High Performance Computing Infrastructure Management Ecosystem Model (HPCI-MEM)

White Paper

APEC Policy Partnership on Science, Technology and Innovation

February 2025

Asia-Pacific
Economic Cooperation

Asia-Pacific
Economic Cooperation

High Performance Computing Infrastructure Management Ecosystem Model (HPCI-MEM)

White Paper

APEC Policy Partnership on Science, Technology and Innovation

February 2025

APEC Project: PPSTI 10 2022A

Produced by
Cheong Lee Sing
Independent Contractor

Project Overseer:
Dr. Manaschai Kunaseth
National Science and Technology Development Agency
Thailand

For
Asia-Pacific Economic Cooperation Secretariat
35 Heng Mui Keng Terrace
Singapore 119616
Tel: (65) 68919 600
Fax: (65) 68919 690
Email: info@apec.org
Website: www.apec.org

APEC#225-PP-03.2

Table of Contents

List of Figures

List of Table

List of Boxes

Preface

This white paper, prepared by Cheong Lee Sing, is part of the services contracted under the Asia-Pacific Economic Cooperation (APEC) Policy Partnership on Science, Technology and Innovation (PPSTI) project titled '*High Performance Computing Infrastructure Management Ecosystem Model (HPCI-MEM) for Sustainable APEC Science and Technology Development*'. This project, proposed by Thailand, is led by the National Science and Technology Development Agency (NSTDA) Supercomputer Center (ThaiSC).

The motivation for this project arises from the observation that, although many APEC economies have invested in large-scale high performance computing (HPC) infrastructures, these facilities often remain underutilized and fail to reach their full potential without a comprehensive ecosystem. Such an ecosystem extends beyond just the HPC systems and their management; it not only supports the commissioning of HPC initiatives but also drives the realization of their benefits for strategic development in Industry 4.0, digital transformation, smart cities and addressing societal challenges. Therefore, the aim of this project is to develop the HPCI-MEM, a model that illustrates the key aspects of such an ecosystem, as well as various limitations experienced by many APEC economies, that impact the utility and effectiveness of HPC facilities.

To achieve this project's aim, interviews were conducted to gather insights and experiences for the model's development. Additionally, a workshop attended by HPC experts, facility operators and senior officials was held from 3-5 April 2024, in Bangkok, Thailand. This collective intelligence, supplemented by feedback from draft reviewers and extensive desk research by the author, is distilled and presented in this white paper.

The target audience for this white paper includes:

- Novice providers and operators of government-supported HPC facilities, seeking conceptual guidance.
- Public policy officers and advisers whose portfolios directly or indirectly involve HPC-related topics, such as science and technology (S&T), higher education, research and development (R&D), innovation, Industry 4.0, digital transformation, smart cities and artificial intelligence.
- HPC users in scientific research, industry and government, utilizing HPC for modeling systems with nonlinear dynamics or chaotic behavior, many-body problems, or multi-scale phenomena, in applications such as scientific discovery, technology R&D, advanced engineering, logistical optimization and strategic decision-making.
- HPC advocates and champions interested in shaping and promoting community-driven HPC policies and initiatives.

This white paper is organized into six chapters, each tailored to specific target audience groups. Chapter 1 is relevant to all four groups. Chapters 2 and 6 are intended for novice providers and operators, public officers and advisers, as well as HPC advocates and champions. Chapter 3 focuses on novice providers and operators, Chapter 4 is aimed at public policy officers and advisers, and Chapter 5 is designed for HPC advocates and champions.

- **Chapter 1: Introduction** lays the foundation by defining HPC, its core capabilities and applications, as well as its socio-economic impact and measurements.
- **Chapter 2: Overview of HPCI-MEM** presents the model, highlighting the dynamic interplay among stakeholders and the HPC community that shapes the utility and effectiveness of HPC facilities.
- **Chapter 3: HPC Facility Setup, Management and Operation** examines the key processes involved in establishing and managing the technical infrastructure of an HPC facility. From a technical perspective, it highlights challenges, complex decision-making aspects, policies, operating procedures and the software tools needed for effective management. On the expertise side, it focuses on human resource management, capability diffusion, and education

and training. Financially, it outlines the budgetary realities, discusses long-term financing, and details various funding models and strategies.

- **Chapter 4: Public Policy for HPC** highlights the essential role of HPC in domestic strategies for Industry 4.0, digital transformation, smart city and addressing societal challenges. It also outlines key considerations for developing holistic public policy for HPC, explains the need for strategic and sustained financing for HPC facilities, and presents three case studies showcasing how Japan; Korea; and the United States have secured sustained financing for their HPC facilities.
- **Chapter 5: Community-Driven Agenda for HPC** distinguishes collective collaboration from cooperation, outlines potential community-driven actions, explores the collaboration and cooperation areas identified during the workshop, and highlights the roles and contributions of user communities within the HPC ecosystem.
- **Chapter 6: Recommendations and Conclusion** offers tailored recommendations based on the functional domains covered in the previous three chapters. It underscores the interdependence of key ecosystem elements, emphasizing that addressing just one aspect is not enough.

Additional supplementary materials exploring the concept of HPC are provided in **Appendix A: Evolution of Computing Systems** and **Appendix B: Exemplars of Supercomputers**.

***

Acknowledgement of Contributors

The development of this white paper was made possible thanks to the invaluable contributions of many individuals and groups.

Workshop Experts and Participants

- Ben Evans, National Computational Infrastructure (NCI), Australian National University (ANU), Australia
- Cristian Gomez, WisecAI, Peru
- Federico C. González Waite, Research and Innovation Center for ICT (INFOTEC), Mexico
- Fumiyoshi Shoji, Operations and Computer Technologies Division, RIKEN[1] Center for Computational Science (R-CCS), Japan
- Guido Alvarez Jefe, National Agency for Research and Development (ANID), Chile
- Jaegyoon Hahm, Center for National Supercomputing Strategy and Policy, Korea Institute of Science and Technology Information (KISTI), Korea
- Jelina Tanya Tetangco, DOST[2] Advanced Science and Technology Institute (DOST-ASTI), the Philippines
- Manaschai Kunasert, National Science and Technology Development Agency (NSTDA), Thailand
- Nirawat Thammajak, Thailand Science Research and Innovation (TSRI), Thailand
- Piyawut Srichaikul, National Electronics and Computer Technology Center (NECTEC), NSTDA, Thailand
- Rattapoom Tuchinda, NSTDA Supercomputer Center, NECTEC, Thailand
- Rifki Sadikin, National Agency for Research and Innovation (BRIN), Indonesia
- Satoshi Matsuoka, RIKEN Center for Computational Science (R-CCS), Japan
- Steven Shiau, Center for High-performance Computing (NCHC), Chinese Taipei
- Tin Wee Tan, National Supercomputing Centre (NSCC), Singapore
- Viwan Jarerattanachat, NSTDA Supercomputer Center, NECTEC, Thailand

Professors Participating as Interviewees

- Ken-ichi Nomura, USC Viterbi School of Engineering, University of Southern California (USC), the United States
- Aiichiro Nakano, USC Viterbi School of Engineering, University of Southern California (USC), the United States
- Osni Marques, Lawrence Berkeley National Laboratory, the United States
- Valentin Plugaru, LuxProvide, High Performance Computing Center, Luxembourg

Event Organizing Staff

- Apinya Kamolsook, Organization Strategy and Policy Management Division, NSTDA, Thailand
- Nucharin Ratchukool, Organization Strategy and Policy Management Division, NSTDA, Thailand
- Papawee Nupason, International Collaboration Division, NSTDA, Thailand
- Pattrawan Sripa, NSTDA Supercomputer Center, NECTEC, Thailand
- Petchpring Chinawongse, International Cooperation and Public Relations Division, NSTDA, Thailand
- Phurithat Supriyathitikul, International Collaboration Division, NSTDA, Thailand
- Siriporn Pansawat, NECTEC, NSTDA, Thailand
- Sronkanok Tangjaijit, International Collaboration Division, NSTDA, Thailand

[1] RIKEN (The Institute of Physical and Chemical Research) is the abbreviation for Rikagaku Kenkyusho

[2] DOST (Department of Science and Technology)

- Suangusa Pullket, International Cooperation and Public Relations Division, NSTDA, Thailand
- Sunutha Phromma, NECTEC, NSTDA, Thailand
- Teerapat Khiangthong, Idext MICE Company, Thailand
- Thanyaporn Hattasingh, Organization Strategy and Policy Management Division, NSTDA, Thailand
- Walailak Kongprajan, NECTEC, NSTDA, Thailand

While the workshop experts and participants, and interviewees provided valuable insights and suggestions, the final content of this report is solely the responsibility of the author. Additionally, the model presented and the recommendations expressed reflect the author's views and do not necessarily represent those of the supporting organizations or agencies.

***

Executive Summary

This white paper, prepared under the Asia-Pacific Economic Cooperation (APEC) Policy Partnership on Science, Technology and Innovation (PPSTI) project titled *'High Performance Computing Infrastructure Management Ecosystem Model (HPCI-MEM) for Sustainable APEC Science and Technology Development'*, addresses the underutilization of high-performance computing (HPC) facilities in APEC economies. The project, led by Thailand's National Science and Technology Development Agency (NSTDA) Supercomputer Center (ThaiSC), aims to develop a comprehensive ecosystem model that supports the effective management and utilization of HPC facilities.

Project Motivation and Goals

Many APEC economies have invested significantly in HPC infrastructure, yet these facilities often fall short of their potential due to a lack of a cohesive ecosystem. This white paper introduces the HPCI-MEM, a model that goes beyond the management of HPC systems themselves, encompassing a broader ecosystem involving stakeholder interactions, public policy and community-driven efforts. The model aims to enhance the strategic benefits of HPC in areas such as Industry 4.0, digital transformation, smart cities and addressing societal challenges. The insights gathered from interviews, a workshop with HPC experts and feedback from draft reviewers form the basis of this white paper.

Target Audience

The white paper is designed for:

- Novice providers and operators of government-supported HPC facilities seeking conceptual guidance.
- Public policy officers and advisors working on portfolios directly or indirectly related to HPC.
- HPC users in research, industry and government, focusing on complex computational problems and strategic decision-making.
- HPC advocates and champions interested in fostering community-driven HPC policies and initiatives.

Content Overview

The white paper is structured into six chapters, each addressing different aspects of the HPC infrastructure management ecosystem:

- **Chapter 1: Introduction** establishes a shared understanding of HPC, its definition, core capabilities, applications and socio-economic impact, setting the stage for the discussions that follow.
- **Chapter 2: Overview of HPCI-MEM** presents the ecosystem model, emphasizing the dynamic interactions between stakeholders and the HPC community that influence the effectiveness of HPC facilities.
- **Chapter 3: HPC Facility Setup, Management and Operation** provides guidance on setting up and managing HPC facilities, addressing technical challenges, decision-making processes, human resource development and financial sustainability, particularly in emerging HPC environments.
- **Chapter 4: Public Policy for HPC** underscores the strategic role of HPC in domestic initiatives like Industry 4.0 and digital transformation, and it highlights the importance of sustained investment and holistic policy frameworks. It includes case studies from Japan; Korea; and the United States to illustrate successful models of sustained HPC facilities financing.
- **Chapter 5: Community-Driven Agenda for HPC** proposes a framework for collaborative and cooperative efforts within the HPC community. It differentiates between collaboration

and cooperation, identifies potential areas of joint action, and outlines the role of user communities in advancing HPC goals.

- **Chapter 6: Recommendations and Conclusion** provides targeted recommendations for enhancing the utility and effectiveness of HPC facilities. It emphasizes the need for a balanced approach that integrates technical, policy and community aspects to maximize the impact of HPC investments.

Key Recommendations

The white paper makes several recommendations to improve HPC facility management, support strategic policy development and foster community-driven initiatives:

HPC Facility Setup, Management and Operation

1. For responsibilities involving *strategic considerations*, establish and document the decision rationale, including constraints, choices and priorities, and update them as needed.
2. Develop and maintain *policies and operating procedures* for managing and operating the HPC infrastructure, ensuring they align with strategic decisions.
3. Utilize *software tools* to effectively implement policies and operating procedures.

Public Policy for HPC

4. Raise awareness of the *critical role of HPC in domestic strategies* for Industry 4.0, digital transformation, smart cities and addressing societal challenges to draw attention to potential policy gaps.
5. Recognize the synergetic relationship between artificial intelligence *(AI), big data and HPC*, and as a result, coordinate the development of shared infrastructure to support all three.
6. Invest in skills development to cultivate a *skilled HPC workforce* and build the intellectual capital necessary for effective HPC utilization.
7. Support innovation by providing direct and indirect *financial assistance to businesses* leveraging HPC for R&D, engineering, logistical optimization and strategic decision-making to enhance their competitiveness.
8. Establish and update *norms and regulations* to ensure interoperability, privacy protection, cybersecurity and compliance with domestic security and export controls, while addressing the operational needs of HPC facilities, such as energy supply, water supply and high-speed internet connectivity.
9. Invest in *HPC facilities* and establish a sustained financing mechanism to ensure consistent support.

Community-Driven Agenda for HPC

10. Design and implement initiatives to address the needs of the HPC community, focusing on education and training, standard-setting, collaborative research, HPC infrastructure integration, knowledge exchange, sharing of research data and computational tools, and policy influence.

Conclusion

The white paper concludes that a comprehensive approach is necessary to fully leverage the potential of HPC infrastructures in APEC economies.

***

Chapter 1. Introduction

This chapter aims to establish a shared understanding of High Performance Computing (HPC) to minimize misunderstandings and ensure clear communication. The concept of HPC has evolved alongside advancements in computing, transitioning from basic calculators to early electronic computers and now to modern personal computers and supercomputers. A detailed overview is provided in **Appendix A: Evolution of Computing Systems**, with examples of notable supercomputers in **Appendix B: Exemplars of Supercomputers**.

HPC can be understood from multiple perspectives: as an activity (performing complex calculations), a process (executing specific computational applications), a domain (encompassing various areas of application) and as a field—a multidisciplinary discipline dedicated to developing and optimizing high-speed computing technologies, infrastructure and methods to address complex challenges. Accordingly, the objectives are to explore the concept of HPC, its core capabilities, its applications, as well as its socio-economic impacts and metrics.

1.1.Defining High Performance Computing (HPC)

High Performance Computing (HPC) is a dynamic concept, illustrated by the evolution of computing systems and the supercomputers described in the appendixes. The technologies and architectures underlying HPC systems have continuously advanced, with performance capabilities making significant leaps from gigaflops to teraflops, then to petaflops and now to exaflops. Moreover, the applications of HPC have expanded substantially—from initial military uses to research across diverse disciplines, evolving into multidisciplinary, mission-based problem-solving.

To encapsulate this dynamic concept, the working definition of HPC for this white paper will focus on the perspective of HPC as an activity. It is the utilization of advanced computational resources and techniques to solve large-scale, complex problems that cannot be efficiently addressed by typical desktop computers and workstations. HPC leverages cutting-edge hardware, sophisticated software and optimized network infrastructure to achieve high processing speeds and analyze vast amounts of data.

1.2.Core Capabilities and Applications Enabled by HPC

1.2.1. Core Capabilities Enabled by HPC

HPC acts as a catalyst for advancements in scientific discovery, technological research and development (R&D), advanced engineering, logistical optimization and strategic decision-making. It does so by unlocking core capabilities essential for modeling complex systems, including nonlinear dynamics or chaotic behavior, many-body interactions and multi-scale phenomena. These capabilities involve accelerating research by bridging the gaps left by traditional empirical and theoretical methods, enabling new use cases through larger-scale and multi-scale integration modeling, and mitigating extensive combinatorial challenges such as parameter sweeps, variations in initial conditions and multiple configurations of model design. Collectively, these capabilities drive the transformative impact of HPC on Industry 4.0, digital transformation, smart cities and the resolution of critical societal challenges.

Accelerating Research

Flywheel Effect. Computational science serves as a crucial bridge between empirical and theoretical science, with HPC extending this role by modeling and simulating larger, more complex systems and at a much faster pace. HPC enables the creation and refinement of computational models that uncover the hidden blueprint of real-world phenomena, enhanced by integrating insights from both empirical data and theoretical predictions. Additionally, it facilitates in-silico simulations, generating data that can be cross-validated with empirical findings and theoretical expectations.

This interconnected process links empirical observations and theoretical science in multiple ways. Empirical data informs the development of theories and guides the construction of computational models, while theoretical predictions shape data collection and model design. HPC simulations generate in-silico data that refine theories and identify areas for focused empirical research. This iterative feedback loop allows empirical findings to validate theoretical and simulated outcomes, accelerating the "re-search" process to gain deeper insights, driving innovation and discovery.

Enabling New Use Cases

More is Different. HPC is essential for enabling new use cases through larger-scale and multi-scale integration modeling. By extending the spatial-temporal range, larger-scale models can capture influences over wider areas and longer-term trends. Enhancing spatial-temporal granularity allows multi-scale models to represent multiple layers of subcomponents—components within components—and their complex interactions across various scales and at different rates.

This combined increase in range and granularity provides a deeper and more holistic understanding of complex systems, revealing insights that simpler models cannot. As a result, HPC enables more accurate and representative modeling of real-world challenges, which often exhibit emergent properties detectable only at broader scales. These properties are driven by mechanisms operating across various scales and rates, where changes in conditions at critical tipping points can lead to shifts in the system’s causal state. By capturing these intricate dynamics, HPC enhances the predictive and forecasting capabilities of computational models, offering a deeper grasp of complex phenomena.

Mitigating Combinatorial Explosion

Concurrent Processing. HPC is vital for finding optimal solutions amid extensive combinatorial possibilities, such as determining ideal parameter values for a model, identifying sets of initial conditions that define outcome boundaries and refining model design configurations. By exploring multiple possibilities concurrently, HPC drastically reduces the time required to identify the best solution within the search space.

This results in a significant time-compression advantage, enabling companies to bring products or solutions to market faster and allowing economies to secure a leadership position. This edge is particularly crucial in zero-sum scenarios and where a first-mover advantage can have profound and lasting implications.

1.2.2. HPC Applications

HPC applications are highly diverse, supporting a vast range of activities across various sectors. They can be viewed from multiple perspectives, including user domains, purposes and specific fields of application. For user domains, HPC serves scientific research organizations, industries, government bodies and specialized agencies. In terms of purpose, it drives scientific discovery, technological R&D, advanced engineering, logistical optimization and strategic decision-making. Its applications span numerous fields, from the natural sciences and engineering to social sciences and the humanities.

HPC Applications in Private Sector

With the rise of Industry 4.0 and the ongoing wave of digital transformation, HPC is no longer confined to traditional high-tech sectors. It has become a critical tool across various industries, driving innovation, enhancing operations and supporting strategic decision-making. Examples of HPC applications across different sectors include:

Resource-Based Industries

- Mining: Conducting geological modeling to locate mineral deposits and simulating mining scenarios for more effective operational planning.
- Agriculture: Modeling climate impacts on crop production, optimizing yields through simulation and applying genomic selection for crop improvement.
- Aquaculture: Simulating disease spread, modeling water quality to prevent mass fish deaths and optimizing nutritional plans to boost aquaculture productivity.

Industrial Sectors

- Manufacturing: Optimizing production processes, managing supply chains and implementing predictive maintenance to reduce downtime and increase efficiency.
- Telecommunications: Enhancing network performance, processing signals and analyzing data transmission to provide improved communication services.
- Energy: Modeling oil and gas exploration for efficient extraction, simulating renewable energy systems to estimate capacity and optimize design, and conducting nuclear simulations for safety and efficiency.

Commercial Fields

- Product Design: Running crash simulations, aerodynamic modeling and electronic design automation to accelerate product development and improve quality.
- Drug Discovery: Utilizing molecular docking, pharmacokinetic modeling and virtual screening to expedite the development of new medicines.
- Market Analysis: Supporting high-frequency trading, analyzing consumer behavior to make informed decisions and implementing fraud detection in financial markets.

Healthcare Sector

- Medical Imaging: Performing simulations for magnetic resonance imaging (MRI), computed tomography (CT) scan analysis and ultrasound modeling to enhance diagnostic capabilities.
- Personalized Medicine: Conducting genomic sequencing, patient data analysis and treatment optimization to tailor healthcare solutions to individual patients.

HPC Applications in Public Sector

With the growing adoption of digital transformation and smart city initiatives, governments are increasingly leveraging HPC to address the complexities of policy-making, administrative efficiency and public service delivery. HPC supports evidence-based policy formulation by offering advanced tools for modeling (e.g., identifying which theory aligns best with available data among conflicting hypotheses), simulations (e.g., future scenario analysis) and data analysis (e.g., risk assessment). In public administration, HPC optimizes internal processes and resource allocation, contributing to more effective governance. Additionally, it enhances the quality, accessibility and efficiency of public services. Examples of HPC applications across these areas include:

Public Policy

- Economic Policy: HPC models economic behavior, simulates market dynamics and forecasts the impact of policy changes on different economic sectors, providing insights for informed decision-making.
- Environmental Policy: It simulates climate change scenarios, evaluates the effects of proposed environmental regulations and assesses policies aimed at reducing emissions for sustainable development.
- Health Policy: HPC uses epidemiological models to predict disease spread and assess the effectiveness of intervention strategies, aiding in the creation of more robust public health policies.

Public Administration

- Urban Planning: HPC models urban development, transportation networks, land use and infrastructure projects. It simulates traffic flows, population growth and resource needs to inform planning decisions, optimize city layouts and address future challenges.
- Disaster Preparedness and Response: HPC simulates potential natural disasters like earthquakes, floods and hurricanes, allowing for the development and management of disaster response protocols, resource allocation, and risk minimization strategies.

Public Services

- Public Transportation: HPC models complex transportation systems, simulating traffic flows and predicting usage patterns to optimize routes, schedules and infrastructure, thereby improving public transit services.
- Public Safety: HPC supports crime pattern analysis, predictive policing and emergency response optimization, enhancing public safety and resource management.

1.3. Economic and Social Impacts of Leveraging HPC and Their Measurements

1.3.1. Economic and Social Impacts

Harnessing the core capabilities of HPC yields significant economic and social benefits across various sectors. In the realms of Industry 4.0, digital transformation, smart cities and addressing critical societal challenges, HPC serves as a catalyst for a wide range of positive outcomes.

