

# Power Delivery Roadmap for AI

By Llew Vaughan-Edmunds, Senior Director, Product Management & Marketing, Navitas Semiconductor & Mattia Magnatta, Student

## Supporting the Exponential Power Demands of the AI Revolution

Artificial intelligence (AI) enables computers and machines to learn from experience, adapt to new inputs, and perform tasks with human intelligence. AI can be narrowed down into three main categories: machine learning, deep learning, and generative AI.

Machine learning (ML) analyzes data to make predictions or decisions without programming for specific tasks. ML models are trained on existing algorithms to recognize patterns and based on those patterns, predict future outcomes.

Deep learning (DL), a subset of ML, uses “neural networks” which are programmed to make decisions similar to the human brain by weighing options, problem solving, and arriving at conclusions. While ML focuses on predicting outcomes based on statistical algorithms, DL identifies patterns and relationships within data, mimicking human decision-making. This approach is used for tasks such as recognizing objects in images or understanding different languages.

Generative AI builds on ML and DL techniques to create new content by learning from existing content. It generates images, text, videos, and other types of material by using patterns learned from previously created human-made data. For example, it can produce original artwork, compose music, or write text that reflects human creativity, based on its training data.

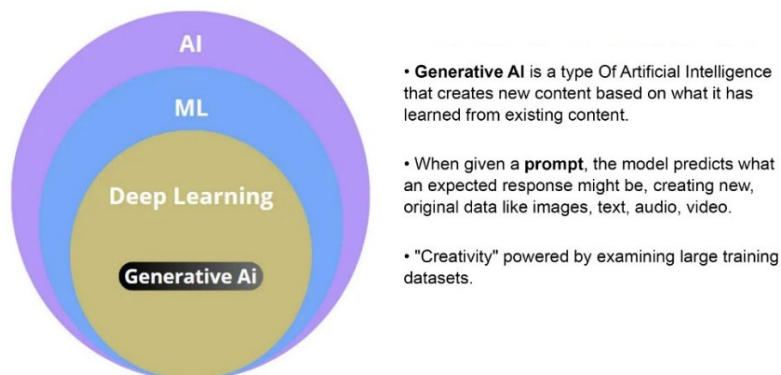
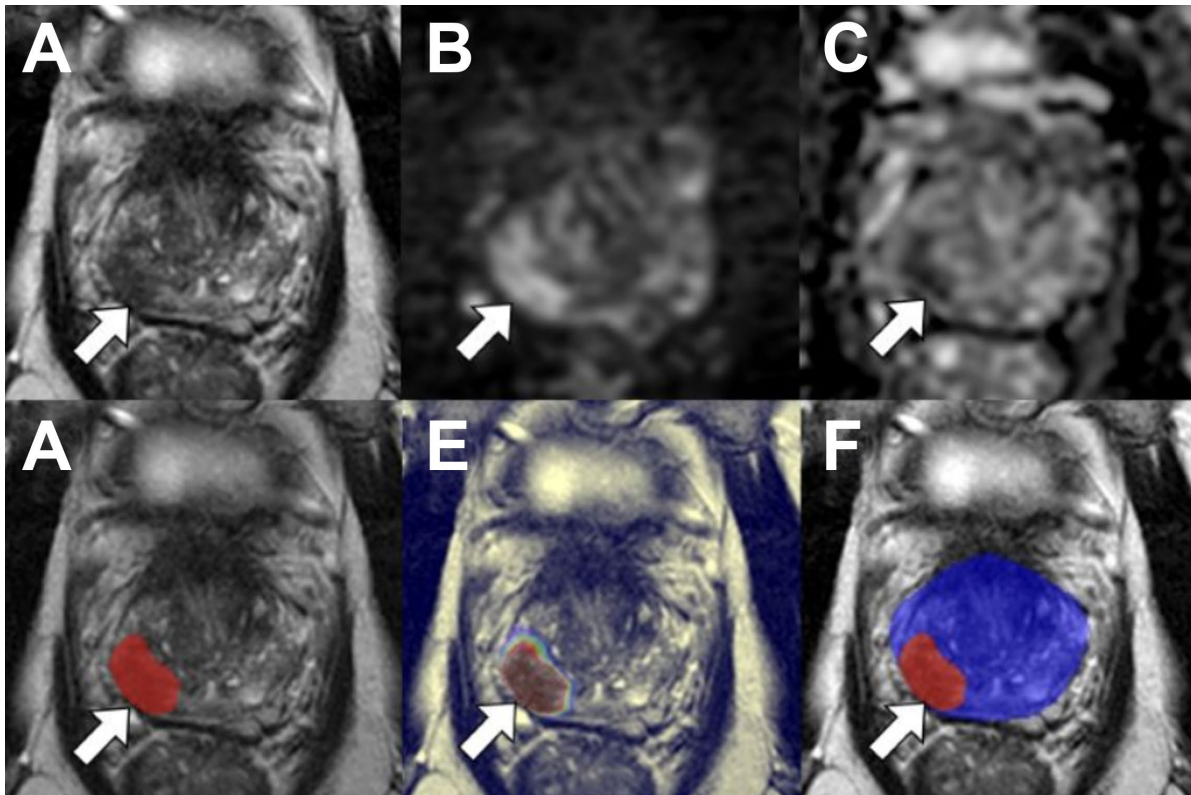


