DOI 10.5162/iCCC2026/P23

Auswirkungen von KI Training und Inferenz auf die Betriebskosten und das Abwärme Potential

Germain-Ray Schmidt, TU Berlin, FG Technologie und Management für Integrierte Energieinfrastrukturen und Geothermie, Straße des 17. Juni 135, 10623 Berlin, Deutschland, germain-ray.schmidt@campus.tu-berlin.de
Prof. Dr. Dr. Tanja Kneiske, TU Berlin, FG Technologie und Management für Integrierte Energieinfrastrukturen und Geothermie, Straße des 17. Juni 135, 10623 Berlin, Deutschland
Fraunhofer IEG, Gulbener Str. 23, 03046 Cottbus

1 Einleitung

Rechenzentren entwickeln sich zu einem der zentralen Treiber des weltweiten Stromverbrauchs. Der jüngste Sonderbericht der Internationalen Energieagentur (IEA) zu „Energy and AI“ schätzt den Strombedarf von Rechenzentren im Jahr 2024 auf rund 415 TWh und erwartet bis 2030 einen Anstieg auf etwa 945 TWh, wobei der Großteil dieses zusätzlichen Verbrauchs auf KI-Workloads und beschleunigte Server (GPUs, TPUs) zurückgeführt wird [1]. Für Deutschland zeigt die Studie „Rechenzentren 2022“, dass der Stromverbrauch deutscher Rechenzentren zwischen 2010 und 2022 von etwa 10 TWh auf 17,9 TWh gestiegen ist und dass gleichzeitig ein technisch nutzbares Abwärme potenzial von mehreren TWh pro Jahr besteht, das bislang nur punktuell genutzt wird [2]. Mit dem Energieeffizienzgesetz (EnEfG) reagiert der Gesetzgeber auf diese Entwicklung und verpflichtet Rechenzentren ab einer IT-Anschlussleistung von 300 kW zu höheren Effizienzstandards sowie zur Prüfung und schrittweisen Nutzung unvermeidbarer Abwärme. Der Bitkom-Leitfaden zum EnEfG fasst die Vorgaben zusammen und macht deutlich, dass neue Rechenzentren künftig Mindestanteile wiederverwendeter Energie nachweisen müssen und Abwärmenutzung damit zu einem verbindlichen Planungsparameter wird [3]. Parallel dazu verändert sich die technische Struktur der Rechenzentren grundlegend: Klassische, CPU-dominierte Infrastrukturen mit typischen Rack-Leistungsdichten von etwa 7–10 kW werden zunehmend durch KI-Cluster mit hohen Anteilen GPU-basierter Server ergänzt, die Leistungsdichten von 30 bis über 100 kW pro Rack erreichen und über lange Zeiträume hoch ausgelastet sind [4]. Experimentelle Messungen von Deep-Learning-Workloads zeigen, dass dabei der überwiegende Teil der IT-Energie in den GPUs umgesetzt wird und nahezu vollständig als Wärme im Rechenzentrum anfällt [5]. In Kombination mit flüssigkeitsbasierten Kühlsystemen eröffnen GPU-dichte KI-Rechenzentren damit neue Optionen für höherwertige Abwärmenutzung, stellen aber gleichzeitig neue Anforderungen an Betrieb, Wirtschaftlichkeit und Einbindung in Energie- und Wärmenetze [4] [6]. Vor diesem Hintergrund untersucht die vorliegende Arbeit, wie sich der Umstieg von CPU-dominierten auf GPU-basierte KI-Hardware auf die Betriebskostenstrukturen sowie das nutzbare Abwärmepotenzial von Rechenzentren unterschiedlicher Größenklassen (klein, mittel, groß, Hyperscale) auswirkt**.** Dazu werden ein CPU-Basisszenario mit überwiegend konventionellen IT-Servern und mehrere GPU-Szenarien mit unterschiedlichen Anteilen KI-beschleunigter Racks definiert. Auf Basis gemessener Leistungsdaten für Training und Inferenz sowie vereinfachter Annahmen zu Kühlung, Abwärmerückgewinnung und Energiepreisen werden die resultierenden Energieflüsse, Abwärmepotenziale und jährlichen Betriebskosten, sowohl mit als auch ohne Einspeisung der Rechenzentrumsabwärme vergleichend analysiert.