Economic benefits include increased financial returns for businesses, accelerated innovation, job creation and overall higher economic productivity. By optimizing processes and enabling advanced research, HPC boosts industry competitiveness and drives economic growth.

Social benefits extend to improved job prospects for graduates, greater intellectual capacity through enhanced research and education, more informed public policy formulation, streamlined public administration, and superior public services. Collectively, these impacts contribute to a more dynamic economy and a higher quality of life for society.

1.3.2. Measuring the Impact of HPC

Quantifying the impact of HPC can be challenging, but two key metrics have been developed to gauge its effectiveness: **Return on Investment (ROI)** and **Return on Research (ROR)**.

Return on Investment (ROI)

A 2020 white paper by Hyperion Research, a spin-out of the International Data Corporation (IDC) analyst team, reported an average ROI of $44 in profits for every dollar invested in HPC, based on an analysis of over 150 use cases worldwide. While the global scope of this data may introduce some sample bias, and the ROI may vary by economy and industry, these findings demonstrate the potential for substantial returns when HPC is strategically deployed.

HPC accelerates research by bridging gaps left by traditional empirical and theoretical methods, enables new use cases through larger-scale and multi-scale integration modeling, and shortens product R&D cycles by using concurrent processing to tackle extensive combinatorial challenges, such as parameter sweeps, variations in initial conditions and different model configurations. Additionally, HPC reduces costs by replacing expensive or hazardous physical experiments with virtual simulations and improves operational efficiency through advanced analytics and process optimization.

Return on Research (ROR)

To capture the broader impact of HPC on innovation, Hyperion Research introduced the concept of *Return on Research (ROR)*. This metric highlights the complexity of measuring financial returns from research, as profitability not only depends on the success of individual projects but also on a business's ability to apply these innovations to optimize production, reduce costs, or create market-leading products.

A 2016 study by Hyperion Research identified 525 innovation outcomes attributed to HPC, ranging from "better products, cost savings, new approaches, discoveries, societal benefits, scientific breakthroughs, and support for research programs".

Beyond individual businesses, the collective impact of HPC-driven innovation boosts overall economic output. The study further demonstrated that, on average, each HPC project generated 25.6 jobs, underscoring the vital role of HPC in job creation and economic development.

***

Chapter 2. Overview of HPCI-MEM

This chapter aims to illustrate that the utility and effectiveness of an HPC facility result from the dynamic interplay among stakeholders, rather than the actions of any single entity. It draws on society's collective knowledge, synthesizing fragmented insights into the High Performance Computing Infrastructure Management Ecosystem Model (HPCI-MEM). This model adopts a holistic view of HPC infrastructure, encompassing not only the facilities themselves but also external components such as internet connectivity, utilities (electricity and water), and extramural data, computational tools and knowledge repositories. It integrates diverse perspectives on the composition of stakeholders within the HPC management ecosystem and establishes the criteria that define the utility and effectiveness of HPC facilities. The chapter further examines how the interactions among these stakeholders and the HPC community influence these criteria.

2.1. Composition of Stakeholder Groups

The HPC infrastructure, which includes HPC facilities substantially invested in by many APEC economies, is embedded within a broader management ecosystem comprising of institutions with overlapping roles, diverse functions and intricate interdependencies. These institutions can be grouped into three primary stakeholder categories:

- Policy makers and funding authorities**,** who set the strategic direction of the economy's HPC agenda and provide the necessary financial resources.
- HPC facility providers and operators, the key agents who actualize the facility's envisioned and inherent capabilities by building or procuring, managing and operating the hardware, software and utilities of high performance computational systems.
- End-users and application developers, the primary beneficiaries who leverage HPC systems for research, industrial applications, product development, public policy-making, and the enhancement of public administration and services.

In addition to these primary stakeholders, other groups play critical intermediary roles, supporting and connecting their functions. They facilitate coordination, regulatory compliance, technological advancement and knowledge sharing. Furthermore, they oversee the coordinated development of the broader HPC infrastructure:

- Regulatory and compliance bodies establish norms and regulations for HPC practices, including technical standards, data privacy, security protocols, usage policies (such as domestic security and export control compliance) and green computing policies. Their goal is to ensure interoperability, privacy protection, security and the compliant use of HPC resources. Separately, they also regulate and coordinate utilities (electricity and water) and internet infrastructure, including cross-border connectivity, to align with the HPC agenda's policies and specifically meet the unique requirements of HPC facilities and their users.
- Research and academic institutions serve as hubs for knowledge creation, dissemination and workforce development. They drive research and innovation in HPC technologies and applications while cultivating the next generation of HPC experts. Additionally, they curate and manage vast datasets, computational tools and knowledge repositories, acting as custodians of these critical resources that benefit researchers, students, developers and industry practitioners.
- Industry partners and commercial vendors act as the bridge between cutting-edge research and its practical adoption in HPC facilities. They also collaborate with government entities in building HPC facilities and serve as suppliers for the procurement of hardware, software and utilities essential to the operation of these facilities.

2.2. Definition and Criteria for Model's Design Purposes

Utility and effectiveness are interdependent. The utility of an HPC facility provides the foundational attributes necessary for its existence, while effectiveness reflects the realization of these attributes through the facility's operations and interactions with stakeholders.

2.2.1. Utility of HPC Facilities

The utility of HPC facilities refers to the intrinsic qualities and attributes that make the facility valuable and functional for its intended user community. It encompasses the facility's fundamental nature**,** existence and the different ways in which it can be accessed, secured and used efficiently. Utility is determined by how well the facility offers accessible, cost-effective, secure and high performance computational resources while also ensuring usability and scalability to accommodate varying user needs.

Criteria capturing the essence and existence of what the facility is:

- Accessibility: Defines the facility's openness and ease of entry for users, encompassing policies, network infrastructure and access mechanisms. It ensures a broad range of users—researchers, students and industry professionals—can connect to and utilize the facility.
- Security: Underpins the facility's trustworthiness by guaranteeing data protection and secure computations. A lack of security compromises the facility's essence, failing to fulfil its role as a secure resource.
- Performance: Central to the facility's existence, as it is designed specifically to perform high-speed computations. The facility's being is characterized by how effectively it delivers computational power.
- Reliability: Defines the facility's capacity to maintain consistent performance and availability. Its existence as a dependable resource hinges on its ability to operate reliably over time, serving its user base consistently.
- Energy Efficiency: Reflects the facility's operational existence within environmental and economic boundaries, ensuring long-term sustainability. It addresses the responsible use of energy resources, reinforcing the facility's ongoing utility.
- Scalability: Describes how the facility can grow and adapt to changing community needs. Scalability determines the flexibility and breadth of the facility's application, ensuring it can evolve to meet increasing or varied user demands.
- Cost-effectiveness: Represents the balance between the computational benefits provided and the costs incurred. This balance is central to the facility's value proposition, ensuring it remains a feasible resource for the community it serves.
- Technological Advancement Enablement: Reflects the facility's role in adopting the latest technologies, and supporting experimental software and hardware. Its existence as a forward-looking, evolving resource depends on its capacity to incorporate and promote new technological advancements.

Criteria for how the facility can be used by its community:

- User-friendliness: Concerns the ease with which users can interact with the facility's resources once access is granted. It involves intuitive design, clear documentation, support services and straightforward workflows, ensuring effective and smooth use of the facility.
- Extramural Integrability: Concerns the facility's capability to connect, interact and integrate with external systems, databases, computational resources and collaborative networks. This includes support for cross-institutional data transfers, integration with cloud services, compatibility with external computational tools and adherence to standards that facilitate external collaborations.

2.2.2. Effectiveness of HPC Facilities

The effectiveness of HPC facilities is the measure of how well the facility fulfils its intended functions and achieves its broader impact on users and society. It encompasses the facility's operational efficiency, maintenance practices, user satisfaction and adherence to regulations. Additionally, effectiveness includes the facility's contributions to scientific research, technological innovation, collaborative efforts and user training, highlighting its role in advancing scientific knowledge and fostering community development.

Criteria capturing how well the facility fulfils its intended function:

- Operational Efficiency: Assesses the facility's ability to optimize its resources and deliver high-quality computational results. It reflects how well the facility performs its core operational tasks, ensuring that resources are used effectively to maximize throughput and productivity.
- Maintenance Effectiveness: Measures the facility's capability to sustain smooth and uninterrupted operations over time. It ensures that the facility remains functional, reliable and up-to-date, thereby supporting its continuous use and longevity.
- User Satisfaction: Indicates the facility's success in meeting user needs and expectations. It is directly linked to how well the facility functions as a user-centered resource, ensuring that users derive maximum value from their interactions with the facility.
- Compliance: Evaluates the facility's adherence to relevant regulations and policies, ensuring legal and ethical operations. Compliance is vital to the facility's responsible functioning, reinforcing its integrity and trustworthiness in the eyes of its users and stakeholders.

Criterion capturing how well the facility fulfils its intended impact:

- Scientific and Technological Impact: Measures the facility's contributions to scientific research and technological advancements. This criterion captures the facility's broader purpose, extending its value beyond computation to drive innovation and support societal progress.

2.3. Relationship between Model's Design Purposes and Stakeholder Functions

The interplay between the utility and effectiveness of HPC facilities can be explained through various criteria, which are influenced by stakeholder functions. This interplay is illustrated in Figure 1.

1. Accessibility (Utility) ↔ User Satisfaction (Effectiveness)

Stakeholder Functions: End-users and application developers leverage the facility, while HPC facility operators manage user access.

Explanation: The facility's accessibility as a resource (utility) directly impacts user satisfaction (effectiveness). If users cannot easily access HPC resources, the facility fails to serve its purpose, regardless of its technical capabilities.

2. Security (Utility) ↔ Compliance (Effectiveness)

Stakeholder Functions: Policy makers set the direction for security policies, regulatory and compliance bodies establish security standards, and HPC facility providers and operators implement security measures.

Explanation: The facility's inherent security (utility) is crucial for maintaining compliance (effectiveness) with legal and regulatory standards. A secure facility not only safeguards data integrity but also ensures confidentiality, access control and protection against cyber threats, meeting regulatory requirements and building trust among users and stakeholders.

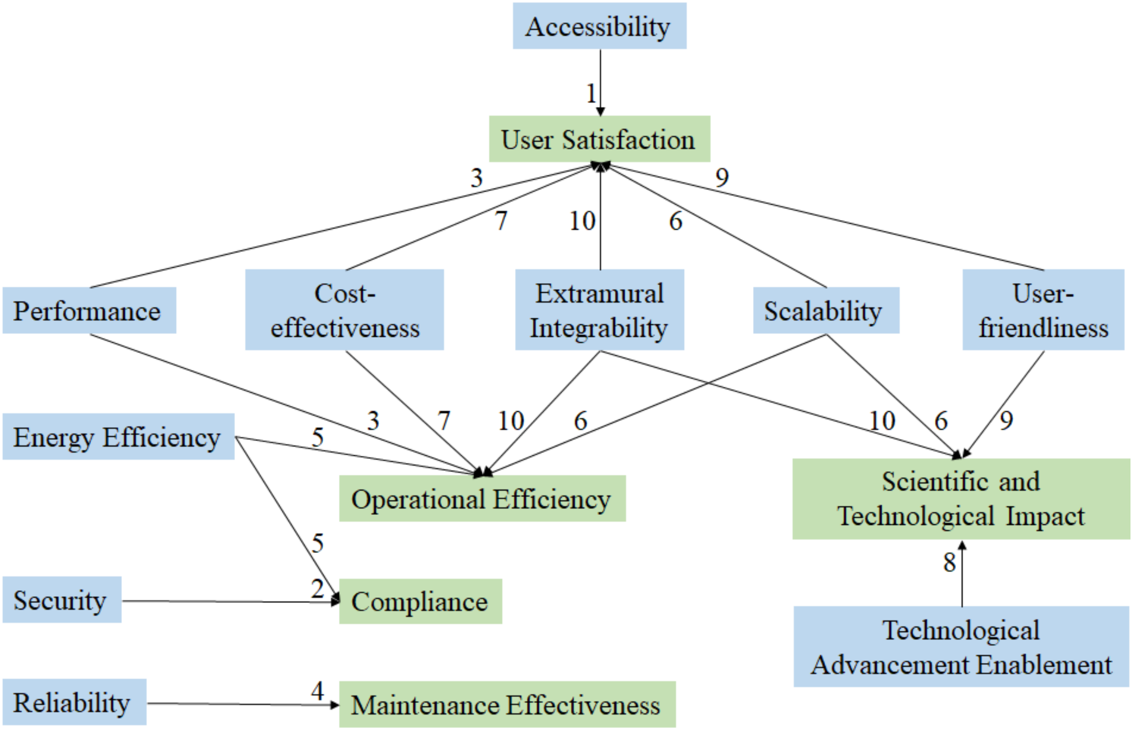

Figure 1: Illustration of Interplay between the Utility and Effectiveness of HPC Facilities. (The criteria for utility are shaded in blue, while those for effectiveness is shaded in green. The numbers correspond to the detailed exploration of these relationships.)

3. Performance (Utility) ↔ Operational Efficiency (Effectiveness) and User Satisfaction (Effectiveness)

Stakeholder Functions: HPC facility providers build or procure high performance systems, operators continually tune and optimize system performance, and funding authorities allocate capital for the initial setup, operational expenses and staffing.

Explanation: The facility's performance (utility) directly influences its operational efficiency (effectiveness) and, in turn, user satisfaction (effectiveness). High performance capabilities enable the rapid and accurate processing of computational tasks, which, when coupled with effective resource management, maximizes system utilization and throughput. This operational efficiency results in faster job completion times, enhancing the user experience.

HPC facility providers set the foundation for this performance through system design and quality. Operators then actualize and maintain peak performance by optimizing how tasks are scheduled, balanced and executed, a process that requires skilled personnel and tacit knowledge. Funding authorities support this by providing the financial resources necessary not just for hardware and upgrades but also for the staff required to ensure the system runs efficiently. Without proper funding and expertise in performance optimization, the facility risks operational inefficiencies, leading to longer job queues and decreased user satisfaction.

4. Reliability (Utility) ↔ Maintenance Effectiveness (Effectiveness)

Stakeholder Functions: HPC facility operators manage ongoing maintenance, industry partners and commercial vendors supply reliable components, and funding authorities provide capital for the initial setup, cover maintenance costs for the initial years and fund operational expenses for maintenance beyond that period.

Explanation: The facility's reliability (utility) is intrinsically linked to effective maintenance practices (effectiveness). Reliability is a defining attribute because it is a core characteristic that the facility must possess to meet user expectations for consistent performance, availability and stability.

It is also an outcome since sustained reliability depends on how well the system is managed and maintained over time. Effective maintenance practices, such as regular upgrades and repairs, directly influence the facility’s ongoing reliability. Without proper maintenance, even a system designed for reliability will experience declines in performance, increased downtimes and potential failures, ultimately compromising its utility.

5. Energy Efficiency (Utility) ↔ Operational Efficiency (Effectiveness) and Compliance (Effectiveness)

Stakeholder Functions: Policy makers set the direction for sustainable energy policies, regulatory and compliance bodies establish green computing regulations, HPC facility providers design the facility for energy efficiency, and operators actively manage energy consumption.

Explanation: The facility's energy efficiency (utility) is crucial for achieving both operational efficiency (effectiveness) and compliance (effectiveness). High energy consumption and excessive heat generation can limit the system's ability to operate at peak performance, leading to increased costs and potential regulatory issues.

Efficient HPC systems dynamically adjust power usage based on workload demands, providing energy only as needed instead of consuming maximum power regardless of the load. This dynamic energy provision minimizes waste, reduces cooling requirements and lowers operational costs. Additionally, by implementing energy-efficient designs and actively managing energy consumption, HPC facility providers and operators not only optimize system performance and enhance computational throughput but also adhere to green computing regulations and sustainable energy policies.

6. Scalability (Utility) ↔ Operational Efficiency (Effectiveness), User Satisfaction (Effectiveness), and Scientific and Technological Impact (Effectiveness)

Stakeholder Functions:

- Funding authorities influence scalability by allocating financial resources for infrastructure expansion, upgrades and maintenance. They provide funding for the procurement of additional hardware, software licenses and support services to enhance the facility's capacity in response to growing demand.
- HPC facility providers design and build the facility with scalability in mind, ensuring that the architecture allows for future expansions, such as adding more compute nodes, storage, or integrating advanced technologies.
- Operators manage and implement the scaling process, including hardware upgrades, software updates and optimization of existing resources to accommodate increasing workloads. They also develop policies for dynamic resource allocation, allowing the system to handle surges in demand efficiently.
- End-users and application developers drive the need for scalability through their growing computational demands. Their feedback and usage patterns guide the facility's development to ensure it meets the changing requirements of research and industrial applications.

Explanation: Scalability (utility) is an intrinsic characteristic of the facility that directly shapes its operational efficiency, user satisfaction, and scientific and technological impact. It reflects the facility's capacity to grow and adapt to changing computational demands, ensuring it remains a versatile and evolving resource.

Scalability (Utility) ↔ Operational Efficiency (Effectiveness): A scalable facility optimizes resource allocation by dynamically adjusting to varying workloads. This adaptability minimizes inefficiencies, ensuring smooth operation regardless of demand and embodies the facility's inherent ability to handle fluctuating computational requirements.

Scalability (Utility) ↔ User Satisfaction (Effectiveness): The facility’s ability to scale directly impacts user experience by offering shorter queue times, reliable access and the flexibility to run increasingly complex tasks. This responsiveness to evolving user needs fosters a favorable perception and continuous engagement.

Scalability (Utility) ↔ Scientific and Technological Impact (Effectiveness): Scalability extends the facility's capability to support advanced research, accommodating extensive datasets and complex models. This capacity to handle a broad range of scientific inquiries amplifies its role in driving innovation and exploration in new scientific frontiers.

7. Cost-Effectiveness (Utility) ↔ Operational Efficiency (Effectiveness) and User Satisfaction (Effectiveness)

Stakeholder Functions:

- Funding authorities play a crucial role in shaping the cost-effectiveness of the facility by providing non-recoverable financial resources for its development, infrastructure and ongoing operations. They also offer research grants and industry subsidies, helping to offset costs, and making HPC resources more accessible and affordable.
- HPC facility operators focus on optimizing resource utilization, managing operational costs and implementing cost-saving strategies, such as energy-efficient operations and effective maintenance, to enhance the facility's economic value and operational efficiency.
- Meanwhile, end-users and application developers evaluate the facility’s cost-effectiveness when deciding to commission HPC projects, taking into account factors like pricing, performance and support services to ensure their research and development needs are met efficiently and within budget.

Explanation: Cost-effectiveness (utility) is fundamental to the facility’s operational efficiency and user satisfaction, representing its ability to deliver high computational value for the resources used.

Cost-Effectiveness (utility) ↔ Operational Efficiency (effectiveness): Efficient resource utilization lowers operational costs, enabling competitive pricing that aligns with the facility's cost-effective nature.

Cost-Effectiveness (utility) ↔ User Satisfaction (effectiveness): When users perceive the facility’s pricing as fair and its performance as valuable, they are more likely to engage, enhancing overall satisfaction.

8. Technological Advancement Enablement (Utility) ↔ Scientific and Technological Impact (Effectiveness)

Stakeholder Functions:

- Funding authorities provide the financial resources needed for research, development and integration of cutting-edge technologies within the HPC facility. They also offer grants and subsidies to support projects that utilize experimental hardware and software.
- HPC facility providers design and build the infrastructure with the flexibility to adopt emerging technologies, ensuring it remains adaptable to future advancements.

- Operators play a crucial role in testing, implementing and maintaining these new technologies, continuously optimizing the system to maximize performance.
- Meanwhile, research and academic institutions, along with industry partners, drive research and innovation in HPC technologies, and collaborate with the facility to test and refine them.
- End-users and application developers engage with the facility to leverage advanced technologies, pushing the boundaries of their research and development efforts.

Explanation: Technological Advancement Enablement (utility) is a fundamental attribute of the facility that underpins its scientific and technological impact (effectiveness). The facility's ability to embrace emerging technologies defines its existence as a critical enabler of advanced research. By supporting cutting-edge computational power and experimental tools, it allows researchers to tackle increasingly complex problems and explore new scientific frontiers. This capability not only extends the range of use cases the facility can support but also drives significant scientific breakthroughs and technological innovations.

9. User-Friendliness (Utility) ↔ User Satisfaction (Effectiveness) and Scientific and Technological Impact (Effectiveness)

Stakeholder Functions:

- Funding authorities allocate resources for user interface development, training programs and support services to enhance user-friendliness. They may also fund the development of intuitive tools and workflows that make the facility accessible to a broader user base.
- HPC facility providers design the facility’s software and hardware environment to be intuitive, with straightforward user interfaces, clear documentation and accessible tools that simplify the user experience.
- Operators maintain user-friendly environments by managing support services, offering user training, creating comprehensive guides and troubleshooting common user issues. They continuously refine the system based on user feedback to streamline interaction.
- End-users and application developers interact with the facility and provide feedback on usability, highlighting areas that require improvements or additional support. Their feedback is crucial for guiding enhancements that make the facility more accessible and effective for diverse research and development needs.

Explanation: User-friendliness (utility) is a core attribute that impacts user satisfaction and scientific impact.

User-Friendliness (utility) ↔ User Satisfaction (effectiveness): An accessible, well-documented facility simplifies interactions, building user confidence and encouraging ongoing engagement.

User-Friendliness (utility) ↔ Scientific and Technological Impact (effectiveness): Lowering technical barriers makes the facility accessible to a wider range of researchers, driving advanced analyses and scientific breakthroughs.

10. Extramural Integrability (Utility) ↔ Operational Efficiency (Effectiveness), User Satisfaction (Effectiveness), and Scientific and Technological Impact (Effectiveness)

Stakeholder Functions:

- Funding authorities provide financial support for developing the infrastructure and tools required to integrate the facility with external systems, data sources and collaborative networks. They may also fund initiatives to adopt interoperability standards that promote seamless data exchange and resource sharing.
- HPC facility providers design the facility to support integration with external systems, ensuring compatibility with various software, data formats and collaborative platforms. They

play a crucial role in building an architecture that facilitates data transfer, cross-institutional workflows and cloud-based interactions.

