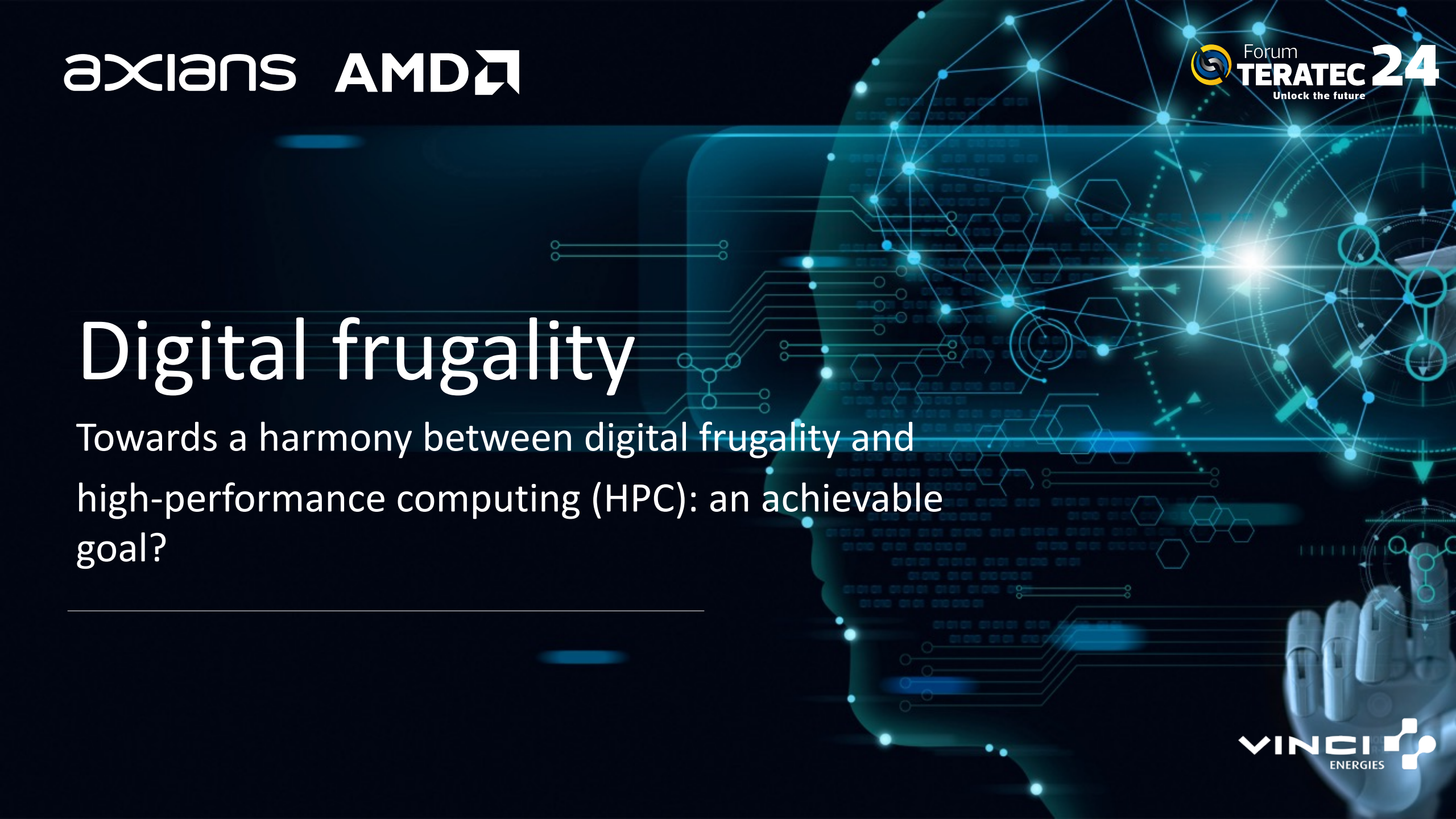


Digital frugality

Towards a harmony between digital frugality and high-performance computing (HPC): an achievable goal?

