Improving energy efficiency of HPC applications using unbalanced GPU power capping

Albert d'Aviau de Piolant, Hayfa Tayeb, Berenger Bramas, Mathieu Faverge, Abdou Guermouche, Amina Guermouche

November 2024

Introduction: Context HPC

Frontier (United States)

- 1.206 ExaFlop/s
- 22.8 MegaWatt (w/o cooling)
- 9 482 AMD Epyc 7713
- 37 888 AMD Instinct MI250X

Consumption problem

In term of power and energy consumption, those supercomputers are really demanding. Frontier is the equivalent of 7 standard wind turbines.

Introduction: Power saving

Hardware solutions:

- DVFS
- UFS
- undervolting
- power capping

Software solutions:

- Scheduling on heterogeneous platform with energy criterion

Introduction: Power capping

Definition

Power capping is a way of limiting power consumption on a device.

- Intel CPUs: intel-rapl
- NVIDIA GPUs: nvidia-smi
- **MUST BE ROOT!**

TDP: **400 Watt** No power cap

GPU 1
Limit: **400 Watt**

TDP: **400 Watt** Power cap 75%

GPU 2
Limit: **300 Watt**

Motivation: Metrics

Performance and energy

- Performance is the number of operations per second.
- Energy is the product of power and time.

Energy efficiency

How well the energy is used.

- flop/Joule
- flop/second/Watt

Motivation: Methodology

Goal

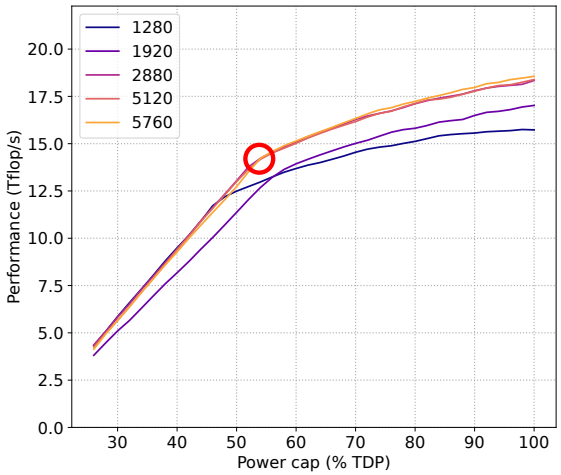
The goal is to see how the GPU behaves with limited power budget. Can we find the power that will give us the best energy efficiency?

- Grid5000
- Nancy - chuc-1
- NVIDIA TESLA A100 40GB
- cuBLAS
- dgemm

- static power capping (NVML) before running the kernel
- multiple standard sizes for gemm
- energy consumption before and after the kernel execution

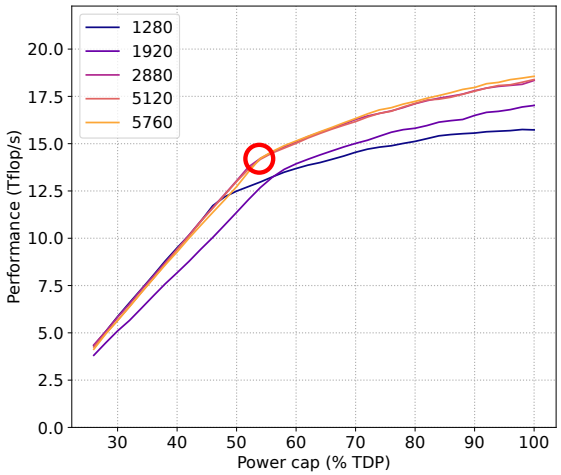
Motivation: Power capping on GPU, double precision

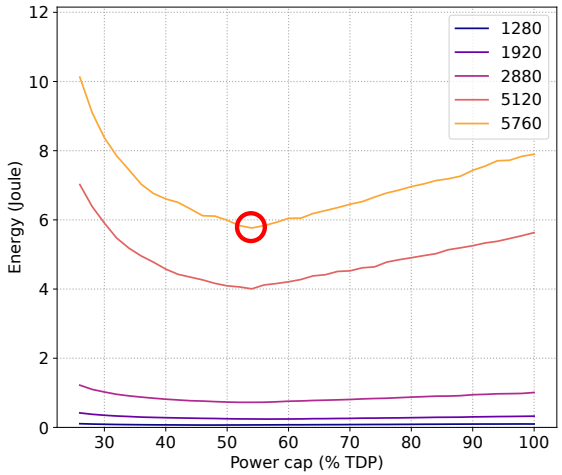
Matrix size comparison: Power capping on NVIDIA A100 (104 to 400 Watt) - cuBLAS gemm (double precision)



Motivation: Power capping on GPU, double precision

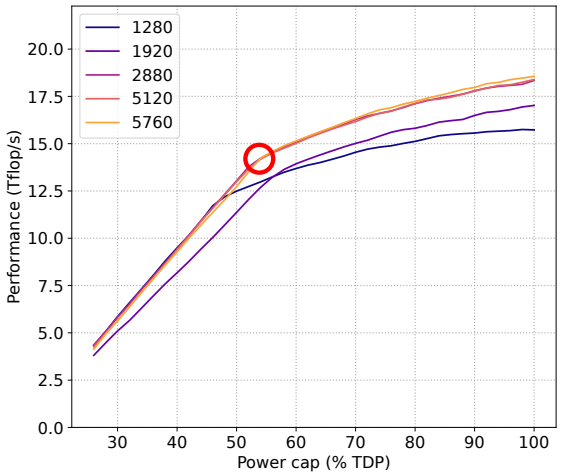
Matrix size comparison: Power capping on NVIDIA A100 (104 to 400 Watt) - cuBLAS gemm (double precision)