Figure 1: AI comprises Machine Learning (ML) to identify patterns, Deep Learning (DL) that mimics behaviors and Generative AI, which creates content based on the ML and DL results<sup>(1)</sup>

Featured in *Electronic Product Design & Test* – November 2024

Markets such as manufacturing, automotive and healthcare will be affected by all subsets of AI. Autonomous driving systems, for example, use ML to analyze data from a vehicle's surroundings, detect driving patterns, road conditions and potential dangers and, based on the data, take appropriate action. Manufacturing will accelerate into Industry 4.0 with AI programs supporting everything from predictive maintenance to enhanced productivity, while logistics will be made more efficient using AI techniques to minimize time spent picking, packing and delivering products from warehouses to customers. In healthcare, while medical professionals may not be replaced by AI, the technology will support accurate, rapid diagnoses that allow them to focus on patient care. Recently, the American Association of Medical Colleges tested the difference between doctor and highly-trained AI diagnoses, noting that: "based on data from thousands of images and sometimes boosted with information from a patient's medical record, AI tools can tap into a larger database of knowledge and can scan deeper into an image and pick up on properties and nuances among cells that the human eye cannot detect."<sup>(2)</sup>



*Figure 2: The top row shows how a doctor would identify a tumor from an MRI image. The bottom row shows how AI identifies the tumor and indicates potential affected areas with color coding<sup>(3)</sup>*

Featured in *Electronic Product Design & Test* – November 2024

## Growth in Data and Data Center Power

In the United States alone 403 million terabytes of data are created each day<sup>(4)</sup>. To put this into context one terabyte is roughly 250,000 professional photos. The need to process, store and transmit this ever-growing volume of data is beckoning in a new era of data centers and hardware. And the fastest growing demand is coming from the need for racks with the capacity to handle data for AI. This is no surprise given that it took five days for ChatGPT to reach one million users and two months to reach 100 million and that a ChatGPT query uses 10 times more data than a regular Google search<sup>(5)</sup>.

This brings with it the significant challenge of meeting demand for power. One query to ChatGPT uses approximately as much electricity as could light one lightbulb for about 20 minutes<sup>(6)</sup> and Goldman Sachs projects that AI will lead to a 160% increase in data center power usage by 2030. The International Energy Agency (IEA) reports that the exponential growth of AI will see data center energy double by 2026, while GridStrategies forecasts the power demand due to data use will double in gigawatt consumption based on a 2023 analysis. And it's not just about energy demand – supporting predicted growth will not only require more power generation but will also create a rise in data center CO2 emissions, which could more than double between 2022 and 2030<sup>(5)</sup>.

