Reducing HPC energy footprint for large scale GPU accelerated workloads

Cray User Group 2023, Helsinki

Gabriel Hautreux <hautreux@cines.fr>
CINES – Montpellier - France

Reducing HPC energy footprint for large scale GPU accelerated workloads

Summary

□ Machine and studies definition

□ Turbo mode usage study

□ GPU frequency capping study

□ Power capping study

□ Conclusion and perspectives

Reducing HPC energy footprint for large scale GPU accelerated workloads

CINES : a national HPC center

➢CINES (National Computing Center for Higher Education)

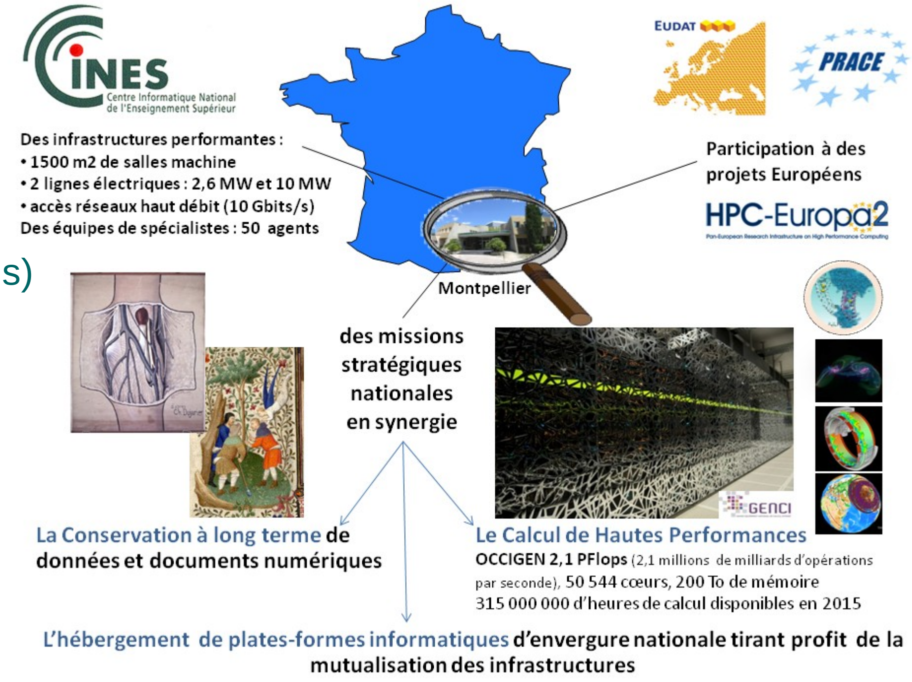
Based in Montpellier(South of France), supervised by French ministry for Higher Education and Research.

Three strategic missions :

- High Performance Computing
- Long term digital preservation
- National hosting entity (servers/platforms)

➢Nationel and European partnerships

- One of the three national centers, with IDRIS(CNRS) and TGCC(CEA)
- Member of PRACE



Des infrastructures performantes :
- 1500 m2 de salles machine
- 2 lignes électriques : 2,6 MW et 10 MW
- accès réseaux haut débit (10 Gbits/s)

Des équipes de spécialistes : 50 agents

Participation à des projets Européens

Montpellier

des missions stratégiques nationales en synergie

La Conservation à long terme de données et documents numériques

Le Calcul de Hautes Performances

OCCIGEN 2,1 PFlops (2,1 millions de milliards d'opérations par seconde), 50 544 cœurs, 200 To de mémoire
315 000 000 d'heures de calcul disponibles en 2015

L'hébergement de plates-formes informatiques d'envergure nationale tirant profit de la mutualisation des infrastructures

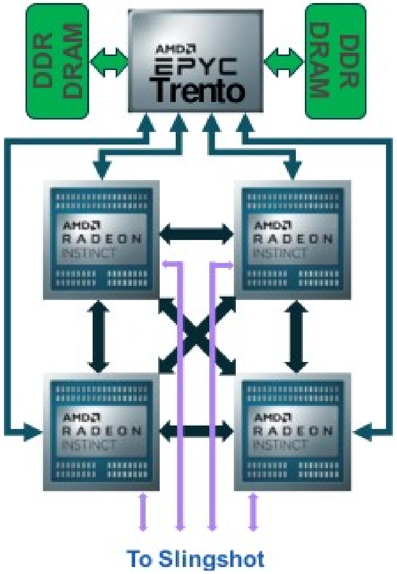
Reducing HPC energy footprint for large scale GPU accelerated workloads