- Operators manage and implement the integration process, including setting up external data access, coordinating with other facilities, and maintaining secure and efficient data exchange protocols. They also provide support services to guide users through the process of connecting and interacting with external resources.
- End-users and application developers rely on the facility's integrability to access external datasets, computational tools and collaborative networks. Their feedback helps identify integration needs and drives the enhancement of extramural capabilities.

Explanation: Extramural Integrability (utility) is a core attribute of the facility, affecting its operational efficiency, user satisfaction and scientific impact by enabling seamless connections with external systems and data sources.

Extramural Integrability (utility) ↔ Operational Efficiency (effectiveness): Integrability streamlines data exchange and resource sharing, reducing delays and errors, thereby optimizing the facility's operational capacity.

Extramural Integrability (utility) ↔ User Satisfaction (effectiveness): An integrable facility aligns with diverse user needs by allowing easy incorporation of external tools and data, simplifying workflows, and enhancing the user experience.

Extramural Integrability (utility) ↔ Scientific and Technological Impact (effectiveness): Access to external tools, datasets and collaborations broadens the facility's research scope, amplifying its role in scientific discovery and technological innovation.

2.4. Potential Roles of the HPC Community and its Interactions with Stakeholders

The HPC community can engage in a range of activities that involve interaction with stakeholders, shaping the utility and effectiveness of HPC facilities. These potential actions, outlined in **Section 5.2: Actions within the Community-Driven Agenda**, include:

- Education and Workforce Development
- Standard-Setting and Best Practices
- Knowledge Exchange and Collaborative Research
- HPC Infrastructure Integration
- Open-Access Data, Computational Tools and Knowledge Repositories
- Advocacy and Policy Influence

Interaction with Policy Makers and Funding Authorities

The HPC community serves as a *knowledge broker and advocate*, influencing policy and funding priorities. By providing data-driven insights, research outcomes and expertise in technological trends, the community guides policy decisions and financial allocations to support the development and growth of HPC infrastructure.

Interaction with HPC Facility Providers and Operators

The HPC community acts as a *standard-bearer and knowledge repository*, offering guidance on best practices, technological standards and operational strategies. This input directly enhances the facility's utility and effectiveness, ensuring it meets evolving performance and interoperability requirements.

Interaction with End-Users and Application Developers

The HPC community functions as a *facilitator and resource provider*, advocating for open-access data, tools and computational resources. It supports users through training programs, workshops and curriculum development, fostering effective engagement with HPC systems and maximizing the utilization of available resources.

Interaction with Regulatory and Compliance Bodies

The HPC community serves as a *consultative authority and policy influencer*, shaping operational practices of HPC facilities to align with regulatory standards. Its guidance ensures facilities operate within data privacy, security and compliance frameworks.

Interaction with Research and Academic Institutions

The HPC community acts as an *educational nexus and innovation catalyst*, driving research and fostering the development of knowledge. It curates training programs and educational content to build the expertise of current and future HPC practitioners, supporting the advancement of research and technological innovation.

Interaction with Industry Partners and Commercial Vendors

The HPC community serves as an *innovation bridge* between research and practical application. By setting standards for hardware integration, software development and system performance, it guides the design of industry products and services. Additionally, it enables knowledge exchange on market needs, technological trends and user requirements, strengthening collaboration between industry and academia to accelerate the adoption of cutting-edge technologies in real-world applications.

2.5. Conceptual Framework of the HPC Infrastructure Management Ecosystem

The various aspects of HPC infrastructure management are further discussed in the following chapters: Chapter 3 focuses on HPC facility setup, management and operation; Chapter 4 addresses public policy for HPC; and Chapter 5 explores a community-driven agenda for HPC. A conceptual framework is presented in Figure 2 to help visualize the interrelationships among the different components of the ecosystem.

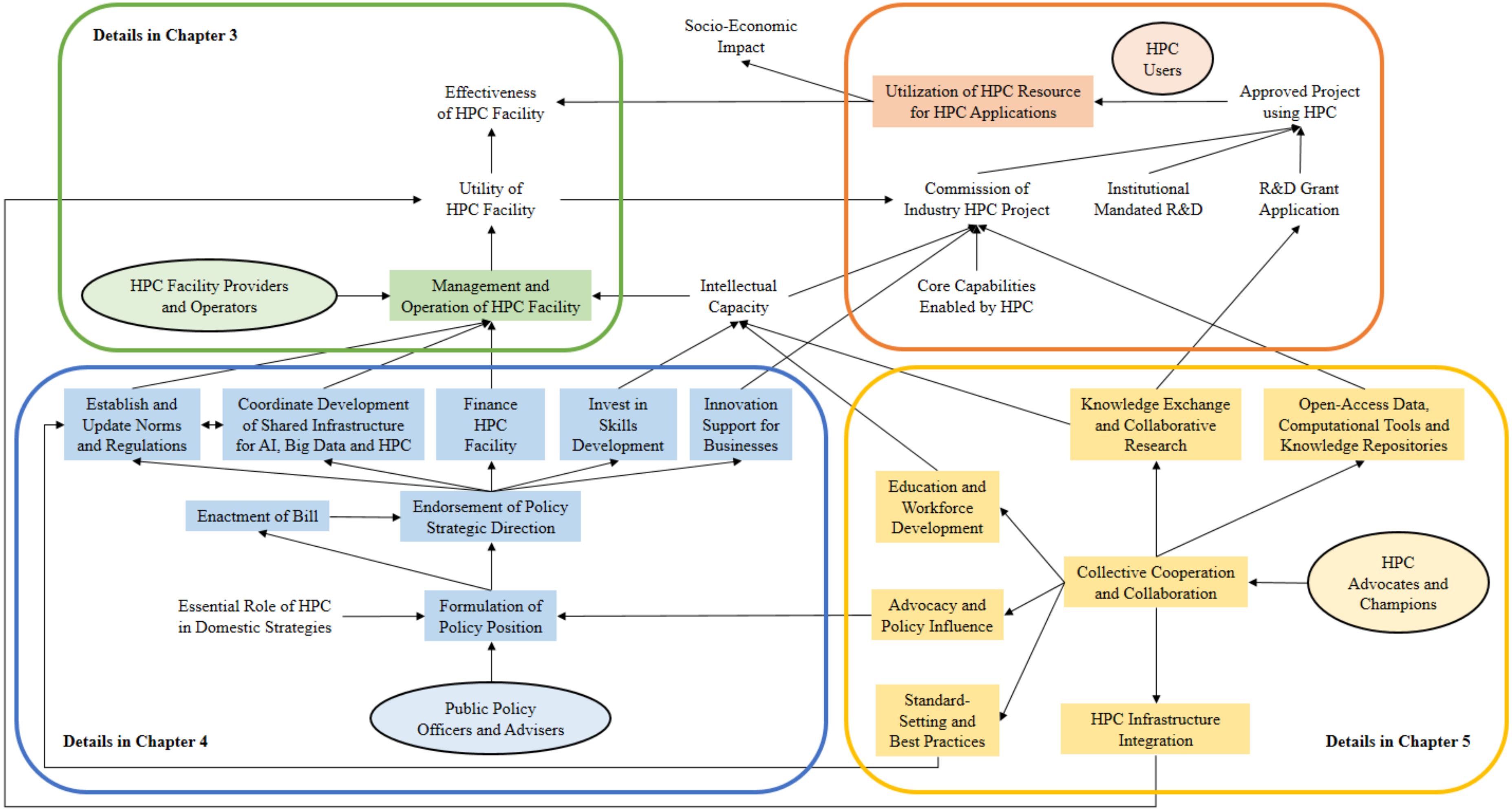
Details in Chapter 3
Effectiveness of HPC Facility
Utility of HPC Facility
HPC Facility Providers and Operators
Management and Operation of HPC Facility
Socio-Economic Impact
Intellectual Capacity
HPC Users
Utilization of HPC Resource for HPC Applications
Approved Project using HPC
Commission of Industry HPC Project
Institutional Mandated R&D
R&D Grant Application
Core Capabilities Enabled by HPC
Establish and Update Norms and Regulations
Coordinate Development of Shared Infrastructure for AI, Big Data and HPC
Finance HPC Facility
Invest in Skills Development
Innovation Support for Businesses
Enactment of Bill
Endorsement of Policy Strategic Direction
Essential Role of HPC in Domestic Strategies
Formulation of Policy Position
Public Policy Officers and Advisers
Details in Chapter 4
Knowledge Exchange and Collaborative Research
Open-Access Data, Computational Tools and Knowledge Repositories
Education and Workforce Development
Collective Cooperation and Collaboration
HPC Advocates and Champions
Advocacy and Policy Influence
Standard-Setting and Best Practices
HPC Infrastructure Integration
Details in Chapter 5

Figure 2: Conceptual Framework of the HPC Infrastructure Management Ecosystem

***

Chapter 3. HPC Facility Setup, Management and Operation

This chapter aims to provide conceptual guidance to novice *HPC facility providers and operators* on the key responsibilities of facility setup, management and operations to achieve optimal utility and effectiveness. It focuses particularly on emerging HPC environments, where constraints and challenges are greater compared to those in mature HPC infrastructures and capabilities.

These responsibilities involve overcoming complex, multifaceted challenges that encompass both technical and strategic considerations. Technically, HPC systems—comprising interconnected hardware and software reliant on utilities—have hidden interdependencies that can lead to unintended consequences, making issues difficult to detect and trace.

On the strategic side, achieving "optimal utility and effectiveness" requires balancing a broad range of factors. Utility considerations include performance, cost, scalability, reliability, accessibility, security, energy efficiency, user-friendliness, adoption and integration of emerging technologies capabilities such as quantum computing, and integration with external systems. Effectiveness, meanwhile, can be defined by operational efficiency, successful maintenance, user satisfaction, compliance, and the system's scientific and technological impact.

These goals must be pursued within the constraints of financial resources, skilled workforce shortages, emerging collaborative and industry networks, geopolitical barriers in accessing advanced HPC hardware and software, changing user demands, and the inherent limits of existing knowledge and control over outcomes of actions taken.

Therefore, this chapter explores the processes involved in establishing and managing the technical infrastructure of an HPC facility. From a technical perspective, it identifies challenges, highlights complex decision-making aspects and offers technical advices. These include mentions of policies, operating procedures and essential software tools for effective management and operation, along with best practices referenced in this white paper. On the expertise aspect, the chapter discusses human resource management, capability diffusion through consultancy services, and education and training. Financially, it outlines the budgetary realities, discusses long-term financing, and details various funding models and strategies.

3.1. Setting Up, Managing and Operating the Technical Infrastructure

3.1.1. Facility Setup and Deployment

The process workflow for *facility setup and deployment* includes defining user needs, developing infrastructure specifications—encompassing the design of the HPC system architecture and the planning of facility utilities—procurement of the HPC system, followed by its installation and testing. This process defines and actualizes the physical infrastructure of the HPC facility, while the operational capabilities will be realized in the next process – *HPC software stack deployment lifecycle*.

Defining User Needs

The fundamental measure of success for an HPC facility is its ability to meet the needs of its target users. This is inherently complex, even before considering challenges like funding shortages and difficulties in recruiting qualified personnel, which are often beyond the control of the HPC facility provider or operator.

This raises the question of defining needs, which in leading HPC economies is addressed through stakeholder engagement. Here, key representatives from industry, public research and administration collaborate to define system requirements based on actual user needs and application use cases.

This approach is particularly effective in economies with developed HPC capabilities, where there is a pool of skilled researchers experienced in integrating HPC into their R&D. These researchers have practical experience using actual HPC technology stacks and deploying HPC workloads. Thus, they are capable of articulating specific hardware architecture and performance requirements that fit their needs.

In economies with emerging HPC capabilities, researchers often lack the expertise to clearly define their hardware architecture and performance requirements, and may only have a partial understanding of their actual needs. As a result, the true value of stakeholder engagement lies in identifying the research questions and goals of prospective and early HPC adopters. This understanding is key to creating a software stack that aligns with their research objectives. For domestic highest-level or university-level HPC facilities in these contexts, the general guideline is to strike a balance between performance, flexibility and usability in system design.

Developing Infrastructure Specification

Designing HPC System Architecture

After gathering system requirements, the next critical step is to develop a detailed technical blueprint for the HPC hardware system. This blueprint typically specifies compute units, data storage systems, network architecture, and key performance metrics such as processing power, memory capacity and input/output (I/O) throughput.

Developing this blueprint involves navigating complex decisions to meet the facility's goal of maximizing utility and effectiveness. Achieving this balance requires weighing multiple factors, including computational performance, scalability, energy efficiency and ease of maintenance, all while staying within the available budget.

Additionally, the blueprint must align with the specific computing needs identified through stakeholder engagement. This ensures that the hardware configuration is optimized for the expected workloads, which may involve balancing central processing unit (CPU) and graphics processing unit (GPU) architectures, fine-tuning data management systems, and ensuring compatibility with the required software stack.

Planning HPC Facility Utilities

HPC systems generate substantial heat, are highly sensitive to humidity fluctuations and require significant power. As a result, robust environmental controls for temperature and humidity, along with a reliable power infrastructure, are essential to ensuring optimal performance and system reliability.

For temperature control, a range of cooling options is available. Air cooling solutions include air conditioning units, hot and cold aisle containment, and high performance fans directly attached to server racks. Liquid cooling methods include direct liquid cooling, immersion cooling, chilled water systems and rear door heat exchangers.

HPC facility operators report that as heat output from modern chips continues to rise, liquid cooling is becoming increasingly popular for its superior thermal management. However, immersion cooling is less favored due to maintenance challenges and compatibility concerns. Air cooling remains a practical choice for less intensive HPC systems.

In addition to temperature regulation, humidity control is critical in HPC facilities to prevent the build-up of static electricity or condensation, both of which can damage sensitive electronic components.

The power infrastructure for HPC facilities includes high-capacity power supplies (with dedicated electrical lines, uninterruptible power supplies and backup generators), power distribution units (including rack-level and high-efficiency units), power conditioning and surge protection systems, and energy management systems, such as dynamic voltage and frequency scaling (DVFS) and power usage effectiveness (PUE) optimization. Operator feedback indicates that the reliability of electricity supply varies by region, with some areas requiring uninterruptible power supply (UPS) systems to support the entire HPC system to avoid downtime during power disruptions.

HPC facility operators' experiences have highlighted that instances of HPC systems disruption, such as those caused by a burst water pipe, have occurred in HPC facilities, and pinpointing the root causes can be challenging. Selecting reliable utility vendors is essential for avoiding such issues and ensuring stability.

Constraints Consideration

The specifications for an HPC hardware system and its associated utilities are fundamentally constrained by the available budget. Allocating the initial capital investment requires complex decision-making to balance spending across hardware, utilities, software licensing and ongoing maintenance costs. It is also vital to ensure that both the HPC hardware and facility utilities incorporate sufficient redundancy to maintain continuous operation in case of hardware, cooling, or power failures. However, when capital expenditure (CAPEX) budgets are limited, redundancy is often one of the first elements to be compromised, which can impact the system’s reliability and resilience.

In addition to budgetary considerations, the *facility resource management policy* plays a crucial role. This policy defines how key resources like power, space and cooling are managed and allocated, ensuring efficient usage and guiding future expansion. It provides clear guidelines on how to expand capacity sustainably while maintaining operational efficiency, energy usage effectiveness and environmental responsibility.

HPC facility operators have observed that discrepancies between projected and actual utilization requirements can lead to critical issues, such as insufficient power supply, inadequate cooling, or unexpected space constraints. These challenges often arise during the installation phase and can prevent the HPC system from reaching its full operational capacity. To avoid these pitfalls, thorough prior planning for both internal and external system integration is essential. This includes addressing the physical placement of internal data servers, budgeting for networking infrastructure to connect these servers with the HPC system, and ensuring adequate network bandwidth and routing to external facilities used by collaborators, partners and users.

Procurement of the HPC System

Following the completion of the detailed technical blueprint, the next steps include issuing a Request for Proposal (RFP), evaluating the received proposals, selecting a vendor or system integrator, negotiating and clarifying contract details, and finally, signing the contract.

Based on insights from HPC facility operators, the procurement process is governed by public laws that emphasize transparency and require a minimum number of proposals for each tender. Typically, co-design of hardware, software and applications is prohibited, although exceptions may be allowed under specific circumstances.

The *procurement and asset management policy* plays a critical role in governing these processes, ensuring that procurement practices are efficient, funds are optimally utilized and assets are managed effectively throughout their lifecycle. This policy includes strategies for managing vendor partnerships, inventory and asset maintenance, ensuring that the acquisition aligns with both legal requirements and the facility's long-term goals.

To support this, HPC facilities can leverage both a **Procurement Management Information System (PMIS)** and an **Asset Management Information System (AMIS)**. The PMIS facilitates tracking the entire procurement cycle—from issuing the RFP to evaluating proposals and finalizing contracts. It enhances transparency, monitors vendor performance and aids decision-making by offering real-time insights into procurement activities. The AMIS, on the other hand, is crucial for tracking assets throughout their lifecycle, managing inventories, warranties and ensuring that assets are maintained or replaced as needed, thus supporting long-term resource allocation and operational efficiency.

In the typical procurement of an HPC system, unlike in co-design, the responsibility for developing a software stack tailored to the diverse needs of the target user audience is deferred to later phases. This approach becomes particularly challenging when gaps arise due to the non-existence or unaffordability of required software.

Installation and Testing of the HPC System

The installation phase follows procurement, and there are numerous potential issues that can arise during installation, such as:

- Human Error: This can range from incorrect hardware assembly and faulty wiring to mislabeling of components, each of which can lead to system malfunctions.
- Hardware Incompatibility and Defect: There may be compatibility issues between different hardware components. Additionally, any component could arrive defective or get damaged during handling, and it may not be immediately apparent.
- Network Misconfiguration**:** A misconfiguration in the network setup, such as incorrect Internet Protocol (IP) addressing or improper routing configurations, can render the system inoperable, severely degraded in performance, or create a non-apparent security loophole.

To address these issues, systematic functionality and performance testing is essential. A robust *Acceptance Testing Procedure* integrates critical knowledge of what needs to be tested, how tests should be conducted and who is responsible for each step. This quality assurance workflow defines the specific types of tests, procedures, and roles of authorities, coordinators and collaborators. It ensures that all hardware, software and network components function correctly and meet the operational standards outlined in the technical blueprint.

To enhance this process, the *acceptance testing software suite* integrates several essential tools, linking the Asset Management Information System (AMIS) with an *issue and incident management tool* and a *documentation and knowledge management tool*.

- Issue and Incident Management Tool: During installation and testing, it is critical to document and track issues such as hardware failures, configuration errors, or software bugs. This tool monitors incidents in real time, assigns responsible personnel, logs actions and ensures timely resolution. Additionally, it maintains a comprehensive history of issues, allowing teams to analyze trends, prevent recurring problems and improve overall system reliability.
- Documentation and Knowledge Management Tool: Effective documentation and knowledge sharing are key throughout installation and testing. This tool centrally stores all relevant information—such as system configurations, testing protocols, technical manuals, vendor specifications and troubleshooting guides—making it easily accessible to the team. It promotes continuity, supports training and serves as a valuable reference for addressing future issues or system upgrades. The tool ensures institutional knowledge is preserved, minimizing risks associated with staff transitions or gaps in expertise.

By integrating these tools with the Acceptance Testing Procedure, HPC facilities can efficiently manage the complex elements that must be rigorously tested. This systematic approach enables early detection and resolution of installation issues, ensuring that the system meets performance metrics and operates according to the specifications in the technical blueprint.

Box 1. Best Practice: Procedure for Acceptance Testing

One of the best practices documented by the Blue Waters project is the Procedure for Acceptance Testing. At Blue Waters, the installation of any new component, including the system's original deployment, follows a strict protocol that includes detailed acceptance planning, extensive testing, defect tracking and formal certifications, supported by multiple levels of coordination and control.

To facilitate the effective implementation of this protocol, the National Center for Supercomputing Applications (NCSA) developed a comprehensive test management Information Technology (IT) system. This system features a test bank, search and acquisition functions, test results databases, and a report generation feature, all accessible through a user-friendly interface. These integrated components allow for the efficient storage, search and recording of test templates and results.

3.1.2. Software Stack Deployment Lifecycle

The process workflow for the *software stack deployment lifecycle* includes developing the software stack architecture, translating it into software stack specifications, acquiring the necessary software—whether through procurement, development, or integration of community-supported codes—followed by software installation and configuration, and finally, system integration, testing and validation. This process transforms the potential of the HPC facility's physical infrastructure into actual operational capabilities.

Developing Software Stack Architecture and Translating into Specifications

Developing a software stack architecture for an HPC system is a complex decision-making process that involves aligning the system's operational goals, user requirements, application needs and the latest available technologies, all within the constraints of the chosen hardware.

This architecture defines and governs the interaction, function and integration of abstraction layers, including the hardware interface layer, system layer, middleware layer and application layer. The integration of these layers and their components enables interactions that create interdependencies, where compatibility is essential. Functions such as modularity, scalability, interoperability and performance optimization, are both key design considerations and emerge from the architecture's implementation of this integration.

This architecture design has to be translated into detailed software stack specifications in order to provide the necessary technical information for guiding procurement, development, configuration and integration of the software.

This process presents challenges, including achieving technical precision and appropriate granularity in the specifications, while also having the foresight to define only those specifications that can be realistically fulfilled—either through available software that meets technical requirements within financial constraints, or by developing custom solutions within the same budgetary limits.

Acquiring Necessary Software

When the software stack specifications are not defined in a way that can be realistically fulfilled, two primary challenges arise: the absence of application codes for specific application areas and use cases relevant to the target user base, and the high costs of licensed software. Typically, in environments with emerging HPC capabilities, the user base consists of individuals who rely on application codes rather than programmers developing their own codes using compilers or scripting languages. Additionally, software at higher abstraction levels, such as application codes, frequently requires paid licenses, further compounding the budget constraints.

In response to these challenges, many HPC facility operators prefer open-source software, which offers greater scalability, flexibility and adaptability to evolving HPC requirements. Open-source solutions, being community-driven, are easier to customize and extend compared to commercial software, which is often slower to adapt and comes with significant licensing costs. For these reasons, commercial licenses are often avoided due to their scalability limitations and high costs, which can become prohibitive in large-scale HPC environments.

