Asian Journal of Innovation and Policy (2024) 13.3:364-375
DOI: https://doi.org/10.7545/ajip.2024.13.3.364

Toward Strategic Demand Management of Public Supercomputing Resources: Policy Gaps and Institutional Improvements

Hyungwook Shim*, Minho Suh**

Abstract This study aims to propose strategic improvements to Korea's supercomputer demand management policy. Through a comprehensive review of the existing legal framework and the operational plans of national and specialized centers, key structural limitations were identified and categorized into three areas: demand forecasting and planning, resource construction, and ecosystem development. To address these challenges, the study proposes a set of policy measures, including the establishment of mid- to long-term supply-demand planning mechanisms, the incorporation of resource quality metrics, resource provisioning based on reserve ratios, and the formalization of supercomputing within the national science and technology classification system. These recommendations collectively provide a foundation for transitioning from a supply-centric to a demand-responsive governance model in high-performance computing policy.

Keywords Supercomputer, HPC, Resource, Demand Management, Strategy

I. Introduction

With the explosive growth of the global artificial intelligence (AI) market, the number and scale of supercomputers have become key indicators of a nation's scientific and technological competitiveness. Over the past five years, data from the Top 500 rankings -an authoritative list of the world's most powerful supercomputers- demonstrate that the total computing resources of leading countries such as the United States, China, Japan, Germany, and France have expanded more than fivefold due to escalating international competition.

Submitted, January 26, 2025; Accepted, May 8, 2025
* Senior Researcher, Division of National Supercomputer, Daejeon, Korea; shw@kisti.re.kr
** Corresponding, Senior Researcher, Division of National Supercomputer, Daejeon, Korea; mhsuh@kisti.re.kr

At the same time, growing environmental concerns, especially those related to greenhouse gas emissions, have cast critical light on the surging energy consumption and carbon dioxide output driven by the rapid expansion of supercomputing infrastructure. In response, many countries and research institutions have formed expert communities and launched targeted technology development initiatives. As a result, next-generation supercomputers are increasingly being equipped with energy-efficient components such as low-power processors and advanced cooling systems. To further encourage sustainable practices, the Green500 ranking has been introduced, evaluating supercomputers based on their environmental efficiency (Tomčala, 2021; Okazaki, 2020; Sîrbu, 2014; Savin, 2019) [1–4].

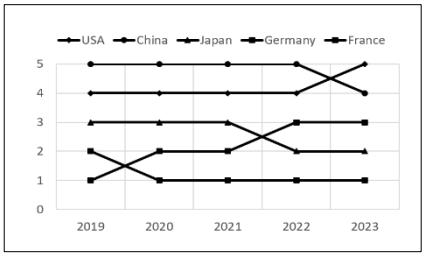

Figure 1. Top 500 (Scale)

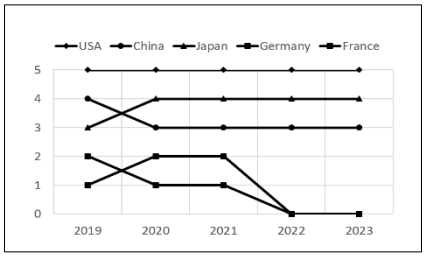

Figure 2. Top 500 (Performance)

Nevertheless, technical limitations remain. As computational performance and the number of users grow, energy consumption tends to rise proportionally. Thus, in addition to expanding physical infrastructure, it is imperative to pursue qualitative advancements in resource efficiency and management.

Recent research has explored more effective strategies for managing supercomputer resources through demand-side management (DSM). These include optimizing existing resource allocation schemes and designing new hierarchical interconnection networks to improve overall system utilization [5–6]. Shim (2023) argues that such demand management approaches can yield benefits comparable to physical expansion, including significant reductions in infrastructure costs. However, implementing effective demand management requires robust governance mechanisms that can assess real-time resource usage and plan for balanced supply, demand, and reserve capacity. Demand management also necessitates strategic planning across short-, mid-, and long-term timeframes. A unified operational roadmap that integrates periodic demand forecasting with tailored resource planning is critical.

