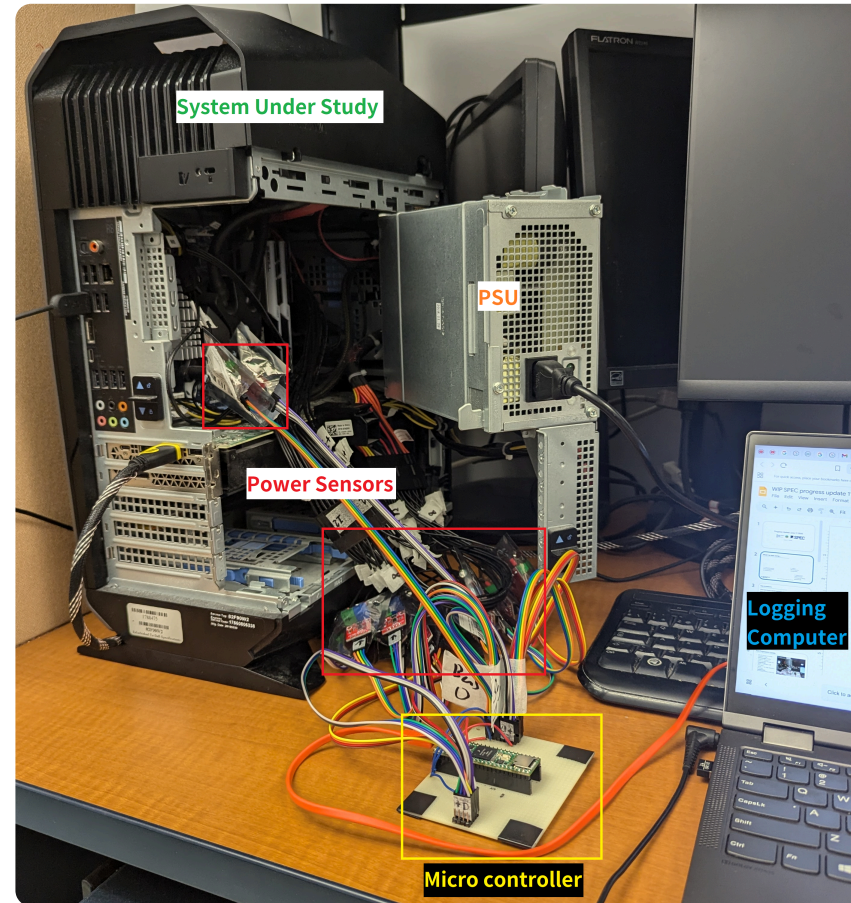


TOOLING MORE TOOLS

INVESTIGATING POWER CONSUMPTION USING POWER
RAIL INTERCEPTS IN GPU, CPU AND BEYOND

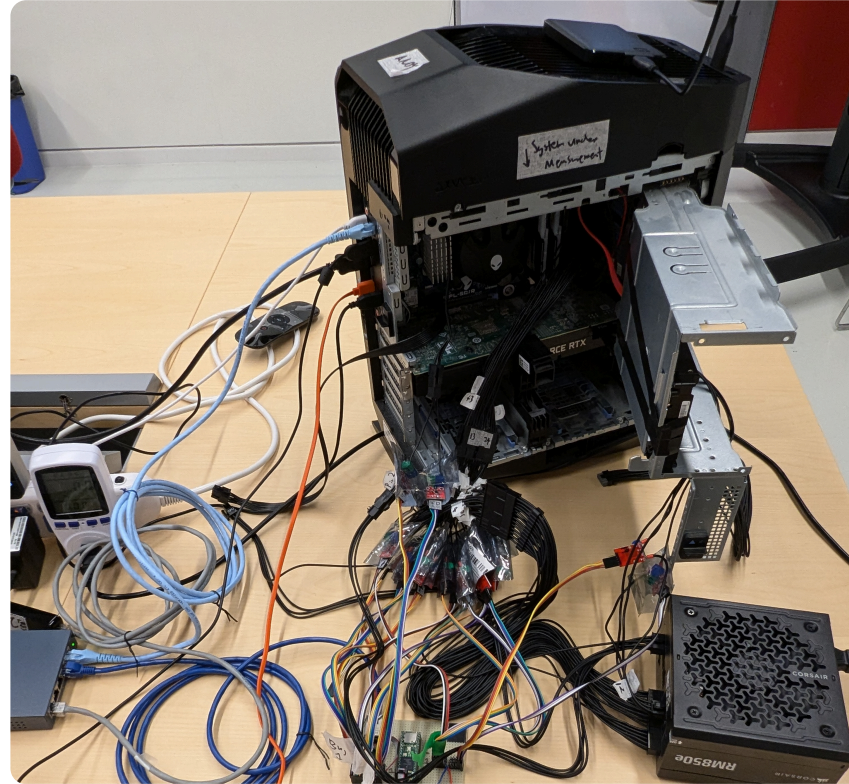
Andy Huo | Mooselab | July 10 2026

RECAP: WHAT WE DID ALREADY IN 2025



Implement a configuration agnostic energy measurement system for component level power attribution
Precision sensor calibrated with laboratory-grade power analyzer orchestrated with a high speed micro controller
Integrated into TraceCompass for multi-level correlation

