Electronic Design

DAVOOD YAZDANI, Executive Vice President and Chief Product Officer (CPO), Endura Technologies

Tearing Down the Power Wall with Integrated Voltage Regulators

IVRs can be placed inside AI chip packages, reducing losses while enabling faster, more precise power delivery.

As AI models grow in complexity, a physical "power wall" is emerging. For decades, engineers have delivered and managed power from a distance, relying on voltage regulators on the motherboard. But as GPUs and other AI accelerators consume ever more power, that separation creates electrical "friction," which wastes energy, generates heat, and slows the system's ability to respond to rapid changes in workload.

Integrated voltage regulators (IVRs) tackle this problem by integrating most of the DC-DC converter into a single chip, including inductors and capacitors required for stable, smooth power delivery. Thanks to this integration, these power ICs enable voltage regulation to move directly into the xPU, whether a CPU, GPU, TPU, or NPU. The result is a shorter path between the voltage regulator and the transistors powered by it. It enables several capabilities that have been difficult to achieve at scale:

- **Reflexive power:** IVRs allow a chip to react to sudden workload spikes in nanoseconds, preventing voltage drops that force engineers to waste power on safety margins.
- **Precise control:** Instead of throttling an entire chip when it nears thermal limits, designers can now manage power core-by-core.

Unlocking a lot of these benefits depends on chip designers and power engineers working more closely together than ever before. But if successful, the IVR could be one of the most significant pivots in power delivery in a generation.

The Power Wall: A Systemic Constraint on Compute

The shift to in-package or on-chip power management is driven by what's now an undeniable reality: Power is now the dominant constraint in modern compute. Today, high-performance server chips operate at power levels that once seemed implausible. The most advanced GPUs have thermal design power (TDP) specifications of more than a kilowatt and peak power demands approaching multiple kilowatts per chip — with both continuing to rise.

These extremes are rapidly exposing the limitations of traditional power delivery networks (PDNs). With kilowatt-class GPUs and other AI accelerators, power engineers are contending not only with massive resistive (or I^2R) losses and extreme heat that comes with delivering very high currents through the PDN — they're also dealing with voltage droop (V_{droop}) that occurs when current isn't delivered fast enough.

Thus, performance gains are no longer gated by transistor count alone, but by the ability to deliver stable and efficient power at very high current loads.

Proximity: Why It Makes a Difference for Voltage Regulation

The first core insight behind the rise of IVRs is that physical distance is the enemy of stability. The problem lies in parasitic impedance that invariably adds up over longer distances. Board-level regulators, often inches away from the processor, must push current through a gauntlet of PCB traces, socket pins, and package interconnects. Each stage introduces parasitic resistance (R) and inductance (L).

In a high-current environment, even a few milliohms of resistance lead to large voltage drops, while inductance limits how quickly the system can respond to bursty AI workloads.

By integrating regulations directly into the package or onto the die, engineers can dramatically reduce PDN impedance. This proximity, now possible with IVRs, enables a level of precision that external components simply can't

match. Moving the IVR closer to the xPU effectively eliminates the electrical bottleneck of the motherboard.

The bottom line is that physical distance creates lag in power delivery. IVRs eliminate that lag by putting the power regulator exactly where computationally heavy workloads like AI take place.

High Bandwidth: The Dual Advantage of Response and Recovery

What stands out about the IVR, however, is its high control-loop bandwidth, which is usually tied to its control architecture and the ability of its power switches to operate frequencies of more than 100 MHz. To understand the importance of this high bandwidth, we need to distinguish between two critical metrics:

- **Transient response:** When a GPU or other AI accelerator suddenly spikes its current demand (di/dt), the voltage tends to undershoot due to the slow transient response of traditional voltage regulator modules (VRMs). This happens frequently due to sudden fluctuations in AI training. IVRs detect and react to these dips in nanoseconds.
  - By keeping undershoots shallow, engineers can lower the overall operating voltage (Vmin), reclaiming power that was previously wasted on safety cushions for the processor. In general, supply voltage undershoots (or overshoots) must be limited to less than 10% (or 0.07 V at 0.7 V_{DD}) to prevent transistor damage during these frequent transient events.
- **Recovery time:** Besides being slow to respond to transient loads, traditional regulators are also slow to recover from them. These components can take over 20 µs to settle after a load transient event. If a second AI workload hits while the voltage isn't settled, the system can crash. An IVR settles and recovers in less than 200 ns *(see figure)*. This ultra-fast recovery ensures that the voltage regulator is ready for the next task almost instantly.

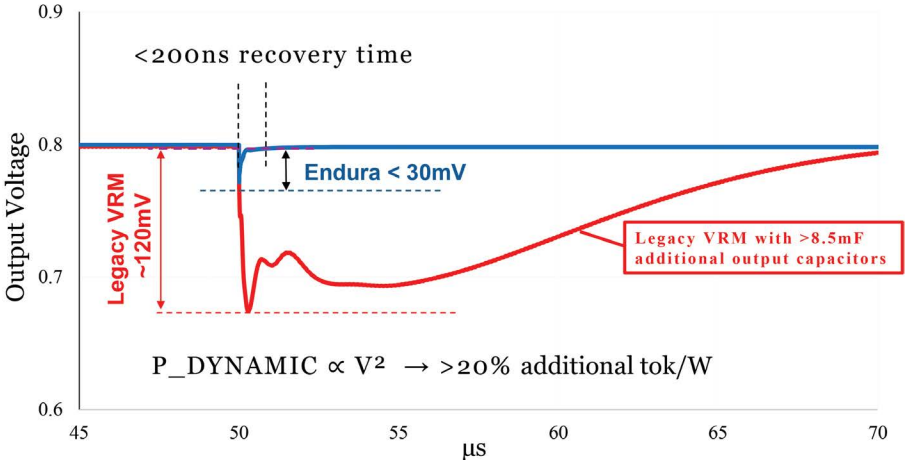
"Micro-Islands" for More Granular Power Management

