Hewlett Packard Enterprise

CUG2024
Diverse Universe
PERTH AUSTRALIA

EVEREST: AN EFFECTIVE AND VERSATILE RUNTIME ENERGY SAVING TOOL

Sanyam Mehta (Ph.D.), Anna Yue, Torsten Wilde (Dr. rer. nat.), Steven Martin

Presenter: Barbara Chapman (Ph.D.)

Mai, 2024

INTRODUCTION

- Power and energy consumption continue to increase worldwide, especially with surge in AI
  - Data center energy consumption has grown 20-40% annually
  - Electricity consumption from data centres, artificial intelligence (AI) and the cryptocurrency sector could double by 2026*

- Power/energy saving opportunities exist for CPUs and GPUs
  - Using Dynamic Voltage and Frequency Scaling (DVFS) at run-time
  - Requires workload characterization to quantify impact on performance
  - Huge opportunity in GPUs
  - Lack of tools to manage CPU/GPUs considering performance/power/energy tradeoffs

*Executive summary – Electricity 2024 – Analysis – IEA: https://www.iea.org/reports/electricity-2024/executive-summary

REQUIREMENTS

- Optimize (reduce) power and/or energy consumption with minimal performance impact
  - Provide a method to allow the specification of a maximum allowed performance loss
- User and application agnostic
  - Users should not need to provide any information about their code
  - Applications should need to be changed
- Hybrid architecture support and vendor agnostic
  - Should work not only on CPUs but also on hybrid (CPU+GPU) architectures
  - Should work on devices from different silicon vendors (Intel, AMD, NVIDIA)
- Should not interfere with applications
  - Method should generate low overhead

BACKGROUND (CPUS VS. GPUS)

Feature	CPUs	GPUs
Memory BW		
Frequency-Power Profile		
Other Opportunities		

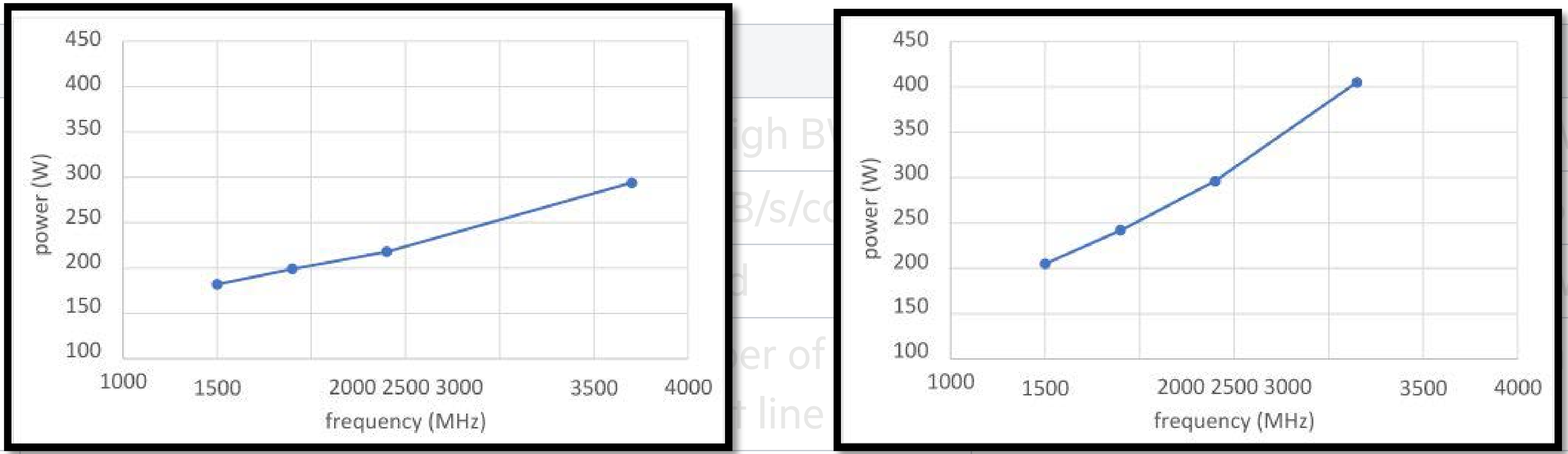
BACKGROUND (CPUS VS. GPUS)

Feature	CPUs	GPUs
Memory BW	Use DDR memory, does not feature high BW	Use HBM now with massive BW
	Latest high-end CPUs provision ~5 GB/s/core	Latest GPUs feature HBM3, >6x BW vs. CPUs
	Applications often memory BW bound	High memory BW reduces application bottleneck
	Implication: While a significant number of routines are memory bound on CPUs and can benefit from reduced clocks, GPUs need a different line of action.	
Frequency-Power Profile		
Other Opportunities		