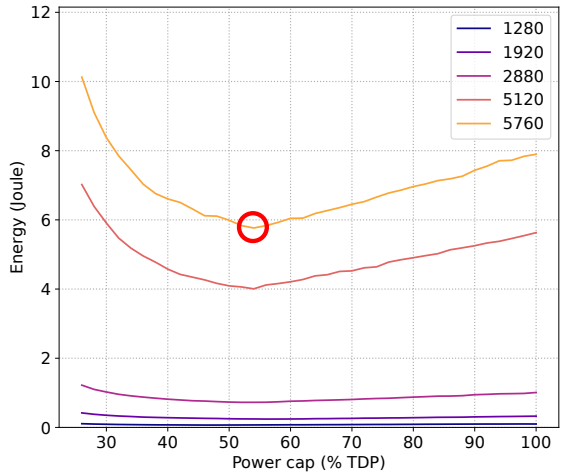


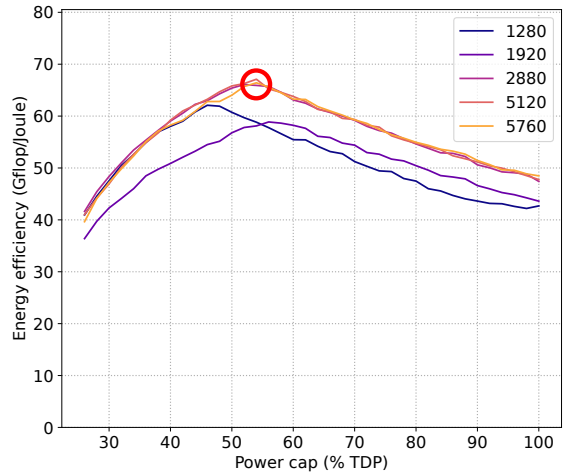


Motivation: Power capping on GPU, double precision

Matrix size comparison: Power capping on NVIDIA A100 (104 to 400 Watt) - cuBLAS gemm (double precision)

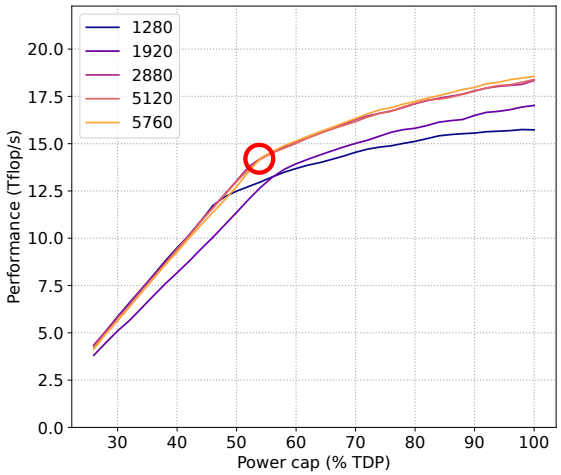


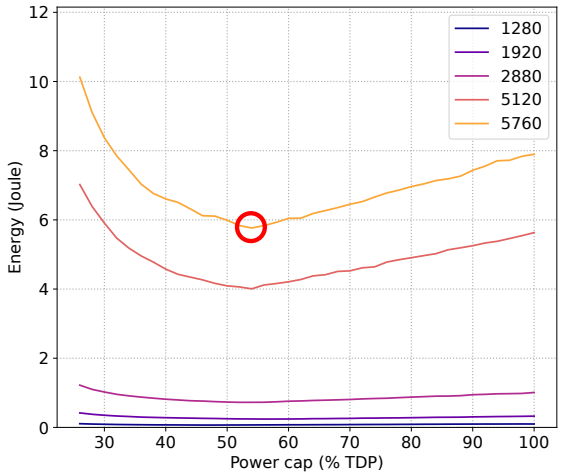


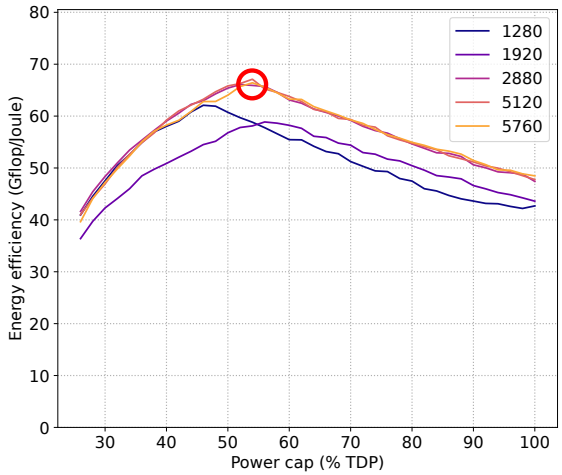


Motivation: Power capping on GPU, double precision

Matrix size comparison: Power capping on NVIDIA A100 (104 to 400 Watt) - cuBLAS gemm (double precision)





