Intelligentes Lastspitzenmanagement in hoch-performanten Rechenzentren

Durch die Kompensation von Lastspitzen im Energiebezug in hoch-performanten Rechenzentren für die massenhafte Nutzung von generativer KI bietet sich die technische Möglichkeit, den teuren Netzausbau hierfür zu optimieren und die Netzstabilität sowie die Netzresilienz zu erhöhen.

Der rasant wachsende Strombedarf für KI-Anwendungen, sowohl beim Training großer KI-Modelle als auch bei deren Nutzung, schafft technische Herausforderungen bei der Energieversorgung – sowohl im Stromnetz und den verfügbaren Netzanschlusskapazitäten als auch hinter dem Netzanschluss beim Energiemanagement und der Energieressourcenverteilung innerhalb von hoch-performanten Rechenzentren.

Der Netzanschluss dieser Rechenzentren ist durch die regulatorischen Gegebenheiten und die Verfügbarkeit von elektrischer Energie und Netzanschlusskapazitäten, gerade in Ballungsräumen, eine wesentliche Herausforderung, die in der VDE Kurzinfo „Netzanschluss für KI“[1] im Detail betrachtet wird. Die massenhafte Nutzung von KI-Werkzeugen, insbesondere generativer KI, stellt vor allem die Netzinfrastruktur vor neue Herausforderungen und somit ist das interne Energiemanagement in diesen Rechenzentren besonders interessant für mögliche technische Optimierungen.

Deshalb beleuchten wir in diesem Beitrag die technischen Gegebenheiten der Systeme der hoch-performanten Rechenzentren und zeigen Lösungsvorschläge für die Kompensation von Lastspitzen auf. Diese ergeben sich aus der Schwankungsbreite des Strombedarfes während der Trainingsphase aber auch während der Inferenzphase durch die massenhafte Nutzung generativer KI. Beide Phasen weisen große Unterschiede in der Energieaufnahme auf und Lösungsansätze sind dringend geboten, um den optimierten und effizienten Ausbau dieser modernen Rechenzentren, hier *Artificial Intelligence Data-Center (AI-DC)* genannt, voranzutreiben.

In diesem Beitrag analysieren wir den dynamischen Energiebedarf und geben Lösungsansätze für ein Lastspitzenmanagement, das auf unterschiedlichen Ebenen in AI-DC technisch umgesetzt werden kann. Für den Bereich großer Rechenzentren für KI-Trainings werden folgende Ansätze zur Stromstabilisierung vorgeschlagen[2]:

- Rein softwarebasierte Ansätze
- Glättung der GPU-Leistung
- Energiespeicherlösungen

Das hier vorliegende Papier behandelt die Frage, inwieweit geeignete hardwaretechnische Ansätze dazu beitragen können, Lastspitzen im Training und auch bei der Inferenz von KI-Modellen zu glätten bzw. zu reduzieren.

Training und Inferenz im Kontext von KI-Modellen

Training und Inferenz sind die zwei zentralen Prozess-Phasen im Lebenszyklus eines KI-Modells. Während beim Training eines neuronalen Netzwerkes das Modell mit Daten und *„Gewichten“* (engl. „Weights“) befüllt wird, werden beim Zugriff auf das trainierte Netzwerk, der Inferenz, Abfragen mittels der vorliegenden Modelle bearbeitet. Beide Phasen bestehen aus klar definierten Schritten, die sich technisch und konzeptionell unterscheiden.

Durch Training „lernt“ das Modell

Beim Training wird ein KI-Modell aus großen Datenmengen so angepasst, dass es Muster, Zusammenhänge und Strukturen erkennt. Die einzelnen Parameter bzw. Weights der neuronalen Netze werden iterativ optimiert. Folgende Prozess-Schritte werden dabei durchlaufen:

- Datensammlung – Sammlung von Rohdaten, z. B. Sprache, Texte oder Bilder
- Datenbereinigung – Entfernung von Fehlern, Duplikaten und nicht konformen Inhalten
- Tokenisierung – die Zerlegung komplexer Daten in kleinere, eindeutige Einheiten (Tokens)
- Modellinitialisierung – Start mit zufälligen „Weights“ der Netzwerkknoten
- Vorwärtsdurchlauf – Modell erzeugt eine Vorhersage
- Fehlerberechnung – Vergleich mit der korrekten Antwort (Loss – Verlustfunktion)
- Backpropagation – „Weights“ werden angepasst
- Iteratives Training – Millionen bis Milliarden Wiederholungen werden durchgeführt