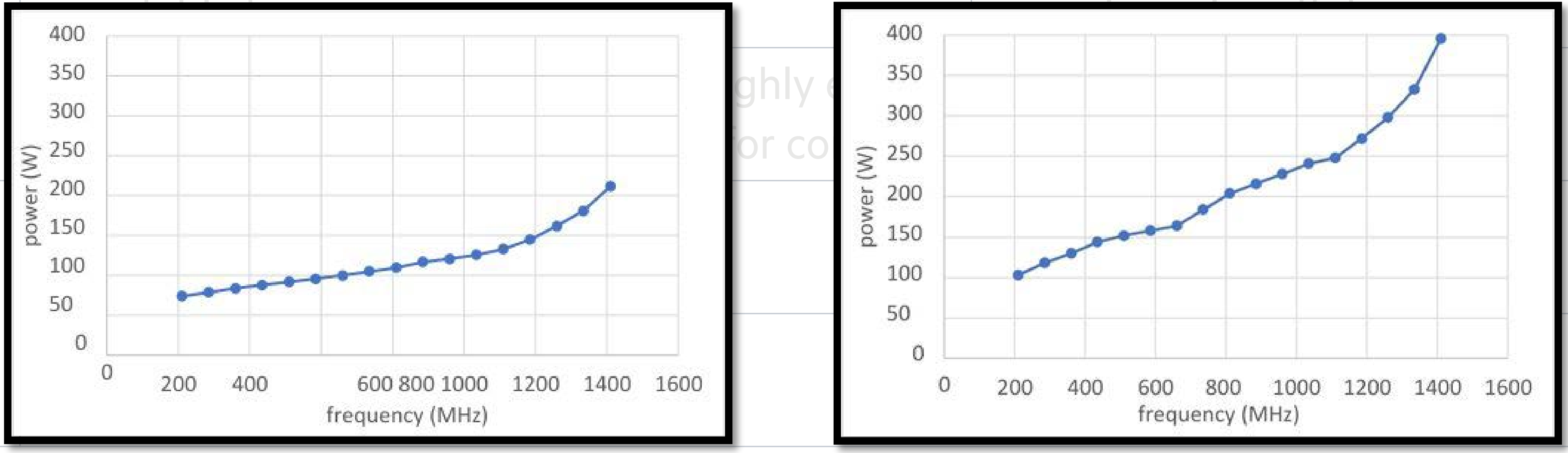
BACKGROUND (CPUS VS. GPUS)

Feature	CPUs	GPUs
Memory BW	Use DDR memory, does not feature high BW	Use HBM now with massive BW
	Latest high-end CPUs provision ~5 GB/s/core	Latest GPUs feature HBM3, >6x BW vs. CPUs
	Applications often memory BW bound	High memory BW reduces application bottleneck
	Implication: While a significant number of routines are memory bound on CPUs and can benefit from reduced clocks, GPUs need a different line of action.	
Frequency-Power Profile	Power increases linearly for memory bound apps, and super-linearly for compute bound apps with frequency	GPUs designed for maximum throughput; power increases super-linearly at higher frequencies
	Implication: unlike CPUs, GPUs are highly energy-inefficient at the top-end of their frequency range – something that could be exploited for considerable energy savings	
Other Opportunities		

BACKGROUND (CPUS VS. GPUS)



CPU frequency-power profile: memory bound (lbm, left) vs. compute bound (imagick, right) application



GPU frequency-power profile: HPC (GROMACS, left) vs. ML (BERT, right) application

BACKGROUND (CPUS VS. GPUS)

Feature	CPUs	GPUs
Memory BW	Use DDR memory, does not feature high BW	Use HBM now with massive BW
	Latest high-end CPUs provision ~5 GB/s/core	Latest GPUs feature HBM3, >6x BW vs. CPUs
	Applications often memory BW bound	High memory BW reduces application bottleneck
	Implication: While a significant number of routines are memory bound on CPUs and can benefit from reduced clocks, GPUs need a different line of action.	
Frequency-Power Profile	Power varies linearly for memory bound apps, and super-linearly for compute bound apps with frequency	GPUs designed for maximum throughput; power increases super-linearly at higher frequencies
	Implication: unlike CPUs, GPUs are highly energy-inefficient at the top-end of their frequency range – something that could be exploited for considerable energy savings	
Other Opportunities	Compute bound applications rarely access data beyond the L2 cache	GPUs might spend considerable time 'waiting' for work from CPUs
	Implication: Additional benefit possible from lowering uncore freq for compute bound phases on CPUs and lowering core freq in applications with low utilization on GPUs.	