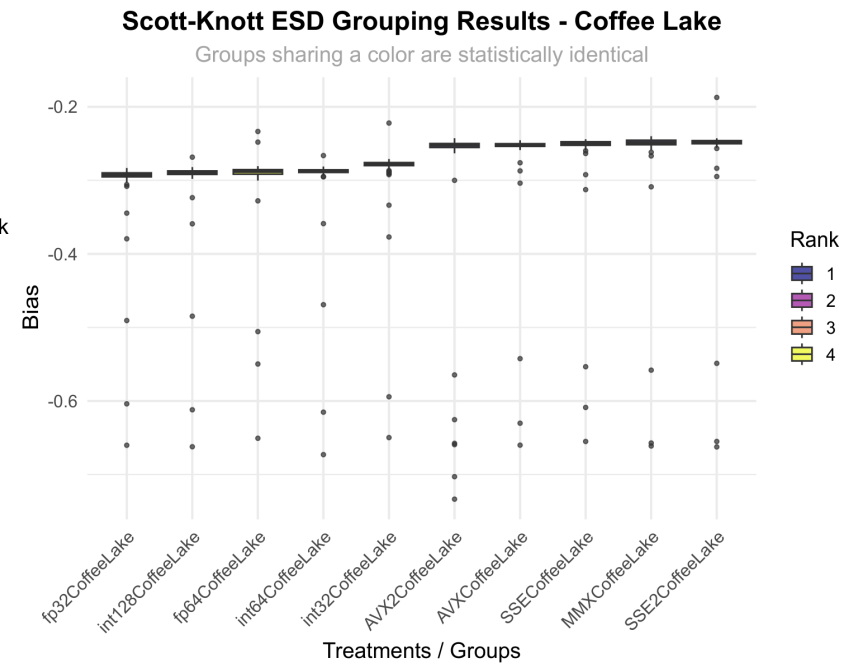
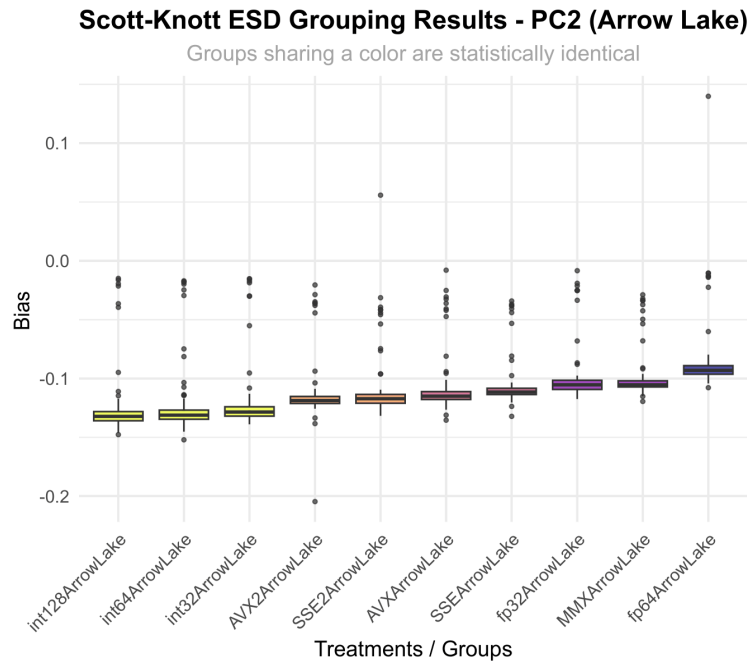
SOME UPGRADES IN 2026



Remote orchestration:

