Energy-based Accounting Model for Heterogeneous Supercomputers

Cristian Di Pietrantonio, Christopher Harris and Maciej Cytowski

Pawsey Supercomputing Research Centre

- Australian Tier-1 Supercomputing Facility
- Headquartered in Perth, Western Australia
- Launched as Pawsey in 2014 with foundations back to 2000
- Critical support for SKA infrastructure, 800km north of Perth
- 50+ staff
- AU$70m capital refresh by Australian Government

CSIRO · Curtin University · ECU Edith Cowan University · Murdoch University · The University of Western Australia · NCRIS National Research Infrastructure for Australia, An Australian Government Initiative · Government of Western Australia

Table of contents

Figure: A quokka. Image credit: Rottnest Island Authority

1. Background
2. Motivation
3. Existing accounting models
4. Pawsey's new model based on energy efficiency
5. Comparison and discussion

Background

- Every year, a supercomputer is shared among thousands of researchers to perform very large calculations.
- Compute capacity of a supercomputer is finite and must be partitioned fairly among users.
- Need to define *what* is given to users and *how*.
  - **What**: traditionally, hourly usage of CPU cores (measured in Service Units, symbol SU).
  - **How**: periodic merit allocation process.
- Then, the cost of a job can be expressed in SU and it is equivalent to the number of CPU cores used multiplied by the elapsed time.
- A supercomputing centre also defines the total capacity of a supercomputer in terms of Service Units.

Motivation

To define a new accounting model for modern heterogeneous (CPUs + GPUs) supercomputers. Why?

- GPUs and CPUs differ in terms of power consumption, compute throughput and use cases.
- Modern supercomputers will host thousands of powerful GPUs and hundred thousands of CPU cores, making this difference even more relevant.
- Energy consumption will increase significantly, so driving users to make an efficient use of resources is critical.
- Finally, a model must take into account usage of other resources, such as RAM or fast storage.

Existing accounting models

Exclusive access node models

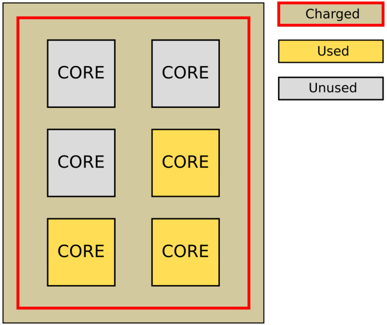


Figure: A typical accounting scheme for a node in exclusive access mode.

Taxonomy revolves around the following two main questions. (1) Can multiple jobs share the same compute node? **Exclusive access**: a job is allocated and charged for entire compute nodes. Examples are the Archer2 system or Pawsey's current system Magnus.

Existing accounting models

Shared access node models

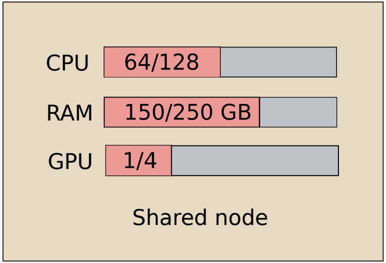


Figure: On a shared node, a job can request a *fraction* of the available resources.

Shared access: resource allocation at sub-node granularity (e.g. CPU cores). For instance, Finnish IT Centre for Science's billing units (BU) are defined for each consumable resource of shared nodes of the Puhti system.

- Each reserved core consumes 1 BU per hour,
- each GB of reserved memory consumes 0.1 BU per hour,
- each reserved GPU consumes 60 BU per hour.

Existing accounting models
CPU architectures

(2) What type of system architecture is modelled?

- **Single CPU** type: all CPU cores are charged the same.
- **Multiple CPU** types: CPU cores may be charged differently based on type.

Examples:

- (NERSC) Cori system, based on Intel Haswell (charged 140SU per node-hour) and Intel KNL (80SU per node-hour) architectures. Use of more energy efficient cores should be rewarded.
- Using peak performance ratio between CPU architectures might mislead researchers to use less energy efficient solutions. An example is introducing 2x multiplier between AVX512 and AVX2 CPU core architectures. Hardly one can achieve a 2x speedup. Moreover, energy consumption does not necessarily increase with use of longer vector instructions.

Existing accounting models

Heterogeneous systems

Heterogeneous supercomputers are systems which are composed of CPUs and accelerators.

- OLCF's Titan was the first GPU accelerated supercomputer.
  - 1 SU = 1 core per hour = 1 Streaming Multiprocessor per hour.
  - 1 hour use of 1 node is 30 SU (16 CPU cores, 14 SM).
- CSC's Puhti accounting model is an interesting heterogeneous model that provides a good introduction to ours. The weight of the GPU usage seems to be derived taking into account the thermal design power (TDP).

A model based on energy efficiency

Pawsey presents an accounting model that targets heterogeneous, shared node supercomputing systems and derives the cost of using GPUs from their energy consumption (TDP).

For any user job, the following quantities are needed:

- the number of compute nodes requested,
- the fraction of CPU cores requested on every node,
- the fraction of GPUs requested on every node,
- the fraction of memory requested on every node, and
- the time it took for the job to complete, expressed in hours.

