Revising the Power Infrastructure for AI Data Centers

Kishore Gagrani

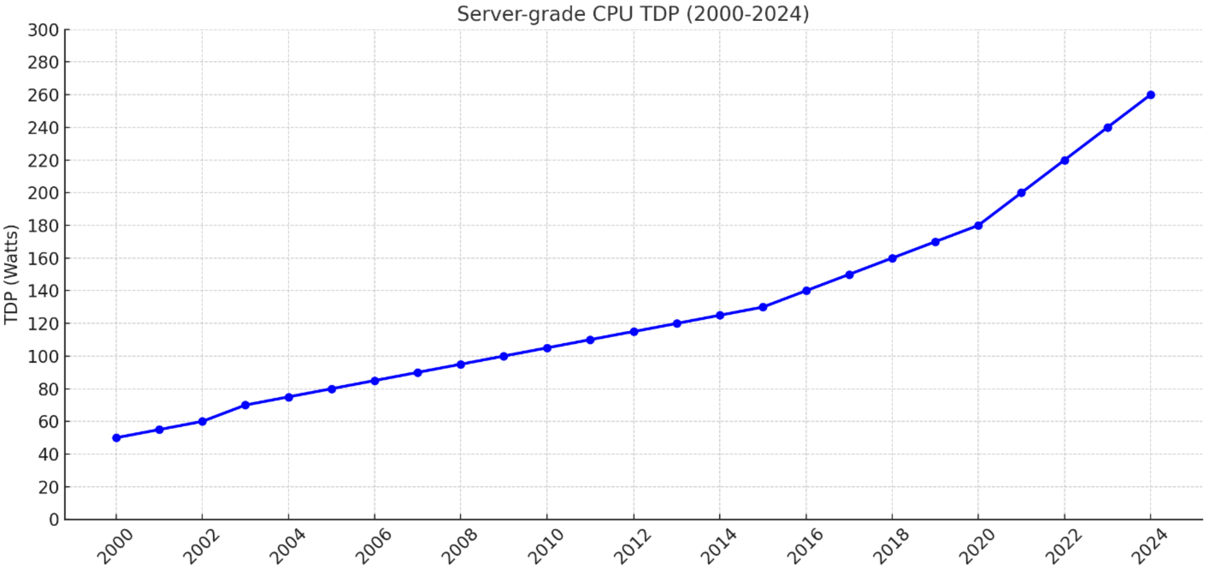
VP of Product Marketing @ Zonit

AI has placed significant pressure on data centers, with operators feeling the strain from growing demands in compute, memory, storage, and network bandwidth. This exponential growth has led to a corresponding increase in the need for power, not just in terms of wattage, but also in terms of resiliency and reliability. This paper addresses these three critical power-related challenges faced by AI-centric data centers.

The Power Trends Driving AI Data Centers

When we look at the latest power trends in the industry, it's clear that the demand for GPUs in AI workloads is the primary driver. The TDP (Thermal Design Power) requirements of both CPUs and GPUs have grown exponentially in recent years, reflecting the increased demands for performance, higher core counts, and the need to accelerate AI workloads.