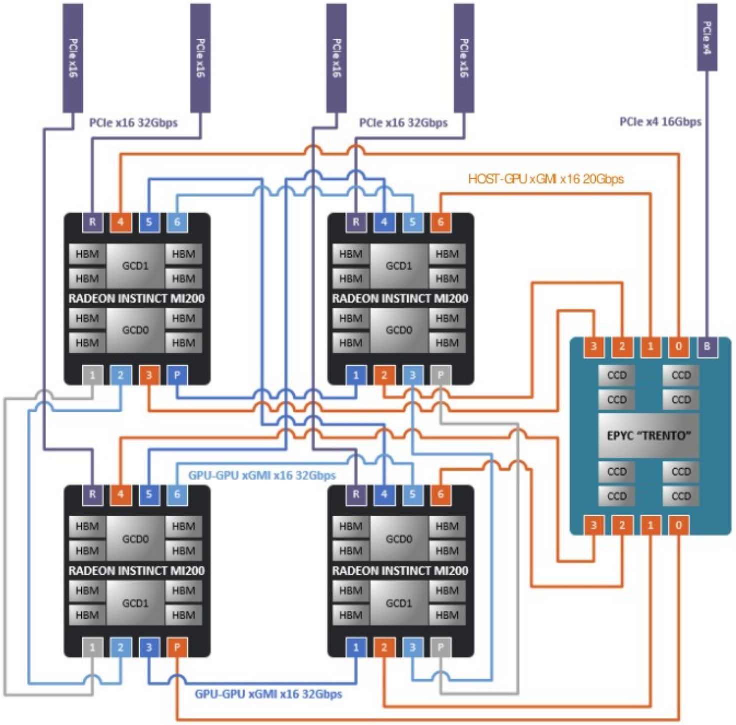
Adastra : enabling exascale technologies

- ❑ HPE Cray EX system
- ❑ AMD GPU + CPU
- ❑ #10 Top 500 (June 22)
- ❑ #3 Green 500 (Nov 22)

Accelerated Partition

- ❑ LUMI/Frontier like system
- ❑ 338 nodes
- ❑ AMD Trento 64 cores, 2.4 GHz, 256 Go DDR4-3200 + 4 GPU AMD MI250X, 4x128 Go HBM2, 4 Slingshot 200 Gbps
  - Infinity fabric
  - ~200Tflops per node





Reducing HPC energy footprint for large scale GPU accelerated workloads

HPC workload

❑ **Scientific workload**

- Based on widely-used codes within the French research community
- **Can run N times a year**
- **Consumes E energy per run**
- **Consumes a total of N*E per year**

