Powering Generative AI Platforms with TCO at the Core

The datacenter of the future – the AI factory – runs on thousands of AI accelerators, each of which consume over 1000 A. Creating a green AI factory starts with designing an efficient voltage regulator (VR) to provide such high currents to these AI accelerators with high power density and efficiency.

By Davood Yazdani, Senior Director Product Marketing, and Paul Yeaman, Director Power Module Design Engineering, Infineon Technologies

Artificial intelligence is driving exponential growth in global data generation, especially with the recent boom of generative AI queries. The energy demand of the chips supporting this enormous data growth have exponentially increased, with each chip approaching >1000 W of thermal design power. Datacenter utility power is reaching its limits as AI servers demand three times higher energy than traditional servers.

Datacenters consume more than two percent of global energy and will consume more. Power solutions and architecture innovations are critical for a measurable impact on global energy savings and better total cost of ownerships (TCOs) for datacenters.

Evolution of processor power requirements

Traditional servers have a processor (<200 A thermal design current, or TDC), coprocessor (<30 A), and memory DIMM (<40 A). The processor is in the board's center, with one side for power and three sides for signals and communication (to the memory and coprocessor).

As computers evolved and high-powered AI-compute applications emerged, CPU/GPU/FPGA vendors and AI start-ups designed faster, powerful ASIC and AI chips. Core-rail current levels doubled to 400 A TDC in 2016, and again four years later to 800 A TDC.

When core rail current increased beyond 200 A, single-sided entry became impractical due to excessive power distribution network (PDN) losses and vendors standardized double-sided entry. It halved the PDN resistance, but PDN losses still grew due to I2R. Vendors integrated the coprocessor and memory with the core processor to increase speed and performance. The coprocessor, instead of being a separate entity, was now located on the core silicon, and high-bandwidth memory (HBM) replaced memory DIMMs. All the power driving the AI processor was physically consolidated into the motherboard's main area. As power increased, the heat was concentrated in a small area.

Processors became larger as well. The industry standard form factor for AI applications is the OCP Accelerator Module (OAM), with typical AI server motherboards accommodating ≤8 OAM modules. An OAM's size (170 × 102 mm) became insufficient for larger AI chip sizes and total power requirements. New high-current AI chips were as large as 110 × 80 mm, leaving limited space for power components due to two large mezzanine bottom connectors. Cooling system design became challenging due to high PCB losses, higher heat density, and component height restrictions.

The challenge

With new AI systems requiring >1000 A concentrated within a small area at the OAM's center, double-sided entry led to multi-sided entry (figure 1), where the resistance scales down by 1/N, N being the number of sides with power. However, this reaches a ceiling, as there are only four sides to a processor and a limited perimeter from which to source current easily.

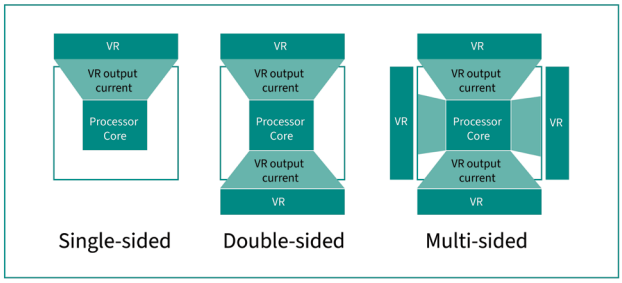

Figure 1: Different methods of voltage regulators (VR) supplying current to the processor core. Each additional side dedicated to supplying power results in more of a challenge for routing signals.

Arranging phases in multiple rows increases PDN resistance as high currents sourced from the back would be routed around the forward stages. Additionally, signal integrity becomes an issue with four-sided entry due to limited places to route sensitive signals while avoiding power planes, which tend to be noisy.

A >1000 A design requires a solution addressing three obstacles:

- increasing power density allowing more current to be sourced from a small space
- increased power integrity allowing power to coexist near high-speed signals
- a low profile enabling power to be sourced from underneath the processor

Modules

Dual-phase power modules incorporate the inductor, discrete capacitors, and power stages of two phases of a multiphase buck regulator onto a substrate to create a single device, deployed in an array as a multiphase system. By this integration of the voltage regulator (VR), a power module decreases a multiphase solution's footprint by 40 percent compared to an equivalent discrete solution.

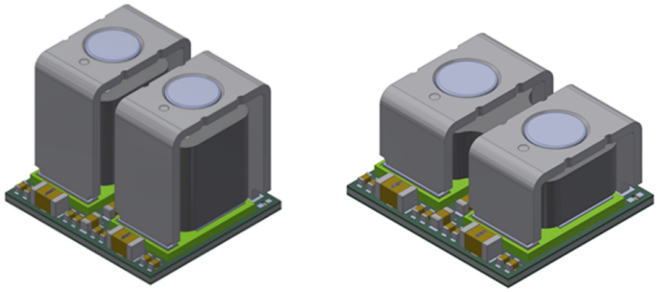

Figure 2: Infineon's dual-phase modules TDM22544D (8 mm tall) and TDM22545D (5 mm tall).

When integrating the inductor and power stage, what should be located on the surface closest to the heatsink?

Since the power stage causes most of the power loss, locating it on top, closest to the heat sink, minimizes the heat conduction path and improves thermal performance. However, a shorter heat conduction path is offset by a longer electrical conduction path, particularly on the input side, leading to additional losses and lower efficiency.