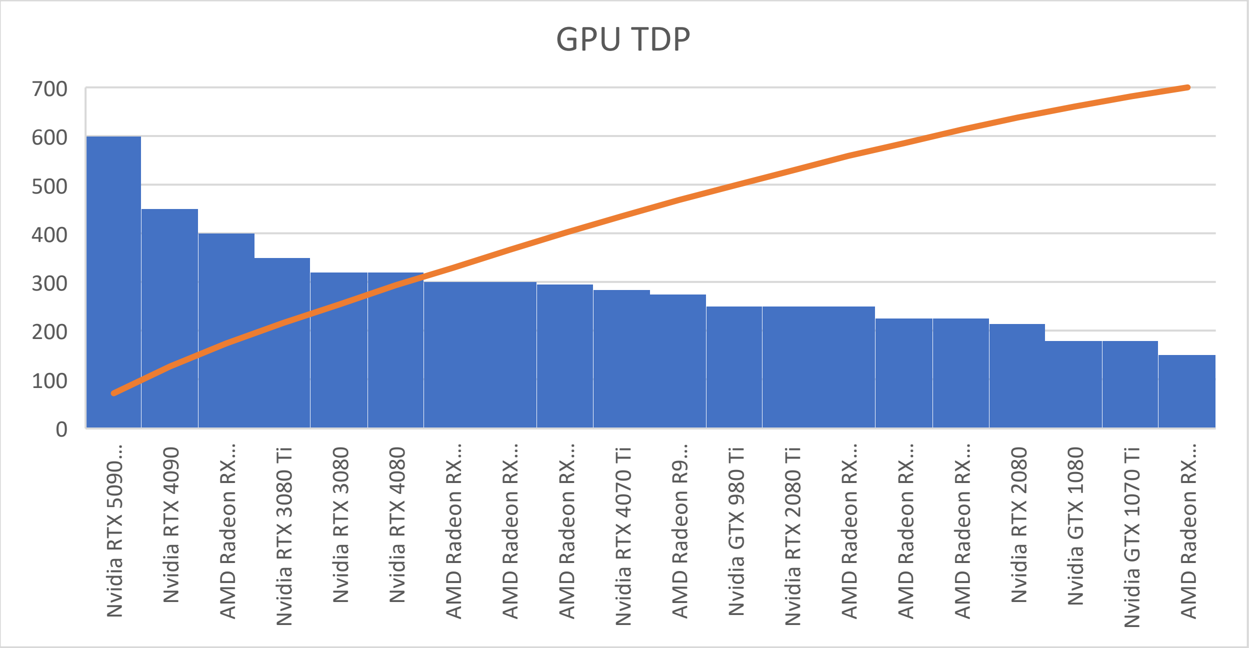
The chart below shows the trend of TDP for server-grade CPUs from 2000 to 2024, highlighting the increasing power requirements driven by performance improvements and the addition of more cores in server processors. Notably, the acceleration in TDP growth started around 2015, with a significant rise in power consumption for high-end models, pushing past 200W for many top-tier CPUs.



So, what happened since 2015

The growth of AI-centric use cases has driven the need for increasingly powerful hardware. As more transistors were packed into silicon to meet these demands, CPU vendors like Intel and AMD pushed the power envelope to deliver more performance. Intel, for instance, complemented these AI use cases by publishing data acceleration libraries, which only further fueled the demand for faster processing and larger-scale computations. However, this only created a higher need for GPUs—the **brains** of AI acceleration.

However, the rise in GPU TDP has been even more dramatic, throwing the power- the **heart** of AI datacenter, requirements completely off the charts.



What does this mean for Data Centers

In simple terms: Higher CPU and GPU TDPs lead to significantly increased power requirements. When multiple high-power CPUs and GPUs are packed into a single server, the total power needed to operate the server at optimal performance levels becomes substantial. If you scale that up to multiple servers within a rack, and then to entire data centers, the total power consumption becomes enormous.

For instance, Dell's XE9640, Lenovo ThinkSystem, and SuperMicro's H13 all support up to 8 AMD MI300X GPUs, with each GPU consuming 750W. When these servers are fully loaded with GPUs, the total TDP can reach up to 6000W per server. Let's put this into perspective:

-20 XE9640 servers in a rack would require around 12,000W of power to operate.

- If there are 500 such racks in a data center, the total power needed just for the servers is around 60,000,000W (or 60 MW).

And this figure doesn't even account for the power needed for networking switches, cooling systems, and other infrastructure components in the data center.

The Growing Demand for Redundant Power

AI workloads today demand continuous, real-time operation. Even a slight interruption in service is unacceptable. To meet the required uptime, a Tier 5 level of uptime is necessary, meaning 99.999% availability. To achieve this, data centers need redundant power sources for each of their racks, effectively doubling the power supply and UPS (Uninterruptible Power Supply) requirements to ensure no downtime.