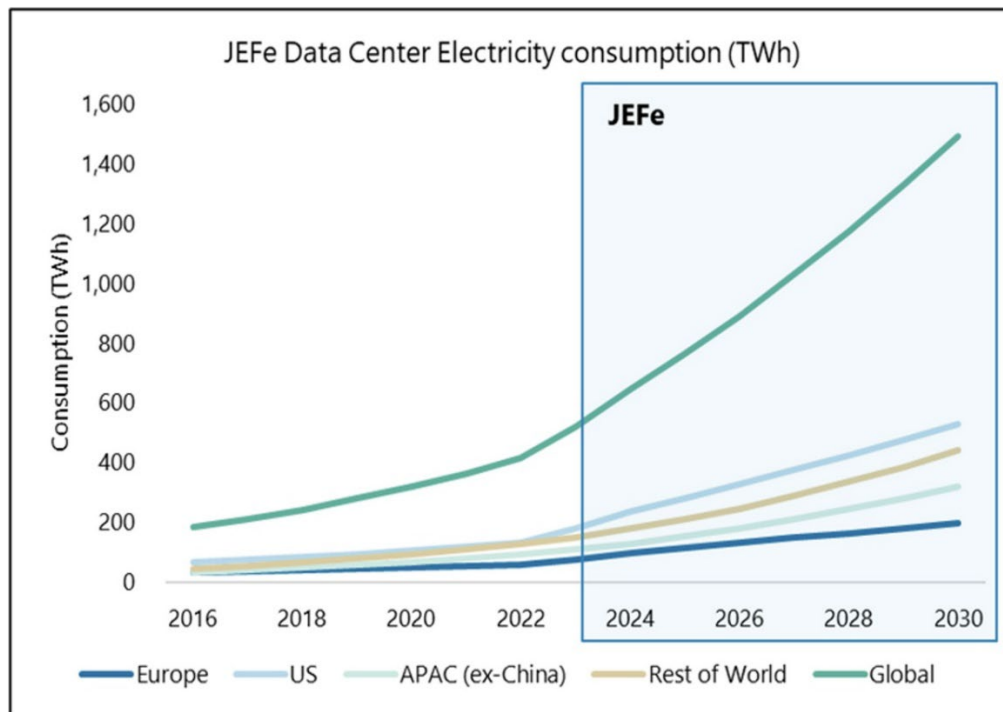


Figure 3: Global Data center electrical power consumption will increase 3x by 2030, due to data-processing and AI processing<sup>(7)</sup>

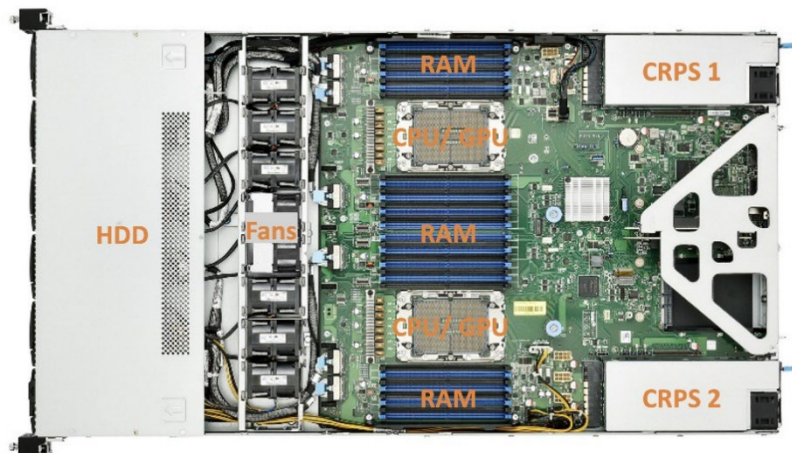
Featured in *Electronic Product Design & Test* – November 2024

The average data center requires 100,000 sq ft and has 10,000 servers, consuming 5 MW. An AI data center of the same size and number of servers requires 30 MW. In each server rack, the traditional power consumed is 5-10 kW /rack for regular non-AI processing and over 60 kW for AI computations. This is because next-generation AI GPUs such as NVIDIA's Blackwell B100 and B200 each demand over 1 kW of power for high-power computation, 3x higher than traditional CPUs. These new demands are driving power-per-rack specifications from 30-40 kW up to 100 kW.