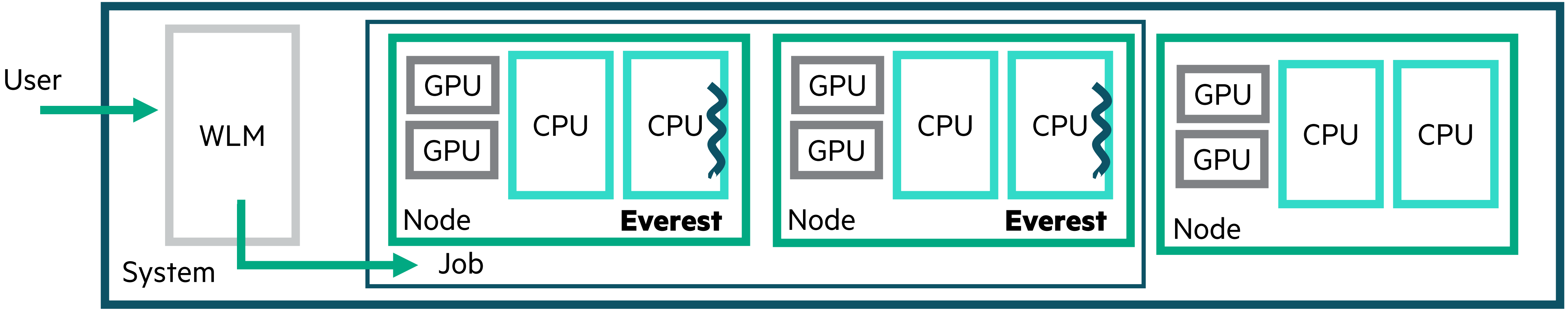
EVEREST

A PROOF OF CONCEPT (POC) PROTOTYPE FOR DYNAMIC ENERGY OPTIMIZATION OF WORKLOADS

EVEREST, AN EFFECTIVE AND VERSATILE RUNTIME ENERGY SAVING TOOL

EVeREST dynamically characterizes workloads with a lightweight and portable algorithm and uses DVFS to achieve power/energy savings while meeting a specified performance guarantee.

- Relies on only 2 metrics that are standard across all architectures
  - CPUs: Instructions Per Second (IPS)
  - GPUs: GPU Utilization



CPU APPROACH

- Goal: Predict application phase performance at different frequencies
- CPU Observations
  - When fully compute-bound, performance will vary proportionally with frequency
  - When fully memory-bound, performance does not change with frequency
  - Express this relationship as a formula:

$$\%CB = 100\% * \frac{\frac{IPS_{high}}{IPS_{low}} - 1}{\frac{Freq_{high}}{Freq_{low}} - 1}$$

$$\%MB = 100\% - \%CB$$

- Thus, when measuring IPS at a high frequency and at a low frequency, one can determine the compute- and memory-boundedness of an individual function (sensitivity analysis)

GPU APPROACH

- Profiling on GPUs associated with significant overhead (1.5x to >3x)
  - Stack walking, kernel serialization
  - Need alternative to directly measure performance used for CPUs

- GPU Observations
  - GPU utilization is a metric directly available without profiling
  - In simple terms, Utilization can be expressed as: (kernel runtime **K**, application wall clock time **WCT**)

