Top 10 Predictions for the Global HPC-AI Community for 2025

March 2025

www.HyperionResearch.com
www.hpcuserforum.com

Earl Joseph, Bob Sorensen, Mark Nossokoff, Jaclyn Ludema, and Tom Sorensen

About Hyperion Research

(www.HyperionResearch.com & www.HPCUserForum.com)

Hyperion Research mission:

- *Hyperion Research helps organizations make effective decisions and seize growth opportunities*
  - *By providing research and recommendations in high performance computing and emerging technology areas*

HPC User Forum mission:

- *To improve the health of the HPC/AI/QC industry*
  - *Through open discussions, information sharing and initiatives involving HPC users in industry, government and academia along with HPC vendors and other interested parties*

© Hyperion Research 2025

The Hyperion Research Team

Analysts

- Earl Joseph, CEO
- Bob Sorensen, SVP Research
- Mark Nossokoff, Research Director
- Jaclyn Ludema, Analyst
- Thomas Sorensen, Analyst

Executive

- Jean Sorensen, COO

Survey Specialist

- Cary Sudan, Principal Survey Specialist

Global Accounts

- Mike Thorp, Sr. Global Sales Executive
- Kurt Gantrish, Sr. Account Executive

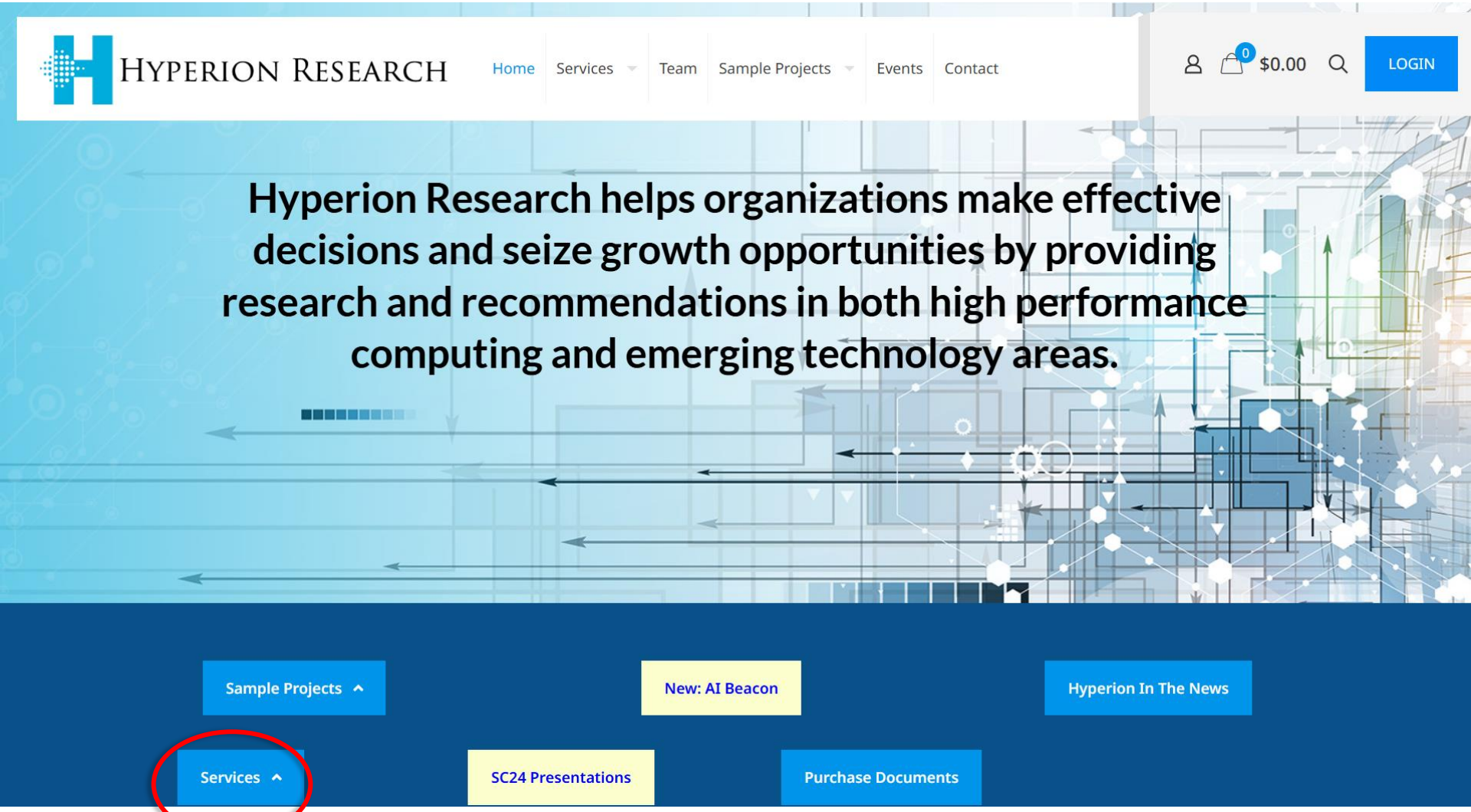
Consultants

- Katsuya Nishi, Japan and Asia
- Kirsten Chapman, KC Associates
- Andrew Rugg, Certus Insights
- Jie Wu, China and Technology Trends
- Mara Jacob, HPC User Forum Support

