High-Density Power for the AI Revolution

By Charles Bailley, Senior Director Global Business Development, Navitas Semiconductor

Safe, powerful, integrated GaN power holds key to efficiency and sustainability

Since it burst into the mainstream with the public availability of ChatGPT and Google Bard, Artificial Intelligence (AI) has dominated the technology news landscape. Indeed, such is the level of 'noise' around AI, even the least technical consumer understands what it is and how it might benefit or impact them.

AI in the industrial and consumer spaces is dominated by powerful, purpose-designed processors from large, established companies like Nvidia and start-ups and scale-ups who are helping to fuel innovation. These increasingly powerful 'superchip' platforms are essential for processing huge volumes of data at speeds that enable AI to fulfill its potential. But implementing these technologies presents challenges. As AI proliferates, the extreme power demanded by processors such as NVIDIA's Grace Hopper H100 super-chip require a two- or three-fold increase - from around 30 to 40 kW per cabinet in current servers to 100 kW or more.

There are physical considerations too. The space afforded to power, processing, and server systems cannot expand without limit, making power density a priority for system architects.

Performance, Efficiency and Size

The carbon footprint of data centers is already significant thanks to the amount of energy needed to power and cool servers that provide the cloud-based services we rely on. The International Energy Agency (IEA) estimates global data center electricity consumption in 2022 at 240-340 TWh, or between one and 1.3% of total global electricity demand. The growth of AI and the increasing amount of data being generated, shared and stored by everything from smartphone apps to connected vehicles will further increase this demand. A spotlight has been thrown on the sector to do more to minimize energy use and maximize efficiency.

While operating expense pressures mean there is a natural desire in industry to improve matters, power supply performance and efficiency targets are also driven by regulations, standards and frameworks. For example, the latest European regulations expect server power supplies to exceed the 80 PLUS 'Titanium' efficiency specification. Certification to this most demanding of the 80 PLUS family of regulations requires guaranteed power supply efficiency between 90% and 96% at loadings from 10% to 100%. Compliance also requires a power factor of at least 0.95 at lower levels - a benchmark that necessitates active power-factor correction.

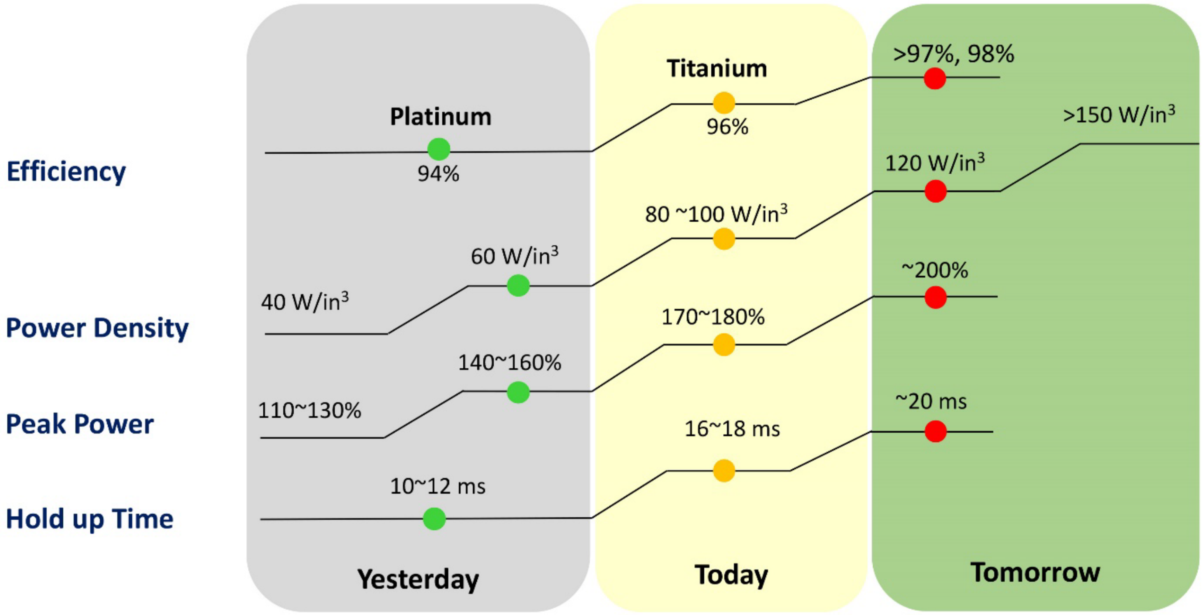

Figure 1: The 80 PLUS certifications for power supplies specify required efficiency levels and other key performance specifications

With so many server farms installed worldwide, physical server rack-size constraints are well-established and hard to change. As a result, the processing power needs of AI now and in the future must be implemented within existing space constraints, with thermal management arrangements compatible to the cooling requirements of legacy (lower) power densities. This means the greater processing and power densities associated with running AI chips must be achieved without creating heat that may be hard to manage without additional, power-hungry colling technologies.

