jh250059

Energy Efficient Operation for Supercomputer Systems

Toshihiro Hanawa (The University of Tokyo)

Abstract

This project explores energy-efficient operation methods for supercomputer systems through Japan–Europe collaboration. In FY2025, the second year, we advanced three themes: low-precision KV-cache techniques for LLM inference (Theme 1); a quantitative analysis of cooling temperature, power, and performance on CLAIX-2023 and of performance-counter measurement overhead on NVIDIA H100/GH200 (Theme 2); and HA-LEOS, a dynamic resource-management framework combining Dynamic Core Binding and UT-Helper, evaluated on stencil codes and Lattice $\mathcal{H}$-matrix (Theme 3).

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 R-CCS): CT, OT; S. Miwa, K. Yoshida, H. Honda (UEC): PM, PA; M. Müller, C. Terboven, C. Wassermann, T. Dollenbacher (RWTH Aachen): PM, M, B; C. Plessl, S. Rohde (Paderborn U): EM, M, OT; H. Huber (LRZ): CT, PM, PA.

(Legend: CT: Cooling Technology, PM: Power Measurement, PA: Performance Analysis, EM: Power/Energy Modeling, M: Modeling, B: Benchmark Code, AT: Autotuning,

LB: Load Balancing, OT: Operation Technology)

2 Purpose and Significance of the Research

This project aims to explore optimal energy-efficient operation methods for supercomputer systems to reduce the operational cost. Energy consumption has become one of the most critical factors in modern HPC operation, and the rapid increase in the power density of recent CPUs and accelerators makes the issue even more pressing.

For the entire three-year period, we aim to (i) measure energy consumption and performance using benchmarks and real applications on a variety of supercomputer architectures and cooling configurations operated by the participating Japanese and European centers, (ii) develop power and energy models that connect application behavior, hardware configuration, and operational parameters such as cooling-water temperature, and (iii) feed back the obtained knowledge into the actual operation of each center, thereby reducing the carbon footprint of HPC services. The findings will be shared among international participants and reflected directly in the supercomputer operations at each center.

In FY2025, as the second year of the project, we focused on (1) extending the analysis of the cooling-temperature/performance/power relationship to newer hardware (Intel Sapphire Rapids and NVIDIA H100 in the RWTH Aachen CLAIX-2023 system), (2) quantifying the overheads of performance-counter measurement tools on NVIDIA GH200/H100 platforms, in order to make in-situ power-aware optimization practically feasible, (3) developing HALEOS, a dynamic resource-management framework that combines DCB and UT-Helper helper threads, and (4) initiating the investigation of low-precision arithmetic — in particular key-value cache quantization for LLM inference — as an energy-reduction lever.

The academic significance lies in the fact that, despite many isolated studies on power and cooling, very few studies have connected end-to-end measurements — application performance, in-band performance counters, hardware temperature, and facility-side cooling parameters — on production-scale supercomputers across multiple sites. The social significance is that the findings will be reflected directly in the operation of the supercomputers at the participating centers, which jointly serve a large fraction of the academic HPC users in Japan and Europe.

3 Significance as JHPCN Joint Research Project

This project is unique in that it is conducted as a collaboration between JHPCN members and members of the NHR (National High-Performance Computing Alliance) in Germany, joined by LRZ. Members from all eight JHPCN centers participate so that information and findings can be shared, and the project is expected to contribute to the design and operation of future computing infrastructures.

The framework requires three properties only available through JHPCN. First, access to a heterogeneous portfolio of large-scale systems — spanning x86, ARM Neoverse/Grace, and A64FX architectures with different cooling philosophies — is essential for generalizable energy modeling. Second, sustained collaboration among center operators, not just researchers, is required to access facility-side cooling parameters and to test operational interventions. Third, those systems can be analyzed using mini-apps and practical applications in interdisciplinary collaboration between computer scientists from the system aspect and computational scientists from the application aspect; JHPCN is well suited for such empirical studies.

