SC25
St. Louis, MO | hpc ignites.

16th International Workshop on

Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems

Characterizing the Impact of GPU Power Management on an Exascale System

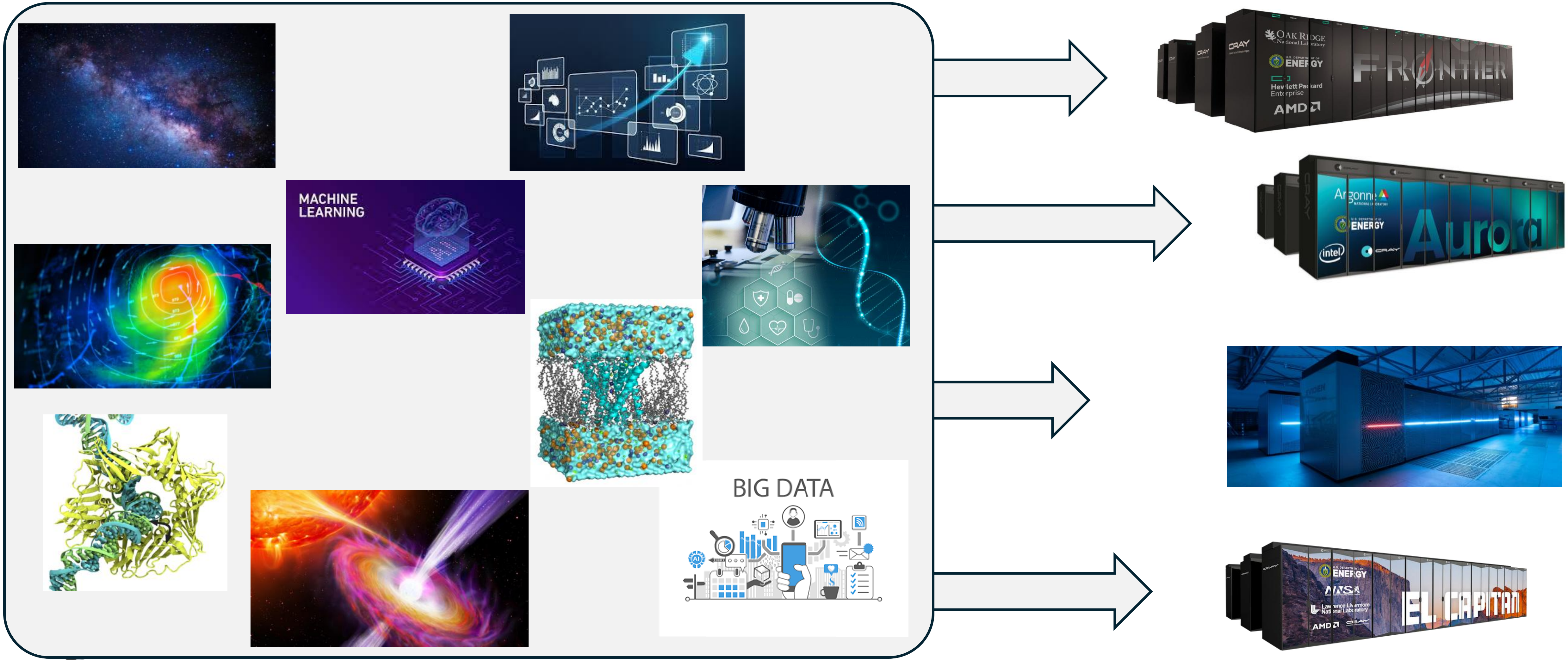
Mariana Costa
Phillipe Navaux
Arthur Lorenzon

Bruno Alvarez
Jordà Polo

Antigoni Georgiadou
James B. White III
Woong Shin
Bronson Messer

GPPD/UFRGS
.Inf INSTITUTO DE INFORMÁTICA UFRGS PPGC
UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
AMD
OAK RIDGE National Laboratory

Motivation


MACHINE LEARNING
BIG DATA
CRAY
OAK RIDGE National Laboratory
ENERGY
Hewlett Packard Enterprise
AMD
FRONTIER
Argonne
ENERGY
intel
CRAY
Aurora
ENERGY
NNSA
Lawrence Livermore National Laboratory
AMD
CRAY
EL CAPITAN

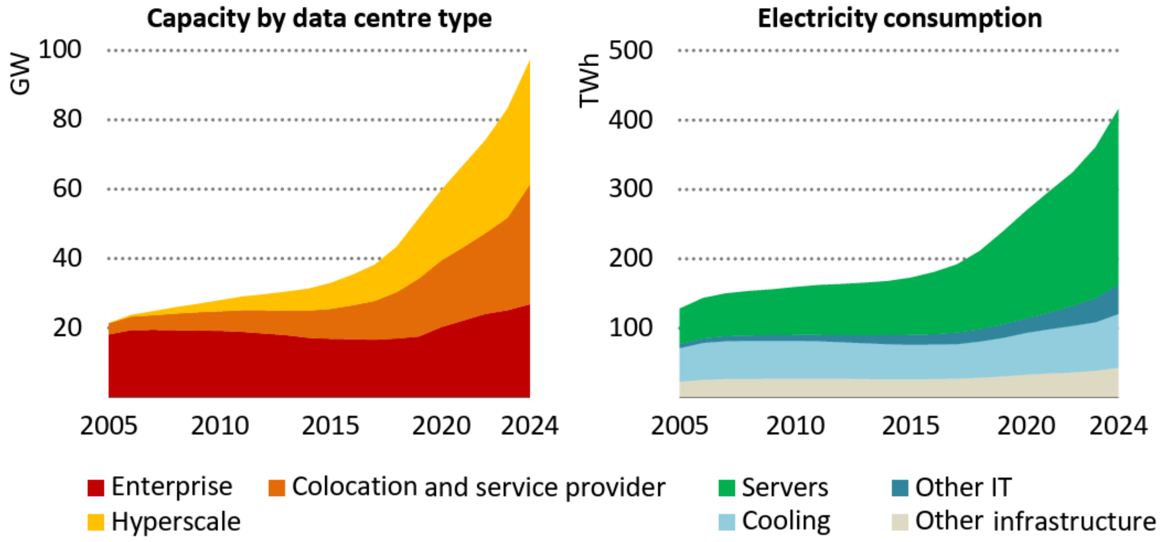
Motivation

Search everything

Global electricity demand to keep growing robustly through 2026 despite economic headwinds

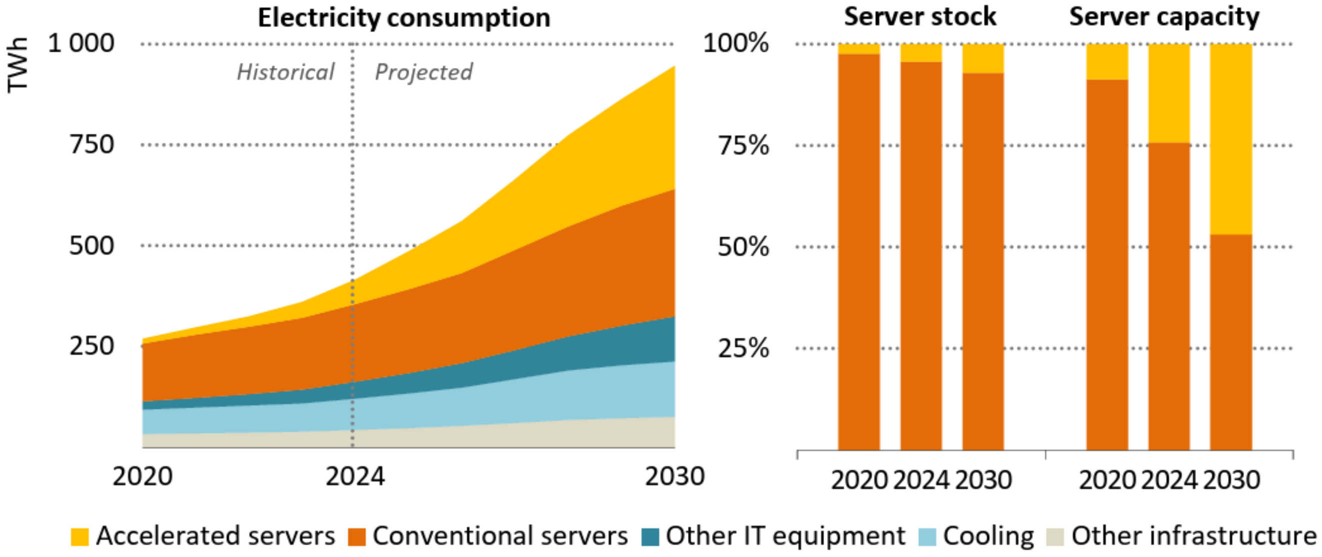
Press release
30 July 2025

Figure 2.3 ▹ **Total data centre electricity consumption by equipment type and data centre type, 2005-2024**



IEA. CC BY 4.0.

After a decade of limited growth, data centre electricity consumption began to accelerate again after 2015

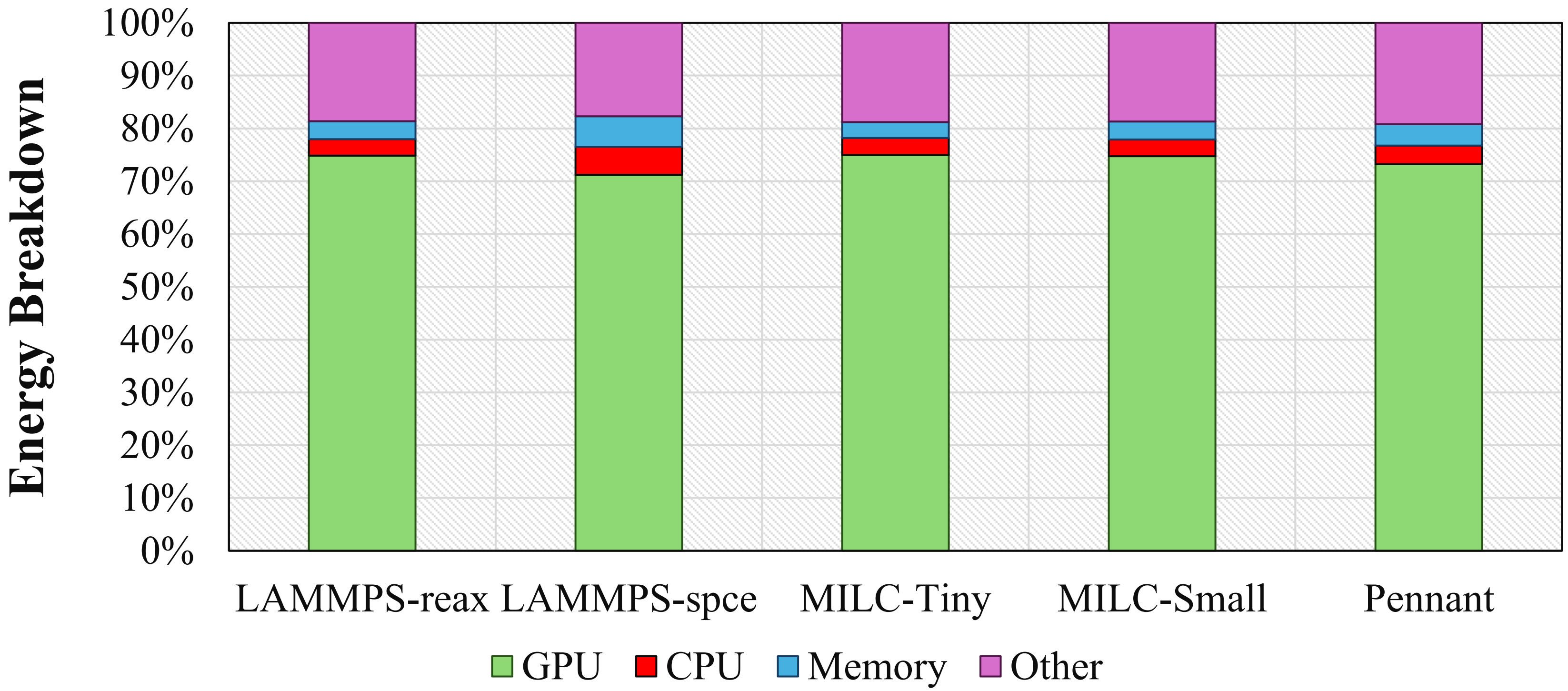


IEA. CC BY 4.0.

Around 70% of the growth in electricity demand from servers between 2025 and 2030 comes from accelerated servers

Motivation

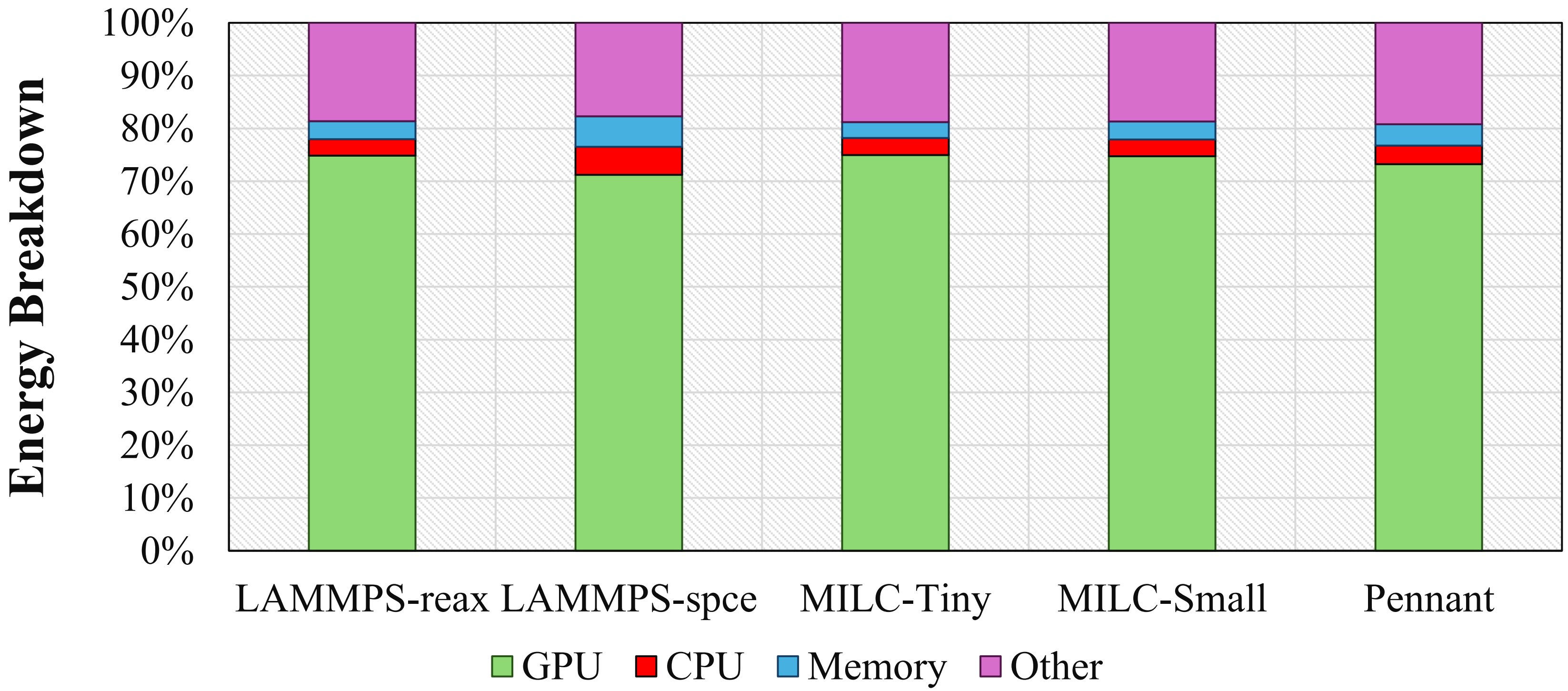
About 70% of system energy is consumed by GPUs (e.g., Frontier)



GPPD/UFRGS

Motivation

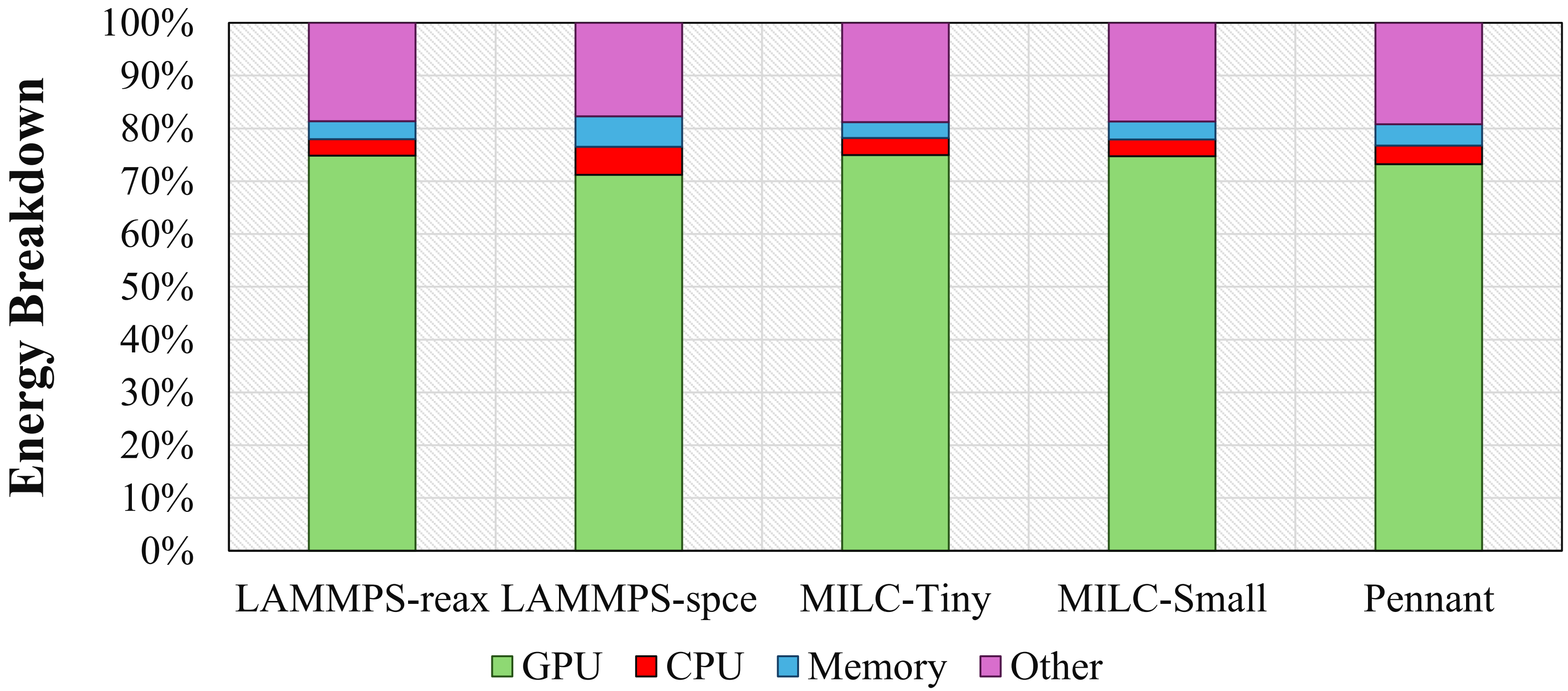
About 70% of system energy is consumed by GPUs (e.g., Frontier)



Improving GPU energy efficiency is key to sustainable HPC

Motivation

About 70% of system energy is consumed by GPUs (e.g., Frontier)