Durch Inferenz „antwortet“ das Modell

Bei der Inferenz nutzt das Modell seine zuvor „gelernten“ Parameter, um Vorhersagen zu erzeugen (ohne dabei weiter zu lernen). Hier werden folgende Prozess-Schritte durchlaufen:

- Eingabe vorbereiten – Tokenisierung des Nutzerinputs
- Vorwärtsdurchlauf – Modell berechnet Wahrscheinlichkeiten für nächste Token
- Sampling – Auswahl der nächsten Tokens
- Dekodierung – Tokens werden zurück umgewandelt, z. B. in Text
- Antwortgenerierung – Modell liefert finalen Output

Inferenz ist der Prozess, bei dem ein zuvor trainiertes Modell ohne Änderung der Weights eine Ausgabe berechnet. Es findet nur ein Vorwärtsdurchlauf statt.

Technisch betrachtet unterscheiden sich Training und Inferenz damit durch ihre Zielsetzungen, aber auch durch ihre mathematischen Abläufe und ihre Hardwareanforderungen.

Beim Training spielen Graphics Processing Unit (GPU)- und Tensor Processing Unit (TPU)-Cluster eine wesentliche Rolle, während bei der Inferenz GPUs, CPUs und Edge Chips die bestimmenden Komponenten sind. Das Training bewegt sich zeitlich im Bereich von Wochen und Monaten, die Inferenz dagegen im Bereich von Millisekunden bis Sekunden.

In der Folge unterscheiden sich Training und Inferenz von KI-Modellen auch systemarchitektonisch, z.B. in der Art, wie Rechenlast, Speicher, Parallelisierung und Hardwareressourcen orchestriert werden.

Beim Training geht es um hochgradig verteilte, synchronisierte Lastverteilung, da massenhaft Daten anhand des verwendeten Modells mehrfach verarbeitet werden.

Hierbei kommt High-Performance-Computing (HPC) zum Einsatz, das durch massive Parallelisierung über tausende GPUs, schnelle Kommunikationsschnittstellen (Interconnects) und komplexe Mechanismen zur Verteilung von Computing-Ressourcen gekennzeichnet ist.

Die Inferenz dagegen ist durch latenzoptimierte Echtzeitalgorithmen und oft dezentrale Ausführung geprägt. Zu den Optimierungszielen zählen hier niedrige Latenz, hoher Datendurchsatz, parallele Anfragenbearbeitung und Wirtschaftlichkeit durch Kostenreduktion.

Energieaufnahme für Training und Inferenz[3]

In Rechenzentren kommt es zu Schwankungen im Energiebedarf beim Training von KI-Modellen und der Inferenz. Diese Lastspitzen sind dynamisch und breit im Spektrum, was eine unterschiedliche Last für das elektrische Energienetz bedeutet. Vor allem das Training stellt einen großen Energiebedarf dar, bei dem die Computing-Knoten unter Voll-Last stehen und Lastspitzen unterschiedlicher Frequenz entstehen.

Doch auch die Inferenz spielt durch die massenhafte Nutzung von KI eine zunehmend größere Rolle, die vor allem das dynamische Verhalten beim Energiebedarf beeinflusst. Faktoren wie Temperaturschwankungen über die Jahreszeiten und die Integration von immer komplexeren Computing-Einheiten mit größerem Energiebedarf in den jeweiligen Ebenen dieser Einheiten, wie dem gesamten AI-DC, den Racks, der Verschaltung auf den Platinen oder den GPU-, TPU-, CPU-Chips stellen den Ausbau und die Auslegung des Netzanschlusses an das Energienetz vor Herausforderungen.

Das Training großer KI-Modelle benötigt Millionen GPU-Stunden. Dafür werden mehr als 5000 GPUs in einem Cluster benötigt.[4] Moderne GPUs (z. B. NVIDIA H100) liegen bei ca. 700 W – 1 kW unter Voll-Last. Bei 5000 GPUs ergeben sich damit 3,5 - 5 MW elektrische Leistung nur für die Recheneinheiten. Mit Kühlung und unter Annahme einer Power-Usage-Effectiveness (PUE) von 1,2 bis 1,5 ergibt sich damit eine Gesamtlast von 4 - 7 MW. Ein einzelner Trainingslauf eines großen Modells kann damit eine Last im Bereich eines kleinen Industrieparks verursachen.