2 Hintergrund und Stand der Forschung

2.1 Entwicklung des Energiebedarfs und Rolle von KI

Der weltweite Energiebedarf von Rechenzentren wird in der aktuellen Literatur auf rund 300-380 TWh im Jahr 2023 geschätzt [6]. Parallel dazu zeigen Analysen zu KI Rechenzentren, dass der zusätzliche Strombedarf durch AI-Workloads bis 2030 auf 200-400 TWh anwachsen könnte, was dann 35-50 % des gesamten Rechenzentrumsstroms entspräche [6]. Damit wird deutlich, dass KI und damit GPU-beschleunigte Systeme künftig einen wesentlichen Anteil am Gesamtenergiebedarf der Rechenzentren ausmachen werden. Für Deutschland zeigt die Borderstep Studie „Rechenzentren 2022“, dass der Stromverbrauch der Rechenzentren zwischen 2010 und 2022 von etwa 10 TWh auf 17,9 TWh gestiegen ist. Davon entfielen 2022 rund 12 TWh auf IT- Hardware und knapp 6 TWh auf Infrastruktur wie Kühlung, USV und Stromverteilung [3]. Zugleich werden in dieser und nachfolgenden Analysen Abwärmepotenziale von mehreren TWh pro Jahr für Deutschland identifiziert, die bei geeigneter technischer und wirtschaftlicher Ausgestaltung in Wärmenetze eingespeist werden könnten [3]. Vor diesem Hintergrund und den gesetzlichen Anforderungen an Effizienz und Abwärmenutzung ergibt sich die Notwendigkeit, die energetischen Auswirkungen neuer KI- /GPU-Rechenzentren systematisch zu bewerten. 2.2 GPU-/KI-Rechenzentren und Leistungsprofile Der Übergang von klassischen CPU-Rechenzentren zu GPU-basierten KI-Rechenzentren verändert sowohl die Leistungsdichten als auch die Lastprofile. In konventionellen Rechenzentren werden Standardracks typischerweise mit 7-10 kW betrieben, während KI-Racks in spezialisierten KI-Rechenzentren 30-100+ kW pro Rack erreichen; in dedizierten KI-Facilities liegt der Durchschnitt über 60 kW pro Rack [1]. Moderne Hyperscale-KI-Rechenzentren wiesen Gesamtleistungen von >100 MW auf und werden teilweise in Richtung Gigawatt-Campi geplant [1]. Diese hohen Leistungsdichten stellen besondere Anforderungen

an Stromversorgung, Kühlung und Netzintegration. Auf Workload-Ebene unterscheiden sich KI-Lasten deutlich von klassischen IT-Workloads. Chen et al. beschreiben KI-Rechenzentren als durch hoch variable und „bursty“ Lastprofile geprägt, insbesondere in der Inferenzphase, während das Training großer Modelle über lange Zeit räume mit sehr hohen und relativ stabilen Leistungsaufnahmen läuft [1]. In einer Synthese aktueller Studien wird der Anteil der KI-Energie grob abgeschätzt auf etwa 60 % Inferenz, 30 % Training und 10 % Datenvorbereitung/Fine-Tuning [6]. Auf der Ebene einzelner KI-Server zeigen experimentelle Messungen von Aquino-Brítez et al., dass bei typischen Deep-Learning-Modellen (CNNs) der Großteil der IT- Energie von der GPU getragen wird: Während des Trainings entfallen rund 61-73 % der Energie auf die GPU, 14– 22 % auf die CPU und 13-17 % auf den Arbeitsspeicher. In der Inferenzphase liegt der GPU-Anteil immer noch bei 49-59%, CPU und RAM teilen sich den Rest [5]. Diese Ergebnisse liefern eine belastbare Grundlage, um leistungsbasierte Profile für GPU-Server zu definieren und für verschiedene Rechenzentrumsgrößen zu skalieren.

