jh240074

Energy Efficient Operation for Supercomputer Systems

Toshihiro Hanawa (The University of Tokyo)

Abstract

1 Basic information

1.1 Collaborating JHPCN centers

- Hokkaido University
- Tohoku University
- The University of Tokyo
- Institute of Science Tokyo
- Nagoya University
- Kyoto University
- Osaka University
- Kyushu University

1.2 Theme area

- Large-scale computational science area

1.3 Project members and their roles

T. Hanawa (U Tokyo): Administration, CT, PM, PA

G. Wellein (FAU) : M, PA

S. Sumimoto (U Tokyo): CT, PM, PA

Y. Miki (U Tokyo): B, PM, PA

T. Shimokawabe, K. Yamazaki, K. Nakajima (U Tokyo): B

R. Ohara (U Tokyo): PM, PA

T. Fukaya (Hokkaido U): B

H. Takizawa (Tohoku U): PM, PA

A. Nomura (Inst. of Sci. Tokyo): CT, PM, PA

T. Endo (Inst. of Sci. Tokyo): B, PM, PA

R. Sakamoto (Inst. of Sci. Tokyo): PM, PA, AT

M. Kawai (Nagoya U ⇒ Tohoku U): LB, PM

T. Katagiri (Nagoya U): B, AT

K. Fukazawa (Kyoto U ⇒ RIHN): B, PM, PA

S. Date (Osaka U): PM, PA

S. Ohshima, T. Nanri (Kyushu U): PM, PA

J. Nonaka, S. Miura, F. Shoji, M. Terai (Riken CCS): CT, OT

S. Miwa, K. Yoshida, T. Kusaba, H. Honda (UEC): PM, PA

M. Müller (RWTH Aachen): PM, M, B

C. Plessl, Stefan Rohde (Paderborn U): EM, M, OT

H. Huber (LRZ): CT, PM, PA

(Legend: CT: Cooling Technology, PM: Power Measurement, PA: Performance Analysis, M: Power/Energy Modeling, B: Benchmark Code, AT: Autotuning, LB: Load Balancing, OT: Operation Technology)

2 Purpose and Significance of the Research

High operating costs, including increased electricity rates, are severe issues for supercomputing center operations. In addition, the carbon emission from supercomputing systems must be managed while computing demand increases. This project aims to explore optimal energy-efficient operation methods for the supercomputer system to reduce the operational cost. For that purpose, the supercomputer systems in operation are measured using benchmarks and real applications, and energy consumption and carbon footprint are modeled according to those results. The knowledge and findings regarding energy-efficient operations will be shared among international participants in this project and they can be reflected directly in the supercomputer operations at each center. Most members are involved in the operation of the supercomputing systems at various sites in Japan, including the computing resources by JHPCN requested for this project, and in Germany. The performance of the supercomputer systems will be able to be compared with that of German supercomputer centers, including the difference in climate and the performance profile of the advanced high-temperature water cooling system.

3 Significance as JHPCN Joint Research Project

This project is unique in that it is conducted as a collaboration between JHPCN members and NHR (National High-Performance Computing Alliance) members in Germany. The members from all the centers in JHPCN participate so that they can share the information and feedback findings. It is expected to contribute to the design and operation of future computing infrastructures. In addition, those systems should be analyzed using mini-apps or practical applications in interdisciplinary collaboration with computer scientists from the system aspect and computational scientists from the application aspect. JHPCN is well suited for such empirical studies.

4 Outline of Research Achievements until FY2023 (Only for continuous projects)

Not applicable

5 Details of FY2024 Research Achievements

Theme 1: Study on the usage of low-precision calculation for energy reduction

In the LLM inference, the relationship between quantization and energy consumption is under investigation. AutoGPTQ, a method for quantizing weights, performs low-bit quantization for each layer according to the importance of each layer. In a method that changes the number of quantization bits according to the importance of each layer, it is expected that the power consumption characteristics will differ for each layer. For that purpose, we prepared to measure the amount of computation, elapsed time, and power consumption for each layer for medium-scale

LLM.