Inferenzen dagegen sind im Vergleich dazu geringer, wenn die Anfragen moderat sind. Zum Beispiel meldete Meta / Facebook Billionen Anfrage pro Tag in ihren Rechenzentren. Genaue Angaben von Zahlen zu Lastspitzen in AI-DC und Servern großer Plattformen, wie Facebook, Amazon und ByteDance liegen nicht vor.

Man unterscheidet Lastspitzen beim Training von KI-Modellen, die im MW-Bereich pro Cluster mit mehreren tausend GPUs liegen. Bei Inferenz kommen einzelne Vorgänge mehrfach vor und liegen im Millisekunden-Bereich. Durch Akkumulation paralleler Prozesse entstehen Lastspitzen über ein großes Frequenzspektrum, bei denen eine Kompensation auf unterschiedlichen Ebenen der Computing-Einheiten vorgenommen werden kann. Dies dient dem Ausgleich und der Pufferung des Bedarfes der Energie, die durch den Netzanschluss in das AI-DC gespeist wird.

Die Schwankungen einzelner GPUs liegen je nach Modell typischerweise zwischen 350 W und 1 kW. Für einzelne Server mit z. B. 4 - 8 GPUs liegen diese Schwankungen zwischen 2 kW und 10 kW.

Energiebedarfe einzelner GPUs:

- *NVIDIA H100* (aktueller Standard für Training von KI-Modellen) - gemessene Energieaufnahme oft zwischen 800 bis1000 W bei FP8/FP16-Training. Wird in großen Clustern mit 8 GPUs pro Server eingesetzt.
- *NVIDIA A100* Energieaufnahme zwischen 450 bis 550 W im Training
- *NVIDIA L40S* häufiger Einsatz bei Inferenz mit Energieaufnahmen zwischen 350 bis 450 W

Lastspitzen pro KI-Server:

Ein typischer Server für den Einsatz in AI-DC enthält 4 - 8 GPUs neben CPU, RAM, Netzteil und Netzwerkkomponenten, wie Switches. Dazu zählen Konfigurationen wie:

- 4 x H100 mit 3 - 5 kW zum Training kleinerer KI-Modelle
- 8 x H100 mit 6 - 10 kW als Standard für das Training großer KI-Modell
- 8 x A100 mit 4 - 7 kW als weitverbreiteter Standard in Cloud-Rechenzentren
- 4 x L40S mit 1,5 - 2,5 kW als Inferenz-Server

Das Lastprofil ist stark schwankend, abhängig von Tageszeit, Nutzeraktivität und Peaks wie beispielsweise Aktivitäten zu viralen Posts oder Events. Lastspitzen können innerhalb von Sekunden bis Millisekunden auftreten. So kann beispielsweise eine Inferenz Hardware für Chatbots morgens 20 % Auslastung beanspruchen und abends 400 % mehr Anfragen verarbeiten. Dabei springt die GPU-Last von unter 100 W auf 350 W.

Training von KI-Modellen: Energieaufnahme pro Server (8 x H100)

Typische Last:

- Training: 6 - 10 kW
- Lastspitzen: 10 - 12 kW kurzzeitig

Das Lastprofil ist stabil mit einer Beanspruchung von 90 bis 100 % GPU der Leistungsaufnahme, hingegen ist die Auslastung der CPU eher moderat bei 10 bis 30 %.

Inferenz: Lastprofil pro Server (4 x L40S oder 8 x T4)

Typische Last:

- Durchschnitt: 1 - 2 kW
- Lastspitzen: 2 - 3 kW insbesondere bei großen Anfragebedarf

Das Lastprofil weist eine hohe Variabilität und ein großes Frequenzspektrum auf, abhängig von der Anzahl der Anfragen (Queries per Second, QPS). Hier ist die Energieaufnahme der CPUs oft höher als beim Training.

