WITH AI INNOVATION COMES POWER CHALLENGES

An interview with Maury Wood, VP Strategic Marketing, Vicor Corporation

1. The cultural revolution that is generative AI (genAI) has been compared to the popularisation of the Internet, in terms of its expected human impact. What are you seeing from your vantage point?

Over the past 18 months, we've seen an extraordinary burst of innovation in the full genAI ecosystem – from processor chips, to dedicated genAI software and supercomputers to speed the development of domain-specific AI applications, to corporate strategy announcements signalling the massive impact genAI is anticipated to make on the worldwide economy for the foreseeable future.

GenAI model training is already driving the highest possible levels of compute performance, storage capacity, and network bandwidth. Some of today's highest performing supercomputers (typically measured by floating point math performance) are dedicated to genAI model training. GenAI is motivating historical levels of new investment in the semiconductor, infrastructure hardware, system software, and network edge sectors, and this investment activity can be expected to spread into embedded AI devices for homes and workplaces.

2. What are some of the downside consequences of the proliferation of genAI?

Well, beyond the existential concerns that have been expressed (!), one key cost of this burst of innovation is rapidly increasing energy usage at the Cloud datacentres that host genAI training and inferencing activities, with some concerning forecasts. For example, The New York Times has compared the expected electrical energy consumed by genAI in 2027 to what Argentina, the Netherlands and Sweden each use in a year. Generative AI model training and inferencing present a mounting power consumption challenge incompatible with societal energy and greenhouse gas reduction objectives.

3. Why is genAI computing so uniquely power hungry?

First, allow me to draw a distinction. When we as individuals use genAI tools, we are making queries against pre-trained large language models (LLMs), and this so-called inferencing activity is not inordinately power intensive and can be performed at the network Edge. On the other hand, the process of training genAI LLMs requires a tremendous amount of compute time (currently measured in months) on the supercomputers I mentioned earlier. These supercomputers use thousands of specialised processors based on graphic processing units (GPUs) and each processor uses a huge number of transistors – 100 billion or more. The training processors are made using the most advanced semiconductor process technologies, such as 4nm CMOS, which leak current during operation. Because the supply voltage of these transistors is 0.7VDD or so, the continuous current demand can be 1,000A or higher, putting continuous power (also known as thermal design power) at 700 watts or more. When you multiply 700W times thousands of processors per genAI supercomputer, times hundreds of Cloud genAI supercomputers worldwide, the aggregate power consumption just skyrockets.

Here is an example. According to Nvidia, OpenAI's GPT-3 with 175 billion model parameters requires about 300 zettaFLOPS (1021 Floating Point Operations per Second), which is 300,000 billion-billion math operations across the entire training cycle. And these model sizes are only going to increase, with trillion-parameter neural network models in development today.

4. Can conventional switch mode power supply architectures meet the genAI power challenge?

Until fairly recently, the racks in datacentres have used 12VDC power distribution. Vicor, among other power system innovators, over the past 10 years, has advocated for the use of 48V DC power in datacentre racks, because (thanks to Ohm's Law) higher voltage yields lower power losses in conductors with non-zero electrical resistance. The adoption of 48VDC power for higher performance computing applications received a major boost in the Open Rack specifications standardised by the Open Compute

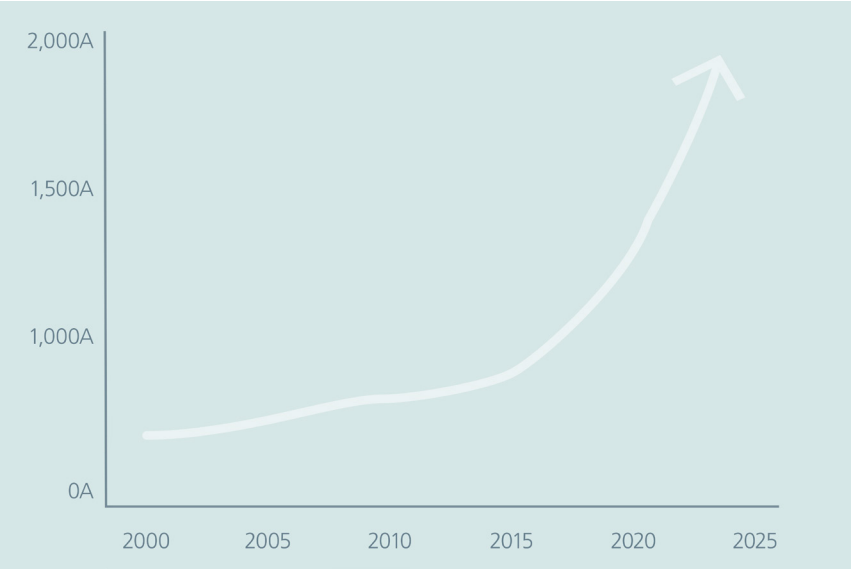
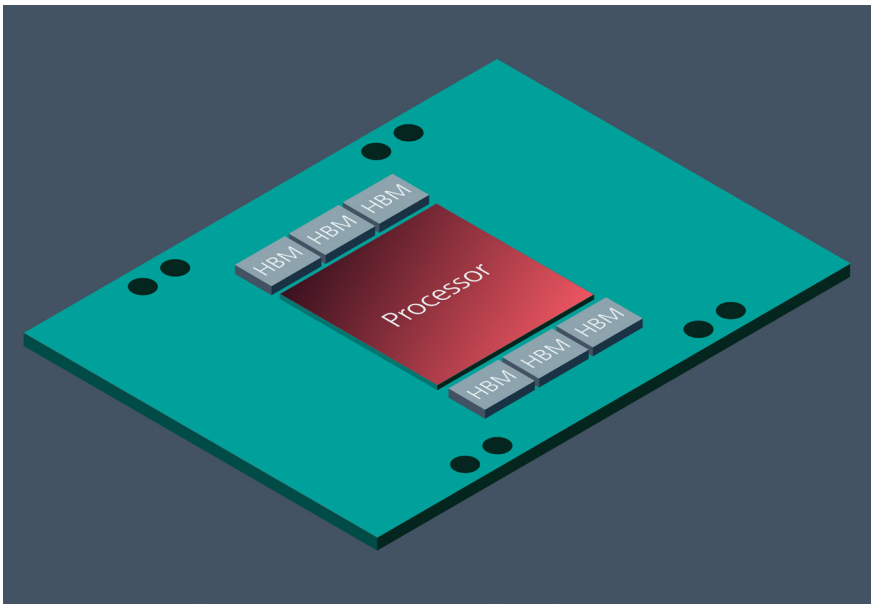