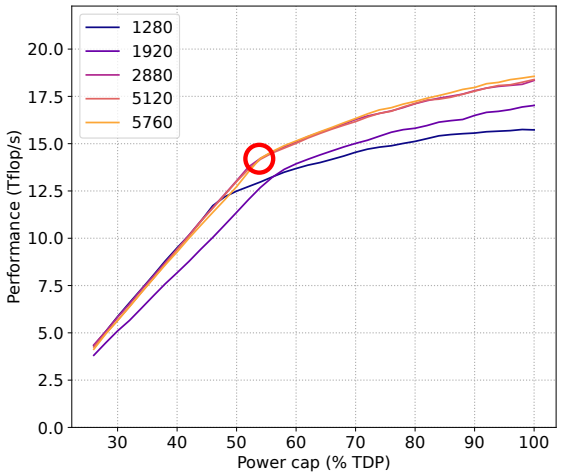


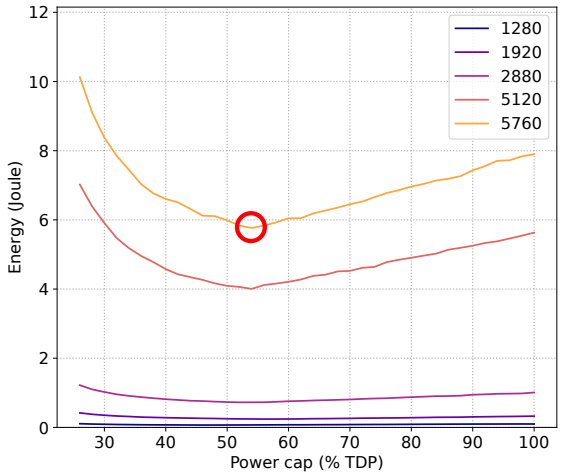
Highlights

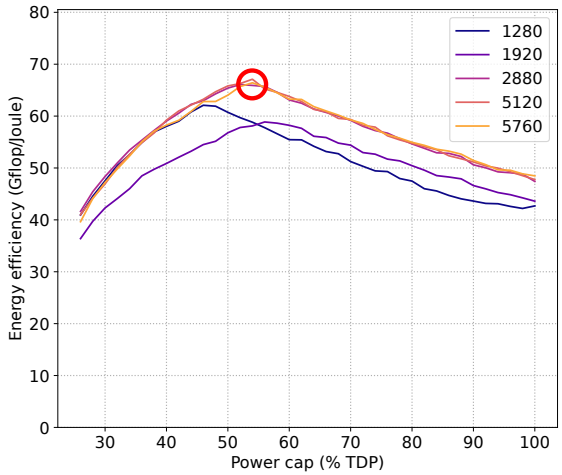
- Best energy efficiency is obtained at 54 % of the TDP (N = 5120).

Motivation: Power capping on GPU, double precision

Matrix size comparison: Power capping on NVIDIA A100 (104 to 400 Watt) - cuBLAS gemm (double precision)





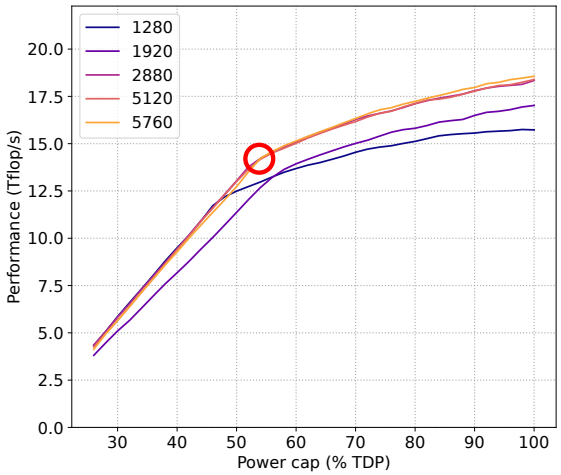


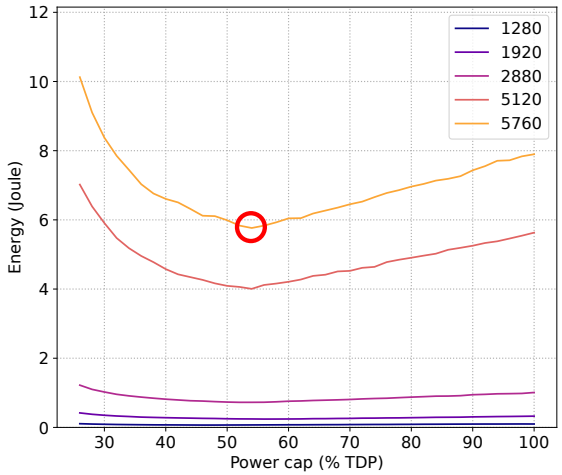
Highlights

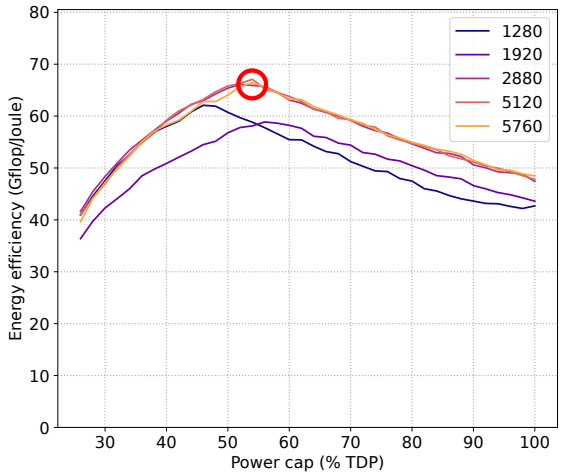
- Best energy efficiency is obtained at 54 % of the TDP (N = 5120).
- Saving of energy efficiency: 28.81 %.