Aufgrund dieser Unterschiede in der Energieaufnahme zwischen Training und Inferenz kommt es in AI-DC zu Schwankungen im Leistungsbedarf, die jedoch für die Auslegung des Netzanschlusses relevante technische Herausforderungen darstellen. Wir wollen im Folgenden darauf eingehen und mittels technischer Methoden auf Kompensationsmöglichkeiten innerhalb von AI-DC hinweisen, mit denen deren Ausbau möglichst weit vorangetrieben werden kann.

Lösungsansätze zur Lastspitzenkompensation mittels Energiespeicher

Neben softwarebasierten Lösungsansätzen, wie sie etwa durch die Veröffentlichung neuer KI-Modelle von Unternehmen wie Anthropic[5] oder DeepSeek[6] vorangetrieben werden, gibt es zahlreiche hardwarebasierte Konzepte, die speziell auf die Herausforderungen von Lastspitzen in Rechenzentren zugeschnitten sind. Besonders effektiv erweisen sich dabei On-Board-Lösungen auf GPU-Ebene sowie in die Racks integrierte Energiespeichersysteme.

Die On-Board-Lösung nahe der GPU zur Leistungsanpassung nutzt GPU-Power-Controller, um mit minimaler Latenz Leistungsschwankungen zu regeln.

Ein Beispiel hierfür ist die Hardware von NVIDIA GB200, die durch die Einstellung eines Minimum Power Floor (MPF) etwa bei 90 % der Thermal Design Power (TDP) die Energieaufnahme reduziert und damit zur Lastspitzenkompensation herangezogen wird. Trotz dieser Maßnahme liegt der Energiemehrbedarf bei ca. 10 %, da die GPU auch in Leerlaufphasen Energie aufnimmt. Eine Herausforderung besteht darin, dass ein höherer MPF-Wert mit größeren Verlusten korreliert. Diese Maßnahme reicht möglicherweise nicht aus, um eine effektive Kompensation zu erreichen.

Ein weiterer Ansatz zur Kompensation von Lastspitzen, ohne den Gesamtenergiebedarf zu steigern, sind Rack-basierte Energiespeichersysteme. Diese Systeme erfüllen alle Anforderungen an eine moderne, skalierbare Infrastruktur. Sie benötigen eine Echtzeit-Lastmessung zur präzisen und latenzoptimierten Regelung, ausreichende Speicherkapazität, um die Schwankungen der Lasten auszugleichen sowie ein hohes dynamisches Verhalten beim Laden und Entladen, um plötzliche Lastspitzen auszugleichen. Die Funktionsweise besteht darin, im laufenden Betrieb, Energie in den unterschiedlichen Komponenten zu speichern und diese während rechenintensiver Phasen, zur Kompensation von Lastspitzen wieder abzugeben.

Simulationen zeigen, dass solche Systeme Lastspitzen ausgleichen können, ohne größere Kapazitäten an Primärenergie bereithalten zu müssen. Der Vorteil liegt darin, dass keine zusätzliche Energie aufgewendet wird, um die nötigen Anforderungsspezifikationen erfüllen zu können und dabei die Spitzenlast reduziert wird, was langfristig Kosten für den Ausbau der Infrastruktur senkt. Dies kann mit vertretbaren Investitionen in Batterien oder Kondensatoren, Platzbedarf in Rechenzentren und Wirkungsgradverlusten durch Lade- und Entladevorgänge technisch effizient umgesetzt werden.

Die beste Strategie zur Kompensation von Lastspitzen kombiniert GPU-basiertes Power Smoothing und Rack-Energiespeicher. On-Board-Lösungen auf GPU-Ebene sind ideal für sofortige, hardwaregestützte Kompensation in kleineren Spannungsbereichen, während Rack-Energiespeicher eine langfristige, skalierbare Lösung ohne zusätzliche Energieressourcen und mit maximaler Flexibilität bieten. Während softwarebasierte Lösungen schnell einsatzbereit, aber mit Leistungseinschränkungen verbunden sind, ermöglichen On-Board-Lösungen und Rack-Energiespeicher die effektivsten und nachhaltigsten Ansätze. Eine ebenen-übergreifende Betrachtung unterschiedlicher Hardware- und Infrastruktur-Komponenten ist dabei entscheidend, um die wachsenden Anforderungen an Energiebedarf bei Training und Inferenz langfristig bedienen zu können.