Thomas BOUISSOU

HPC Consultant

Thomas.bouissou@axians.com

— *PhD. Theoretical Quantum Chemistry*

— *Expertise: installation of HPC solutions, benchmarking, audit, design, […]*

Aurélien ORTIZ

HPC Architect / Expert

aurelien.ortiz@axians.com

— *PhD. Computer Science (Grid computing & distributed systems)*

— *Expertise: installation of HPC solutions, benchmarking, audit, design, […]*

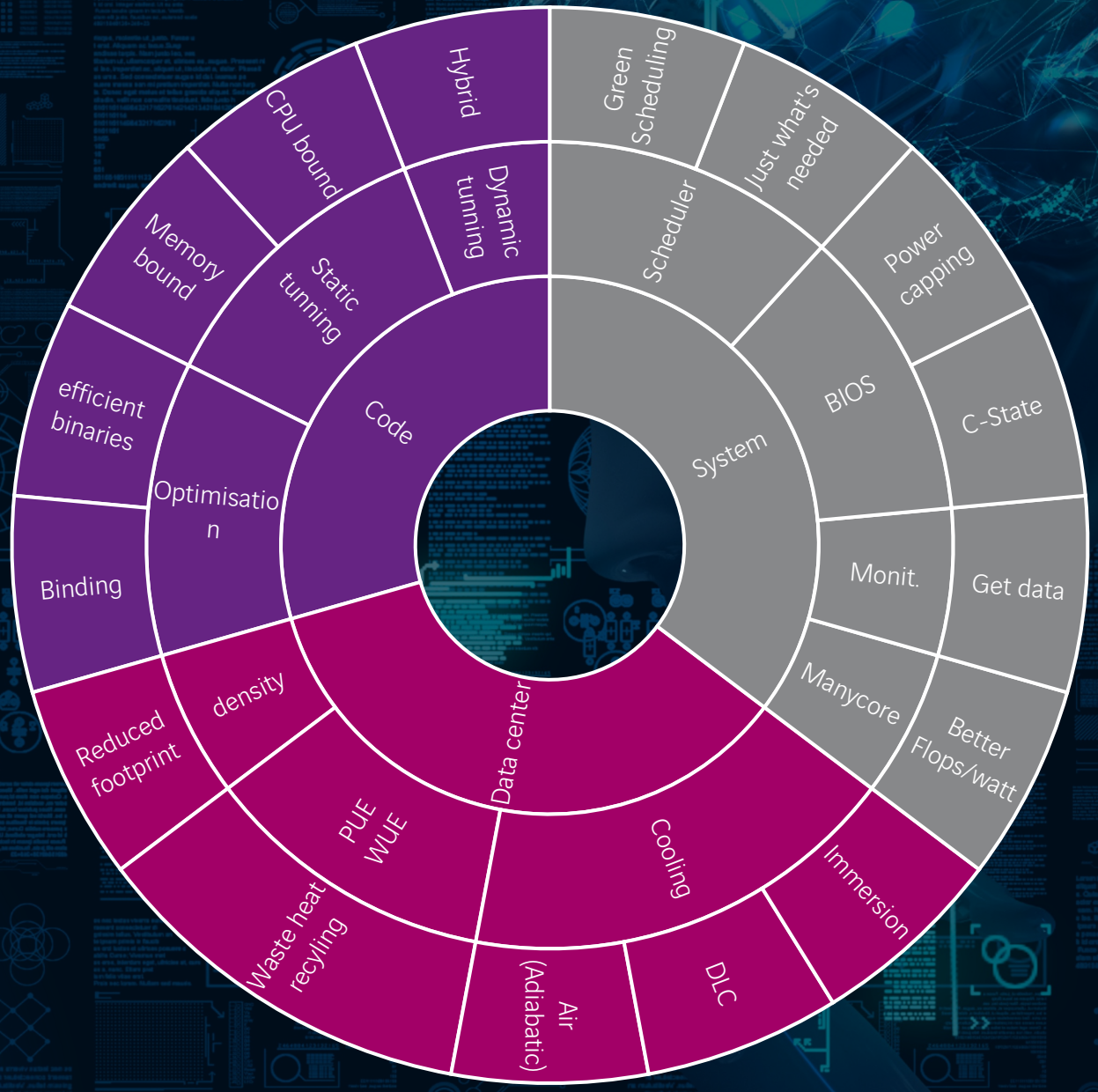