❑ **Reported results**

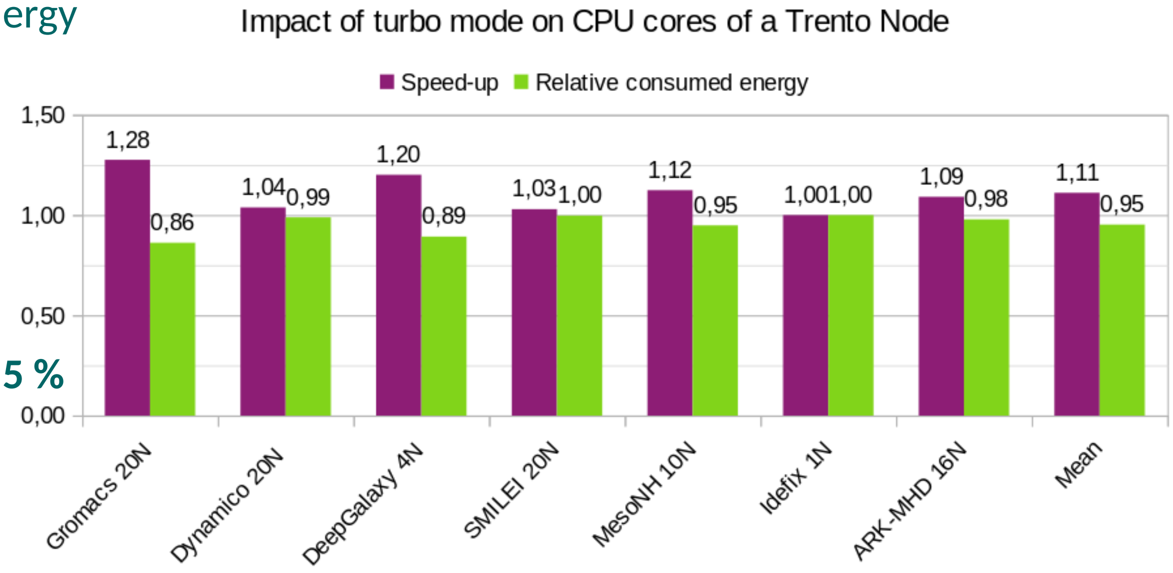
- Time To Solution (**TTS**) using « time » command
- Energy To Solution (**ETS**) using Slurm (energy at nodes level)

Reducing HPC energy footprint for large scale GPU accelerated workloads

CPU Turbo mode usage study

❑ **Run performed turbo mode on/off on CPU**

- Turbo off as base for speed-up and relative consumed energy



❑ **Speed-up : 11 %**

❑ **Energy gain for 1 workload : 5 %**

❑ Over the year, we will then run 1.1*N times our workload, using each time 0.95*E of energy.

$$1.1*0.95*N*E=1.045*N*E$$

❑ **The energy gain per workload implies a global increase of 4,5 % of energy consumed per year**

-

Reducing HPC energy footprint for large scale GPU accelerated workloads

Frequency and power capping studies

❑ **CINES developed a small tool for the energy study (ERIS)**

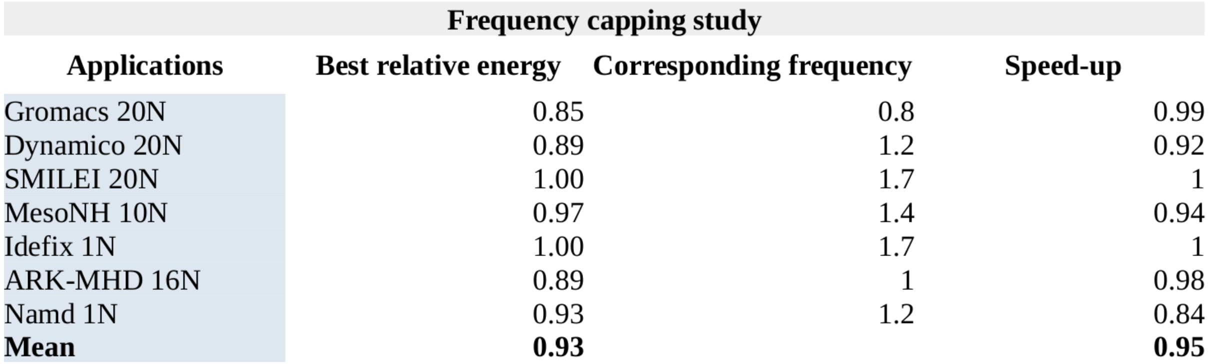
- Cap frequency and power using different values on the same nodelist to minimize noise
- Frequency values : [0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7] GHz
- Power capping values : [300, 350, 400, 450, 500, 560] W (per GPU)

❑ **Slurm plugins to cap frequency and power**

- Frequency : using rocm-smi
- Power : GPU device driver values

Frequency capping study

- ❑ **Applications run for all defined frequency**
  - Find the best frequency to minimize energy