Motivation: Power capping on GPU, double precision

Matrix size comparison: Power capping on NVIDIA A100 (104 to 400 Watt) - cuBLAS gemm (double precision)







Highlights

- Best energy efficiency is obtained at 54 % of the TDP (N = 5120).
- Saving of energy efficiency: 28.81 %.
- Drop of performance: 22.93 %.

Study conclusion

Conclusion

Faster is not equivalent to being energy efficient on GPU!

Now, what is the plan?

- Multi-GPU configurations
- Operations instead of kernels
- Existing schedulers

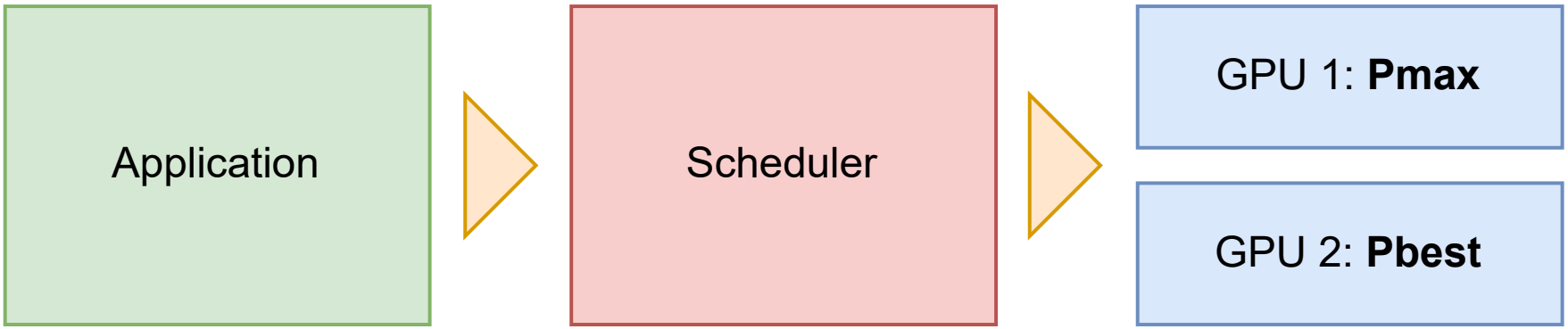
Power capping and scheduling

We would like to have a node with a heterogenous set of GPUs.

- Some GPUs for performance
- Some GPUs for energy efficiency
- Some GPUs at the minimum power cap

Goal

We want to analyse a State-of-the-Art scheduler with different configurations on an HPC application.



Experimental set-up: applications

Chameleon

Chameleon is a framework which provides routines to solve dense general systems of linear equation for HPC.

StarPU

StarPU is a software tool design for HPC. It provides the runtime and the schedulers for Chameleon.

- dmdas - based on performance models

Experimental set-up: hardware & configurations

chuc-1

- Lille
- 4 x SXM4 **A100**
- DGEMM
  Pbest: **54 %**
- Pmax: 400 Watt
- Pmin: 100 Watt

grouille-1

- Nancy
- 2 x PCIe **A100**
- DGEMM
  Pbest: **78 %**
- Pmax: 250 Watt
- Pmin: 150 Watt

chifflot-8

- Lille
- 2 x PCIe **V100**
- DGEMM
  Pbest: **62 %**
- Pmax: 250 Watt
- Pmin: 100 Watt

Configuration glossary

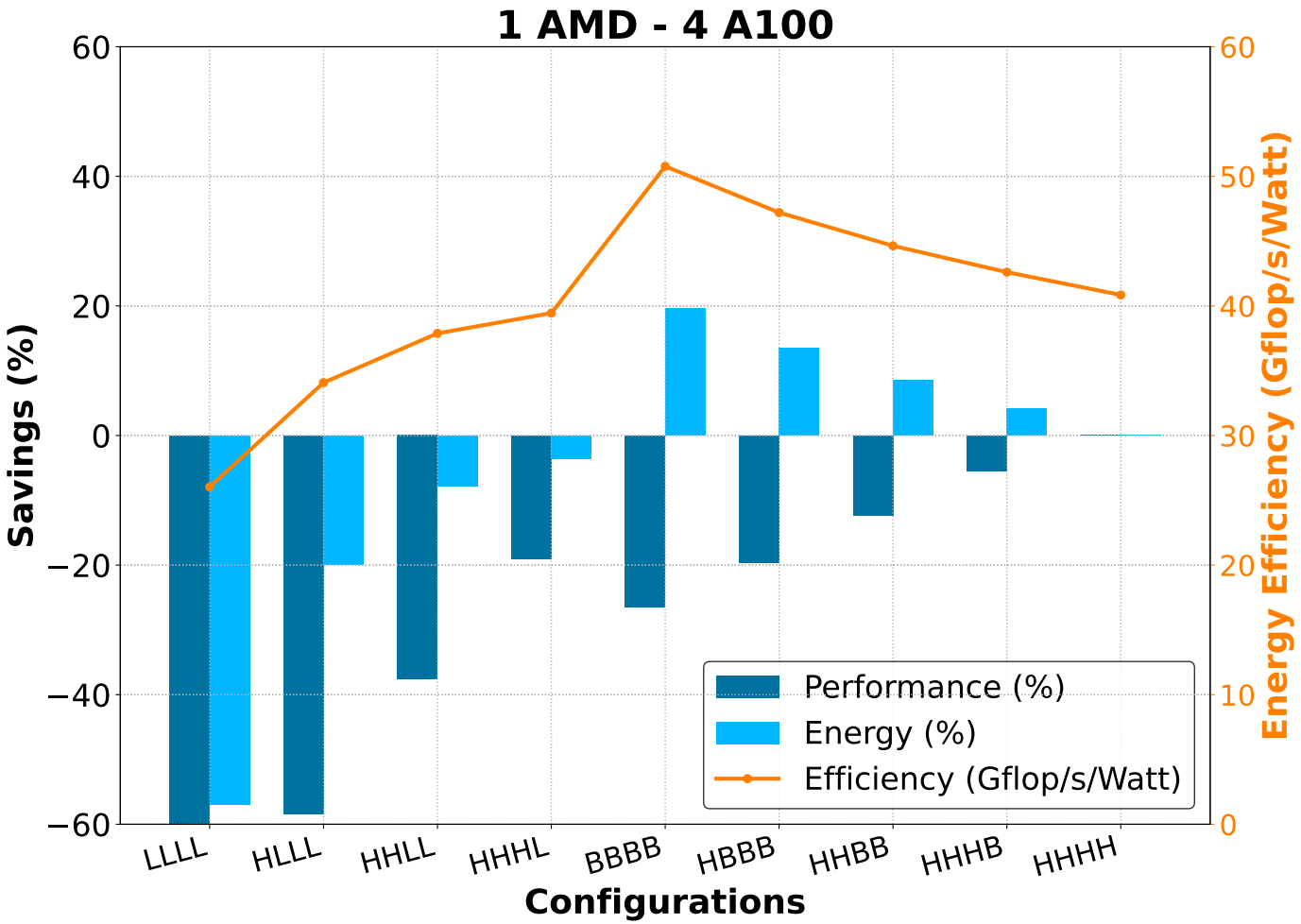
We tested different configurations according to the architecture. A letter represents one GPU in the configuration:

