VISHAY I NTERTECHNOLOGY, INC.

www.vishay.com

The DNA of tech.

AI Servers | White Paper

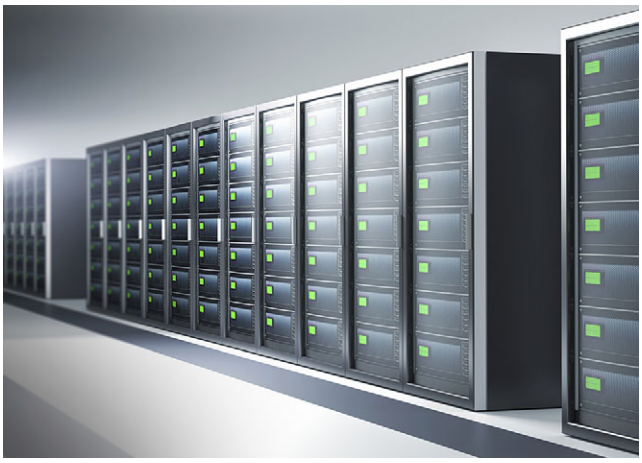
Managing Power Density, Thermal Limits, and High Speed Interconnects from Data Centers to the Edge

By Justin Chen and Yichang Yang

SUMMARY

As AI infrastructure pushes server power and power density higher, data center designers are increasingly constrained by heat, efficiency, and power integrity. This article outlines the key hardware challenges behind modern AI servers and the component technologies that address them – advanced power MOSFET approaches (including packaging aimed at better thermal performance), integrated power stage solutions that reduce parasitics, and supporting passives such as low ESR polymer capacitors and high performance inductors for fast transient response and high frequency conversion. It also touches on system needs around protection and robustness, backup power architectures, and high speed optical connectivity where signal integrity at very high frequencies becomes critical – framing how these building blocks help enable reliable, scalable AI platforms in the data center and beyond.

POWER DENSITY GROWTH DRIVES THERMAL, PROTECTION, AND MEASUREMENT REQUIREMENTS



Generative AI and increasingly large models are driving rapid growth in GPU cluster size, which in turn is raising server power demand and power density. Mainstream server output power has moved from roughly 3 kW (pre-2023) to about 6 kW, while volumetric power density has increased from about 50 W to 60 W/in^3 to more than 100 W/in^3. Roadmaps point to further increases to approximately 113 W/in^3 and potentially as high as 130 W/in^3, forcing the industry to rethink power delivery, cooling, and component selection.

As power density climbs, thermal management becomes a primary design constraint. More effective cooling reduces power losses and enables higher power density within the same footprint, creating a reinforcing cycle in which lower losses reduce heat, which in turn improves achievable efficiency.

In high power server power stages, MOSFET on-resistance ($R_{DS(on)}$) and cooling capability are central to overall efficiency – particularly on power distribution boards (PDBs) converting 48 V to 12 V, where designers must handle high currents (> 100 A) and then distribute low voltages down to point of load rails (from 12 V toward 1 V). Traditional PowerPAK® SO-8 or LFPAK package approaches primarily support single-sided cooling, and their higher thermal resistance can limit performance in space- and airflow-constrained environments.

To address this bottleneck, double-sided cooling MOSFET packaging can provide a more direct thermal path. The PowerPAK DC family, for example, uses top and bottom metal pads to conduct heat simultaneously into PCB copper and a cooling interface, with a reported $R_{\theta JC}$ of about 0.8 K/W – roughly 30 % to 50 % lower than conventional PowerPAK packages. In addition to supporting higher current and temperature rise, this architecture can align well with liquid-cooled server designs. Combined with the TrenchFET® Gen IV process, the result is reduced conduction and switching losses, which is particularly relevant for 3 kW to 10 kW AI servers and high current, high frequency PDB applications.

System robustness and signal integrity also become more critical as power density rises. Transient surge suppression devices, including TVS solutions such as the XClampR® family, are used to clamp fast transient events in dense AI power architectures.

WHITE PAPER

Managing Power Density, Thermal Limits, and High Speed Interconnects from Data Centers to the Edge

For GPU cards and edge AI accelerator boards, integrated power stages (DrMOS) can further improve power conversion by combining high side and low side MOSFETs with a driver IC in one package, reducing parasitic inductance, improving switching behavior, and enhancing transient response while simplifying PCB layouts.

Finally, accurate current and voltage measurement must keep pace with the move to higher currents and, in some proposed network architectures, higher bus voltages. Low resistance shunt solutions such as the WSK1216 – offering resistance values down to 200 μΩ with a 4-terminal construction – help minimize heat generation while enabling more accurate high current sensing in a compact footprint. For future high voltage measurement needs, divider networks such as the CDHV provide a single-chip solution rated up to 3000 V with resistance ratios from 250:1 to 500:1. Integrating both resistors on the same alumina substrate improves ratio tracking by ensuring both elements experience similar temperature changes, while also conserving PCB area.

ENSURING POWER SUPPLY STABILITY IN HIGH RELIABILITY AI SERVERS

AI servers place unusually high demands on power stability under high speed computing conditions. At switching frequencies above 500 kHz, core compute devices – GPUs, CPUs, and ASICs – can draw sharply changing currents on millisecond and even microsecond time scales. These fast transients can create significant voltage and current ripple, directly impacting system stability and conversion efficiency. Low ESR capacitors with high capacitance density in compact packages are therefore essential for absorbing ripple and supporting rapid transient response. Polymer tantalum capacitors such as vPolyTan™ devices are designed to help stabilize the supply rail during load steps, reducing noise susceptibility while maintaining voltage regulation.