2.2 Abwärmepotenziale und Technologien zur Wärmerückgewinnung

Da nahezu die gesamte in Rechenzentren aufgenommene elektrische Energie letztlich in Wärme umgewandelt wird, stellen Rechenzentren konzentrierte Abwärmequellen dar. Hao et al. schätzen, dass typischerweise etwa 40 % des Stroms direkt in der IT-Hardware, weitere 40 % in den Kühlsystemen und die verbleibenden 20 % in sonstiger Infrastruktur umgesetzt werden und dass diese Energie nahezu vollständig als Niedertemperatur-Abwärme anfällt [2]. Die Nutzbarkeit dieser Abwärme hängt stark von der Kühltechnologie und dem resultierenden Temperaturniveau ab. Luftgekühlte Rechenzentren liefern üblicherweise Ablufttemperaturen im Bereich von etwa 25-40 °C, während flüssiggekühlte Systeme 40-80 °C erreichen können [2]. In einer umfassenden Bewertung von Abwärmerückgewinnungstechnologien schlagen Hao et al. einen exergiebasierten Effizienzindikator vor, der unterschiedliche Nutzungsoptionen) auf eine vergleichbare Basis „äquivalenter elektrischer Energie“ abbildet [2]. Die relativen Exergieeffizienzen liegen typischerweise im Bereich von etwa 7,5–11,5 % für direkte Heizungsanwendungen, 4,2-12,8 % für Kältebereitstellung und 3,0-13,2 % für Stromerzeugung. Diese Ergebnisse verdeutlichen, dass GPU-dichte KI Rechenzentren mit Flüssigkühlung nicht nur höhere spezifische Leistungen aufweisen, sondern aufgrund höherer Rücklauftemperaturen auch ein größeres nutzbares Abwärmepotenzial bieten können als klassische, luftgekühlte CPU-Rechenzentren. Genau an dieser Schnittstelle setzt die in dieser Arbeit entwickelte Szenario- und Methodikbetrachtung an.

3 Methodik

Die Arbeit verfolgt einen szenariobasierten**,** bottom-up Ansatz: Ausgehend von typisierten Rechenzentrumsgrößen werden CPU- und GPU-Lasten, Abwärmepotenziale und Betriebskosten für verschiedene Hardware-Mixe modelliert und miteinander verglichen. Die Methodik gliedert sich in fünf Schritte: Definition der Systemgrenzen und Rechenzentrumsgrößen, Szenariodefinition, Lastmodellierung von CPU- und GPU-Servern, Modellierung von PUE und Abwärmenutzung sowie Betriebskostenberechnung inkl. Sensitivitätsanalyse.

3.1 Systemgrenzen und Rechenzentrumsgrößen

Die Systemgrenze umfasst jeweils ein Rechenzentrum inklusive IT-Hardware (Server, Storage, Netzwerk) und technischer Infrastruktur (Stromversorgung, Kühlung, Nebenverbraucher). Vorketten (z. B. Herstellung der Hard ware) sowie nachgelagerte Netze (Fernwärmeverteilung) werden nicht bilanziert. Es werden vier idealtypische Rechenzentrumsgrößen betrachtet, die sich an der in Bitkom beschriebenen Strukturmodellen und Leistungsbereichen orientieren, aber um noch eine Hyperscale-Größe ergänzt wurde.[3]: kleines RZ mittleres RZ, großes RZ, Hyperscale-RZ.

Für jede Größenklasse werden eine typische Rackanzahl und eine mittlere IT-Leistung pro Rack festgelegt (kW/Rack). Diese Werte werden aus Bandbreiten für klassische und KI-Racks abgeleitet und so gewählt, dass sie im Rahmen der in [1], [3], [6] beschriebenen Größenordnungen liegen.

3.2 Szenariodefinition: CPU-Basisszenario und GPU-Szenarien

Für jede Rechenzentrumsgröße wird zunächst ein **CPU**-Basisszenario definiert: ausschließlicher Einsatz konventioneller IT-Server (CPU-dominierte Racks),ntypische Rack-Leistungsdichten klassischer Rechenzentren (8 kw/ Rack) überwiegend luftbasierte Kühlung, keine Abwärmenutzung

Darauf aufbauend werden mehrere GPU-Szenarien entwickelt, in denen der Anteil der GPU-basierten KI-Hardware an der gesamten IT-Leistung stufenweise steigt Die Szenarien sind so ausgelegt, dass: ein definierter Anteil klassischer IT-Server in allen Szenarien erhalten bleibt (nicht-KI-Workloads), die Gesamt-IT-Leistung der jeweiligen Größenklasse vergleichbar bleibt,

3.3 Lastmodellierung von CPU- und GPU-Servern

Die elektrische Leistungsaufnahme der IT-Komponenten wird getrennt für CPU- und GPU-Server modelliert.

CPU-Server

Für CPU-Server wird mangels detaillierter Lastprofile ein konstanter mittlerer Leistungswert pro Server angenommen, der eine realistische Jahresauslastung klassischer Rechenzentren abbildet. Dieser Wert wird aus Literaturangaben zu typischen Server- und Auslastungsniveaus abgeleitet [3], [6] und je

DOI 10.5162/iCCC2026/P23

Rechenzentrumsgröße skaliert. Zeitliche Schwankungen (z. B. Tagesgang) werden in dieser ersten Modellierung nicht explizit berücksichtigt; die Einflüsse werden in der Sensitivitätsanalyse diskutiert.