In FY2025, this scheme produced concrete cross-site outcomes: the cooling-temperature analysis was carried out on the CLAIX-2023 system at RWTH Aachen; the OTF-CPT critical-path tool from RWTH-HPC was coupled with the DCB library and is now usable on the Japanese systems; a joint paper with the University of Bologna and RIKEN R-CCS on automated power-management knob configuration was accepted at CCGrid 2026; and the project's view on energy-efficient HPC operation was presented as an invited talk at the 15th European HPC Infrastructure Workshop.

4 Outline of Research Achievements until FY2024

This project is a continuous project. FY2024, the first year, established the methodological foundations on which the FY2025 work builds.

On Theme 1, the relationship between LLM-inference quantization and energy consumption was scoped. AutoGPTQ, which performs low-bit quantization for each layer according to per-layer importance, was identified as the representative target, and the measurement infrastructure for elapsed time, operation count, and per-layer power consumption on a medium-scale LLM was prepared.

On Theme 2, a multi-vendor GPU status-monitoring framework was constructed that runs on an unused CPU core in parallel with the main computation, covering NVML for NVIDIA, the ROCm SMI library for AMD, and Level-Zero Sysman APIs for Intel. Using N-body gravity calculation as the workload, performance and energy efficiency were measured on NVIDIA H100 SXM, NVIDIA GH200, AMD Instinct MI210, and Intel Data Center GPU Max 1100. Saturated GPU temperatures of NVIDIA H100/GH200 were lower than those of the AMD and Intel cards, reflecting the higher cooling capability of the SXM/OAM form factor. AMD MI210 exhibited clock-frequency throttling that grouped runs into distinct frequency bands depending on the arithmetic mode. The Japanese and German members also agreed on a common power/energy monitoring framework. Since RWTH Aachen and other German systems already use FAU's LIKWID, the project decided to develop a Grace-Hopper version of LIKWID for Miyabi as a vehicle for cross-site comparison.

On Theme 3, an initial implementation of

Dynamic Core Binding (DCB) was demonstrated to reduce both execution time and energy consumption by shrinking the core count of lighter-loaded processes; the DCB library was applied to the $\mathcal{H}$-matrix application in collaboration with the JHPCN project jh240072. In addition, a large-scale A64FX analysis was carried out on Fugaku and Wisteria/BDEC-01 Odyssey, which revealed that node-level power efficiency on A64FX is largely independent of the application running — a property unique to A64FX. On the basis of this finding, a variation-aware method that ranks nodes by efficiency and shuts down the least-efficient groups was shown to increase computational capability under specific power constraints . A complementary study of the A64FX "Power Knob" (clock frequency, number of active floating-point units, and core states), using eight microbenchmarks together with PMU counter values, demonstrated that the optimal Power-Knob configuration can reduce energy consumption when the configuration is selected based on application-specific PMU summary metrics.

Together, these FY2024 results provided the GPU status-monitoring infrastructure, the DCB implementation, and the A64FX power-modeling methodology on which the FY2025 themes build.

5 Details of FY2025 Research Achievements

5.1 Theme 1: The usage of low-precision calculation for energy reduction

In LLM inference, the relationship between numerical precision and energy consumption was investigated, with particular attention to key-value (KV) cache techniques. Building on the FY2024 layer-wise quantization work based on AutoGPTQ, we surveyed and prototyped several low-precision KV-cache representations on top of recent inference engines and quantified their effect on throughput, accuracy, and end-to-end energy. Long-context inference is increasingly memory-bound rather than compute-bound, so shrinking the KV-cache footprint translates directly into reduced DRAM/HBM traffic and therefore reduced energy per token. The investigation has clarified the trade-off space between memory pressure and arithmetic precision, and identified the regimes — batch size, context length, and model scale — in which low-precision KV-cache delivers a clear net energy reduction without an unacceptable loss of output quality. This forms the basis for the FY2026 measurements on the GH200-class systems, which combine large unified memory with high HBM bandwidth and are particularly well suited to long-context, memory-bound inference workloads.

5.2 Theme 2: Analysis corresponding among cooling, energy consumption, and performance

An effective way to reduce the operating cost of supercomputer systems is to lower the cooling cost (PUE) by raising the cooling-water temperature. However, higher operating temperatures may increase CPU/GPU leakage power and may activate DVFS to keep the system within the TDP envelope, both of which can degrade application performance and erase the cooling-side energy gain. Quantifying this trade-off on the latest hardware is therefore essential.