© Hyperion Research 2025

Visit Our Website: www.HyperionResearch.com

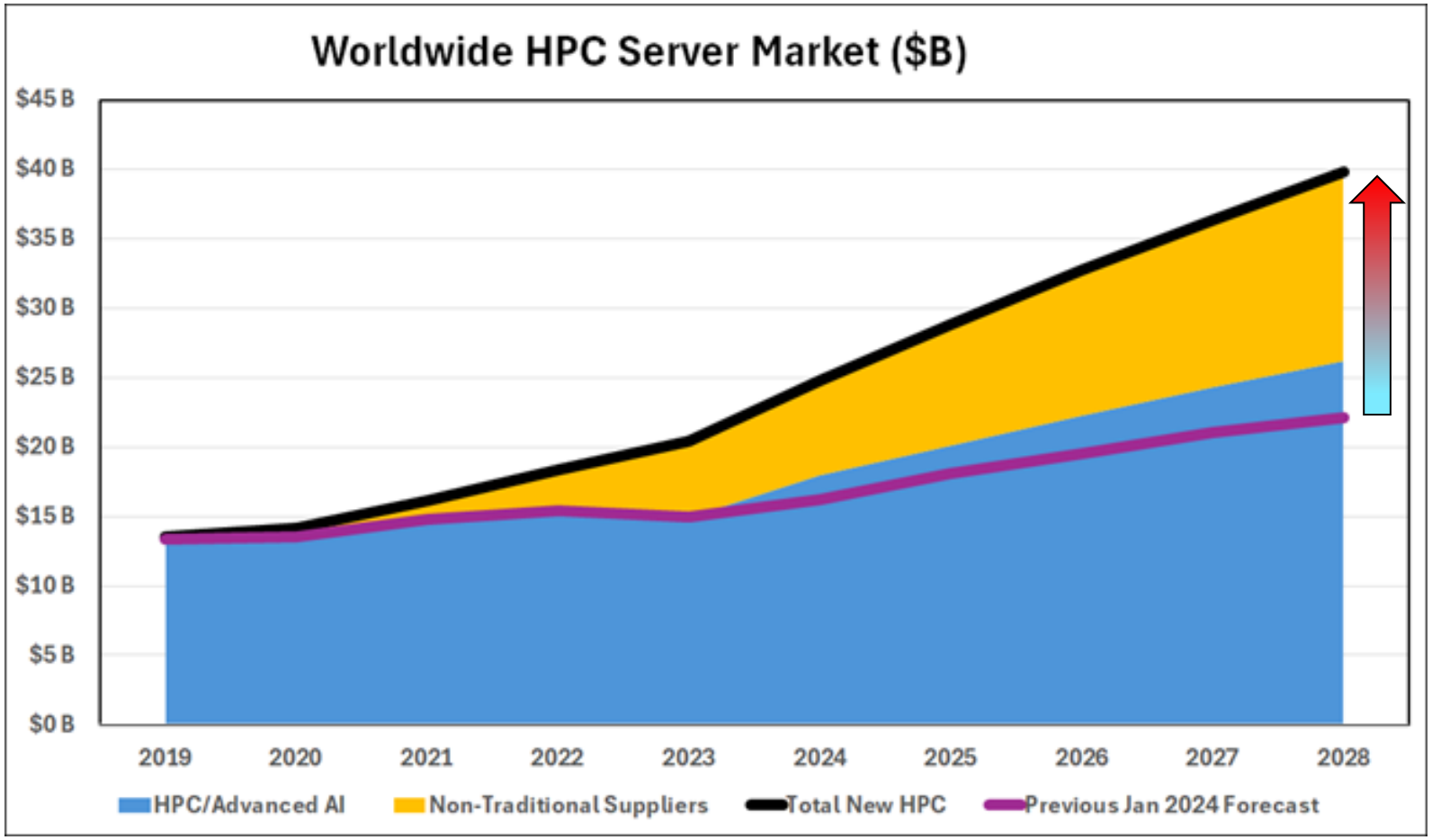
Twitter: @HPC_Hyperion



© Hyperion Research 2025

Updated View of the On-Prem Server Market

- *Hyperion Research just announced a 36.7% increase in the HPC/AI server market size (now growing at 15% CAGR)*
- *Added tracking of non-traditional AI/HPC suppliers*



© Hyperion Research 2025

The Overall HPC/AI Market in 2024

2024 HPC/AI Spending is projected to reach $51.9 billion ($US)

2024 HPC/AI Total Spending

Worldwide Technical Computing/HPC Spending	
	2024
Traditional HPC/AI Suppliers	$17.9
Non-Traditional Suppliers	$7.5
Storage, Software, Service	$17.8
HPC Cloud Spending	$8.7
Total HPC/AI	**$51.9**
Source: Hyperion Research, Oct. 2024	

- **$25.4 billion in on-premises servers**
- **$8.7 billion in spending to run HPC/AI workloads in the cloud**

© Hyperion Research 2025

Hyperion Research's 2025 Predictions

1. There will be a resurgence of the human element within adopting and integrating AI.
2. As efforts to adopt and integrate AI gain traction among industry leaders, new use cases, optimization, regulatory developments, and ROI will become a new focus for users.
3. The rapid rise of compute requirements for large language model training runs will begin to slow with a shift in emphasis on smaller and more efficient models using more focused training data sets.
4. HPC end users, particularly those with major investments in legacy codes built on 64-bit floating-point data formats, will begin to explore in earnest the increasing performance capabilities of mixed and low precision hardware, a growing trend currently driven by the compute demands of the AI-centric processor and accelerator space.
5. Users will more fully embrace the idea of “continuum computing”, incorporating the cloud as a viable tool in conjunction with (or instead of) their on-premises infrastructure as they become more sophisticated identifying the infrastructure required to achieve their desired outcomes.
6. Multiple factors will accelerate users to use CSP resources, including AIaaS and GPUaaS providers, to meet their compute needs. Factors include supply chain sluggishness, shortened hardware release cycles, and sustainability goals.
7. Interest in on-premises quantum computing will increase, with several leading HPC sites announcing on-premises QC acquisitions.
8. High performance computing centers, AI datacenters, and hyperscalers will increasingly create demand for innovation in energy provisioning.
9. Sovereign efforts in multiple areas (e.g., processors, semiconductor manufacturing, LLMs) continue but will not, yet, have material impact on the global supply chain in 2025.
10. The lack of HPC and AI talent will become a greater issue and constraint for all but the largest sites and hyperscalers.

© Hyperion Research 2025

Humanity Strikes Back!

