Hyperion Research Market Update

April 2025

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

www.HyperionResearch.com
www.hpcuserforum.com

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*

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

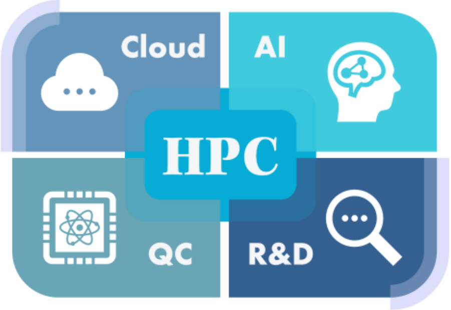
- Mike Thorp, Sr. Global Sales Executive
- Kurt Gantrish, Sr. Account Executive
- Brian Eccles, Client Services Specialist

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

Example Research Areas

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

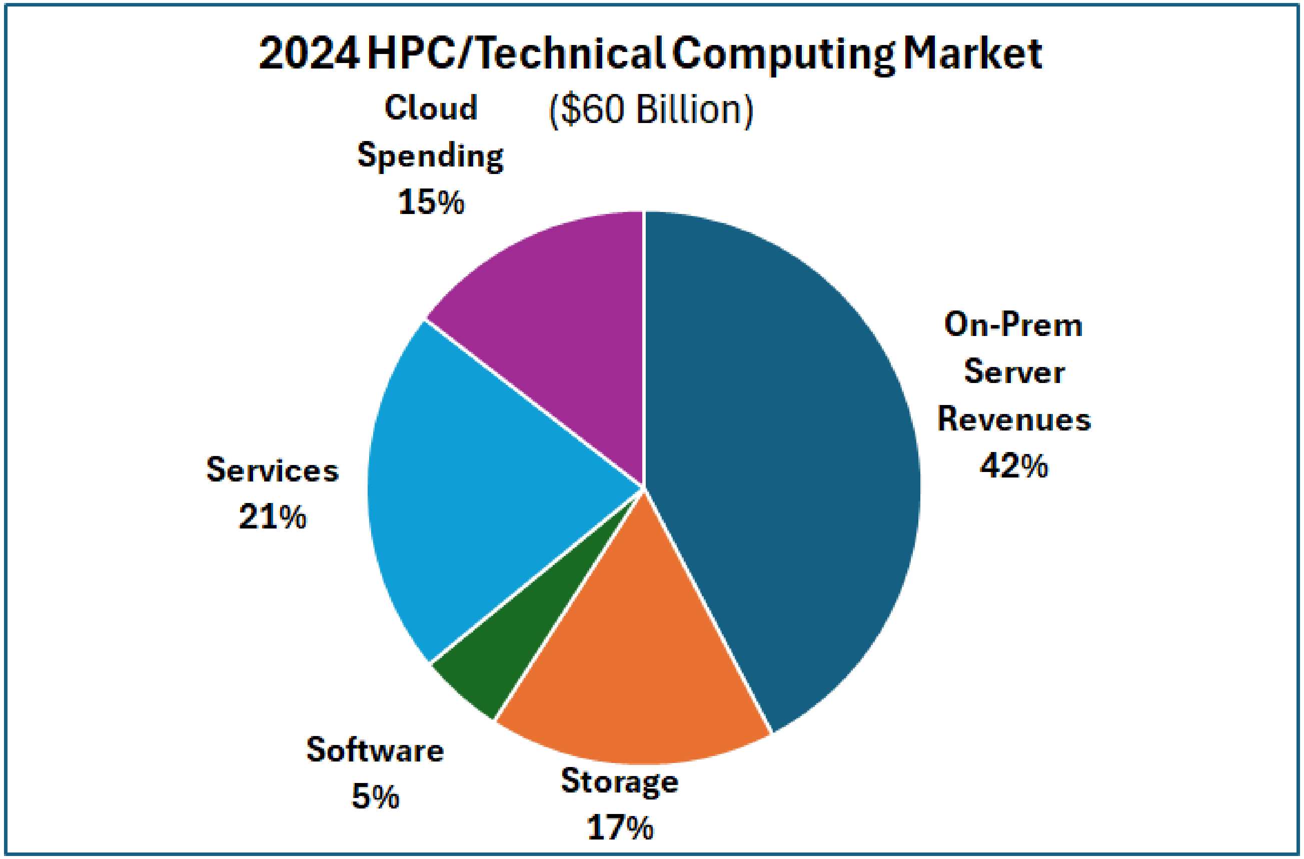


- Traditional HPC
- AI, ML, DL, LLMs, Graph
- Cloud Computing
- Storage & Data
- Interconnects
- Software & Applications
- ROI and Scientific Returns from HPC
- Power & Cooling
- Tracking all Processor Types & Growth rates
- Quantum Computing
- R&D and Engineering -- all types
- Edge Computing
- Supply Chain Issues
- Sustainability

HPC/AI Market Update

2024 Was a Strong Growth Year

The highest growth in over two decades (23.5%)!



- **23.4% growth in on-premises servers**
- **21.3% growth in the use of clouds**
- **Over $60 billion in total spending**