Theme 2: Study on analysis corresponding among cooling, energy consumption, and performance

Monitoring GPU status (clock frequency, GPU temperature, and power usage on GPU) during performance measurements would provide helpful information. Status monitoring during performance measurements should not interfere with the primary workload but should be synchronized with the main routine. Therefore, we observe GPU status using an unused CPU core for the main computation (List. 1).Adopted interfaces for the monitoring are NVML (NVIDIA Management Library) for NVIDIA GPU, ROCm SMI (System Management Interface) library for AMD GPU, and Level-Zero Sysman APIs (Application Programming Interfaces) for Intel GPU. Although NVML and the ROCm SMI library provide a power counter, the Level-Zero Sysman API only returns a monotonic energy counter, and users must estimate the average power by taking multiple snapshots. We monitored GPU power directly and integrated it into total energy on NVIDIA and AMD GPUs. We recorded energy consumption at multiple timesteps and converted it to GPU power in each epoch on Intel GPU.

Figure 1 shows the measured performance (upper panels) and energy efficiency (lower panels) as a function of the number of N-body particles N. Figure 2 displays the time evolution of GPU status (temperature, power, and clock frequency from top to bottom) during the gravity calculation in $N = 16\,777\,216$. The GPU temperatures shown in the top panels saturate after a gradual increase in the first ~ 100 s. NVIDIA H100 and NVIDIA GH200 show saturated GPU temperatures over $\sim 10\,^{\circ}$C lower than those of AMD and Intel GPUs. It reflects the difference of form factors: SXM (and OAM: OCP (Open Compute Project) Accelerator Module) is designed to achieve higher performance using much more electricity than PCIe cards, and therefore, the cooling capability is also higher. Accordingly, NVIDIA H100 and NVIDIA GH200's TDP (700 W and 1000 W[*1]) is more than twice that of AMD MI210 and Intel PVC1100 (300 W). The average power consumption in NVIDIA H100 reaches 638 W at $N = 33\,554\,432$ for SYCL (`icpx`), which is the driving force behind the highest performance (Fig. 2e). The higher power utilization would result

Listing 1 How to monitor GPU status.

```
static bool repeat;
repeat = true;
#pragma omp parallel num_threads(2)
{
#pragma omp single
  {
#pragma omp task
    {
      while (repeat) {
        observe_GPU();
        sleep();
      }
    }
#pragma omp task
    {
      run_simulation_on_GPU();
      repeat = false;
    }
#pragma omp taskwait
  }
}
```