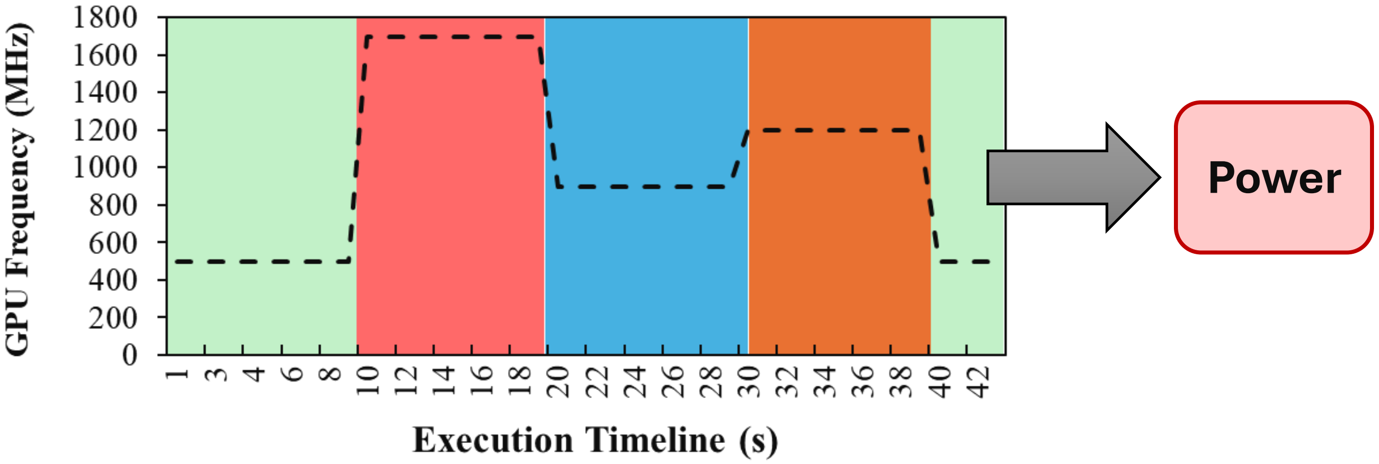
Improving GPU energy efficiency is key to sustainable HPC

Power management strategies can be employed to optimize energy efficiency

GPPD/UFRGS

Motivation: Power Management

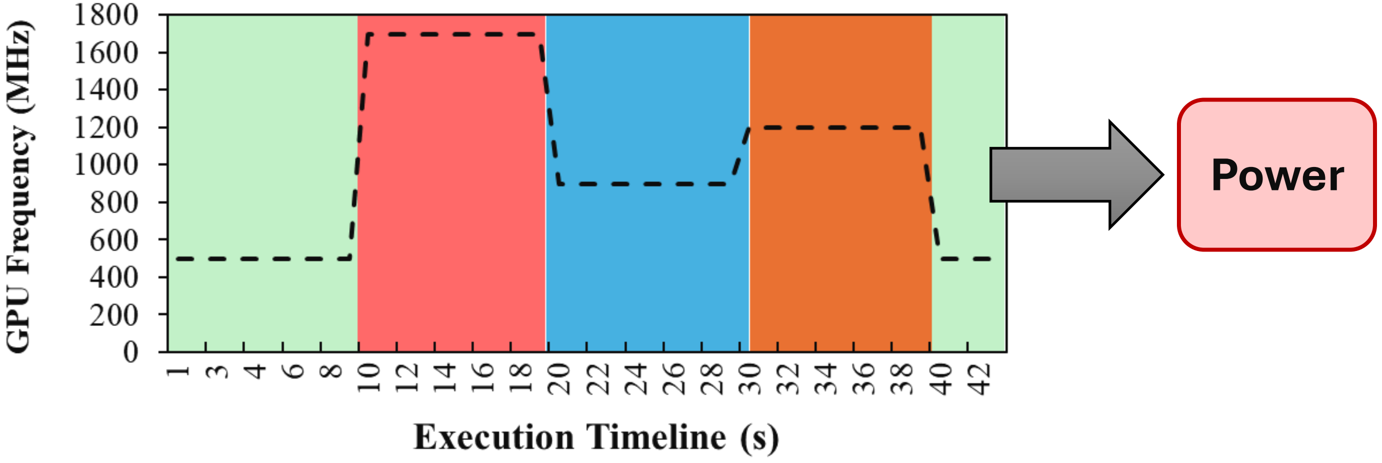
- **Frequency Capping (DVFS):**
    - Limits GPU clocks to a fixed upper bound
    - Proactive/static control
    - Predictable performance, easy to reproduce



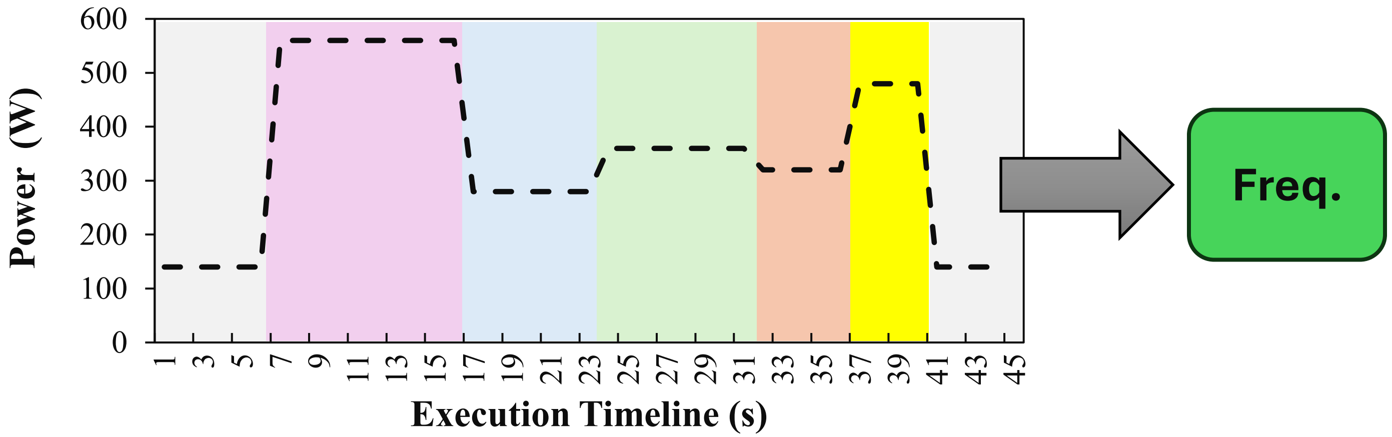
Power

Motivation: Power Management

- **Frequency Capping (DVFS):**
  - Limits GPU clocks to a fixed upper bound
  - Proactive/static control
  - Predictable performance, easy to reproduce



- **Power Capping:**
  - Sets a maximum GPU power budget
  - Reactive/adaptive control
  - Frequency adjusts with workload intensity
  - Less predictable

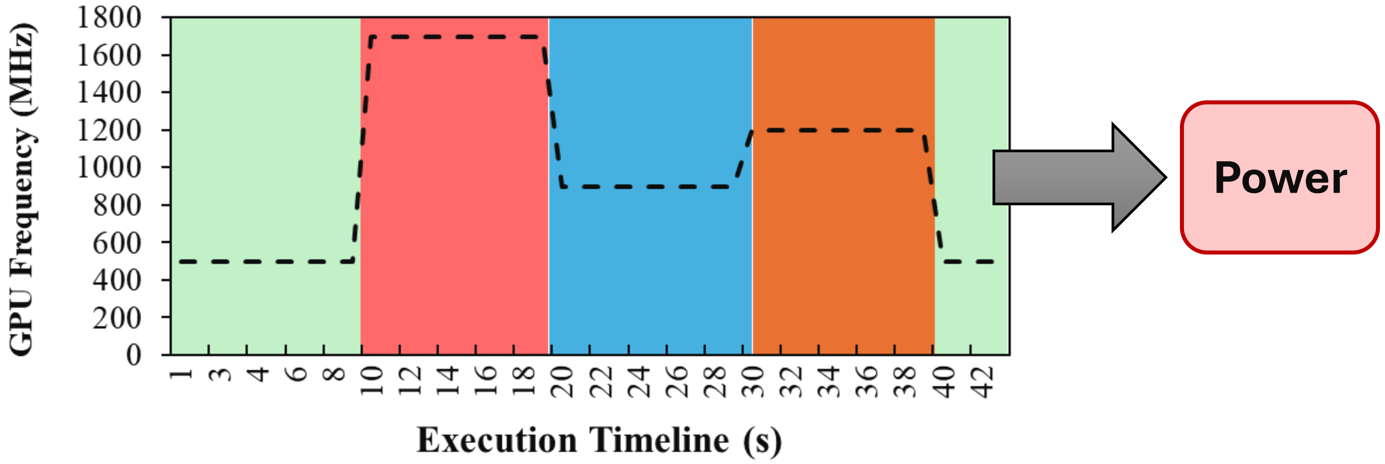


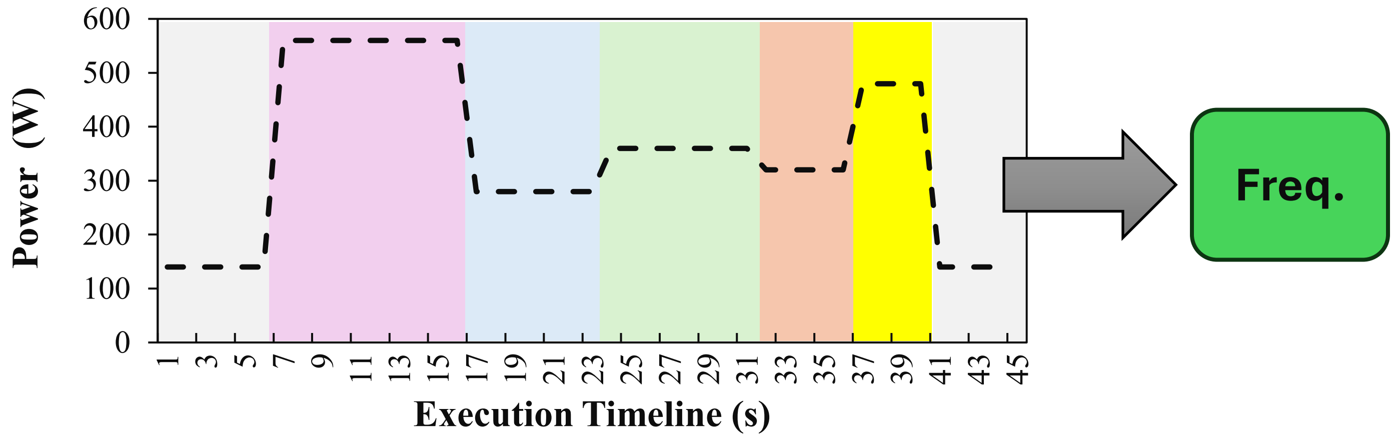
Motivation: Power Management

- **Frequency Capping (DVFS):**

Which power management strategy delivers the best performance-energy efficiency for a given application?

  - Frequency adjusts with workload intensity
  - Less predictable





Contributions

- Performance and energy benchmarking of power and frequency management strategies

- Characterize how these techniques affect:
  - Runtime
  - Energy-to-solution
  - Energy efficiency

Agenda

- Methodology
- Evaluation
- Concluding Remarks

Agenda

- Methodology

Methodology: Benchmarks

- Seven GPU-accelerated applications:
  - **Cholla:** Memory-intensive astrophysics hydrodynamics code
  - **HACC:** Cosmology simulation sensitive to memory bandwidth and compute throughput
  - **Kripke:** Particle-transport proxy stressing memory access and spatial patterns
  - **LAMMPS:** Compute-intensive molecular dynamic code for materials simulation
  - **Pennant:** Unstructured-mesh hydrodynamics proxy sensitive to memory and compute balance
  - **PortUrb:** Urban flow simulation limited by memory bandwidth and large-array operations
  - **QuickSilver:** Monte Carlo transport proxy with irregular control flow and latency-bound access

Methodology: Benchmarks

- Benchmarks exhibit different behaviors of FLOPs/byte, FLOPs/s, and L2 Cache HIT

	FLOPs/byte	FLOPs/s	L2 Cache Hit (%)
Cholla	0.62	6.58E+11	37.51
HACC	215.04	4.67E+12	84.02
Kripke	0.10	3.90E+10	62.68
LAMMPS	3.41	5.71E+12	51.78
Pennant	0.67	5.57E+11	45.11
PortUrb	11.45	1.08E+12	61.32
QuickSilver	1.83	1.61e+10	74.72

Methodology: Benchmarks

- Benchmarks exhibit different behaviors of FLOPs/byte, FLOPs/s, and L2 Cache HIT

	FLOPs/byte	FLOPs/s	L2 Cache Hit (%)
Cholla	0.62	6.58E+11	37.51
HACC	215.04	4.67E+12	84.02
Kripke	0.10	3.90E+10	62.68
LAMMPS	3.41	5.71E+12	51.78
Pennant	0.67	5.57E+11	45.11
PortUrb	11.45	1.08E+12	61.32
QuickSilver	1.83	1.61e+10	74.72

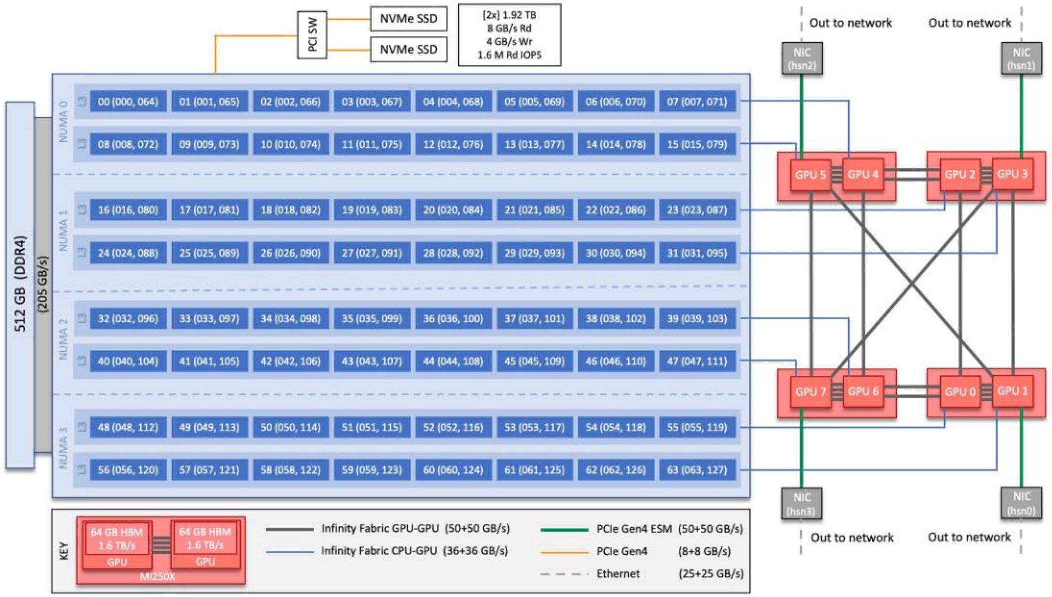
Methodology: Benchmarks

- Benchmarks exhibit different behaviors of FLOPs/byte, FLOPs/s, and L2 Cache HIT

	FLOPs/byte	FLOPs/s	L2 Cache Hit (%)
Cholla	0.62	6.58E+11	37.51
HACC	215.04	4.67E+12	84.02
Kripke	0.10	3.90E+10	62.68
LAMMPS	3.41	5.71E+12	51.78
Pennant	0.67	5.57E+11	45.11
PortUrb	11.45	1.08E+12	61.32
QuickSilver	1.83	1.61e+10	74.72

Methodology: Target Architecture

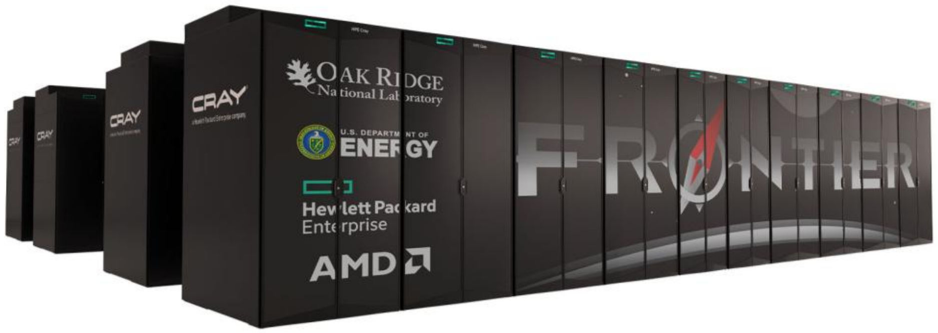
- 1 – 32 nodes from Frontier Supercomputer
  - 1x 64-core AMD Optimized 3rd Gen EPYC CPU
  - 4x AMD MI250X, each with 2 GCDs (total of 8 GCDs per node)



- GPU Operating Frequency:
  - 31 levels: 500 MHz, 540 MHz, … 1700 MHz.

- GPU Power Capping:
  - 21 levels: 140W, 160W, … 560W.

- Compilation process
  - AMD ROCm hipcc 6.2.4
  - Flags → `-O3` and `--ofload-arch=gfx90a`
  - Modules → `craype-accel-amd-gfx90a` and `rocm/6.2.4`



Methodology: Evaluated Metrics

- **Performance (Perf)**
  - FOM (when available)
  - Total execution time

- **Performance per Watt (Perf/Watt)**
  - Ratio between performance and the average GPU power draw during execution

- **Getting power, energy, and other system-level metrics:**
  - ***Omnistat***, an open-source, low-overhead monitoring tool
  - Configured using a runtime control file to enable the collection of core GPU metrics via the system-management interface
  - Aggregated metrics on a per-Slurm-job basis across all compute hosts.
  - Metrics collected every 0.1s
  - 10 executions, with std deviation < 2%

Agenda

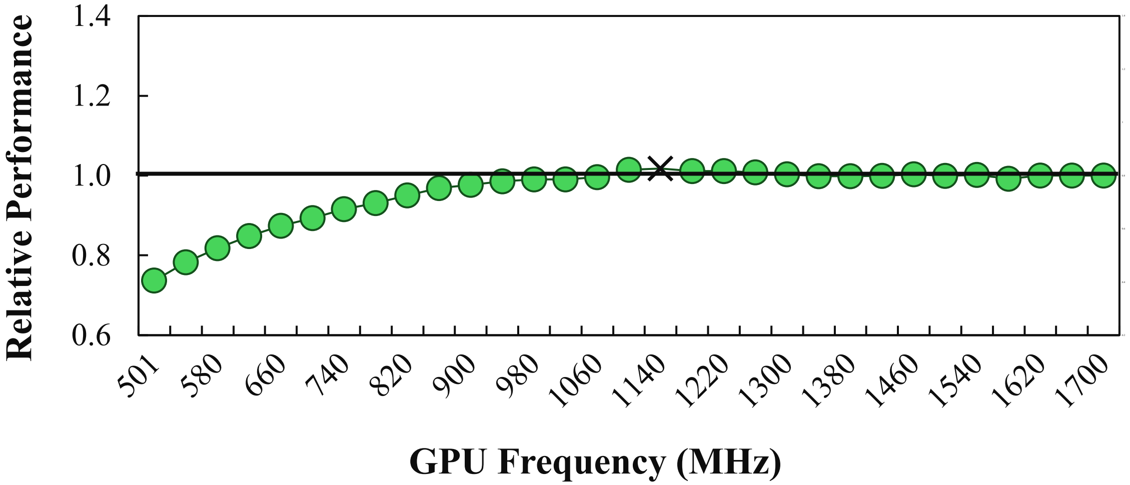
- ~~Methodology~~
- Evaluation

Evaluation

- What applications benefit more from Frequency Capping?