Teensy, SUS programs, Logger can all communicate to each other over Ethernet

WHAT WE FOUND

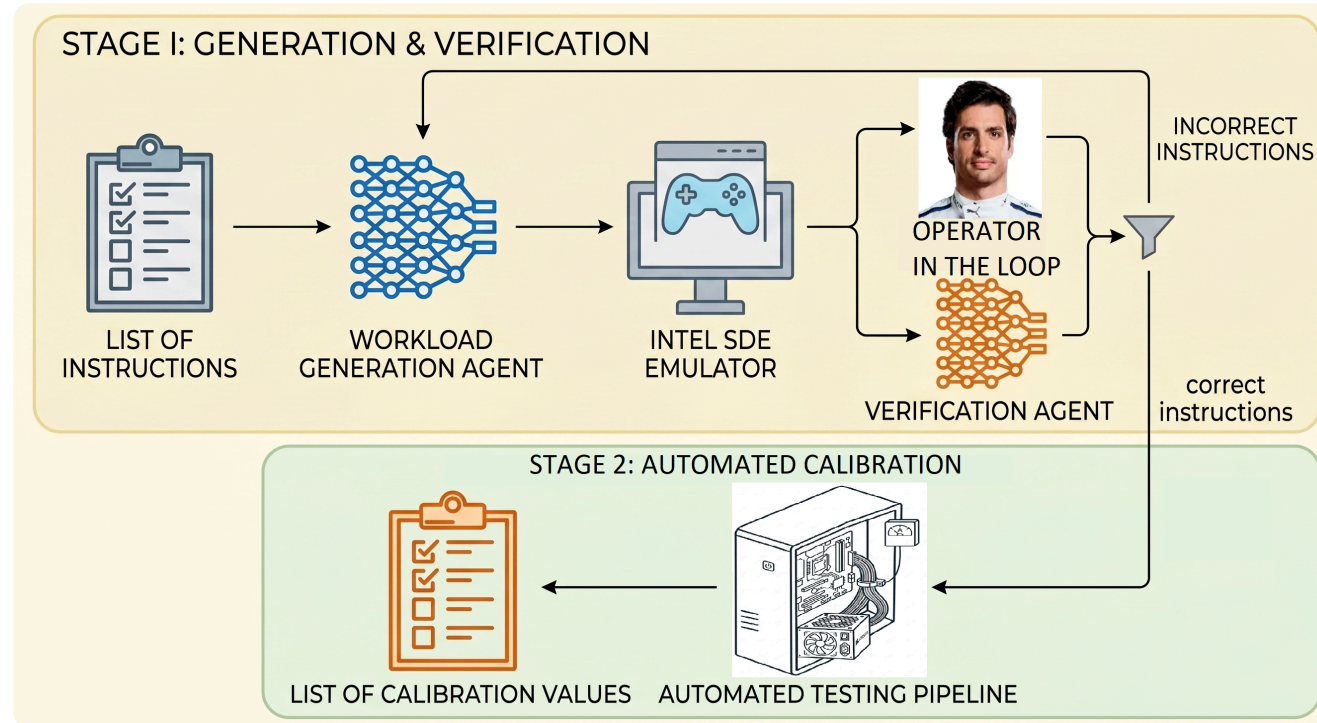


RAPL underestimations are idiosyncratic:
Bias varies depend on Generation, Architecture, Instructions
Complicates calibration

CHALLENGES

- The search space of all CPU instructions is huge
- Instructions vary based on Compiler setting, hardware configuration and Cpu generation.
- How do we produce a CPU-agnostic calibration for RAPL?

THE SOLUTION: AUTOMATED PIPELINE FOR PER-INSTRUCTION RAPL CALIBRATION



By leveraging a semi-automated pipeline, calibration can be obtained with reasonable effort.

VERIFICATION: TRUST, BUT VERIFY THE WORKLOAD

- LLM + human review each workload
- Intel SDE: verify intended instructions are compiled
- Maximum trust and minimal labor

TOOLING MORE TOOLS



RTX 5090 = washing machine
Yet its energy reporting interface is not particularly well studied.

WHAT IS NVML?



NVIDIA Management Library

Software Interface for reporting Power, usage and other GPU telemetry

GPU DIRECTION I: NVML ACCURACY

- Map power distribution within NVIDIA GPUs and correlate with physical rails
- Tensor core vs. CUDA core energy cost per FLOP, and how faithfully NVML reflects it
- Test NVML accuracy under bursty workloads (short kernels defeat coarse sampling)

GPU DIRECTION II: OPTIMIZATION STRATEGIES AND ENERGY

- Characterize NVML per-FLOP accuracy as a function of instruction set
- Effect of structured sparsity on energy efficiency — real savings or only speed?
- Whether NVML accuracy itself shifts under sparse vs. dense execution

GPU DIRECTION III: SYSTEM-LEVEL ENERGY MODELS

- Capacity-to-power curve: relate GPU occupancy / utilization to actual power drawn
- Is GPU power energy-proportional, or is there a high fixed floor?
- Cost of data movement on a per-GB basis across the hierarchy (L2, HBM, NVLink, host)

SPECIAL THANKS 🤝🤝

- Professor Heng Li (Polytechnique Montreal)
- Matthew Khouzam (Ericsson)
- The Ericsson community
- The Polytechnique community

CHECK OUT MY LTTNG MCP FOR YOUR LLM AGENTS



QUESTIONS?



Thank you for listening!
Follow-up: hhuo2000@gmail.com