Cyril LAURIE

Sr. Manager EMEA FAE

cyril.laurie@amd.com

— *Managing AMD team of Technology experts in EMEA*

FINDING A PATH TO ENERGY EFFICIENCY



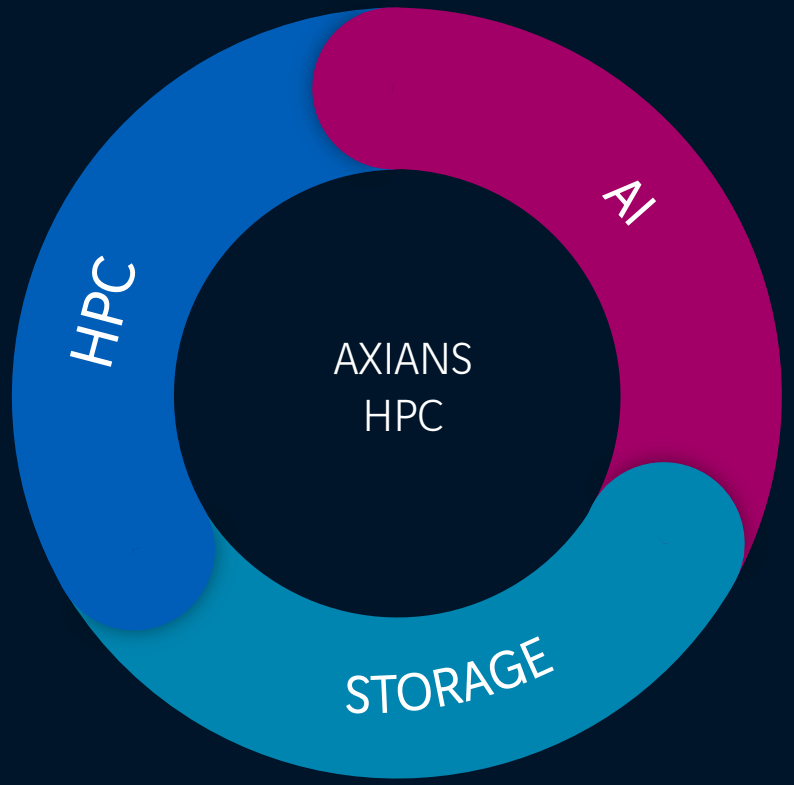
Axians HPC ...

... who are we?

What we do

High Performance Computing

Compute clusters
Application integration
Batch system tuning
...



Artificial Intelligence

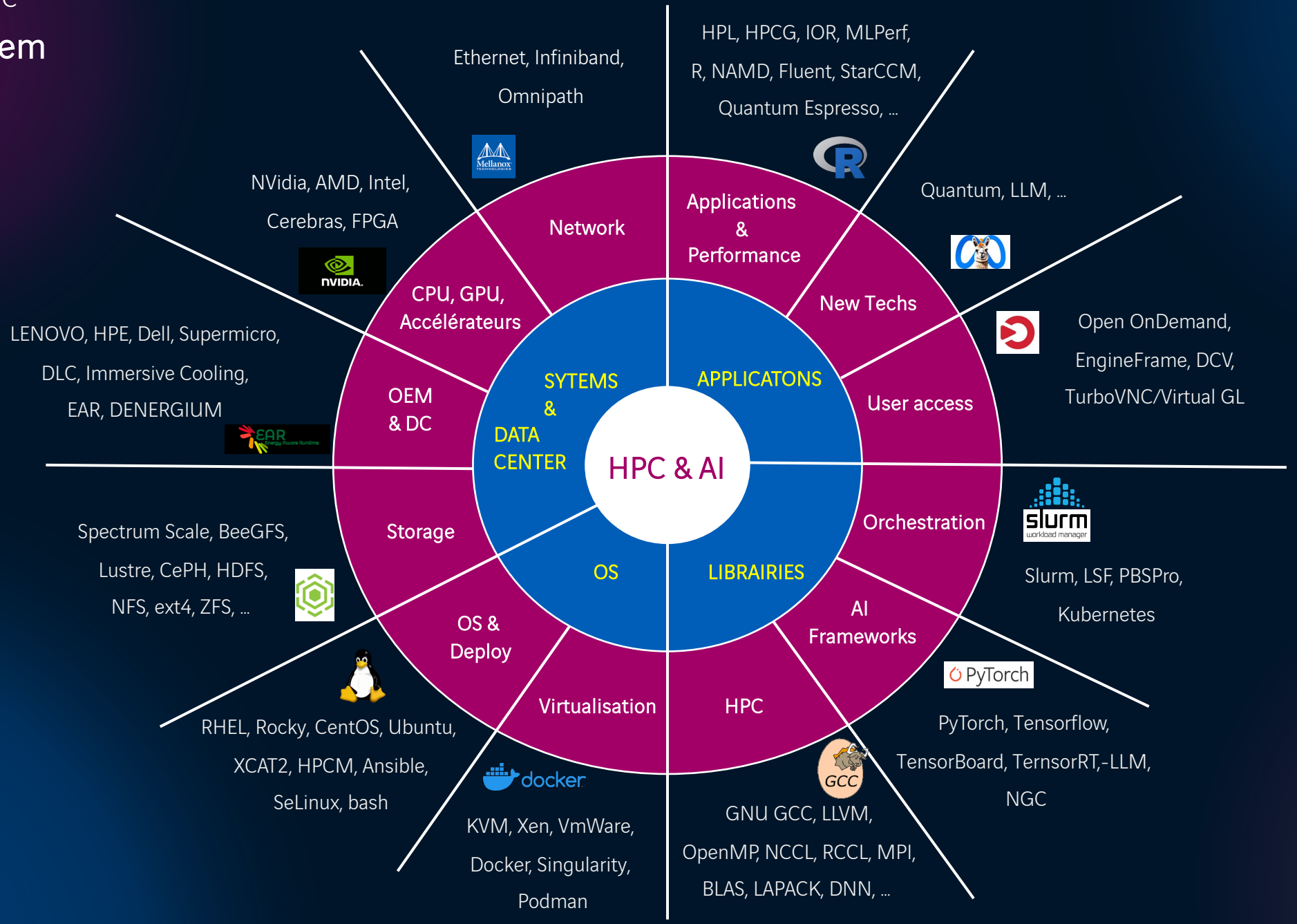
Deep Learning Training
LLM Tuning & Inference
...