Evaluation: Apps that benefit more from Freq. Capping

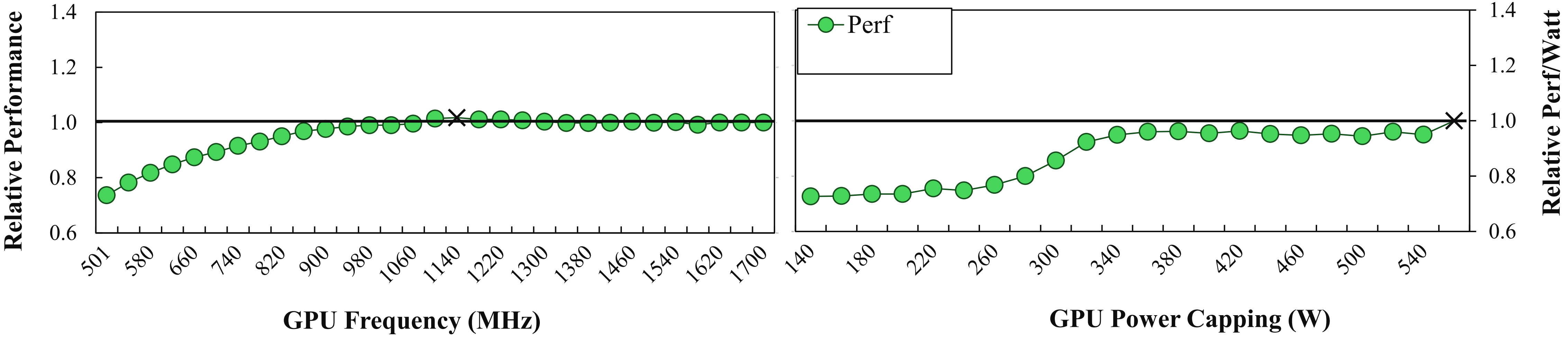
Cholla
1 node



GPPD/UFRGS

Evaluation: Apps that benefit more from Freq. Capping

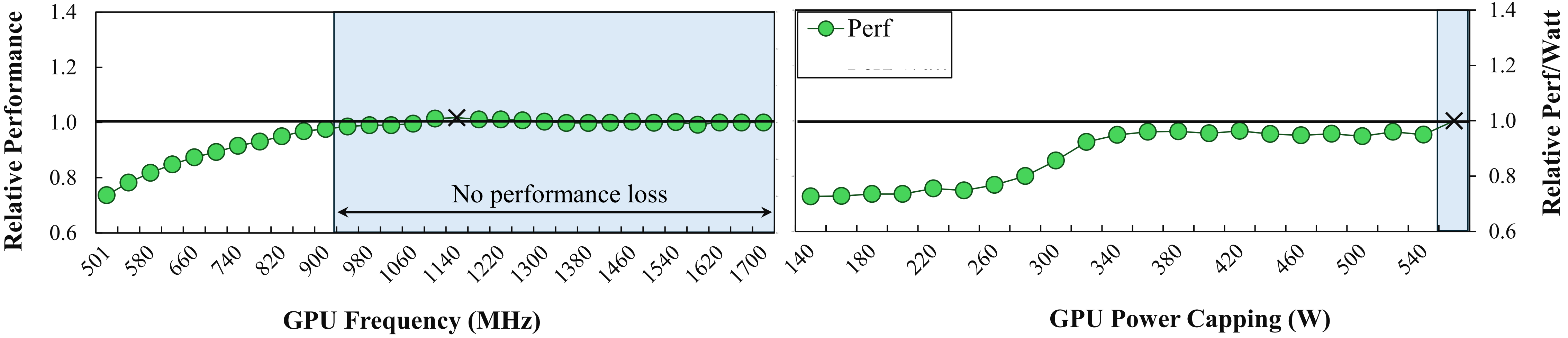
Cholla
1 node



GPPD/UFRGS

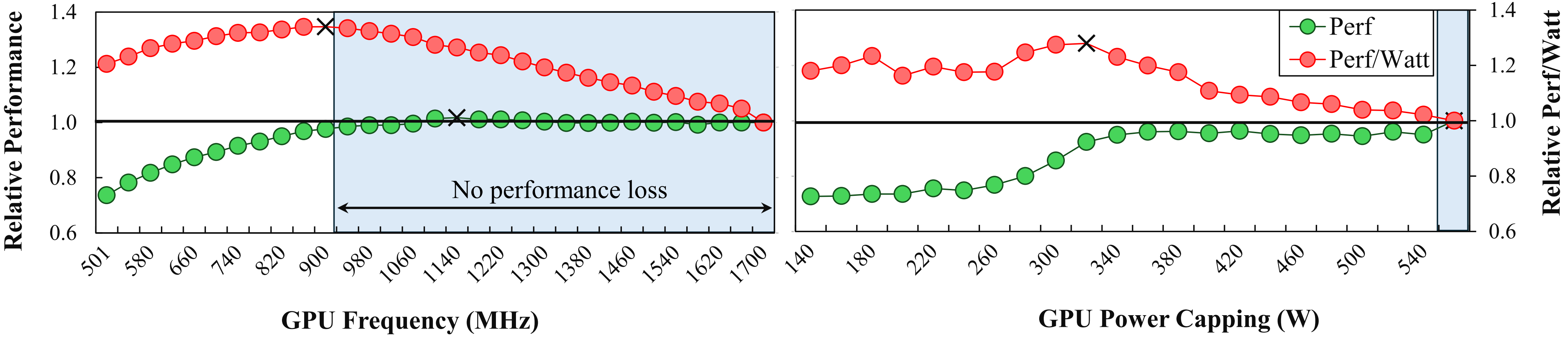
Evaluation: Apps that benefit more from Freq. Capping

Cholla
1 node



Evaluation: Apps that benefit more from Freq. Capping

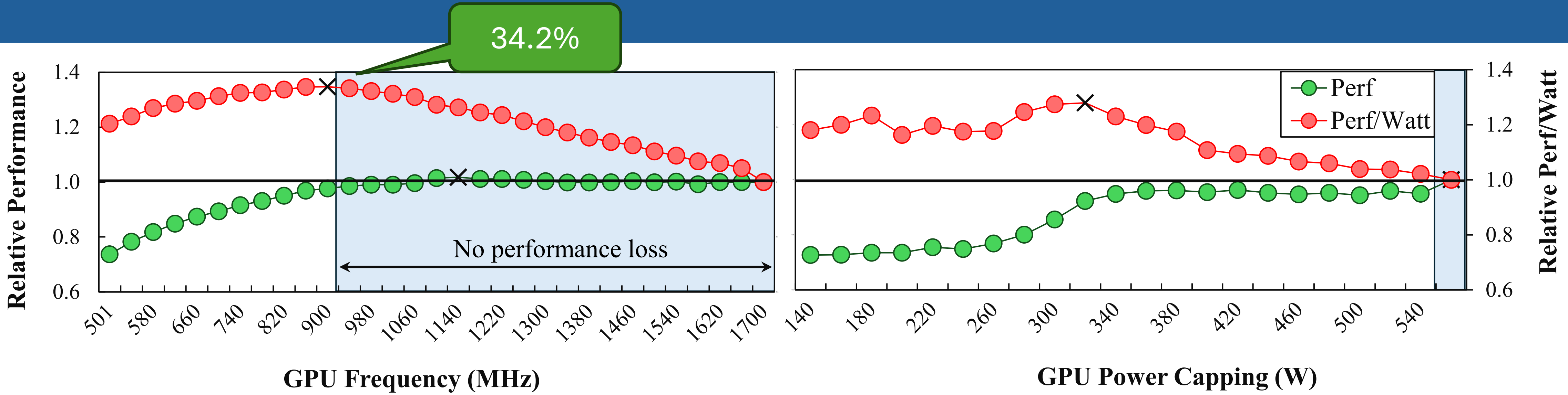
Cholla
1 node



GPPD/UFRGS

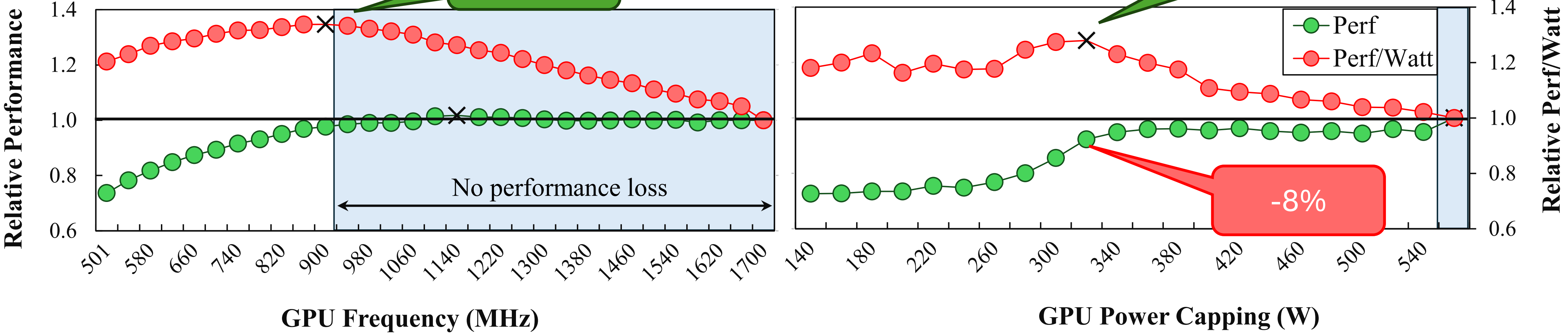
Evaluation: Apps that benefit more from Freq. Capping

Cholla
1 node



Evaluation: Apps that benefit more from Freq. Capping

Cholla
1 node



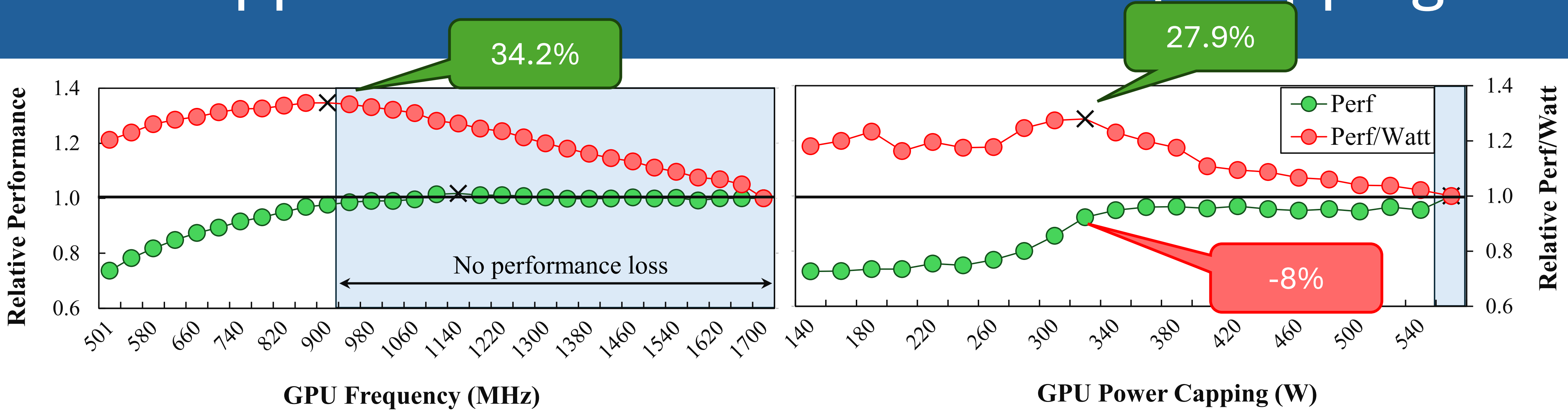
34.2%

27.9%

-8%

Evaluation: Apps that benefit more from Freq. Capping

Cholla
1 node



34.2%

27.9%

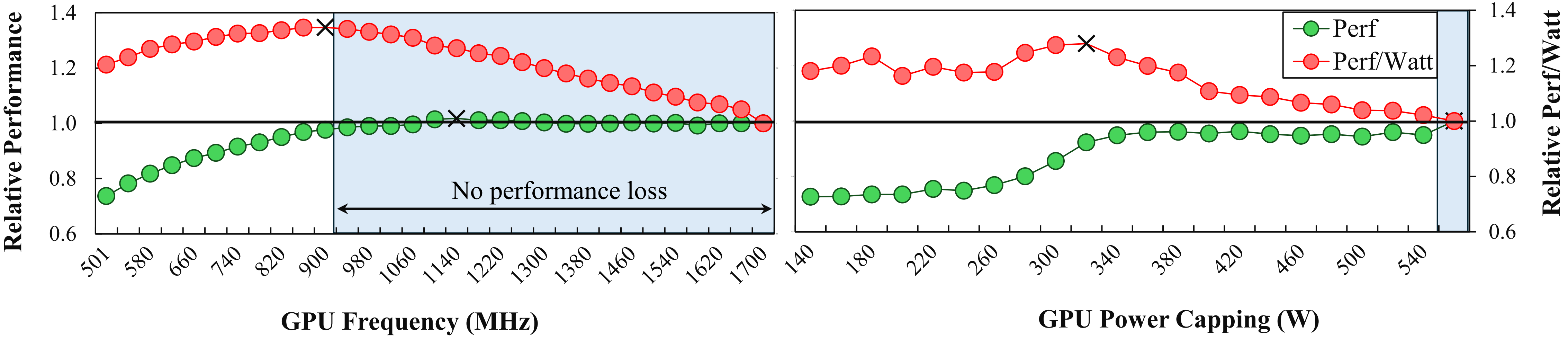
-8%

Energy efficiency increases when operating below peak frequency, as memory-controller and leakage power drop
Frequency capping matches power draw to memory bandwidth limits, eliminating unnecessary dynamic power

