Electronic Components Challenges in AI Power Management

Tomas Zednicek
president, EPCI European Passive Components Institute
tom@passive-components.eu

Abstract

The rise of artificial intelligence (AI) has brought about a technological revolution, with its impact felt across various industries. However, this advancement comes with a significant energy cost, particularly within data centers that power AI operations.

A decade and a half ago, server CPUs and GPUs operated on a few hundred watts of power. Today, the Nvidia H100 GPU, used for training large language models, operates at a thermal design power (TDP) of 700W, and future processors are expected to exceed 1kW. This starkly contrasts the human brain's energy efficiency of approximately 20W.

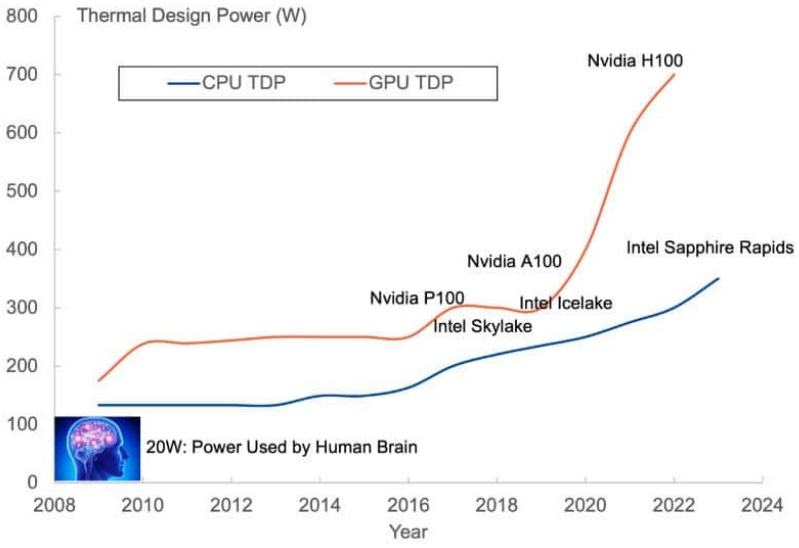

Figure 1.: Computer power requirements; source [1]

The impact of millions of users on data centers' energy usage is substantial with AI deployment expected to increase energy consumption, potentially exceeding 4% of global energy by the end of the decade, raising concerns about power generation becoming a bottleneck in AI advancement.

The Evolution of Power Management in Data Centers

Power management is a critical concern in data centers that directly impacts thermal management and overall efficiency. One of the key advancements is the development of new power supply topologies. These innovative designs are crucial for enhancing power delivery and reducing waste. The transition from traditional 12V to 48V power distribution allows for a substantial reduction in power losses.

GaN and SiC wide band gap-based power switches can withstand higher temperatures, voltages, and frequencies compared to conventional silicon semiconductors. This capability translates into reduced cooling requirements and the potential for smaller passive components, which can be positioned closer to the processor to cut down losses in circuit board wiring.

Furthermore, addressing 'parasitic losses'—the power dissipated as heat due to the Equivalent Series Resistance (ESR) of capacitors, and losses in inductors' windings and magnetic

cores—is essential. By focusing on reducing these losses, data centers can conserve energy and lower cooling demands.

By moving power supply components, such as Power Management Integrated Circuits (PMICs) and passives, closer to the processor chip, engineers can significantly reduce resistive and inductive losses. This not only curtails parasitic power losses but also enhances voltage control during high current load transients, which are common in high-performance processors.

Capacitors for AI Power Management

48V rails are becoming increasingly common in data centers and automotive applications. This shift necessitates the use of capacitors that can handle higher voltages. Class I “stable dielectric” MLCC ceramic capacitors are now available with a 50V rating, while Aluminum and Tantalum polymer capacitors have been developed to withstand 63/75V. These enhancements allow them to be used in various converter topologies, catering to specific capacitance and ESR requirements.

A notable innovation in this field is the advancement in energy density of class I MLCC dielectrics. For example, U2J dielectric [2] is offering approximately 2.5 times the capacitance of the C0G, along with stable temperature and voltage characteristics. These capacitors are particularly beneficial in power applications where high ripple current handling and stability are crucial.

The use of Tantalum and Aluminum Polymer capacitors, alongside traditional wet Aluminum Electrolytic capacitors, in bulk decoupling and filtering applications, demonstrates the versatility and adaptability of these components. With the availability of 63 and 75V ratings, they are well-suited to meet the demands of modern 48V voltage rails.

In the continuous operation of data centers, where CPUs and GPUs often operate at temperatures close to 90°C, the reliability of components is paramount. A single failure can result in significant financial losses, particularly during intensive tasks such as AI training runs.