Storage

Parallel File Systems (GPFS, Weka, ...)
Object Storage
...

Conception → **Build & Expertise** → **Support**

Ecosystem


HPC & AI
SYTEMS & DATA CENTER
APPLICATONS
OS
LIBRAIRIES
Network
Applications & Performance
CPU, GPU, Accélérateurs
New Techs
OEM & DC
User access
Storage
Orchestration
OS & Deploy
AI Frameworks
Virtualisation
HPC
Ethernet, Infiniband, Omnipath
HPL, HPCG, IOR, MLPerf, R, NAMD, Fluent, StarCCM, Quantum Espresso, ...
NVidia, AMD, Intel, Cerebras, FPGA
Quantum, LLM, ...
LENOVO, HPE, Dell, Supermicro, DLC, Immersive Cooling, EAR, DENERGIUM
Open OnDemand, EngineFrame, DCV, TurboVNC/Virtual GL
Spectrum Scale, BeeGFS, Lustre, CePH, HDFS, NFS, ext4, ZFS, ...
Slurm, LSF, PBSPro, Kubernetes
RHEL, Rocky, CentOS, Ubuntu, XCAT2, HPCM, Ansible, SeLinux, bash
PyTorch, Tensorflow, TensorBoard, TernsorRT,-LLM, NGC
KVM, Xen, VmWare, Docker, Singularity, Podman
GNU GCC, LLVM, OpenMP, NCCL, RCCL, MPI, BLAS, LAPACK, DNN, ...

HIGH DENSITY ARCHITECTURE…

~70 W → ~350 W

Processeurs

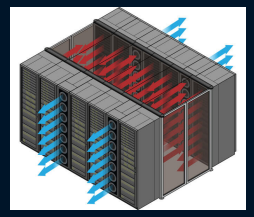
~150 W → ~700 W

GPU

~300 W → +1000 W

Serveur

Consommation par rack de calcul (kW)



~ 15kW ~ 20 kW ~ 30 kW ~ 40 kW + 50 kW

...NEED ADVANCED INTEGRATION

axians

INTEGRATION

SOFTWARE (+Open Source)

redhat

IBM Spectrum Scale

xCAT

influxdb
chronograf
telegraf
Grafana

+

DATA CENTER

What tools do we use …

… to measure energy?

IPMITOOL

- Server power consumption
- Requires root privileges (sudo is an option…)
- Integration with Monitoring systems (Grafana/Telegraf)
- Integration with Slurm
  - Best option for user reading
  - A huge limitation, however…

```
root@node01:~ # ipmitool dcmi power reading

    Instantaneous power reading:                   430 Watts
    Minimum during sampling period:                428 Watts
    Maximum during sampling period:                432 Watts
    Average power reading over sample period:      430 Watts
    IPMI timestamp:                           Wed May 22 12:25:23 2024
    Sampling period:                          00000001 Seconds.
    Power reading state is:                   activated
```

SLURM limitation?

- Only relevant for exclusive jobs…

```
axians@node01:~ # sbatch -w node02 -N 1 -n 127 --mem=500G stress_cpu.sbatch
axians@node01:~ # sbatch -w node02 -N 1 -n 1 --mem=500G stress_cpu.sbatch
Submitted batch job 806
Submitted batch job 807
```

```
axians@node01:~/stress # sacct -j 806 --format=JobID,JobName,AllocCPUs,CPUTime,MaxRSS,elapsed,ConsumedEnergy
JobID           JobName  AllocCPUS    CPUTime     MaxRSS    Elapsed ConsumedEnergy
------------ ---------- ---------- ---------- ---------- ---------- --------------
806          stress_cpu        127   02:11:14              00:01:02
806.batch        batch         127   02:11:14      8836K   00:01:02         51.43K
```

```
axians@node01:~/stress # sacct -j 807 --format=JobID,JobName,AllocCPUS,CPUTime,MaxRSS,elapsed,ConsumedEnergy
JobID           JobName  AllocCPUS    CPUTime     MaxRSS    Elapsed ConsumedEnergy
------------ ---------- ---------- ---------- ---------- ---------- --------------
807          stress_cpu          1   00:01:02              00:01:02
807.batch        batch           1   00:01:02      1280K   00:01:02         51.43K
```