Power Supply Topologies and Design

The tough demands of the enterprise and hyperscale markets have driven the formation of the Open Compute Project. This respected community (including big names like Facebook, Intel, Google, Microsoft, and Dell) seeks to define the Common Redundant Power Supply (CRPS) specification to ensure modularity for easy replacement and maintenance, and flexibility to address different and emerging applications for IT ecosystems. As well as redundancy, CRPS mandates that rack power must continue to be provided in a 1U (40 mm) x 73.5 x 185 mm form factor.

With server power demand set to increase towards 100 kW, pressure to meet 80 PLUS 'Titanium' specifications and requirements to comply with the OCP-defined CRPS specification, the challenge for power designers is significant. A closer look at a typical compliant power supply provides some insights into the technology evolution that is needed.

The most commonly used topology for data center server power supplies employs silicon MOSFETs and has a boost power factor correction (PFC) stage followed by an LLC resonant converter. Increasing power demands associated with AI workloads requires multiple outputs when using LLC transformers to minimize conduction losses of secondary windings and synchronous rectifiers (SRs). Combining windings and SRs in parallel can achieve lower winding resistance but the more pronounced termination losses this causes (as switching frequency increases) negatively impacts overall system efficiency.

Using multiple transformers in parallel instead of windings and SRs can help by lowering AC-related conduction losses and leakage flux resulting in reduced leakage inductance. This approach can also simplify terminal design. The flipside is that multiple transformers introduce larger core losses and a larger magnetic size. Deploying matrix transformers can be a way to mitigate these losses, but using traditional wire-based transformers in any form means the physical size of the converter becomes an issue in an operating environment where power is increasing but space is not.

PCB-based planar transformers can provide a space-saving and cost-effective answer to the challenges described, but due to the limitation on the number of PCB windings possible within the circuit board structure, they can only be used when the switching frequency is high enough to support a minimum number of turns.

A fundamental issue is that silicon has reached its performance limits in terms of suitability for many new and emerging applications. In high-density CRPS applications, swapping out silicon MOSFETs with GaN devices allows much higher switching frequencies and supports the use of planar transformers. GaN technology also helps designers to keep switching losses within the scope of 80 PLUS Titanium specifications, as well as supporting improved robustness, higher power densities.

However, it is important to choose the right type of GaN device. Discrete GaN FETs, for example, have a relatively fragile gate which requires hard to achieve control and solutions in which they are used can experience problematic high-side and low-side ringing and shoot-through currents. A much better - and simpler - alternative can be realized by selecting integrated GaN devices.

Integrated GaN

Monolithically-integrating a GaN gate driver circuitry on the same chip as the GaN FET avoids the issues of using discrete GaN devices, helping designers properly control gate voltage as well as keeping component count down. But the latest integration levels go far beyond that to additionally incorporate control, sensing and protection. These next-generation integrated GaN platforms are setting new industry benchmarks in efficiency, density and reliability for demanding applications.

Take, for example, the Navitas GaNSafe™ family.

Conceived specifically to meet the needs of AI-based data centers and other high-power, fast growth applications such as EVs, solar and battery energy storage systems (BES), GaNSafe provides the switching speed, efficiency and power density that addresses the challenges of 80 PLUS Titanium and the framework and IT ecosystem defined by the OCP. Supplied in cool-running, surface-mount TOLL packages, these ICs have in-built algorithms for dead-time control and a suite of additional safety features needed for high-power applications.