- H: (High) power cap set to 100 %.
- B: (Best) power cap set to the value that provides the best energy efficiency.
- L: (Low) power cap set to the minimum power cap possible.

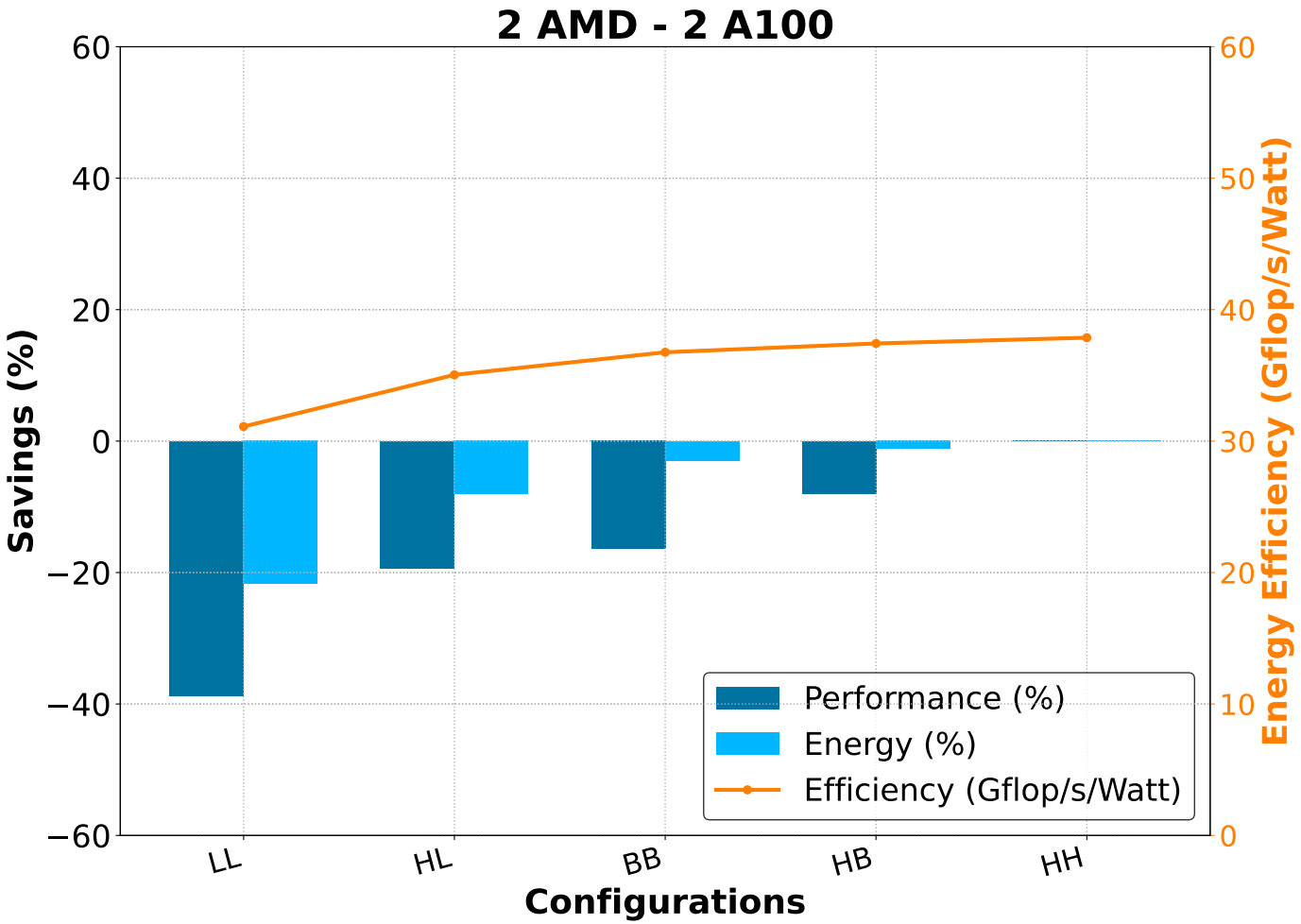
Example

HHHB: 3 GPUs set to 100 % of the TDP, 1 GPU set to the best power cap.