[*1] including CPU (NVIDIA Grace) and memory

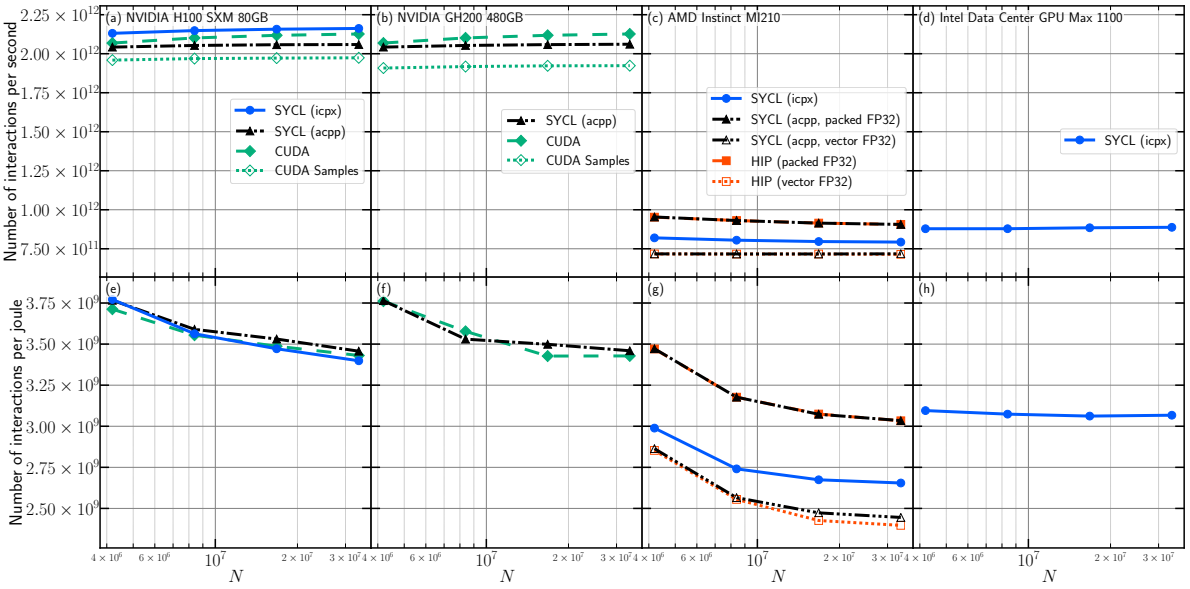

Fig. 1 Performance and energy efficiency during gravity calculation. The upper panels show the number of processed interaction pairs per second (the best performance in ten measurements) as a function of the number of N-body particles N. The lower panels exhibit the number of processed interaction pairs per joule (the best score in ten measurements) as a function of the number of N-body particles N.

from higher core usage with shorter execution time, which is also the reason for the slightly higher temperature than CUDA and `acpp` (Fig. 2a). CUDA consumes 629 W and 621 W at $N = 33\,554\,432$ on NVIDIA GH200 and NVIDIA H100, respectively. As for AMD MI210, the monitored clock-frequency profiles are characterized as three groups (Fig. 2k): (1) packed FP32 cases drop the clock frequency around 1595 MHz (1580 MHz at $N = 33\,554\,432$) with the shortest elapsed time, (2) clock frequency for SYCL (`icpx`) also drops down to 1646 MHz (1638 MHz at $N = 33\,554\,432$) and with a longer execution time, The same groups exist in the power domain (Fig. 2g), and only vector FP32 cases have a tiny gap from TDP of 300 W (290 W and 295 W for SYCL (`acpp`) and HIP). On the other hand, the temperature profiles are similar except for the total execution time (Fig. 2c). Clock frequencies in all other cases, including NVIDIA GPUs and Intel PVC1100 results, are always constant at the boost clock, implying the GPU temperature is sufficiently cooled and there is enough power supply. In fact, the average Intel PVC1100 power usage is 290 W at $N = 33\,554\,432$; therefore, there is a 10 W gap between the TDP of 300 W.

Theme 3: Study on parameterizing optimal configuration based on application performance

We are investigating the effectiveness of Dynamic Core Binding (DCB) for load balancing to reduce energy consumption. DCB works by reducing the number of bound cores