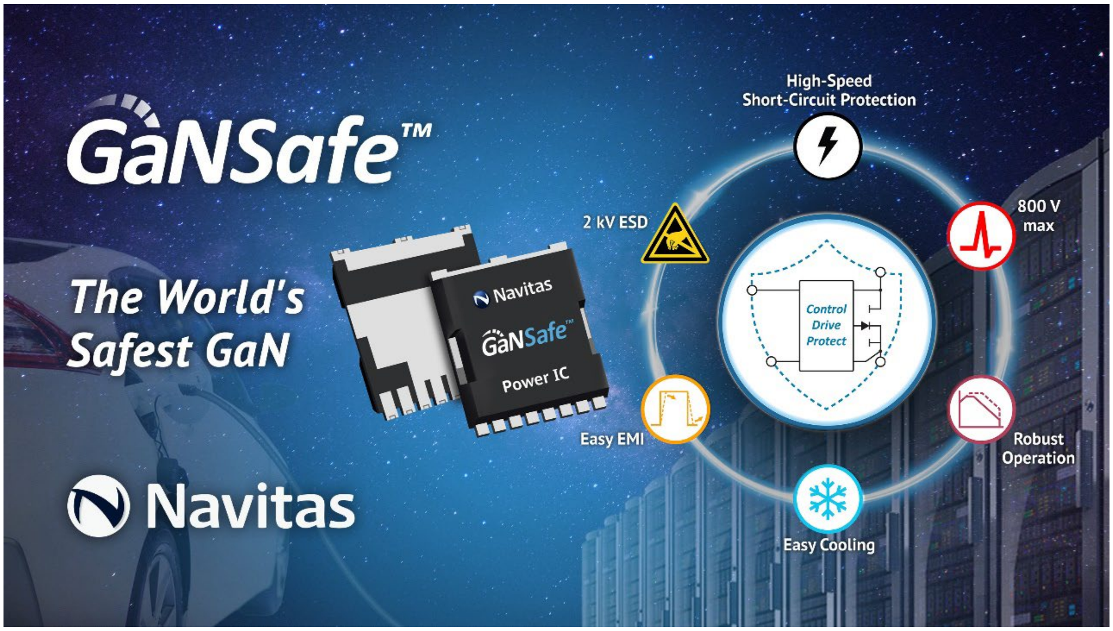

Figure 2: GaNSafe Power ICs achieve the power densities required to support AI applications and help designers with a range of important protection features

Certainly, it's hard to overstate the importance of protection, safety and robustness for devices deployed in data center servers, where issues associated with device, module or system failure can be costly or even dangerous, with a major impact on the infrastructure they support. That's why GaNSafe has high-speed protection built-in, including with autonomous 'detect and protect' that acts within 50 ns. ESD protection - not normally part of the specification of discrete GaN transistors - is also included to protect against events up to 2 kV. Further protection against extraordinary application conditions comes from a 650 V continuous and 800 V transient voltage rating. In addition, programmable turn-on and turn-off speeds ease the job for designers trying to meet EMI regulatory requirements.

GaN in Action – Application Example

An AI data center server power supply built using devices such as GaNSafe can achieve significantly better performance and support enhanced system safety and reliability versus a unit that utilizes discrete silicon or silicon carbide (SiC) devices.

To demonstrate potential performance improvements, Navitas has created a reference design for a 54 V AC-DC data center AI/GPU server power supply in a CRPS185 format using Navitas' GaNSafe and GeneSiC technologies. This design demonstrates that power density can be increased from less than 100 W/in^3 to 138 W/in^3, which translates to a power output improvement from 3.2 kW to 4.5 kW (Figure 3).

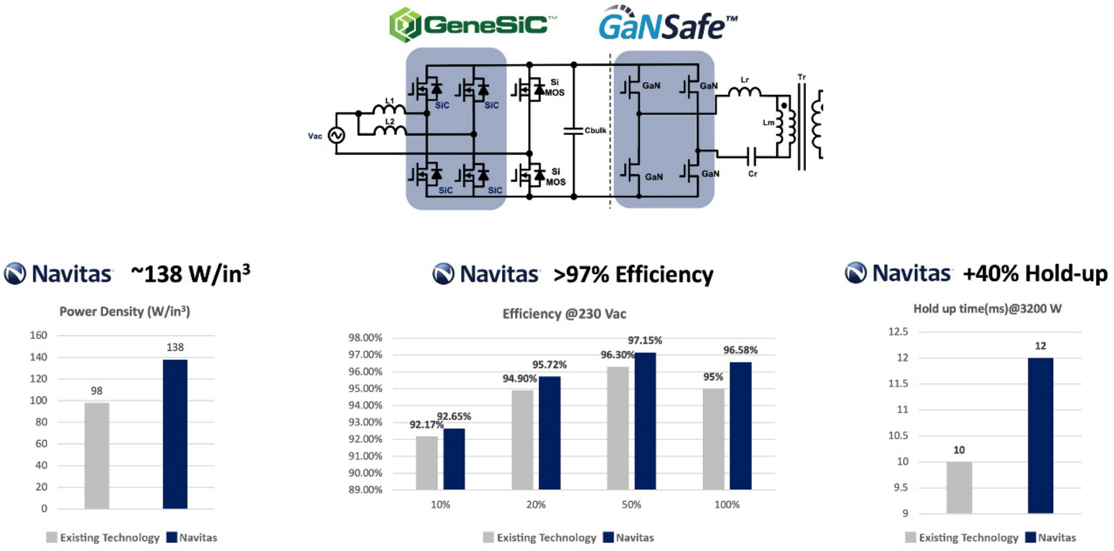

Figure 3: 54 V AI/GPU server reference design performance improvements in comparison to existing 3,200 W CRPS solution

In addition to improved power, efficiency improves to over 97%, reducing waste power and mitigating the need for additional thermal management provision. Finally, hold-up time goes up by around 40%, contributing to reduced risk of data loss or hardware damage in the event of power interruptions.

Conclusion

Delivering the increased processing and capacity demanded by mainstream AI and machine learning (ML) applications requires new approaches to power delivery and management to ensure performance while addressing design, commercial and legislative requirements related to size, efficiency and sustainability. Implementing these new approaches means moving away from conventional silicon and discrete GaN semiconductors to GaN devices that feature high levels of integration. By choosing ICs that combine control, switching, sensing and protection into a single device, for example, it is possible to develop highly efficient power supplies with significantly higher power ratings within existing form factors.