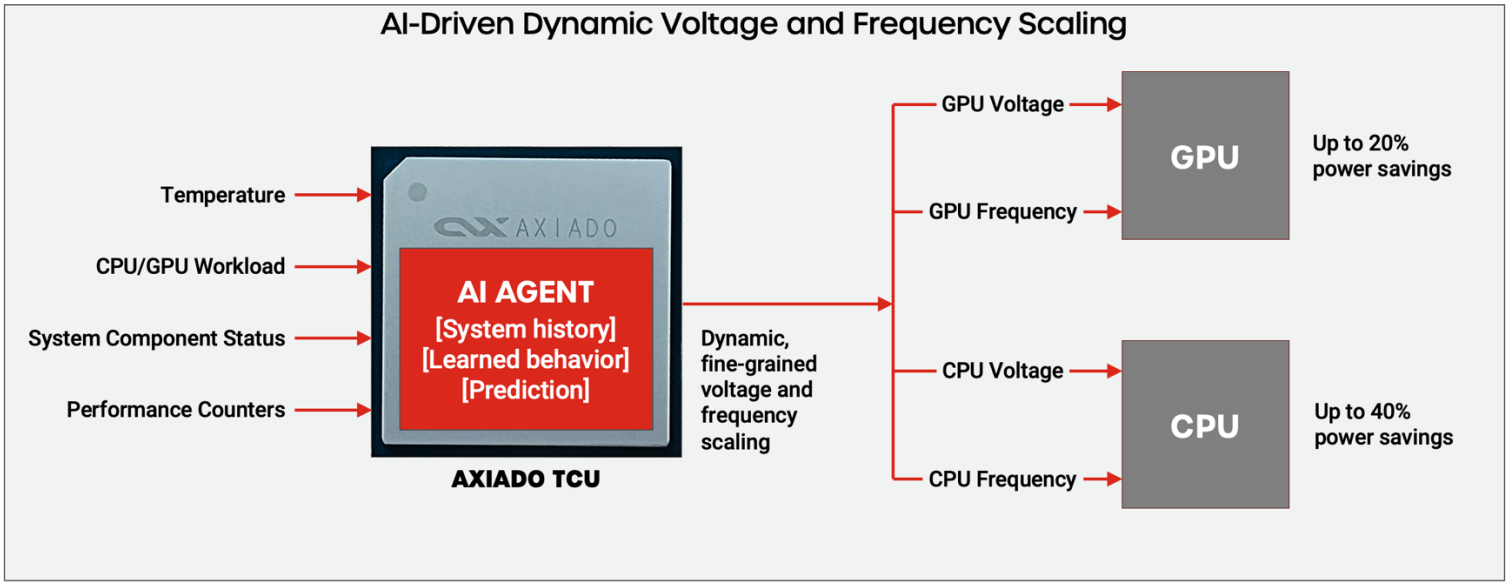
Axiado AI Agents Enhance Rack Power Efficiency, Savings up to 148 MWh/Year/Rack

AI-DRIVEN DYNAMIC THERMAL MANAGEMENT & DYNAMIC VOLTAGE AND FREQUENCY SCALING OPTIMIZE COOLING, CPU & GPU OPERATION

Axiado offers AI agents running in trusted silicon on the AX3080 Trusted Control/Compute Unit (TCU), an integrated hardware solution that is available as a complete system-in-package or as part of an OCP-compliant DC-SCM or Nvidia IFF control module that can plug into any compatible system. These AI agents can perform many advanced tasks to make modern compute infrastructure more secure, efficient, and resilient. One compelling use case is using AI agents to optimize air cooling and CPU/GPU operational settings to save power, an important goal as data centers strive to lower costs and become more sustainable. AI-driven Dynamic Thermal Management (DTM) uses AI agents to monitor system parameters, workloads, and temperature and dynamically determine optimal fan speeds for air cooling while maintaining safe operating conditions. This can save up to 50% of the power consumed by air cooling. AI-driven Dynamic Voltage and Frequency Scaling (DVFS) uses AI agents and proprietary algorithms to control the operating frequency and voltage of CPUs and GPUs to reduce power without impacting performance. This can save up to 40% of CPU power and up to 20% of GPU power on representative workloads based on measured experimental data in real systems. Multiple AI agents can run in parallel on the Axiado TCU, enabling DTM and DVFS to run simultaneously for maximum power savings. Using both DTM and DVFS agents, a GB200 server can save up to 8.2 MWh/year. A standard NVL72 rack with 18 servers would save up to 148 MWh/year, for a cost savings of up to $15,000 per year per rack.

SUMMARY

- AI Agents running on the Axiado TCU can optimize air cooling and CPU/GPU operational setting.
- AI-driven Dynamic Thermal Management saves up to 50% in air cooling costs.
- AI-driven Dynamic Voltage and Frequency Scaling saves up to 40% of CPU power and 20% of GPU power without impacting performance.
- Total power savings of 8.2 MWh/year for a GB200 server based on measured experimental data. For an NVL72 rack with 18 servers, the total savings would be up to 148 MWh/year.
- Assuming a power cost of $0.10/kWh, a gigawatt AI factory with 7,500 racks would save up to $110 million/year.



Proprietary DVFS techniques use AI agents running on the TCU to save power without impacting performance