INTELLIGENCE UPDATE

Erratic power profiles of AI clusters: the root causes


Daniel Bizo

27 Sep 2024

Ever since ChatGPT-3 arrived with a bang in late 2022, generative AI has dominated discussions among those working in the data center industry. Most of the attention has been directed at the high power density of hardware for AI, as well as the sheer size of the infrastructure needed to run some of these workloads. However, there is a much less publicized issue around generative AI training: its potentially erratic power profile.

In discussions with vendors and data center operators, Uptime Intelligence has come across several independent reports of AI training clusters producing big amplitude power swings. What makes these power events unusual is their scale — they are large enough to create rapid and large variations in the electrical load on UPS systems, the grid or engine generators.

The root cause of the phenomenon is in the design of modern processors, in this case, GPUs. On the silicon level, spikes in power consumption are well documented and extensively researched. For the past couple of decades, the performance of multicore processors has been hindered by thermal power limitations rather than the absolute switching speeds that the circuitry could achieve (limited by signal propagation delay). Modern many-core processors are able to drive much higher clock speeds (e.g., 5 GHz across all cores) without thermal design power (TDP) limitations. Such limits are in place, both to match products to various market needs and, for the highest-performing products, to prevent hotspots on the silicon from overheating. However, TDP does not define the absolute maximum power consumption of a chip.

Transient extremes: GPUs in lockstep

There are several types of power excursions possible with modern silicon that differ in their magnitude and length. Most common are the result of opportunistic, performance-boosting mechanisms that are built into virtually all modern compute silicon to temporarily extract more performance when the temperature allows it. These excursions are relatively moderate (typically up to a 25% increase in power consumption) and can last tens of seconds.

The build-up of these incremental power events is fast but relatively gradual as it happens in

© COPYRIGHT 2024 UPTIME INSTITUTE. ALL RIGHTS RESERVED.

voltage-frequency steps (tens of microseconds for each step), often governed by the operating system kernel (that can respond in milliseconds). The climb-down is orders of magnitude slower because it is thermally dictated. Such events do not pose a significant risk to power distribution components.

Silicon can also produce a different kind of power excursion that is much deeper and more rapid, near instantaneous. Some instruction sequences create conditions for a sudden burst of transistor activity, such as a parallel (vectorized) operation on a dataset that has just been loaded into a processor cache. When the code is multithreaded, the same event can happen simultaneously across several processor cores due to the synchronization effect of data load from memory, sequential dependency on other computations, or other explicit synchronization events in the code that make core executions align. This creates an inrush of current that can result in a transient (up to hundreds of microseconds) power spike that is two to three times the TDP rating of a processor or GPU (e.g., 350 W).

Silicon power delivery networks and well-designed system boards are built to handle these extreme power events, including guarding voltage levels against sags and surges. Modern server processors and GPUs can drive currents as high as several hundreds of amps (A) each — the next-generation chips, which are already in high-volume production for general availability in 2025, will be in excess of 1,000 A. GPUs are more susceptible to this behavior than general-purpose processors, owing to their parallel compute-oriented design, which enables them to simultaneously exercise a much larger proportion of the silicon before silicon controls intervene to rein back power.

Notably, extreme power excursion events are common and well-documented in gaming PCs equipped with powerful GPUs that test the capabilities of the system board and power supply unit (PSU). If the PSU is sized incorrectly or uses sub-par quality components (e.g., voltage regulators or capacitors), the computer may crash or shut down during a transient power surge driven by the GPU (as high as 300 W to 400 W above their TDP rating), as many PC enthusiasts have learned the hard way.

Although mis-sizing and quality issues in power electronics are much less likely in enterprise-grade servers, the novel nature of generative AI training still creates an issue for data center facilities. When many GPU-heavy servers run the same massively parallel model training workload, synchronization events will make a large number of GPUs operate practically in lockstep (for performance, not fault tolerance), even if only temporarily. This can create conditions for macro-level power events across several servers.

These software-induced power transients can surge up to 150% of the steady-state maximum power level of a compute cluster, Uptime Intelligence learned during discussions with suppliers. For a larger AI training cluster with hundreds or thousands of GPUs, this will mean near instantaneous (climbing in tens of microseconds) spikes of hundreds of kilowatts, followed by similarly dramatic drops, all showing up on the UPS load. These transients can occur frequently, with some operators reporting that macro-level power excursion events happen several times a second, which can, albeit not always, create large and rapid swings in currents.

© COPYRIGHT 2024 UPTIME INSTITUTE. ALL RIGHTS RESERVED.

Activity lulls during checkpoints

AI training workloads also exhibit another type of volatile power behavior: the load regularly drops, albeit not as fast as transients, to a fraction (e.g., 10% to 20%) of the typical steady-state maximum power for several seconds. After the lull, the ramp to full power is steep, which often results in an overshoot.

This can happen every few minutes, Uptime Intelligence understands, and the phenomenon is likely due to data checkpointing and data loading events, during which computation is interrupted to write the model's state onto persistent storage. Frequent checkpoints, particularly for larger training clusters with long training runs, are needed to ensure that any hardware component failures, which happen relatively often at scale, do not result in too much wasted training time. AI training systems are typically unable to tolerate any failure during runtime and need restarting from the last known good state.

Often, load drops are followed by transient surges because compute threads are unleashed at the same time. Together, these two types of power events mean that some AI training clusters will see the power load go from 10% to 20% of the steady-state maximum to 150% in a matter of a second before dropping back to 100% followed by several transients and another checkpoint cycle.

This combination of events may put even higher undue stress on upstream electrical equipment, such as breakers and static switches, but crucially on the UPS system and the engine generators. In upcoming reports, Uptime Intelligence will discuss the options that data center operators have to cushion the impact on facility equipment.

© COPYRIGHT 2024 UPTIME INSTITUTE. ALL RIGHTS RESERVED.

ABOUT THE AUTHOR

Daniel Bizo

Daniel serves as Research Director at Uptime Institute. Over the past 15 years, he has covered the business and technology of enterprise IT and infrastructure in various roles, the past ten years as an industry analyst and advisor. His research includes sustainability, operations, and energy efficiency within the data center, on topics like emerging battery technologies, thermal operation guidelines, and processor chip technology.

dbizo@uptimeinstitute.com

© COPYRIGHT 2024 UPTIME INSTITUTE. ALL RIGHTS RESERVED.

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions.

With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.

© COPYRIGHT 2024 UPTIME INSTITUTE. ALL RIGHTS RESERVED.