An Evaluation of Reducing Power Consumption in Taiwania 3 Supercomputer

Kuan-Chih Wang
National Center for High-performance Computing
Taiwan
2203071@narlabs.org.tw

Chin-Hung Li
National Center for High-performance Computing
Taiwan
oscarli@narlabs.org.tw

Te-Ming Chen
National Center for High-performance Computing
Taiwan
gavin@narlabs.org.tw

Steven Shiau
National Center for High-performance Computing
Taiwan
steven@narlabs.org.tw

1. INTRODUCTION

As a result of the ongoing compound global energy and recession-inflation crises, the rising electricity cost presents an unforeseen challenge for HPC system operators like NCHC.

Built in late 2020, Taiwania 3 is NCHC's current in-service HPC system consisting of 900 CPU compute nodes. The average system utilization is about 75% of the maximum capacity. However, we observe that the system utilization exhibits distinct temporal variability of both diurnal and seasonal scales. Furthermore, even when compute nodes are idle, the CPUs still operate at the all-core turbo frequency, which unnecessarily wastes energy. This finding motivates us to investigate and pursue additional opportunities in reducing energy consumption without disrupting our users.

2. METHODS

We implement the energy saving measures from two aspects:

1. Advanced BIOS Configuration Tuning

We discovered that at vendor-supplied BIOS defaults, which prioritize raw performance above all else, the CPU core C-states are disabled entirely. The principle here is to allow idle CPU cores to transition to higher C-states, resulting in better energy saving[1].

Additionally, for C0 (i.e., normal working) state, we also modified BIOS settings to prioritize efficiency (performance per watt, PPW) instead of raw performance.

Table 1. BIOS configuration tuned for energy saving and efficiency.

Socket Configuration
- Power/Performance Profile = High Performance
- Advanced Power Management Configuration
- Hardware PM State Control
- Hardware P-State = Native
- CPU C State Control
- Autonomous Core C-State = Enable
- CPU - Advanced PM Tuning
- Energy/Performance Bias
- Energy/Performance Bias = Balanced Performance

2. Enabling System Sleep

For always-on server systems, features such as system sleep may receive less or even no validations from vendors. In our case, we discovered that all compute nodes of Taiwania 3 cannot perform any kind of system sleep due to a crash in the RAID controller kernel driver. A reboot is required to recover.

After debugging and resolving the kernel driver issues, idle compute nodes are now able to be put into Suspend-to-Idle (i.e., ACPI S0) sleep state[2], and awakened on demand with the help of Slurm job scheduler[3].

In the next section, CPU core C-state residency is read by `turbostat` program[4], which uses hardware data from model specific registers (MSR) provided by CPU. Power consumption data are read from power supply sensors.

3. RESULTS

At vendor-supplied BIOS defaults, the idle power consumption is 180 W/node. The CPU core C-state residency indicates that idle CPU cores spend >99% of the time in C1 state. When in C0 state, all CPU cores are clocked at the all-core turbo frequency of 3.3 GHz.

With Method 1, a 53% reduction in idle power consumption (down to 84 W/node) can be achieved, which translates to estimated saving of 62K NT$/mon. In contrast, the CPU core C-state residency indicates that idle CPU cores now spend >99% of the time in C6 state. When in C0 state, each CPU core is underclocked or overclocked individually depending on workloads.

Finally, with Method 1+2, a 65% reduction in idle power consumption (down to 63 W/node) can be achieved, which translates to estimated saving of 75K NT$/mon.

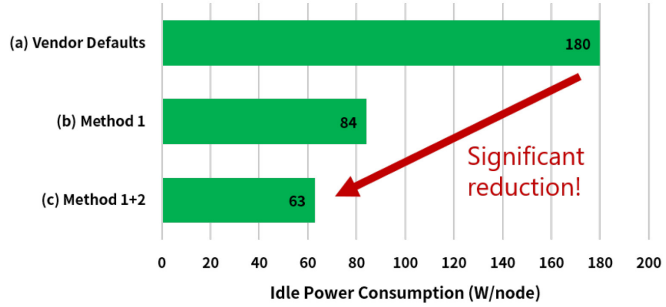

Figure 1. Idle power consumption under various conditions.

Table 2. Statistics of CPU core C-state residency during idle.

	C0 Residency	**C1 Residency**	**C6 Residency**
Vendor Defaults	0.07%	99.93%	0%
Method 1	0.08%	0.75%	99.16%

Currently, we have deployed the changes to one rack of compute nodes. We plan to continue monitoring and eventually roll out to all compute nodes.

REFERENCES

[1] Intel Corporation. *Power Management - Technology Overview*. https://builders.intel.com/docs/networkbuilders/power-management-technology-overview-technology-guide.pdf

[2] Intel Corporation. *Linux Kernel User's and Administrator's Guide - System Sleep States*. https://www.kernel.org/doc/html/v5.15/admin-guide/pm/sleep-states.html

[3] SchedMD. *Slurm Workload Manager - Slurm Power Saving Guide*. https://slurm.schedmd.com/power_save.html

[4] Len Brown. *turbostat - Report processor frequency and idle statistics*. https://manpages.debian.org/testing/linux-cpupower/turbostat.8