Energy Use and Charging

Birds of a Feather, CUG 2023

Juan Herrera, Alan Simpson, Andy Turner, EPCC, University of Edinburgh
Martin Bernreuther, Björn Dick, HLRS
Torsten Wilde, HPE
Sridutt Bhalachandra, Norm Bourassa, NERSC, Lawrence Berkeley Laboratories
Maciej Cytowski, Cristian Di Pietrantonio, Pawsey Centre

Schedule

- 1705-1715: Introduction and context
- 1715-1745: Questions and discussion
- 1745-1750: Wrap up and sign up for working group

Principles

- Anyone can ask questions and anyone can answer
- Use polite, respectful and inclusive language
- Appreciate that different sites/systems/organisations are operating under different constraints and may have different priorities

Why are HPC centres interested in energy?

- Total Cost of Ownership of HPC centres used to be dominated by capital costs
    - …but energy costs may now make up a significant fraction
- Motivations vary between sites, and may include:
    - Reducing running costs
    - Reducing energy use, particularly at peaks of demand
    - Reducing carbon emissions
    - Improving integration between HPC centres and energy grids (power demand control)
    - Tuning of cooling infrastructure to reduce overheads
    - Educating and enabling users to be energy-aware
    - Understanding and improving instrumentation
    - Fair attribution of actual costs
    - Increasing importance as we move to Exascale

Discussion topics

1. Instrumentation and measurement of energy and attribution to users/jobs
2. User control of job energy use - e.g. processor frequency and scheduling
3. Charging/allocation that includes energy use (compute & infrastructure)
4. Raising awareness and training users to understand energy use
    a. Incentivising users to be more energy efficient
5. Impact of different HPC service operating environments

Potential additional topics

1. Energy-efficient design of future technologies
2. Software stack for monitoring and controlling energy/power

Potential working group

- Use this BoF to capture questions/problems of interest across attendees
- Report on BoF - keen for any volunteers from attendees to help write/review the report
- Working group would work to produce outputs that address (or make progress against) the questions and problems identified
    - Quarterly online meetings to discuss topics of interest and progress
    - Possible submissions to future conferences/events
- Add name and contact email address to sign up sheet at end of session to be included in the working group

EPCC Perspective

- Node level energy use data seems to be a good proxy for cabinet energy use
- Reducing the CPU/GPU frequency can lead to better energy efficiency but it depends on the software in use
  - Need to have the setting under user control so they can change if required
- Hard to motivate users to be energy efficient if there is no link between energy and charging/allocations
  - Charging should reflect the cost of the system - energy cost very dependent on location
  - What about variable energy overheads (e.g. cooling energy costs)?
- Users do not generally have a good handle on how much energy they are using - need a way to report this and place it in a context they can understand
  - e.g. How much average household energy use is their HPC system use equivalent to?