The HPC/AI Market Should See Growth in 2025

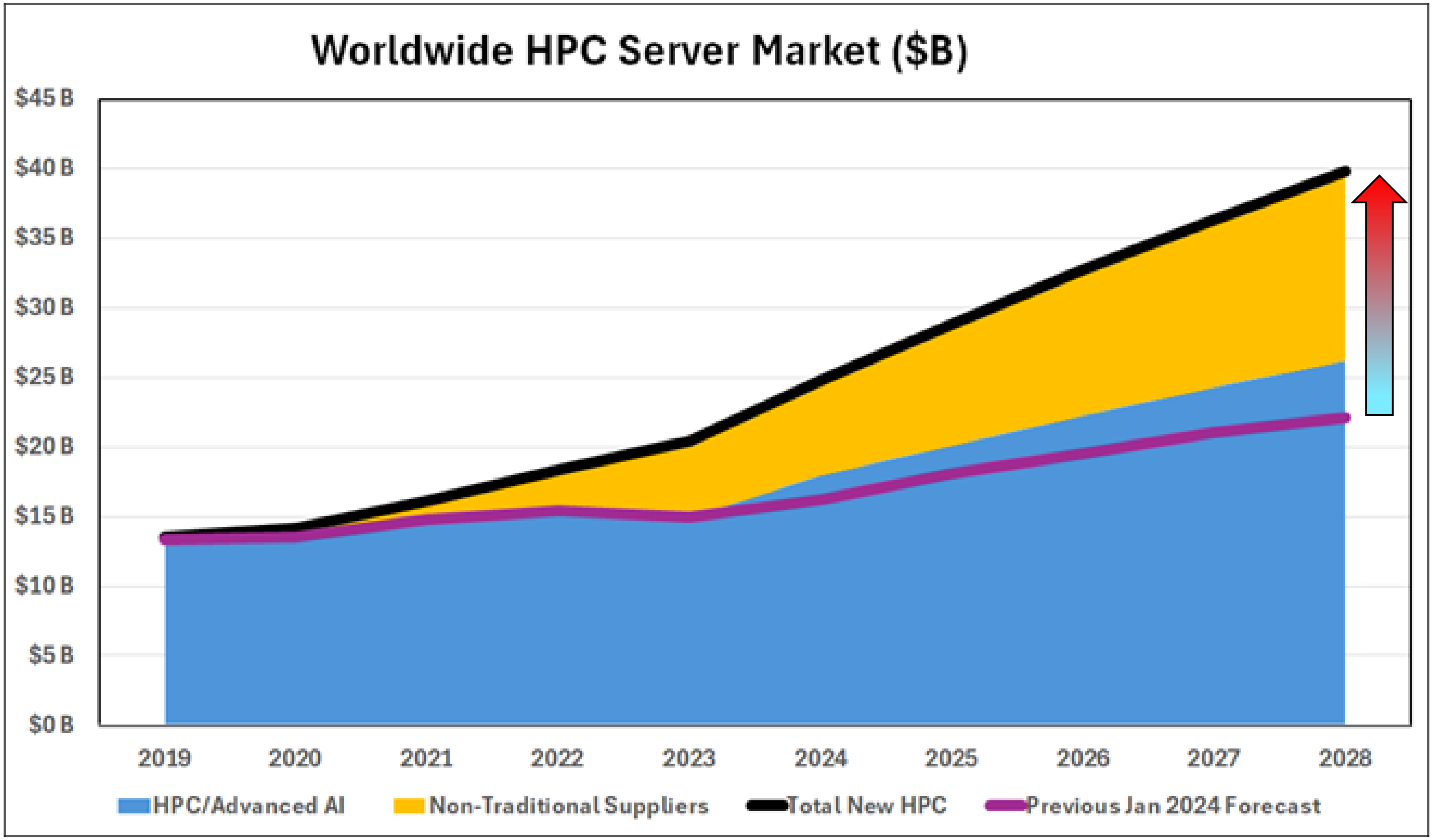
… but there are some major concerns

- **The global economic situation and changing trade rules could have a major impact to IT build outs in 2025**
- **Supply chain issues are still impacting installations (e.g., GPUs)**
- **Exascale system acceptances are seeing delays**
- **The lower end of the on-premises market continues to struggle**

- **Growth drivers include:**
  - New use cases especially in AI/LLMs/Generative AI/Smarter AI are providing new areas for users to advance their research
  - Countries and companies around the world continue to recognize the value of being innovative and investing in R&D to advance society, grow revenues, reduce costs, and become more competitive