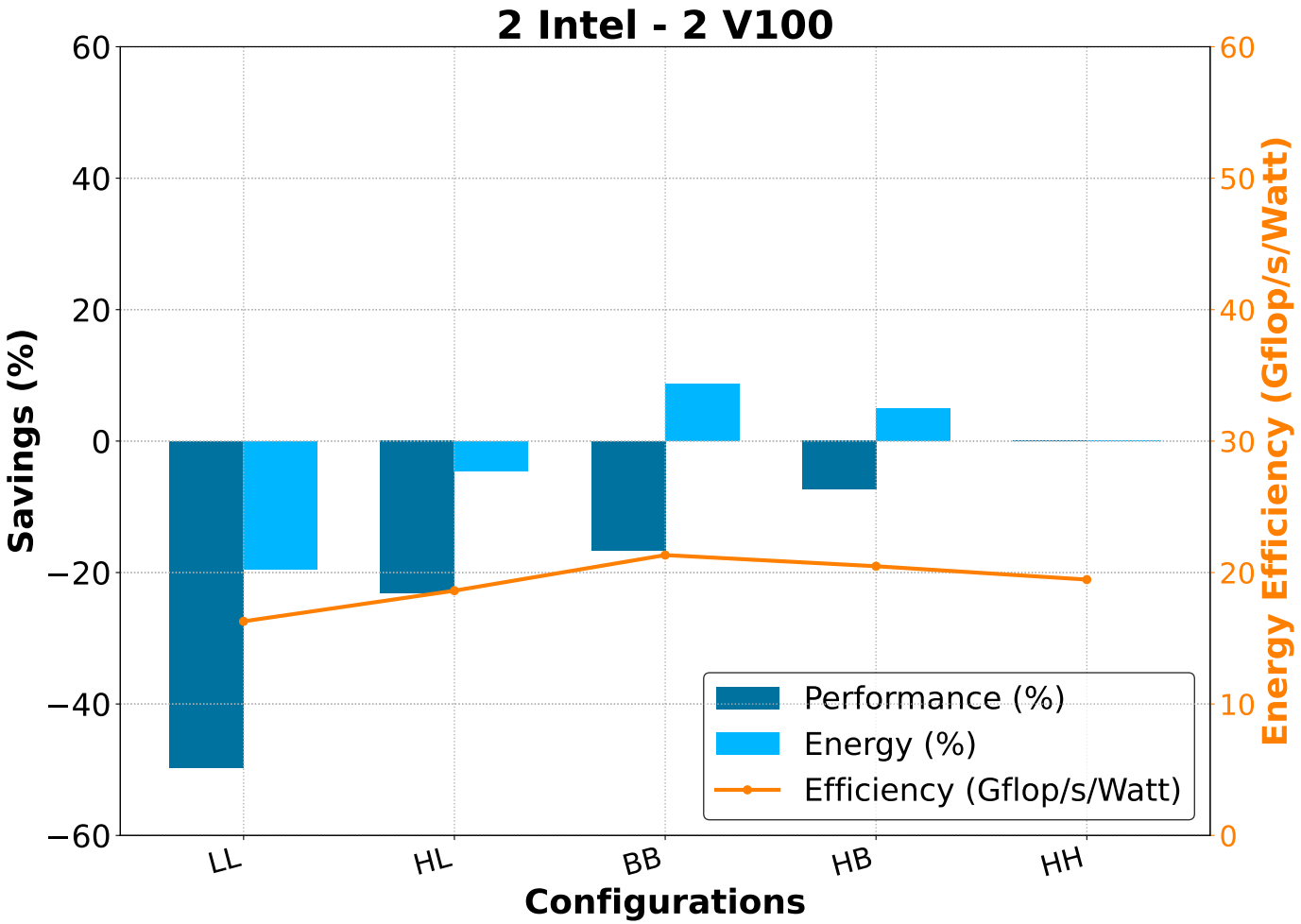
Results for scheduling under power capping