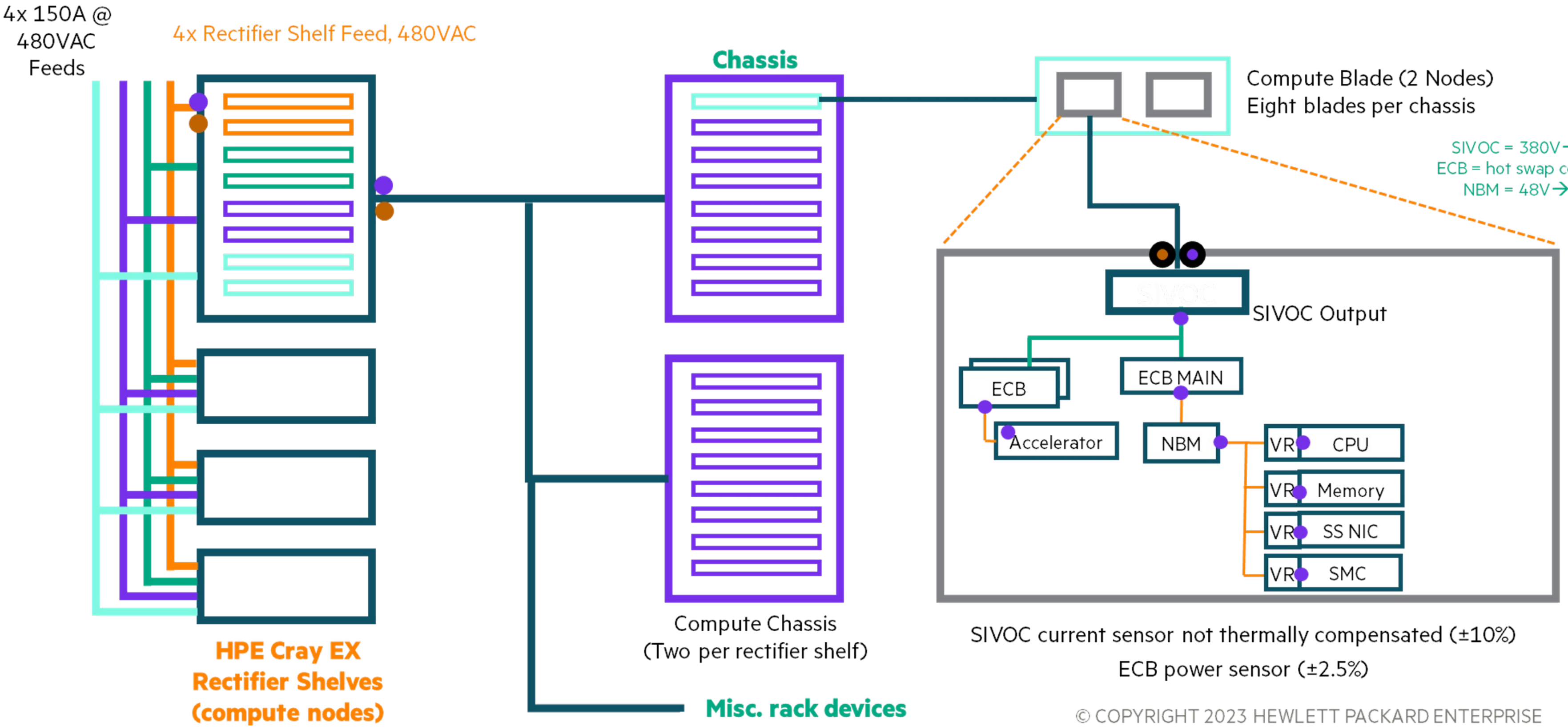
HLRS Perspective

HLRS HAWK (HPE Apollo 9000)
german Tier-1 system

- Preconditions:
  - power limits and high energy costs (energy availability variations?)
  - society/politics appreciate reduced power usage and sustainability improvements (also lower carbon footprint and get certificates to testify the progress)
- Targets:
  - measuring & accounting of power consumption per job
    - New metrics needed
    - also include shared resources or even Total Cost of Ownership?
  - Operate hardware with lowered power limits for certain workloads? (selective undervolting)
    - Understanding and adjustment of tunables like “power capping” to optimize energy usage of codes (this usually also changes the roofline resp. machine to code balance)
      ...not only for a single node but also for parallel runs
    - Offering batch queues with fixed settings for selected job classes? Dynamical steering of tunables? User involvement?
    - Optimization of user codes (also) with respect of energy efficiency
    - TCO: slower job throughput slows down amortization; Idling nodes should always run on low power
- Needs to achieve the targets?
  - modern CPUs and GPUs already offer tunables
  - Hardware and Software to collect & analyze data (ELK, TimescaleDB; cooperation with HPE)
  - Profiling tools
  - How to achieve optimal control for a dynamic tunable steering?

