An Evaluation of Reducing Power Consumption in Taiwania 3 Supercomputer

Kuan-Chih Wang*[1], Chin-Hung Li*, Te-Ming Chen*, Steven Shiau*[2]

1: 2203071@narlabs.org.tw | 2: steven@narlabs.org.tw |

*: NCHC National Center for High-performance Computing

NARLabs National Applied Research Laboratories

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.
- NCHC observes distinct **temporal variability** (diurnal/seasonal scales) in the **HPC system utilization** of **Taiwania 3**. Furthermore, even when compute nodes are idle (i.e., not allocated for jobs), the **CPUs** still operate at the **all-core turbo frequency**, which unnecessarily wastes energy.
- NCHC investigates and pursues **additional opportunities** in **reducing energy consumption** without disrupting users.

Hardware Specifications of

台灣杉三號 TAIWANIA 3

- Built in late 2020
- **900 compute nodes**
- CPU
  - Dual-socket Intel Xeon Platinum 8280, **56 Cores**
- Memory
  - Samsung DDR4-2933 ECC RDIMM, **384 GB**
- Storage
  - Local: Intel DC P4610 NVMe SSD, **3.2 TB**
  - Global: IBM Spectrum Scale (GPFS), **9.4 PB**
- Network
  - Mellanox ConnectX-5 Ethernet, **25 Gb/s**
  - Mellanox ConnectX-6 InfiniBand HDR100, **100 Gb/s**

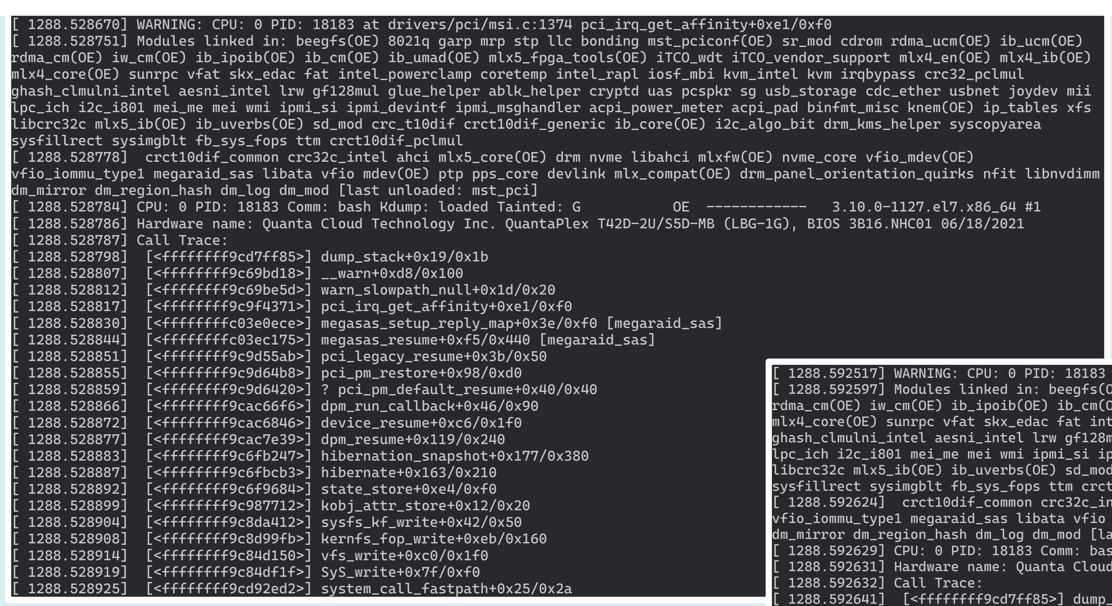
Methods

1. Advanced BIOS Configuration Tuning
   - Allow **idle CPU cores** to transition to **higher C-states**, resulting in better energy saving.
   - For **C0** (i.e., **normal working**) **state**, 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
   - Debug the RAID controller **kernel driver issues** that eventually lead to a system crash when attempting to suspend/resume. All compute nodes of Taiwania 3 are affected. This is resolved by **manually upgrading** the problematic kernel driver.
   - Idle compute nodes are put into **Suspend-to-Idle** (i.e., **ACPI S0**) sleep state, and **awakened on demand** with the help of Slurm job scheduler.
   - Stability and Usability Tests
   - Hundreds of consecutive suspend/resume cycles are performed to confirm that the problem is resolved.