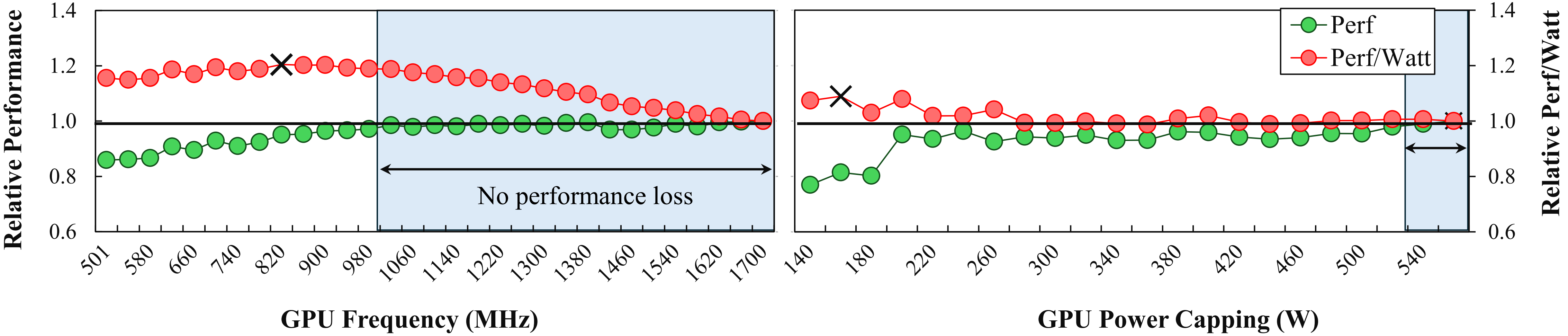
GPPD/UFRGS

Evaluation: Apps that benefit more from Freq. Capping

Cholla 1 node



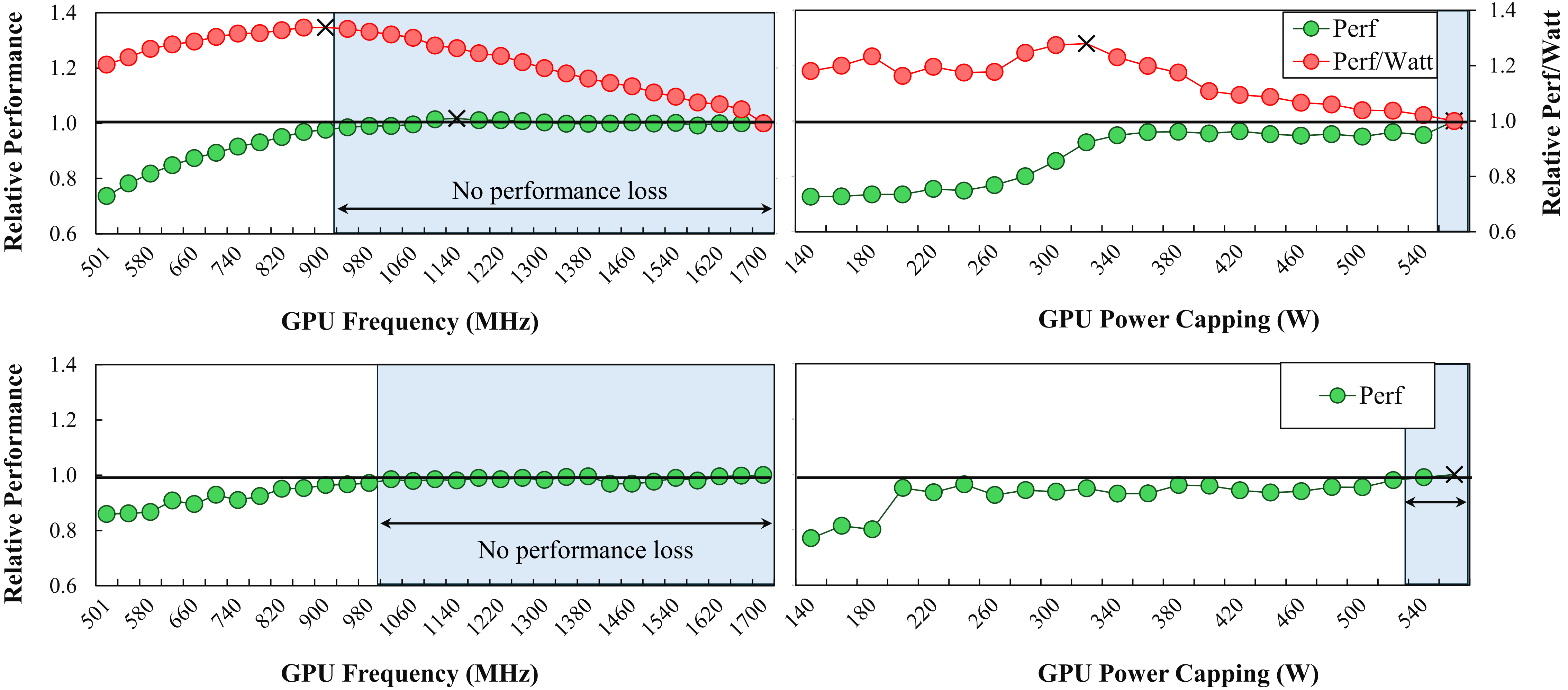
Cholla 32 node



Evaluation: Apps that benefit more from Freq. Capping

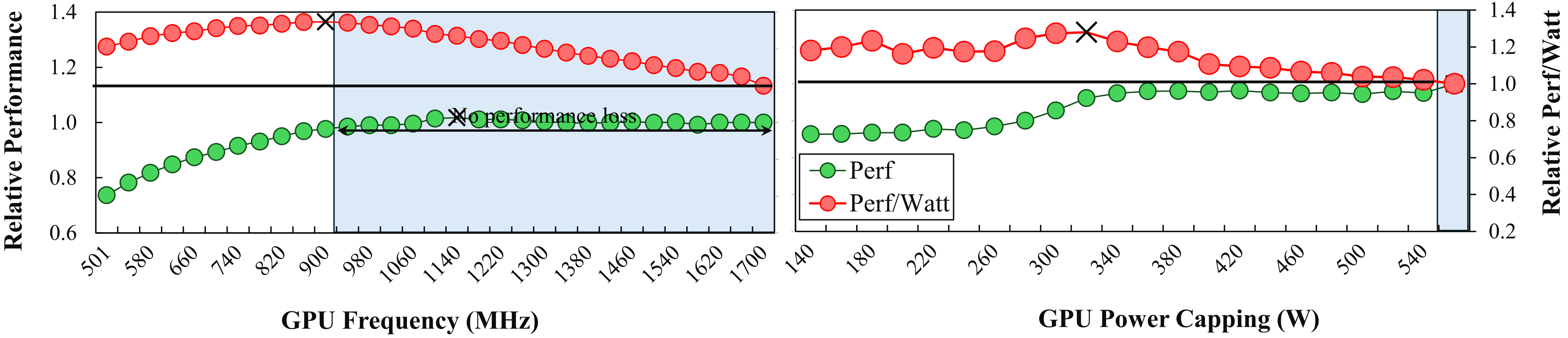
Cholla
1 node

Cholla
32
node

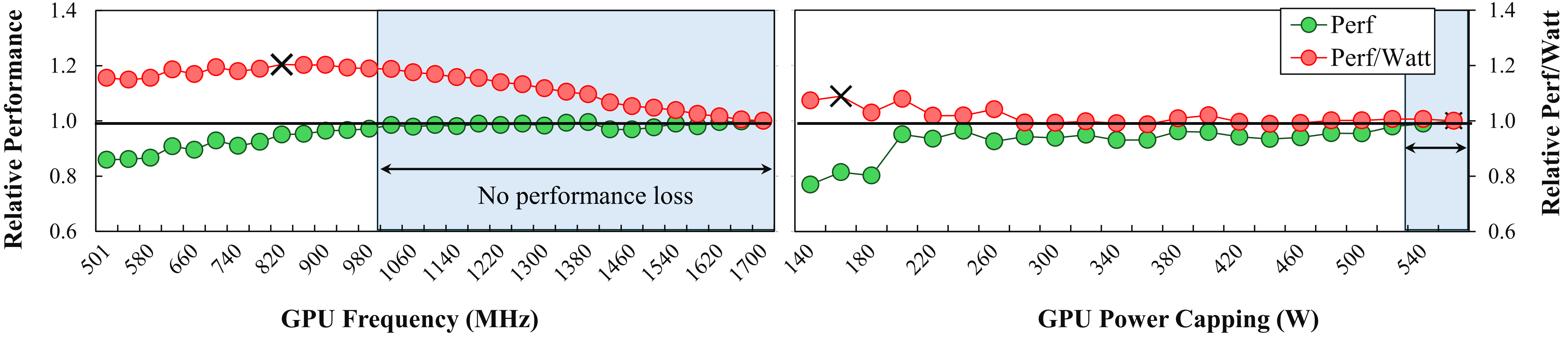


Evaluation: Apps that benefit more from Freq. Capping

Cholla 1 node



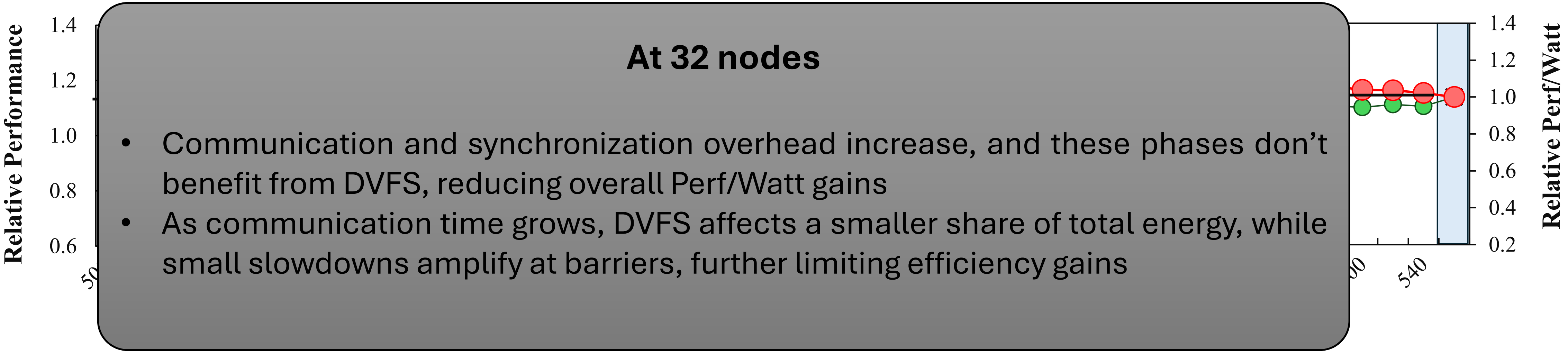
Cholla 32 node



GPPD/UFRGS

Evaluation: Apps that benefit more from Freq. Capping

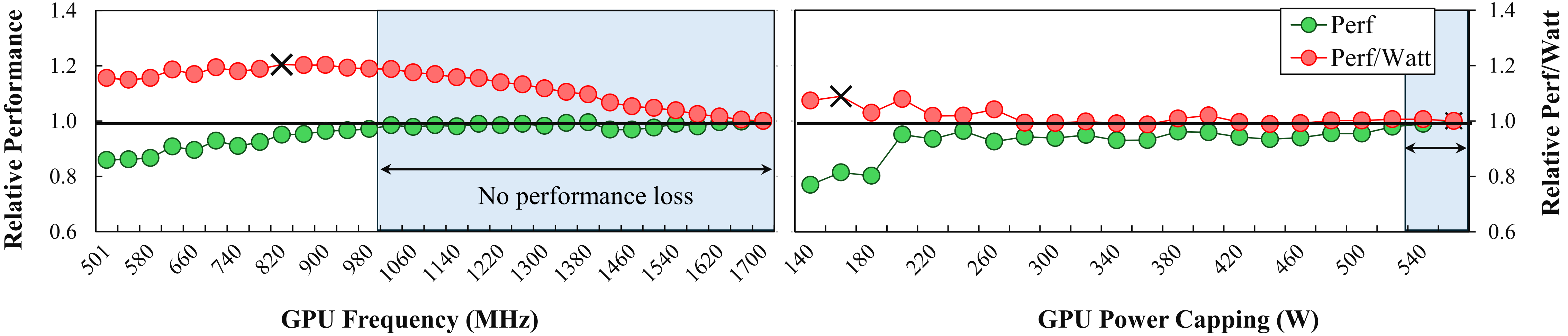
Cholla
1 node



At 32 nodes

- Communication and synchronization overhead increase, and these phases don’t benefit from DVFS, reducing overall Perf/Watt gains
- As communication time grows, DVFS affects a smaller share of total energy, while small slowdowns amplify at barriers, further limiting efficiency gains

Cholla
32
node

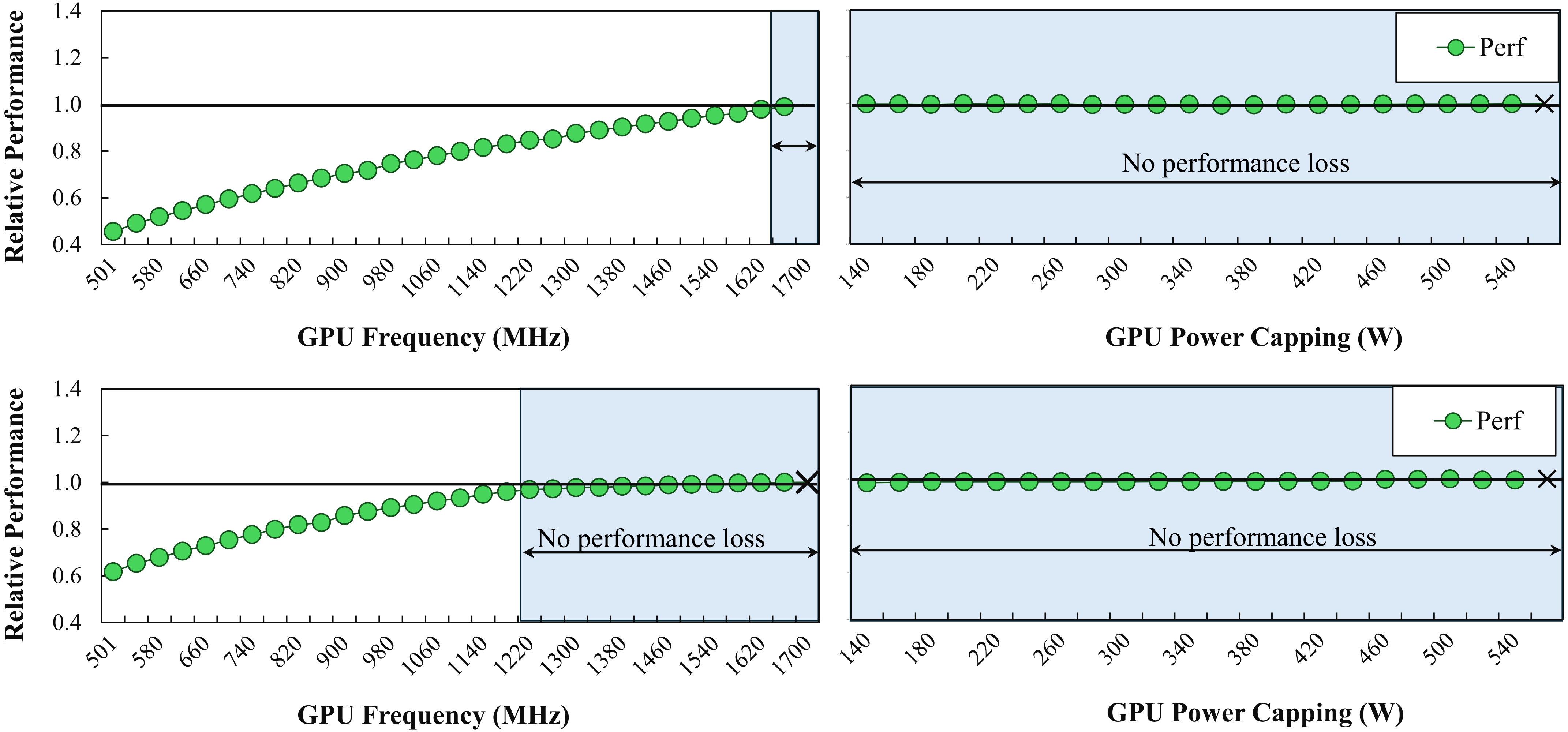


GPPD/UFRGS

Evaluation: Apps that benefit more from Freq. Capping

LAMMPS
1 node

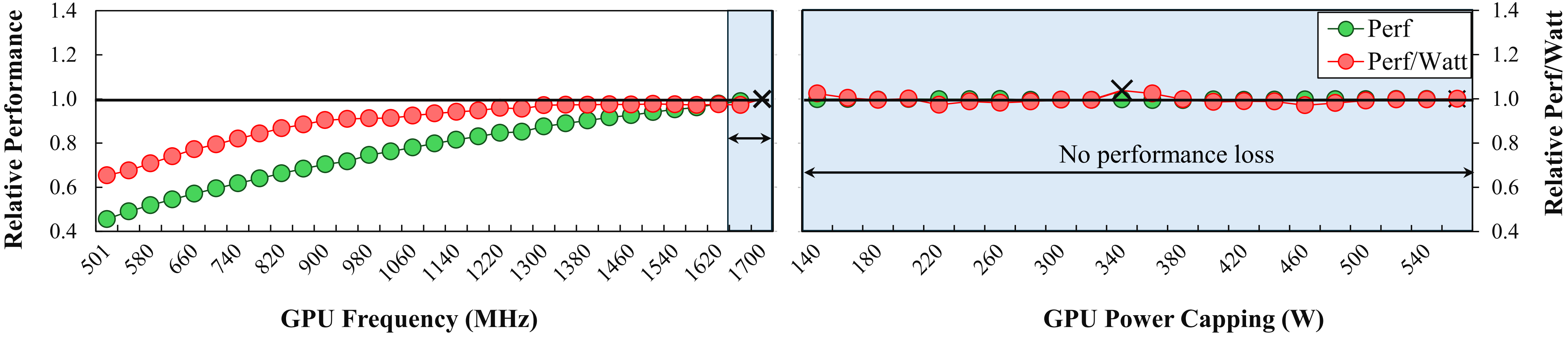
LAMMPS
32 node



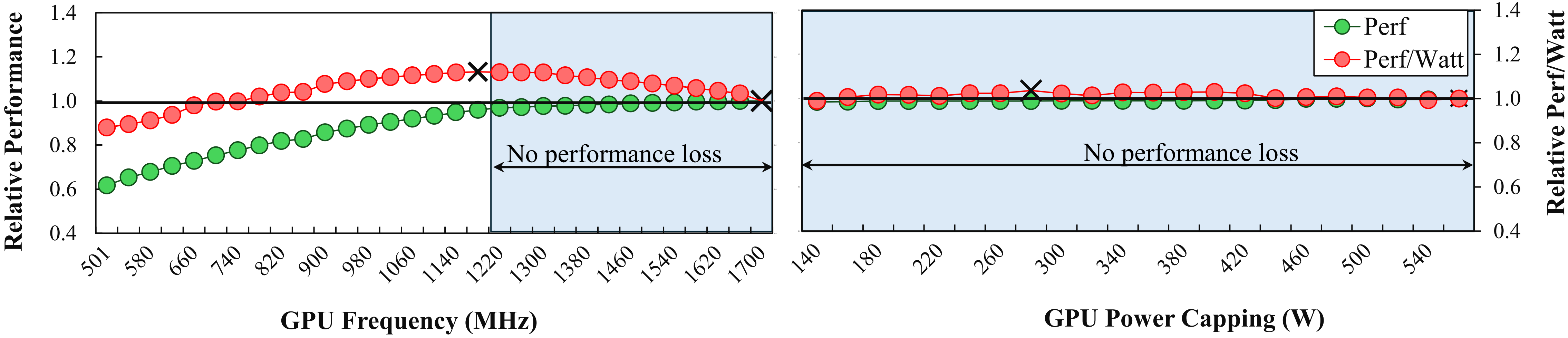
GPPD/UFRGS

Evaluation: Apps that benefit more from Freq. Capping

LAMMPS
1 node



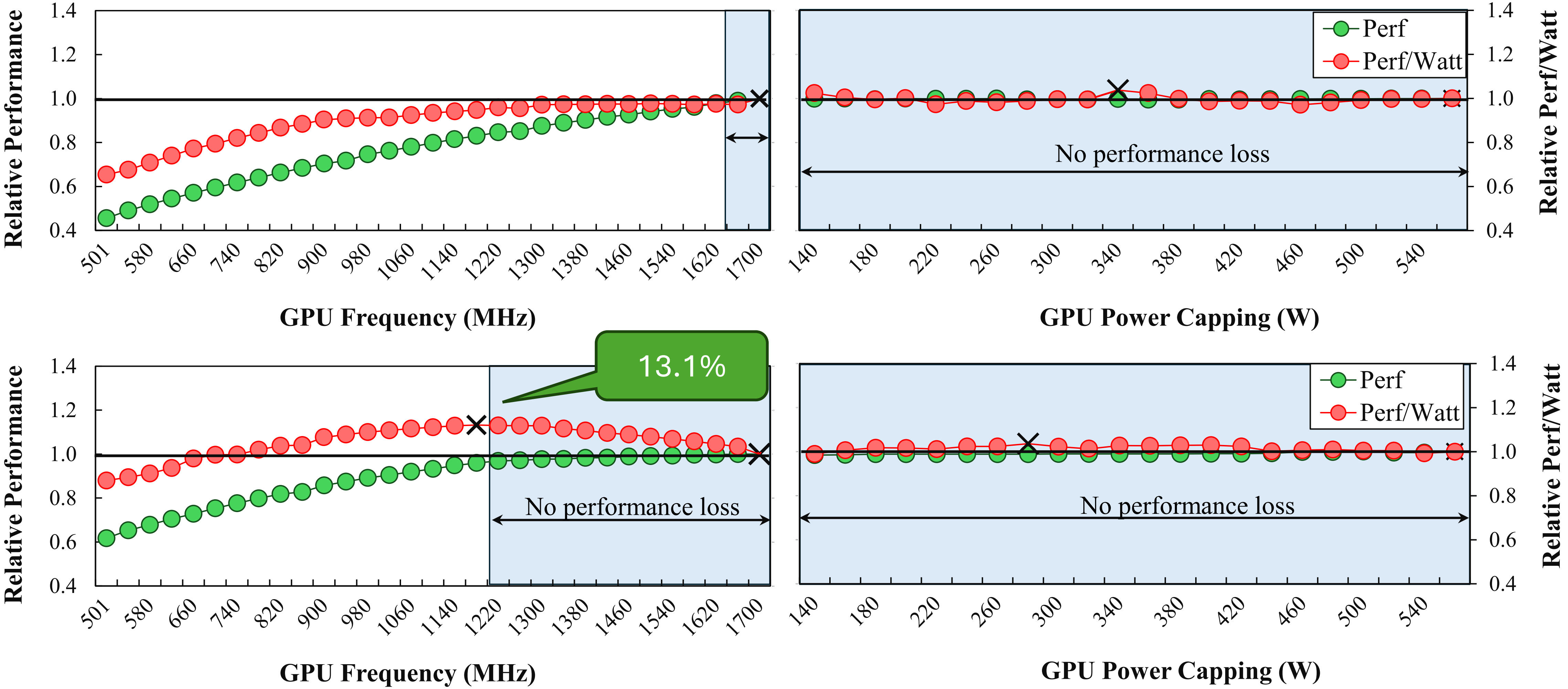
LAMMPS
32 node



Evaluation: Apps that benefit more from Freq. Capping

LAMMPS 1 node

LAMMPS 32 node



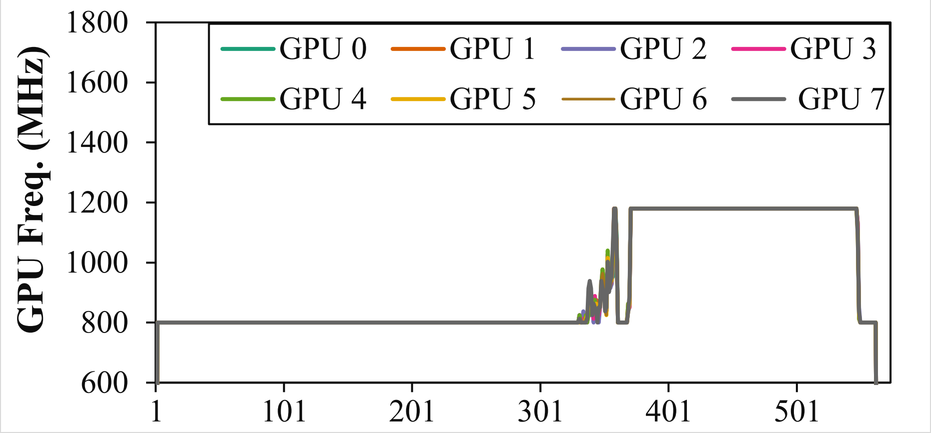
At scale, LAMMPS is bandwidth limited: fixed lower clocks cut power without hurting runtime, so DVFS beats power caps on Perf/Watt