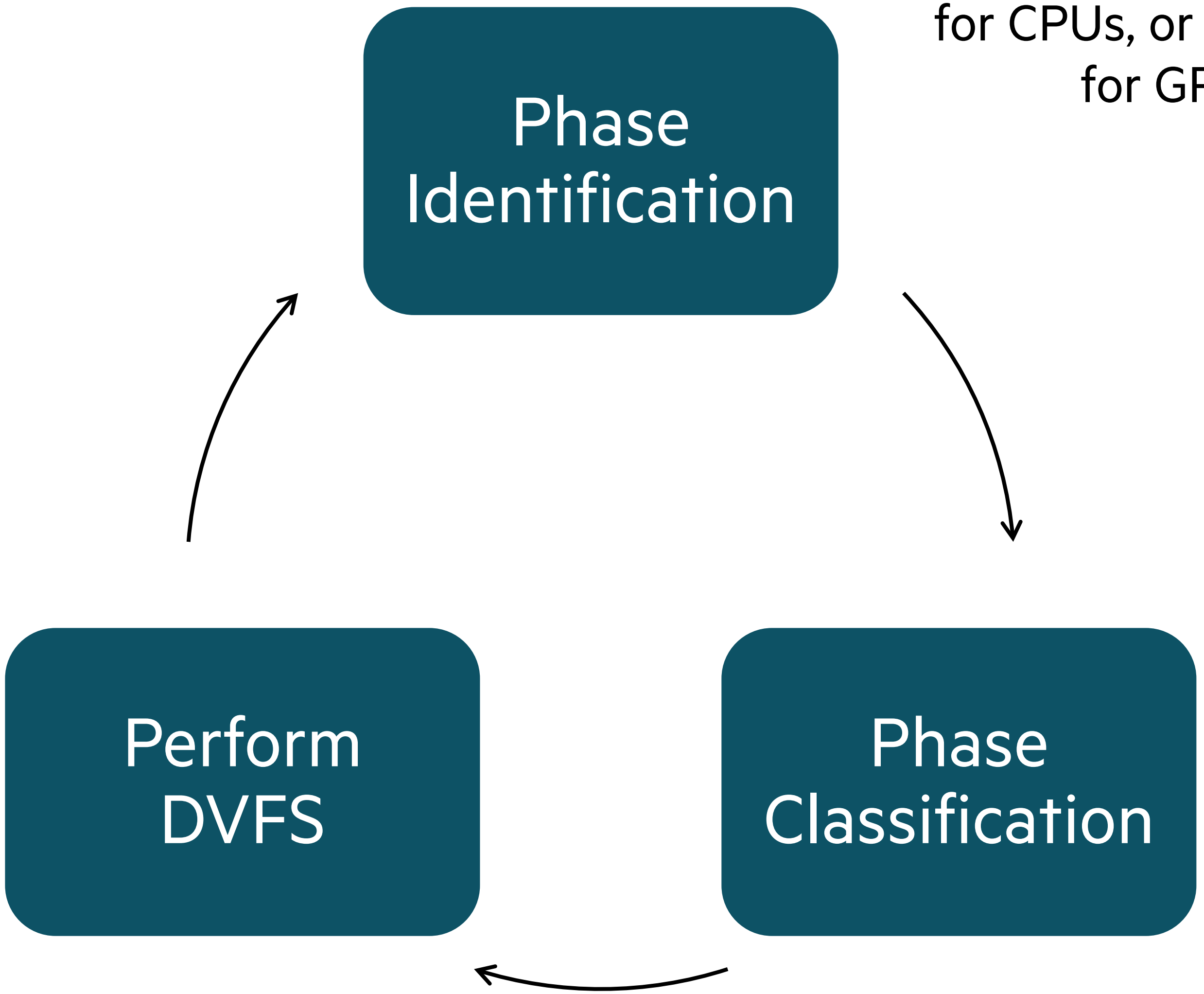
$$Utilization = \frac{K}{WCT}$$

  - Many applications overlap GPU kernel execution with CPU code or memory transfers between device and host.
    - Application performance may become limited by either the CPU or the memory transfer time and not the GPU.
    - When clock reduces, K increases. If Utilization also increases proportionally to K, it implies WCT is independent of GPU clock.

- Thus, like CPUs, if we measure Utilization at a high frequency and at a low frequency, then we can predict application performance.