A better design keeps the power stage on the motherboard side of the module, minimizing both conduction and parasitic losses. In this configuration, the inductor is located between the power stage and the heatsink, but still able to effectively cool the power stage by thermally coupling it to the heatsink.

Infineon's dual-phase modules (figure 2) utilize a proprietary inductor-on-top design for improved thermal performance with better efficiency (figure 3). This design maximizes heat-conduction to the top surface from the key heat-generating areas of the power stage below it, through the inductor.

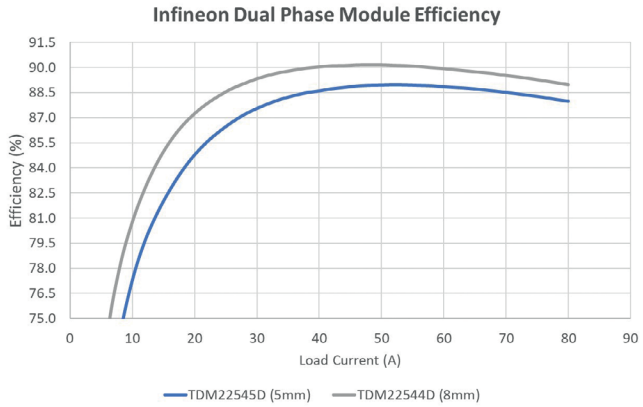

Figure 3: Infineon dual-phase module efficiency at 13.5 V input, 0.8 V output, 800 kHz.

The power module also adds a vital electrical component to the VR: the substrate. The substrate isolates the switch-node from the motherboard, rendering all input and output currents into the module at a constant DC voltage. This improves signal integrity by eliminating switching voltages from the motherboard and protecting sensitive signals from noise-coupling.

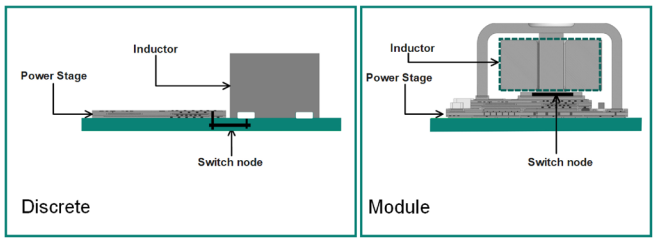

Figure 4: Implementing the switch-node inside the module reduces impedance as well as parasitics that cause noise and limit performance.

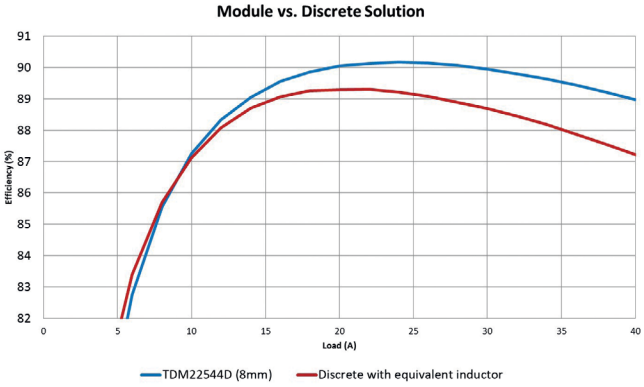

Figure 5: Efficiency comparison between a single phase of a module and a single phase of a discrete solution at 13.5 V input, 0.8 V output, 800 kHz.

At higher switching frequencies, switch-node parasitics limit efficient switching of the VR. Integrating the switch-node into the module (figure 4) minimizes its parasitic impedances for a more efficient operation at higher switching frequencies (figure 5). High switching frequency VRs use smaller inductors and less output capacitance, increasing power density.

Once the switch-node is off the motherboard, the module can now be located under the processor. This minimizes PDN losses as the high current flows vertically through the thickness of the motherboard, instead of laterally traversing the package to the core. This reduces the current path by ~30 times, significantly reducing power losses.

PDN study

Dual-phase power modules are critical to high-power GPU systems due to effective power density and signal integrity management without reducing system efficiency. Locating modules closest to the processor and utilizing all the available mounting area, achieving >2000 A becomes feasible by reducing PDN losses.

Assume a 2000 A OAM with 90 µΩ resistance from the north and south sides of the processor core. Sourcing 100 percent of the current from both sides yields 180 W of PDN loss, or >10 percent of the total processor loss (assuming 0.8 V core voltage). Sourcing 60 percent of that current from underneath the processor at 18 µΩ, PDN losses drop by 70 percent to 50 W, or ~3.1 percent of the total processor loss.

Since typical large-scale deployments consist of ~100,000 processors, saving 130 W/processor translates into megawatts for a datacenter. This means millions of dollars saved over the system's lifetime.

Improving power efficiency at the core yields significant energy savings. Leveraging decarbonization-based technology improves each power conversion stage starting with AC power entry. Service providers can leverage Infineon's XDP™ controllers, dual-phase power modules, combine them with AI capabilities in datacenters, to deliver superior power-conversion efficiency, enhanced flexibility and energy-efficiency in system-design, and the best TCO for datacenters.

Conclusion

Increasing power density without compromising on efficiency is key to presenting a path to an efficient and green 2000 A GPU system and beyond. Modules enable this path and open new avenues for being able to further integrate the multiphase VR while also moving it closer to the processor. Infineon's patented inductor technology carries both output current and heat from the module and enables best-in-class system performance with simpler construction. Combining Infineon's Trench device technology with this patented magnetic technology paves the way for true Vertical Power Delivery (VPD) architectures that enable AI processors to achieve the best possible performance and lowest TCO.

www.infineon.com