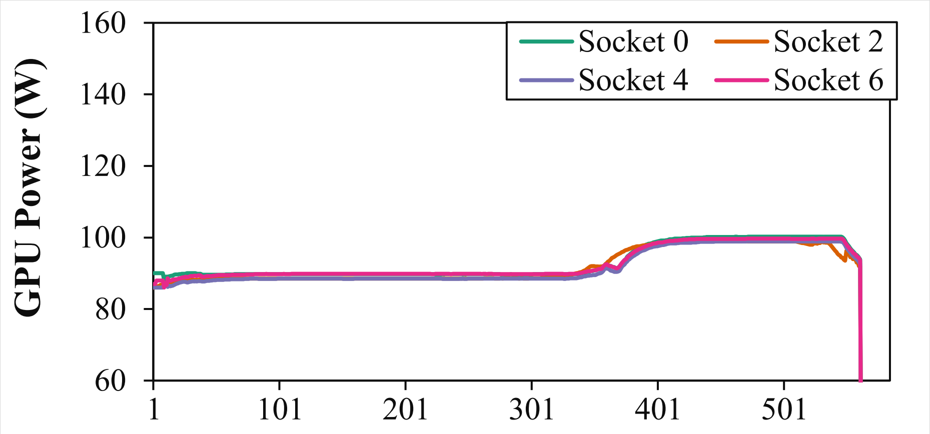
GPPD/UFRGS

Evaluation: Apps that benefit more from Freq. Capping

LAMMPS
32 node

Freq. Capping at 1180MHz



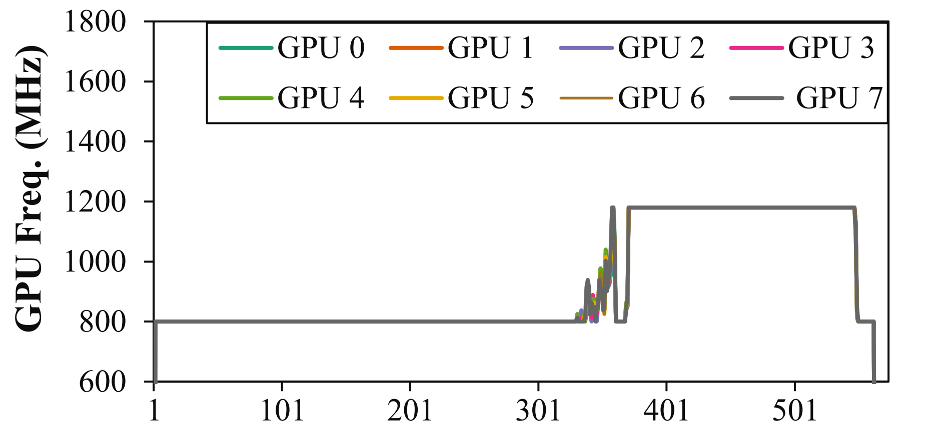


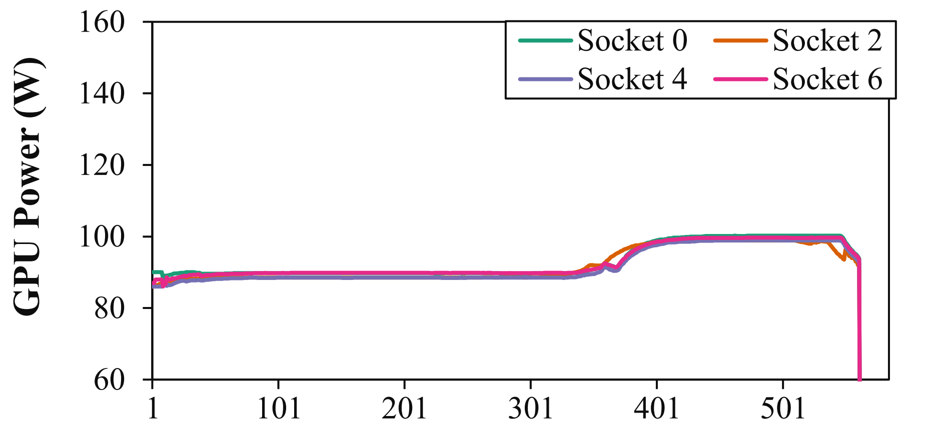
GPPD/UFRGS

Evaluation: Apps that benefit more from Freq. Capping

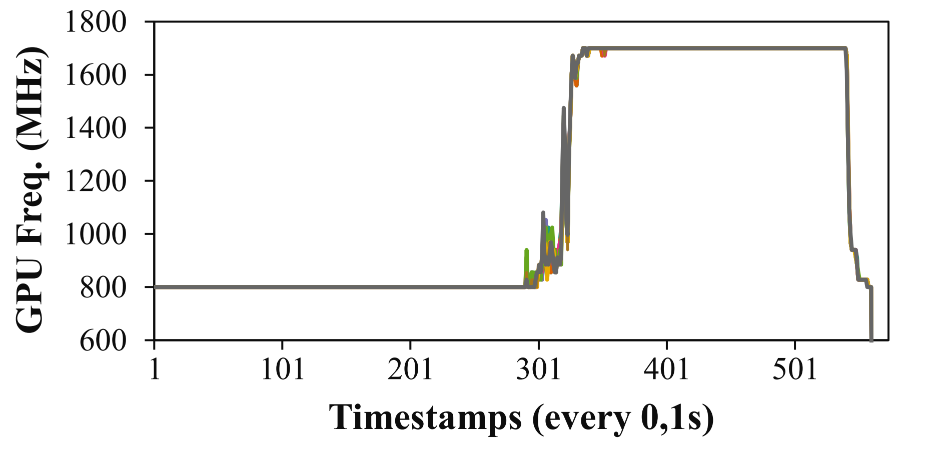
LAMMPS
32 node

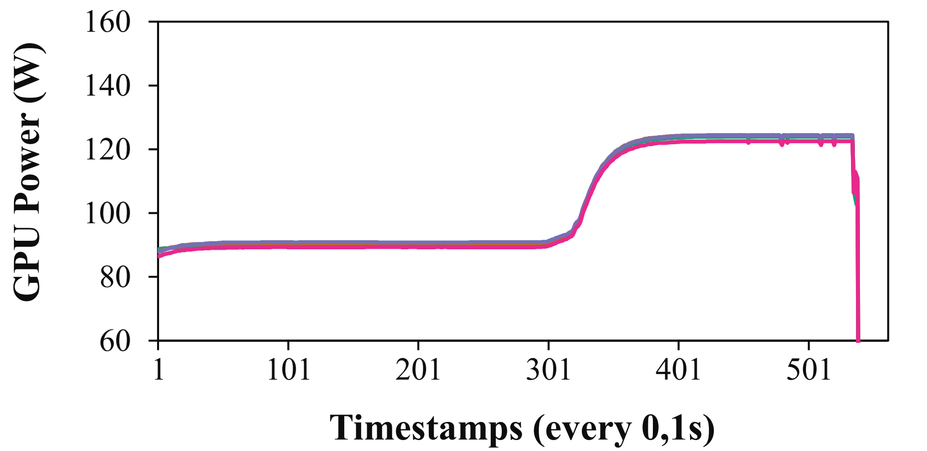
Freq. Capping at 1180MHz





Power Capping at 280 W

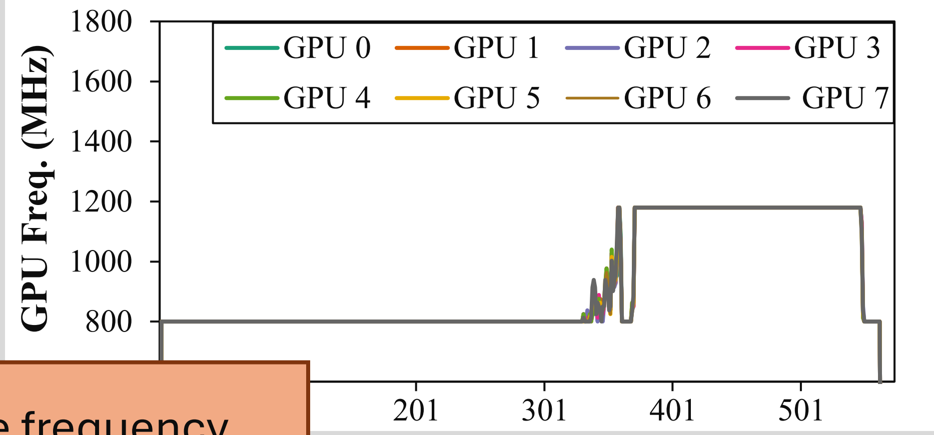


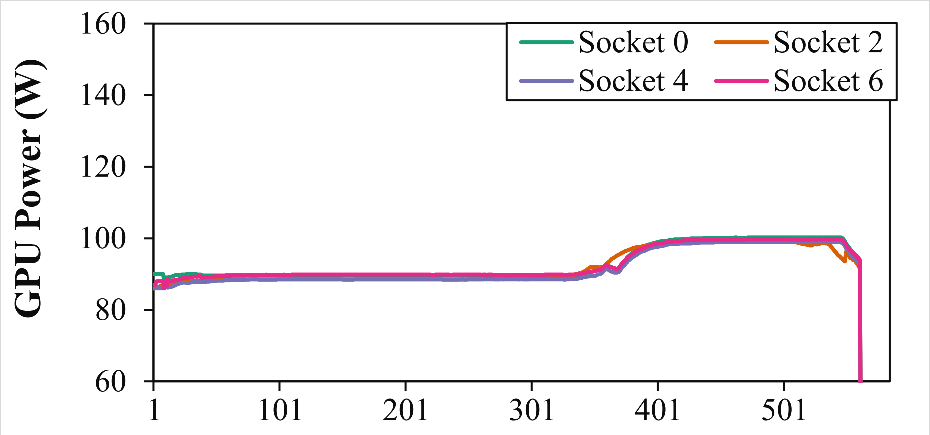


Evaluation: Apps that benefit more from Freq. Capping

LAMMPS
32 node

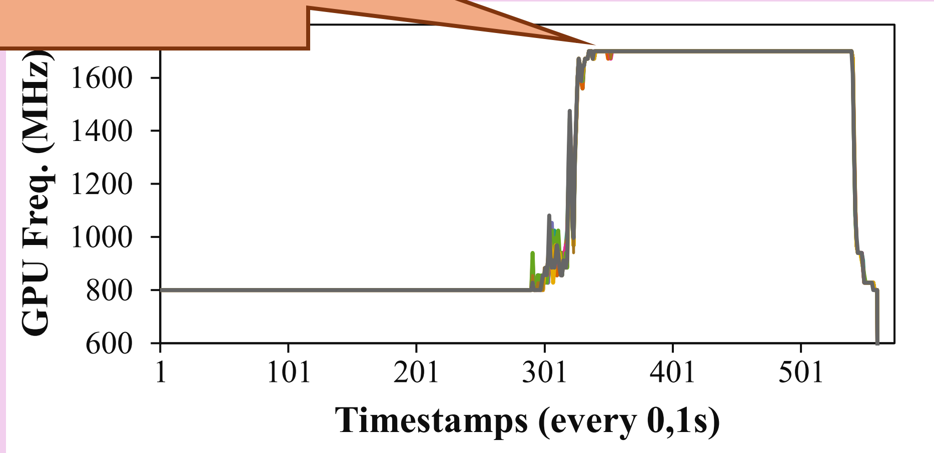
Freq. Capping at 1180MHz

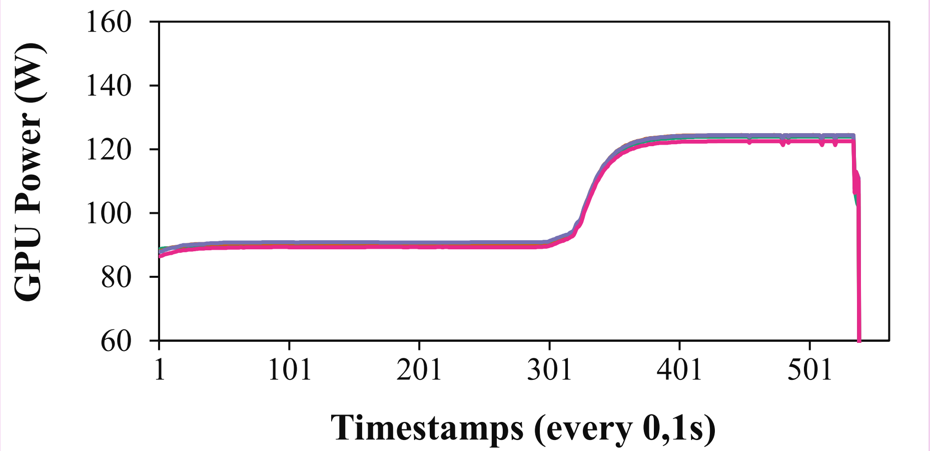




Allowing GPU to increase frequency yelded no performance gain, only higher power and energy waste.

Power Capping at 280 W





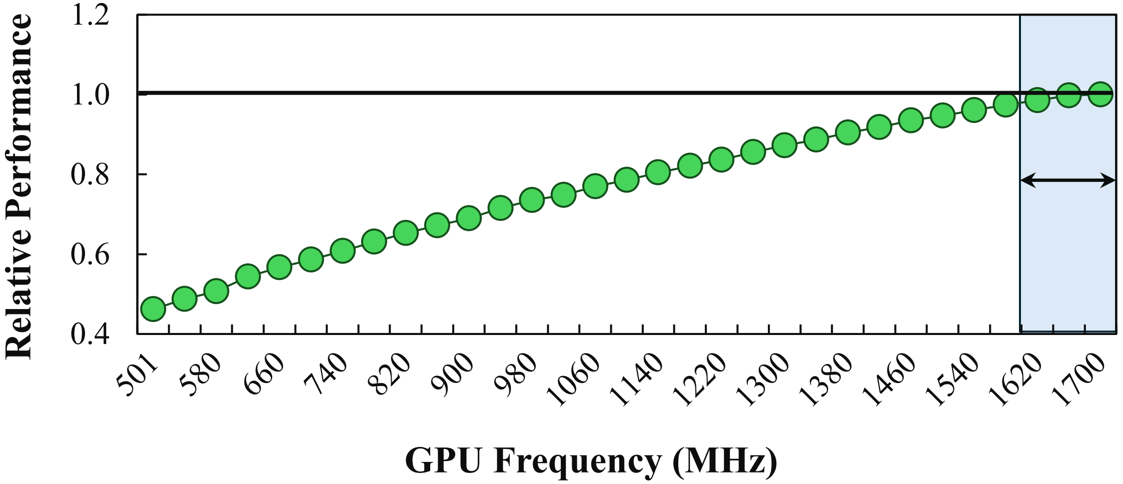
Evaluation

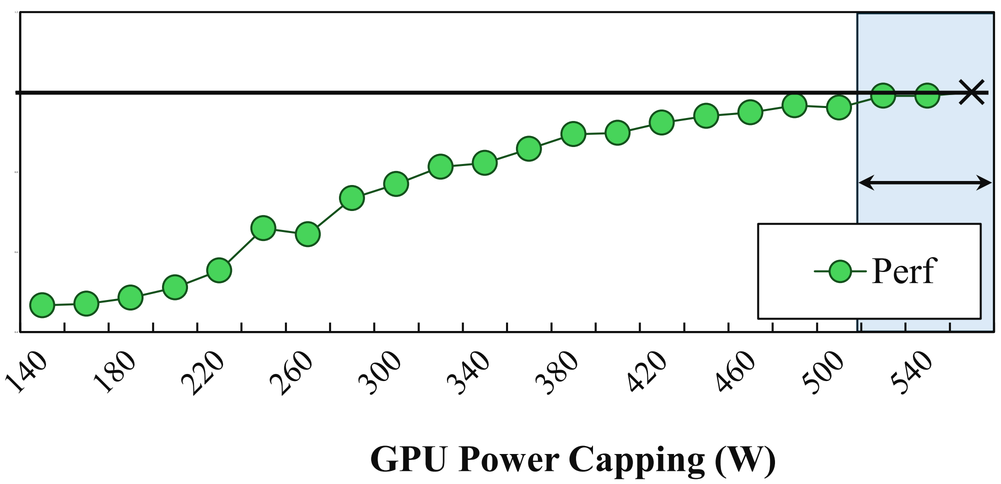
- ~~What applications benefit more from Frequency Capping?~~

- What applications benefit more from Power Capping?

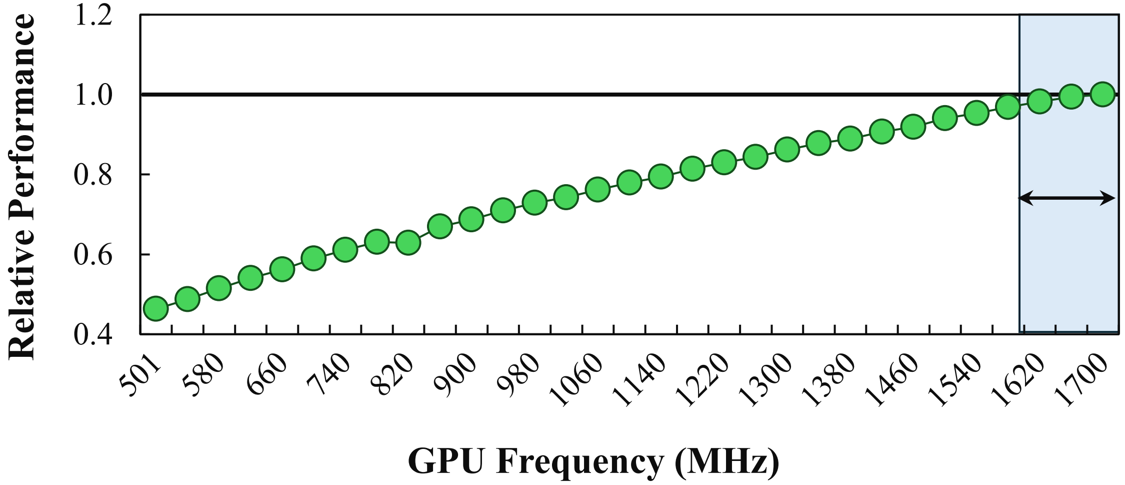
Evaluation: Apps that benefit more from Power Capping