EAR (Energy Aware Runtime)

- Use ipmitool to read the power consumption
- Store jobs in a MySQL Database
- Consolidate values after job completion
  - *Therefore, not available immediately at the end of the job!*
- 2 modes:
  - Monitoring only
  - Energy optimization
- Sometimes buggy…

How it works?

- Non-exclusive jobs are handled using CPU/Core allocation ratio

```
axians@node01:~ # sbatch -w node02 -N 1 -n 127 --mem=500G stress_cpu.sbatch
axians@node01:~ # sbatch -w node02 -N 1 -n 1 --mem=500G stress_cpu.sbatch
Submitted batch job 806
Submitted batch job 807
```

```
axians@node01:~ # eacct -j 806
    JOB-STEP USER       APPLICATION      POLICY NODES AVG/DEF/IMC(GHz) TIME(s)    POWER(W) GBS     CPI   ENERGY(J)    GFLOPS/W IO(MBs) MPI%  G-POW (T/U)     G-FREQ  G-
UTIL(G/MEM)
    806-sb   axians     stress_cpu       NP     1     2.79/1.90/---    62.00      709.11   ---     ---   43965        ---      ---     ---   ---             ---     ---
```

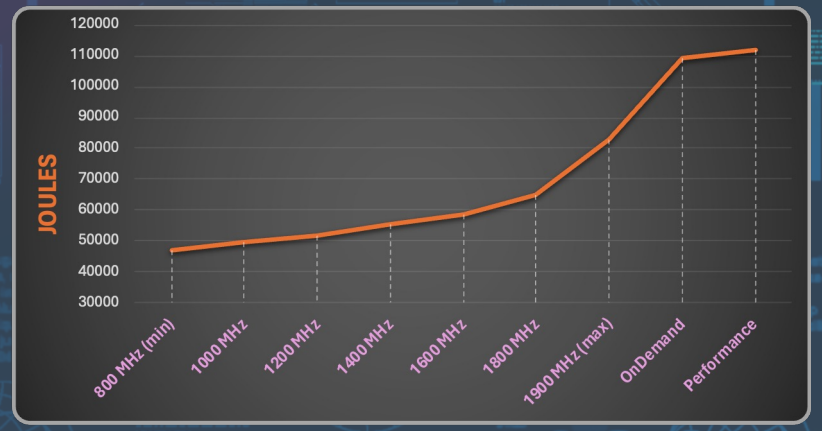
```
axians@node01:~ # eacct -j 807
    JOB-STEP USER       APPLICATION      POLICY NODES AVG/DEF/IMC(GHz) TIME(s)    POWER(W) GBS     CPI   ENERGY(J)    GFLOPS/W IO(MBs) MPI%  G-POW (T/U)     G-FREQ  G-
UTIL(G/MEM)
    807-sb   axians     stress_cpu       NP     1     2.79/1.90/---    62.00      50.00    ---     ---   3100         ---      ---     ---   ---             ---     ---
```

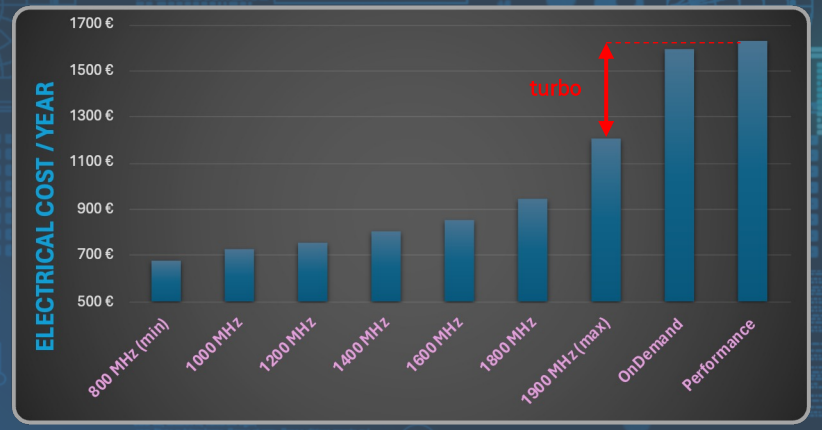
Best practice ...