To further streamline software inventory management, tools like an Asset Management Information System (AMIS) can be valuable. AMIS helps track software assets, manage licenses, facilitate licensing compliance and optimize procurement decisions, enabling HPC facility operators to maintain cost efficiency and operational flexibility. By leveraging such systems, organizations can better manage both open-source and licensed software stacks, ensuring alignment with technical and budgetary requirements.

One effective approach to overcoming the challenges of acquiring licensed software is demonstrated at Blue Waters, where the focus is on supporting community codes.

Box 2. Best Practice: Support of Community Codes

At Blue Waters, the emphasis is on supporting community codes rather than maintaining centralized binaries of pre-built applications. Staff members assist in documenting the build process of these community codes. Although Blue Waters does not provide access to source code or pre-built binaries, support is offered for porting and building these applications.

Software Installation and Configuration

In the context of an HPC software stack, the challenges of software installation and configuration involve correctly setting up the stack by managing complex dependencies, configuring parallelism tools, optimizing for specific workloads and ensuring robust security measures.

The installation process can be enhanced by leveraging *automation and orchestration tools* to streamline routine tasks, manage workflows, and enable efficient software deployment and scaling. Configuration can be optimized by implementing a *software configuration management policy* through *configuration management tools*, which automate the configuration, deployment, management and maintenance of the software stack, ensuring alignment with user requirements and performance objectives.

Early testing during installation and configuration is essential for identifying and resolving issues such as software incompatibilities, dependency conflicts and integration problems. This early testing establishes a stable foundation for subsequent phases of system integration.

To effectively manage the testing process, utilizing an *issue and incident management tool* can help track and address problems. Additionally, the *HPC Software TRL (Technology Readiness Level) model*, developed under the European Exascale Software Initiative (EESI), can be used to assess the software's maturity and readiness. Learnings from the testing process and readiness assessments should be captured and organized using a *documentation and knowledge management tool* to support future improvements and ensure a smooth deployment process.

Box 3. Useful Methodology: HPC Software TRL

Methodology: The HPC Software TRL model developed as one of the work tasks of the European Exascale Software Initiative (EESI) provides a systematic framework for assessing the maturity levels of software components, including application codes, libraries, frameworks, development tools and programming model implementations. This methodology is crucial for establishing a common language for profiling and evaluating readiness, which ensures that only sufficiently mature components are deployed.

Criteria: Assessment is based on documentation, support, availability, coverage, portability, scalability, performance and quality, with detailed attributes for each to facilitate comprehensive evaluations.

System Integration, Testing and Validation

After the initial installation and configuration phase, the focus shifts to system integration, testing and validation to ensure the entire software stack works cohesively with the hardware infrastructure. The goal is to optimize the software stack for the HPC environment, validate performance and confirm scalability across the system.

Various workflows can be applied for system integration, testing and validation, including incremental bottom-up, continuous integration/continuous deployment (CI/CD), parallel and big bang approaches. The choice of workflow depends on factors such as risk tolerance and the complexity of the HPC software stack, and should be documented in the operating procedure for *system integration testing*.

Comprehensive system-wide testing should be conducted before full-scale deployment to ensure the software stack functions cohesively and scales effectively across the HPC environment. Ongoing post-deployment testing is also necessary to maintain performance and reliability as the software stack evolves.

3.1.3. System Resource Allocation

Resource allocation optimization connects hardware failure risk management on the supply side with system resource allocation on the demand side. Hardware failure risk management focuses on ensuring system reliability and performance, while resource allocation aims to ensure the efficient distribution of resources to meet demand effectively.

While system resource allocation is managed through job scheduling, creating a clear and direct connection, resource allocation operates at multiple levels, such as programmatic allocation, user group quotas and project selection, which may not be as immediately apparent.

- Programmatic allocation directs financial support and institutional backing to specific programs or initiatives aligned with overarching goals, such as domestic strategies.
- User group quotas define how resources are distributed across various groups, such as industry, academia and the HPC facility's internal teams. Quotas may be set as caps, offering simplicity for HPC facility operators, or as ranges, providing flexibility in meeting policy mandates across institutions.
- Project allocation evaluates project submissions within supported programs or initiatives, determining eligibility for HPC resources and setting initial resource caps for approved projects. Additionally, proposals from paying users, whether subsidized or fully funded, are considered for HPC system resources, with resource caps defined based on their financial contribution and the facility's policies (where applicable, as some facilities do not accommodate paying users). The actual granting of resources depends on job scheduling policies, job characteristics and assigned priorities.

- Job scheduling determines how, when and what resources and software licenses are allocated to individual jobs.

While these components form the framework for system resource allocation, carrying out this process is more complex than it seems. Challenges arise not only from technical issues and competing goals but also from human factors. Stakeholders, each with their own interests, create competition, and risk aversion often results in over-requests. Additionally, power dynamics interact with governance and administrative rules, complicating decisions-making around programmatic allocation, project allocation and job scheduler policies.

Project Allocation

Projects are considered for HPC resources through various pathways, including competitive access via open calls for proposals and programmatic initiatives, where they are evaluated based on merit and alignment with the facility's objectives. Collaborative partnerships with research institutions, universities and industry also contribute projects, often aligning research goals and benefiting from long-term resource access.

Paid access allows fully funded or subsidized users, typically from industry, to submit proposals, with resource allocation determined by financial contributions and facility policies. Government-driven projects, aligned with domestic priorities, receive dedicated allocations, while educational programs and cross-border collaborations provide additional access routes.

In all cases, HPC centers follow a project allocation process where users apply for resources, and proposals are evaluated based on scientific merit, strategic importance, or financial contributions.

The administrative, evaluative and managerial tasks for project allocation include:

Call for Proposals, Programmatic Initiatives and Submission Management

- Developing Open Calls for Proposals and Programmatic Initiatives: Crafting detailed guidelines and requirements for both competitive access proposals and specific programmatic initiatives. Ensuring alignment with the HPC facility's strategic goals and scientific priorities, including domestic or institutional initiatives.
- Managing Proposal Submissions: Creating submission systems for easy entry of project proposals under both open calls and programmatic initiatives. Ensuring smooth communication with applicants throughout the process.
- Publicizing Calls for Proposals and Programmatic Opportunities: Promoting both open calls and programmatic initiatives to relevant researchers, institutions and industry partners to ensure a broad pool of applicants.

Evaluation and Selection Process

- Establishing Evaluation Criteria: Defining the standards by which proposals are judged, such as scientific merit, innovation, alignment with strategic goals and potential impact.
- Forming Review Panels: Assembling expert panels from academia, industry and government to assess submitted proposals.
- Coordinating Proposal Reviews: Managing the review process to ensure consistent evaluations based on established criteria.
- Selecting Projects for Approval: Facilitating review panel discussions to determine which projects should be recommended for HPC resources.

Project Resource Allocation and Prioritization

- Defining Resource Caps: Setting limits on the HPC resources (e.g., CPU hours, memory) that approved projects can access, based on their needs and the facility's capacity.

- Setting Job Priorities: Establishing priority tiers to ensure that critical or high-impact projects receive appropriate priority in job scheduling.

These responsibilities are carried out either directly by the HPC facility provider or in coordination with external committees or stakeholders. Regarding the evaluation and selection process, HPC facility providers note that approaches can vary by economy—some facilities use external review committees, while others rely on internal decision-making processes.

The decisions made regarding which projects are allocated resources (on ledger), as well as the resource caps and job priorities set for each project, place significant demands on the job scheduling system. These demands include:

- Overcommitted Resources: Approved projects may collectively request more resources than are immediately available.
- Enforcing Resource Caps: Ensuring that projects do not exceed their allocated resources (e.g., CPU hours, memory).
- Limited Software Licenses: Coordinating jobs that require limited software licenses adds complexity to scheduling.
- Conflicting Priorities: Some projects are deemed more critical due to their strategic importance, requiring job scheduling to prioritize certain jobs over others. This can lead to delays for lower-priority projects and challenges in maintaining fairness.
- Job Preemption: Higher-priority jobs may require preempting lower-priority ones, adding additional strain on scheduling efficiency and system utilization.
- Time-Sensitive Jobs: Certain projects may have strict deadlines, especially for industry partners or government-related initiatives, requiring the job scheduler to accommodate these time-sensitive tasks without unduly delaying others.

The interaction between project allocation and job scheduling underscores the importance of strategic resource management. Balancing immediate resource demands, long-term fairness and system efficiency is a complex task, requiring both human decision-making and automated scheduling algorithms to work together effectively.

Job Scheduling

Job schedulers are software tools that implement the administrator's selected scheduling policies to manage the allocation of system resources. These policies govern when jobs are executed, which resources they utilize and how efficiently those resources are optimized. The scheduling policy specifies which techniques are applied from a range of possible options to achieve several key sub-goals, including:

- Job Prioritization and Preemption: The policy selects a specific technique (e.g., queue-based sorting, job ranking, wait time) to determine job priority, establishing their execution order and whether lower-priority jobs can be preempted by higher-priority ones.
- Resource Allocation and Limits: The policy defines the technique (e.g., resource limits, fairshare, routing) for controlling the amount of resources assigned to jobs, as well as the technique (e.g., job limits per project, user, or group, round-robin queues) for limiting the number of jobs a user or group can run.
- Time Slot Allocation: The policy specifies time slots during which particular jobs are allowed to run, ensuring resources are reserved for those jobs within a defined time window.
- Job Placement Optimization: The policy specifies placement optimization settings for how virtual nodes are organized, how jobs are distributed and how resources are assigned for efficient system performance.
- Resource Efficiency Optimization: The policy specifies techniques (e.g., backfilling, tracking dynamic resources, avoiding highly loaded nodes) to improve throughput, minimize job turnaround time and maximize resource efficiency.
- Overrides: The policy allows for manual intervention to bypass standard scheduling behavior.

Selecting a job scheduling policy involves more than just picking techniques; it requires balancing technical and subjective considerations, while also addressing specific demands and constraints. Administrators must manage trade-offs between objective factors like resource efficiency and job wait times, along with considerations of fairness, licensing constraints and the unique needs of the system.

Subjective Factors:

- Job Priority versus Fairness: Prioritizing certain jobs based on importance or urgency is inherently in conflict with the goal of distributing resources equitably across users and projects.
- Short-Term Efficiency versus Long-Term Fairness: Some scheduling approaches optimize for immediate resource usage, while others distribute resources more evenly over the long term.

Objective Factors:

- Maximizing Resource Utilization versus Minimizing Job Wait Time: Maximizing resource usage and minimizing job wait times can result in different outcomes depending on system load and job characteristics.
- Throughput versus Latency: Job scheduling may focus on processing a higher number of jobs overall (throughput) or reducing the time jobs spend waiting in the queue (latency).
- Static versus Dynamic Resource Allocation: Static allocation locks resources for the duration of a job, while dynamic allocation adjusts resources as they become available.
- Energy Efficiency versus Performance: Lowering energy consumption can affect system performance and job schedulers handle the trade-offs between these two factors.

Specific System Demands:

- Small Jobs versus Large Jobs: Small jobs can be processed quickly and efficiently, while large jobs require more resources and have different scheduling requirements.
- Resource Contiguity versus Fragmentation: Some jobs benefit from contiguous resource allocation, while others can operate efficiently with fragmented resources.

Licensing Constraint:

- License Management: Software license availability can influence job scheduling, as jobs requiring specific licenses must wait until they are available. This dependency can cause delays or bottlenecks in job scheduling and may require reallocation of resources to other projects or jobs in the queue.

The complexity of job scheduling can lead to extended queue times for certain types of jobs, particularly those with specific characteristics that make them more challenging to schedule. These characteristics include:

- Resource-Intensive Jobs: High CPU, memory, or specialized hardware (e.g., GPUs) demands can cause delays due to limited availability.
- Large Jobs: Jobs requiring many or contiguous resources may wait longer for the right combination of resources to become available.
- Low-Priority Jobs: Jobs with lower priority rankings are often pushed back in favor of higher-priority tasks, resulting in longer wait times.
- Long-Running Jobs: Jobs with long execution times may be delayed in favor of shorter jobs that optimize system throughput.
- License-Dependent Jobs: Jobs requiring specific software licenses may be delayed by limited license availability.
- Preemptible Jobs: Jobs that can be interrupted by higher-priority tasks may experience extended wait times as they restart.
- Specialized Resource Needs: Jobs requiring specific nodes, environments, or configurations may wait until those specific resources are free.

Extended queue times can delay the timely execution of project milestones for approved projects and may also affect the evaluation and selection process during project allocation, as resource availability and scheduling efficiency become critical factors in determining project feasibility and timelines.

License Management

While many HPC facility operators prefer open-source software, and it is possible for HPC centers to function without relying on licensed software by exclusively deploying software packages or applications that do not require licensing, this section offers guidance for operators who either use or are considering the use of licensed software in their facilities.

Managing software licenses in an HPC environment involves two key aspects: acquiring the appropriate licenses and ensuring operational compliance with licensing terms.

Acquiring the Appropriate Licenses:

- Understanding License Types and Limitations: Software licenses vary in their restrictions. For example, some are tied to specific hardware (node-locked), while others allow shared use across multiple systems (floating). A comprehensive understanding of these different license models and their limitations is essential to ensure that licensing aligns with the facility's operational needs and compliance requirements.
- Planning for Peak Demand: Software usage often fluctuates, particularly during peak periods like major project deadlines. Anticipating these demand spikes and planning accordingly can help ensure that sufficient licenses are available, avoiding bottlenecks and delays in processing.
- Tracking License Usage for Future Planning: Monitoring license usage over time can provide valuable insights into usage patterns, helping to inform future decisions. It may reveal whether additional licenses are necessary or if adjustments in usage strategies can optimize existing licenses. This approach can also prevent overpaying for underutilized licenses.
- Procuring the Necessary Licenses: Once needs and trends are understood, securing the right number and type of licenses becomes more straightforward. This may involve purchasing new licenses, renewing existing ones, or expanding current license agreements, all depending on the workload and operational needs.
- Bring Your Own License (BYOL): Users bring their own pre-purchased software licenses to the HPC environment and the center facilitates the installation and execution of the software on its infrastructure.

Ensuring Operational Compliance with Licensing Terms:

- Monitoring License Availability: Using tools that track license availability in real time helps ensure that jobs requiring specific licenses are not delayed due to license shortages. This is particularly useful for preventing downtime caused by licensing issues.
- Incorporating Licensing into Job Scheduling: While job schedulers like Slurm or OpenPBS efficiently manage system resources, they typically do not handle software licenses natively. However, certain commercial job schedulers provide integrated license management, ensuring that jobs requiring licenses do not exceed available resources. This helps maintain compliance with licensing agreements and avoids potential legal complications.
- Managing Software Inventory and Licensing Changes: Keeping an up-to-date inventory of licensed software is essential, as updates or new versions may introduce changes to licensing terms. In some cases, software that was previously license-free may now require a license. Staying informed about these changes ensures compliance and helps avoid issues related to outdated or evolving license requirements.

In addition to internal management, it is crucial that users are also aware of the licensing terms for any commercial software they utilize. Clear communication about these terms helps prevent unintended license violations and ensures smoother operations. By effectively managing software licenses, HPC facilities can optimize resource use, minimize delays and maintain compliance, ultimately improving overall operational efficiency.

3.1.4. Hardware Failure Risk Management

Hardware failures refer to malfunctions or breakdowns in physical components that disrupt normal system operations. These failures can severely impact the performance, reliability and availability of systems, and are often caused by defects, wear and tear, environmental factors, or improper handling.

The types of hardware failures include:

- Component Malfunctions: Failures in critical components such as hard drives, CPUs, memory, or power supplies, leading to system crashes or operational disruptions.
- Wear and Tear: Over time, hardware components naturally degrade, resulting in unexpected shutdowns, reduced performance, or complete failure.
- Connectivity Issues: Failures in network hardware (e.g., routers, switches, cables) or connection points that cause communication breakdowns, affecting system and network availability.
- Environmental Damage: Physical damage caused by environmental factors such as excessive heat, moisture, or power surges, which can lead to hardware failure or permanent damage.
- Improper Handling or Installation: Damage caused by mishandling, incorrect installation, or inadequate maintenance, which can shorten the lifespan of hardware components and lead to failure.

Managing hardware failure risk involves combining sound judgment for subjective decision-making in strategy development with deep technical and administrative expertise to effectively operationalize the strategy into practical actions.

Developing Hardware Failure Risk Strategy

Developing a hardware failure risk strategy involves addressing challenges related to malfunctions or breakdowns in physical components within complex systems. It requires subjective decision-making due to the unpredictable nature of failures and conflicting priorities. Key factors include:

Nature of the Problem:

- System Complexity: High performance computing (HPC) systems are intricate, with many interdependent hardware components. This complexity makes it difficult to predict how a hardware failure in one area may impact the entire system.
- Unpredictability of Failures: Hardware failures, such as component malfunctions or connectivity issues, are often unexpected. The likelihood and impact of these failures can change over time, making prediction and prevention more difficult.
- Rapid Technological Changes: Constant advancements in hardware introduce new risks that may not have been anticipated, making it essential to continuously update risk management strategies.

Nature of the Solution:

- Performance versus Reliability: Decision-makers often face trade-offs between optimizing system performance and ensuring hardware reliability. For instance, pushing hardware components to their performance limits can increase the likelihood of failure.
- Cost versus Risk Mitigation: Mitigating hardware failures through redundancy, preventive maintenance, or hardware upgrades can be expensive. Decision-makers must balance the costs of these strategies with the potential financial impact of downtime or failures.
- Conflicting Priorities: Stakeholders from different departments (e.g., IT, finance) often have conflicting priorities. IT may prioritize hardware stability, while finance may focus on minimizing costs. Balancing these priorities requires careful consideration.

The process of systematically developing a hardware failure risk strategy consists of three key phases: risk identification, risk assessment and the planning of risk mitigation strategies.

In the risk identification phase:

- The process begins with an *asset inventory* to identify and catalogue all critical hardware components, leveraging the Asset Management Information System (AMIS), as discussed in **Section 3.1.1: Facility Setup and Deployment**. This step establishes the foundation for determining which components are essential and which are more prone to failure.
- A *failure mode analysis* is then conducted to examine potential failure points for each hardware component, such as power supply issues, overheating, or disk drive malfunctions.
- Additionally, an *environmental and operational assessment* is performed to evaluate how factors such as temperature, humidity and operational loads (e.g., workload, uptime demands) may affect hardware reliability.

In the risk assessment phase:

- The *likelihood of hardware failures* is evaluated based on several factors, including the age of components, historical failure data and environmental conditions. This likelihood evaluation helps determine the probability of hardware failures.
- Following this, an *impact analysis* is conducted to assess the potential consequences of hardware failures on system operations, including downtime, data loss and business continuity.
- After completing both the likelihood and impact assessments, the *risks are prioritized* to focus on the most critical risks that require immediate mitigation.

In the final phase, planning risk mitigation strategies:

- A high-level plan for *preventive measures* is developed. This phase overlaps with discussions in **Section 3.1.1: Facility Setup and Deployment**, regarding decisions related to facility utilities for environmental controls, hardware system specifications for redundancy and procurement processes for vendor support agreements. Preventive actions may include creating proactive maintenance schedules and implementing environmental controls, such as cooling systems, to minimize the risk of hardware failure.
- A *redundancy strategy* is also formulated to ensure that critical hardware components have backups in place, maintaining business continuity in the event of a failure.
- The need for *support agreements* is assessed, determining whether vendor agreements like warranties or service level agreements (SLAs) are required to secure timely responses to hardware failures.
- Finally, *risk avoidance plans* are devised, focusing on replacing high-risk components that are nearing the end of their life cycle or showing signs of failure, thereby proactively mitigating potential disruptions.

Establishing Hardware Failure Risk Management Practices

Operationalizing the hardware failure risk strategy involves implementing preventive measures, monitoring systems, responding to failures and continuously improving practices. This process also includes defining policies, establishing clear operating procedures and utilizing software tools to effectively manage hardware risks.

The implementation of preventive measures includes several key actions:

- Hardware Preventive Maintenance: Implement a maintenance schedule that ensures regular checks, inspections, cleaning and replacement of aging hardware components in HPC systems and facility utilities, including environmental controls and power infrastructure.

- Environmental Controls: Maintain and monitor environmental controls such as temperature regulation, humidity control and dust management to improve hardware longevity and minimize the risk of failure.
- Redundancy Deployment: Deploy backup hardware components, such as servers and power supplies, to ensure system continuity in the event of hardware failure.
- Preventive Maintenance Policy: Establish a policy outlining the responsibilities for hardware preventive maintenance, assigning system administrators tasks like monitoring system health, adhering to maintenance schedules and managing environmental controls. The policy should also define collaboration with facility teams for inspections, cleaning and hardware replacements.

Key Practices for Monitoring and Detection:

- Monitoring and Alert Systems: Install and configure tools to track hardware health in real-time, monitoring metrics such as disk health, CPU temperature and power supply status. These systems should provide early warnings to prevent failures and include automated alerts that notify system administrators of issues like overheating, performance degradation, or impending disk failures.
- Software Tools: Leverage specialized tools such as *hardware health monitoring tools*, *environmental monitoring tools* and *alert management tools* to continuously monitor and manage hardware performance.
- Monitoring Response Procedure: Establish clear procedures for responding to monitoring alerts, guiding system administrators in addressing alerts based on severity, escalating issues when necessary and resolving hardware problems promptly.

Key Practices for Incident Response:

- Failure Detection Protocols: Establish protocols to quickly identify and isolate hardware failures, minimizing downtime. This includes diagnosing the root cause of the problem (e.g., power supply failure, disk crash) and taking corrective actions.
- Repair and Recovery Procedures: Document step-by-step procedures for repairing failed hardware and restoring system functionality, including replacing failed components and testing systems after recovery to ensure proper operation.
- Hardware Failure Response Policy: Develop policies that establish clear protocols for responding to hardware failures, including response times, escalation procedures and administrator responsibilities during the recovery process.
- Data Recovery Protocol: Establish protocols for data recovery in the event of hardware failure, including restoring from backups, Redundant Array of Independent Disks (RAID) recovery, or other methods, to efficiently recover lost data and restore system functionality as quickly as possible.