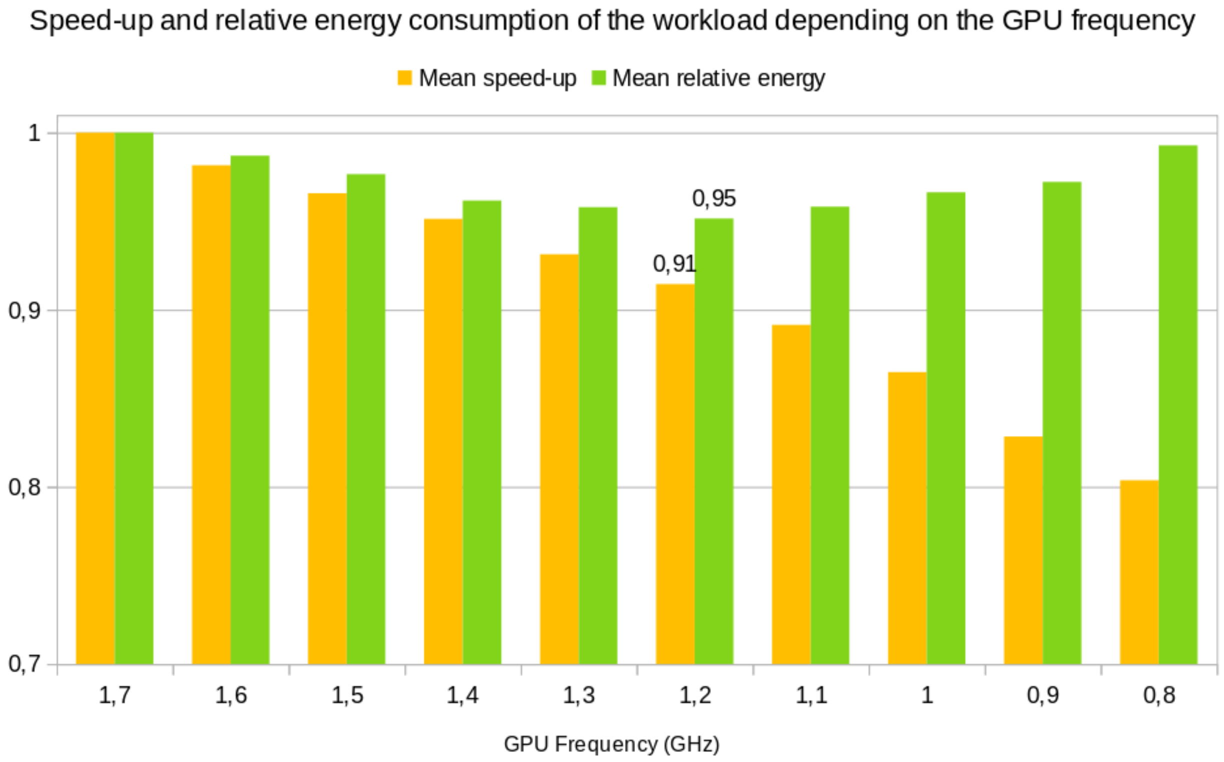
Frequency capping study			
Applications	**Best relative energy**	**Corresponding frequency**	**Speed-up**
Gromacs 20N	0.85	0.8	0.99
Dynamico 20N	0.89	1.2	0.92
SMILEI 20N	1.00	1.7	1
MesoNH 10N	0.97	1.4	0.94
Idefix 1N	1.00	1.7	1
ARK-MHD 16N	0.89	1	0.98
Namd 1N	0.93	1.2	0.84
Mean	**0.93**		**0.95**

- ❑ **This specific configuration is called « Fine-Tuned »**
  - Energy gain : 7 %
  - Speed-up : -5 %
- ❑ **Energy gain greater than performance loss**