HIGH-LEVEL CONTROL FLOW

Using instruction pointer for CPUs, or utilization for GPUs



Characterize how performance affected by change in frequency

Using performance target and classification

EVEREST RESULTS

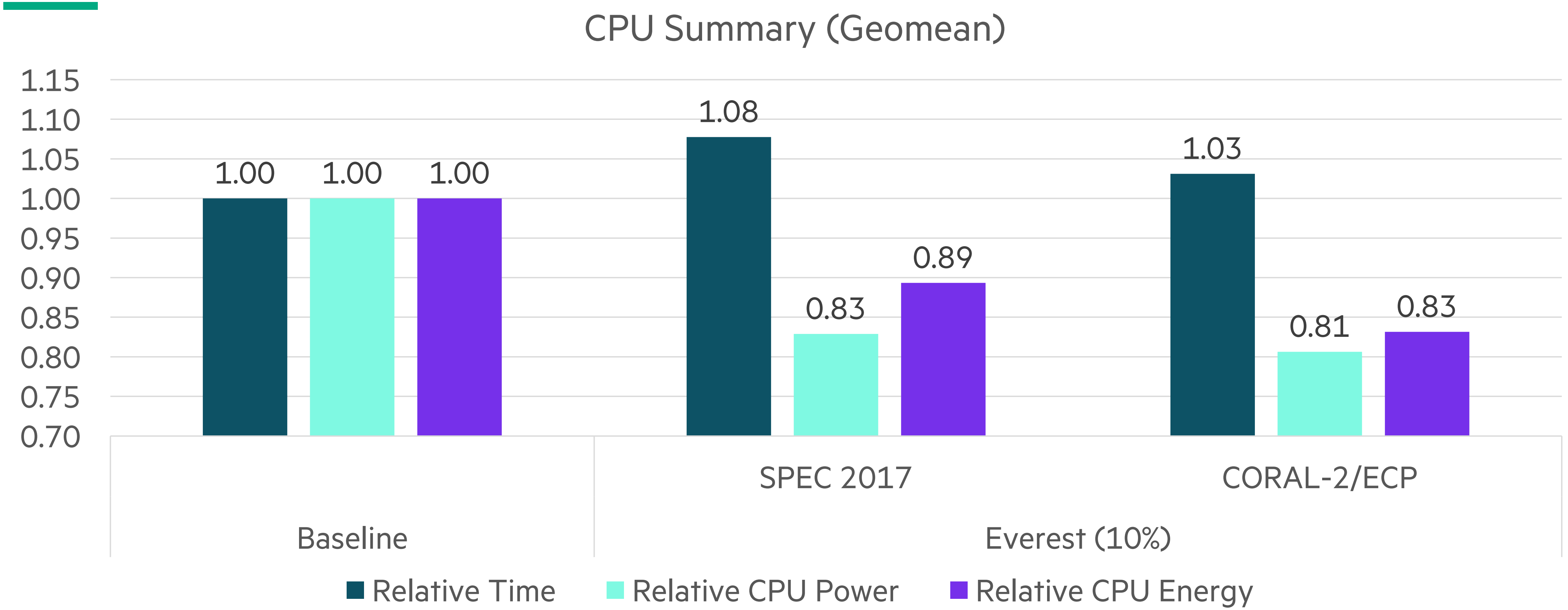
RESULTS

- Evaluation on latest generation CPUs and GPUs – Intel Sapphire Rapids, AMD Genoa, NVIDIA A100
  - 27 CPU apps: 22 from SPEC 2017, 5 from CORAL-2/ECP
  - 6 (9) GPU apps: 3 (6) from HPC, 3 from AI/ML

- Evaluated at different levels of acceptable performance loss (5%, 10%, and 20%)

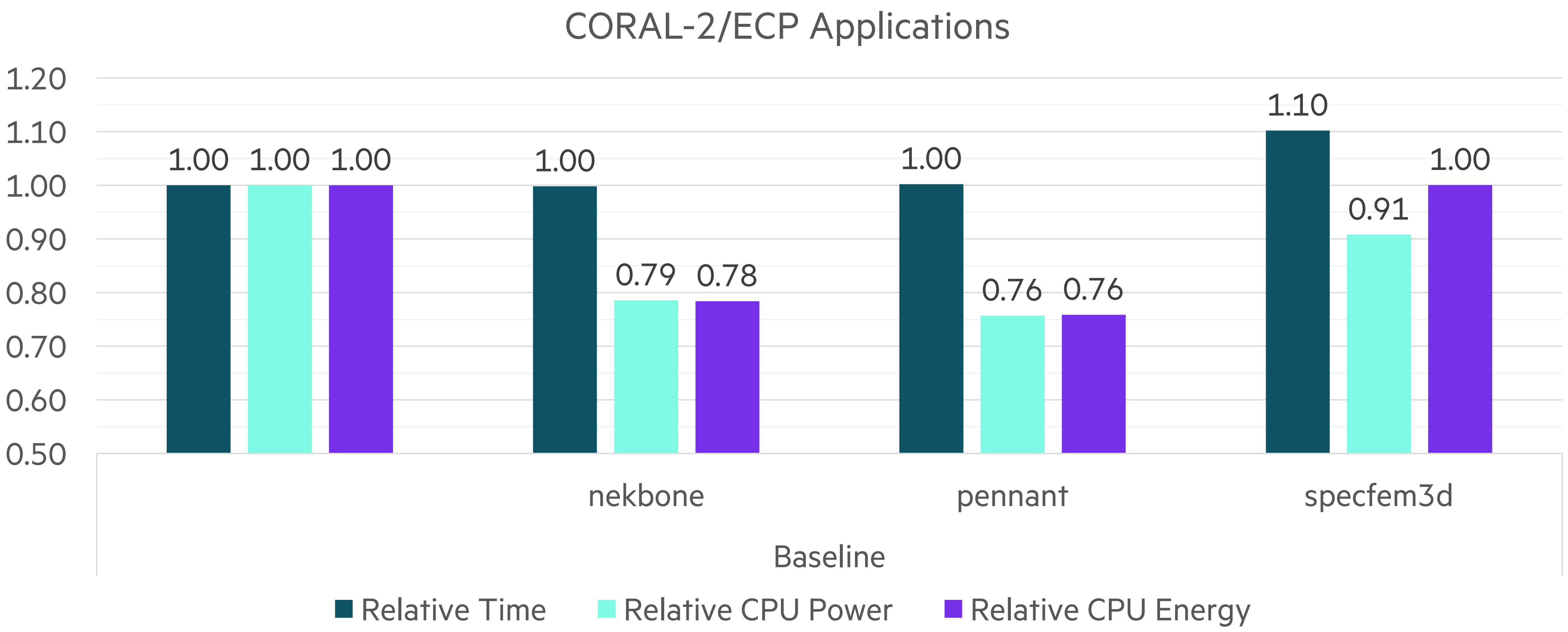
- Usage:
  - User submits job with additional parameter for acceptable performance loss
  - srun .... --use-everest:pd ...
    - Users can specify the maximal performance reduction they are willing to incur
    - Does not require modifying the application source
    - Does not depend on a specific compiler and MPI