Continuous Improvement through Post-Failure Review and Predictive Monitoring:

Key Practices for Post-Failure Review and Documentation:

- Incident Documentation: Thoroughly document each hardware failure, including its cause, resolution steps, recovery time and system impact.
- Lessons Learned: Conduct post-incident analysis to identify areas for improvement in hardware failure risk management, such as adjusting maintenance schedules, upgrading hardware, or refining response procedures.
- Updating Preventive Measures: Based on post-failure reviews, adjust preventive maintenance practices, including modifying routines or replacing aging hardware components earlier to prevent future failures.
- Post-Failure Documentation and Review Procedure: Establish clear procedures for incident documentation and post-failure reviews, outlining how to capture critical information, share insights with relevant teams and update risk management practices as needed.

Key Practices for Predictive Failure Monitoring and Risk Reassessment:

- Predictive Failure Monitoring: Continuously monitor hardware performance to detect early warning signs of potential failures, such as declining disk performance or overheating, enabling proactive intervention.
- Software Tools: Utilize specialized tools such as *predictive maintenance tools* and *performance monitoring tools* for ongoing monitoring and performance analysis.
- Risk Reassessment and Adjusting Risk Management Strategies: Regularly reassess hardware risks, particularly as systems age or operational demands change, and adjust risk management strategies to address emerging risks or shifts in hardware performance.
- Hardware Predictive Failure Monitoring and Risk Management Policy: Define policies specifying the responsibilities of system administrators for monitoring potential hardware failures using predictive tools and conducting regular risk reassessments. These policies should ensure that early warning signs of hardware degradation are identified and addressed promptly.

3.1.5. Cybersecurity Risk Management

Managing cybersecurity risk for an HPC center is more appropriately viewed as a wicked problem, rather than as a simplistic collection of isolated preventive measures, continuous monitoring, incident response and recovery practices. Several key factors contribute to this complexity:

- Dynamic Risks: Evolving threats and the need for continuous adaptation.
- System Complexity: Interconnected components and dependencies like utilities add risk.
- Performance Trade-offs: Security measures can impact performance.
- Regulatory Challenges: Traditional Information and Communication Technology (ICT) regulations often misalign with HPC needs.
- Diverse Stakeholders: Balancing varying security expectations complicates strategy development.

Thus, effectively managing cybersecurity risks for an HPC center requires a comprehensive and tailored approach that integrates security into every aspect of the system’s design and operation, while accounting for the unique characteristics of HPC environments.

- This requires sound judgment to balance performance-security trade-offs, a thorough understanding of legal and regulatory frameworks to ensure compliance, and strong mediation and communication skills to address the diverse needs of stakeholders facilitating the thoughtful development of a cybersecurity strategy and operational practices.
- It also requires deep technical and administrative expertise to implement proactive threat modeling, secure system architecture, performance-optimized security protocols and continuous monitoring.

Developing Cybersecurity Strategy

The process of systematically developing a cybersecurity strategy for an HPC facility consists of three key phases: assessment and analysis, needs and expectations alignment, and requirements and policy development.

Assessment and Analysis: The first phase focuses on gaining a deep understanding of the current state of the HPC facility's assets, risks and cybersecurity landscape.

- Asset Identification: Catalogue all critical assets—hardware, software, data and network infrastructure—using the Asset Management Information System (AMIS), as detailed in **Section 3.1.1: Facility Setup and Deployment**. This step establishes a foundation for understanding the role of each asset within the facility and determining its specific security requirements.
- Risk Assessment: Identify potential threats, vulnerabilities and associated risks. This includes external threats (e.g., malware, cyberattacks), internal threats (e.g., insider misuse, system

misconfigurations) and third-party risks (e.g., supply chain vulnerabilities, vendor security weaknesses).

- Threat Landscape Evaluation: Analyze the evolving threat landscape specific to HPC environments, including risks introduced by the use of cutting-edge hardware and community-developed software. This evaluation accounts for emerging vulnerabilities and sophisticated attack vectors that are unique to HPC infrastructures.

Needs and Expectations Alignment: This phase is crucial for aligning the cybersecurity strategy with the operational, performance and security needs of various stakeholders within the HPC environment. The goal is to ensure that security measures address the needs of all parties without compromising system performance.

- Stakeholder Engagement: Engage key stakeholders, including researchers, system administrators, regulatory bodies and funding agencies, to gather insights into their cybersecurity requirements and compliance needs. Facilitate collaboration to resolve competing interests, such as balancing the demands for high computational performance with stringent data security.
- Performance, Expectation and Security Trade-offs: Assess acceptable trade-offs between security and system performance, considering the performance sensitivity of HPC environments. It is essential to determine how much performance can be compromised to enhance security. Additionally, evaluate trade-offs between security and user expectations, such as access control and virtualization technologies. Align cybersecurity goals with the facility's operational objectives, including protecting sensitive data, ensuring compliance and minimizing downtime.

Requirements and Policy Development: This phase defines cybersecurity requirements and establishes high-level policies to address them.

- Requirements Specification: Define the cybersecurity requirements based on the trade-off decisions made in the previous phase, balancing the need to protect the facility from assessed cybersecurity risks while maintaining the performance efficiency and features expected by users. These requirements are also shaped by the need to comply with relevant cybersecurity regulations.
- Policies Creation: Develop policies that outline high-level, strategic actions aimed at meeting the defined cybersecurity requirements

Operationalizing Cybersecurity Strategy

Operationalizing a cybersecurity strategy for an HPC facility involves translating defined cybersecurity requirements and policies into practical, actionable steps that integrate security measures throughout the system's lifecycle. This proactive approach ensures security is embedded into the architecture, setup and ongoing operations of the HPC environment, rather than treated as an afterthought. While this approach shares similarities with traditional IT security practices, HPC environments introduce unique challenges—such as performance demands, specialized hardware, and open-source and self-developed software—that require tailored solutions.

These distinct considerations are comprehensively addressed in *NIST Special Publication 800-223: High-Performance Computing Security: Architecture, Threat Analysis, and Security Posture*. To structure these solutions effectively, the National Institute of Standards and Technology (NIST) publication introduces a lexicon and reference architecture for HPC systems, dividing them into four functional zones: the High-Performance Computing Zone, the Data Storage Zone, the Access Zone and the Management Zone. This division facilitates the description and targeted application of security measures across the various parts of the HPC infrastructure, allowing for tailored security strategies based on each zone's specific needs. Key distinct practices in HPC environments include:

- ScienceDMZ Architecture for Data Transfer: Rather than using a traditional firewall, HPC environments often rely on the ScienceDMZ architecture for data transfer nodes to avoid the performance bottlenecks imposed by firewalls. This specialized architecture is designed to optimize high-speed data transfers while maintaining security.

- Network Segmentation for Access Control: HPC environments implement access control through network segmentation, separating management networks, high performance networks and auxiliary networks. This segmentation ensures that different types of traffic are isolated and properly managed, enhancing both security and performance.
- Compute Node Sanitization: Ensuring the proper sanitization of compute nodes between jobs is a critical security practice in HPC. This helps prevent residual data or processes from being accessed by unauthorized users or subsequent jobs, maintaining data integrity and confidentiality.
- Securing the Software Supply Chain: With a strong reliance on open-source and custom-built software, HPC environments face unique challenges related to the software supply chain. Managing dependencies, and conducting vulnerability testing and code audits are critical to ensuring that open-source or self-developed software components do not introduce security risks.
- Secure Diskless Booting Image: In diskless booting HPC environments, boot images may contain sensitive information, such as Secure Shell (SSH) keys. It is crucial to implement measures that secure these images from unauthorized access.
- Container Security: As containerized applications become more common in HPC environments, ensuring their security is crucial. It is important to verify that containers and their dependencies are sourced from trusted providers. Leveraging tools designed to audit container contents helps ensure the integrity and security of containerized workloads.

In addition to the distinct cybersecurity practices specific to HPC environments, many security controls are universally applicable across various infrastructures. A key reference for these controls is the ISO/IEC[3] 27002 standard, *Information Security, Cybersecurity, and Privacy Protection – Information Security Controls*, which provides comprehensive guidance on a broad range of security controls. Below are some practices that are relevant to both HPC environments and traditional IT security:

Access Control and Authentication:

- Role-Based Access Control (RBAC): RBAC restricts user access to resources based on their roles, ensuring that privileges are aligned with specific responsibilities to prevent unauthorized access.
- Multi-Factor Authentication (MFA): MFA adds an additional layer of security during authentication by requiring more than just a password, reducing the risk of unauthorized access.
- Remote Administration Controls: Limit remote administrative access to critical systems by enforcing access restrictions, such as Media Access Control (MAC) address filtering, to prevent unauthorized remote access.
- File System Permissions: Enforce strict permissions to control who can access or modify files, ensuring that only authorized users can interact with sensitive data.

Data Protection and Privacy:

- Encryption: Encryption of data at rest and in transit safeguards sensitive information from unauthorized access or interception.
- Data Integrity Protections: Techniques like checksums and hashing ensure data integrity throughout its lifecycle, preventing unauthorized alterations.
- Privacy-Preserving Technologies: Tools such as data anonymization, obfuscation and differential privacy help protect sensitive data while allowing it to be used for analysis and research.

[3] ISO/IEC (International Organization for Standardization/International Electrotechnical Commission)

System and Software Maintenance:

- Software Updates and Patch Management: Regular updates and security patches help address vulnerabilities. Automating this process reduces the risks posed by outdated software.
- Firmware Updates: Regular firmware updates close security gaps in hardware, preventing potential exploitation.

Monitoring and Incident Response:

- Intrusion Detection/Prevention Systems (IDS/IPS): IDS/IPS monitor network traffic for suspicious activity and respond to potential threats, helping to prevent unauthorized access.
- Security Information and Event Management (SIEM): SIEM systems aggregate logs, analyze real-time security events, detect anomalies and enable swift incident response.
- Incident Response and Recovery: A well-defined incident response plan and recovery protocols are crucial for mitigating security incidents and restoring normal operations quickly.

Security Testing:

- Penetration Testing: Routine penetration testing identifies vulnerabilities before they can be exploited, helping organizations strengthen their defenses against malicious actors.

HPC facility operators emphasize that cybersecurity is a critical priority across all centers, with measures such as network monitoring, automated updates and two-factor authentication commonly implemented. However, compliance standards can vary significantly, with government-level security typically being more stringent than that of research institutions. Network security strategies also differ between centers; some utilize firewalls, while others rely on IP-restricted access, dedicated data transfer nodes and additional countermeasures. Cybersecurity budgets also varies, with some centers designating specific portions to enhance security. Notably, HPC facility operators point out that a fully developed disaster recovery infrastructure is not always considered essential for research-focused facilities.

Some notable best practices at Blue Waters include their security model, approach to software updates and security patching, and their bug tracking process.

Box 4. Best Practice: Security Model

At Blue Waters, a multi-pronged security model is implemented. All traffic is monitored by a network intrusion detection system, and SSH access is fully logged with keystroke capture to detect anomalous behavior. Two-factor authentication is used to mitigate the risk of fraudulent credentials. Privileges cannot be escalated, even by administrators, when accessing the system through user access points. Privilege propagation is unidirectional across multiple server-client subsystem hierarchies. Access to the system is controlled through membership in Lightweight Directory Access Protocol (LDAP) groups, which are governed by Privileged Access Management (PAM) access control measures.

Box 5. Best Practice: Approach to Software Updates and Security Patching

At Blue Waters, software updates are deployed using a two-step approach: initially, functionality and performance testing is conducted on a test and development system, followed by a secondary test on the full system after deployment. The protocol for security patching mandates quick action, with critical vulnerability patches typically applied and the system rebooted within twenty-four hours of receipt from the vendor.

Box 6. Best Practice: Bug Tracking

At Blue Waters, bugs are recorded using issue and project tracking software that also manages all system problem reports. If a bug is traced to a vendor-supplied component, the issue is escalated to the vendor, and both the vendor case number and the bug number are documented in the ticket. An automated service seamlessly updates the bug's status in the vendor's system directly into Blue Waters' tracking software. Status updates may include stages such as awaiting correction, correction received, installed and tested, and problem resolved. Additionally, Blue Waters monitors and reports on both vendor and internal responsiveness, tracking metrics such as time to human response and time to resolution.

3.1.6. User Support

In providing user support, especially when facing a potentially high ratio of users to support staff, it is crucial to keep users informed, invest in scalable support solutions and robust training programs, and continuously refine these measures. Without such proactive steps, HPC facility operators risk entering a capacity over-burden death spiral, where escalating demands on user support exceed the organizational capacity, leading to a progressive deterioration of service quality and operational failure.

HPC facility operators have categorized user support into three distinct levels: basic IT and administrative support, technical and domain-specific support, and scientific collaboration for complex projects.

- Across all levels, strategies include establishing regular feedback loops and periodically reviewing the capacity and scalability of support measures in place.
- For basic IT and administrative support, strategies include implementing automated ticketing systems, developing self-service knowledge bases and FAQ (frequently asked questions) portals, deploying artificial intelligence (AI) enabled virtual assistants, and regularly updating training materials and conducting staff training sessions.
- For technical and domain-specific support, strategies involve implementing advanced monitoring and reporting tools to effectively track resource usage and optimize system performance, employing internal collaborative problem-solving platforms to document and share collective organizational knowledge, establishing and nurturing a user community through community forums and workshops, and conducting specialized staff training programs.

3.2. Managing, Diffusing and Building Expertise and Capacity

3.2.1. Human Resource Management

Staffing public HPC centers presents a universal challenge for HPC facility operators. Key roles typically include system engineering, user support, business operations and promotional activities. Positions in facility operations and IT engineering are particularly difficult to fill due to high market demand and the specialized licenses required. Moreover, specialized positions such as domain experts in HPC and AI are challenging to attract and retain due to the lower salaries typically offered in the public sector compared to the private sector.

To address recruitment and retention challenges, HPC centers might consider these strategic approaches:

1. Developing a pipeline of progressive internship, training and educational programs, specifically tailored to the center's operational needs and aligned with the refresh rate of the center’s workforce.

2. Implementing a strategy to engage staff in continuous organizational learning and improvements, thereby enhancing operational efficiency and employee engagement.
3. Incorporating intrinsic motivation into job roles to boost job satisfaction and commitment.

HPC facility operators observe that remote work has emerged as a trend post COVID-19 (coronavirus disease of 2019) and offers a viable solution to attract talent. While outsourcing and managed services traditionally help address staffing challenges, they are not suitable for niche roles like domain experts. Additionally, generative AI and chatbots are proposed as alternatives to fill certain positions, especially in user support roles.

3.2.2. Diffusing and Building Expertise and Capacity

In leading HPC economies, some operators deliver comprehensive consultancy services that include:

- Technical Consultancy: This service focuses on technology extension, searching and assessing technologies, and facilitating technology transfer.
- Project Consultancy: Assistance spans from support for HPC project call applications and project design to overall project management.

In addition to their core activities, some operators engage in education and training as part of cross-border collaborations or academic partnerships to cultivate a skilled workforce. This includes:

- Curriculum Development: Designing educational programs that integrate HPC concepts and skills, tailored to diverse educational levels to ensure a continuous pipeline of proficient professionals.
- Workshops and Webinars: Conducting a series of workshops, webinars and training sessions that equip researchers, students and professionals with essential HPC skills and knowledge, crucial for advancing their capabilities.
- Certification Programs: Offering certification programs that provide formal recognition of HPC expertise, enhancing career prospects.

From the perspective of an economy, the ultimate goals of these efforts are to amass a critical mass of intellectual capital within the economy and to capitalize on this intellectual capacity through industry and public research laboratories for economic and social impact. Accordingly, the metrics for evaluating education and training should encompass not only activity-based measures, such as the number of trained students, but also impact metrics. Furthermore, the objective of HPC facility operators providing consultancy services should be to facilitate the industry's absorption of this capacity, rather than merely generating financial returns to compensate for shortfalls in public funding for the HPC center.

3.3. Budgetary Reality, Long-Term Financing and Near-Term Funding

3.3.1. Budgetary Reality in Environments with Emerging HPC Capabilities

The primary differences between economies with emerging HPC capabilities and those leading in the field lie in their intellectual capital, and the strength of their collaborative and industry networks. Leading HPC economies benefit from a well-established pool of specialized talent and strong connections between businesses and academia, which enable effective knowledge sharing and collaboration. This makes it easier for them to realize the economic and social benefits of their HPC investments, securing political and industry support.

In these leading economies, funding for domestic highest-level HPC centers is viewed as a strategic investment, with public funds covering capital expenditure (CAPEX) and most of the operational expense (OPEX). This approach allows academic researchers to access HPC resources without the burden of strict cost recovery protocols.

In contrast, economies with emerging HPC capabilities often face budgetary constraints that hinder the development of essential infrastructure and talent. Limited funding affects critical aspects such as system redundancy, vendor support and management, resulting in operational challenges for HPC centers.

Even with political support for HPC initiatives, budget constraints typically force compromises to be made. CAPEX funding often falls short, limiting the ability to procure HPC systems that meet performance needs or to provide the necessary support for continuous, reliable operations.

Moreover, constrained budgets and the emphasis on high returns, as highlighted by sources like Hyperion Research, often push policymakers toward implementing cost recovery for OPEX. For an HPC center, this can result in budget shortfalls, forcing cuts in maintenance, the build-up of technical debt and difficulties in retaining skilled staff—further undermining the performance and reliability of HPC systems.

Cost recovery-focused policies also limit the development of intellectual capacity. Many small and mid-sized businesses, as well as academic researchers, struggle to afford HPC resources. Additionally, there is a shortage of research leaders who have both the foresight to pre-allocate budgets and the expertise to accurately estimate the costs of required HPC usage. These challenges impede the accumulation of intellectual capital needed for HPC projects to generate significant economic and social benefits.

3.3.2. Sustained Financing

The long-term sustainability of domestic highest-level HPC centers hinges on securing genuine political buy-in and understanding from policymakers, who need to recognize which actions will advance or hinder the HPC agenda.

The economic and social impacts of successful HPC applications are substantial. Leading HPC economies experience significant returns in terms of financial gains, innovation, job creation and increased economic output, alongside improved employment prospects for graduates, enhanced societal intellectual capacity, better public services, more informed policymaking and optimized public administration.

However, the successful implementation of HPC use cases critically depends on the availability of specialized talent. Building this intellectual capital is a time-intensive process that is also subject to the constraints of time-compression diseconomy, which suggests that rushing this process can lead to diminished returns. The distinct advantage of economies with advanced HPC capabilities lies in their well-developed intellectual capital.

Policy makers need to understand that metrics for assessing the development of this intellectual capital should vary across different stages. For economies with emerging HPC capabilities, initial metrics should focus on the HPC facility operator’s ability to provide reliable access to advanced computing resources. As the intellectual capital develops, success metrics should evolve to measure how well the HPC centers meet the needs of their specific user communities and, eventually, the impact of successful HPC use cases.

Thus, for HPC facility providers, it is crucial to address misconceptions about the requirements for realizing these benefits while advocating for the potential advantages of HPC investments. Continual public funding is essential for the sustainability of HPC centers, and withdrawing funds prematurely can waste the developing intellectual capital.

Additionally, it is vital to capitalize on any successful HPC use cases—whether they occur serendipitously, through prioritized research projects with strong potential outcomes, or via cross-border collaborations. Publicizing these successes can build credibility, potentially enhanced by endorsements from reputable third parties like consulting companies. Creative approaches such as featuring HPC in popular media, like dramas or television shows, can also effectively raise public awareness and demonstrate the impact of the centers.

Implementing governance structures and advisory groups for HPC centers can provide valuable external perspectives, insights and connections, enhancing collaboration and the visibility of operations.

3.3.3. Funding Models and Strategies

A determined HPC facility operator facing a funding shortfall can advocate for top-slicing—a mechanism where the HPC center receives a predetermined percentage or specific amount from the budgets of various funding entities—as well as government grants and subsidies. Additionally, they can explore alternative funding models, such as:

Usage-Based and Service Charges Model:

- Compute Product Usage-Based Charge: Charges based on actual usage of resources like processing time or storage, aligning costs with consumption but requiring robust monitoring systems. The pricing should consider operational costs and market value, with potential inclusion of capital expenses for certain users.
- Service-Level Based Fees: Fees linked to guaranteed service levels (e.g., uptime and support), providing premium revenue but requiring consistent service quality.
- Data Management Fees: Charges for data management services, offering a steady revenue stream but necessitating significant infrastructure and competitive pricing.
- Technical Extension Services Charge: Fees for specialized technical services like consulting or training, leveraging technical expertise for additional revenue but requiring skilled personnel and facing competitive pressures.

Academic Partnership Model:

- Infrastructure Sharing and Co-Investment: Collaborations with academic institutions to pool resources for HPC infrastructure, reducing individual financial burdens and enhancing access to advanced capabilities.
- Educational Partnerships: Joint educational initiatives with academic institutions, enhancing skills and generating revenue through enrolments or sponsorships.
- Joint Research Grants: Collaborative efforts with academic partners to secure research grants, covering project costs and potentially leading to shared revenue from new technologies or intellectual property.

Industry-Collaboration Model:

- Co-Investment in HPC Infrastructure: Partnerships with companies to co-invest in HPC resources, reducing costs and aligning infrastructure with industry needs.
- Membership Programs: Establishing membership arrangements where companies financially support the HPC center in exchange for benefits like priority resource access and networking opportunities.
- Joint R&D Projects: Collaborative projects with private sector companies to tackle industry-specific challenges, develop new products, or enhance existing technologies, with potential for commercialization and shared revenue.

These models offer a range of strategies for HPC centers to secure funding and sustain operations through innovative and collaborative approaches.

Additionally, targeted marketing efforts aimed at recipients of public policy instruments can enhance conversion rates. Table 1 organizes the functional scope, public policy purview areas and public policy instruments, along with their associated funding recipients.

