This will make it the most powerful AI supercomputer in the world with over 90 exaFLOPS of performance says Nvidia."

[7] Die Adressierung der steigenden Bedarfe im KI-Bereich an Hochleistungsrechenressourcen erfolgt durch einen kontinuierlichen Ausbau und eine stetige Weiterentwicklung der Rechnerinfrastruktur der NHR-Zentren im Rahmen der zwischen Bund und Ländern vereinbarten, gemeinsamen Förderung und im Zuge der regulären Reinvestitionszyklen.

EuroHPC hat zu Beginn des Jahres 2024 einen Aufruf für den **Aufbau eines ersten von zwei geplanten Europäischen Industrierechnern** gestartet. Diese Rechner können für Forschung ebenso wie für die wettbewerbliche Produktentwicklung eingesetzt werden. Die Europäische Kommission übernimmt hierbei 35 % der Investitionskosten bis zu einem Gesamtbudget in Höhe von 35 Mio. €. Für den voraussichtlich Ende 2024 oder Anfang 2025 erfolgenden Aufruf für den zweiten Industrierechner wird dieses Budget erheblich erhöht auf ein Gesamtbudget bis zu 130 Mio. €. Die weiteren Kosten für die Beschaffung der Systeme sowie die vollständigen Betriebskosten muss ein Europäisches Industriekonsortium aufbringen. Der Vorteil der Industrierechner gegenüber bestehenden EuroHPC-Systemen besteht darin, dass das Konsortium die Rechnerarchitektur und die Zugangsregelungen für ihr Kontingent entsprechend seinem Bedarf auswählen kann.

Die Einreichungsfrist für Interessenbekundungen für den ersten Industrierechner ist abgelaufen. Derzeit erfolgt bei EuroHPC die Auswertung. Inwiefern sich dieses erweitere Angebot und die tatsächliche Nachfrage insbesondere von KMU und Startups treffen, wird ausgewertet, um zu beurteilen, ob die Maßnahmen ausreichend und sachgerecht sind. Dann kann auch entschieden werden, ob eine Bewerbung aus Deutschland für den zweiten Europäischen Industrierechner forciert werden sollte.

- Das Beihilferecht ermöglicht die wettbewerbliche Nutzung der öffentlich finanzierten Ressourcen in geringem Umfang (max. 20% der Gesamtkapazität) zu marktüblichen Preisen. Bislang ist am GCS-Standort in Stuttgart schon eine kommerzielle Nutzung möglich. Eine Verbreiterung des Industriezugangs an den GCS Standorten für die kommerzielle Nutzung wird angestrebt.
- Daneben sollen die KI-Forschungs- & Wissenschaftscommunity besser mit den bestehenden Rechnerinfrastrukturen vernetzt und die **Zugangsbedingungen für KI-Anwender deutlich verbessert werden**. Mehrere Initiativen von EuroHPC zielen darauf ab:
    - Die **Rechenzeitvergaberegeln** (Access Policy) von EuroHPC wurden im Dez. 2023 modifiziert. Es wurde ein **privilegierter Zugang** (Special Access) im Umfang von bis zu 20 % der Rechenzeit der EuroHPC-Systeme für KI-Simulationen durch Wissenschaft und Wirtschaft, insb. Start-ups und KMU, reserviert. Anträge können **alle zwei Monate** gestellt werden (für reguläre Projekte alle sechs Monate) mit einem **vereinfachten Auswahlverfahren** (Bewertung der Relevanz, technische und administrative Checks; kein vergleichendes Peer-Review-Verfahren). Die neuen Regelungen wurden zum Jahresbeginn 2024 umgesetzt[8].
    - Mit dem Aufbau eines **European Support Centre for AI** soll eine zentrale EU-weite Anlaufstelle geschaffen werden, die die nationalen KI-Servicezentren ergänzt. Mit einem Projektstart ist in Q2/2024 zu rechnen. Das Support Centre bietet fachliche Beratung für KI-Anwender aus Wissenschaft und Wirtschaft, insbesondere auch für KMU und Start-ups.
    - Um die Bereitstellung von Methoden, Programmumgebungen und Software Stacks zu beschleunigen und um den Zugang von KI-Nutzern auf HPC-Systeme deutlich zu vereinfachen wird ein europaweites Projekt von EuroHPC gefördert. Der Projektaufruf erfolgt im Verlauf von 2024.
- Die vorgenannten Schritte zielen insbesondere auf Nachfrager mit dezidiertem KI-Rechenbedarf auf HPC-Systemen ab. Um den Bedürfnissen von Nutzern aus Wissenschaft und

[8] Die Rechenzeitvergaberegelungen von EuroHPC sowie Informationen über die Bewerbungsmöglichkeiten sind abrufbar unter: https://eurohpc-ju.europa.eu/access-our-supercomputers/access-policy-and-faq_en.

der Wirtschaft, insbesondere dem Mittelstand, zu entsprechen, die Innovationsvorhaben vorantreiben und nur daraus abgeleiteten Bedarf an KI-Lösungen und entsprechender HPC-Rechenzeit haben wird eruiert, ob Kooperationsmodelle mit Serviceanbietern (z.B. Forschungseinrichtungen) realisiert werden können.