Reducing HPC energy footprint for large scale GPU accelerated workloads

Frequency capping study

❑ Workload behavior regarding frequency



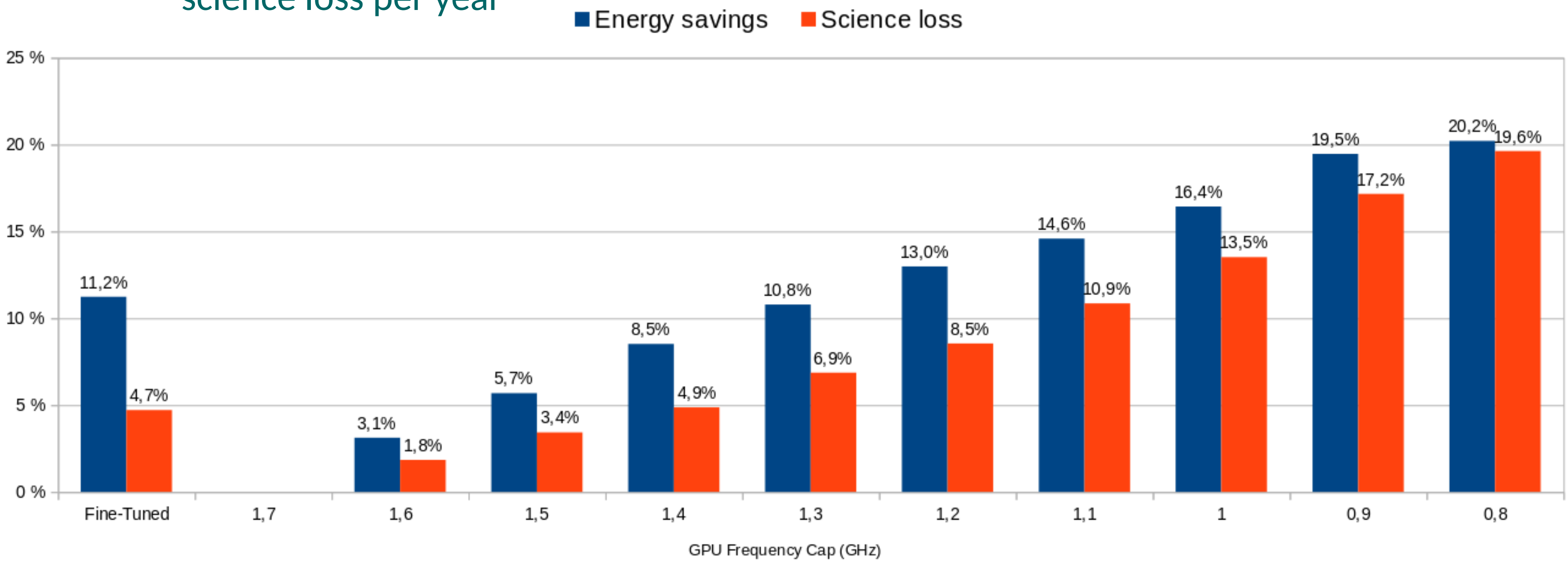
❑ Sweet spot at 1,2GHz, 5 % energy gain per workload but 9 % of performance loss

Reducing HPC energy footprint for large scale GPU accelerated workloads

Frequency capping study

❑ **Impact on the global energy consumption <u>per year</u>**

- Apply the previous values over the year (N*E), performance loss per workload becomes science loss per year



- The 1,2GHz value enables us 13 % energy savings over the year, by losing « only » 8,5 % of science

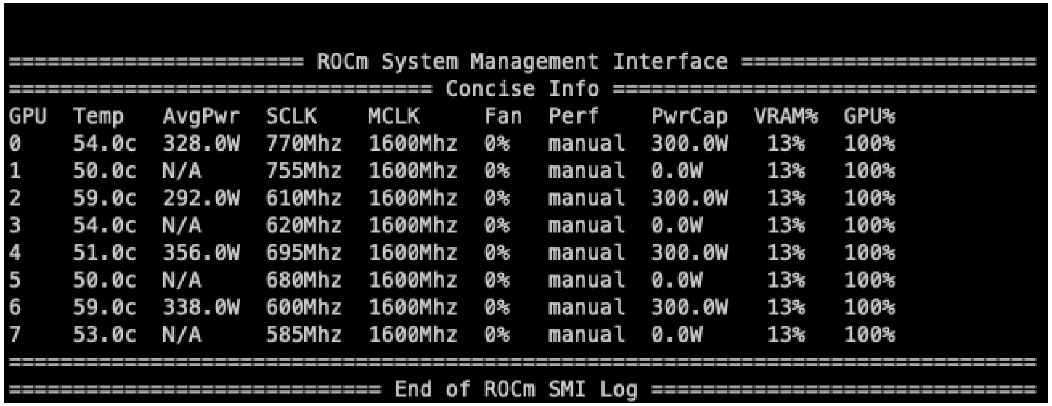
- Fine-Tuned is better in performance, not in energy

