



PROFILE



Linkedin
LINKEDIN



AGENTIC AI SYSTEMS

AGENTIC AI SYSTEMS | TECHNICAL PUBLICATION 202610-AASPUB-0004

Consumer GPU Power-Curve Optimisation and Active Thermal Management

Experimental observations of 70-75% GPU power limits, 1-5% performance degradation and workload-aware fan cooling on selected AMD and NVIDIA consumer GPUs

Author: Reddy (Eddie) Jamisola

Organisation: Agentic AI Systems

Publication date: 20 October 2026

Document number: 202610-AASPUB-0004

Classification: Applied AI Infrastructure Engineering

Abstract

Agentic AI Systems presents an experimental approach to improving power efficiency and thermal management for consumer-grade graphics processing units (GPUs) used in artificial intelligence computation. Reported observations from selected NVIDIA GeForce RTX 3080, NVIDIA GeForce RTX 3090 and AMD Radeon RX 6800 cards indicate that operating at approximately 70-75% of the standard GPU power limit can correspond to approximately 1-5% compute-performance degradation under the tested configurations and workloads.

The associated cooling policy commands supported GPU fans to 100% during active compute, retains that speed for 30 seconds after processing completes, and then restores automatic fan control. Results can vary between GPU models and individual chips of the same model due to manufacturing-related electrical differences commonly called the silicon lottery. These experimental ranges are not universal guarantees.

1. Introduction

Consumer GPUs are widely used for local language-model inference, image and video generation, audio processing and distributed AI workloads. Maximum GPU power does not necessarily provide maximum efficiency: near the upper region of a voltage-frequency curve, extra electrical power may deliver only modest gains in processing throughput. This study describes reduced power targets and workload-aware cooling as an integrated operating policy.



PROFILE



LINKEDIN



AGENTIC AI SYSTEMS

2. Selected Experimental Hardware

Manufacturer	GPU model	Video memory
NVIDIA	GeForce RTX 3080	10 GB
NVIDIA	GeForce RTX 3090	24 GB
AMD	Radeon RX 6800	16 GB

These cards define the stated experimental scope. No per-card benchmark breakdown is asserted here; the ranges below represent the reported observations across selected configurations.

3. Reported Power and Performance Ranges

Operating parameter	Observed range or policy
Configured GPU power limit	Approximately 70-75%
Relative compute performance	Approximately 95-99% of baseline
Performance degradation	Approximately 1-5%
Fan speed during compute	100% command where supported
Post-compute cooling	30 seconds at full fan speed
After cooldown	Restore automatic fan control

A configured power limit of 70-75% does not establish that actual GPU or wall-socket energy consumption falls by 25-30%. Actual electrical power and energy must be measured for each workload.

4. Silicon Lottery and Chip-to-Chip Variability

The silicon lottery describes differences between semiconductor chips arising from manufacturing variability. Two cards of the same model may require different voltages for a given clock speed, reach different stable frequencies under a power cap, and display different power, temperature or throughput results. Board design, firmware and cooling solutions can add further variation.

For example, two RTX 3090 cards could retain different fractions of baseline throughput at the same nominal 70% power target. Likewise, two RX 6800 cards may have different stable voltage-frequency characteristics. These are illustrations of possible variability, not measured head-to-head comparisons in this paper. A common model name does not guarantee an identical power-efficiency curve.



PROFILE



Linkedin
LINKEDIN



AGENTIC AI SYSTEMS

5. Power-Curve Optimisation

The proposed approach selects a reduced GPU power target and evaluates whether useful workload throughput is retained. The reported 70-75% operating range and 1-5% degradation suggest that the tested configurations can occupy a favourable part of their power-performance curve. A stable and efficient target must nevertheless be identified for each physical GPU and workload, rather than assumed to apply to all cards.

6. Active Fan Control During Compute

When a qualifying GPU compute workload begins, the controller commands supported cooling fans to operate at 100% speed. The intent is to increase available airflow during sustained computation and help manage thermal accumulation. Actual temperature effects depend on cooling design, airflow, ambient conditions, driver capabilities and the GPU workload.

7. Thirty-Second Post-Compute Cooldown

At workload completion, the controller maintains full fan speed for another 30 seconds to address residual heat in the heatsink and surrounding components. It then returns fan management to the GPU automatic mode. If new computation begins during cooldown, the controller should keep active cooling enabled and restart the cooldown when processing ends. The 30-second duration is a control policy, not proof that every GPU reaches a specific temperature in that time.

8. Integrated Operating Sequence

Idle: GPU fans remain under automatic control.

Compute detection: A qualifying AI workload starts.

Active compute: Use the individually validated power target and command supported fans to 100%.

Cooldown: Maintain full fan speed for 30 seconds after the final active workload completes.

Recovery: Return fan control to automatic; a fail-safe should restore automatic control if the controller terminates unexpectedly.

9. Performance-per-Watt Interpretation

Useful computational throughput per watt provides one efficiency measure. As a purely illustrative calculation, retaining 97% of baseline throughput at 75% of measured baseline GPU power would improve relative performance per watt by about 29%. This is not an experimentally measured improvement claimed for the listed cards. Full evaluation should record throughput, elapsed time, actual GPU power, total workload energy and any extra fan energy consumption.



PROFILE



LinkedIn
LINKEDIN



AGENTIC AI SYSTEMS

10. Experimental Caveats and Limitations

The experiments were conducted on selected NVIDIA GeForce RTX 3080, RTX 3090 and AMD Radeon RX 6800 consumer GPUs. The observed results may differ on other GPU models and between individual cards of the same model. Variance can arise from silicon lottery, GPU architecture, chip revision, board and cooler design, firmware, drivers, voltage-frequency settings, ambient temperature, system airflow and workload composition.

The reported 70-75% power limit and approximately 1-5% performance degradation are observed ranges, not guaranteed outcomes. The observations do not establish a separate result for each listed model or demonstrate universal applicability across AMD or NVIDIA product families. Full-speed fans may increase noise, fan electricity use and mechanical wear; manual fan and power controls are not supported on all GPUs. Any deployment requires hardware-specific stability, thermal and energy validation.

11. Applications and Validation

Potential applications include local AI inference, concurrent model serving, image and video generation, audio processing and distributed consumer-GPU compute nodes. For reproducibility, future validation should record the exact card and firmware, driver, ambient temperature, workload, power target, measured electrical power, throughput, temperatures, fan speed and cooldown behaviour, with repeated baseline and optimised runs.

12. Conclusion

Selected NVIDIA RTX 3080, RTX 3090 and AMD RX 6800 consumer GPUs have shown reported operating ranges in which reduced power targets preserve much of baseline AI compute throughput. Combining a 70-75% power target with full-speed fans during compute and a 30-second post-compute cooldown offers a practical experimental policy for power-aware thermal management. The silicon lottery and other hardware-specific variables mean that optimisation must be measured and validated per card, rather than presented as a universal GPU guarantee.

© 2026 Agentic AI Systems - Intelligence. Automation. Advantage.