Netzanschluss und Lastspitzen-Management auf Hardwareebene

Im Gegensatz zu konventionellen Cloud-Rechenzentren sind AI-DC aufgrund geringerer Anforderungen an Latenz und Netzknoten nicht mehr zwingend an die FLAP-D-Standorte, also die fünf wichtigsten Standorte für europäische Rechenzentren in Frankfurt, London, Amsterdam, Paris und Dublin, oder generell urbane Ballungsräume gebunden. Ausschlaggebend sind heute insbesondere die Verfügbarkeit hoher Netzanschlussleistungen und geeignete Flächen für den Ausbau der AI-DC. Die Leistungsanforderungen steigen deutlich: In Europa werden derzeit AI-DC Projekte mit Netzanschlussleistungen von bis zu 900 MW entwickelt.

Die Leistungs- und Kapazitätswerte von im Bau befindlichen Batterie-Großspeichern (BESS) erreichen ebensolche Größenordnungen (z.B. RWE Gundremmingen mit 400 MW / 780 MWh) sowie in UK Coalburn (1 GW / 2 GWh) und Thorpe Marsh (1,4 GW / 3,1 GWh), für deren Netzanschlussbegehren die Netzbetreiber zunehmend umfangreiche technische Anforderungen verlangen. Große AI-DC Projekte werden in punkto Anforderungen hier zunehmend gleichbehandelt.

Diese Entwicklungen verdeutlichen, dass sich sowohl die Anforderungen an die Energieinfrastruktur als auch die Standortwahl für AI-DC grundlegend verändern.

Die Verfügbarkeit hinreichend großer elektrischer Energiemengen und Netzanschlusskapazitäten wird zunehmend zum entscheidenden Wettbewerbsfaktor für die Ansiedlung von AI-DC in Europa.[7]

Integration von Energiespeichern:

Den genannten Lastschwankungen in AI-DC kann auf verschiedenen Ebenen durch eine Zusammensetzung aus verschiedenen Energiespeichern entgegengewirkt werden. Hierfür ist eine hierarchische Unterteilung in Campus-, Data Hall-, Rack- oder Server-Ebene notwendig.

Campus-Ebene

Batterie-Großspeicher (BESS) unterstützen die systemdienliche Nutzung erneuerbarer Energien durch Energiespeicherung, Lastverschiebung und Spitzenlastkappung. Angebunden auf der 110kV- oder 380kV-Ebene sind solche Anlagen jedoch nicht geeignet, schnelle Laständerungen, wie sie durch die GPU-Auslastung auf Server- oder Rack-Ebene entstehen, ortsnah zu kompensieren. Die Belastungsspitzen innerhalb von AI-DC würden sich hier über sämtliche, dazwischenliegende Komponenten im Stromsystem übertragen.

Für sogenannte Tier-3-Rechenzentren sind bereits Notstromversorgungen für 72 Stunden gefordert, die den Betrieb beim Netzausfall sicherstellen. Diese werden heute fast ausschließlich durch Diesel-Generatoren bereitgestellt. Zunehmend wird jedoch geprüft, ob man BESS mit mehreren Stunden Laufzeit einsetzen kann, um einen Großteil der Netzausfälle abdecken zu können sowie zusätzliche Leistung bereitzustellen und gleichzeitig Erlöse im Energie- und Regelleistungsmarkt zu erzielen. Die in den letzten Jahren stark gesunkenen Batteriekosten machen dieses Konzept wirtschaftlich attraktiv und reduzieren gleichzeitig den Verschleiß der Dieselgeneratoren. Das reduziert ebenso die CO_2-Emissionen der AI-DC.

Data Hall-Ebene

Unterbrechungsfreie Stromversorgungen (USV) dienen der Überbrückung kürzerer Netzausfälle und Netzstörungen, wie z.B. Spannungsschwankungen (Fault Ride-Through, FRT). USV-Anlagen überbrücken typischerweise 5–15 Minuten und in besonders kritischen Rechenzentren auch bis zu 45 Minuten Voll-Last.