1. *There will be a resurgence of the human element within adopting and integrating AI*

- **New emphasis on the importance of human oversight, collaboration, and ethical decision-making**
- **Humans will play a crucial role in interpreting AI predictions, validating AI results, and providing subject matter expertise**
- **Key players in the AI industry are increasingly favoring "human-in-the-loop" designs**
  - Investment in training programs to upskill their workforce
  - More user-friendly AI tools - complements human skills, creativity, and ethics rather than replacing human input
  - Enhanced reliability and accountability of AI systems
  - Using AI to make humans more productive (vs. replacing them)

© Hyperion Research 2025

AI Maturity Brings New Questions

2. *As efforts to adopt and integrate AI gain traction among industry leaders, new use cases, optimization, regulatory developments, and ROI will become a new focus for users*

- **HPC/AI integrators have come to expect:**
  - Robust return on investment
  - New levels of efficiency
  - Effective regulatory guidelines
- **As AI integrated systems become the norm, the effectiveness and limitations of the technology will become better understood**
- **Aspirant goals will be realized for many users, but some may face costly challenges of unexpected severity such as:**
  - High cost of upkeep
  - Continual education of in-house expertise
  - Management of regulatory demands

© Hyperion Research 2025

LLM Training Needs a Reboot

3. The rapid rise of compute requirements for large language model training runs will begin to slow with a shift in emphasis on smaller and more efficient models using more focused training data sets

- **Current LLM training requirements 10^{26} total training operations**
  - Projections call for an increase of two to three order of magnitude in the next few years (10^{28} to 10^{29})
  - This is out of reach for all but the most aggressive, well-funded organizations: e.g., Anthropic, OpenAI, Telsa, Meta, Google
- **The mainstream HPC world will instead focus on less demanding LLMs or small language model training**
  - Requires less total compute, perhaps three to four orders of magnitude less
  - Based on training data sets that are smaller, more disciplined or subject focused, appropriately curated, and perhaps even proprietary to a targeted end use or end users

© Hyperion Research 2025

Continued Debate on Precision vs. Performance

4. *HPC end users, particularly those with major investments in legacy codes built on 64-bit floating-point data formats, will begin to explore the increasing performance capabilities of mixed and low precision hardware*

- **Many AI applications do not need 64-bit floating-point formats**
  - They often require only 32-bit, 16-bit, 8-bit or even lower floating point or integer schemes
- **GPU designers are increasingly optimizing their chip and core designs to take advantage of this trend**
  - Configuring hardware to offer increased computational performance with lower memory overhead for these mixed and lower precision AI jobs
- **Creating opportunities/concerns for traditional HPC end users**
  - Performance on lower precision is growing when compared with counterpart gains for 64-bit floating point
  - Potentially leaving future processors underpowered for some traditional science and engineering applications or forcing major, if not complex, HPC end user rewrites of existing legacy codes

© Hyperion Research 2025

Mastering the Cloud-On-Prem Continuum

5. *Users will more fully embrace the idea of "continuum computing", incorporating the cloud as a viable tool in conjunction with (or instead of) their on-premises infrastructure*

- **Optimized Resource Allocation**
  - Align infrastructure with workload-specific demands
  - Enable cost-effective and outcome-driven computing strategies
- **Enhanced Efficiency and Agility**
  - Dynamically shift resources between cloud and on-premises
  - User ability to respond rapidly to changing business needs and priorities
- **The ability to add or access new technologies more quickly**
- **Advancing Orchestration Tools**
  - New tools to simplify transitions across hybrid environments
  - Ensure interoperability and minimizes disruption

© Hyperion Research 2025

The Neo-Cloud Rises

6. *Multiple factors will accelerate users to use CSP resources, including AIaaS and GPUaaS providers, to meet their compute needs*

- **Acceleration of Cloud Adoption for AI Workloads**
  - AIaaS and GPUaaS providers ("neo-clouds") offer instant access to state-of-the-art hardware
  - Supply chain delays and frequent hardware refresh cycles drive demand for cloud-based solutions
- **Faster Access to Cutting-Edge Technology**
  - Expensive GPUs with yearly iterations encourage low-commitment cloud adoption
  - Rapid compute access accelerates AI/ML/DL integration/time-to-market
  - Supply chain uncertainty hinders smaller on-premises build-outs
- **Diversification of Application-Specific Hardware**
  - CSPs appeal to organizations in pilot, testing, and pre-production phases
  - Specialized AI data centers focus on refined service models over traditional CSPs (e.g., AWS, Google, Microsoft)
- **Sustainability as a Catalyst for Change**
  - Organizations avoid costly upgrades (e.g., liquid cooling) while reducing their carbon footprint
  - CSPs innovate energy management practices, promoting renewable energy and green architectures