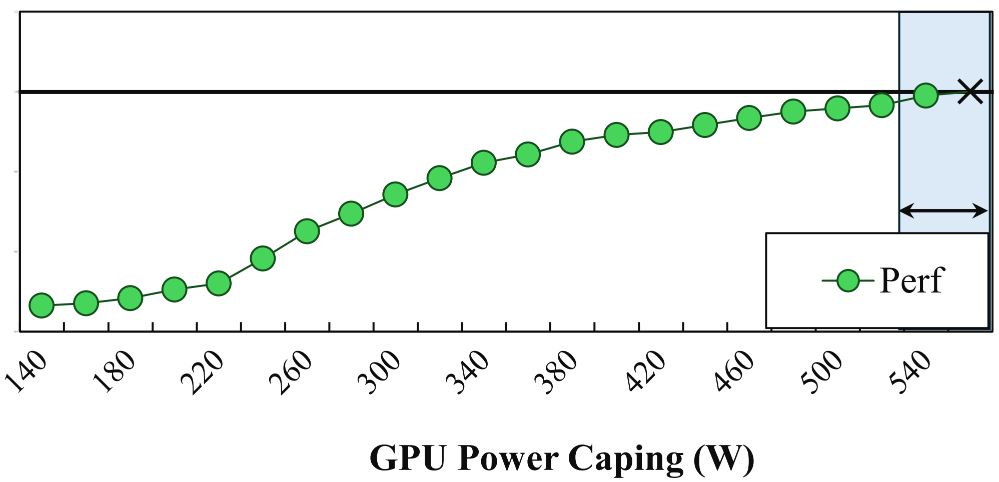
HACC
1 node





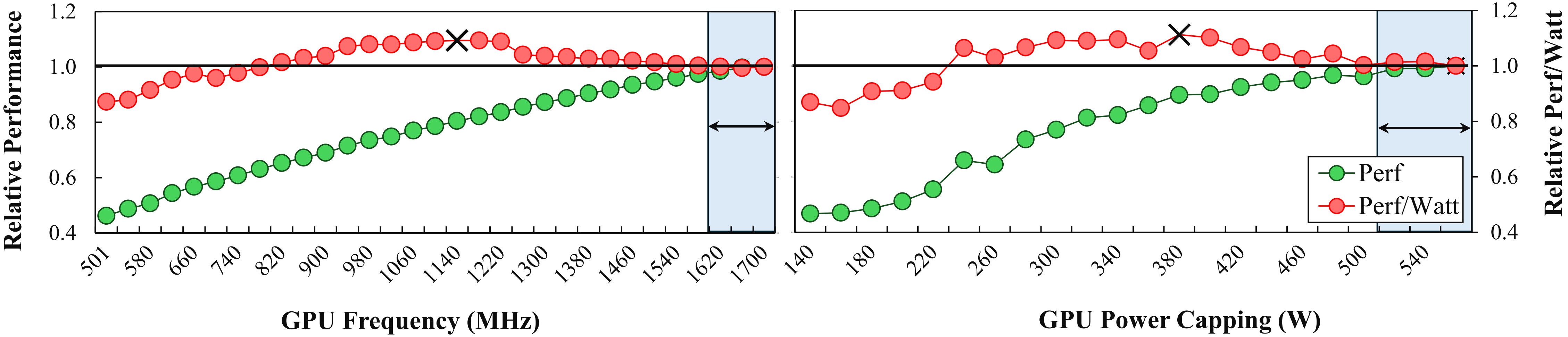
HACC
32
node



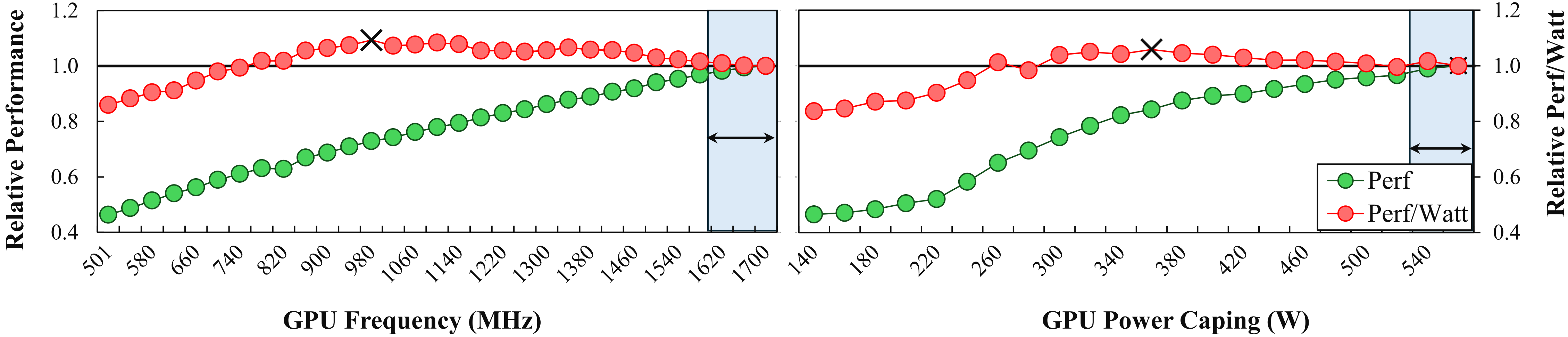


Evaluation: Apps that benefit more from Power Capping

HACC
1 node



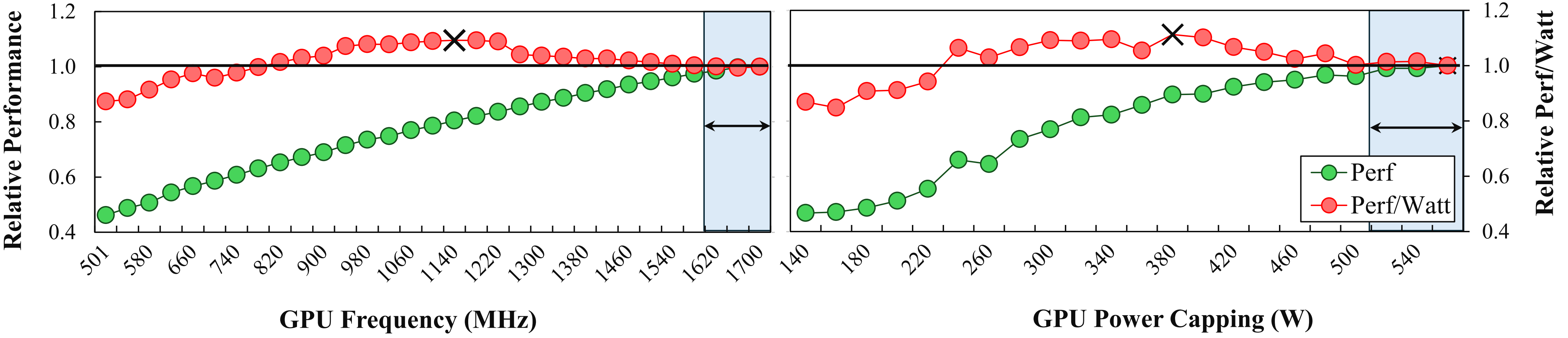
HACC
32
node



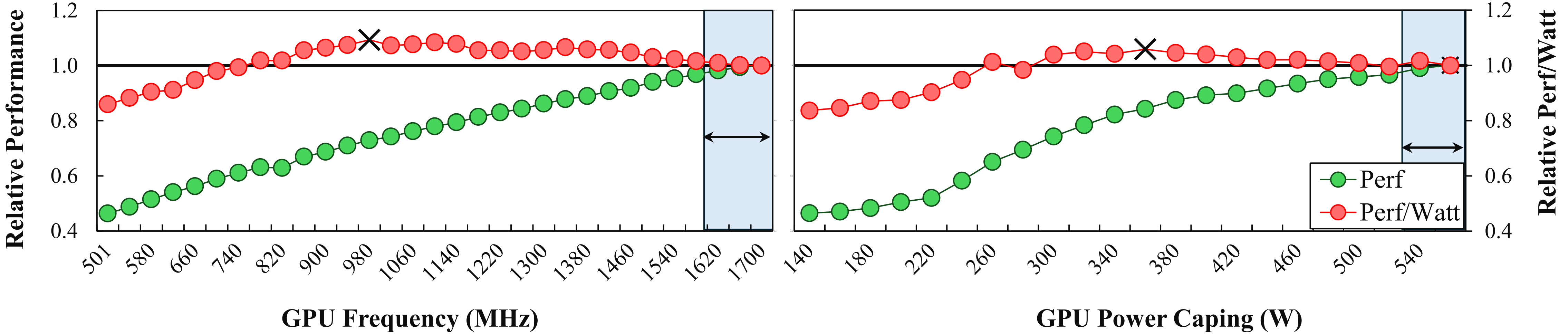
GPPD/UFRGS

Evaluation: Apps that benefit more from Power Capping

HACC
1 node



HACC
32
node



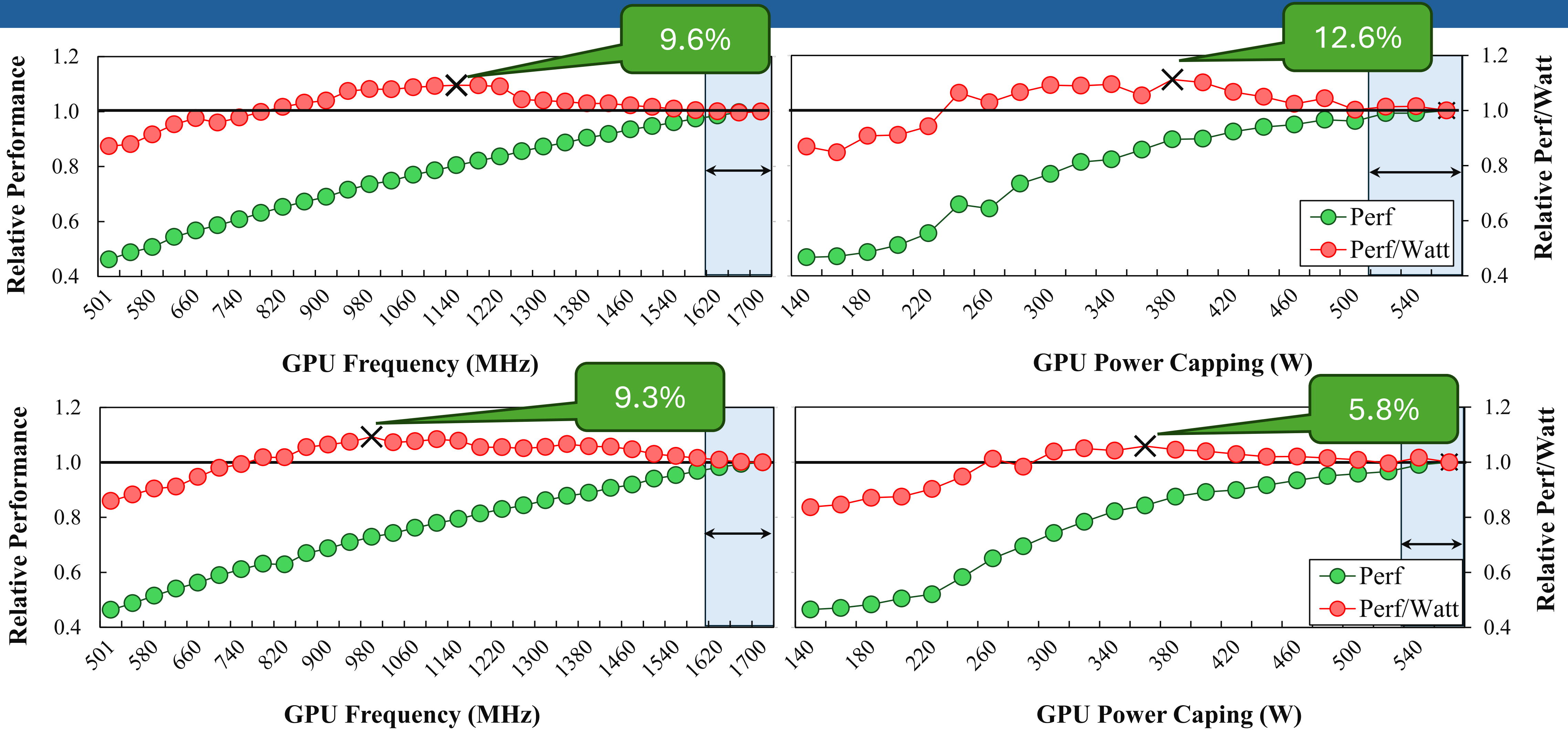
HACC keeps GPU utilization stable at scale, so power and frequency capping drive the hardware to similar voltage-frequency states, resulting in equivalent performance-energy behavior

GPPD/UFRGS

Evaluation: Apps that benefit more from Power Capping

HACC 1 node

HACC 32 node



9.6%

12.6%

9.3%

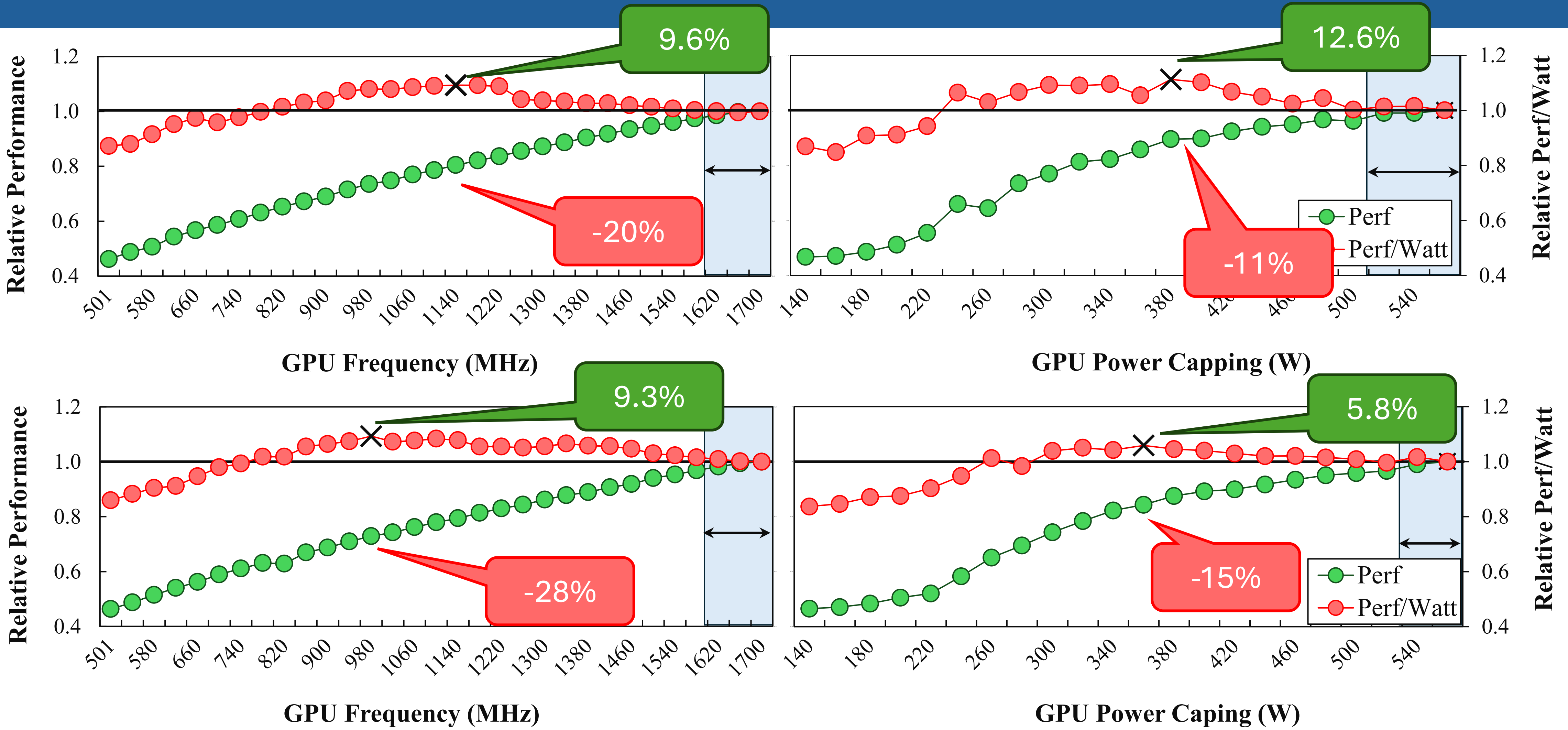
5.8%

HACC keeps GPU utilization stable at scale, so power and frequency capping drive the hardware to similar voltage-frequency states, resulting in equivalent performance-energy behavior

Evaluation: Apps that benefit more from Power Capping

HACC
1 node

HACC
32
node



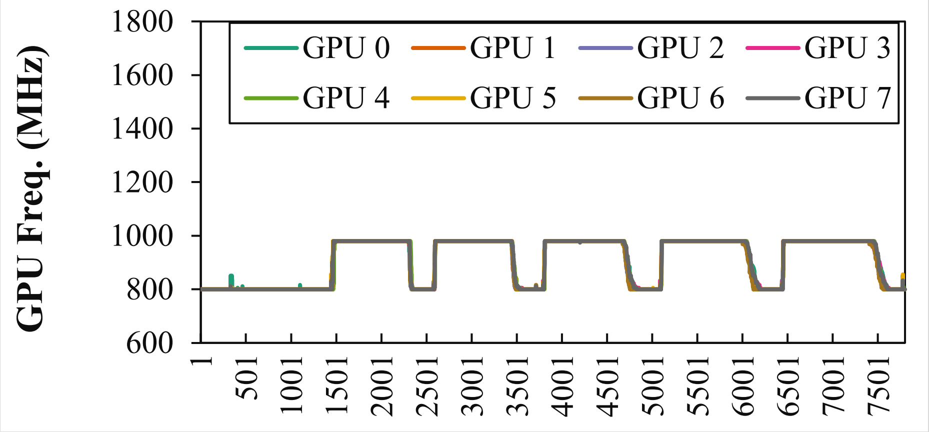
HACC keeps GPU utilization stable at scale, so power and frequency capping drive the hardware to similar voltage-frequency states, resulting in equivalent performance-energy behavior