Results for scheduling under power capping

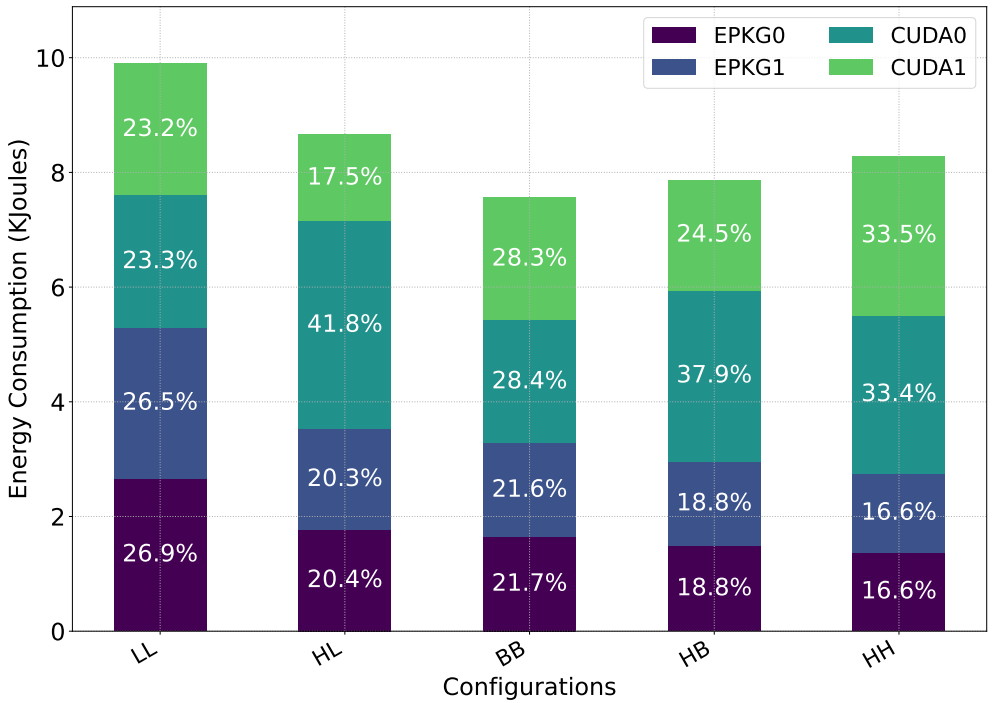


Results for scheduling under power capping



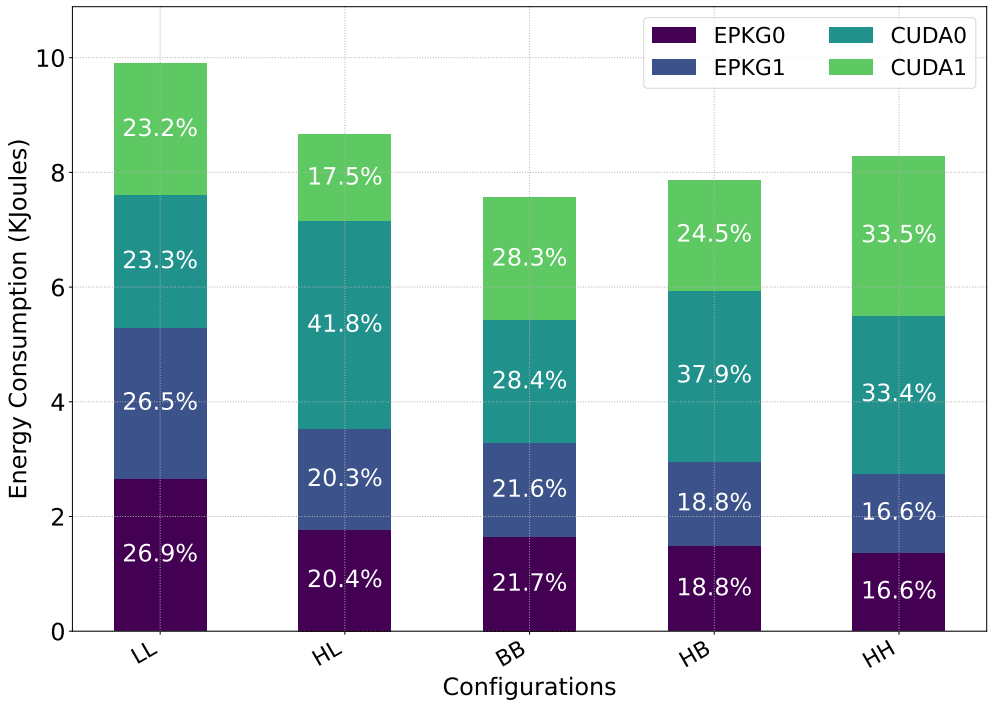
Power cap on CPU

What is the CPU's energy consumption ?



Power cap on CPU

What is the CPU's energy consumption ?