With power-hungry GPUs, power delivery and cooling become critical aspects to these next generation AI server racks and ensuring scaling of these high-density data centers. Data Center Dynamics reports that 60% of companies are looking to improve their racks so that they can handle the increased heat that comes with more data.

A typical server rack consists of CPUs/GPUs, RAM, HDD, cooling fans, and common redundant power supplies (CRPS). There are typically two CRPS for each server rack, which allows for hot-swapping capabilities, so the server can remain online through power outages and electrical maintenance. Each CRPS is connected to a separate circuit so the server can continue working in the event of a trip. Power redundancy reduces downtime for businesses that rely heavily on their servers<sup>(8)</sup>.


The CRPS come in various form factors, which measure the same width and height (73.5 mm x 40 mm), but vary in length (CRPS185 = 185 mm, CRPS265 = 265 mm, OCP = 600 mm). Specifications are developed and defined by the Open Compute Project (OCP), whose members include Dell, Meta, Google, Intel and Microsoft. By having these common standards, the industry can maximize interoperability and therefore simplify upgrades and minimize downtime.



*Figure 4: Two CRPS provides back-up power ensuring uninterrupted electrical power, in case of a primary power source failure. As GPUs demand more power, the CRPS increases in power density*

*Featured in Electronic Product Design & Test – November 2024*

As GPUs demand up to 3x power, the CRPS must also provide increased power delivery. However, CRPS have standard form factors, therefore more power must be delivered in the same volume. This increase in power density becomes a significant challenge when using silicon power devices due to their performance limits. Additionally, efficiency regulations are in place to ensure highest efficiencies, resulting in minimizing data center energy use and reduction in CO<sub>2</sub> emissions. The 80 PLUS Titanium standard requires efficiencies above 96% at half-load and has been adopted by the European Union to align with its EcoDesign Directives, mandating AI data centers operating in the EU.



	Efficiency					
Load	80 Plus	Bronze	Silver	Gold	Platinum	Titanium
10%	-	-	-	-	-	90.00%
20%	80.00%	81.00%	85.00%	88.00%	90.00%	94.00%
50%	80.00%	85.00%	89.00%	92.00%	94.00%	96.00%
100%	80.00%	81.00%	85.00%	88.00%	91.00%	91.00%

*Figure 5: 80 PLUS Titanium standard requires that PSUs are 90% efficient at 10% of load, 96% at 50% and 91% at 100% of load with a 230 V input*

Silicon is the common semiconductor material used in power devices for today's PSUs. However, power density and efficiency requirements of next-generation, high-power-density PSUs for AI applications are exceeding silicon's limits. As a result, the industry is turning to wide bandgap materials such as gallium nitride (GaN) and silicon carbide (SiC) to meet power demands as effectively and as efficiently as possible. GaN and SiC FETs can switch at higher frequencies than legacy silicon devices. They also have lower internal resistance and are less affected as temperature rises. This makes them ideal for systems that focus on higher power densities and efficiencies.

*Featured in Electronic Product Design & Test – November 2024*



Transitioning systems from silicon to GaN and SiC FETs allow much higher switching frequencies which enable reductions in size, weight, and system cost of passive components such as inductors and capacitors. Due to the higher efficiency, less energy is wasted resulting in cheaper electricity costs. Also, the operating temperature of the devices reduce, allowing for smaller heatsinks and less fans. SiC is fast switching and has excellent conductivity properties which allows for high in-rush currents and power cycling. Navitas' latest Gen-3 Fast MOSFETs utilize 'trench-assisted planar' technology which delivers lowest  $R_{DS(ON)}$  shift versus temperature, enabling lowest conduction and switching losses at elevated temperatures. GaN FETs have the highest switching frequency capabilities, however, have relatively fragile gates, which are required to switch on and off the device. To maximize the switching frequency and to eliminate the gate sensitivity, Navitas monolithically-integrated a GaN gate driver circuit on the same chip as the GaN FET, named as GaNFast, GaNSense, and GaNSafe power ICs, each with tailored features and protection for specific applications such as CRPS for AI Datacenters.