Inductor performance is equally critical, particularly as AI server power stages push into the 500 kHz to 1 MHz range. Designers must balance high magnetic energy density with small form factors, which makes molded inductors a common choice. In these applications, inductors must handle high voltage and high current simultaneously while keeping losses low. Core material selection is pivotal: it must support high frequency switching with minimal core loss, while also offering high magnetic flux density and temperature tolerance to prevent saturation during peak current events – conditions that can otherwise lead to voltage droop, instability, or downstream MOSFET stress and damage.

Molded powdered-iron inductors such as the IHLP® series are engineered to address these requirements by combining compact sizes with shielding characteristics that help suppress flux leakage and reduce EMI. Low DCR supports efficiency, while magnetic material design aims to preserve inductance stability across switching frequencies from hundreds of kHz into the MHz range. Higher saturation current (I_{sat}) further supports stability under peak loads, aligning well with dense power architectures in AI servers. For additional EMI mitigation, the IHLE® variant adds an integrated e-shield to reduce radiated emissions that can interfere with nearby densely packed circuitry. For emerging multiphase topologies such as TLVR, newer inductor options – including the IHTL and IHSR series – target very high frequency multiphase converters using inductances below 200 nH, with lower DCR intended to further improve efficiency and power density.

Resistor technologies also play a key role in long term reliability, especially as next-generation AI data centers adopt harsher operating conditions, including higher temperatures and liquid-cooling approaches that may use specialized chemical fluids. Sulfur-resistant thick film options such as the RCA family (including the RCA, RCA-HP, RCA-LS, RCA-SR, and CRCW-SR) are designed to reduce the risk of sulfur-driven resistance drift or open circuits over time. Where higher precision is required, thin film solutions such as the TNPW series provide sulfur resistance alongside tighter tolerance (down to 0.1 %) and lower TCR (down to 10 ppm).

Because the cost of core motherboard components – particularly CPUs and GPUs – is high, system builders increasingly prioritize reliability and service life. While traditional servers are often depreciated over roughly 4 to 5 years, AI training systems are commonly expected to remain in service for 5 to 10 years. In that environment, component strategies that emphasize efficiency, thermal management, and long term stability help support extended platform lifecycles and provide a more resilient foundation for scaling AI compute workloads.

Managing Power Density, Thermal Limits, and High Speed Interconnects from Data Centers to the Edge

STRENGTHENING BACKUP POWER AND OPTICAL INTERCONNECTS IN AI DATA CENTERS

Backup power architectures in data centers are evolving to better accommodate the fast, high amplitude load transients associated with AI servers. In addition to the broad use of lithium-ion battery backup units (BBUs), future designs are expected to add power conversion systems (PCS) alongside BBUs to respond more effectively to instantaneous load demands. Energy storage capacitors with high energy density, low self-discharge, and long service life can support these architectures by buffering short duration events and improving overall stability. ENYCAP™ energy storage capacitors, including EDLC solutions such as the 235 EDLC-HVR, are positioned for use in BBU designs, with the added benefit of low maintenance requirements that can help reduce operating overhead.

Complementary signal chain components – such as optocouplers and isolation amplifiers – can further improve voltage and current sensing within BBU systems, supporting more robust monitoring and intelligent management.

In parallel, optical connectivity is advancing rapidly as AI training and inference require massive data exchange among GPUs. Optical modules are moving from 400 Gbps to 800 Gbps and trending toward 1.6 Tbps, with single-channel data rates reaching 200 Gbps. At these speeds, signal integrity margins shrink dramatically: above roughly 50 GHz, small impedance discontinuities can increase bit error rates. High frequency thin film resistors can help maintain impedance control at RF bandwidths. CH Series thin film resistors, for example, are specified to preserve near-ideal resistive behavior from 10 Ω to 500 Ω across very wide bandwidths – up to 70 GHz – supporting stable performance in high speed optical modules.

Power protection remains a parallel priority across these subsystems. Advanced eFuse solutions provide fast, precise protection against overcurrent, overvoltage, and short-circuit events with accurate current limiting. Compared with traditional fuses or discrete protection approaches, eFuses can simplify implementation, reduce board space requirements, and improve safety and serviceability. Together, RF-capable thin film resistors and integrated protection devices help enable the stable, high speed, and robust operation required across both backup power and optical communication modules in next-generation AI data center infrastructure.

EXTENDING AI INFRASTRUCTURE ADVANCES TO AI PCs AND THE EDGE

The design pressures reshaping AI data centers – higher power density, tighter thermal margins, faster load transients, and more demanding signal integrity – are now influencing the next wave of AI platforms. As AI PCs and workstations gain momentum and edge AI deployments expand, many of the same fundamentals apply: power delivery must remain stable under dynamic compute loads, thermal performance must be engineered into increasingly compact form factors, and high speed interconnects must preserve signal integrity as bandwidth continues to rise.

Meeting these requirements favors component choices that reduce losses, improve cooling effectiveness, and sustain performance over long operating lifetimes. High performance MOSFETs that support improved thermal paths, low ESR capacitors that strengthen transient response, inductors engineered for high frequency conversion with low DCR, and resistors capable of maintaining predictable behavior at very high frequencies all contribute to the same objective: enabling more computing capability within constrained space, power, and cooling envelopes.

At the same time, the most effective designs treat these elements as part of an integrated system rather than isolated parts. Approaching AI platforms from a whole-system perspective – spanning power conversion, protection, sensing, backup power, and high speed communication – helps designers balance efficiency, thermal management, robustness, and signal integrity in a coordinated way. As AI continues to expand beyond hyperscale environments, the component-level strategies that address today's data center challenges provide a practical foundation for scaling performance and reliability across AI PCs, workstations, and edge computing systems.