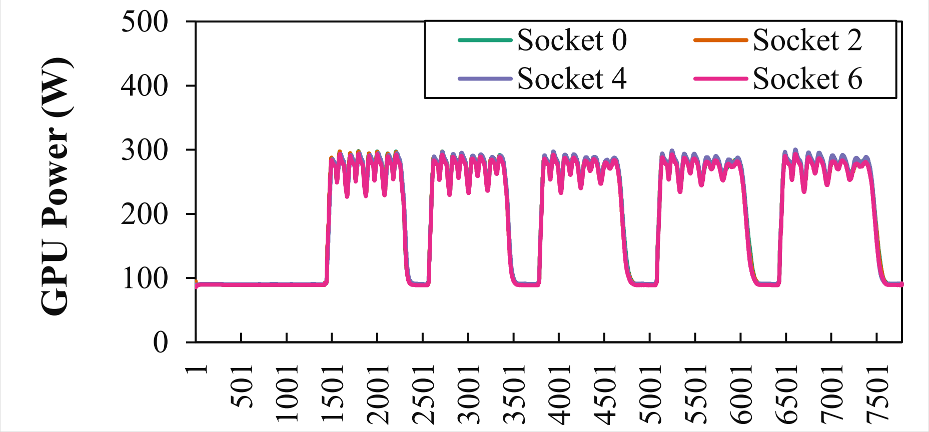
GPPD/UFRGS

Evaluation: Apps that benefit more from Power Capping

HACC
32
node

Freq. Capping at 980MHz

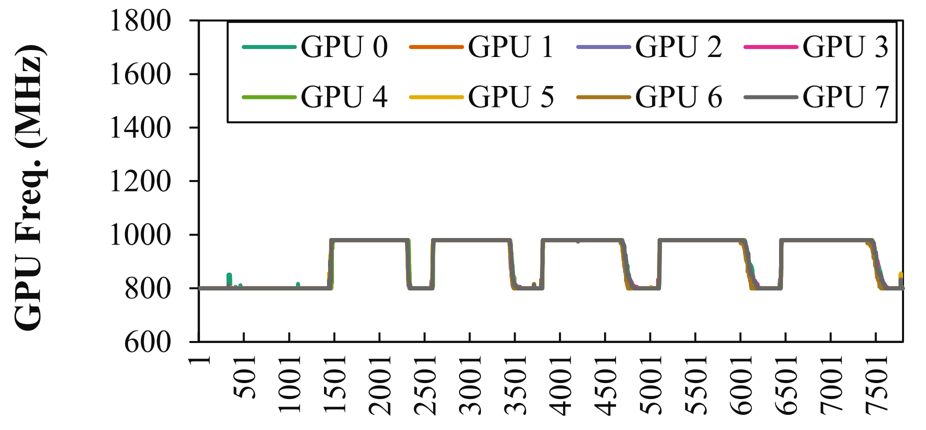


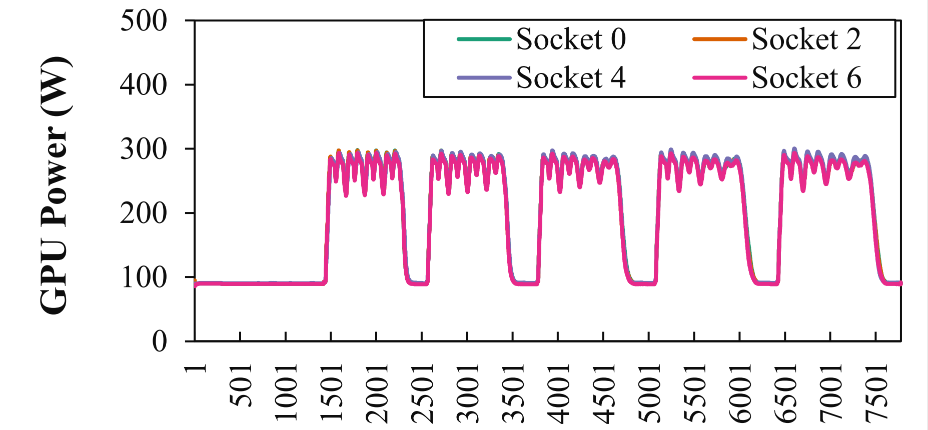


Evaluation: Apps that benefit more from Power Capping

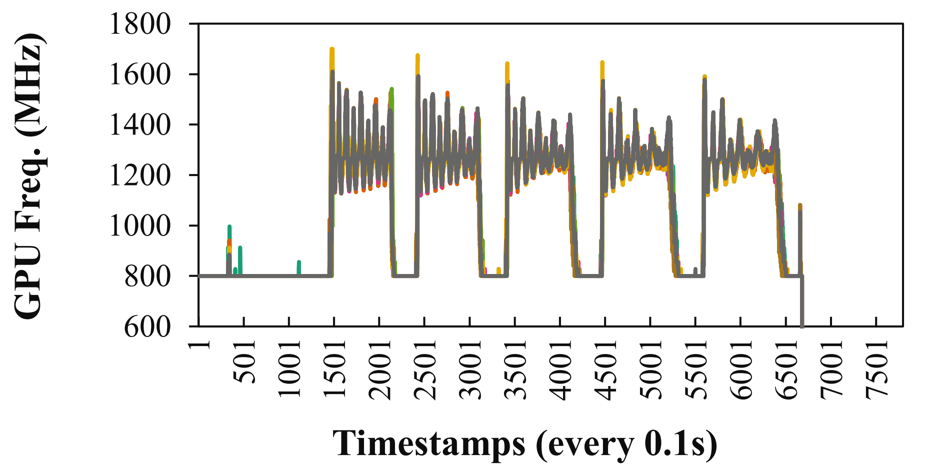
HACC 32 node

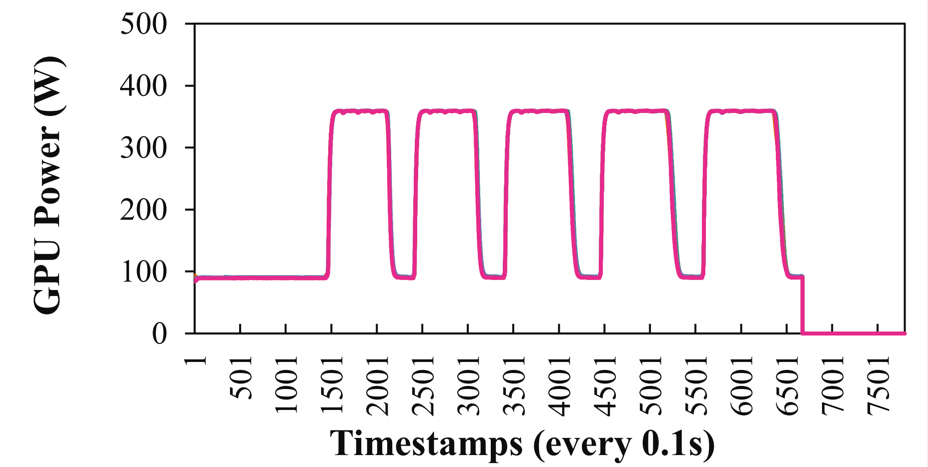
Freq. Capping at 980MHz





Power Capping at 360 W



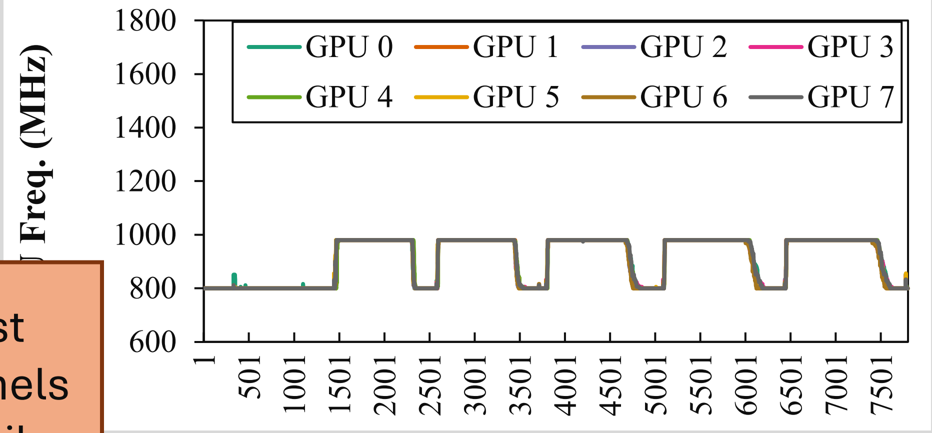


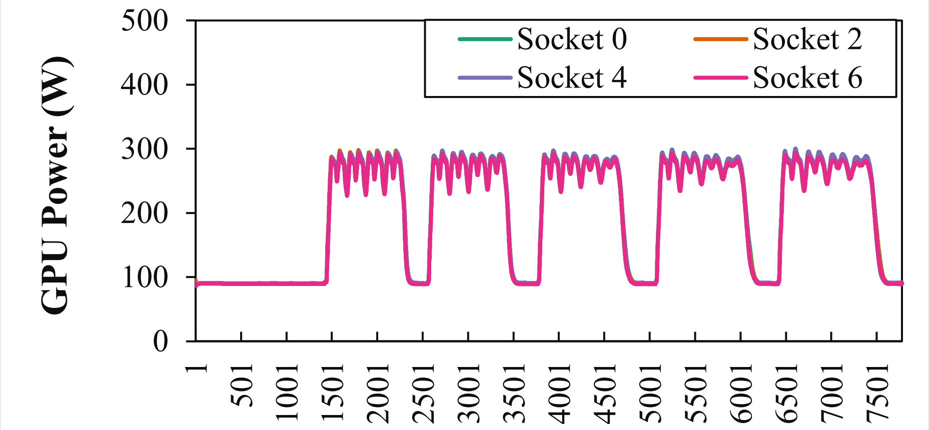
GPPD/UFRGS

Evaluation: Apps that benefit more from Power Capping

HACC
32
node

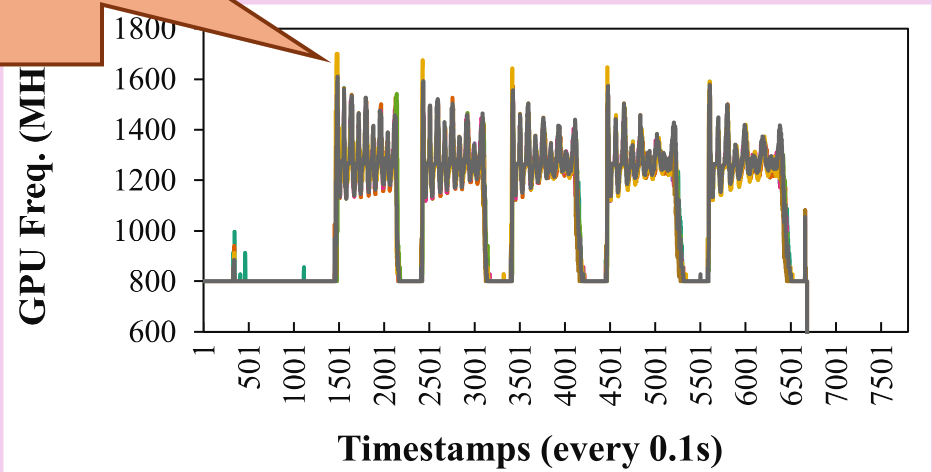
Freq. Capping at 980MHz

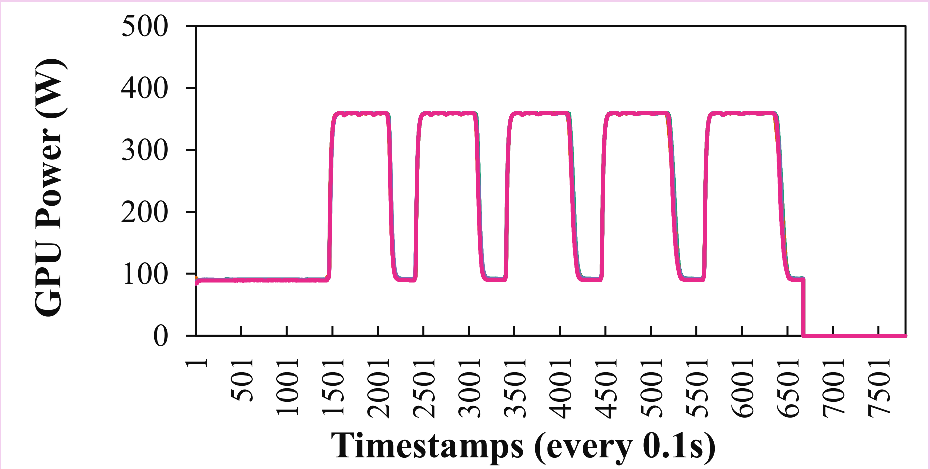




Allowing the GPU to boost frequency during burst kernels improved performance while reducing overall energy consumption

Power Capping at 360 W





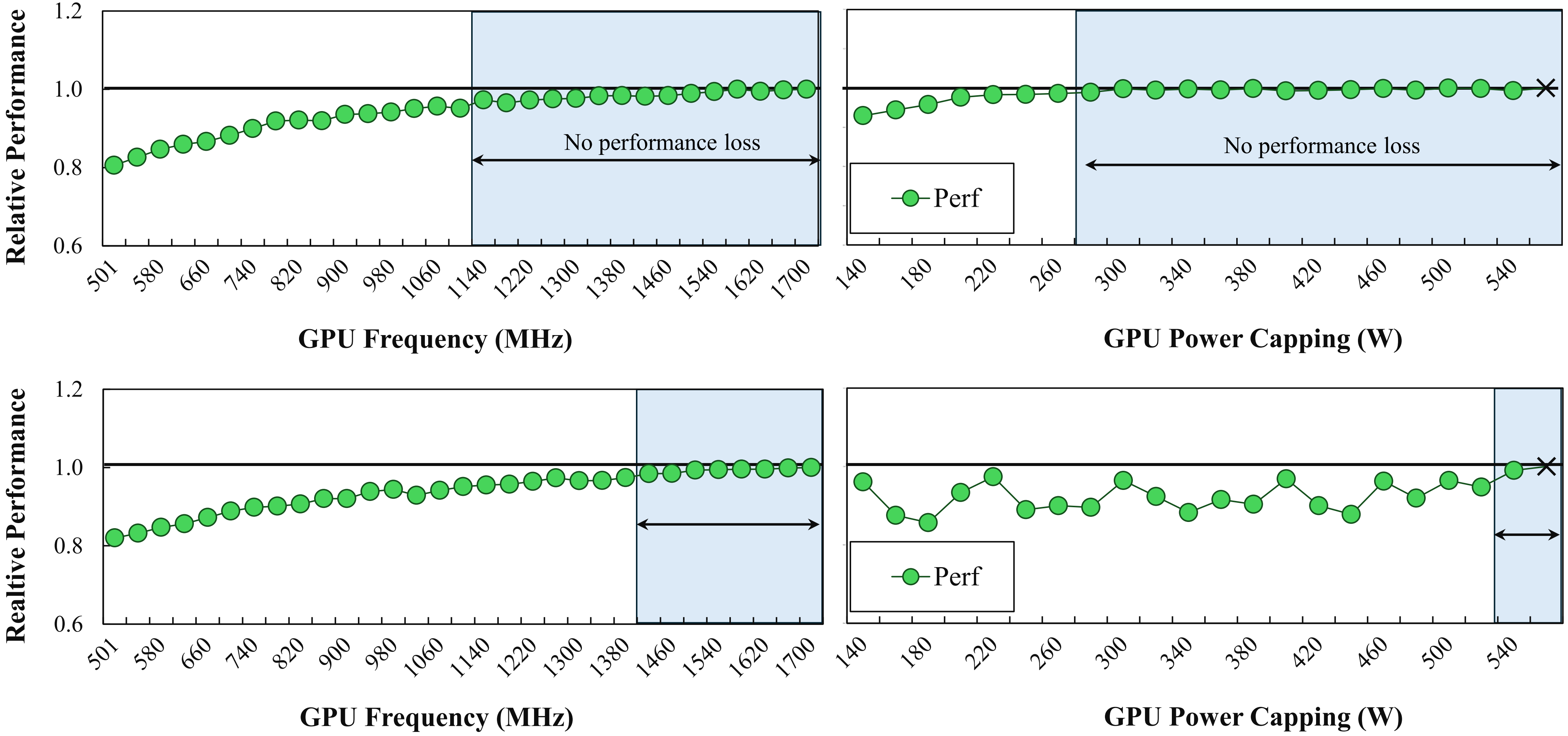
Evaluation

- ~~What applications benefit more from Frequency Capping?~~
- ~~What applications benefit more from Power Capping?~~
- What applications are impacted in similar ways?

Evaluation: Apps with similar behavior

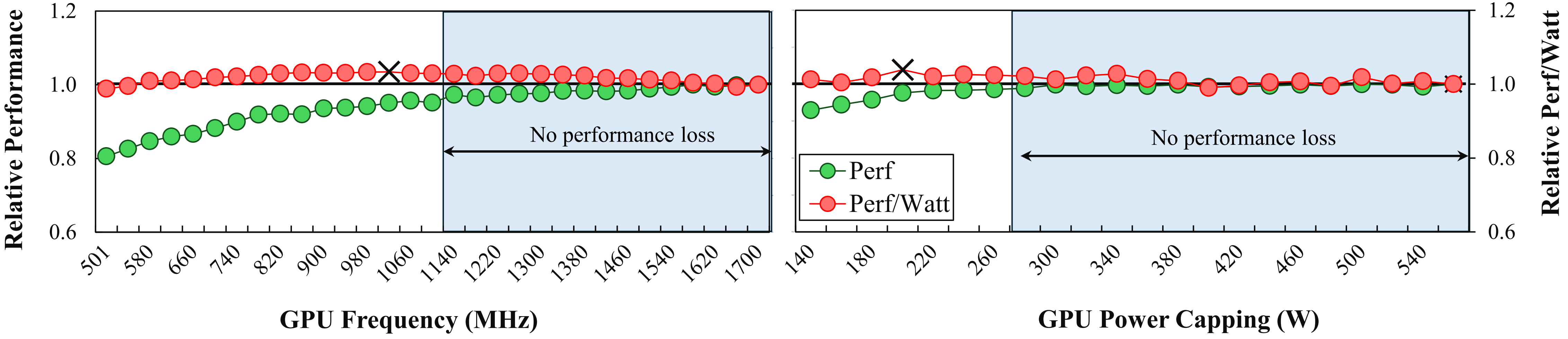
Kripke
1 node

Kripke
32
node

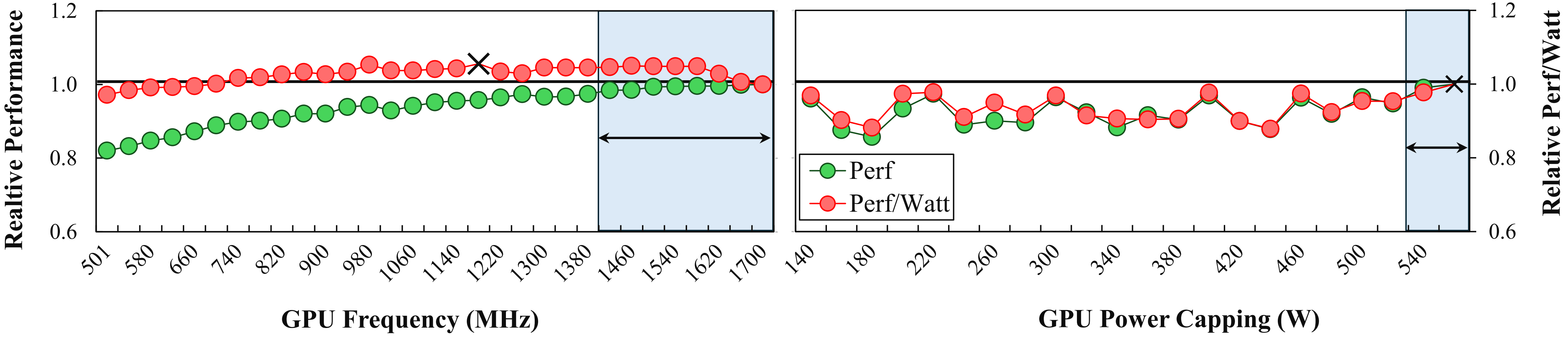


Evaluation: Apps with similar behavior

Kripke
1 node

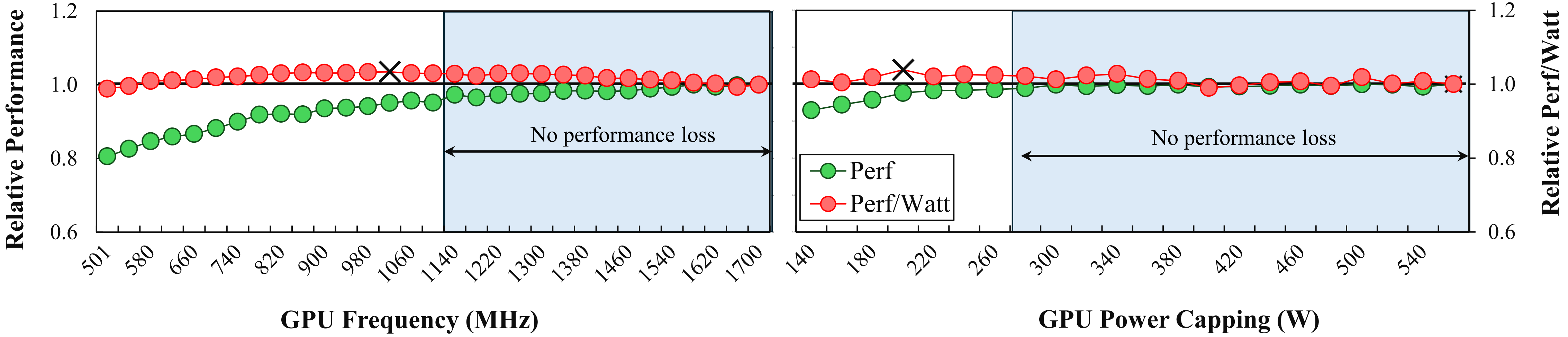


Kripke
32
node

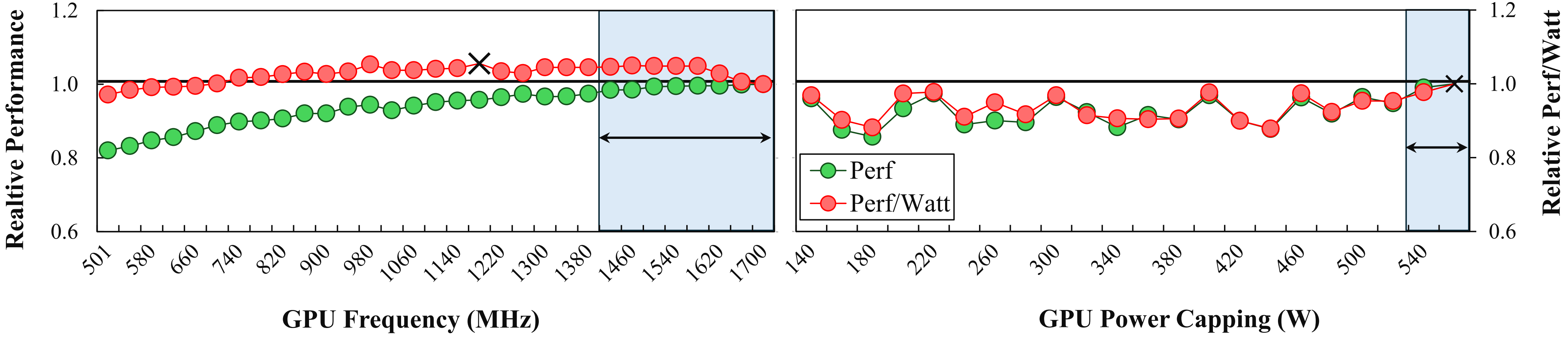


Evaluation: Apps with similar behavior

Kripke
1 node



Kripke
32
node



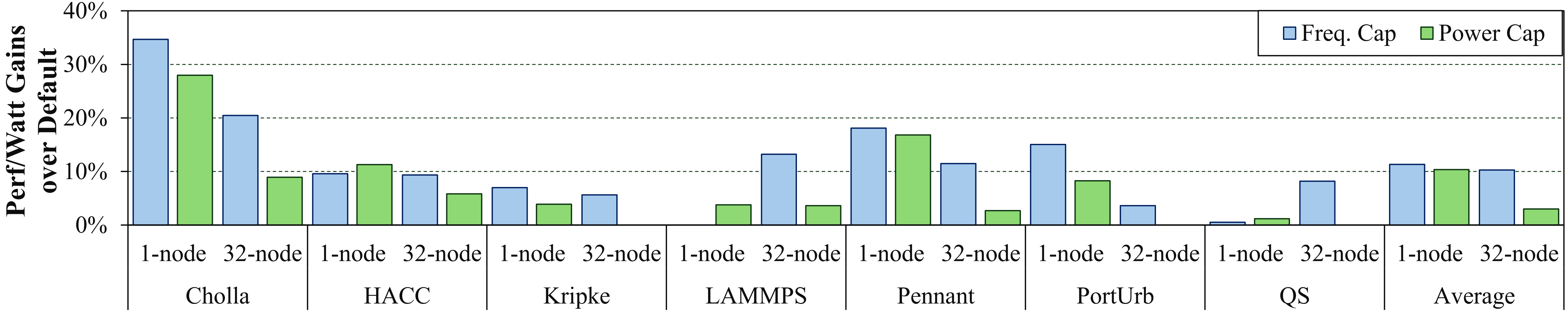
GPPD/UFRGS

Low AI (0.1 FLOPs/byte) → performance is limited by memory access latency rather than compute throughput
Lowering frequency or constraining power yields comparable effective frequencies, power draw, and Perf/Watt