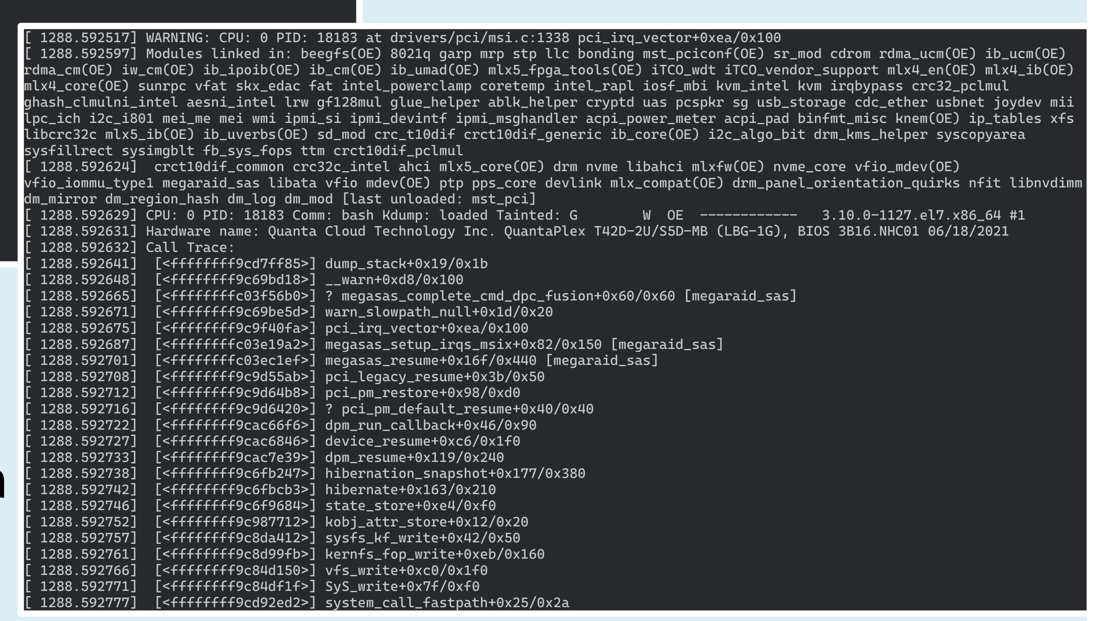


Figure 1. Sample kernel logs showing a system crash when attempting to suspend/resume on a compute node of Taiwania 3.

Results

a. 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.

b. 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** instead. When in **C0 state**, each CPU core is **underclocked** or **overclocked individually** depending on workloads.

c. 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**.

- Performance impact is around **1-2%** for LINPACK. Transition latency to/from sleep state is around **20-30 sec**.
- 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.

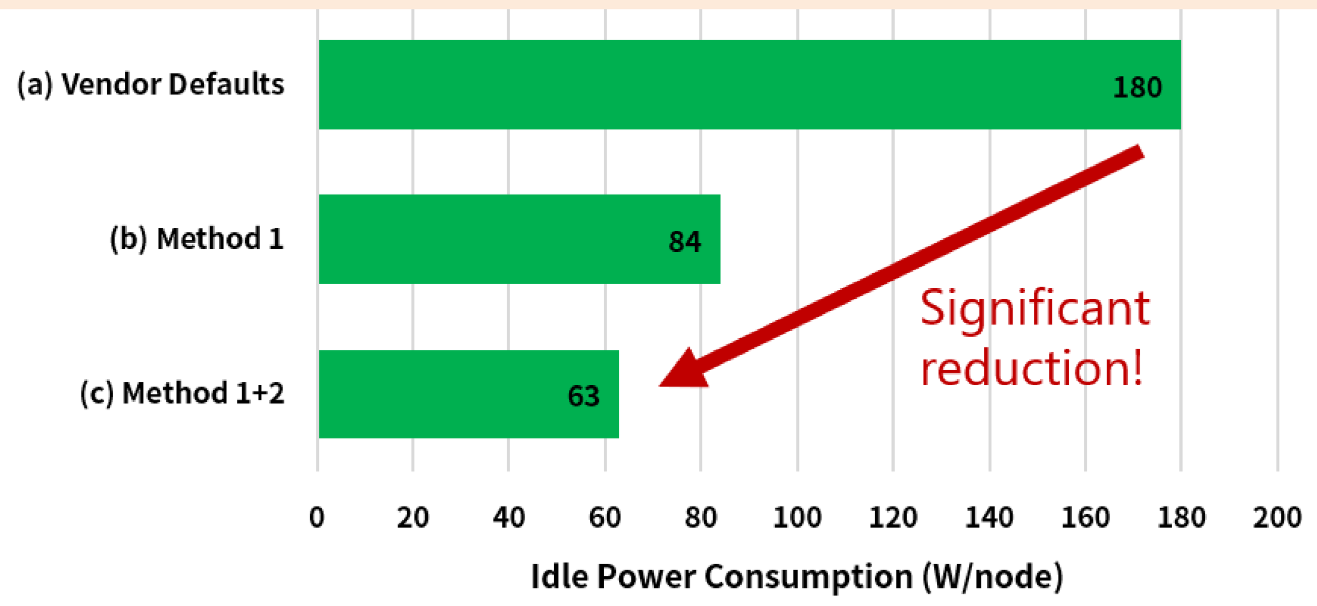


Figure 2. Idle power consumption under various conditions.

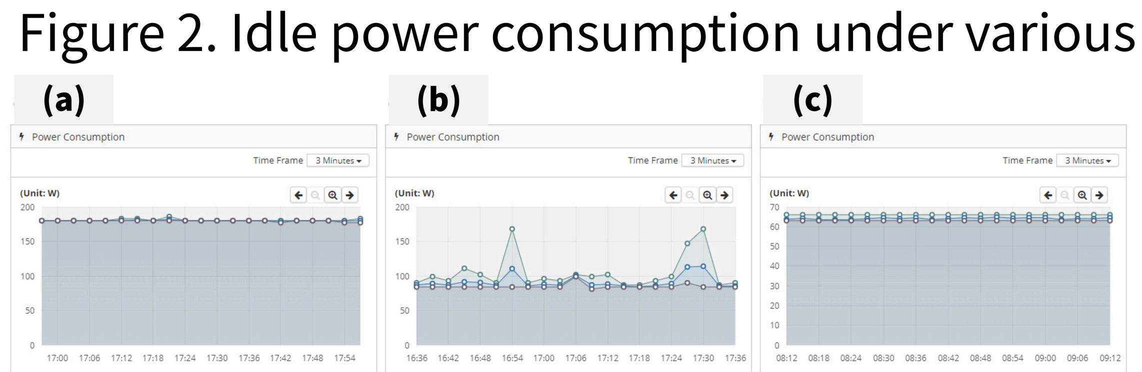


Table 2. Statistics of CPU core C-state residency during idle.

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

- 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