To reevaluate previous findings on this trade-off on newer hardware and under different cooling constraints, we reproduced the analysis on the CLAIX-2023 cluster at RWTH Aachen, the new Intel Sapphire Rapids / NVIDIA H100 SXM5 system that replaced CLAIX-2018 as the German baseline and that uses direct liquid cooling with serial flow paths — a configuration well suited to controlled side-by-side experiments.

The dual-socket Intel Sapphire Rapids nodes were measured at a low temperature between 46 °C and 49 °C and a high temperature between 50 °C and 53 °C. By leveraging the series connection within the direct-liquid-cooling loop, the same socket could be measured at the two temperature levels with a mean difference of 4 K. As in the previous study, higher CPU temperatures led to a lower clock frequency (mean: −0.32 %), which in turn lowered the measured performance (mean: −0.47 %), resulting in a higher energy consumption (mean: +0.47 %).

The same analysis was extended to the GPU partition of CLAIX-2023, whose nodes feature four NVIDIA H100 SXM5 GPUs cooled by direct liquid cooling with two pairs of GPUs in series. A low temperature between 59 °C and 67 °C and a high temperature between 67 °C and 77 °C were measured, resulting in a mean temperature spread of 10 K. The trend matched the CPU results but was more pronounced on the GPU side: higher GPU temperatures led to a lower clock frequency (mean: −1.12 %), which in turn lowered the measured performance (mean: −1.12 %), resulting in a higher energy consumption (mean: +1.27 %). The CPU and GPU results together indicate that the previously reported direction holds on Sapphire Rapids and H100, with GPUs roughly 2–3× more sensitive than CPUs to a given temperature change.

In parallel, we analyzed the runtime and power overheads of performance-counter measurement tools on NVIDIA H100 and GH200 platforms. CUPTI, Nsight Compute CLI (`ncu`), PAPI, and LIKWID were compared on the per-event and per-application level. The study revealed substantial differences in both intrusiveness and observed power-signal stability among the tools, with the underlying CUPTI Profiling API identified as a major source of measurement artifacts. Focusing on the two most practical tools, LIKWID and PAPI, we replaced their internal use of the CUPTI Profiling API with the newer CUPTI Range Profiling API. With this modification, the power-fluctuation amplitude during DGEMM runs

was reduced 35.9 % on the NVIDIA H100 GPUs of the University of Tokyo Miyabi system and Hokkaido University Grand Chariot 2 system, and the previously observed counter-value anomalies were eliminated, yielding substantially more stable measurements across the two platforms. The methodology has been accepted for publication at HIPS/IPDPS 2026 ("Mitigating Power Fluctuation in Performance Counter Measurements on NVIDIA H100 GPU"). A complementary study on A64FX, which uses the on-package performance counters to characterize application power on Fugaku-class hardware, was presented at the EESP Workshop in conjunction with ISC-HPC 2025 (Hamburg) in June 2025, reported as a poster at xSIG 2025 in August 2025, and accepted as a peer-reviewed paper in the LNCS HPC proceedings ("Analysis of Application Power Characteristics Using Performance Counters on A64FX"). Dedicated results on GH200 power-performance optimization were reported at IPSJ SIG-HPC.

Continuing the multi-vendor GPU energy-efficiency comparison initiated in FY2024, the analysis was extended to NVIDIA GH200, NVIDIA B200, and AMD MI300A using a direct-method N-body gravity benchmark in CUDA, HIP, SYCL, Kokkos, and Solomon. On B200 (Kokkos), the 1.12× speedup over GH200 closely tracked the 1.13× theoretical peak ratio, indicating high performance portability across NVIDIA generations, and B200 led GH200 in energy efficiency by 7.20 % (TSMC 4NP vs. 4N). MI300A reached 2.45×10^{12} interactions/s with HIP packed FP32 but the speedup over GH200 (1.14×) fell well short of the 1.83× theoretical peak ratio due to clock-frequency throttling caused by insufficient power delivery. Performance-portable SYCL and Kokkos matched or exceeded vendor-specific CUDA/HIP across all platforms. These results were reported at IPSJ SIG-HPC (Vol. 2026-HPC-203).