Despite the urgency of these needs, most leading nations-including Korea-have yet to establish fully functional policies for managing supercomputing demand. In Korea, existing legal frameworks merely suggest identifying

resource demand and securing capacity in broad terms. To address this gap, institutional reform is essential-including the development of explicit supply-demand projections, long-term infrastructure planning, and the creation of predictive models to support data-driven policy formulation.

II. Literature Review

In light of the growing demand for supercomputing infrastructure, effective demand management has emerged as a critical issue. However, the academic discourse surrounding supercomputer demand policies remains limited. This chapter seeks to address this gap by reviewing relevant literature from adjacent fields such as cloud computing, electricity demand forecasting, and IT resource management. The selected studies were not chosen for their direct application to high-performance computing (HPC) policy, but rather for the conceptual and methodological insights they offer into resource planning, governance frameworks, and predictive modeling—elements that are increasingly vital in shaping supercomputing policy in both operational and institutional dimensions.

Hamzaoui et al. (2020) provided a comprehensive survey of resource utilization scheduling techniques in cloud environments, particularly emphasizing energy efficiency. Their work presents a detailed taxonomy of scheduling methods, distinguishing between reactive and proactive approaches, and addresses the complexity of managing virtualized and containerized infrastructure. Although this context differs technically from HPC, the study's emphasis on multi-level classification and adaptive scheduling directly parallels the challenges of managing diverse supercomputing workloads, where flexible and hierarchical resource coordination is essential. Similarly, Mir et al. (2020) reviewed various electricity demand forecasting techniques used in low- and middle-income countries, offering valuable insights into the temporal and structural modeling of demand. Their findings show that time series approaches dominate long- and medium-term forecasting, while AI methods are preferred for short-term accuracy. By analyzing frequently used demand determinants, such as population size, GDP, weather, and load data, the study highlights the importance of multi-variable and time-sensitive models. These insights are highly relevant for forecasting computational demand in national HPC systems, which also require the integration of macroeconomic and scientific drivers in their planning frameworks. Expanding on the use of AI, Jeyaraj et al. (2021) proposed a deep learning model for residential energy forecasting that combines a copula-based Hankel matrix transformation with a novel pooling deep neural network (PDNN). The predictive accuracy and adaptability of their approach underscore the potential of machine learning in modeling dynamic and complex demand environments, characteristics shared by supercomputing infrastructures

operating under heterogeneous workloads. Beyond the technical domain, Quichiz et al. (2017) examined IT Demand Management (ITDM) from an organizational perspective through a systematic review. They identify the critical roles of governance, strategic alignment, and leadership in ensuring the success of demand-side initiatives. Importantly, the study warns that the absence of institutional integration can render even the most technically robust policies ineffective. This insight is particularly relevant in Korea's current HPC governance, where fragmentation between central and specialized centers has impeded coherent policy implementation. Finally, Madhoo et al. (2015) presented a forward-looking experiment involving IoT-based residential energy management within the TRESCIMO initiative. By leveraging machine-to-machine (M2M) platforms and real-time data feedback, the project enabled users to understand and control their energy consumption behavior. This example demonstrates the effectiveness of user-centric, data-driven governance systems-offering a valuable model for supercomputing centers aiming to enhance operational transparency and responsiveness through real-time usage monitoring and user engagement.

Together, these studies, though drawn from distinct fields, highlight several recurring themes that are directly applicable to the management of supercomputing resources: the importance of accurate and multi-horizon demand forecasting; the need for integrated governance structures that align technical and institutional layers; the effectiveness of adaptive scheduling and resource classification; and the role of user visibility and feedback in refining demand-side strategies. The overarching implication is clear: managing supercomputing demand is not merely a technical challenge, but a multi-faceted policy problem that requires a coordinated, interdisciplinary approach.

Given the relative novelty of demand management as a policy concern in the field of supercomputing, prompted largely by the rise of AI-driven large-scale simulation and analysis, the development of institutional frameworks must precede technical implementation. These findings thus support the central thesis of this study: that the future of national supercomputing capacity lies not only in hardware expansion but in the strategic and sustainable management of demand.