Reducing HPC energy footprint for large scale GPU accelerated workloads

Power capping study

❑ **Issues with power capping**

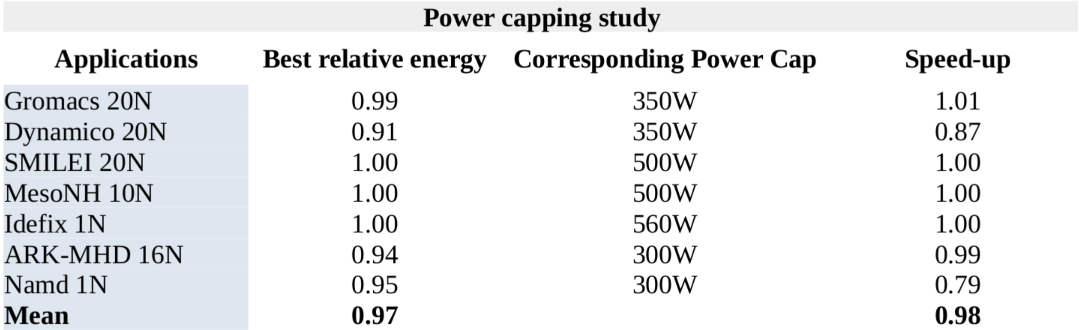
- Cap is not « hard »
- Spikes seen :
  e.g. 356W while 300W cap



```
======================= ROCm System Management Interface ========================
================================= Concise Info ==================================
GPU  Temp   AvgPwr  SCLK    MCLK     Fan  Perf    PwrCap  VRAM%  GPU%
0    54.0c  328.0W  770Mhz  1600Mhz  0%   manual  300.0W   13%   100%
1    50.0c  N/A     755Mhz  1600Mhz  0%   manual  0.0W     13%   100%
2    59.0c  292.0W  610Mhz  1600Mhz  0%   manual  300.0W   13%   100%
3    54.0c  N/A     620Mhz  1600Mhz  0%   manual  0.0W     13%   100%
4    51.0c  356.0W  695Mhz  1600Mhz  0%   manual  300.0W   13%   100%
5    50.0c  N/A     680Mhz  1600Mhz  0%   manual  0.0W     13%   100%
6    59.0c  338.0W  600Mhz  1600Mhz  0%   manual  300.0W   13%   100%
7    53.0c  N/A     585Mhz  1600Mhz  0%   manual  0.0W     13%   100%
=================================================================================
============================= End of ROCm SMI Log ===============================
```

❑ **Run performed for all the power values**

- Max energy gain : 9 %
- Global energy gain : 3 %
- Impact on perf : 2 %



Power capping study

Applications	Best relative energy	Corresponding Power Cap	Speed-up
Gromacs 20N	0.99	350W	1.01
Dynamico 20N	0.91	350W	0.87
SMILEI 20N	1.00	500W	1.00
MesoNH 10N	1.00	500W	1.00
Idefix 1N	1.00	560W	1.00
ARK-MHD 16N	0.94	300W	0.99
Namd 1N	0.95	300W	0.79
Mean	**0.97**		**0.98**

❑ **Due to limited gain and uncertainty on the capping, we prefered to drop this study for now**

Reducing HPC energy footprint for large scale GPU accelerated workloads

Conclusion

❑ **Make sure to enable turbo mode on your Trento nodes !**

- Better performances per run, less power consumption per run
- Expect a global rise for your annual bill…

❑ **Frequency capping is more reliable than power capping (for now)**

- Hope the methodology presented can be reproduced

❑ **Metrics for decision making at the political level**

- Know we know how much science do we loss by reducing our energy footprint

❑ **The per-app frequency defining approach is the most efficient**

- But requires a lot of time
- Classification of applications could be considered for better decision

Q/A

Thank you for your attention.

hautreux@cines.fr