Functional scope	Public policy purview areas	Public policy instruments	Funding recipients
Research and development (R&D)	Science and technology (S&T)	Grants (Research / R&D / collaboration)	Public research entities
Innovation skills and human capital	Higher education	Grants (Research / R&D / collaboration)	Public higher education entities
	Labor	HPC-AI workforce up-skilling programs	Implementing entities for programs
Enterprise development and innovation	• Enterprise • Industrial • Innovation	• HPC-AI internship programs • Grants (R&D / innovation) • Loans (R&D / innovation) • Tax incentives • Startup incubators • Startup accelerators	Businesses
Public services	• Meteorology • Health • Transportation • Etc.	Programs where HPC-AI usage is appropriate for public service's mission delivery	Public services entities

Table 1: Functional Scope, Public Policy Purview Areas and Instruments with Associated Funding Recipients

***

Chapter 4. Public Policy for HPC

This chapter aims to address the latent need for a deeper understanding of high performance computing (HPC) among public policy officers and advisers—a need that often goes unrecognized but is crucial for effective policy formulation. It focuses on enhancing their awareness of the critical role of HPC in domestic strategies for Industry 4.0, digital transformation, smart cities and addressing societal challenges. Additionally, the chapter seeks to improve policymakers' ability to develop effective strategies that leverage HPC across diverse policy portfolios, including science and technology (S&T), higher education, research and development (R&D), innovation, and artificial intelligence (AI).

The chapter emphasizes the importance of adopting strategic and sustained financing for HPC facilities, rather than relying on competitive funding, where HPC facilities are not specifically targeted but are merely eligible to compete for research funds to establish, operate and upgrade. To support these objectives, this chapter highlights the critical role of HPC in domestic strategies, outlines key considerations for developing holistic public policy for HPC, explains the need for strategic and sustained financing for HPC facilities, and presents three case studies showcasing how Japan; Korea; and the United States have successfully achieved sustained financing for their domestic HPC facilities.

4.1. Essential Role of HPC in Domestic Strategies

HPC serves as a cornerstone of domestic strategies in Industry 4.0, digital transformation, smart cities and addressing societal challenges. It accelerates innovation for technology developers, enables optimization and augment strategic decision-making ability for businesses that adopt these technologies, enhances urban planning and management, and provides powerful tools to address complex societal issues.

Industry 4.0

In the landscape of Industry 4.0, two key players emerge: innovators and adopters. Innovators are businesses or organizations that develop cutting-edge technologies, focusing on research and development to push technological boundaries. Adopters, on the other hand, integrate these innovations into their operations to transform and optimize their processes.

For innovators, HPC is indispensable for conducting advanced research and development (R&D). It accelerates progress by bridging gaps left by traditional empirical and theoretical methods, enables new possibilities through large-scale and multi-scale integration modeling, and addresses complex combinatorial challenges, such as parameter sweeps, variations in initial conditions and multiple model configurations (refer to **Section 1.2.1: Core Capabilities Enabled by HPC** for details).

Digital Transformation

Digital transformation involves converting different forms of information into digital formats, making data more manageable, analyzable and shareable. The role of HPC in this transformation is twofold:

- Logistical Optimization: When digitalized data relates to logistics—such as supply chain data, transportation and inventory management—HPC enables companies to analyze and optimize these processes. Its computational power allows for complex simulations that can streamline supply chains, improve route planning and enhance inventory control, leading to greater efficiency and cost reductions.
- Strategic Decision-Making: Beyond operational improvements, HPC also plays a pivotal role in strategic planning. By processing and analyzing vast amounts of digital data, HPC allows businesses to extract insights, identify trends, simulate potential scenarios and make data-

driven decisions. This supports long-term planning, market strategy adjustments and risk management, making digital transformation not just about efficiency but also about strategic agility.

Smart Cities

HPC is vital in the development of smart cities, where the integration of digital technologies with urban infrastructure leads to smarter and more efficient urban management:

- Urban Planning and Simulation: HPC enables cities to model complex urban systems, such as traffic flows, energy consumption patterns and waste management processes. These simulations help planners design more efficient, sustainable urban spaces, allowing cities to optimize resources and reduce their environmental footprint.
- Real-Time Data Processing: Smart cities generate enormous amounts of real-time data from internet-of-things (IoT) sensors, smart grids and transportation networks. HPC processes these data at high speeds, allowing cities to quickly adapt to changes, such as rerouting traffic to ease congestion or adjusting energy distribution in response to demand. This real-time capability is crucial for maintaining the functionality and responsiveness of smart city systems.

Addressing Societal Challenges

Computational modeling is another area where HPC plays a crucial role in tackling societal challenges like climate change and public health crises. These challenges are often conceptualized as many-body problems, which involve understanding and analyzing the interactions among numerous entities.

- For example, in *climate modeling,* HPC can simulate how atmospheric particles interact with each other and with environmental factors, providing insights into how different interventions might impact climate outcomes.
- In public health, HPC enables *epidemiological modeling*, simulating the spread of diseases to predict potential outcomes and design effective intervention strategies.

By treating these challenges as many-body problems, HPC allows researchers to explore various scenarios and test potential solutions before implementing them in reality. This helps in developing effective strategies, whether it is for reducing carbon emissions or improving public health outcomes.

4.2. Considerations for Developing Holistic Public Policy for HPC

A holistic public policy for HPC ensures that the benefits of HPC are widely distributed and fully utilized, addressing the needs of different sectors, fostering collaboration, investing in infrastructure and skills, and aligning HPC efforts with broader domestic goals. It aims to create an environment where HPC can thrive and drive innovation, while also considering issues like accessibility, security and environmental sustainability. Such an approach maximizes the potential of HPC to contribute to economic growth, scientific advancement and the resolution of complex societal challenges.

Recognizing Synergetic Relationship between AI, Big Data and HPC

Public policy could greatly benefit from recognizing the synergetic relationship between AI, Big Data and HPC, and fostering collaboration that harnesses the unique strengths of each technology to accelerate innovation.

AI and HPC complement each other in model creation. AI relies on learning-based approaches, such as machine learning and deep learning, which train models on large datasets to recognize patterns and make predictions. Meanwhile, HPC uses non-learning approaches, like equation-based modeling (applying mathematical equations to describe system behaviors) and agent-based modeling (simulating interactions among individual entities). While AI learns from data to refine its models, HPC focuses on simulating and analyzing complex systems through direct computation, making them ideal partners for solving diverse modeling challenges.

AI workloads involve training models to identify patterns, make predictions and automate decision-making, requiring large datasets and significant computational power, especially for deep learning. Big Data workloads focus on the collection, storage, processing and analysis of massive volumes of structured and unstructured data, extracting insights and identifying trends. HPC workloads handle complex simulations and computational modeling, using high-speed processors to solve intricate scientific, engineering and mathematical problems. Together, these technologies manage a variety of data-driven and computationally intensive tasks, each enhancing the capabilities of the others.

Given the similarities in the computing systems required to support AI, Big Data and HPC workloads, a policy approach that promotes the coordinated development of shared infrastructure—including computing systems, internet connectivity and supporting utilities—can offer significant benefits. This approach would optimize the use of computational resources, reduce costs and accelerate progress across these interconnected fields.

Investing in Skills Development

The ability of businesses to fully capitalize on HPC relies heavily on the availability of specialized talent. Accumulating intellectual capital to a critical mass, however, is a time-intensive process. It is also subject to the constraints of time-compression diseconomy, which suggests that expediting this process can lead to diminished returns.

In this context, universities and research institutions play a crucial role in cultivating the expertise needed. Their efforts, guided by public policy on science and technology (S&T), research and development (R&D), and higher education, are essential for ensuring a steady pipeline of skilled professionals.

Box 7. Tiered Classification of HPC Systems

One approach to organize HPC systems into categories is the tiered classification. Reflecting the operational capacity, scale and intended applications of an HPC system, these systems, typically supported by government, can be organized into a tiered classification:

- Tier-0: Intergovernmental HPC systems designed for the most demanding global scientific and engineering challenges.
- Tier-1: Domestic highest-level HPC systems aimed at supporting an economy's strategic scientific research and industrial applications.
- Tier-2: University-based HPC systems, publicly or privately owned, primarily used for a diverse range of research activities.
- Tier-3: HPC systems located at public research laboratories, tailored to meet specific scientific and engineering needs of those facilities.

These entities do more than train individuals; they align skills with the specific needs of various HPC tiers, from basic research to advanced application development. This structured approach ensures that each tier functions effectively and cohesively. Consequently, it creates a pipeline of increasingly competent individuals, the most skilled of whom may enter the industry or collaborate on projects with it, thereby fostering a dynamic interplay between academia and the commercial sector.

Tier-0 (Intergovernmental HPC Systems): Utilizing Tier-0 systems necessitates a cross-border consortium of top-tier researchers and engineers who specialize in complex, large-scale computational projects. Tier-0 systems act as hubs for pioneering scientific discoveries and technological breakthroughs, requiring the highest levels of computational power and collaborative research. Such collaborations serve as fertile grounds for the creation of new knowledge, spiraling through the mixing of diverse ways of thinking and proven practices. This newly generated knowledge can then be cascaded down to enrich the domestic intellectual ecosystem, enhancing both the capacity and capabilities at lower tiers. This process not only fosters global scientific leadership but also fortifies the domestic intellectual capital by integrating global innovations and expertise.

Tier-1 (Domestic Highest-level HPC Systems): Effective utilization of Tier-1 domestic highest-level HPC systems relies on the expertise of leading domestic scientists and engineers who focus on critical research areas aligned with the economy's strategic interests in science and technology. The research and development conducted within these systems often drive innovations that can transform industries, boost competitiveness and security, and address societal challenges. Projects undertaken at this tier serve as vital links to industries and domestic agencies, facilitating the cross-diffusion of knowledge and fostering a comprehensive understanding of the technical and operational challenges inherent in practical applications.

Tier-2 (University-based HPC Systems): University-based HPC systems play a pivotal role in academic research and education across a wide array of disciplines. These systems integrate computational with empirical and theoretical research, and are instrumental in developing new computational methods, exploring innovative research questions and providing practical training to students. By enabling hands-on experience with advanced computing resources, Tier-2 systems help cultivate the next generation of scientists and engineers. The intellectual capital developed here is not only critical for academic progression but also vital for industry innovations, as it bridges theoretical knowledge with real-world applications. Collaborations between universities and industries facilitated by these HPC systems can lead to the commercialization of research, turning academic insights into marketable products and services.

Tier-3 (Public Research Laboratory HPC Systems): Public research laboratory HPC systems are tailored to address specific scientific and engineering challenges specific to their mission-critical contexts. Tier-3 systems are designed for high performance tasks that require robust and secure computational support. The specialization and intensity of research conducted here demand a deep understanding of domain-specific challenges and a high degree of technical skill. The outcomes of Tier-3 research typically have direct implications for policy and industry practices, enhancing the economy's capabilities in critical areas and contributing to societal well-being.

Supporting Innovation through Direct and Indirect Financial Support for Businesses

The application of HPC in R&D, engineering, logistical optimization and strategic decision-making can significantly boost business competitiveness. However, many small and mid-size enterprises (SMEs) face challenges in accessing these resources due to high costs.

Public policy can play a crucial role by offering grants, subsidies and tax incentives targeted at SMEs, enabling them to invest in HPC technology or gain access to shared HPC facilities. These measures help to lower the financial barriers, reduce risks and make advanced computing more accessible to smaller businesses, fostering innovation and growth.

Establishing and Updating Norms and Regulations

A holistic public policy for HPC requires the continuous development and refinement of regulations that address technical standards, data privacy, security protocols and the responsible use of HPC resources. This includes creating guidelines to ensure interoperability, privacy protection and cybersecurity, as well as compliance with domestic security and export control requirements.

Additionally, effective policy must encompass the coordination of critical infrastructure—including energy supply, water supply and high-speed internet connectivity—to support the specific operational demands of HPC facilities. This ensures that HPC systems are not only compliant and secure but also have the robust infrastructure needed for optimal performance.

Investing in HPC Facilities

HPC facilities serve as the enabling infrastructure that makes HPC policies actionable and impactful. Investing in these facilities establishes the computational backbone essential for driving advanced research, innovation and economic growth.

Such investments empower industries and research institutions to fully harness the potential of HPC across diverse fields, from scientific discovery to industrial optimization, thereby ensuring that HPC policies achieve their intended outcomes.

4.3.Need for Strategic and Sustained Financing of HPC Facilities

Given the high costs, congestion risks and rapid obsolescence of HPC facilities, strategic and sustained financing is necessary, rather than reliance on general competitive funding models. Targeted investment is critical because HPC facilities underpin HPC policies and drive innovation. Consistent support ensures that governments can fully harness the benefits of HPC, protect their investment in intellectual capital, and maintain a competitive and effective research infrastructure.

Challenges of HPC Facilities:

- High Costs: HPC facilities require substantial investments to establish, operate and upgrade, due to the need for cutting-edge hardware, specialized infrastructure and skilled personnel. Both initial investment and ongoing maintenance are costly.
- Congestibility: Limited capacity can lead to congestion, where high demand results in too many users competing for resources. This can reduce efficiency and slow down access, negatively impacting the quality and timeliness of research.
- Rapid Obsolescence: HPC technology evolves rapidly, making today’s state-of-the-art hardware and software obsolete within a few years. Continuous investment is essential to keep facilities up to date and competitive on a global scale.

Strategic Importance of HPC Facilities:

- Despite these challenges, HPC facilities are vital infrastructure that make HPC policies actionable and effective. They are essential for advanced research, innovation and the implementation of domestic strategies in areas such as Industry 4.0, digital transformation, smart cities and addressing societal challenges.
- HPC facilities contribute directly to competitiveness. Having modern, accessible HPC infrastructure is a key factor in an economy’s ability to drive innovation and maintain a competitive edge in the global research landscape.

Limitations of Non-Targeted Competitive Funding:

- Non-targeted competitive funding refers to general research funds that all projects and facilities can compete for, without specific provisions for HPC. While this approach supports various research activities, it fails to guarantee that HPC facilities will receive the sustained and substantial funding needed for their development and upkeep.
- Given the high costs and strategic importance of HPC facilities, relying solely on non-targeted funding is inadequate. It risks underfunding critical infrastructure, leading to delays in upgrades, capacity limitations and an inability to keep pace with technological advancements.

Need for Strategic and Sustained Financing:

- To ensure the effective implementation of HPC policies, a more strategic and sustained financing approach is required. This involves allocating dedicated funding streams specifically for the establishment, operation and continuous upgrading of HPC facilities.
- A strategic funding approach acknowledges the long-term value of HPC facilities in driving research and innovation. It ensures that these facilities operate at the necessary scale and sophistication to meet the economy's research and development objectives, maintaining their effectiveness over time.

Box 8. Approaches to Sustained Financing of HPC Infrastructure

The approach to financially supporting HPC infrastructure (HPCI) varies across different economies. Both the United States and Korea have enacted legislative acts focused on HPC. Other than the European Union, no additional economies have enacted laws that in effect ensure the financial sustainability of their HPCI.

Regardless of a legislative mandate, government officials may choose to act on HPC matters. Their policy tools—also known as policy instruments—which provide both direct and indirect funding for HPCI, typically include provision of research infrastructure, pre-commercial procurement, public-private partnerships, and research and development (R&D) grants.

Details on public policies related to these instruments can be sourced from legislative acts, executive orders, strategies, master plans, roadmaps and implementation plans. These policies often form part of broader frameworks that encompass science and technology (S&T), higher education, research and development (R&D), and innovation. They may be integrated into holistic policy frameworks of the Industrial 4.0, digital transformation and artificial intelligence or exist specifically as comprehensive policies on HPC.

4.4.Case Study: The United States

The High Performance Computing Act of 1991 stands as a pivotal piece of legislation, establishing a coordinated federal program that has proven highly effective. It has played a crucial role in maintaining and boosting funding for high performance computing (HPC) facilities across various federal agencies throughout the United States.

This federal program, originally established as the High Performance Computing and Communications (HPCC) program, has evolved over time. It is now known as the Networking and Information Technology Research and Development (NITRD) program. This program has been refreshed, revised and extended through The Next Generation Internet Research Act of 1998, America COMPETES Act of 2007, Cybersecurity Enhancement Act of 2014, and American Innovation and Competitiveness Act of 2017.

For the fiscal year (FY) 2023, the government budget, as enacted in the appropriation bills, allocated USD 1.907 billion to support the NITRD program's component area of High Capability Computing Infrastructure and Applications. This allocation marks a significant increase from the USD 917.9 million provided in FY2013.

This sustained financing has enabled public research laboratories in the United States to receive and operate three exascale supercomputers: the Frontier supercomputer at Oak Ridge National Laboratory (ORNL), delivered in 2021; the Aurora 21 supercomputer at Argonne National Laboratory (ANL), installed in June 2023; and the El Capitan supercomputer at Lawrence Livermore National Laboratory (LLNL), with installation commencing in 2023.

4.4.1. Legislation

As a case study, it is essential to thoroughly examine the circumstances that enabled the HPC agenda to attain legislative attention, be enacted into law and consistently secure annual funding for HPCI.

Raising Agenda Profile to Garner Legislative Attention

Before the bill was introduced for legislative consideration, the HPC agenda had already gained significant traction in public policy circles. Advocacy by the scientific and research community, along with influential reports, raised awareness in Congress. Coincidentally, the public was galvanized by a controversial claim in the 1990 book titled '*The Japan That Can Say No'*, co-authored by Akio Morita and Shintaro Ishihara.

Reports, functionally similar to *policy position papers*, were published by the Office of Science and Technology Policy (OSTP) under the Executive Office of the President (EOP), proposing an HPC strategy and its implementation plan:

- The November 1987 report by the Federal Coordinating Council for Science, Engineering and Technology (FCCSET) committee, titled '*A Research and Development Strategy for High Performance Computing*', provided a strategy for HPC development. It essentially set forth the policy problem—"U.S. high performance computer industry leadership is challenged by government-supported research and development in Japan and Europe"— and established the policy goal of "maintaining high performance computing leadership".
- The September 1989 report from the FCCSET Committee on Computer Research and Applications, titled '*The Federal High Performance Computing Program*', detailed the implementation plan for this strategy. Layering on top of activities to be performed and roles of actors, it presented a problem-solving approach that focused collaboration and cooperation efforts on "Grand Challenges", of which twenty were summarized.

Additional influential reports which built the case for the *policy rationale* included:

- The December 1987 report prepared by the Subcommittee on Science and Engineering Computing of the FCCSET Committee on Computer Research and Applications, titled '*The U.S. Supercomputer Industry*', among its findings, highlighted issues such as the "U.S. supercomputer leadership is threatened" and "Impacts on the supercomputer industry are not thoroughly considered in formulating trade policy".
- The September 1989 background paper by the Office of Technology Assessment, titled '*High Performance Computing and Networking for Science'*, explored key issues concerning the federal role in supporting HPC facilities, and in developing a research and education digital connectivity network.
- The March 1991 report prepared by Gartner Group for the US Department of Energy and Los Alamos National Laboratory, titled '*High Performance Computing and Communication: Investment in American Competitiveness*', estimated the economic impact of the proposed Federal HPCC Program.

Furthermore, the following quote from page 5 of '*The Japan That Can Say No*', despite its oversimplification[4] of the complex issue of leadership in the semiconductor chips industry, galvanized *public opinion*:

> "…, if Japan stopped selling them the [semiconductor] chips, there would be nothing more they could do. If, for example, Japan sold [semiconductor] chips to the Soviet Union and stopped selling them to the U.S., this would upset the entire military balance."

[4] It did this by focusing on specific domains within the semiconductor industry while generalizing their impact on broader military dynamics.

Navigating Politics to Advance the HPC Bill

Building on the groundwork laid by influential reports and public opinion stirred by the provocative claim of dependency on Japan's critical computer chips, the bill was strategically crafted as a bipartisan agenda. It was designed with a policy rationale that commanded high legislative priority and positioned as a coordinated federal program to garner broad support. Influential lawmakers sponsored it as well.

Political Common Ground**:** The bill’s policy rationale, invoking prosperity and security, attracted bipartisan support. It framed “advances in computer science and technology” as vital for “the Nation’s prosperity, national and economic security”, among others, while highlighting that the United States’ leadership in “the development and use of high performance computing for national security, industrial productivity, and science and engineering” “is being challenged by foreign competitors”.

Legislative Priority**:** Security remained high on the legislative agenda through early 1991. Although the fall of the Berlin Wall in November 1989 symbolized the decline of the Soviet Union, the Cold War had not officially ended by January 1991, when the bill was introduced. The Soviet Union was formally dissolved later, in December 1991.

Broad Support**:** The bill proposed a coordinated federal program that established a collaborative framework involving federal agencies, academia and the private sector. It aimed to maintain leadership in HPC by extending and synergizing existing activities, thus securing support from implementing entities. Additionally, the program sought to streamline the implementation of current policies by eliminating duplications, thereby reducing public expenditure. This cost-saving strategy is designed to appeal to broader interests in Congress by balancing targeted spending with fiscal responsibility.

Bill’s Sponsorship**:** Furthermore, the bill was introduced as a bipartisan effort led by Senator Al Gore, a prominent figure and former candidate for the Democratic presidential nomination in 1988, along with the support of eighteen other senators. This collective effort provided the political influence necessary to advance the agenda.

Maintaining Budgetary Attention to Secure Annual Funding

Before securing the necessary political consensus for its enactment, the bill and the HPCC program were meticulously crafted to align with established budgetary processes. This strategic alignment significantly enhanced the likelihood of securing annual government funding through appropriations to individual agencies and departments for the program's component areas, notably for the HPCI.

The bill's strategic design mandates that each federal agency and department involved in the program identify funds for its component areas in their annual budget submissions. Importantly, since the program’s inception, ‘High Performance Computing Systems’—which later evolved into ‘High Capability Computing Infrastructure and Applications’—has always been one of the component areas.

Thus, this strategic alignment establishes a *mechanism* that enables both the President and Congress to scrutinize the annual budget request allocations for each component of the HPCC program across all participating federal agencies and departments. It also provides these agencies and departments with an invaluable annual opportunity to submit and justify their budget requests, particularly for HPCI, which requires sustained financing.