Die oben beschriebenen Maßnahmen werden die Angebotslage in Deutschland verbessern. Darüber hinaus sind weitere Maßnahmen denkbar:

- **Verbesserung der Informationslage** zu vorhandenen KI-Rechenkapazitäten und deren Nutzungsbedingungen, insb. durch bessere Vernetzung bestehender Angebote und Einrichtung einer Plattform als „**One-Stop-Shop**“
- **Aufbau von Rechenkapazitäten** (teil-) finanziert durch die öffentliche Hand, wie durch die LEAM-Machbarkeitsstudie vorgeschlagen. Der **Aufbau eines Public-Private-Partnership** zu KI-Infrastruktur wäre prinzipiell denkbar, aber unter anderem aufgrund rechtlicher Rahmenbedingungen wohl komplex und **kaum kurzfristig umsetzbar**. Dem Bund sind hier durch das Beihilferecht erhebliche Grenzen gesetzt. Die Herausforderung bestünde zudem darin, die notwendige finanzielle Beteiligung der Wirtschaft herbeizuführen, wenn zugleich staatliche Investitionen avisiert sind. Vor dem Hintergrund knapper Haushaltsmittel sowie teuren Marktpreisen und kurzfristig schwer zu erreichender Verfügbarkeit von Hardware für KI-Anwendungen sind Verbesserungen hierdurch kaum kurz- bis mittelfristig zu erreichen.
- Initiativen auf Länderebene (insb. Baden-Württemberg und Nordrhein-Westfalen), die den Aufbau eines KI-Ökosystems inkl. Rechenzentrum für kommerzielle Zwecke zum Ziel haben: Während Baden-Württemberg mit dem Innovationspark KI („ipai“) auf private Mittel, insb. der Dieter-Schwarz-Stiftung, zurückgreifen kann, plant Nordrhein-Westfalen die Nutzung von Mitteln aus dem Strukturwandelfonds für die Braunkohle-Regionen. Beide Vorhaben sind allerdings noch in einem frühen Planungsstadium und mit Unsicherheiten behaftet.

7. Zusammenfassung

Trotz der, vor allem im Forschungsbereich, bereits gut ausgebauten Recheninfrastruktur in Deutschland, indiziert die Analyse einen steigenden Bedarf an KI Rechenkapazitäten in Deutschland sowohl für die Forschung als auch für Unternehmen (vor allem KMUs). Da das Bild bei der wettbewerblichen Nutzung schwerer zu erheben ist, soll hier eine ausführlichere Analyse erstellt werden.

Gerade im vorwettbewerblichen Bereich hat die Bundesregierung bereits etliche Maßnahmen gestartet, um dem steigenden Bedarf entgegen zu wirken. Dazu zählt der Aufbau von GPU-Hardware, u.a. durch die EuroHPC Supercomputer, aber auch an nationalen KI-Servicezentren. Der aktuell im Bau befindliche neue Supercomputer Jupiter wird rund 24.000 GPUs bereitstellen und damit zu den größten KI-Rechnern der Welt gehören. Im Zuge dieser Maßnahmen wird auch die wettbewerbliche Nutzung im Rahmen des Beihilferechts geprüft.

Im Bereich der Serviceunterstützung hat das BMBF mit den KI Servicezentren eine Anlaufstelle zur Unterstützung für die KI-Anwender aus Forschung und Unternehmen geschaffen.

Im Bereich des einfacheren Zugangs wurden erste Maßnahmen ergriffen, wie bspw. der Special Access auf EuroHPC-Systemen und dem European Support Centre for AI.

Die Maßnahmen werden überwacht und der Bedarf analysiert. Weitere Aufstockungen, wenn notwendig, sind bei vielen der Maßnahmen möglich.

Impressum

Herausgeber

Bundesministerium für Bildung und Forschung (BMBF)
Referat Künstliche Intelligenz
Referat Elektronik und autonomes Fahren; Supercomputing
53170 Bonn

Bundesministerium für Bundesministerium für Wirtschaft und Klimaschutz (BMWK)
Referat Grundsätze der nationalen und internationalen Digitalpolitik, Digitalisierung und Nachhaltigkeit
Referat Start-ups, Digitale Vernetzung, Digital Hub Initiative
11019 Berlin

Stand Mai 2024

Text: BMBF und BMWK