© Hyperion Research 2025

Quantum Computing Gaining On-Prem Traction

7. Interest in on-premises quantum computing will increase, with several leading HPC sites announcing on-premises QC acquisitions

- **A growing number of QC vendors currently offer on-premises options**
  - Including QuEra, IBM, D-Wave, Quantinuum, and IQM, augmenting their cloud-based portal access offerings
  - Some installations are already on the books
  - Most recently, Microsoft/Atom Computing announcement
- **QC end users, particularly those in the HPC space, increasingly will be looking to on-premises QC installations**
  - Help their efforts in HPC/QC integration
  - Support bare metal access for QC software developers
  - Mitigate time of flight delays with cloud-based models
  - Ensure that critical data and applications remain safely protected through internal cybersecurity controls

© Hyperion Research 2025

Advanced Computing Demands Advanced Power

8. *High performance computing centers, AI datacenters, and hyperscalers will increasingly create demand for innovation in energy provisioning*

- **The most advanced computer hardware, especially AI-related, is ushering in a new phase of energy demands**
  - Impacting data centers overall, and individual rack designs
- **New innovations in energy provisioning will be required to meet heightened demand, influencing:**
  - Compute/data center site assessment
  - Greater focus on energy efficiency in hardware design
  - Ongoing efforts to continue to meet sustainability goals
  - Forays into novel or underutilized power sources (i.e., nuclear)

© Hyperion Research 2025

Minimal Market Impact from Sovereign Investments in 2025

9. Sovereign efforts in multiple areas (e.g., processors, semiconductor manufacturing, LLMs) continue but will not, yet, have material impact on the global supply chain in 2025.

- **Conversations and investments continue relative sovereign investments**
- **Drivers**
  - Regional influence and control over technology ownership
  - Government regulation and compliance
- **Multiple technology areas**
  - Processors (e.g., EPI)
  - Semiconductor manufacturing (e.g., Middle East, India)
  - LLMs
- **Economic and market impact will lag investments**
- **Current global political climate will have significant impact**

© Hyperion Research 2025

Dearth of HPC-AI Talent to Continue

10. The lack of HPC and AI talent will become a greater issue and constraint for all but the largest sites and hyperscalers

- **Market dynamics continue to contribute to scarcity of HPC-AI talent**
  - Aging HPC workforce
  - New and specialized HPC and AI hardware increase the complexity of systems and how to use them
  - Modern AI-related software and programming environments
- **Organizations with smaller budgets and modest IT needs may have difficulty competing with those that have larger budgets and a wider variety of challenges and opportunities to offer associates**
- **Increasing sovereignty issue looms large**

© Hyperion Research 2025

In Summary

© Hyperion Research 2025

Hyperion Research's 2025 Predictions

1. There will be a resurgence of the human element within adopting and integrating AI.
2. As efforts to adopt and integrate AI gain traction among industry leaders, new use cases, optimization, regulatory developments, and ROI will become a new focus for users.
3. The rapid rise of compute requirements for large language model training runs will begin to slow with a shift in emphasis on smaller and more efficient models using more focused training data sets.
4. HPC end users, particularly those with major investments in legacy codes built on 64-bit floating-point data formats, will begin to explore in earnest the increasing performance capabilities of mixed and low precision hardware, a growing trend currently driven by the compute demands of the AI-centric processor and accelerator space.
5. Users will more fully embrace the idea of “continuum computing”, incorporating the cloud as a viable tool in conjunction with (or instead of) their on-premises infrastructure as they become more sophisticated identifying the infrastructure required to achieve their desired outcomes.
6. Multiple factors will accelerate users to use CSP resources, including AIaaS and GPUaaS providers, to meet their compute needs. Factors include supply chain sluggishness, shortened hardware release cycles, and sustainability goals.
7. Interest in on-premises quantum computing will increase, with several leading HPC sites announcing on-premises QC acquisitions.
8. High performance computing centers, AI datacenters, and hyperscalers will increasingly create demand for innovation in energy provisioning.
9. Sovereign efforts in multiple areas (e.g., processors, semiconductor manufacturing, LLMs) continue but will not, yet, have material impact on the global supply chain in 2025.
10. The lack of HPC and AI talent will become a greater issue and constraint for all but the largest sites and hyperscalers.