Electronic components must be designed for long operational lifetimes, withstanding up to 125°C, along with enhanced humidity resistance.

Inductors for AI Power Management

Inductors, play a key role in high-efficiency power supplies found in most modern data centers. They come in various form factors and core materials, each with its own set of advantages. Metal composite inductors, for example, can handle high saturation currents and offer a gradual inductance roll-off with increasing current. However, they can suffer from core losses within the metal material. On the other hand, single-turn ferrite core inductors, or Power Beads, boast very low losses, making them suitable for high-power applications. These inductors do have a 'hard' saturation point, where inductance drops sharply beyond the saturation current limit. Its range from 47 nH to 230 nH and are designed for use in point-of-load power converters near processors to minimize board resistive losses. They can handle up to 53 A and have a maximum DC resistance rating of just 0.32 mOhms, minimizing losses and heat dissipation.

An innovative approach to power conversion is the Trans-Inductor Voltage Regulator (TLVR), which utilizes a dual-winding inductor. This design is gaining traction in high-current, high-efficiency multi-phase converters, showcasing the continuous evolution of power components in data centers.

The push for smaller, more efficient power converters has led to the development of inductors with higher saturation currents and lower core losses, capable of operating at high switching frequencies without overheating.

While the inductance of a ferrite inductor will vary with age, the metal composite core material is free from any effects of aging and it can be considered the best choice, to guarantee a system that functions over its entire specified lifetime at elevated temperatures environment.

Compared to ferrite inductors, metal composite inductors have a much higher energy density, which leads to a size reduction of 30% – 50% for comparable current specifications that serve the trend for downsizing high-current power circuits.

The Future of Power Delivery in AI Applications

As artificial intelligence (AI) applications become more advanced, the power delivery solutions required to support them are undergoing significant evolution. The next generation of Graphics Processing Units (GPUs) is expected to consume over 1000W per processor, necessitating the supply of multiple kilowatts of power per rack. This substantial power consumption brings forth the challenge of heat dissipation, which traditional air cooling methods may not be able to manage effectively.

Another trend is the miniaturization of power conversion components, moving them closer to, or directly onto, the chip. This shift is particularly important for inductors, which are typically the tallest components on a circuit board. The drive towards smaller components has led to the development of more compact modules, some of which are being integrated into complete System-in-Package (SiP) modules. The future may see components directly attached to the chip or embedded within the circuit board, further reducing the spatial footprint and potentially improving power efficiency.

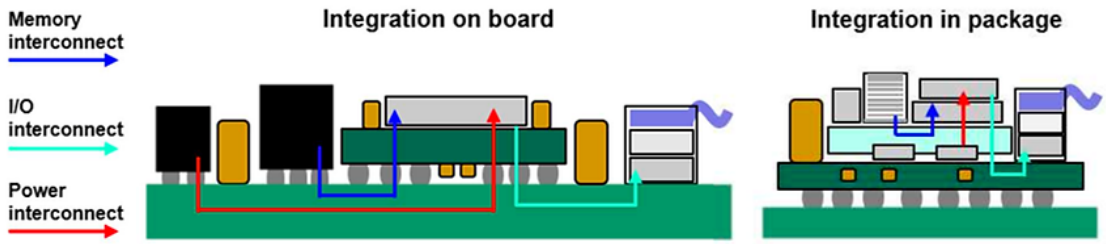

Figure 2. Component migration SiP downsizing trend; source: TSMC

Semiconductor technology is also advancing at a rapid pace. Backside power delivery is emerging as a key feature in the roadmaps of major industry players for next-generation nodes. This development could significantly impact overall power delivery schemes. The progression from 2.5D to 3D packaging continues, with chiplets becoming increasingly common. Additionally, glass interposers are emerging as a potential alternative or complement to traditional silicon interposer technology, offering new possibilities in semiconductor design and functionality.

These advancements in power delivery and semiconductor technology are critical for supporting the growing demands of AI applications. They not only promise to enhance the performance and efficiency of AI processors but also pose new challenges and opportunities for innovation in thermal management and component integration.

References

[1] P.Lessner.,“Artificial Intelligence--Why Now?“,https://passive-components.eu/phil-lessner-yageo-cto-on-artificial-intelligence-why-now/

[2] D. Patel, et al., “,” retrieved from: AI Datacenter Energy Dilemma – Race for AI Datacenter Space March 13, 2024.

[3] AVNET blog, „AI is impacting passive and interconnect design“., https://passive-components.eu/understanding-ai-passive-components-and-interconnects/