On the German side, RWTH Aachen extended its GPU monitoring infrastructure to address the limitations of the existing nvidia-smi-based sampling, which was found to lack the granularity required for accurate load and energy models. The new infrastructure leverages NVIDIA Management Library (NVML) GPU Performance Monitoring (GPM) metrics, introduced in NVML v520, which expose continuously integrated hardware counters for SM utilization, SM occupancy, and NVLink throughput, without the operational complexity of DCGM. A lightweight open-source collector (NVML-GPM-Collector) was released, and a post-processing methodology was introduced for deriving CUDA-kernel-level utilization metrics. Analysis of 9,468 production jobs of the CLAIX-2023 GPU partition showed that the SM utilization has the strongest correlation with the GPU power consumption (Pearson 0.95, MAPE 15.40 %), outperforming the instantaneously sampled nvidia-smi GPU utilization (Pearson 0.79, MAPE 32.27 %) and the SM occupancy (0.84, 26.07 %). The relationship confirms that GPU power depends more directly on the number of active SMs than on the number of active

warps. This work was published as "Leveraging NVML GPM for NVIDIA GPU Monitoring" at SCA/HPCAsia Workshops 2026.

Additionally, ITC (RWTH Aachen) presented its measurement and monitoring setup for the PUE, with particular emphasis on the proportional attribution of energy consumption and cooling load of components shared between different compute partitions. This enables the data center's location-specific carbon footprint to be attributed to individual core hours of user applications, building on the user-centric carbon footprinting methodology of Wassermann et al. (IEEE Cluster Workshops 2024).

5.3 Theme 3: Parameterizing optimal configuration based on application performance

This year we designed and evaluated HALEOS, the Hardware-Adaptive Load-Efficient Optimization System. HALEOS combines two cooperating mechanisms: Dynamic Core Binding (DCB), which dynamically adjusts the number of computational cores assigned to each process according to runtime load conditions, and UT-Helper, which utilizes the unused or spare cores generated by DCB to execute communication, I/O, and selected computations in the background. Through this cooperation, HALEOS aims to improve node-level resource utilization from both the performance and the energy-efficiency perspectives.

The core implementation of UT-Helper is based on POSIX threads, with core affinity set by DCB and with control mechanisms for synchronization, launching, and sleep management. An issue identified mid-year was that the helper-thread launch overhead was on the order of tens of milliseconds rather than the expected tens of microseconds, occurring whenever the other cores were heavily loaded with computation. The implementation was therefore revised so that helper threads are launched once during initialization and kept idle until required, eliminating the per-region launch cost.

Using stencil computations as a target workload, we compared conventional execution, MPI-based asynchronous communication and I/O, and background execution using UT-Helper. The results confirmed that MPI asynchronous mechanisms do not always sufficiently hide communication and I/O overhead, particularly for small-message communication and for I/O operations that internally require CPU processing. UT-Helper, in contrast, provided more stable overlap by executing communication and I/O on helper threads independently of the main computation, and we further confirmed that UT-Helper can be applied to selected halo computations so that surrounding processing proceeds without directly blocking the main loop.

We also evaluated HALEOS under intentionally introduced load imbalance. DCB reallocated cores according to the workload of each process, and UT-Helper used the resulting spare resources for background processing; this coordinated execution improved the overall efficiency of computation, communication, and I/O. In addition, the power-saving effect of HALEOS was confirmed: by suppressing unnecessary active-core us-

age and selectively utilizing spare resources, HALEOS improved energy efficiency while maintaining or improving application performance.