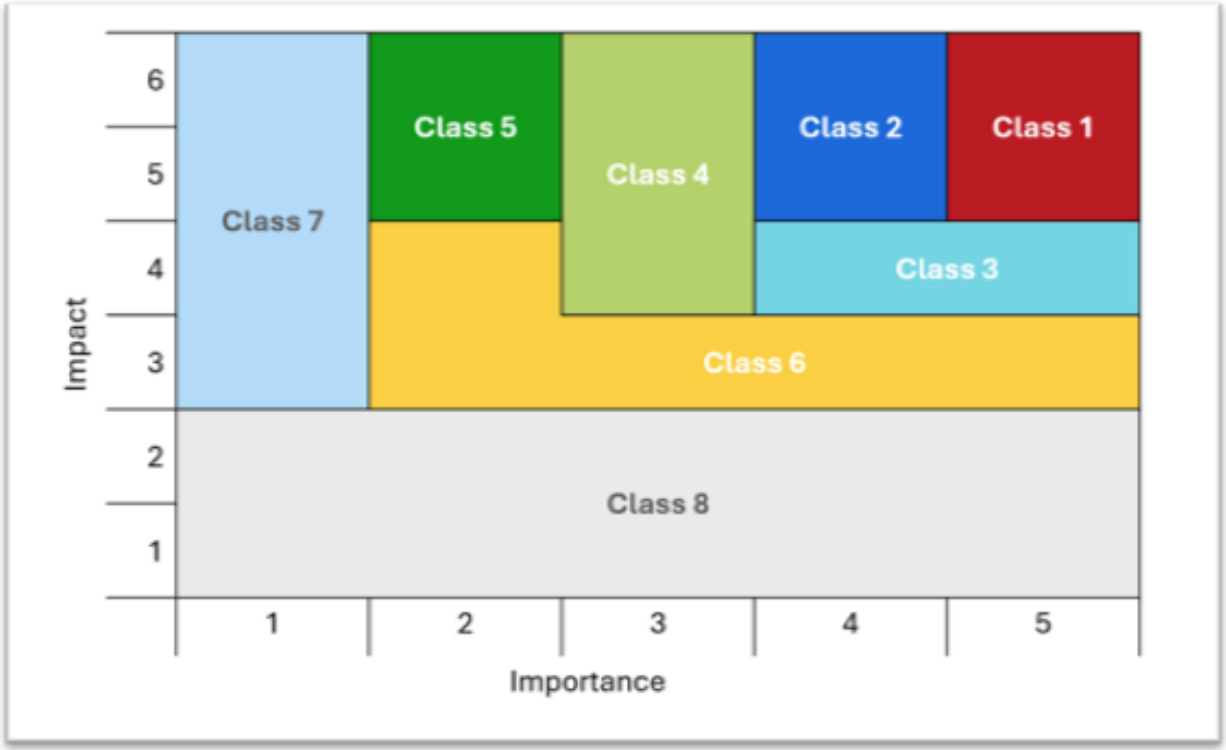
© Hyperion Research 2025

A New Way to Show the Value of Leadership Computing

Using two scales: innovation importance level, and how broadly impactful are the results