... from an HPC user perspective

axians

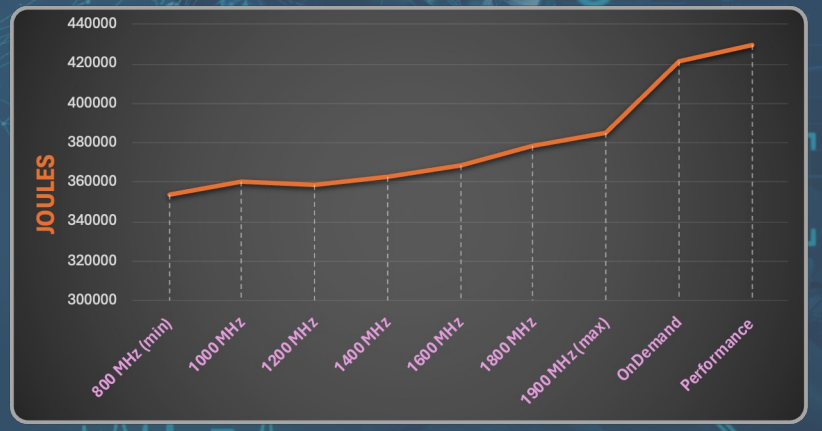
stress-ng --fma (cpu only)

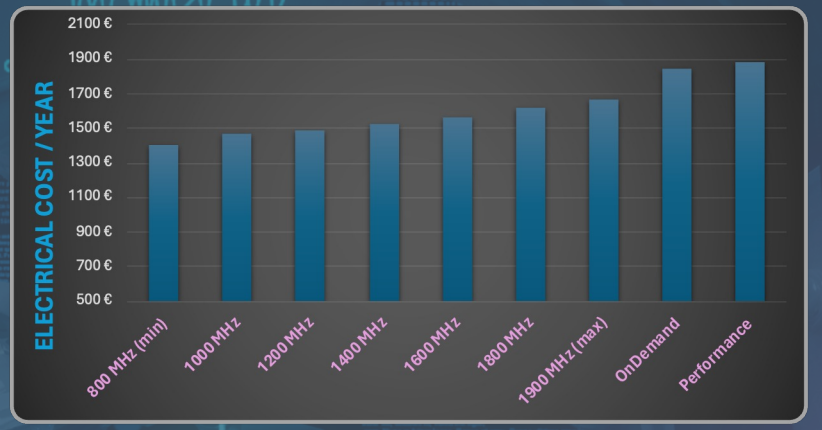




1KWh = 0.20€

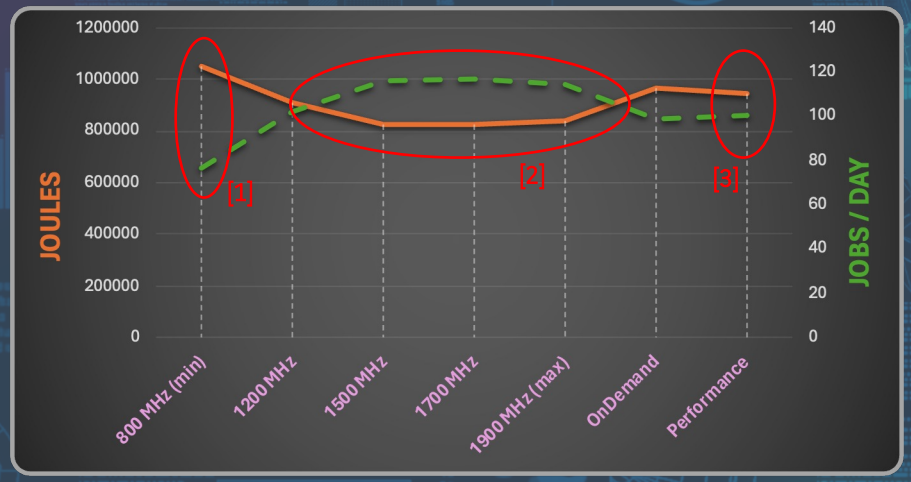
likwid triad_avx512_fma (cpu+memory)