CPU RESULTS



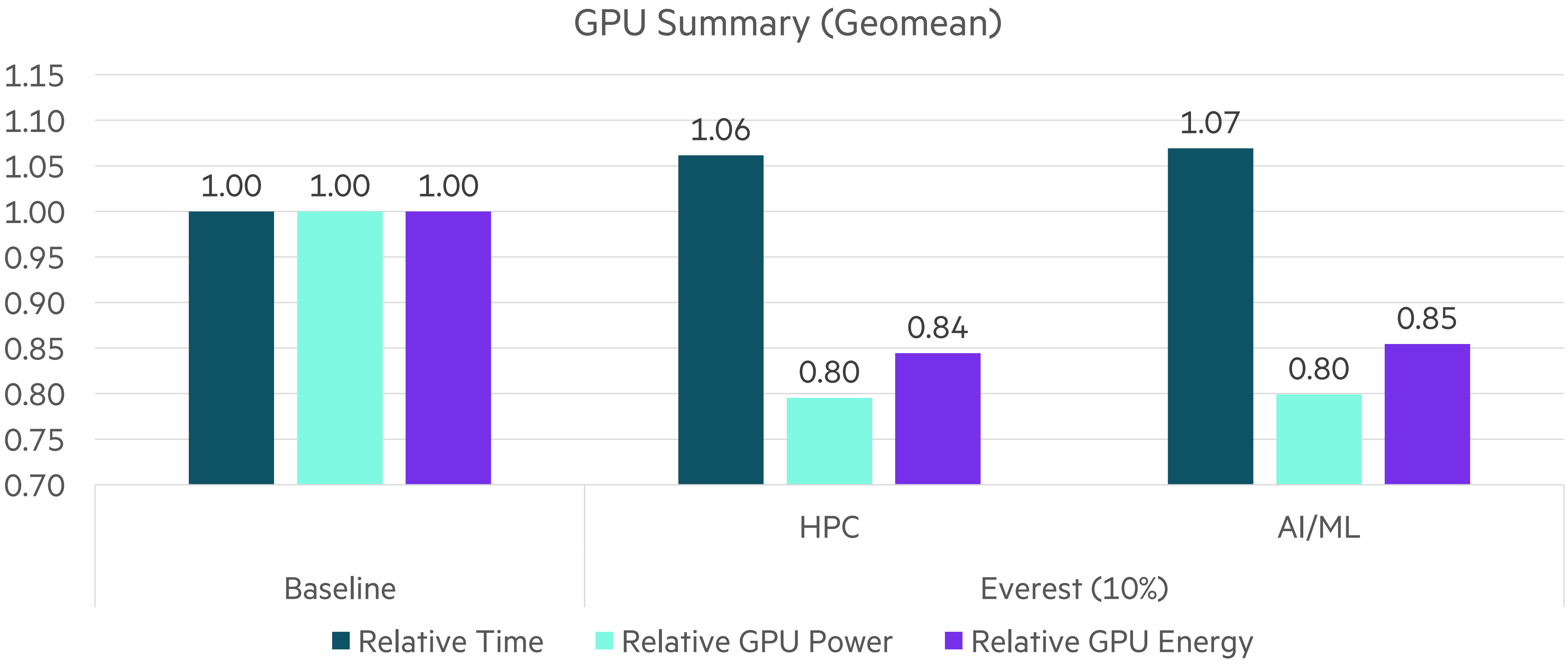
Memory-bound applications provide opportunities for 20-30% energy savings at minimal performance loss, while compute-bound applications can still achieve power savings proportional to the acceptable performance loss.