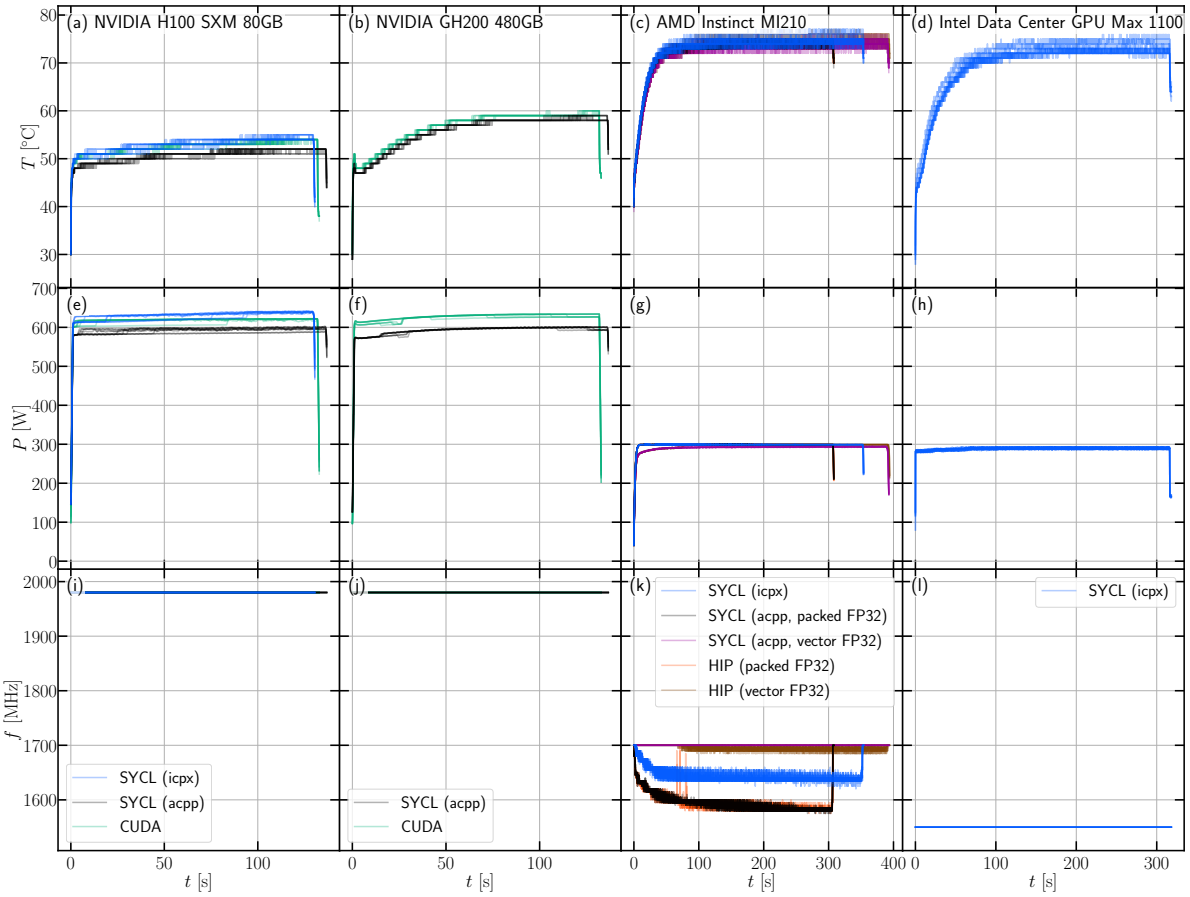

Fig. 2 GPU status during gravity calculation. The results of ten measurements in $N = 16\,777\,216$ are plotted by individual lines as a function of time ($t = 0$ means the launch of the kernel function (gravity calculation)). (a-d) GPU temperature. (e-h) Power consumption by GPU. (i-l) Clock frequency of GPU.

for processes with lighter loads, focusing on the heaviest load. This has already shown reductions in both execution time and energy usage. Moving forward, we will evaluate DCB's performance on other applications and explore auto-tuning methods to further enhance energy efficiency and computational performance.

We are studying the effectiveness of Dynamic Core Binding(DCB) for load balancing on reducing energy consumption. For reducing the energy consumption with DCB, based on the process that has the largest load, DCB reduces the number of bound cores for the other processes. In practice, execution time and energy consumption of application has been reduced. We will evaluate the effectiveness of DCB with other applications and also try to improve the energy and computational performance with the auto-tuning approach.

In A64FX supercomputers, using data from 12,289 nodes in Fugaku and 6,144 nodes in Wisteria/BDEC-01 Odyssey, we analyzed the impact of applications on power-

efficiency variation. The study found that power efficiency of compute nodes is mostly independent of the applications running, a unique feature of A64FX. Based on this, the authors propose a method to enhance computational capability by grouping and shutting down the least power-efficient nodes. This approach increased system performance by up to 13% in Fugaku and 21% in Wisteria-O under specific power constraints.

Based on this finding, a variation-aware method is proposed to optimize system performance under power constraints by ranking and grouping nodes by efficiency. Shutting down the least efficient groups increases computational capability by up to 13% in Fugaku and 21% in Wisteria-O under specific power limits.

On the other hand, the A64FX processor includes "Power Knob" capabilities that reduce power usage by adjusting specific hardware functions, such as clock frequency, the number of active floating-point units, and core states. This study investigates the correlation between these power knobs and application energy consumption using performance monitoring unit (PMU) counter values on the A64FX processor. By measuring the energy consumption of the eight microbenchmarks, we demonstrate that the optimal power knob configuration can reduce energy usage by up to 53.8%. Additionally, we collected all PMU event counter values and identified events that exhibited remarkable changes in response to power knob adjustments. Based on these observations and the characteristics of each application, we selected representative events most closely linked to application behavior. We then derived summary metrics to determine the optimal power knob settings. Using these metrics, we classified applications according to their power characteristics, demonstrating that optimal power knob configuration can be selected based on application-specific tendencies.

6 Self-review of Current Progress and Future Prospects

This project is planned to be conducted over three years. During the first year, we will mainly share and exchange information from the results of previous studies and perform basic performance measurements using many systems with different types of architectures, different cooling conditions, and different operation parameters. Through the exchange of expertise gained from operating supercomputers, German and Japanese participants will collaborate on performance analysis and modeling, including aspects like architecture and operational parameters such as cooling, starting in the second year and beyond.