GPU-Server

Die Leistungsaufnahme der GPU-Server während Training und Inferenz wird aus experimentellen Messdaten von Aquino-Brítez et al. abgeleitet [5].

Aus den im Sensors-Paper angegebenen Energieverbräuchen und Laufzeiten für verschiedene Deep-Learning-Modelle auf NVIDIA-GPUs werden für jede Architektur mittlere Leistungen für Training Inferenz bestimmt. Für das Jahreslastprofil eines GPU-Servers wird ein vereinfachtes Duty-Cycle-Modell verwendet, das den Anteil von Training, Inferenz und Idle-Betrieb abbildet. Basierend auf Literaturangaben zum relativen Energieanteil von Training und Inferenz in KI-Rechenzentren (ca. 30 % Training, 60 % Inferenz, 10 % weitere KI-Aufgaben [1], [6]) werden für jedes Szenario

$$\alpha_{\text{train}}, \quad \alpha_{\text{infer}}, \quad \alpha_{\text{idle}}$$

die Zeitanteile
festgelegt. Die mittlere elektrische Leistung eines GPU - Servers ergibt sich zu

$$\bar{P}_{\text{GPU}} = \alpha_{\text{train}} P_{\text{train}} + \alpha_{\text{infer}} P_{\text{infer}} + \alpha_{\text{idle}} P_{\text{idle}}.$$

Unter der Annahme linearer Skalierung wird diese Serverleistung mit der Anzahl GPU-Server bzw. GPUs pro Rack multipliziert und so auf die jeweilige Rechenzentrumsgröße hochgerechnet. Die Gesamt-IT-Leistung ergibt sich aus der Summe der CPU- und GPU-Teilleistungen.

3.4 PUE, Abwärmepotenzial und Wärmerückgewinnung

Die energetische Infrastruktur wird über den Power Usage Effectiveness (PUE) modelliert. Für das CPU-Basisszenario wird je Rechenzentrum ein PUE Wert von 1,4 festgelegt. Für GPU-Szenarien mit höherer Leistungsdichte und Flüssigkühlung werden, im Einklang mit der Literatur zu modernen KI-Rechenzentren, reduzierte PUE-Werte im Bereich 1,1-1,3 angenommen [1], [2], [6]. Die Gesamtleistung des Rechenzentrums ergibt sich damit aus: $P_{\text{total}} = P_{\text{IT}} \cdot \text{PUE}$

Da die aufgenommene elektrische Energie im Rechenzent rum nahezu vollständig in Wärme umgewandelt wird, wird diese Leistung als Abwärmestrom interpretiert. Für die tatsächlich nutzbare Wärmemenge wird ein technologieabhängiger Wirkungsgrad der Wärmerückgewinnung eingeführt, der sowohl die Kühltechnik (Luft vs. Flüssig kühlung) als auch Verluste in Wärmeübertragern und Speichersystemen abbildet [2].

$$E_{\text{Wärme, nutzbar}} = P_{\text{IT}} \cdot \eta_{\text{WRG}} \cdot h_{\text{Betrieb}}$$

Für die Nutzung der Abwärme wird in dieser Arbeit eine Wärmepumpenlösung zur Einspeisung in ein Fernwärmenetz unterstellt, da diese in der Literatur als zentrale Option für die Kopplung von Rechenzentren und Wärmenetzen identifiziert wird [2]. Die dafür notwendige zusätzliche elektrische Energie der Wärmepumpe wird über angenommene Jahresarbeitszahlen modelliert.

3.5 Betriebskostenmodell und Sensitivitätsanalyse

Zur Bewertung der wirtschaftlichen Effekte wird ein vereinfachtes Betriebskostenmodell auf Basis jährlicher Energieflüsse verwendet. Investitionskosten (CapEx) der IT- und Kühltechnik werden in dieser ersten Betrachtung nicht detailliert bilanziert; der Fokus liegt auf den laufen den Energiekosten (OpEx) und möglichen Erlösen aus Wärmeeinspeisung.

