

Energy-Aware AI Workloads on Heterogeneous HPC Clusters via Adaptive GPU Frequency Control

Background

AI workloads place growing energy demands on HPC data centers. While low- and mixed-precision techniques reduce computational cost at the numerical level, often at the cost of model accuracy, the GPU SM frequency is rarely optimized for energy-efficiency. Reducing GPU SM frequency lowers power draw but slows workloads; below a certain point the added runtime outweighs the power savings. Finding the efficient operating region requires adaptive control that responds to the workload as it runs.

Aims

This work presents an adaptive energy optimization controller for AI workloads on HPC clusters, finding and maintaining the optimal energy-time operating point across heterogeneous GPU hardware and parallelism configurations.

Method

The controller adjusts the GPU SM frequency cap at runtime using a contextual bandit on observed GPU telemetry: utilization, power draw, SM frequency, throughput. For training, a calibration sweep during the first steps identifies the energy valley, setting a safe frequency floor. The bandit then operates within this range, minimizing energy per step, which accounts for the runtime cost of lower frequencies. The controller runs as a separate process requiring no framework modifications, no pre-trained models, and only standard GPU telemetry. The same approach extends to inference workloads by substituting latency for throughput, optimizing the energy-latency trade-off under a response time target

Results

Validated on heterogeneous GPU hardware including NVIDIA_RTX_PRO_6000_Blackwell and H200_NVL, across multi-gpu parallelism configurations covering models from 7B to 72B parameters, the controller achieves 10-39% energy reductions across training and inference workloads, with savings directly configurable via a time budget and no impact on model quality or outputs. This amounts to a savings of 3.5-9 kWh per 12-hour window on a 4-GPU setup, with savings scaling further on large HPC clusters with hundreds or thousands of GPUs. GPU temperatures dropped by 3-14°C, suggesting potential reductions in facility cooling load worth further investigation.

Conclusion

Adaptive runtime SM frequency control delivers measurable energy reductions across heterogeneous GPU hardware and AI workload types without offline profiling or framework modifications, complementing precision-level optimizations on modern HPC

architectures. A defined time budget gives a direct lever to trade runtime for energy savings across AI workloads.