CPU HIGHLIGHTS – MPI WORKLOADS



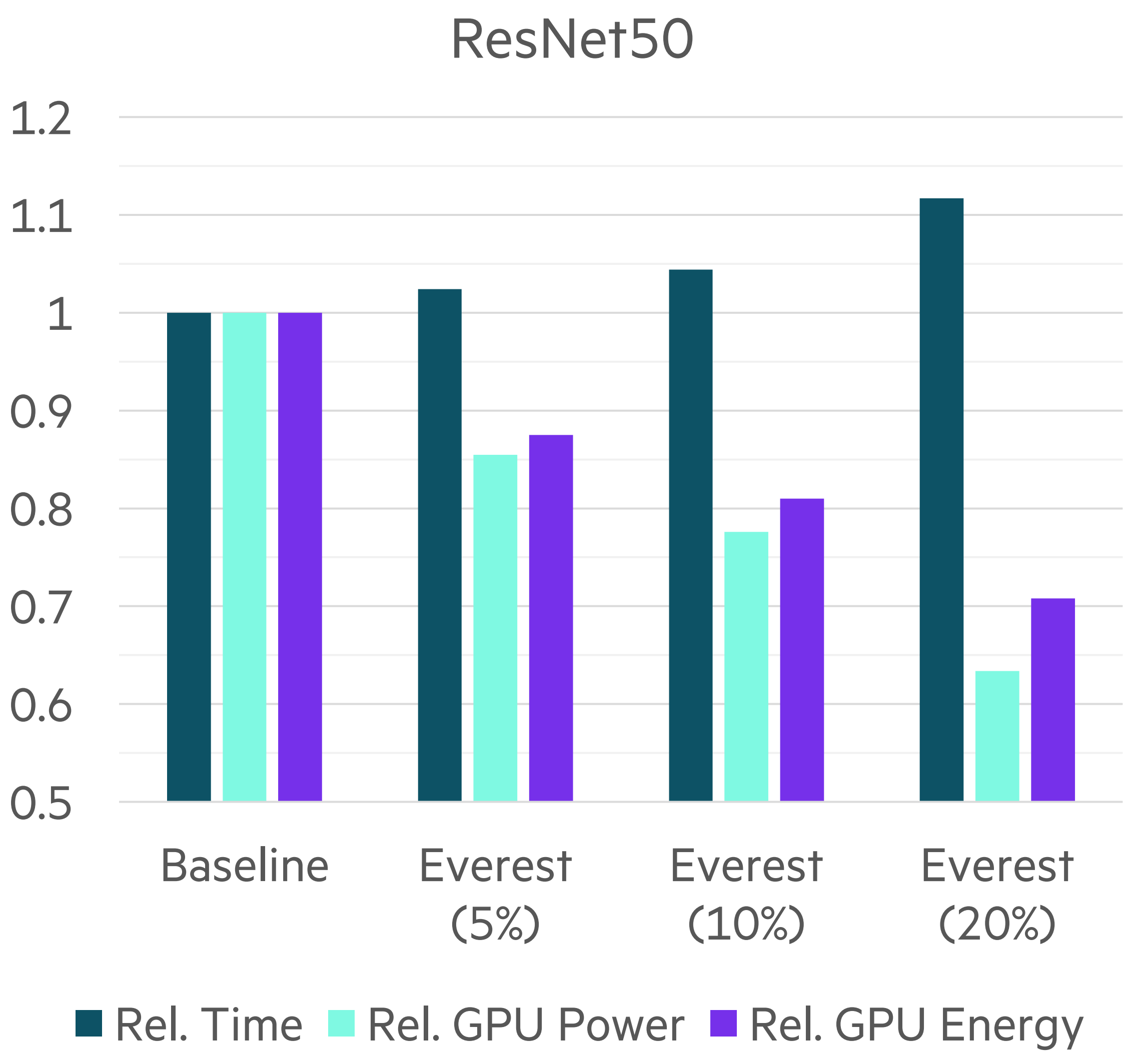
Everest can exploit any opportunities to save power and energy during intensive communication phases.