4.4.2. Policy Instruments

When discussing the evolution of procurement practices since the initiation of the HPCC program, it is useful to categorize HPC systems into two distinct types:

- Leadership Systems: The leading-edge high capability computers that will enable breakthrough science and engineering results for a select subset of challenging computational problems. These are problems that have been unsolvable with currently available computing resources.
- Production Systems: Computers that address the challenging computational problems that require high-end computational resources but do not require access to the extraordinary Leadership Systems.

Since the initiation of the HPCC program, federal agencies and departments have developed various strategies for procuring HPC resources:

- Self-Operated HPC Systems: Some entities chose to procure and manage their own HPC systems, with a particular focus on Production Systems built from commercial-off-the-shelf (COTS) components through public procurement. This approach is recognized for its cost-effectiveness and aligns with the policy instrument of funding the *provision of research infrastructure*.
- Negotiated Resource Agreements: Other entities have pursued agreements for the use of existing HPC resources, allocating costs to *R&D grant* programs they administer. This strategy has allowed for the utilization of existing resources without incurring the full costs associated with system procurement and operation.

The enactment of the Department of Energy High-End Computing Revitalization Act of 2004 introduced additional procurement strategies that expanded the landscape, particularly emphasizing the procurement of Leadership Systems:

- Procurement of R&D Services: This strategy entails acquiring R&D services specifically aimed at developing HPC hardware and software, facilitated through *pre-commercial procurement*. This approach encourages innovation and development ahead of commercial availability.
- Integrated Co-Design and Development: This integrated strategy, utilized in the Exascale Computing Initiative (ECI) in 2016, co-develops hardware components (such as processors, memory systems and interconnects), software layers (including operating systems, compilers and middleware) and applications (like scientific simulations and data analytics) concurrently. Supported by *public-private partnerships*, this approach fosters collaboration between government and industry to drive technological breakthroughs in HPC.

4.4.3. HPC Policy Documents

The key HPC policy documents are listed below for reference.

High Performance Computing:

- Strategy: (November 1987) A Research and Development Strategy for High Performance Computing
- Implementation plan: (September 1989) The Federal High Performance Computing Program
- Legislative act: The High Performance Computing Act of 1991

High-End Computing:

- Roadmap: (June 2003) The Roadmap for the Revitalization of High-End Computing
- Legislative act: Department of Energy High-End Computing Revitalization Act of 2004
- Implementation plan: (May 2004) Federal Plan for High-End Computing

NITRD Program:

- Strategic plan: (July 2012) The Networking and Information Technology Research and Development (NITRD) Program 2012 Strategic Plan

National Strategic Computing Initiative:

- Strategic plan: (July 2016) National Strategic Computing Initiative Strategic Plan

Future Advanced Computing Ecosystem:

- Strategic plan: (November 2020) Pioneering the Future Advanced Computing Ecosystem: A Strategic Plan
- Roadmap: (May 2022) Future Advanced Computing Ecosystem Strategic Plan FY2022 Implementation Roadmap

4.5. Case Study: Korea

The enactment of the National Supercomputing Promotion Act of 2011 positioned Korea as the second economy globally to legislate specifically on the theme of high performance computing (HPC). This law established a comprehensive legal framework designed to harness the underexplored socio-economic potential of HPC and facilitate resource mobilization for supercomputing infrastructure.

Following the law's enactment, the project 'Super Korea 2020' was initiated, aiming to develop Korea's fifth-generation supercomputing infrastructure. Approved in 2015, this project led to the operational launch of Nurion, the fifth-generation supercomputer, in 2018 at the Korea Institute of Science and Technology Information (KISTI), Korea's designated National Supercomputing Center.

Building on this momentum, the National Supercomputing Innovation Strategy (2021-2030) was launched in 2021, outlining plans for the sixth and seventh generations of supercomputing infrastructure. The development of the sixth-generation supercomputer is underway, with its initial operations, originally scheduled for 2023, being delayed. The seventh-generation supercomputer is slated to begin operations in 2028.

For this case study, the core inquiry examines how Korea, which operates under a presidential government system yet lacks a direct counterpart to the United States' Office of Science and Technology Policy (OSTP)—established by Congress in 1976—successfully transitioned a significant socio-economic research agenda into legislative act and subsequent budgetary allocations for HPC infrastructure.

4.5.1. Legislative Action

The origins of Korea's legislative foray into supercomputing unfold as a narrative richly imbued with both serendipity and strategic evolution, illustrating a vivid interplay between luck and path dependency.

Dr Jysoo Lee's Leadership: In 2004, Korea's journey took a pivotal turn with the appointment of Dr Jysoo Lee as the director of KISTI Supercomputing Center, a position he held from 2004-2006 and then again from 2009-2012. At the time, the full scope of his impact could not be fully anticipated. His tenure marked the beginning of a transformative era for Korea's HPC. Dr Lee's leadership extended beyond traditional roles, catalyzing profound changes within the HPC community in Korea. His contributions included growing Korea's HPC community and advocating for supportive legislative measures. This advocacy and community development, recognized in hindsight, earned him the SupercomputingAsia HPC Leadership/Achievement Award in 2022, highlighting his significant and lasting impact on the economy's technological landscape.

The United States Legislative Influence: Parallel to Dr Lee's initiatives, a critical development was happening in the United States. In June 2003, a seminal workshop orchestrated by an interagency task force produced '*The Roadmap for the Revitalization of High-End Computing*'. This document, pivotal in its own right, provided an HPC policy blueprint that captured the attention of policymakers and technologists alike, eventually leading to the Department of Energy High-End Computing Revitalization Act of 2004. This legislative move, unknowingly, set a global reference that would echo across continents.

Korean Parliamentary Inspection Directive**:** Inspired by the United States' strides in HPC legislation, Korea's legislative dynamics were set into motion. A parliamentary inspection during the 250th regular session of Korea's National Assembly in 2004 catalyzed a critical directive—a study for legal and systematic measures to vitalize the economy's supercomputing. This action transcended mere procedural formalities; it was a momentous move that bridged the chasm between a significant socio-economic research agenda and legislative deliberations. By leveraging developments in the United States, Korea was propelled to define its own legislative milestones in supercomputing.

KISTI Instrumental Role**:** This call for a study initiated a series of detailed inquiries and legislative deliberations that spanned several years. Central to this process was the office of KISTI, which, under the leadership of Dr Lee at the Supercomputing Center, played a crucial role. Through extensive policy studies and HPC community building, this office was instrumental in advancing the legislative agenda on HPC.

After years of meticulous groundwork and advocacy, the collective efforts culminated in the enactment of the National Supercomputing Promotion Act in 2011. This act was not merely a legislative victory but also marked a historical milestone of significance for Korea's strategic, path-dependent journey in HPC. A journey that was shaped by a blend of blind luck—the unexpected guidance from the United States' legislative precedent—and purposeful luck, stemming from laborious, competent and concerted efforts by all involved.

4.5.2. Budgetary Allocation

Korea's evolution from passing critical legislation to achieving substantial government funding for developing fifth and sixth-generation supercomputers exemplifies a path-dependent trajectory, driven by strategic foresight and diligent execution.

Diligent Execution of Legislative Mandate: Under a legislative mandate to develop a five year Master Plan, the crafting of the first National Supercomputing Master Plan (2013-2017) took a year to complete. This endeavor required overcoming the challenges of understanding complex technical details and achieving consensus among a diverse array of stakeholders.

Strategic Foresight on Funding Justification: Instead of merely upgrading existing capabilities based on the economy's third-generation supercomputers, the plan introduced ambitious strategic goals. These included the vision to position Korea among the top seven in supercomputing by 2017 and the objective to secure a top ten global ranking for the economy's supercomputer. These targets played a crucial role in securing justification for a significant increase in future budget allocations necessary for robust implementation.

Diligent Execution of Master Plan: This initial phase culminated in the late 2012 endorsement of the master plan, paving the way for the 'Super Korea 2020' project. This project aimed to procure and operate the fifth-generation supercomputer with a proposed budget of USD 200 million, which received approval in 2015. With the operational launch of the fifth-generation supercomputer, Nurion, in 2018, Korea achieved a significant milestone, placing Nurion 11th globally on the Top 500 list of June 2018—a remarkable leap from having no public-funded supercomputer entries in 2016 and 2017.

Continuation of Trajectory: Attention then shifted towards sustaining and expanding this technological frontier. The National Supercomputing Innovation Strategy (2021-2030), finalized in 2021, outlined the sequential development of the economy's sixth and seventh-generation supercomputers, aiming for them to rank among the top five globally. The project proposal for the sixth-generation supercomputer, developed under this new strategy, was approved in 2022, ensuring substantial government funding.

4.5.3. HPC Policy Documents

The key HPC policy documents are listed below for reference.

Legislative Acts:

- National Supercomputing Promotion Act of 2011: English translation by the Korean Law Translation Center is titled "*Act on Utilization and Fostering of National Super-Computers*"
- Enforcement Ordinance of the National Supercomputing Promotion Act: English translation by the Korean Law Translation Center is titled "*Enforcement Decree of the Act on Utilization and Fostering of National Super-Computers*"

Strategy:

- National Supercomputing Innovation Strategy (2021-2030) [in Korean]

Master Plan:

- Third National Supercomputing Master Plan (2023-2027) [in Korean]

4.6. Case Study: Japan

Japan's supercomputers have consistently achieved top rankings in the Top500 list. The Numerical Wind Tunnel (NWT) secured the number one spot in November 1993 and maintained it from November 1994 through December 1995 following an upgrade. The Computational Physics by Parallel Array Computer System (CP-PACS) was the leading supercomputer in November 1996. The Earth Simulator (ES) dominated from June 2002 to June 2004. The K Computer topped the list in both June and November 2011. Most recently, the supercomputer Fugaku held the number one position from June 2020 through November 2021.

In contrast to economies like the United States and Korea, which have specific legislation supporting high performance computing (HPC), Japan does not have comparable statutory provisions. This distinction raises important questions about the dynamics enabling effective public policy and substantial government funding in Japan, evident in the development and support of its latest supercomputer, Fugaku. This case study aims to delve into these underlying factors.

4.6.1. HPC Public Policy

The development of the Fugaku supercomputer illustrates a meticulously orchestrated combination of public policy formulation, strategic planning and collaborative innovation. This serves as a testament to the synergy between Japan's robust public institutions and its profound societal intellectual capacity.

Public Policy Formulation

Whole-of-Government Science and Technology Policy Guidance**:** Under its parliamentary governance system, Japan manages its science and technology strategy through the Council for Science, Technology and Innovation (CSTI), which is an integral part of the Cabinet Office. CSTI orchestrates overarching policy direction and has marked HPC as a strategic priority. This priority is clearly laid out in the Third Science and Technology Basic Plan (FY2006–FY2010), where next-generation supercomputing technology is identified as crucial for the economy's advancement. The Fourth Science and Technology Basic Plan (FY2011-FY2015) further emphasizes HPC, affirming the government's commitment to advancing Japan's technological capabilities in this critical area.

Policy Measure Design

Academia-Led Strategic Planning**:** The foundation for the Fugaku supercomputer was laid in August 2010 with the initiation of the workshop on Strategic Direction/Development of High Performance Computers (SDHPC). Spearheaded by Yutaka Ishikawa and involving top academic institutions like the University of Tokyo, University of Tsukuba, Tokyo Institute of Technology and Kyoto University, this initiative epitomized the collaborative spirit of Japan's academic sector. The workshop was not just a technical endeavor but also a cultural reflection, resonating with Japan's deep-rooted Confucian values that emphasize collaboration for societal good. The primary objectives were to foster academic collaboration, integrate diverse individual viewpoints, define system requirements and explore technological innovations for future supercomputing needs.

Ministry Coordinated Strategic Planning**:** In April 2011, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) established the Working Group on the Study of Future HPC Technology R&D under the Council for HPCI Plan and Promotion. This group was pivotal in shaping the future of Japan's supercomputing initiatives. It comprised two specialized subsidiary groups: the Application Working Group, focused on potential applications of supercomputing technologies, and the Computer Architecture/Compiler/System Software Working Group, which evolved from the earlier SDHPC workshop. These groups worked diligently to align technological research with practical applications, culminating in the submission of the *Report on Future HPCI Technology Development*, the *HPCI Technology Roadmap White Paper* and the *Computational Science Roadmap White Paper* in March 2012.

Ministry Commissioned Feasibilities Studies**:** Between April 2012 and March 2014, the MEXT spearheaded detailed feasibility studies on advanced HPC. These studies, conducted by four specialized teams, analyzed societal and scientific demands to develop an R&D roadmap for applications targeted for 2020, explored the potential of next-generation 'general-purpose' supercomputers with many-core architecture, assessed the viability of exascale heterogeneous systems with accelerators, and investigated multi-vector core architecture with enhanced memory bandwidth. The cumulative insights from these studies were encapsulated in the *Computational Science Roadmap* and the *Report of the System Study Working Group on the Next Flagship System*, providing pivotal guidance for the development of Japan's supercomputer Fugaku.

Policy Measure Approval

CSTI Facilitated Budget Approval: The CSTI played a crucial role[5] in securing budget approval for the Flagship 2020 Project. This significant R&D project, managed by the MEXT, aimed to develop the Fugaku supercomputer as the successor to the K computer. Following the comprehensive evaluation of CSTI in fiscal year (FY) 2013, funding was obtained, and the project commenced in FY2014.

Policy Measure Implementation

Public-Private Partnership for Integrated Co-Design and Development: The development of the supercomputer Fugaku under the Flagship 2020 Project exemplified a strategic integration of hardware, software and application co-design, facilitated through a public-private partnership. This

[5] *Source: White Paper on Science and Technology 2014, Part 2: Measures Implemented to Promote Science and Technology, Chapter 1: Development of Science and Technology Policy*

approach not only streamlined the development process but also leveraged the collective expertise and intellectual capacity of multiple stakeholders, embodying the Skokunin kishitsu (“craftsman's spirit”). The result was the creation of Fugaku, a supercomputer with robust general-purpose capabilities, designed to support a wide array of research and engineering applications and accessible to a diverse user base.

4.6.2. HPC Policy Documents

For reference, the key documents shaping HPC public policy and evaluations are listed below:

Basic Plans – Sources for HPC Public Policy:

- Third Science and Technology Basic Plan (2006-2010) [retrieved from archive.org]
- Fourth Science and Technology Basic Plan (2011-2015) [in Japanese, retrieved from archive.org]

White Papers – Sources for CSTI Evaluation Results of the Flagship 2020 Project:

- *White Paper on Science and Technology 2014, Part 2: Measures Implemented to Promote Science and Technology, Chapter 1: Development of Science and Technology Policy*
- *White Paper on Science and Technology 2015, Part 2: Measures Implemented to Promote Science and Technology, Chapter 1: Development of Science Technology Policy*
- *White Paper on Science and Technology 2019, Part 2: Measures Implemented to Promote Science and Technology, Chapter 1: Development of Science Technology*

***

Chapter 5. Community-Driven Agenda for HPC

This chapter aims to propose an outline for a community-driven agenda that acts as an informal yet coordinated platform for collective collaboration and cooperation. It distinguishes collective collaboration from cooperation, outlines potential community-driven actions, explores the collaboration and cooperation areas identified during the workshop, and highlights the roles and contributions of user communities within the HPC ecosystem.

5.1. Distinguishing Collective Collaboration from Cooperation

Collective collaboration is defined by interdependency and a shared purpose that aligns with the strategic goals of all parties, with effective relationship management and mutual benefits being essential for sustaining the partnership.

Cooperation, on the other hand, involves independent contributions that are aligned towards a shared goal but without the same level of direct interaction or shared creative process. It allows each party to work independently while aiming for a common outcome.

Box 9. Two Examples Illustrating the Definition of HPC Collaboration

For collaboration to succeed, a clearly defined common purpose that aligns with the strategic goals of all parties is essential. Effective relationship management is also critical, ensuring smooth interactions and the swift resolution of conflicts. Moreover, it is imperative that each collaborator recognizes mutual benefits; this shared perception of value is crucial for sustaining the partnership and driving collective success in the cross-border arena.

An illustrative example of collaboration occurs when economies with advanced HPC capabilities share their resources with those with emerging HPC capabilities. For the latter, the benefits include accelerated research timelines and access to cutting-edge computational power. In return, the former gain opportunities for cost recovery and new research partnerships. These collaborations also generate compelling success stories, which can be leveraged to secure future funding opportunities.

Another example is when economies with advanced HPC capabilities donate decommissioned but still powerful HPC systems to educational and research institutions in economies with emerging HPC capabilities. Through a rigorous selection process, recipient institutions are matched based on their research needs and capacity-building goals. Once selected, experts from the donating economies assist with the transportation, installation and setup of the HPC systems. The initiative includes a comprehensive training program to maximize the use of the donated systems, enhancing the recipients' research capabilities and fostering their engagement in the global scientific community.

The benefits of such donations are multifaceted. Recipient institutions gain access to advanced computational resources that allow for complex simulations and large data processing, opening up new research avenues. Donor economy benefit from efficient resource utilization, fostering goodwill and strengthening diplomatic ties, while also creating opportunities for collaborative research. Success stories from the initiative serve as compelling evidence to secure further funding, demonstrating the tangible outcomes of enhanced educational and research activities, and supporting the program’s expansion.

5.2. Outline of Potential Community-Driven Actions

These potential actions illustrate how the HPC community could leverage both collaboration and cooperation to achieve shared objectives:

Education and Workforce Development: The community, through research institutions, universities and training programs, cultivates a skilled HPC workforce, ensuring a steady supply of professionals capable of managing HPC infrastructure, developing applications and leveraging HPC for various purposes.

Standard-Setting and Best Practices: The community collaboratively develops industry standards, guidelines and benchmarks for areas like HPC system performance, software development, data management, cybersecurity and energy efficiency. These standards ensure interoperability and efficiency across different HPC systems, benefiting both infrastructure providers and end-users.

Knowledge Exchange and Collaborative Research: Through knowledge-sharing forums and joint research efforts, the community drives innovation, strengthens partnerships and accelerates advancements in the HPC field. This collaboration fosters a culture of learning and co-creation.

HPC Infrastructure Integration: The community plays a crucial role in coordinating resources, ensuring inclusive access to HPC resources, enhancing resilience during service disruptions and optimizing HPC facilities' ability to meet varying demands.

Open-Access Data, Computational Tools and Knowledge Repositories: By creating and maintaining open-access repositories, the HPC community ensures that digitalized data, tools and knowledge are accessible to everyone, supporting equitable access and encouraging further research and development.

Advocacy and Policy Influence: Through advocacy efforts, members of the HPC community promote expert insights and contribute to the knowledge base that informs policy documents, such as HPC policy papers, strategies, roadmaps and implementation plans. Their advocacy helps to shape policy directions, ensuring that HPC developments align with broader societal and economic needs.

5.3. Workshop-Identified Collaboration and Cooperation Areas

Areas suggested during the accompanying workshop for the development of this white paper include collaboration in quantum computing and cooperation in talent development, data sharing, transfer, connectivity, security and management, as well as environmentally sustainable computing.

5.3.1. Quantum Computing

Quantum computing leverages quantum mechanical phenomena—superposition (existing in multiple states simultaneously), entanglement (interconnected states) and interference (manipulating probabilities)—to perform calculations that are beyond the capabilities of classical computers. It is an emerging computing technology that offers new computational approaches, which complement and synergize with classical von Neumann HPC systems.

The quantum computing landscape is shaped by a diverse array of competing models and hardware technologies, each with unique approaches, and distinct advantages and challenges in harnessing quantum mechanics for computation. Key models include adiabatic quantum computing, quantum annealing and topological quantum computing, while primary hardware technologies encompass superconducting qubits, ion traps and photonics.

- Adiabatic quantum computing involves gradually evolving a quantum system from a simple initial state to a more complex one that represents the solution to a problem. The system remains in its lowest energy state (ground state) throughout this process, requiring slow, careful adjustments to avoid transitions to higher energy states.
- Quantum annealing is particularly effective for solving optimization problems. It uses quantum mechanics to find the lowest energy configuration of a system, which often corresponds to the optimal solution for certain computational challenges.
- Topological quantum computing seeks to encode qubits in topological states of matter, making them more resistant to errors. This is achieved through the braiding of particles known as anyons, with changes in the braiding patterns altering the state of the qubits.
- Superconducting qubits are built using superconducting circuits that conduct electricity without resistance at extremely low temperatures, enabling fast gate operations and making them a popular choice for building scalable quantum systems.
- Ion trap quantum computing uses charged atoms (ions) held in place by electromagnetic fields. These ions are manipulated with lasers to encode and process qubits, offering high precision and long coherence times.
- Photonic quantum computing uses particles of light (photons) as qubits, manipulating them with beam splitters, phase shifters and interferometers. This approach allows for quantum operations at room temperature and is well-suited for integration with optical communication networks.

This diversity is reminiscent of the early days of computer networking, where multiple protocols vied for dominance until Transmission Control Protocol/Internet Protocol (TCP/IP) emerged as the standard. Such historical parallels suggest that the path to a dominant quantum computing standard is still unfolding. Despite the enthusiasm surrounding quantum computing, HPC facility operators recommend a degree of caution, given its nascent stage and the unproven practical applicability of many quantum solutions.

Therefore, the field of quantum computing is currently in the pre-competitive stage, focused on fundamental research and the development of foundational technologies, making it well-suited for collective collaboration. Collaborative efforts in this phase can include:

- Joint research projects through consortia that bring together HPC facility operators, universities, government labs and companies to advance quantum algorithms and technologies.
- Shared quantum computing facilities that provide researchers from various organizations with access to quantum computing resources.
- Interoperability and scheduling protocol development to ensure seamless integration and efficient coordination between quantum and classical computing systems.

5.3.2. Talent Development

Collective cooperation in HPC is pivotal for addressing shared challenges in capacity building and talent circulation. The field of HPC is critically reliant on fostering a robust pipeline of skilled professionals to meet growing demands, particularly in the context of the artificial intelligence (AI) boom. Innovations in human capacity building are necessary not only to sustain the current levels of technological advancement but also to propel them forward. Programs focused on future planning and capacity building are crucial.

To bridge these gaps, initiatives such as the ASEAN HPC School have emerged, supported by partners like the EU; Japan; and Korea, showcasing a successful model of regional talent development. Furthermore, creating professional organizations that offer accreditation and certification—similar to successful models like Cisco Academy and Nvidia Academy—could standardize and elevate professionalism within the HPC community. Increasing mobility through scholarships and professional opportunities are also essential, promoting inward mobility and ensuring the HPC sector remains vibrant and globally connected.