Fractional node usage

A model based on energy efficiency

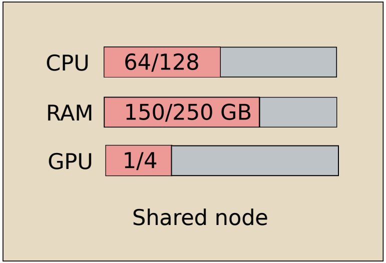


Figure: On a shared node, a job can request a *fraction* of the available resources.

What our model does is the following.

- For each allocated compute node, considers the largest fraction of allocated resource type.
- Accumulates such values.
- Multiplies the result for the elapsed time and a weight w. The interesting part is how w is defined.

CPU partition

A model based on energy efficiency

For CPU-only nodes, the weight w is the number of CPU cores available on that node.

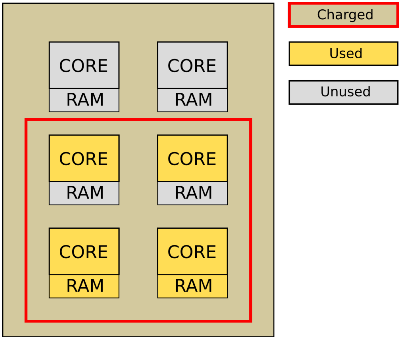


Figure: Scenario 1 ($w = 6$). The job uses more cores ($\frac{2}{3}$) than RAM ($\frac{1}{3}$), in proportion. It is charged $t \cdot 6 \cdot \frac{2}{3} = 4t$.

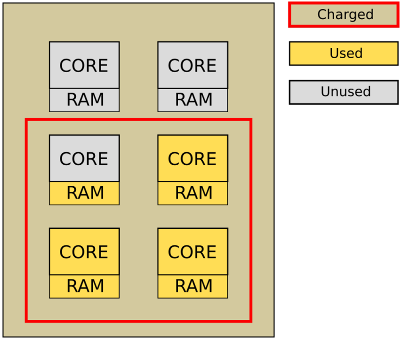


Figure: Scenario 2 ($w = 6$). The job uses more RAM ($\frac{2}{3}$) than cores ($\frac{1}{2}$), in proportion. It is still charged $t \cdot 6 \cdot \frac{2}{3} = 4t$.

GPU partition

For nodes equipped with GPUs, w is defined as the number of CPU cores scaled by the ratio between GPU and CPU TDP values.
For instance, on a node with 4 A100 GPUs (400W TDP) and dual Xeon Gold 18-core processors (150W TDP), we would have

$$
\begin{aligned}
w &= \frac{\sum_j E_g^j}{\sum_i E_c^i} C \qquad (1)\\
&= \frac{(4 \cdot 400)}{(2 \cdot 150)} 36 \qquad (2)\\
&\approx 192 \qquad (3)
\end{aligned}
$$

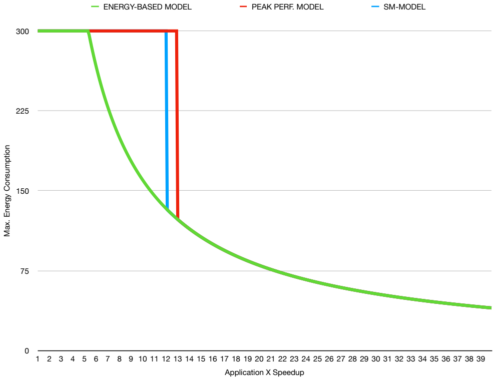
This results in a GPU node hour costing 192 SU.

Comparison of Accounting Models

Application	Perf. Ratio	SM Based	Perf. Based	Energ. Based
FUN3D	41	3.42	3.17	7.69
RTM	32	2.67	2.47	6
SPECFEM3D	105	8.75	8.11	19.69
AMBER	153	12.7	11.8	28.68
GROMACS	23	1.92	1.78	4.31
LAMMPS	59	4.92	4.56	11.06
NAMD	36	3.00	2.78	6.75
Relion	12	1.00	0.93	2.25
GTC	53	4.42	4.09	9.94
MILC	108	9.00	8.34	20.25
Chroma	99	8.25	7.64	18.56
Quantum Expresso	13	1.08	1.00	2.44
ICON	15	1.25	1.16	2.81

Table: Cost ratios between CPU and GPU charge using various accounting models for one hour benchmarks.

Comparison of Accounting Models



- Assume an application X runs in 1 hour on a CPU node with a certain cost in SU.
- For several speedup values, compute the cost in SU of running the application on a GPU node, using the 3 accounting models.
- Choose the node type with the least associated cost and compute energy consumption.
- Plot energy consumption of a job run as a function of possible speedup provided by GPUs.

Conclusion

- We reviewed the existing accounting models and found that most are not suitable for the new generation of heterogeneous supercomputers.
- In the presented model, researchers optimising their service units usage will also optimise the cost and energy consumption.
- Service units are now directly linked with power consumption, that is, the operational cost of a supercomputing centre.
- We penalise inefficient energy consumption, not inability of reaching GPU peak performance (while still being able to benefit from accelerators).