HPE Perspective - System Measurements

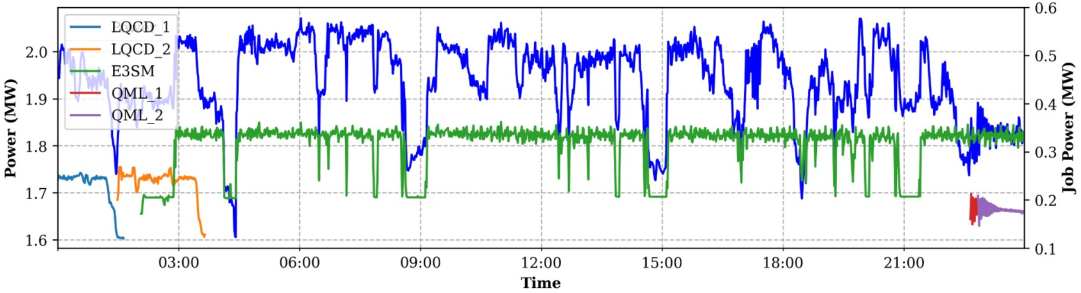
● Power
● Energy
● Calculated



SIVOC current sensor not thermally compensated (±10%)

ECB power sensor (±2.5%)

NERSC Perspective



- Job & Application Power profiling is hard!
    - Requires considerable investment in monitoring infrastructure and analytical methods development
- Quantifying facility energy costs can use PUE, iTUE or TUE metrics
    - Using a Trailing 12-month (TTM) Average is likely the fairest method
        - Simple calculation
            - Job Power (kW) * Job Time (hrs) * TTM PUE = Job Energy (kWh)
    - Using an instantaneous metric reading during Job runtime is more complicated
        - Same simple calculation, but more difficulty with time synchronization of metering data.
        - Many factors can affect fairness
            - Scheduler efficiency
            - Load balancer efficiency
            - Cooling season affects performance metric value, therefore increased energy use
- Power profiles for applications and job types, will facilitate LBNL’s participation in California AutoDR programs

Pawsey Centre Perspective

Energy-based accounting model for Heterogeneous Supercomputers

- An increasing number of supercomputing facilities are based on heterogeneous architectures with CPUs and accelerators
- Accounting models using the core hour as the base unit of measure need to be redefined to provide a charging rate for workloads running on GPU partition
- Pawsey has proposed a new accounting model that, while retaining the core hour as base unit of measure on Setonix's CPU partition, bases the GPU charging rate on relative energy consumption [1].
- It is simple, and it slightly changes the meaning of supercomputing accounting and service units:
  Researchers optimising their service units usage on Setonix also optimise the cost and energy consumption of their computational jobs, workflows and projects. Therefore, researchers are encouraged to choose the most energy efficient solution for their science.

[1] "Energy-based Accounting Model for Heterogeneous Supercomputers", Cristian Di Pietrantonio, Christopher Harris, Maciej Cytowski, CoRR, abs/2110.09987, 2021, https://arxiv.org/abs/2110.09987

Resources used	Service Units
	Setonix CPU: 128 AMD Milan cores per node GPU: 4 AMD MI250X GPUs per node
1 CPU core / hour	1
1 CPU / hour	64
1 CPU node / hour	128
1 GPU / hour	128
1 GPU node / hour	512

Useful links - software and tools

- Slurm pm_counters plugin: https://slurm.schedmd.com/cray.html
  https://github.com/SchedMD/slurm/tree/master/src/plugins/acct_gather_energy/pm_counters
- Global Extensible Open Power Manager (GEOPM): https://geopm.github.io/

Useful links - papers and reports (1)

- Energy-based Accounting Model for Heterogeneous Supercomputers, Cristian Di Pietrantonio, Christopher Harris, Maciej Cytowski, CoRR, abs/2110.09987, 2021, https://arxiv.org/abs/2110.09987
- Understanding power variation and its implications on performance optimization on the Cori supercomputer, Sridutt Bhalachandra; Brian Austin; Nicholas J. Wright, International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS), 2021 https://doi.org/10.1109/PMBS54543.2021.00011
- The imperative to reduce carbon emissions in astronomy, Stevens, A.R.H., Bellstedt, S., Elahi, P.J. et al. Nat Astron 4, 843–851 (2020). https://doi.org/10.1038/s41550-020-1169-1
- Energy Usage on ARCHER2 and the DiRAC COSMA HPC services, Alastair Basden, Andy Turner, https://doi.org/10.5281/zenodo.7128628

Useful links - papers and reports (2)

- Energy-based charging on the ARCHER2 HPC service, Alastair Basden, Andy Turner, https://doi.org/10.5281/zenodo.7702105