ICeGaN to Enable 100kW/Rack Data Centres

By Di Chen, Director of Business Development & Technical Marketing, Cambridge GaN Devices

It could be argued that the power requirement for data centres is getting out of hand. If you look back even just a few years, the accepted demand to drive CPUs and typical servers was 10 to 15 kilowatts per rack. But now – driven by the voracious appetite for high-performance GPUs used in generative AI – the expectation is that 100kW/rack will be needed…indeed some customers are talking about 200kW/rack. Couple this massive order-of-magnitude increase in power required with the need to meet international mandatory and advisory efficiency guidelines and the scale of the challenge becomes impressive.

Artificial intelligence had been discussed for many years, but when ChatGPT launched in late 2022, it grabbed the world's consciousness and all of a sudden became a topic of conversation that was no longer limited to geek circles but widely discussed on mainstream TV, in bars, and across ordinary dinner tables. Generative AI is able to create new content - text, images, audio – which some see as unleashing a huge burst of human creativity and others perceive as career-ending. Whatever your view, the genie is out of the bottle and won't be going back in. Generative AI not only requires power-hungry GPUs, it also demands that other system elements – CPUs, logic, memory –keep up. And it is this new application that is driving the huge upsurge in demand for more and more powerful data centres.

Efficiency and Power Density

The issue really comes down to this: IT equipment users are familiar with energy efficiency standards such as 80 Plus Bronze/Silver/Gold/Titanium which were introduced to drive efficiency levels up and wasted energy levels down. Now we have the Open Compute Project, and the ORV3 spec which demands end-to-end efficiency conversion levels as high as 97.5%. To put that into context, assuming two stages in the AC to DC power supply, PFC and LLC, to achieve 97.5% will require conversion efficiencies of over 99% in the PFC stage, and near 98.5% for the LLC stage.

It is possible, using complex power architectures, to achieve such high levels of conversion efficiency using high-performance silicon MOSFETS, but the topologies required will mean that a lot of components are required, and extra thermal management solutions – huge heatsinks and even water cooling systems - will be needed, resulting in a decrease in power density. This is exactly the opposite of what is required: to deliver more power per rack, much greater power

densities are needed. Today, we might see six 3kW power supplies on a 1U shelf, giving 18kW, or 15kW if one unit is required for redundancy. To deliver 100kW at current power density levels, the power supply units alone would occupy 6U, leaving very little for the IT systems themselves.

But things are moving quickly. Only shortly after the current ORV3 spec was released, the latest ORV3 HPR (High Power Rack) spec has now increased the power level to 5.5kW per PSU at 48W/in^3, or a total of 33kW for the power shelf. Next-generation PSUs will target 8 to 10kW each reaching the limit for the forced air-cooled solution, and the efficiency target will eventually reach 98%. If the PSU spec goes to 10kW and its size stays same as in the HPR spec, the power density requirement will be 90W/in^3! This demand for power density and efficiency of the server rack is unprecedented.

Today, the best of the silicon-based mainstream datacenter power supplies will have a power density of 60-70W/in^3. But they are only 96% efficient. PSUs employing silicon switches that will meet the ORV3 97.5% target will be 40% larger, with a power density of around 50 W/in^3. It's a trade-off: either you meet the power density, or you achieve the efficiency levels – not both. Another way to address the issue is to look at silicon carbide switches. Whilst these solutions can achieve the required efficiency levels, again the designs tend to be complex and therefore larger and more costly.

Enter Gallium Nitride

This situation has persuaded more and more designers and power architects to turn to gallium nitride (GaN) power solutions for data centres. GaN devices exhibit much lower switching losses than silicon for two main reasons. First, GaN's Figure of Merit, Q_g*$R_{DS(on)}$, is an order of magnitude better (smaller) than silicon, and second, because GaN switches have no body diode, they have no reverse recovery charge, Q_{rr}. Some commentators have described GaN as the near-perfect switching technology. This means that GaN power converter systems can hit 97.5% - and even higher in the future – using simple topologies with fewer components.

Let's look at that in more detail, considering a 3kW AC/DC Bridgeless Totem pole PFC design. Using GaN in Continuous Current Mode (CCM), requires only one half-bridge employing two GaN switches, plus the other half-bridge which can use a silicon MOSFET since it's at line frequency and therefore slow switching. With the inductor that totals just four components. Because silicon MOSFETs have a body diode, hard switching is impractical, therefore our Totem Pole PFC example fabricated using silicon MOSFETs could not use CCM. Instead, a CRM (Critical Conduction Mode) design would be required, usually coupled with interleaving. This would

mean that two phase legs (each phase includes 1x half bridge) would be needed, and many more components would be required than the four needed for the GaN solution.