A second strand of Theme 3 was a transparent dynamic core allocation method that removes the need for application source-code modification. In the original DCB usage model, users had to call the DCB library explicitly and supply load information manually. To remove this requirement, the DCB library was integrated with RWTH Aachen's On-The-Fly Critical-Path Tool (OTF-CPT), which tracks per-thread runtime split into time spent in MPI, time spent in OpenMP, and useful computation in user code, via the standardized OMPT and PMPI tool interfaces. The combined library can be transparently injected into an unmodified user application via `LD_PRELOAD`: the per-process load is estimated as the accumulated useful computation inside OpenMP parallel regions, which is then fed to the DCB rebalancing routine. Two operation modes are supported: a user mode in which rebalancing is triggered manually via the standard `MPI_Pcontrol` interface, and a fully automatic mode in which rebalancing is triggered by intercepting blocking collective MPI calls. The system calls in the `sched.h` affinity API were identified as the most time-consuming part of the rebalancing path, and heuristics for skipping unbeneficial rebalance points are under development to reduce this overhead. Open questions for future evaluation include the retention time of historic load information as a tuning parameter for aggressive vs. conservative rebalancing strategies, and the utility of additional proxy metrics for the load balance in real-world applications.

The transparent DCB method was evaluated on Lattice $\mathcal{H}$-matrix, a real application with significant process-level load imbalance. The results confirmed that the method improves performance and reduces total energy consumption without modifying the application source code, both on a single node and in a multi-node environment, and the implementation is now being further evaluated with a set of proxy applications.

Overall, this year's Theme 3 work showed that runtime resource control, background processing, and transparent load estimation can be combined to improve both performance and energy efficiency on modern many-core CPU systems. The HALEOS effort is being aligned with the European OTF-CPT critical-path tool released by RWTH-HPC during this fiscal year, which lets us identify the regions in MPI+OpenMP applications where dynamic load balancing is most beneficial.

5.4 Cross-cutting outcomes

A joint paper with the University of Bologna and RIKEN R-CCS, "Automated Configuration of Power-Management Knobs for Optimal HPC Job Executions", was accepted at CCGrid 2026. The paper ties together the per-application performance-counter analysis from Theme 2 and the runtime resource-management framework from Theme 3.

At the operational and policy level, the project's view on energy-efficient HPC operation was presented as an invited talk at the

15th European HPC Infrastructure Workshop (“Update on Japanese Supercomputer Centers and S+D+L+I concept”) and was reflected in a domestic paper on the operation of the next-generation national HPC infrastructure (“次世代計算基盤の運用に向けて”, IPSJ SIG-HPC).

Beyond the publications listed above, project members were also involved in shaping international community activities on energy-efficient HPC. The performance-counter results from Theme 2 were first presented at the EESP Workshop (Energy Efficiency with Sustainable Performance) held in conjunction with ISC-HPC 2025 in Hamburg. At SupercomputingAsia / HPCAsia 2026 in Osaka, a project member served as a panelist of the BoF “Energy Efficiency for High Performance Computing and AI Datacenters” and as a Program Co-Chair of the DREAM Workshop on Data Reduction and Energy-Aware Data Movement.

6 Self-review of Current Progress and Future Prospects

The FY2025 plan called for (a) reproduction of cooling/temperature/power studies on the latest hardware, (b) initial implementation of a dynamic-resource framework with helper-thread offload, and (c) initial investigation of low-precision arithmetic for energy reduction. All three deliverables were achieved. (a) was carried out on the CPU and GPU partitions of CLAIX-2023, with two additional contributions: a performance-counter overhead study on H100/GH200 yielding a 35.9 % reduction of power-fluctuation amplitude, and an A64FX power-characterization study. (b) progressed from concept to HALEOS implementation and evaluation, plus an originally-unplanned transparent DCB mechanism (OMPT/PMPI) that improves performance and reduces energy on Lattice $\mathcal{H}$-matrix without source-code modification. (c) reached the survey-and-prototype stage with a clear scope for FY2026; effort was partially redirected toward the GH200/H100 overhead study where peer-reviewed venues opened up.

The proposal for the third year has been accepted. FY2026 work will move Theme 2 from descriptive analysis to a predictive energy model validated across CLAIX-2023, Miyabi/Wisteria, and LRZ; extend HALEOS into a more automatic runtime platform with UT-Helper as a runtime-observation vehicle feeding back into DCB, and broaden the transparent DCB evaluation beyond stencil and Lattice H-matrix; and measure the prototype low-precision KV-cache on GH200, tied to the Theme 2 model. The combined results will be summarized as the project’s final deliverable in early FY2027.