Peer Reports and Industry Trends

So, what are data center operators saying about these challenges? According to a recent report from IDC, 766 qualified data center operators were asked about their priorities for 2024. Four of the top ten priorities directly relate to AI and its power requirements:

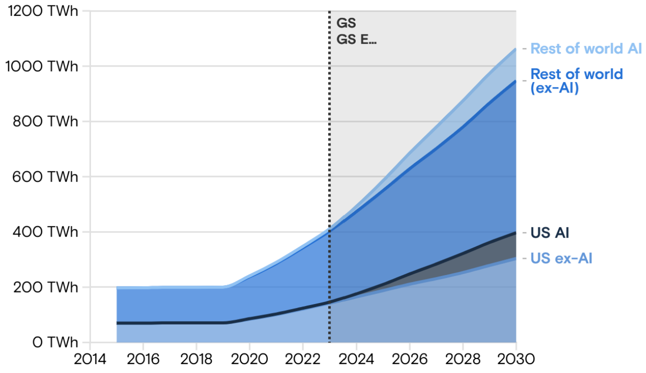
- Prepare Infrastructure for Gen AI
- Assure Resiliency and Availability
- Address Power Scarcity
- Improve Environmental Sustainability

Environmental Impact: A Growing Concern

One significant concern arising from AI's power demands is its environmental impact. According to IDC's analysis, by 2027, AI consumption could result in approximately 26 million tons of CO2 emissions—roughly 15% of total global PCF (Product Carbon Footprint).

The PCF calculation is indicative of growth in power demand, a recent publication by CISCO in collaboration with Goldman Sachs puts the growth to more than 1000 TWh by the year 2030.

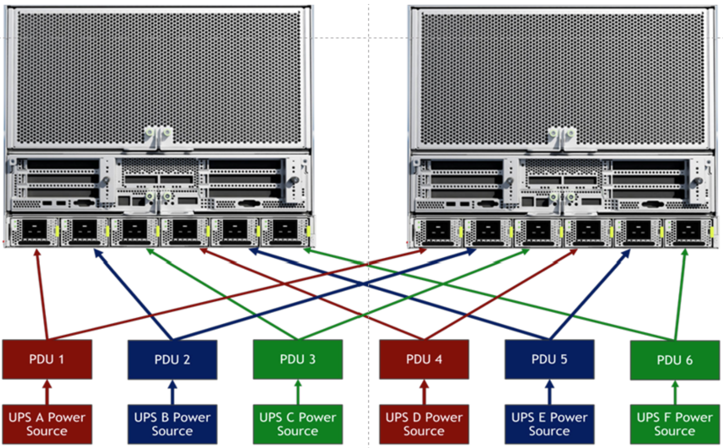
Data center power demand



Source: Masanet et al. (2020), Cisco, IEA, Goldman Sachs Research

Goldman Sachs

This growth report by the way doesn't even account for NVIDIA's AI Data Center power recommendation. NVDIA recommends redundant power infrastructure for the AI servers. An NVIDIA recommended power architecture is shown below.



Power Infrastructure recommendation from NVIDIA DGX B200 Spec

While this provides an excellent 2N, for the rack to have 100% redundant source of power redundancy but there are major challenges, the cost of power infrastructure that can run in millions, added operational power consumption cost due to higher PUE resulting from inefficient use of UPSs and increase in data center floor space.

The Two Main Issues: Resiliency and Cost

Where does all this lead us? The growing power requirements of AI-driven workloads present two key challenges for data centers:

1. Resilient and Redundant Power Sources

As AI workloads become more power-hungry, data centers need to ensure redundant power sources to guarantee continuous operation. Without proper backup and uninterruptible power solutions, downtime becomes inevitable, which leads to lost revenue and damage to reputation.

For the reference, here is a quick at a glance view of AI server vendors and their power supply redundancy capabilities

Server	Power supplies	Redundancy Matrix
Dell XE9680	6 2600 W Power Supplies	5+1 Redundancy
NVIDIA DGX H100	6 3300 W Power Supplies	4+2 Redundancy
NVIDIA DGX B200		5+1 Redundancy
ASUS ESC N8A-E12	6 3000 W Power Supplies	4+2 Redundancy

SuperMicro GPU SuperServer SYS-821GE-TNHR	6 3000 W Power Supplies	4+2 Redundancy
SuperMicro GPU SuperServer SYS-820GH-TNR2	6 3000 Power Supplies	4+2 Redundancy
Gigabyte G593 Series (ALL MODELS)	6 3000 W Power Supplies	4+2 Redundancy
Hypertec TITAN GS670R-G6	6 3000 W Power Supplies	4+2 Redundancy
HPE Cray XD670	6 Delta Titanium 3000 W Power Supplies	N+2 Redundancy
Small Form Factor Servers		
Gigabyte G493 Series (ALL MODELS)	4 3000 W Power Supplies	3+1 Redundancy
Dell XE8640	4 2800 W Power Supplies	3+1 Redundancy
ASA Computers 4U 8X GPU AI SERVER	4 3000 W Power Supplies	3+1 Redundancy