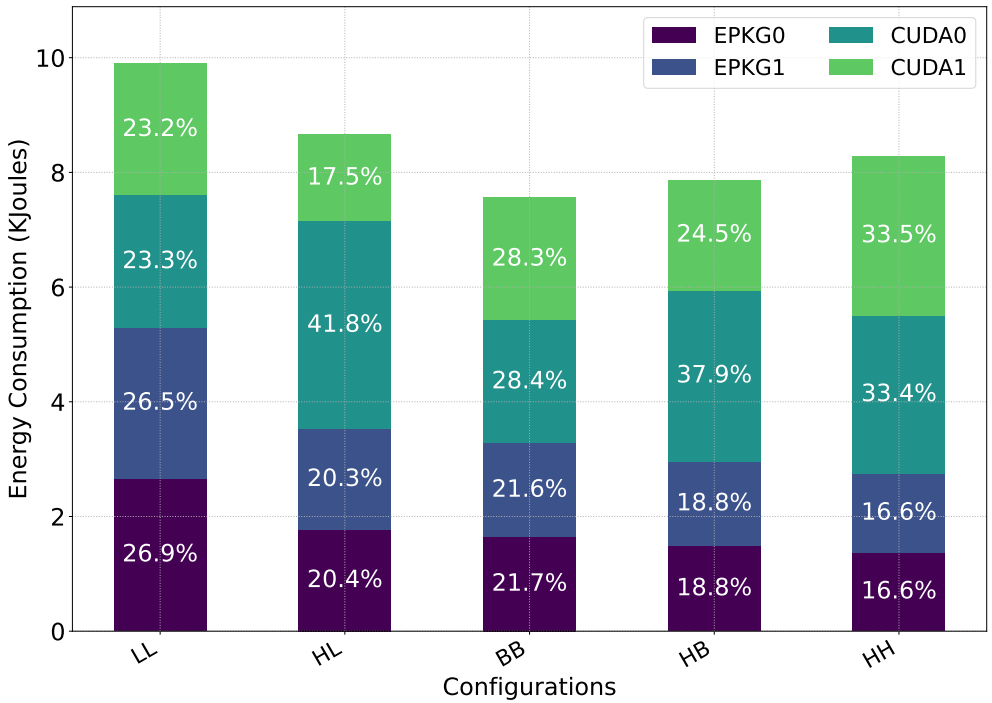


Reasons to power cap CPUs

- GPUs handle much more tasks than CPUs
- CPUs are not really used on compute intensive operations.

Power cap on CPU

What is the CPU's energy consumption ?



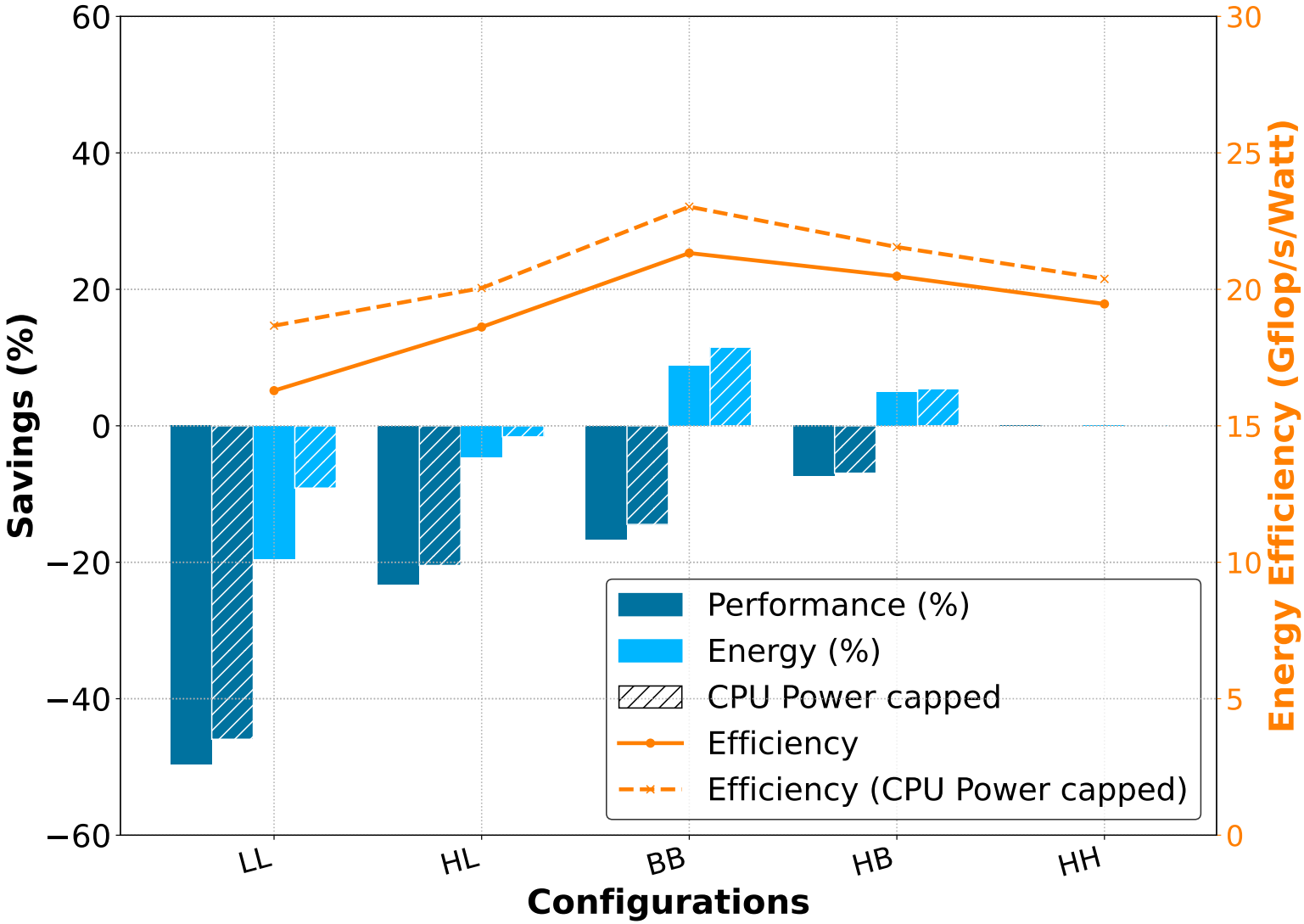
Reasons to power cap CPUs

- GPUs handle much more tasks than CPUs
- CPUs are not really used on compute intensive operations.

Constraint on CPU Power capping

Only on Intel CPUs: we have to use the V100 platform.

Results with CPU power capping



Conclusion

- Faster does not mean more energy efficient
- All GPUs at Pbest: energy efficiency improvement: 24.3%
  Slowdown: 26.4%
- CPU power capping: energy efficiency improvement: 8 % with no drop of performance.

Future works

- Dynamic power capping
- Mixed precision for energy efficiency
- Enhance existing scheduler