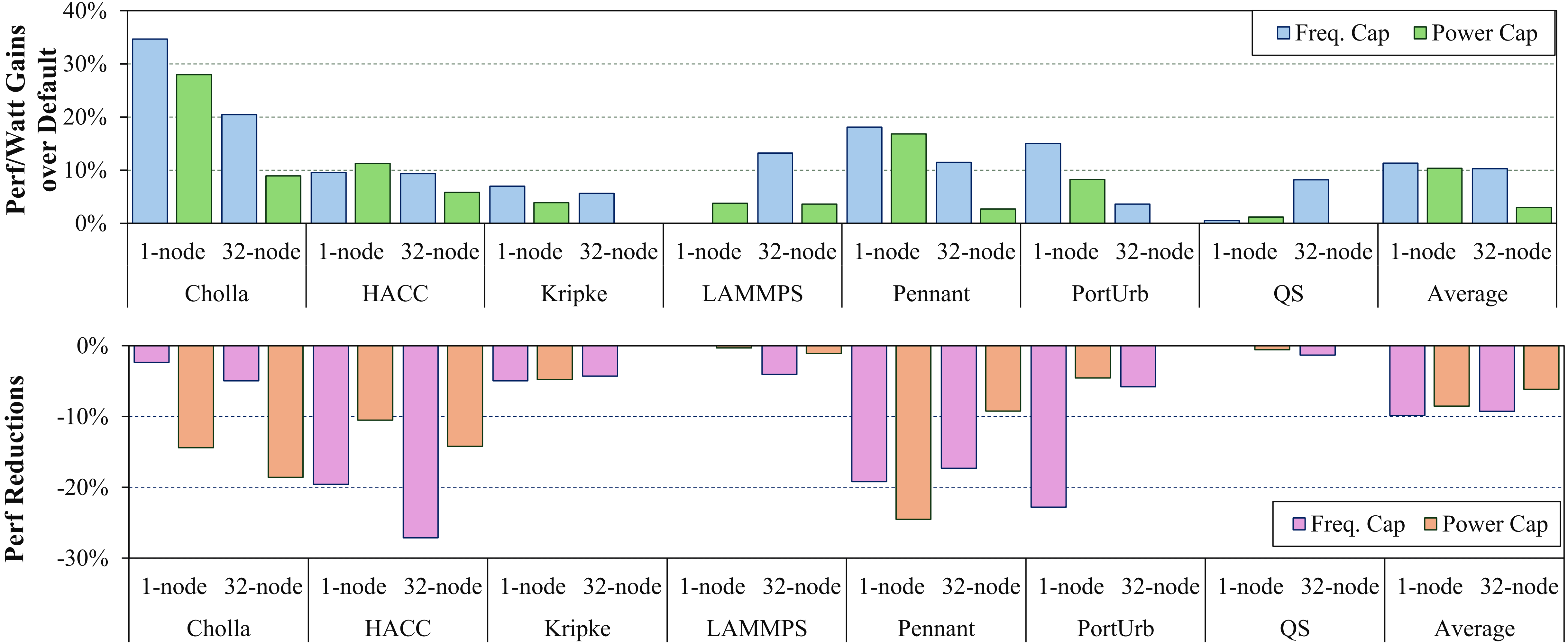
Evaluation

- ~~What applications benefit more from Frequency Capping?~~
- ~~What applications benefit more from Power Capping?~~
- ~~What applications are impacted in similar ways?~~
- What is the impact on the performance of applications?

Evaluation: What is the impact on the Performance?

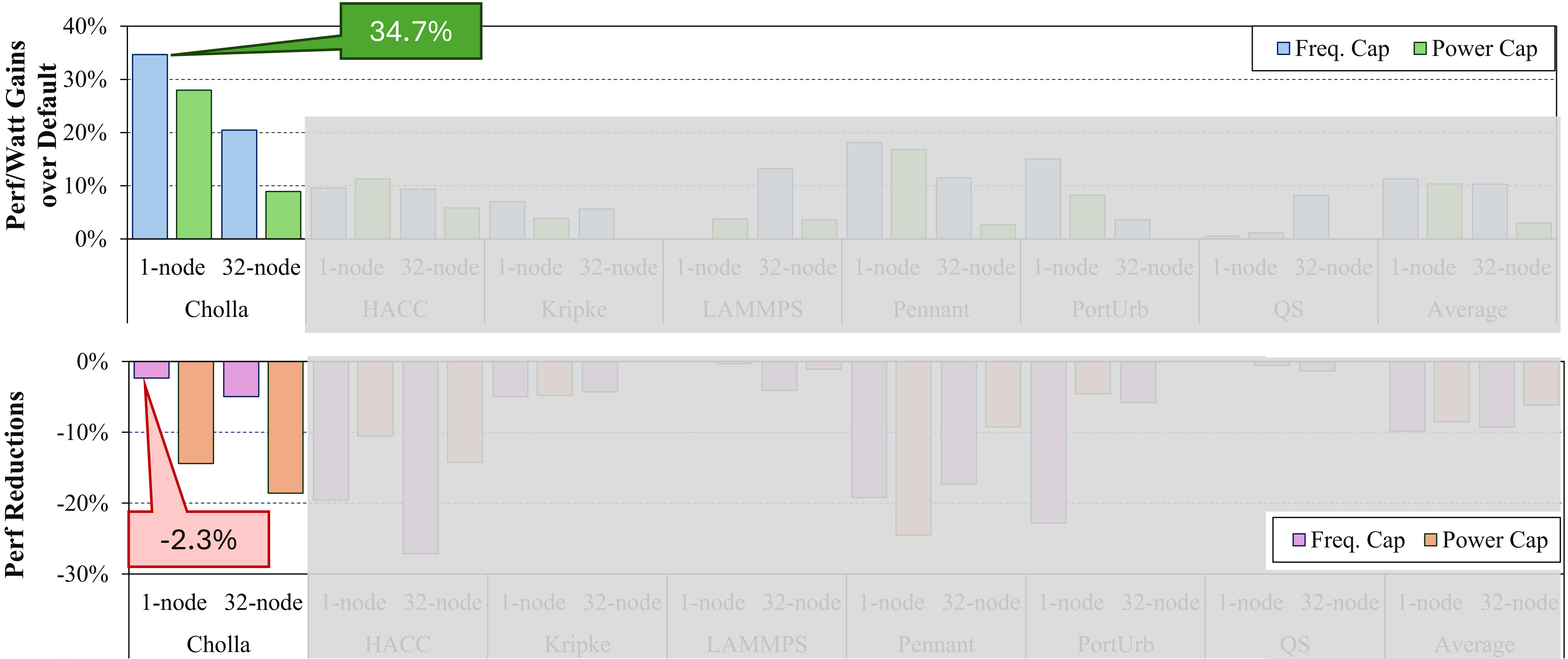


Evaluation: What is the impact on the Performance?

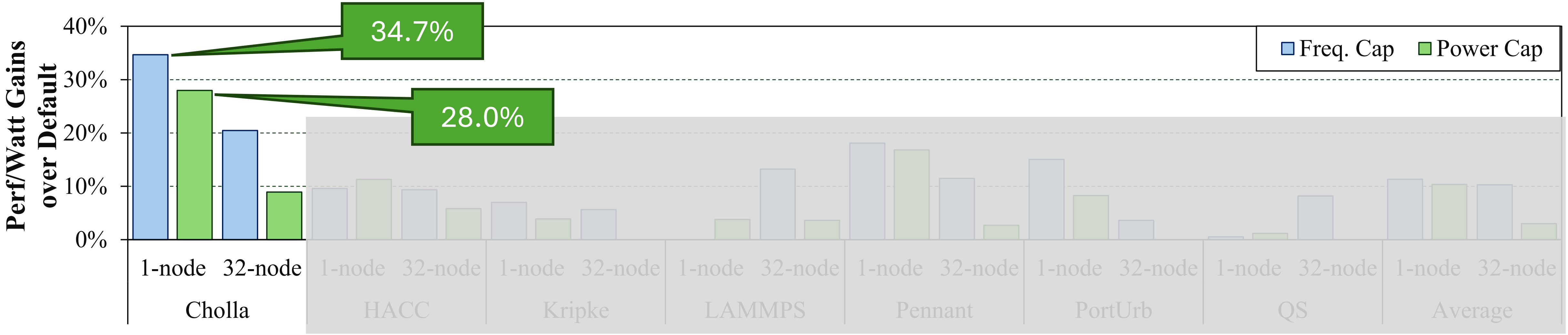


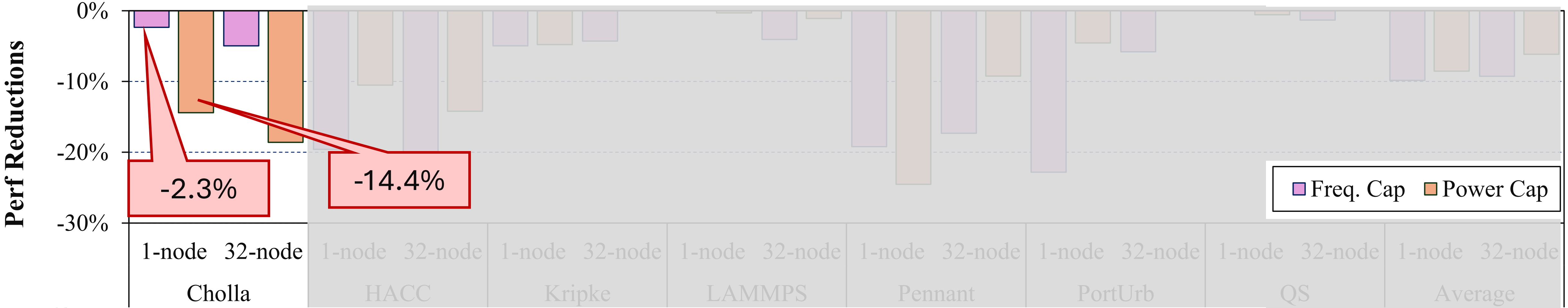
GPPD/UFRGS

Evaluation: What is the impact on the Performance?

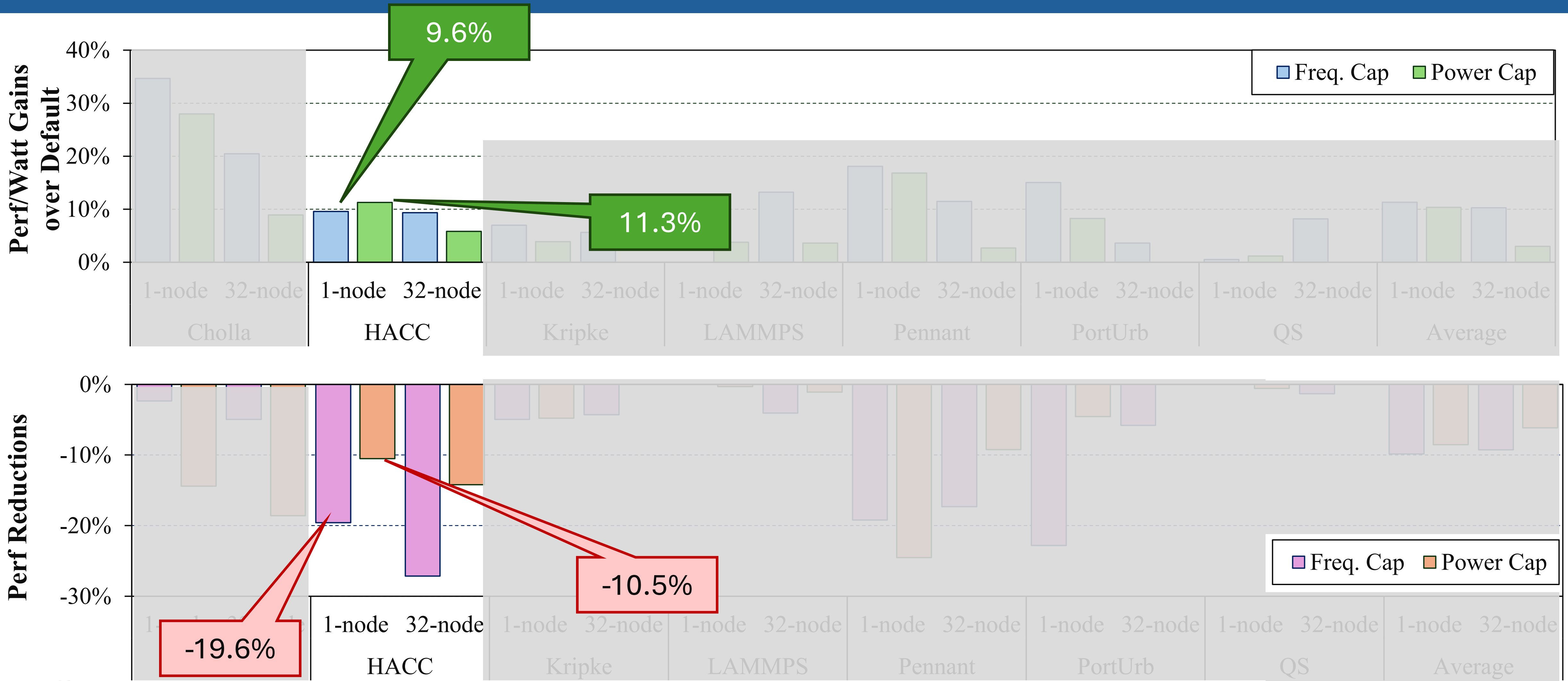


Evaluation: What is the impact on the Performance?

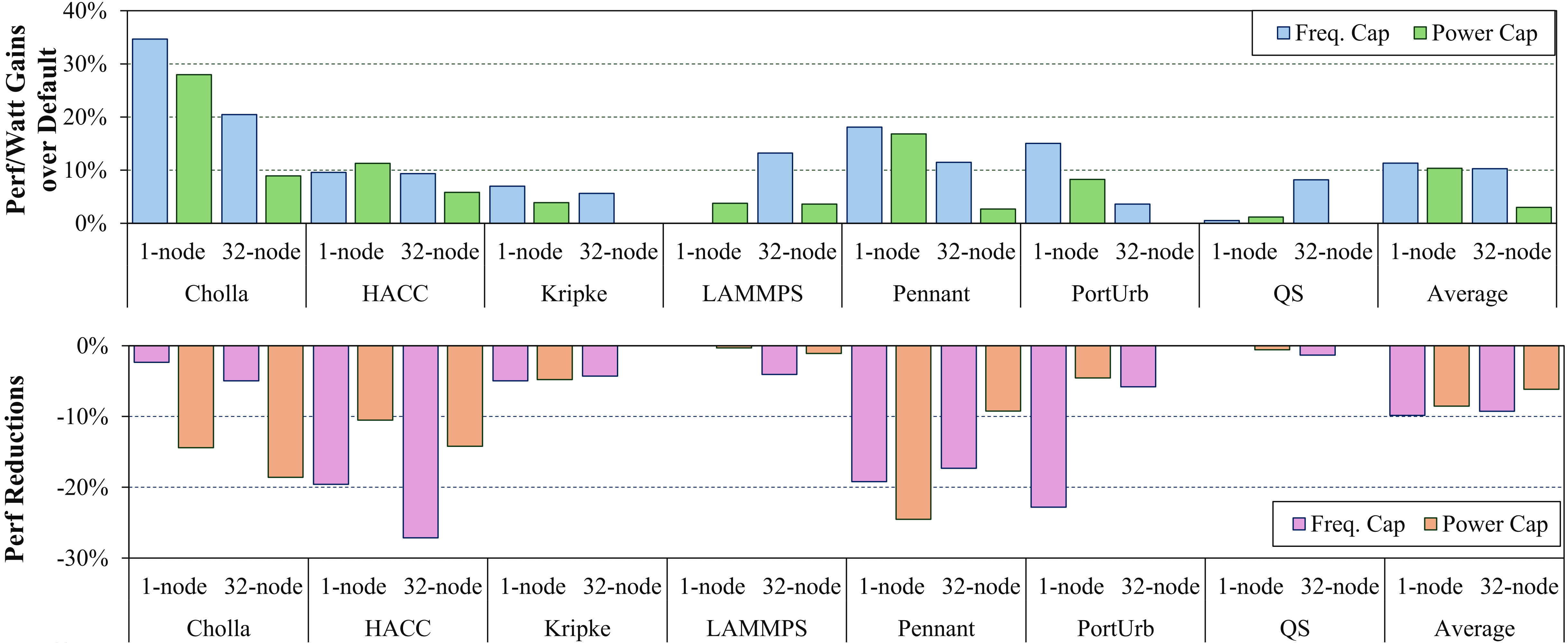




Evaluation: What is the impact on the Performance?



Evaluation: What is the impact on the Performance?



GPPD/UFRGS

Evaluation

- ~~What applications benefit more from Frequency Capping?~~
- ~~What applications benefit more from Power Capping?~~
- ~~What applications are impacted in similar ways?~~
- ~~What is the impact on the performance of applications?~~
- Discussion

Evaluation: Discussion

- **Practicality and Applicability**
  - Both power and frequency capping can improve energy efficiency without modifying application code.
  - Suitable for production workflows and system-level energy policies (runtime)

Evaluation: Discussion

- **Practicality and Applicability**
  - Both power and frequency capping can improve energy efficiency without modifying application code.
  - Suitable for production workflows and system-level energy policies (runtime)

- **Technique Behavior and Trade-offs**
  - Frequency Capping was more effective at large scale and for compute/memory-resilient applications
  - Power capping was better for bursty workloads with variable utilization.
  - Aggressive throttling should be avoided as it increases time-to-solution and reduces overall efficiency.

Evaluation: Discussion

- **Practicality and Applicability**
  - Both power and frequency capping can improve energy efficiency without modifying application code.
  - Suitable for production workflows and system-level energy policies (runtime)

- **Technique Behavior and Trade-offs**
  - Frequency Capping was more effective at large scale and for compute/memory-resilient applications
  - Power capping was better for bursty workloads with variable utilization.
  - Aggressive throttling should be avoided as it increases time-to-solution and reduces overall efficiency.

- Both methods are effective, but their benefit is workload dependent.

Evaluation: Discussion

- **Practicality and Applicability**
  - Both power and frequency capping can improve energy efficiency without modifying application code.
  - Suitable for production workflows and system-level energy policies (runtime)
- **Technique Behavior and Trade-offs**
  - Frequency Capping was more effective at large scale and for compute/memory-resilient applications
  - Power capping was better for bursty workloads with variable utilization.
  - Aggressive throttling should be avoided as it increases time-to-solution and reduces overall efficiency.
- Both methods are effective, but their benefit is workload dependent.
- Power management is a non-invasive, cost-effective path toward sustainable Exascale HPC.

(Ongoing) Future Works

- Understand the impact of hardware metrics on the power profile of each application

- Devise a model to automatically define the best power management technique

- Model a tool to change GPU frequency/power capping according to the active kernel
    - Implications on MPI communication, barriers, etc. etc..

Aknowledgments

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Advanced Scientific Computing Research programs in the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

This study was partially financed in part by the CAPES - Finance Code 001, FAPERGS - PqG 24/2551-0001388-1, and CNPq.

SC25

St. Louis, MO | hpc ignites.

16th International Workshop on

Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems

Characterizing the Impact of GPU Power Management on an Exascale System

Mariana Costa

Phillipe Navaux

Arthur Lorenzon

Bruno Alvarez

Jordà Polo

Antigoni Georgiadou

James B. White III

Woong Shin

Bronson Messer

GPPD/UFRGS

.Inf INSTITUTO DE INFORMÁTICA UFRGS PPGC

UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

AMD

OAK RIDGE National Laboratory