GWDG Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Energy efficiency optimization for HPC operation

Sebastian Krey, Christian Boehme, Julian Kunkel

15.02.2023

Outline

1 Infrastructure
- Data center temperature

2 HPC operation
- Monitoring power consumption
- IDLE nodes
- Powercapping

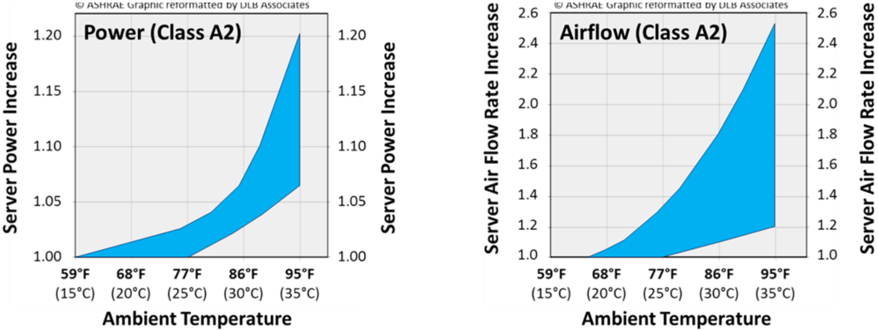
3 Benchmarks

Data center temperature

- What supply air temperatures are in use?
- What temperature limits does the equipment have?
- ASHRAE temperature limits
  - A1 up to 32°C
  - A2 up to 35°C
  - A3 up to 40°C
  - A4 up to 45°C
- ASHRAE Recommended up to 27°C
- Most servers are A2, some configurations A3
- Vendor surveys: permanent usage with 27°C no problem
- High efficiency data centers at Google 27°C, at Intel 32°C

Data center temperature

Effect of supply air temperature on airflow requirements and power consumption



- Airflow and total power increase with temperature
- Fan power increases to the cube of the fan speed (RPM)

Source: ASHRAE_TC0909_Introduction_and_Overview_09_Jan_2019.pdf

Data center temperature

Own experiments

- Air cooled compute nodes: up to 24°C flat fan curve, around 26°C high fan speeds → 24.5°C supply air
- DLC cooled compute nodes: up to 27°C flat fan curve → 27.5°C supply air
- Storage systems are more difficult, very rough fan control, reaction on cold aisle and hot aisle temperature changes → both temperatures have to be controlled.
- Storage supply air 24°C and return air not above 32°C

Monitoring power consumption

- IPMI sensors (often only PSU)
- IPMI DCMI (not on all systems available)
- Intel Node Manager (only via FreeIPMI and Intel Data Center Manager)
- Metered PDUs
- Official metering from infrastructure group (often difficult to integrate in HPC system monitoring)
- RAPL (not always interpretable results)
- Integration in Slurm to provide users Wh per job (easy integration for DCMI)

IDLE nodes

Easiest power save method is shutdown of IDLE nodes:

Emmy: 10 air cooled racks, 11 water cooled racks

- One air cooled rack of IDLE nodes: 1.9% powersave
- One water cooled rack of IDLE nodes: 7.4% powersave
- Power save potiential:

  80% system load : about 5.5%
  60% system load : about 15%

Enabling C-states C1E and lower for powered IDLE nodes has a major effect on IDLE power consumption (up to 40%).
C1E and lower effect memory and network latencies → Disable in job prolog.

Powercapping

- Emergency plans for long term power supply problems
- How much electricty is needed for an operational ready system without compute (users can login, access their data and compute could be started)
- Powersave potential of different performance/power limit methods:
  - Turbo off
  - Frequency limit
  - Custom TDP
  - Platform based power capping
- How do these different methods effect different workloads (compute, memory, communication intensive)
- TCO optimized procurements in the last years limit possible perf/watt gains

Minimal Powerconsumption

Minimal energy for an operational ready cluster system:

Can be reached wihin a few minutes in case of power supply problems.

CARO:
- Login-, admin nodes, network, BMC: 54 kW
- Storage: 10 kW
- plus cooling

Emmy: Total 120 kW
- Computeracks (edge switches, BMCs): 31 kW
- Service nodes and director switches 22 kW
- Storage 25 kW
- Cooling 42 kW

Turbo off/Frequency limit

Turbo off
- Very easy to implement
- Performance degradation workload dependent
- For Emmy power conumption reduction around 11% (8% air cooled, 14% water cooled nodes)
- Affects also the high speed interconnect

Frequency limit
- Above base clock only possible for Intel CPUs
- Fine granular adjustment (200MHz)
- Less throttling for AVX2 and AVX512 codes
- Good balancing of performance variation between nodes
- For Emmy from 1% at 3.0GHz to 11% at 2.4GHz power consumption reduction

Custom TDP

- Linear power/perf relation up to certain point, then stronger power increase
- Power capping can move performance optimized CPUs into linear range
- Custom TDP provides a few (2-3) vendor defined levels
- Custom TDP for Intel Skylake and Cascadelake CPUs only effective for Non-Turbo operation
- Affects only CPU, not mainboard, network, storage etc. → Strong limit on CPU reduces performance per watt
- For CARO (AMD Rome 7702) 165 W and 200 W result in 5.0 GFlops/W HPL
- Emmy: combination of Turbo off and cTDP reduces perf/watt by 8-15% in Gromacs

Power capping

Intel Node Manager allows global or event based platform power limiting. Similar features for AMD based systems exist probably but are unknown to me.

- Power limit affects all components of system
- Fine granular limiting possible
- Control via BIOS, FreeIPMI or Intel Data Center Manager
- Main control based on dynamic frequency limits for the CPU
- AVX2 and AVX512 clocks (lower) will also be limited
- Problem: Same power limits, different CPU clocks (production variance)
- HPC requires same performance on all nodes to minimize load imbalances
- Sometimes hickups of the controller resulting in bad performance

Benchmarks

Ongoing benchmarks with different tools from our benchmark suite

- Gromacs (highly optmized compute intensive)
- OpenFOAM (memory bandwidth intensive)
- BQCD (communication latency sensitive)

First results: Gromacs

Turbo off
- Air cooled nodes: No performance impact, 3% power save
- Water cooled nodes: 3% performance loss, 8% power save

Power cap.
- Air cooled nodes: performance/watt increase for 500W and 450W limit
- Water cooled nodes: performance/watt increase for 800W, 750W reduces perf/watt

cTDP/Turbo off Not beneficial, up to 15% perf/watt loss