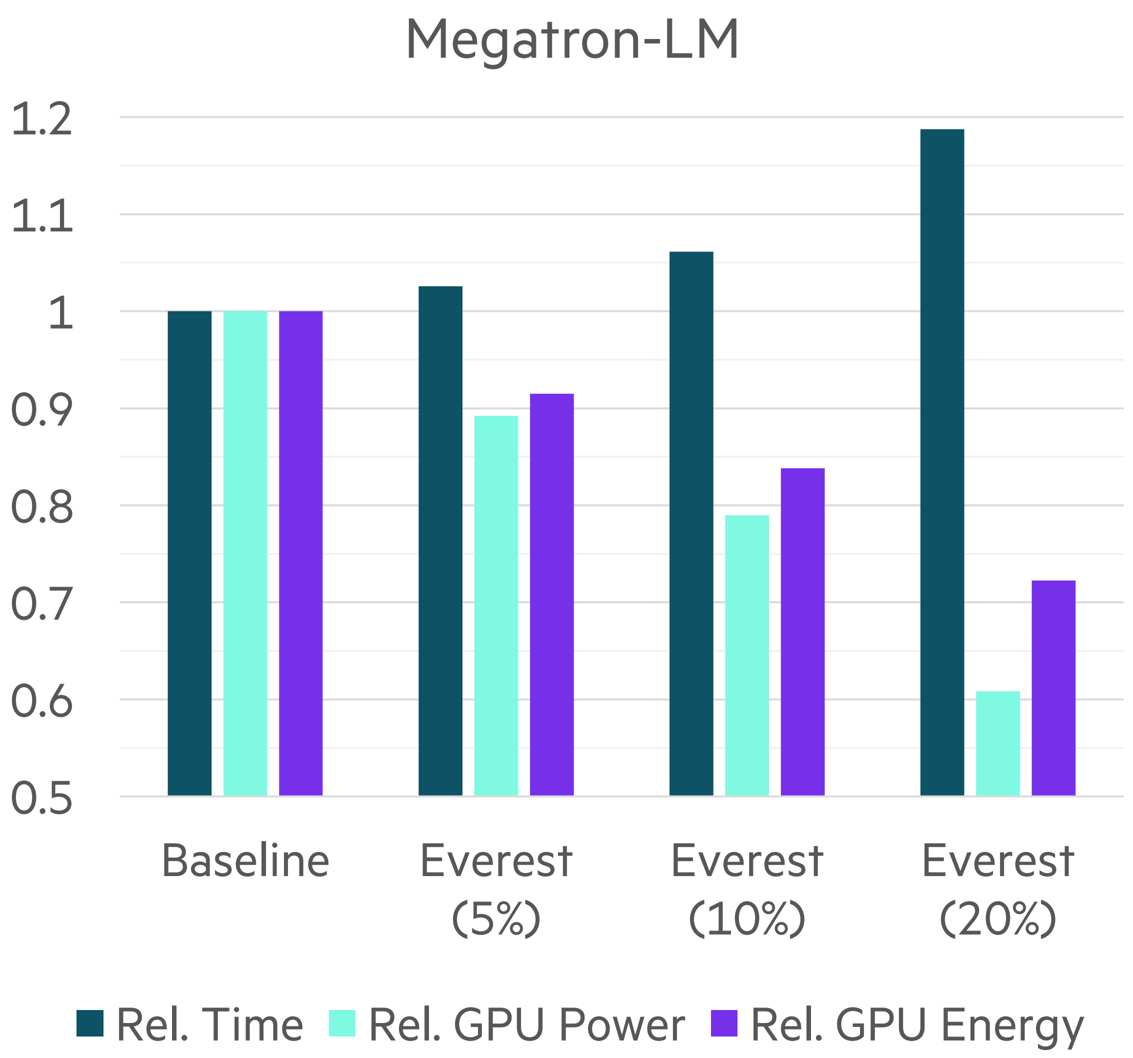
GPU RESULTS



Everest can provide significant power and energy savings for GPUs.

GPU HIGHLIGHTS – AI WORKLOADS





CONCLUSION

- Lightweight solution for dynamic optimization of application according to power/performance/energy tradeoffs
  - Huge opportunity with GPUs

- Compute vendor agnostic
  - Works for CPUs and GPUs of different vendors

- Portable
  - runtime-only, integration with user code not required

- Phase awareness
  - Can extract maximum power/energy savings without requiring user input

- Opportunity for collaboration and influencing product roadmap

THANKS

SANYAM.MEHTA@HPE.COM
WILDE@HPE.COM