Heutige USV-Lösungen in Data Centern sind von den Servern / Racks aus Sicherheitsgründen räumlich getrennt. Rund 70 % der heutigen USV-Systeme verwenden Blei-Säure-Batterien, wobei sich der Markt jedoch zunehmend in Richtung Lithium-Ionen-Technologie entwickelt. Diese eignet sich besonders für 800 V DC-Architekturen, da moderne Batteriespeicher (BESS) bereits heute auf 800 V DC und künftig auf bis zu 1.500 V DC ausgelegt sind.[8]

Der USV wird die elektrische Leistung über 400/480 V AC zugeführt, in der USV gleichgerichtet für den Anschluss der USV-Batterie und anschließend auch zum Zwecke der Verbesserung der Netzqualität wieder in 400/480 V AC umgewandelt und den Racks zugeführt.

Es wäre aus Effizienzgründen deutlich besser, direkt aus der USV eine DC-Verteilung zu den Servern / Racks auf der Spannungsebene 700 / 800 V DC vorzunehmen, was zwei Wandlerstufen spart und damit Effizienzgewinne von 2-4 % erwarten lässt.

Rack-Ebene

Das Rack wird heute von der dreiphasigen 400/480 V AC aus der USV versorgt, die im Rack mittels eines aktiven Gleichrichters mit Power-Factor-Compensation (PFC), wie 2-Level oder 3-Level Topologien, z.T. mit Totem Pole, zunächst in eine Gleichspannung von 600/700 V DC gewandelt wird.[9]

Auf dieser Spannungsebene ist oft ein geschalteter Zwischenkreiskondensator installiert, der damit stärker entladen werden kann und so besser zur Stabilisierung beiträgt.[10]

Die Gleichspannung wird über einen isolierenden, resonanten DC/DC Wandler (meist LLC, aber auch DAB-SRC, ein- und drei-phasig) in 48 V DC gewandelt.

Als Leistungshalbleiter für die PFC-Stufe und die Primärseite des resonanten DC/DC Wandler werden SiC-MOSFETs oder GaN HEMTs eingesetzt, die eine hohe Taktfrequenz und damit eine kompakte Bauweise bei sehr guter Effizienz erlauben.

Für die Nivellierung der Fluktuationen auf Rack-Ebene sind Energiespeicher notwendig, die über eine hohe C-Rate verfügen, bei denen also das Verhältnis aus Leistung und Energieinhalt sehr groß ist. Neben den üblichen Kondensatoren, die ohnehin Bestandteil der PFC und LLC-Wandlertopologien sind, können hierfür sinnvollerweise Supercaps / Doppelschichtkondensatoren eingesetzt werden, deren C-Rate bei bis zu 1000 liegt. Damit decken Supercaps Lastfluktuationen im Sekunden-Bereich ab.

Abschätzungen zeigen, dass man für ein Rack mit 32 kW Spitzen-Leistungsaufnahme einen Supercap-Speicher mit ca. 40 F Kapazität benötigt, um bei Inferenz eine weitgehende Nivellierung des Leistungsbezugs durch das Rack zu erzielen. Der Supercap-Speicher wird im einfachsten Fall über einen nicht-isolierenden Hochsetzsteller an die 48 V DC angeschlossen, was eine einfache Integration in bestehende Stromversorgungen erlaubt.

Neben der Nivellierung des Leistungsbezugs weist die Integration von Supercap-Speichern in die Rack-Stromversorgung weitere Vorteile auf:

- Kostengünstigere Dimensionierung der PFC sowie der Resonanzwandler (LLC-Converter), da sich das Verhältnis aus Spitzen- und Durchschnittsleistung reduziert.
- Der Supercap-Speicher kann bei Komplettausfall die Rack-Stromversorgung für einige Sekunden übernehmen. Daraus ergeben sich Freiheitsgrade im Zusammenspiel mit der unterbrechungsfreien Stromversorgung (USV).

Es ist technisch möglich, dass Supercaps als Komponenten Hochleistungsbatterien (20 bis 25 C Rate) ablösen, was zu einer weiteren Entlastung der Sammelschienen-, USV- und Mittelspannungsanlagen sowie der Transformatoren führen könnte.

Die heute auf Rack-Ebene basierende Spannungsversorgung mit 48 V DC stößt aufgrund des stark steigenden Leistungsbedarfs von KI-Rechenzentren an ihre technischen Grenzen. Daher zeichnet sich ein Übergang zu 800 V DC ab, wie er unter anderem von der Firma NVIDIA propagiert wird.