Traditional power delivery relies on bulk regulation, which involves a single voltage rail feeding hundreds of cores. This creates a lowest common denominator effect: If one core is struggling with a heavy workload or if it's hampered by process variation, the entire chip must be over-energized to maintain stability. This wastes massive amounts of energy across the rest of the die and potentially leads to other problems.

IVRs enable the transition to power what can be called "micro-islands," revolutionizing silicon real estate through several different mechanisms:

- **Per-core IVRs:** Using IVRs, it's possible to partition the SoC into dozens of independent voltage islands. For example, a high-demand tensor block for AI acceleration can receive 0.85 V for peak throughput, while adjacent SRAM caches can stay at a stable 0.75 V to prevent bit-flips and idle I/O lanes drop into deep sleep, where they operate on 0.4 V.
- **Thermal precision:** Traditional power management is a blunt instrument in the sense that hitting a thermal limit forces the performance of the entire chip to be throttled. With IVRs, the system gains more granular control over the power profile, reducing the voltage of the specific block where the hot spot is occurring. The rest of the chip can then maintain its clock speed.
- **Mitigating silicon variation:** Due to manufacturing variances, no two cores are physically identical. IVRs allow the system to "tune" the voltage for each core's unique physical characteristics. You no longer need to set a single supply voltage in the VRM based on the weakest core on the processor; you can optimize every millimeter of silicon for its specific performance-per-watt sweet spot.

Another way to think about the difference: Traditional power management is like a giant light switch for an entire building, while IVRs give every room its own dimmer switch.



The plot compares Endura's IVR and a legacy VRM in terms of transient response to a 1,000-A load step at 75 A/ns. (Credit: Endura Technologies)

AVFS and IVR: On-Chip Intelligence and Off-Chip Muscle

IVRs also play directly into on-chip power management techniques such as adaptive voltage and frequency scaling (AVFS). This is the high-speed control logic that uses in-situ monitors (critical path monitors) embedded directly into the logic paths. These sensors act like an early-warning system, detecting exactly how much timing margin remains before the silicon fails due to heat, aging, or voltage droop.

The IVR is the execution engine for AVFS. Without an IVR, the AVFS's insights are useless because the hardware can't react fast enough. In a traditional setup using legacy voltage regulators, by the time a command to adjust voltage travels to the motherboard and back, the workload has already changed, potentially leading to a crash. The IVR enables a closed-loop system where the AVFS system identifies even a 10-mV opportunity for efficiency, and the IVR executes it in nanosecond intervals.

This allows the chip to react in real-time to the AI model's computational bursts, reclaiming dark silicon that was previously unusable due to power constraints.

Beyond the Rack: Powering the Next Generation of Robotics

While IVRs are becoming a core building block of AI data centers, they also have the potential to play in the world of physical AI, specifically robotics.

In these situations, power integrity is synonymous with physical safety and reliability. The fast transient response of an IVR functions like a human reflex, ensuring a robot's navigation system never suffers a brownout or reset during a sudden, high-torque motor movement.

Meanwhile, the fast recovery time enables fluid movement. As a result, a robot or drone can process intense bursts of data from cameras, radar, and LiDAR without the electrical noise or voltage drops causing computational latency.

For autonomous robots to interact with their surroundings smoothly and safely, their processors need nanosecond-scale reflexes that only IVRs can provide.

1. The Ecosystem Shift: A Unified Co-Design Model

The shift to IVRs isn't as easy as placing them inside the processor's package. They represent a fundamental shift in the semiconductor business model and the usual approach to chip design. Instead of working separately, processor architects, IVR engineers, and package engineers must collaborate from the first line of RTL code. Logic and power can no longer be designed in silos.

The tools used by engineers are also changing to break down the barriers to co-design. The semiconductor industry is moving beyond static analysis toward AI-augmented multi-physics platforms. While foundational simulation engines exist, the next generation of tools must bridge the gap between power, heat, and logic — automating the design of hundreds of independent power domains that are now too complex for manual human optimization.

The other piece of the puzzle is a standard interconnect. Industry adoption of the Universal Chiplet Interconnect Express (UCIe) is crucial. It acts as the common language for in-package connectivity, allowing different chiplets to seamlessly share a unified power and data fabric.

The Future of Compute Comes Down to How It Gets Power

We're entering an era where power delivery is a first-class architectural discipline, as critical to the success of an AI chip as the instruction set. As AI models continue to grow, the ability to deliver massive currents with nanosecond precision and surgical granularity will determine which architectures survive. IVRs are the key to unlocking even more performance. Without them, compute will be capped by physics.

Davood Yazdani, executive VP and CPO, oversees engineering, product, and operations at Endura, shaping the company's vision for AI power solutions. He brings decades of leadership in power management from Infineon, Renesas, and Intersil, and holds a Ph.D. in Electrical Engineering.