Navitas announced its own [AI Power Roadmap](#) in March 2024, showcasing next-generation data center power solutions for the growing demand in AI and high-performance compute (HPC) systems. The first design was a GaNFast-based 3.2 kW AC-DC converter in the CRPS form factor, which enabled a 40% increase in power density size vs. the equivalent legacy silicon approach and easily exceeded the 80 Plus Titanium efficiency benchmark.



Figure 6: Navitas' latest 4.5 kW CRPS demonstrates how GaNSafe™ power ICs and GeneSiC Gen-3 'Fast' (G3F) SiC MOSFETs can enable the world's highest power density and efficiency solution

Featured in [Electronic Product Design & Test](#) – November 2024

The 4.5 kW AI power system reference design has a peak efficiency above 97% and, at 137 W/inch<sup>3</sup>, making it the world's highest power density AI PSU. At the heart of the system is an interleaved CCM totem-pole PFC topology using SiC G3F MOSFET G3F45MT60L (650 V 40 mΩ, TOLL). The 650 V G3F SiC MOSFETs have been optimized for the fastest switching speed, highest efficiency, and increased power density demanded by applications such as AI data center power supplies. The 'trench-assisted planar' technology provides high-speed, cool-running performance that ensures up to 25°C lower case temperatures and up to 3x longer life than alternative SiC products.

For the LLC stage, NV6515 (650 V, 35 mΩ, TOLL) GaNSafe™ Power ICs provide integrated power, protection, control, and drive in an easy-to-use, robust, thermally-optimized TOLL power package. These ICs offer extremely low switching losses, with a transient-voltage capability up to 800 V, and other high-speed advantages such as low gate charge ( $Q_g$ ), output capacitance ( $C_{oss}$ ), and no reverse-recovery loss ( $Q_{rr}$ ). As power density increases, next-gen GaN and SiC enable sustainability benefits, specifically CO<sub>2</sub> reductions due to system efficiency increases and 'dematerialization'.

## Summary

AI, which includes machine learning (ML), deep learning (DL), and generative AI, mimics human-like tasks, behaviors, and intelligence and is altering the way sectors like automotive, manufacturing, and healthcare conduct business. As a result, the AI market is booming with significant investments including Amazon investing over \$100B in AI-focused data centers over the next 10 years<sup>(9)</sup>.

The growth in AI is significantly driving-up data generation and processing, resulting in over 3x more power delivery for high-power density AI GPU servers. Silicon CRPS are not able to meet these higher power density requirements at 80 PLUS Titanium efficiency standards.

Next generation GaN and SiC power devices offer superior switching and performance compared to legacy silicon FETS, which enables the continuation of power density and efficiency in CRPS to support the growth of AI into the industry.

*Featured in Electronic Product Design & Test – November 2024*

## References

1. Source: Introduction to Generative AI
2. Boyle, P. (2024, March 28). Is it cancer? Artificial intelligence helps doctors get a clearer picture. AAMC. <https://www.aamc.org/news/it-cancer-artificial-intelligence-helps-doctors-get-clearer-picture>
3. Is it cancer? Artificial intelligence helps doctors get a clearer picture | AAMC
4. Duarte, F. (2023, March 16). Amount of Data Created Daily (2024). Exploding Topics. <https://explodingtopics.com/blog/data-generated-per-day>
5. AI is poised to drive 160% increase in data center power demand. (2024). <https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand>
6. AI brings soaring emissions for Google and Microsoft, a major contributor to climate change. (2024, July 12). NPR. <https://www.npr.org/2024/07/12/g-s1-9545/ai-brings-soaring-emissions-for-google-and-microsoft-a-major-contributor-to-climate-change>
7. How Data Centers Are Shaping the Future of Energy Consumption (2024). Jefferies. <https://insights.jefferies.com/the-big-picture/how-data-centers-are-shaping-the-future-of-energy-consumption>
8. Components of a Server (2024). Kirbtech. <https://kirbtech.com/components-of-a-server/>
9. Amazon to Invest \$100B in AI Data Centers Over Next Decade. CRE (2024). <https://www.credaily.com/briefs/amazon-to-invest-100-billion-usd-in-ai-data-centers-over-next-decade/>

*Featured in Electronic Product Design & Test – November 2024*