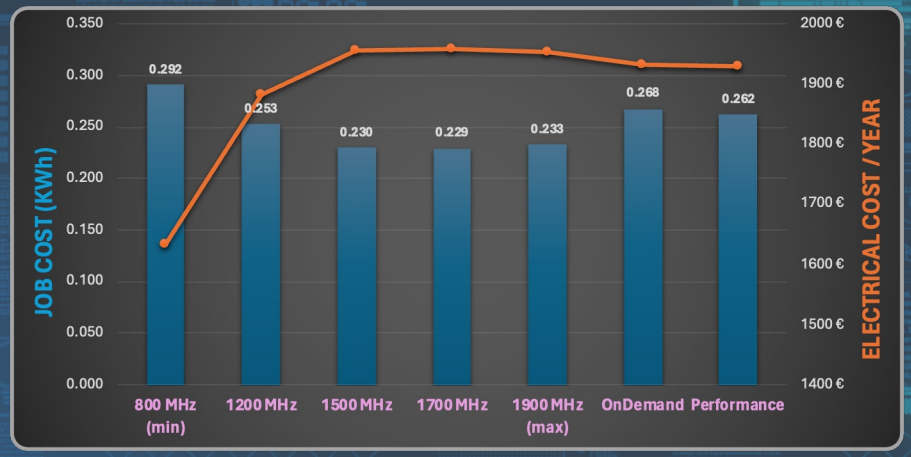




axians

NPB-CG (Conjugate Gradient)





[1] Job took too much time to complete
[2] Sweet spot:
- Lowest energy consumtion
- Highest number of jobs per day

[3] What we generally do on HPC clusters...

JOB COST (KWh)

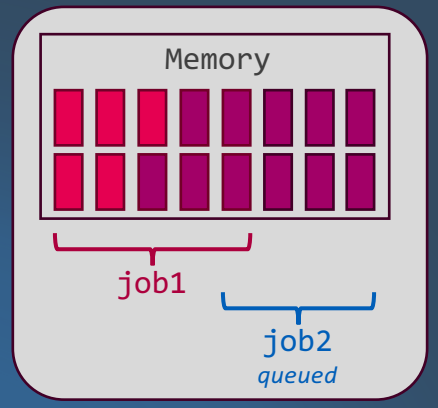
$$\frac{Energy\ (Joules)}{3.6 * 10^6}$$

axians

Resource allocation

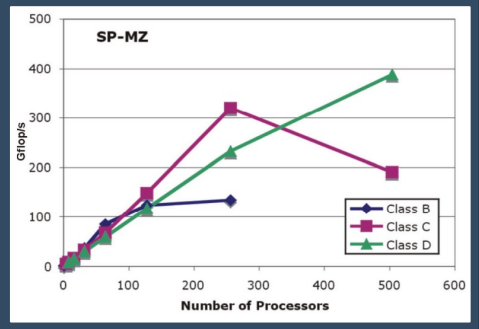
Memory request

- Try not to request more memory than necessary



Scalability

- Try not to request more cpu than necessary



Code optimization(s)

Compilation

- -O2, -O3
- -march (avx512, ...)
- [...]

Math libraries

- Blas, Lapack
- MKL
- Blis, Flame
- FFTW

CPU Binding

- Job scheduler
- MPI libraries
- OpenMP affinity
- Numactl
- [...]