III. Domestic Demand Management Policies and Systems

In Korea, the demand management of national supercomputing resources is governed primarily through mid- to long-term strategic planning frameworks and supporting legal instruments. The Ministry of Science and ICT is mandated to formulate the Master Plans for Fostering of National Supercomputing every five years. These master plans outline national strategies related to the

acquisition, distribution, and shared utilization of supercomputing resources. The most recent version, the 3rd Master Plan, categorizes its objectives into four strategic directions: innovation support by application fields, enhanced accessibility to supercomputers, technological sovereignty, and the establishment of a sustainable ecosystem connecting industry, academia, and research institutes (Table 1) (Ministry of Science and ICT, 2023)

Table 1. 3rd Master Plan (Main Direction and Promotion Strategy)

Direction	Strategy
1. Innovation supported by the application field	1.1. Advancement of Utilization Support System
	1.2. Innovative usability and creation
	1.3. Industrial utilization support
2. Strengthening access to supercomputing	2.1. Expansion of supercomputing infrastructure
	2.2. Establishment of national joint utilization service system
3. Leap forward as a technological powerhouse	3.1. Promotion of technology independence in supercomputing
	3.2. Establishment of the foundation for industrial growth
4. Establishment of Industry-University-Institute Ecosystem Foundation	4.1. Nurturing excellent talents with expertise
	4.2. Expansion of supercomputing manpower
	4.3. Formation of research base

The plan encompasses key strategies such as the establishment of specialized centers to facilitate domain-specific utilization, the development of a national joint service platform for resource sharing, the construction of auxiliary systems (pilot infrastructures) for emerging demands, and initiatives to foster public demand through the dissemination of R&D outcomes. Collectively, these measures are intended to advance resource accessibility, usability, and national technological competitiveness.

In terms of legal underpinnings, the Act on Utilization and Fostering of National Supercomputers (hereafter referred to as the Supercomputer Act) plays a central role. Article 11 of the Act mandates that the government endeavor to secure the highest level of computational resources in accordance with evolving technological demands. Furthermore, Article 12-2 provides for the active promotion of national R&D projects when necessary to ensure the timely acquisition of advanced supercomputing capabilities. These provisions position the government as a proactive agent in resource acquisition and supply-side expansion.

However, such a policy orientation has led to a supply-driven model of demand management that prioritizes the expansion of physical resources based on projected demand rather than adaptive responsiveness to dynamic usage patterns. While this approach allows for centralized, rapid response to strategic

initiatives, it poses limitations in terms of operational efficiency, especially when demand fluctuates in unforeseen ways or when user needs become more heterogeneous.

Contrasting this supply-centric approach, public institutions that operate supercomputing resources, including national and specialized centers, are increasingly incorporating demand-side considerations into their planning. Article 9 of the Supercomputer Act assigns the national center the responsibility for forecasting supercomputing demand at the national level. Accordingly, the center conducts annual quantitative assessments of potential user demand to determine appropriate system specifications in terms of performance and scale. The corresponding enforcement decree delegates similar roles to specialized centers operating across seven strategic fields, including materials and nano, weather and climate, life sciences, and autonomous systems.

Each of these specialized centers is tasked with identifying domain-specific users and designing annual operational plans based on survey-derived demand estimates. Despite these responsibilities, the legal framework currently grants only high-level mandates such as research and demand forecasting, without specifying detailed operational mechanisms. As a result, the effective implementation of demand management remains constrained by the absence of concrete institutional instruments, such as standardized forecasting procedures, clear definitions of demand scope, and actionable infrastructure planning protocols.

To bridge this gap, institutional reinforcement is required. This includes formalizing implementation tools that guide how demand is assessed, how resources are allocated, and how reserve capacity is planned for. Without such foundational mechanisms, even well-intentioned policies risk falling short of their strategic objectives, as they lack the precision and agility necessary to meet the evolving needs of the supercomputing ecosystem.