Server-Ebene

Die 48 V DC werden anschließend über weitere DC/DC-Wandlerstufen im Server bis auf eine Spannung von 1 V unmittelbar an der GPU heruntergesetzt. Eine einzelne GPU kann hierbei 1000 A Strom aufnehmen bei einer Leistung von 1 kW, was extreme Anforderungen an die Kühlung stellt.

Spezielle Kondensatoren fangen schnelle Transienten von Lastspitzen im Mikrosekundenbereich ab.

Ein neuer Trend scheinen „Sidecar-Racks“ zu sein, bei denen ein zusätzliches Serverrack, zur Versorgung bereitgestellt wird. Dies wird mit Wechselstrom in Hochspannungs-Gleichstrom bei 800V umgewandelt, bevor dieses an das AI-DC-Rack angeschlossen wird. Man verspricht sich hieraus eine effiziente Stromversorgung und geringere Verluste. Hierbei kommen Technologien wie GaN zum Einsatz, um die Stromumwandlung zu optimieren und die Integration zu vereinfachen.

Fazit

Die massenhafte Nutzung von KI-Werkzeugen, insbesondere generativer KI, stellt den Ausbau von AI-DC vor besondere Anforderungen. Der Netzanschluss an die Primärenergieversorgung dieser Rechenzentren ist durch die regulatorischen Gegebenheiten und die Verfügbarkeit von Energie – gerade in Ballungsräumen – eine wesentliche Herausforderung. Um hier einen Ausbau von AI-DC auf kürzeren Zeitskalen zu ermöglichen und zu beschleunigen, beleuchten wir in diesem Beitrag die Optimierungsmöglichkeiten in der technischen Umgebung dieser AI-DC.

Wir stellen fest, dass es zwischen dem Training von KI-Modellen und der Inferenz Unterschiede in der transienten Analyse im Energiebedarf gibt. Zudem gibt es zwischen dem Primäranschluss und auf der Ebene der Computing-Knoten der GPUs und CPUs einen großen, noch nicht fest definierten Unterschied an verwendeten Spannungen. Hiernach richten sich aus unserer Betrachtung die Kompensationsmöglichkeiten der Schwankungen im Energiebezug der AI-DC. Die Leistungsstufen von AI-DC-Racks sind mehrere hundert Kilowatt groß und drohen teilweise, die traditionelle 48V-Stromverteilungsarchitektur zu überlasten. Um dies zu bewältigen, vollzieht man den Übergang zu einer 800V-Verteilung, um Leistungsverluste und Kupferbedarf zu reduzieren. Wesentliche Komponenten sind hierbei Gleichstromwandler, bei denen von 800V auf 48V heruntergesetzt wird. Damit wird eine Kompatibilität mit bestehenden Systemkomponenten gewährleistet.

In dieser Entwicklung steht die Leistungselektronik auf Basis von GaN (Galliumnitrid) und SiC (Siliziumcarbid) für mittlere Spannungen im Fokus. Damit können verlustarm höhere Schaltfrequenzen und eine effiziente Leistung auf der Primärseite ermöglicht werden und die Technologie kann in Deutschland gefertigt werden. Die DC/DC-Wandlerstufe, die im Open-Loop-Betrieb mit konstanter Frequenz arbeitet, liefert eine präzise 48V-Ausgangsspannung mit einem Wirkungsgrad von 98 %.

Die aktuelle Forschung zielt darauf ab, die Leistungselektronik weiter zu optimieren, indem die Anzahl der Wandlungsstufen reduziert wird, wobei die GaN-Technologie für Niederspannung zusätzliche Effizienzsteigerungen verspricht. Ein aktueller Trend liegt hierbei in den sogenannten „Sidecar-Rack“-Systemen auf Basis von GaN für mittlere Spannungen.[11]

Langfristig müssen Rechenzentren eine mehrstufige Speicherarchitektur nutzen: Spezielle Kondensatoren auf Serverebene für den Millisekunden-Bereich, Supercaps für den Ausgleich von Lastschwankungen im Sekunden-Bereich, USV-Batterien für die sofortige Überbrückung im Minuten-Bereich, Lithium-Ionen-BESS für mehrere Stunden sowie Langzeitspeicher (ESS) wie Eisen-Oxid- oder Vanadium-Flow-Batterien für eine Versorgung über mehrere Tage. Diese Kombination bietet sowohl eine hohe Leistungsfähigkeit als auch lange Autonomiezeiten und wird die zukünftige Stromversorgung von hochperformanten Rechenzentren, den sogenannten AI-DC, prägen.