Figure 1. The progression of genAI training processor peak current requirements has no end in sight

Figure 2. Conceptual accelerator module (AM) showing the GPU-based processor and supporting high bandwidth memories (HBMs) is the building block for genAI



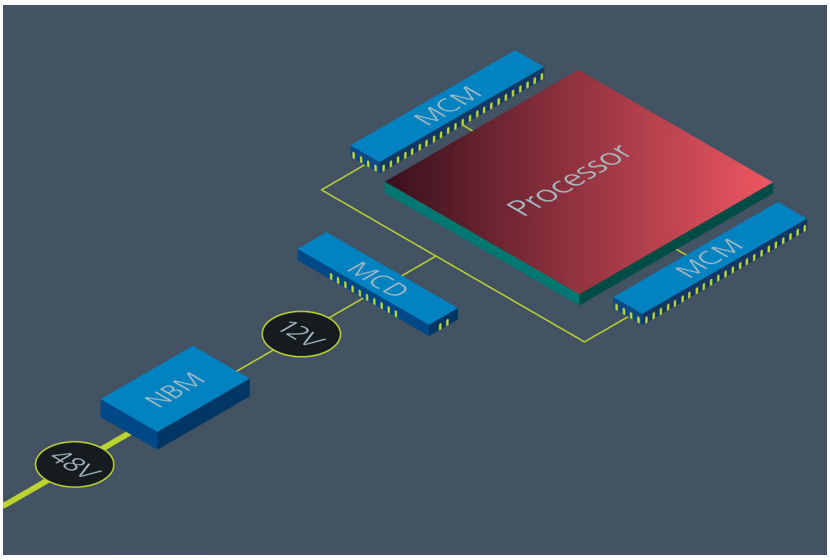


Figure 3. The Vicor factorised power architecture (FPA) with thermally-adept modular current multipliers (MCM) at the point-of-load is ideal for genAI training applications

Project. In early genAI power distribution architectures, this nominal 48VDC supply is converted into an intermediate bus voltage at the accelerator module, with this intermediate DC signal feeding multiphase trans-inductor voltage regulators (TLVRs), an approach that has hard limits in terms of scalability and current density.

5. Why is the TLVR approach to power delivery for genAI processors inadequate?

The printed circuit board (PCB) space available on the accelerator modules used with genAI training processors is extremely limited, meaning that the power delivery subsystems for these processors must have very high power density (W/mm2) and current density (A/mm2). Conventional power supplies simply cannot achieve the power and current density required to both supply the needed current and fit within the available PCB area. The power components for GenAI training processors must also meet the dynamic performance demands caused by load-step transients. Again, conventional power delivery approaches are not optimally suited for those demands.

Additionally, the components in a genAI power delivery architecture must be highly thermally adept. Whether the genAI system is liquid- or air-cooled, the power components must have high thermal conductivity and packaging that can withstand extraordinarily high levels of thermal cycling over its operating life. More recent genAI accelerator modules use a factorised power architecture, with the point-of-load converters utilizing current multiplication, such as those innovated by Vicor.

6. How is Vicor technology improving genAI power delivery?

Unique Vicor power modules are moulded and then plated using electroless nickel immersion gold (ENIG) processing. The moulded construction ensures mechanical rigidity and environmental robustness across temperature, humidity and vibration. The plated exterior enables high-yield surface-mount assembly, which provides an ideal thermal conductor for forced-air or liquid cooling using cold plates.

Vicor power modules incorporate a proprietary Sine Amplitude Converter (SAC) circuit topology, which uses zero voltage switching (ZVS) and zero current switching (ZCS) technology to minimise switching noise and spurious radiated emissions and maximises DC-DC conversion efficiency. Vicor also uses high-frequency MOSFET switching to reduce the physical size of the highly integrated modules. Finally, Vicor's point-of-load components for AI/HPC applications are very thin (≤1.7mm) and deliver scalable current output in a family of PCB-compatible footprints.

GenAI will, without doubt, remain the most power-intensive and thermally-challenging application in the modern computing world for the foreseeable future. Vicor will continue to innovate to meet the escalating power delivery requirements of this exciting new business opportunity.