IV. Policy Improvement Plan

In Korea, the demand management of national supercomputing resources is governed primarily through mid- to long-term strategic planning frameworks and supporting

1. Overview

Supercomputing resources subject to demand management are largely public assets, managed under national frameworks. To enhance the effectiveness of demand management, improvements are needed in forecasting methods, the

setting of quantitative and qualitative usage targets, and resource planning strategies that align supply with expected utilization. In particular, demand forecasting must extend over a horizon of at least five years, consistent with the typical replacement cycle of supercomputing infrastructure. Quantitative goals should be established by considering projected user numbers alongside available resource capacity, operational efficiency, and economic feasibility.

Resource construction, in this context, involves the strategic planning of physical capacity, including new installations and the phasing out of outdated systems, based on forecasted demand. Finally, the surrounding ecosystem must be improved through refined policies on pricing, access, and service design, ensuring both the retention of existing users and the facilitation of new demand.

2. Demand Forecasting and Goal Setting

At present, the national center is legally required under Article 9 of the Supercomputing Act to conduct annual demand forecasts, the results of which are included in the operational plans submitted each January. However, current practices remain largely supplier-driven: demand is defined by fixed-scale surveys targeting expected users for specific collaborative or technical programs, such as high-precision simulations or parallelization support. This approach does not actively manage demand but rather allocates resources according to static expectations.

To address this, a mid- to long-term supply-demand plan must be institutionalized. Given that the construction of a supercomputer requires a minimum of two years, including system design, procurement, installation, and testing, and that the average operational lifespan is around five years, a seven-year planning window is essential for accurate resource provisioning. As illustrated in Figure 3, demand varies dynamically throughout the lifecycle of each generation of supercomputing systems, and identifying the directional trend, whether increasing or decreasing, is crucial for sizing future installations. If demand trends upward, capacity must be expanded; if downward, investments can be optimized accordingly.

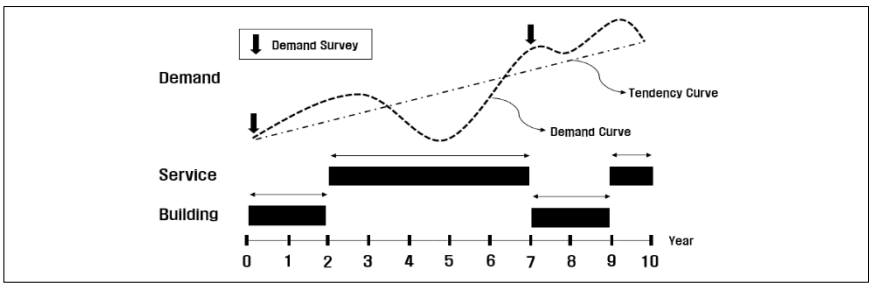

Figure 3. Full-cycle demand management

Furthermore, operational plans should be disaggregated and managed on a daily, weekly, and monthly basis. Resource usage data from 2023 indicates that the majority of jobs completed at the national center had runtimes under one hour. Without fine-grained planning, CPUs and nodes remain idle, decreasing overall efficiency. Meanwhile, the longest recorded jobs extend beyond 360 hours, underscoring the need for flexible resource allocation on a longer time horizon.

These forecasting efforts must be supported by appropriate theoretical modeling techniques, tailored to the specific time frames involved. Additionally, the concept of resource quality should be introduced into planning. In the context of modern computational science, where AI applications have become ubiquitous, this implies a shift toward GPU-centric infrastructure and services that align with performance expectations. Here, quality refers not only to the technical specifications of resources but also to the extent to which they fulfill users' actual computational requirements.

3. Resource Construction Strategy

The upcoming sixth-generation national center is expected to feature a 600-petaflop (PF) system. This new platform will pursue improvements in three areas: infrastructure expansion, service stabilization, and the application of optimized scheduling and in-house software solutions for heterogeneous, accelerator-based architectures. These developments mark a turning point toward more sophisticated HPC operations.

Nonetheless, the scale of construction must be informed by demand forecasting results—an area in need of substantial improvement. Current demand surveys are conducted only on existing users during the business planning phase, resulting in limited foresight and potential mismatch between capacity and actual utilization throughout the system’s lifecycle.