GaN reduces complexity in other ways too. Because it is much more efficient, much less cooling is required since much less heat is generated. Also, because GaN can routinely switch at twice or three times the frequency of silicon, the size of the passives required for input filtering can be reduced, and also the size of the magnetics.

However, it goes further than just swapping out silicon MOSFETs for GaN. Because of A.I.'s aggressive power demand and energy-saving requirements, designers are starting to get creative (it's what they do!) and rethinking the whole design of power systems for data centre server racks. Currently, the usual OCP standard approach operates with a 200-277 V AC input into the rack, after which the power supplies convert AC to DC, followed by a DC-to-DC conversion to a low voltage bus or 48 volts. But now people are questioning the wisdom of performing this conversion inside the rack since we are going to need to squeeze 200 kilowatts in one rack and every single conversion stage creates losses. Perhaps instead, it might be better to have a centralized AC to DC stage outside of the rack. This approach would see the three-phase power converted to HV DC and distributed using a DC busbar. And then the 400V DC power is fed into the rack, requiring only one stage high voltage conversion inside the rack.

Whether is AC or DC feed power rack system, GaN is also a perfect fit on the primary side of the DC/DC converters employed in data centre applications. Using LLC as example, GaN enables >2x higher resonant frequency on LLC and increases the switching frequency into >500kHz range. GaN has much lower and linear C_{OSS} making the LLC ZVS design much easier than SJ MOS. With GaN solution, the LLC transformer and resonant tank can be much smaller and more efficient, improving the overall efficiency and power density.

But even with DC input to the rack, if the rack power keeps increasing as it is today, the loss on the 48V DC bus will be too large to be ignored. In future, a single-stage, direct 380V-to-12V GaN-based DC/DC converter could be used to replace the two-stage (400V to 48V LLC and 48V-12V IBC) conversion technique. Such a compact DC/DC brick could be located directly on the server shelf for optimum efficiency (HV DC direct into the server board) to maximize the power density, with the help of liquid immersion cooling. This approach shows that GaN can really help to further push the power density.
In summary, GaN can push the AC/DC PSU power density into the range of >100W/in^3 and at same time it can achieve 98% efficiency target. It is the optimum solution for the next gen high power rack PSU.

ICeGAN™ for data centers

So we can see that GaN technology, in general, offers significant intrinsic benefits of efficiency, power density and cost in data centre power supply design. ICeGaN, Cambridge GaN Devices' (CGD) HEMT technology, is particularly suited to this application for several reasons. Firstly, of course, is the well-documented 'ease-of-use' attribute. ICeGaN HEMTs include driver interface circuitry within the IC package, so devices can be driven using standard silicon MOSFET drivers – the customer does not have to change their entire driver system to switch from a silicon MOSFET solution to GaN.

Perhaps more significantly is that because ICeGaN HEMTs include a Miller Clamp on the gallium chip, they do not need a negative voltage rail to prevent shoot-through and dv/dt immunity problems. Many other competitors recommend -3V or even -6V. That can cause a significant increase in reverse conduction voltage drop – from, say, 1.7V for ICeGaN parts, to 4.7V or (worst case) 7.7V. That difference can result in a significant increase in dead time reverse conduction losses As well as the overall efficiency of the system, this may require extra cooling and the increase injunction temperature will affect reliability.

Finally, there is a very practical benefit that ICeGaN delivers, again because of the included Miller Clamp. Without the Miller Clamp, designers are forced to locate drivers as close as physically possible to the GaN HEMT to minimize parasitics. Otherwise, there can be ringing, overshoot and false turn-on events, even if the negative three volts is applied. This can cause challenges as every customer's layout is different, and they definitely don't want to make changes just to accommodate the GaN device. Sometimes, for example, they want to mount the switch vertically, or maybe keep the driver off the board altogether. CGD has demonstrated that with ICeGaN the drivers can even be connected to the switches using long leads and TO packaging without impairing the system's functionality. This is because in ICeGaN, the Miller Clamp is used to turn off the devices instead of relying entirely on an external driver. So, the customer does not need to care about how far they position the driver from the ICeGaN switch.

Conclusion

The voracious demands of generative AI have driven the anticipated power demand per server rack up by almost an order of magnitude. Addressing that increase, power system designers are looking for new architectures that meet not only the power density demands but also the energy efficiency requirements of international standards organisations. Gallium nitride technology plays a part in whether we stick to conventional rack power systems or adopt new

approaches. ICeGaN from CGD addresses the requirements with an innovative, easy-to-use and power-efficient implementation.