2. The Escalating Cost of Power

The cost of power is also rising at an alarming rate. As shown in the chart, a data center requiring 5MW of power can face an annual power cost of up to $1.9 million. The cost of power is now a major factor in the operational budget of AI-centric data centers.

So, what can be done

Fortunately, there are several infrastructural steps that AI data centers can take to minimize downtime, ensure availability, and reduce the costs of power consumption. These include:

- Optimizing Power Distribution: Using intelligent power distribution units (PDUs) to efficiently manage power and ensure backup systems are always ready.

- Energy-Efficient Cooling: Implementing advanced cooling technologies to reduce the overall power footprint. Even bringing better PUE for Liquid Cooling.

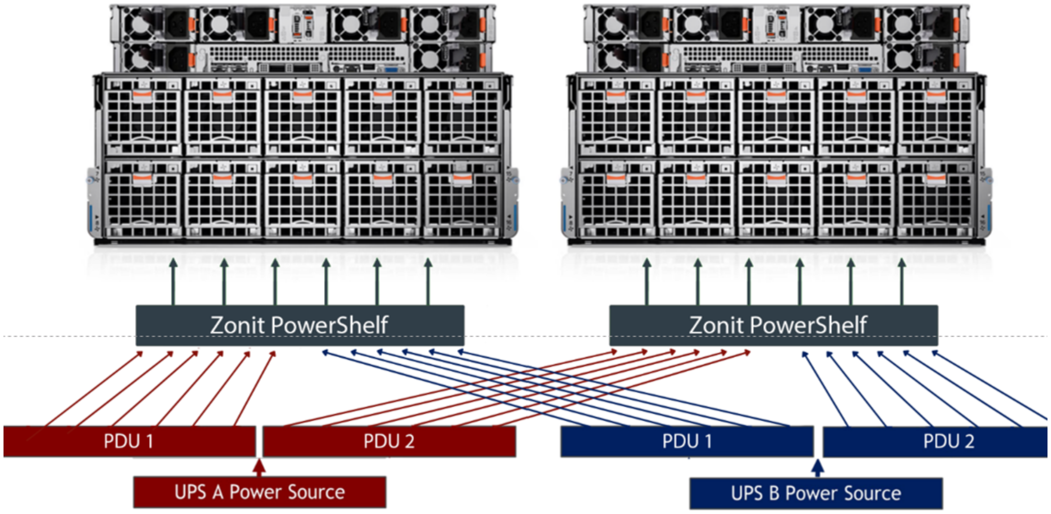
- Renewable Energy Sources: Incorporating solar, wind, or other renewable energy sources to offset the growing energy consumption and reduce carbon footprints.

- Power Usage Effectiveness (PUE) Optimization: Reducing power consumption in non-compute areas (like cooling and lighting) can have a significant impact on overall energy usage.

By taking these steps, AI data centers can help ensure that their services remain available, scalable, and sustainable in the face of ever-growing demands for compute power.

Investing in quality power cables, transfer switches, PowerTray, PowerShelf and PowerSwitch from Zonit is a great place to start.

For instance, the above NVIDIA's recommended power architecture can easily be transformed in to below newer architecture that has potential to increase POE, reduce infrastructure and power operating cost and yet provides a lot more rack space in the data center. A Win-Win approach.



Additionally, regular maintenance and testing of your power infrastructure can help identify potential issues before they become major problems.

Conclusion

AI workloads are driving a paradigm shift in data center power infrastructure. As the need for higher-performance GPUs and CPUs grows, the challenges of power consumption, resiliency, and cost become more critical. To meet the demands of the future, data centers will need to evolve their power infrastructure to ensure that AI workloads can be handled efficiently and sustainably—without compromising uptime or increasing costs. Let's ensure that as the power demands of AI rise, we rise to the occasion with smarter, greener, and more resilient power solutions.