To resolve this, construction planning must incorporate a reserve ratio, the portion of total resources held in reserve for system flexibility and national-level contingencies such as emergency simulations or disaster response. Since 2022, specialized centers in designated domains have been required to share portions of their resources for public use. As summarized in Table 2, these centers maintain varying annual shared resource ratios across fields such as space, disaster, life sciences, and autonomous driving. Therefore, the national plan must account for cumulative resource availability, including shared contributions from specialized centers.

Table 2. Shared Resource Ratios of Specialized Centers (2023)

Fields	Raio(%)
Space	30
Nuclear Fusion/Accelerator	12
Disaster	25
Life/Health	31
Meteorology/Climate/Environment	10
Material/Nano	5
Autonomous-driving	14

4. Ecosystem Development and Institutional Infrastructure

An equally critical component of policy improvement lies in fostering the human and institutional infrastructure necessary to sustain advanced demand management practices. As of 2023, the National Science and Technology Standard Classification System lacks an official classification for supercomputing technologies. While related categories exist, such as "Information and Communication Convergence Platform Technology" (EE0805) and "Computer Manufacturing" (Y08), there is no classification that captures the full scope of technologies underpinning HPC, including cloud computing and data-intensive platforms.

By contrast, the European CERIF taxonomy includes distinct classifications for supercomputing, such as "Computer Science, Numerical Analysis, Systems and Control (P170)" and "Computer Technology (T120)." To advance domestic expertise, Korea must introduce an official technical classification for supercomputing and develop structured training pathways to cultivate specialists in demand forecasting, infrastructure planning, and resource optimization.

Lastly, an institutional mechanism must be established to govern the approval and oversight of mid- to long-term supercomputing supply and demand plans. At present, the Ministry of Science and ICT reviews annual operational plans for both the national and specialized centers. For specialized centers, this review includes evaluation by advisory panels based on established operational guidelines. A dedicated committee or working group should be formed to independently assess the feasibility and appropriateness of the national center's plans as well, ensuring alignment with strategic objectives and efficient use of public resources.

5. Summary Comparison Between Existing and Improved Policies

This section summarizes the key differences between the existing demand management policy and the proposed improvements suggested in this study (Table 3). The current policy framework has primarily focused on a supplier-centered approach, emphasizing resource expansion and basic quantitative demand identification. However, it lacks a systematic and responsive demand management mechanism in terms of actual operations and utilization. In contrast, the improvement strategy proposed in this study focuses on refining demand forecasting, improving resource quality, and strengthening institutional foundations.

Table 3. Comparative Summary of Existing and Proposed Supercomputer Demand Management Policies

Category	Existing Policy (Supplier-Centered)	Improved Policy (Proposed in This Study)
Demand Forecasting System	Annual quantitative survey-based	Mid-to-long-term forecasting using demand trend models
Supply-Demand Planning	Resource expansion prioritized	Balanced planning based on forecasted demand
Operational Planning	Annual-level operational planning	Daily/weekly/monthly demand-driven operational management
Resource Composition Strategy	CPU-centric, no quality consideration	Shift to GPU-centric systems, quality defined by performance fulfillment
Use of Reserve Resources	Not specified	Reserve ratio introduced for flexible and stable operations
Expertise and Institutional Basis	Lack of experts, no technical classification system	Establishment of a classification system and training of demand management specialists

This comparative analysis highlights that the proposed policy shifts from a reactive resource expansion model to a strategic and demand-responsive system. The improved policy emphasizes qualitative enhancement of resources, continuity in planning, and institutional capacity-building. Future efforts should focus on translating these proposals into legal and administrative frameworks, supported by follow-up research to facilitate practical implementation.

V. Conclusion

This study proposed a set of institutional improvements for enhancing the demand management of supercomputing resources within Korea's public sector. Recognizing the limitations of the current supply-driven policy framework, we emphasized the need for both mid- to long-term and short-term supply-demand planning mechanisms. These mechanisms should be grounded in robust forecasting methodologies and supported by the clear articulation of demand management goals. Moreover, we introduced the concept of resource quality, defined by the degree to which computational services meet user requirements, as a critical dimension in optimizing resource allocation and infrastructure performance.