Für jedes Szenario werden zunächst die jährlichen Stromkosten ohne Abwärmenutzung berechnet. Im Fall mit Abwärmenutzung werden zusätzlich die Stromkosten der Wärmepumpe berücksichtigt und Erlöse aus der Einspeisung der nutzbaren Wärme in ein Fernwärmenetz angesetzt. Aus der Differenz werden die relativen Kostenvorteile bzw. -nachteile der Abwärmenutzung für CPU- und GPU-Szenarien abgeleitet. Da zentrale Parameter (Strom- und Wärmepreise, PUE-Werte, Jahresarbeitszahlen) mit Unsicherheiten behaftet sind, ist eine Sensitivitätsanalyse vorgesehen. In dieser werden die genannten Größen in plausiblen Bandbreiten variiert und die Auswirkungen auf Abwärmepotenziale und Betriebskosten untersucht. Dies ermöglicht eine robuste Einordnung, ob und unter welchen Rahmenbedingungen GPU-basierte KI-Rechenzentren durch Abwärmenutzung wirtschaftliche Vorteile gegenüber CPU-dominierten Basisszenarien bieten können.

4 Erwartete Ergebnisse und Beitrag

Auf Basis der beschriebenen Szenarien wird für jede Rechenzentrumsgrößenklasse und jeden GPU-Anteil zu nächst die jährliche IT-Leistung und Gesamtenergie ermittelt. Es ist zu erwarten, dass mit zunehmendem GPU-An teil die absolute IT-Leistung und damit das theoretische Abwärmepotenzial deutlich steigen, insbesondere in den großen und Hyperscale-Szenarien. Gleichzeitig führen niedrigere PUE-Werte in flüssiggekühlten GPU-Szenarien dazu, dass der relative Anteil der IT-Leistung an der Gesamtleistung zunimmt, während der Energiebedarf der Infrastruktur anteilig sinkt. Für die Abwärmenutzung wird erwartet, dass GPU-dominierte Szenarien aufgrund höherer Leistungsdichten und höherer Rücklauftemperaturen der Kühlmedien größere nutzbare Wärmemengen für Fernwärmesysteme bereitstellen können als CPU-dominierte Basisszenarien. Insbesondere für große und Hyperscale-Rechenzentren dürften sich bei geeigneter Auslegung der Wärmerückgewinnung und ausreichend hoher Wärmenachfrage signifikante Einspeisemengen und damit

relevante Erlöspotenziale ergeben. Im Betriebskostenvergleich werden voraussichtlich zwei gegenläufige Effekte sichtbar: Einerseits führen höhere IT-Leistungen in GPU-Szenarien zu steigenden Stromkosten, andererseits können verbesserte PUE-Werte und Erlöse aus der Wärmeeinspeisung die spezifischen Betriebskosten pro IT-Leistung senken. Die Arbeit soll aufzeigen, in welchen Größenklassen und bei welchen GPU-Anteilen Abwärmenutzung dazu beitragen kann, die zusätzlichen Energiekosten von KI-Rechenzentren teilweise zu kompensieren. Der Beitrag der Untersuchung liegt damit in einem konsistenten Modellrahmen, der

(i) den Übergang von CPU- zu GPU-basierten Rechenzentren entlang verschiedener Größenklassen abbildet,

(ii) empirisch basierte Leistungsprofile von KI-Workloads integriert und

(iii) Energiebilanz, Abwärmepotenziale und Betriebskosten unter regulatorischen Rahmenbedingungen gemeinsam betrachtet. Dies liefert eine Grundlage für Planungs- und Standortentscheidungen von Betreibern sowie für energiewirtschaftliche Bewertungen von KI-Rechenzentren als Wärmequelle.

[1] IEA: Energy and AI, 2024

[2] Borderstep: Rechenzentren 2022, 2022

[3] Bitkom: Rechenzentren in Deutschland, 2024

[4] Chen, X.; Wang, X.; Colacelli, A.; Lee, M.; Xie, L.: Electricity Demand and Grid Impacts of AI Data Centers: Challenges and Prospects., 2025

[5] Aquino-Brítez, S.; García-Sánchez, P.; Ortiz, A.; Aquino-Brítez, D.: Towards an Energy Consumption Index for Deep Learning Models: A Comparative Analysis of Architectures, GPUs, and Measurement Tools., 2025 [6] IEA: Data Centre Energy Use: Critical Review of Models and Results, 2025

[7] Hao, Y.; Zhou, H.; Tian, T.; Zhang, W.; Zhou, X.; Shen,Datacenters waste heat recovery technologies: Review and evaluation.