FIGURE 1

Innovation Class Map

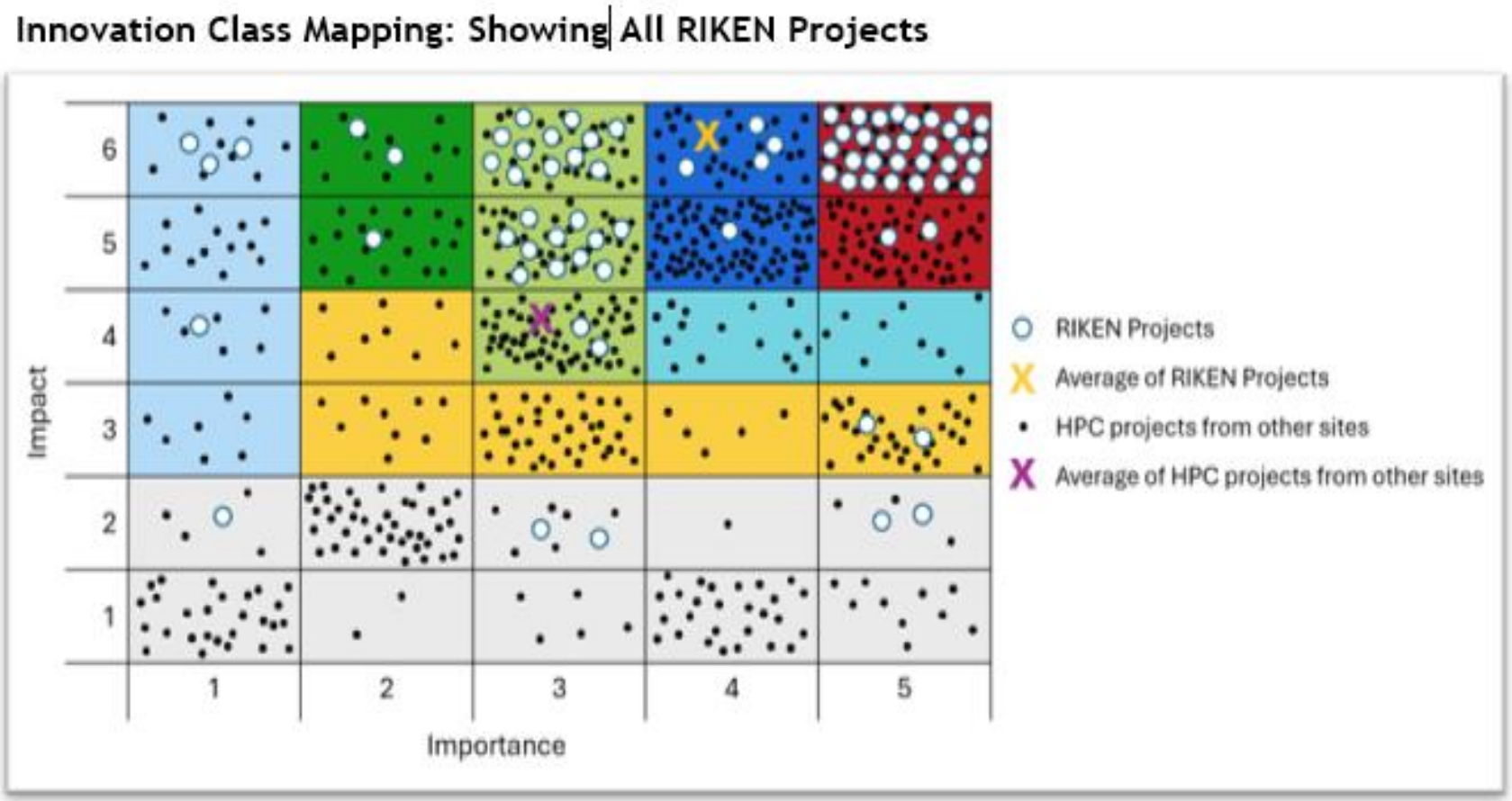


Source: Hyperion Research, 2024

© Hyperion Research 2025

A New Way to Show the Value of Leadership Computing - RIKEN

An example from a 2024 study compared to 650 other projects

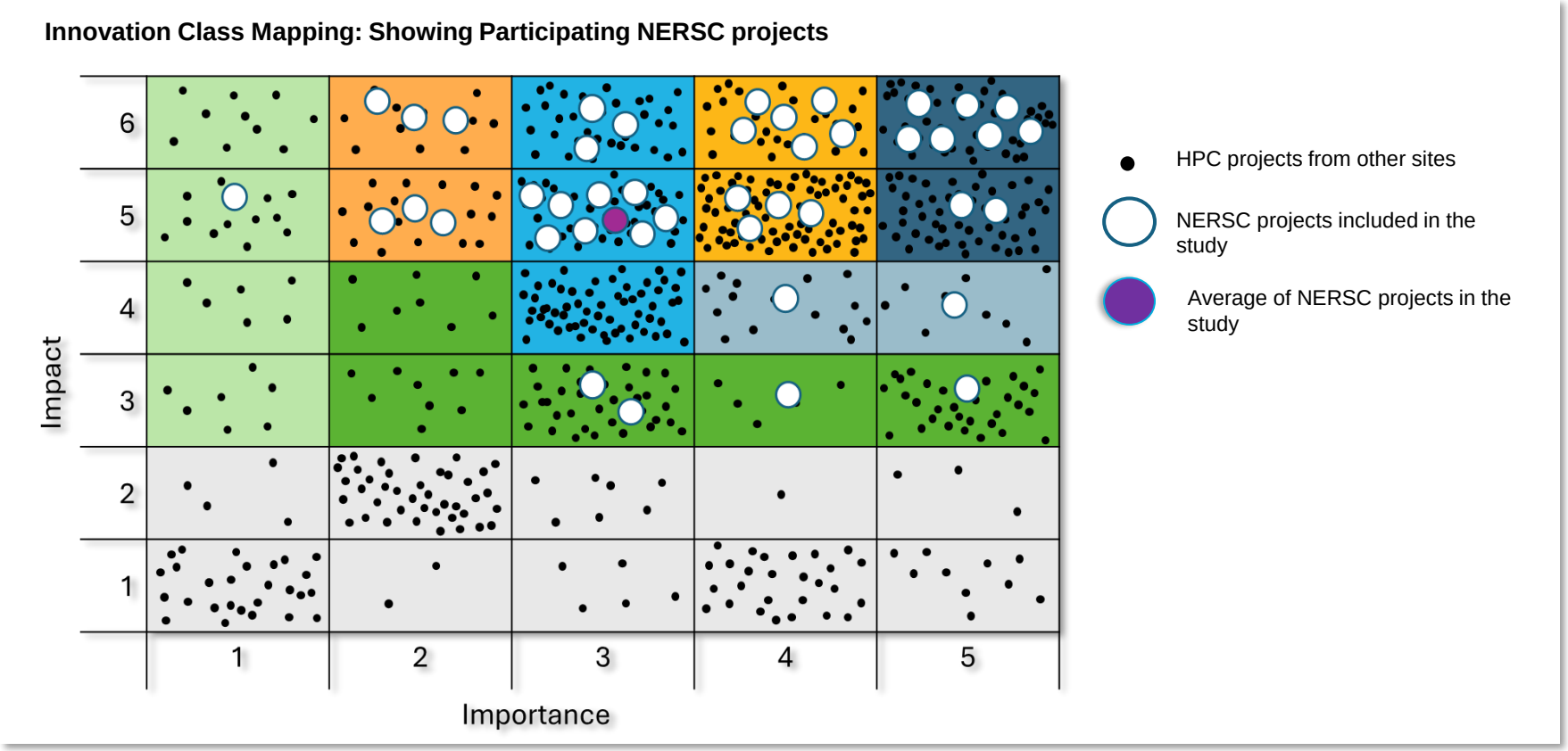


Source: Hyperion Research, 2024

© Hyperion Research 2025

A New Way to Show the Value of Leadership Computing - NERSC

An example from a 2024 study compared to 650 other projects



Source: Hyperion Research, 2024

© Hyperion Research 2025

HYPERION RESEARCH

QUESTIONS?

Questions or comments are welcome!

Please contact us at:
info@hyperionres.com

© Hyperion Research 2025