5.3.3. Data Sharing, Transfer, Connectivity, Security and Management

Data plays a pivotal role in the HPC ecosystem, particularly with the increasing reliance on machine learning and AI. The integration of data into HPC systems emphasizes the need for making data immediately analyzable alongside real-time CPU processing. Initiatives like the BioMirror project for bioinformatics and the collaboration between the San Diego Supercomputing Center and Rutgers University, which established a Protein Databank mirror site in Singapore, illustrate the global efforts to manage and share critical research data.

Managing large-scale data across sovereign borders presents significant challenges, including the need for high-speed data transfer capabilities and cost-effective storage solutions. Regional data hubs have become crucial in addressing these challenges, especially in data-sparse regions like South Africa and Asia. For instance, the Data Mover Challenge (DMC) organized at the SuperComputing Asia Conference highlights efforts to improve data transfer protocols, exploring technologies like RDMA[6] over Converged Ethernet (RoCE v2) to facilitate faster and more reliable data exchanges.

Data security remains a top priority, with ongoing debates on the necessity and methods of encrypting large data transfers. Advances in technologies like quantum key distribution (QKD) and post-quantum cryptography (PQC) are crucial, especially as economies like China; Japan; and Korea develop their QKD network infrastructures. With the recent approval of four classes of PQC algorithms by the National Institute of Standards and Technology (NIST), the HPC community is actively exploring their implementation to strengthen the security of data exchanges.

Effective data management extends beyond mere storage and transfer to encompass issues like data citation, licensing and the establishment of provenance chains. The concept of data librarianship is gaining traction, transforming data handling into a more structured and recognized discipline within HPC. This shift is crucial as the demand for rigorous data management practices increases, driven by scientific journals and research funders who require detailed data stewardship plans.

5.3.4. Environmentally Sustainable Computing

The imperative for environmentally sustainability computing has never been more critical, particularly in economy like Singapore where the introduction of a carbon tax and climate-related financial disclosures pose significant challenges. Data centers, especially in hot and humid climates, have a considerable responsibility to mitigate their environmental impact.

To address these challenges, it is essential to focus on enhancing data center technologies that improve Power Usage Effectiveness (PUE) and optimize water usage. Implementing solutions such as advanced cooling systems, energy-efficient hardware and renewable energy sources can significantly reduce the carbon footprint of these facilities. Recommendations include adopting best practices for environmentally sustainable operations, and continually innovating in response to evolving environmental standards and expectations.

5.4. Roles and Contributions of User Communities in HPC

User communities act as a central platform for collaborative and cooperative efforts that are of interest or relevance to HPC users. These efforts include:

[6] RDMA (Remote Direct Memory Access)

- Standard-Setting and Interoperability: User communities collaborate to establish standards that ensure interoperability between diverse HPC systems and software. This includes creating technical guidelines that enable different systems to work seamlessly together, facilitating data sharing and integration across platforms.
- Community Code Development: User communities often co-develop open-source codes and software tools that address common challenges in HPC, fostering a spirit of shared innovation and resource pooling.
- Training and Skills Development: User communities organize training programs and skill-building workshops to up-skill HPC users, ensuring that both newcomers and experienced professionals stay up-to-date with the latest technologies and best practices.
- Knowledge Sharing: These communities provide a space for exchanging knowledge about HPC tools, software, best practices and emerging developments. This sharing of expertise helps keep users informed about cutting-edge advancements and promotes the adoption of new methods.
- Peer-to-Peer Support: User communities facilitate peer-to-peer support, where members can seek advice, share troubleshooting tips and resolve technical challenges together, fostering a culture of mutual assistance.
- Networking Hub: User communities serve as a networking hub, enabling potential partners to connect, discuss collaborative opportunities and pool resources for large-scale research projects that benefit all participants.
- Policy Input: User communities can gather feedback from their members and present a unified perspective to policymakers, influencing decisions that affect the growth, funding and accessibility of HPC resources.

While the value of flourishing user communities is evident, the challenge of effectively building and sustaining these communities lies in the careful management of their interaction dynamics. Successful communities must not only attract a large number of users but also the right kind of users—those who contribute positively and sustain community growth.

A central aspect of this endeavor is to cultivate a ‘pass it on’ culture of reciprocal altruism, where members are motivated to pay forward the aid they have received and to engage in mutual support, rather than focusing on immediate personal gain. Establishing clear obligations and expectations is crucial, as is nurturing social norms and bonds that promote and strengthen cooperation and support among members.

Another critical aspect of research involves recognizing it as a series of repeated searches across numerous haystacks—the ‘know-where’—in the hope of discovering the elusive needle. Equally important is possessing the ‘know-how’—the right methodologies to effectively locate the needle. Given the overwhelming number of haystacks encountered in complex problems, it is advantageous to foster a collaborative environment. By aiding one another, members can expedite the process of mapping out their individual R&D efforts, identifying which haystacks are likely dead-ends and which offer promising leads. Additionally, cultivating a shared toolkit of methodologies enhances the community's ability to efficiently navigate through various haystacks. This collaborative approach not only accelerates individual projects but also enriches the collective knowledge base, driving innovation and success in complex research endeavors.

***

Chapter 6. Recommendations and Conclusion

6.1.Recommendations

This white paper intends to illustrate the key aspects of an ecosystem model that extends beyond HPC systems and their management, focusing on factors that significantly influence the utility and effectiveness of HPC facilities. Based on this model, the following recommendations are provided to enhance the utility and effectiveness of HPC facilities for the realization of their benefits:

HPC Facility Setup, Management and Operation

1. For responsibilities involving *strategic considerations*, establish and document the decision rationale, including constraints, choices and priorities, and update them as needed.
2. Develop and maintain *policies and operating procedures* for managing and operating the HPC infrastructure, ensuring they align with strategic decisions.
3. Utilize *software tools* to effectively implement policies and operating procedures.

Public Policy for HPC

4. Raise awareness of the *critical role of HPC in domestic strategies* for Industry 4.0, digital transformation, smart cities and addressing societal challenges to draw attention to potential policy gaps.
5. Recognize the synergetic relationship between artificial intelligence *(AI), big data and HPC*, and as a result, coordinate the development of shared infrastructure to support all three.
6. Invest in skills development to cultivate a *skilled HPC workforce* and build the intellectual capital necessary for effective HPC utilization.
7. Support innovation by providing direct and indirect *financial assistance to businesses* leveraging HPC for R&D, engineering, logistical optimization and strategic decision-making to enhance their competitiveness.
8. Establish and update *norms and regulations* to ensure interoperability, privacy protection, cybersecurity and compliance with domestic security and export controls, while addressing the operational needs of HPC facilities, such as energy supply, water supply and high-speed internet connectivity.
9. Invest in *HPC facilities* and establish a sustained financing mechanism to ensure consistent support.

Community-Driven Agenda for HPC

10. Design and implement initiatives to address the needs of the HPC community, focusing on education and training, standard-setting, collaborative research, HPC infrastructure integration, knowledge exchange, sharing of research data and computational tools, and policy influence.

6.2.Conclusion

In summary, this white paper explores the multifaceted nature of HPC and its role in driving socio-economic impact.

- It provides an overview of the HPCI-MEM model, emphasizing the dynamic relationships between stakeholders and their influence on the utility and effectiveness of HPC facilities.
- The technical, human and financial challenges in HPC facility setup, management and operation are thoroughly examined, highlighting the need for expertise and long-term financing strategies.
- The paper also underscores the critical role of HPC in domestic strategies such as Industry 4.0 and digital transformation, emphasizing the importance of holistic public policy and sustained investment.

- Additionally, the community-driven agenda is outlined, stressing the value of collaboration and cooperation.

The concluding chapter presents tailored recommendations and emphasizes the interdependence of key elements within the HPC ecosystem, advocating for a balanced approach that addresses all facets to fully leverage the potential of HPC.

***

Appendix A: Evolution of Computing Systems

This supplementary material traces the evolution of computing systems from basic calculators to modern supercomputers.

Calculators

Initially, the term ‘computing’ referred to the process of humans performing calculations using mechanical desk calculators. This method evolved through several technological stages: first, to vacuum tube-based calculators, then to models powered by transistors. These were followed by calculators that utilized integrated circuits and ultimately, the technology advanced to semiconductor chips, with computations performed by microprocessors.

The development of the electromechanical automatic calculator, MARK I (IBM[7] Automatic Sequence Controlled Calculator), served as a precursor to modern electronic computers. This device was developed as part of the United States' military efforts and initially deployed in 1944 to expedite the calculations required for ballistic projections during World War II.

Early Electronic Computers

Following this, the ENIAC (Electronic Numerical Integrator and Computer), a fully electronic general-purpose computer based on vacuum tubes, was developed and completed in 1945 for military applications. This was succeeded by the UNIVAC (Universal Automatic Computer), the first commercially available computer, also based on vacuum tubes, which was introduced in 1951.

The industry then witnessed a significant technological shift from vacuum tubes to transistors. IBM commercialized the first fully transistor-based computer, the IBM 7090, with the first system delivered in 1959, which marked a new era in scientific computing marketplace.

Subsequently, the transition from transistor-based to integrated circuit-based computers led to a segmented commercial market. In 1964, Control Data Corporation (CDC) released the CDC 6600, the world's fastest computer at the time, targeting the supercomputing sector. That same year, IBM launched the System/360, a versatile mainframe capable of handling a broad spectrum of enterprise applications. Following closely in 1965, Digital Equipment Corporation (DEC) introduced the PDP-8, positioning it as a cost-effective minicomputer for smaller businesses and educational institutions.

Modern Personal Computers

The transition from integrated circuit-based computing systems to architectures centered on discrete semiconductor components, such as microprocessors and memory chips, was marked by two significant introductions by Intel. The Intel 1103 1K DRAM (Dynamic Random Access Memory), introduced in 1970 and the Intel 8080 microprocessor, released in 1974, catalyzed this shift. These developments played a pivotal role in rapidly expanding the personal computer (PC) market, albeit in a fragmented manner.

However, a more consequential milestone occurred in 1981 with the launch of the IBM PC Model 5150, which utilized the Intel 8088 microprocessor featuring the x86 architecture. The strategic decision of IBM to employ an open architecture with modular discrete components and to license its operating system non-exclusively led to the unforeseen creation of a global mass market for IBM-compatible PCs.

[7] IBM (International Business Machines Corporation)

The mass market for PCs enabled economies of scale for commoditized computer components, which significantly lowered PC prices. This, coupled with intense market competition, necessitated accelerated performance improvements in each successive generation to drive repeated sales. These advancements reduced the financial burden on higher education institutions for acquiring computing systems, thereby transforming research by introducing the third paradigm of scientific discovery—computational science. Furthermore, the continuous generational enhancements in computer performance have expanded their application range and capabilities.

Modern Supercomputers

Meanwhile, supercomputers designed for high computational performance have continued to evolve, enabling the resolution of yesterday's unsolvable problems. Cray Research, a pioneer in the field, led the industry from the mid-1970s to the early 1990s. This company introduced several notable models, including the Cray-1 in 1975, Cray-2 in 1985, Cray C90 in 1991 and Cray T90 in 1995.

However, by the late 1980s and early 1990s, the supercomputer industry encountered significant challenges due to the advanced capabilities of commoditized microprocessors, including x86 architectures. Additionally, the advent of commodity microprocessors featuring Reduced Instruction Set Computer (RISC) architecture catered to supercomputing needs by effectively balancing performance with energy efficiency.

Furthermore, developments in high-speed interconnects, the establishment of the initial Message Passing Interface (MPI) standard in 1994 and the emergence of an open-source software stack have facilitated the assembly of low-cost Beowulf clusters from commodity computer parts. Subsequently, tutorials for building these clusters became available starting in 1995. Additionally, the introduction of the initial Open Multi-Processing (OpenMP) standard in 1997 further enhanced parallel computing capabilities.

Despite these advancements, the development and installation of supercomputers designed for HPC have continued. Notable examples include:

- The IBM Blue Gene series, which emerged in the 2000s.
- Japan's custom-built Fugaku supercomputer, which began installation at the RIKEN Center for Computational Science in December 2019.
- In the United States, the custom-built Frontier supercomputer was delivered to Oak Ridge National Laboratory (ORNL) in 2021. It was followed by the Aurora 21 supercomputer at Argonne National Laboratory (ANL) in June 2023 and the El Capitan supercomputer at Lawrence Livermore National Laboratory (LLNL), with installation starting in 2023.

These developments have led to the creation of two distinct market segments: one focused on systems built from commercial-off-the-shelf (COTS) components and the other dedicated to the development of supercomputers for HPC. This division influences public policy decisions regarding procurement strategies. These strategies include the direct acquisition of COTS-based systems, the procurement of R&D services for HPC hardware and software, and public-private partnerships that concentrate on the integrated design of hardware components, software layers and applications.

***

Appendix B: Exemplars of Supercomputers

This supplementary material provides exemplars of supercomputers. Supercomputers have undergone significant evolutionary milestones, marked by substantial increases in processing power, advancing from gigaflops (billion of floating-point operations per second) to teraflops (trillion of floating-point operations per second), then to petaflops (quadrillion of floating-point operations per second) and now reaching exaflops (quintillion of floating-point operations per second).

Gigaflops**:** Launched in 1985 by Cray Research, the Cray-2 supercomputer was a commercial product, with a total of 27 units sold. It achieved an initial performance of 1.9 gigaflops, becoming the first to break the gigaflop barrier. The Cray-2 featured an innovative liquid immersion cooling system to efficiently manage heat from its densely packed integrated circuits. Additionally, its unique memory architecture enabled rapid access times, significantly enhancing its performance capabilities.

Teraflops: Developed under the United States' Accelerated Strategic Computing Initiative (ASCI) for simulating nuclear weapon testing, ASCI Red supercomputer built by Intel Corporation was installed at Sandia National Laboratories in late 1996. It was the first supercomputer to exceed one teraflop on the LINPACK[8] benchmark. Powered by thousands of Intel microprocessors, ASCI Red featured a massively parallel processing architecture that became foundational to the concept of parallelism in HPC, allowing for multiple calculations or processes to be performed simultaneously.

Petaflops: The Roadrunner supercomputing system, which encompassed both the supercomputer hardware and its applications, was developed through a collaboration between IBM and Los Alamos National Laboratory for simulating nuclear weapon testing. Completed in 2008, Roadrunner achieved an initial performance of 1.026 petaflops on the LINPACK benchmark, becoming the first supercomputer to reach this performance milestone. It featured a pioneering hybrid design, utilizing two different processor architectures, which laid the groundwork for what would become known as the concept of heterogeneous computing—a system that integrates more than one type of processor or core. Additionally, this collaboration led to a novel methodology in integrated hardware-software-applications design, where hardware components, software layers and applications are developed concurrently.

Exaflops: As supercomputers increasingly support machine learning and artificial intelligence applications, which often require only single or further-reduced precision in floating-point calculations, performance metrics have evolved. Distinctions are now made between the classical double precision used in the LINPACK benchmark and the newer mixed precision of floating-point operations used in the HPL-AI[9] benchmark.

Mixed Precision Exaflops: Developed as part of Japan's Flagship 2020 project through a collaboration between Fujitsu and RIKEN, the Fugaku supercomputer was completed in 2020. It became the first supercomputer to surpass exaflops performance on the HPL-AI benchmark. The supercomputer featured broad-based capacity and applicability. Its capacity stemmed from an architecture that employed a single type of custom-designed processor for supercomputing, thereby eliminating the communication overhead, latency and bandwidth constraints typically found in mixed-processor environments such as CPU-GPU configurations. Its applicability arose from an integrated design approach, where hardware components, software layers and applications were developed concurrently.

Double Precision Exaflops: Developed as part of the United States' Exascale Computing Initiative (ECI), the Frontier supercomputer at Oak Ridge National Laboratory (ORNL) became operational in 2022. With a performance of 1.102 exaflops on the LINPACK benchmark, it achieved the milestone of becoming the first exascale supercomputer. Frontier supercomputer featured a unique accelerated

[8] LINPACK: Originally used as abbreviation for linear equation software package

[9] HPL-AI (High Performance LINPACK for Accelerator Introspection)

computing node architecture that overcame the prevailing challenges of communication, memory and energy constraints, enabling unprecedented scaling in computational performance. It also employed an integrated design approach, where hardware components, software layers and applications were developed concurrently, enabling is broad applicability.

***

References

Chapter 1

A. Norton and E. Joseph, “HPC Investments bring High Returns”, Hyperion Research, Sponsored by Dell Technologies and Intel Corporation, July 2020. Retrieved from https://www.delltechnologies.com/asset/en-us/products/ready-solutions/industry-market/hyperion-hpc-investment-brings-high-returns.pdf

The US Department of Energy, “IDC DOE ROI Research Update”, December 2016. Retrieved from https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/201612/IDC_DOE_ROI_Research_Update_12-16-2016.pdf

Chapter 3

European Exascale Software Initiative, “Final Report on Operational Software Maturity Level Technology”, May 2015. Retrieved from http://www.eesi-project.eu/wp-content/uploads/2015/05/EESI2_D6.1R_Final-report-on-operational-software-matutity-level-technology.pdf

International Organization for Standardization, “ISO/IEC 27002:2022 Information Security, Cybersecurity, and Privacy Protection – Information Security Controls”, ISO, 2022.

S. Lathrop, C. Mendes, J. Enos, B. Bode, G. Bauer, R. Sisneros and W. Kramer, “Best Practices for Management and Operation of Large HPC Installations”, Concurrency and Computation: Practice and Experience, *31*(16), e5069, 2019.

Y. Guo, R. Chandramouli, L. Wofford, R. Gregg, G. Key, A. Clark, C. Hinton, A. Prout, A. Reuther, R. Adamson, A. Warren, P. Bangalore, E. Deumens and C. Farkas, “High-Performance Computing Security: Architecture, Threat Analysis, and Security Posture”, National Institute of Standards and Technology, Gaithersburg, MD, NIST Special Publication (SP) NIST SP 800-223, 2024. https://doi.org/10.6028/NIST.SP.800-223

***

Glossary of Abbreviations and Acronyms

AI	Artificial Intelligence	
AMIS	Asset Management Information System	
ANL	Argonne National Laboratory	(the US)
APEC	Asia-Pacific Economic Cooperation	
ASCI	Accelerated Strategic Computing Initiative	(the US)
BYOL	Bring Your Own License	
CAPEX	Capital Expenditure	
CDC	Control Data Corporation	
CI/CD	Continuous Integration / Continuous Deployment	
COTS	Commercial-Off-The-Shelf	
COVID-19	Coronavirus Disease of 2019	
CP-PACS	Computational Physics by Parallel Array Computer System	(Japan)
CPU	Central Processing Unit	
CSTI	Council for Science, Technology and Innovation	(Japan)
CT	Computed Tomography	
DEC	Digital Equipment Corporation	
DMC	Data Mover Challenge	
DRAM	Dynamic Random Access Memory	
DVFS	Dynamic Voltage and Frequency Scaling	
ECI	Exascale Computing Initiative	(the US)
EESI	European Exascale Software Initiative	(the EU)
ENIAC	Electronic Numerical Integrator And Computer	(the US)
EOP	Executive Office of the President	(the US)
ES	Earth Simulator	(Japan)
EU	European Union	
FAQ	Frequently Asked Questions	
FCCSET	Federal Coordinating Council for Science, Engineering and Technology	(the US)
FY	Fiscal Year	
GPU	Graphics Processing Unit	
HPC	High Performance Computing	
HPCC	High Performance Computing and Communications	(the US)
HPCI	HPC Infrastructure	
HPCI-MEM	High Performance Computing Infrastructure Management Ecosystem Model	
I/O	Input / Output	
IBM	International Business Machines corporation	
ICT	Information and Communication Technology	
IDC	International Data Corporation	
IDS/IPS	Intrusion Detection / Prevention Systems	
IP	Internet Protocol	
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission	
IT	Information Technology	
KISTI	Korea Institute of Science and Technology Information	(Korea)
LDAP	Lightweight Directory Access Protocol	

LLNL	Lawrence Livermore National Laboratory	(the US)
MAC	Media Access Control	
MARK I	IBM Automatic Sequence Controlled Calculator	
MEXT	Ministry of Education, Culture, Sports, Science and Technology	(Japan)
MFA	Multi-Factor Authentication	
MPI	Message Passing Interface	
MRI	Magnetic Resonance Imaging	
NCSA	National Center for Supercomputing Applications	(the US)
NIST	National Institute of Standards and Technology	(the US)
NITRD	Networking and Information Technology Research and Development	(the US)
NSTDA	National Science and Technology Development Agency	(Thailand)
NWT	Numerical Wind Tunnel	(Japan)
OpenMP	Open Multi-Processing	
OPEX	Operational Expense	
ORNL	Oak Ridge National Laboratory	(the US)
OSTP	Office of Science and Technology Policy	(the US)
PAM	Privileged Access Management	
PC	Personal Computer	
PMIS	Procurement Management Information System	
PPSTI	Policy Partnership on Science, Technology and Innovation	
PQC	Post-Quantum Cryptography	
PUE	Power Usage Effectiveness	
QKD	Quantum Key Distribution	
R&D	Research and Development	
RAID	Redundant Array of Independent Disks	
RBAC	Role-Based Access Control	
RDMA	Remote Direct Memory Access	
RFP	Request For Proposal	
RIKEN	The Institute of Physical and Chemical Research	(Japan)
RISC	Reduced Instruction Set Computer	
RoCE v2	RDMA over Converged Ethernet version 2	
ROI	Return On Investment	
ROR	Return On Research	
S&T	Science and Technology	
SDHPC	Strategic Direction / Development of High Performance Computers	(Japan)
SIEM	Security Information and Event Management	
SLAs	Service Level Agreements	
SMEs	Small and Mid-Size Enterprises	
SSH	Secure Shell	
TCP/IP	Transmission Control Protocol / Internet Protocol	
ThaiSC	NSTDA Supercomputer Center	(Thailand)
TRL	Technology Readiness Level	
UNIVAC	Universal Automatic Computer	
UPS	Uninterruptible Power Supply	

***