Führende Rechenzentrumsbetreiber setzen daher zunehmend auf eine mehrstufige Systemarchitektur, in der netzseitige Batteriespeicher mit leistungsfähigen USV-Systemen kombiniert werden. Die konkrete Auslegung hängt vom Standort und der Nutzungsanforderung an das Rechenzentrum (Training oder Inferenz) sowie den Vorgaben der Netzbetreiber ab.

Die Integration von erneuerbaren Energiequellen im Ausbau der Netze verläuft deutlich langsamer als der derzeit rapid steigende Leistungsbedarf der AI-DC.[12] Um eine bestmögliche, gemeinsame Entwicklung und zuverlässige Integration zu ermöglichen, ist die Koordination über mehrere technische Ebenen – von der Netzanbindung über „on-board“-Lösungen bis hin zur CPU bzw. GPU – notwendig.

Da es sich um unterschiedliche Spannungsebenen bei den genannten Systemkomponenten handelt, ist ein Standardisierungsprozess sicherlich von Vorteil, damit man die Baugrößen der einzelnen Kompensationsvorrichtungen adäquat auslegt und den Markt dafür öffnet.

[1] Dudek, D. et al. (2026). VDE Kurzinfo *Netzanschluss für KI*. (letzter Zugriff: 13.07.2026)
[2] Microsoft, OpenAI, NVIDIA (2025). Power Stabilization for AI Training Datacenters. https://arxiv.org/abs/2508.14318v2
[3] LiquidWeb. A100 vs H100 vs L40S: A simple side-by-side and how to decide [2508.14318] Power Stabilization for AI Training Datacenters
[4] KI Bundesverband (2025). KI-Infrastruktur für das Training großer Modelle in Deutschland.
[5] Handelsblatt (10.06.2026). Anthropic veröffentlicht neues Hochleistungsmodell „Fable“.
[6] DeepSeek (2026). https://www.deepseek.com/en/ (letzter Zugriff: 13.07.2026)
[7] R. Bucher (2026) A New Market for BESS & E-Statcom: Smoothing AI Data-Center Load Patterns, The smarter e Europe conference, Munich, https://www.thesmartere.com/talk/use-of-battery-storage-systems-to-optimize-the-grid-connection-capacity-of-data-centers
[8] G.J. Hackett (2020) Safety aspects of Li-ion batteries used in mission-critical UPS batteries for data centres", Saft, https://saft.com/en/media-resources/knowledge-hub/white-papers/safety-aspects-li-ion-batteries-used-mission-critical-up
[9] https://www.st.com/content/dam/OLM%20Email%20Marketing/2025/Asia%20Pac/Event/taiwan-techday-2025/PE_01_ORv3_5.5kW_AI_Server_Power_Platform.pdf (letzter Zugriff: 13.07.2026)
[10] Sam Abdel-Rahman: “12 kW PSU for AI Servers featuring 113W/in3 with integrated Peak Shaving and Hold-up Functionalities”, PCIM 2026
[11] P. Scalia et al. (2025). Renesas Electronics White Paper: Power Architecture Evolution in Data-Centers - i.R.d. Open-Compute-Project.
[12] R. Bucher (2024) BVES Investor Summit: De-risking the high voltage grid access for BESS & data centers by early contractor involvement, Umweltforum.

Autoren (alphabetisch):

Dr. Ralf Bucher, H&MV Engineering

Dr.-Ing. Damian Dudek, VDE ITG

Prof. Dr.-Ing. Gerd Griepentrog, TU Darmstadt

Gareth Hackett, MBA, H&MV Engineering

Prof. Dr. sc. Andreas Ulbig, RWTH Aachen + Fraunhofer FIT

Dr. phil. nat. Matthias Wirth, VDE ITG

Kontakt:
VDE Verband der Elektrotechnik
Elektronik Informationstechnik e.V.
Merianstraße 28
63069 Offenbach am Main
Tel. +49 69 6308-360
damian.dudek@vde.com