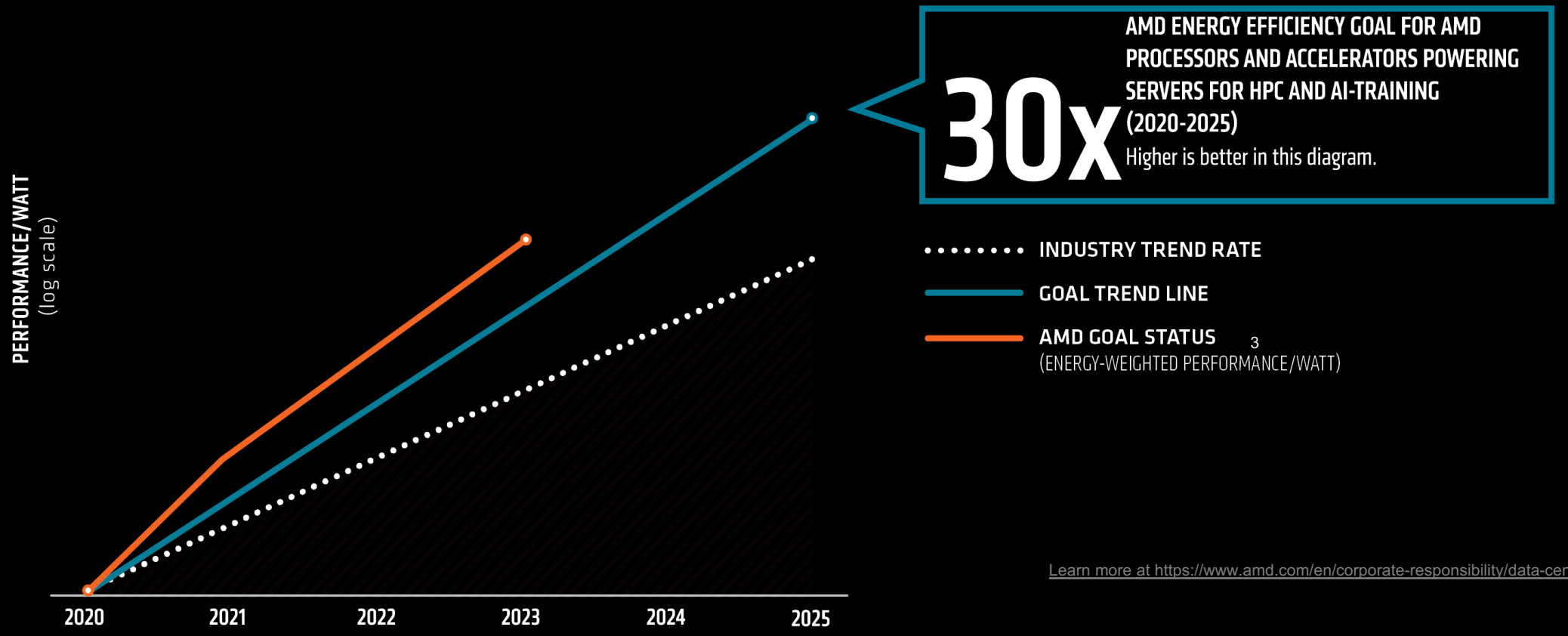
Processor Architecture

- AMD
- INTEL
- ARM

Advancing data center sustainability

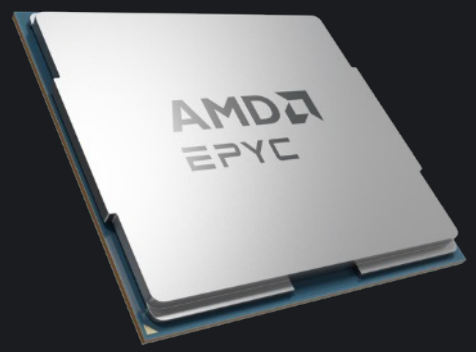
The AMD “30x25” goal is to deliver 30x more energy efficiency for our acerated compute nodes powering servers for AI-training and HPC (2020-2025).[2] The goal represents:

- 2.5x acceleration of the industry trends from 2015-2020 (measured by worldwide energy consumption for these computing segments)
- 97% reduction in energy use per computation from 2020-2025



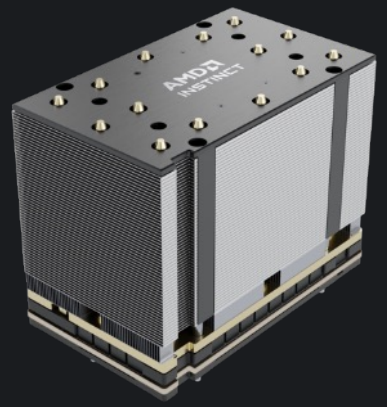
Learn more at https://www.amd.com/en/corporate-responsibility/data-center-sustainability

AMD | Data Center Solutions



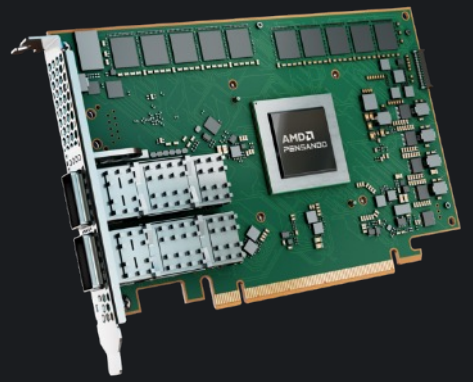
Server CPU Family

AMD EPYC



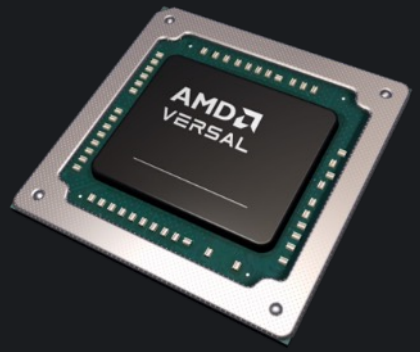
AI and HPC Accelerator

AMD INSTINCT



Networking and DPU

AMD ALVEO AMD PENSANDO



FPGA and Adaptive SoC

AMD VERSAL AMD ALVEO

AMD delivers the broadest technology portfolio to the data center

axians

Intégrateur de vos solutions IA & HPC
le service sur mesure en plus !

AMD ARROW IBM Gold Partner NVIDIA PURESTORAGE

Retrouvez-nous sur le

STAND D04