In the domain of resource construction, the study highlighted the importance of determining system scale based on demand projections, while also introducing the reserve ratio as a necessary planning parameter for enhancing system resilience and flexibility. Additionally, the study stressed the urgent need to strengthen the supporting ecosystem, including the formal recognition of supercomputing within the national science and technology classification system and the cultivation of domain-specific experts through structured government initiatives.

To our knowledge, this is the first study in Korea to systematically examine demand-side policy for supercomputing at the national level. Its significance lies in offering a holistic assessment of the current policy landscape, spanning legal mandates, operational strategies of national and specialized centers, and institutional governance, and in presenting concrete, implementable pathways for improvement. Given the nascent stage of policy discourse in this area, the findings presented here are expected to serve as a foundational reference for future academic and policy-oriented work.

Nevertheless, this study is not without limitations. The domestic policy environment is still devoid of a mature conceptual framework for demand management, and numerous barriers hinder the transition from a supplier-centered expansion model to a demand-responsive governance system. These barriers include a relatively underdeveloped HPC ecosystem, a shortage of qualified experts, and a structural dependency on foreign technologies. Future research should aim to address these challenges through empirical validation of proposed models, the development of implementation roadmaps, and collaborative policy experimentation.

In moving forward, the practical realization of demand-responsive supercomputing policy will require not only technical innovation but also institutional transformation-one that enables Korea to shift from a reactive to a strategic mode of national HPC governance.

Acknowledgment

This research was supported by Korea Institute of Science and Technology Information (KISTI). (No. K25L2M1C1)

References

Al Faisal, Faiz, MM Hafizur Rahman, and Yasushi Inoguchi. "HFBN: An energy efficient high performance hierarchical interconnection network for exascale supercomputer." IEEE Access 10 (2021): 3088-3104.

Bartolini, Andrea, et al., "Unveiling Eurora—Thermal and power characterization of the most energy-efficient supercomputer in the world." 2014 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2014.

Hamzaoui, Ikhlasse, et al., "A survey on the current challenges of energy-efficient cloud resources management." SN Computer Science 1 (2020): 1-28.

Jeyaraj, Pandia Rajan, and Edward Rajan Samuel Nadar. "Computer-assisted demand-side energy management in residential smart grid employing novel pooling deep learning algorithm." International Journal of Energy Research 45.5 (2021): 7961-7973.

Madhoo, H., et al., "Future internet concepts for demand management." 2015 International Conference on the Domestic Use of Energy (DUE). IEEE, 2015.

Ministry of Science and ICT (2023), 3rd Master Plans for Fostering of National Supercomputing.

Mir, Aneeque A., et al., "A review of electricity demand forecasting in low and middle-income countries: The demand determinants and horizons." Sustainability 12.15 (2020): 5931.

Okazaki, Ryohei, et al., "Supercomputer Fugaku Cpu A64fx realizing high performance, high-density packaging, and low power consumption." Fujitsu Technical Review (2020): 2020-03.

Quichiz, Luis Palacios, and Sussy Bayona Oré. "It demand management in organizations: a review." Proceedings of the 8th International Conference on Computer Modeling and Simulation. 2017.

Savin, G. I., et al., "Joint supercomputer center of the Russian Academy of Sciences: Present and future." Lobachevskii Journal of Mathematics 40 (2019): 1853-1862.

Shim, Hyungwook, and Jaegyoon Hahm. "A study on demand management plans for National Supercomputer resources." Technology in Society 75 (2023): 102376.

Sîrbu, Alina, and Ozalp Babaoglu. "Power Consumption Modeling and Prediction in a Hybrid CPU-GPU-MIC Supercomputer (preliminary version)." arXiv preprint arXiv:1601.05961 (2016).

Tomčala, Jiří. "Supercomputer power consumption prediction using machine learning, nonlinear algorithms, and statistical methods." Journal of Physics: Conference Series. Vol. 2090. No. 1. IOP Publishing, 2021.