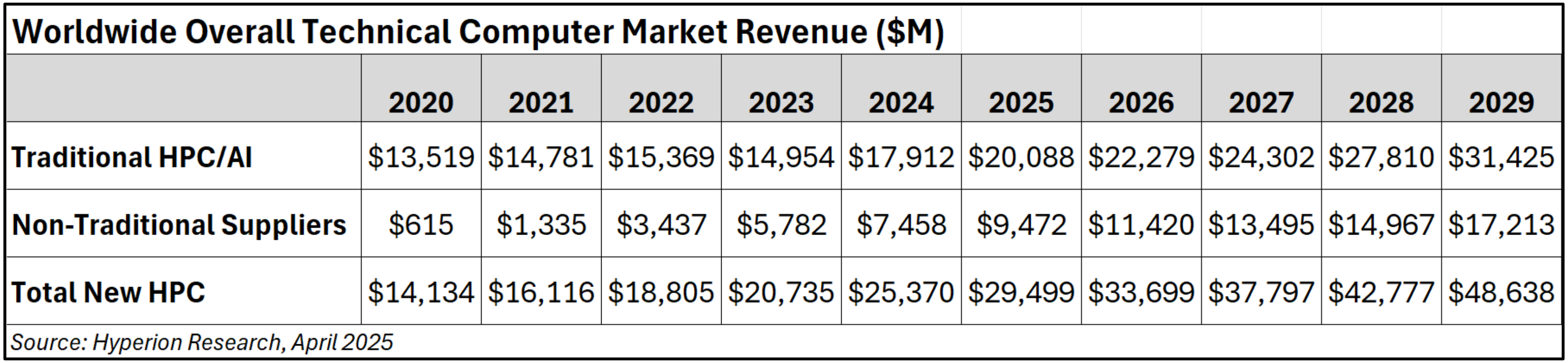
Updated View of the On-Prem Server Market

- ***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***



Updated View of the HPC/AI Market

On-prem HPC/AI servers are projected to exceed $48 billion in 2029



Worldwide Overall Technical Computer Market Revenue ($M)

	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029
Traditional HPC/AI	$13,519	$14,781	$15,369	$14,954	$17,912	$20,088	$22,279	$24,302	$27,810	$31,425
Non-Traditional Suppliers	$615	$1,335	$3,437	$5,782	$7,458	$9,472	$11,420	$13,495	$14,967	$17,213
Total New HPC	$14,134	$16,116	$18,805	$20,735	$25,370	$29,499	$33,699	$37,797	$42,777	$48,638

Source: Hyperion Research, April 2025

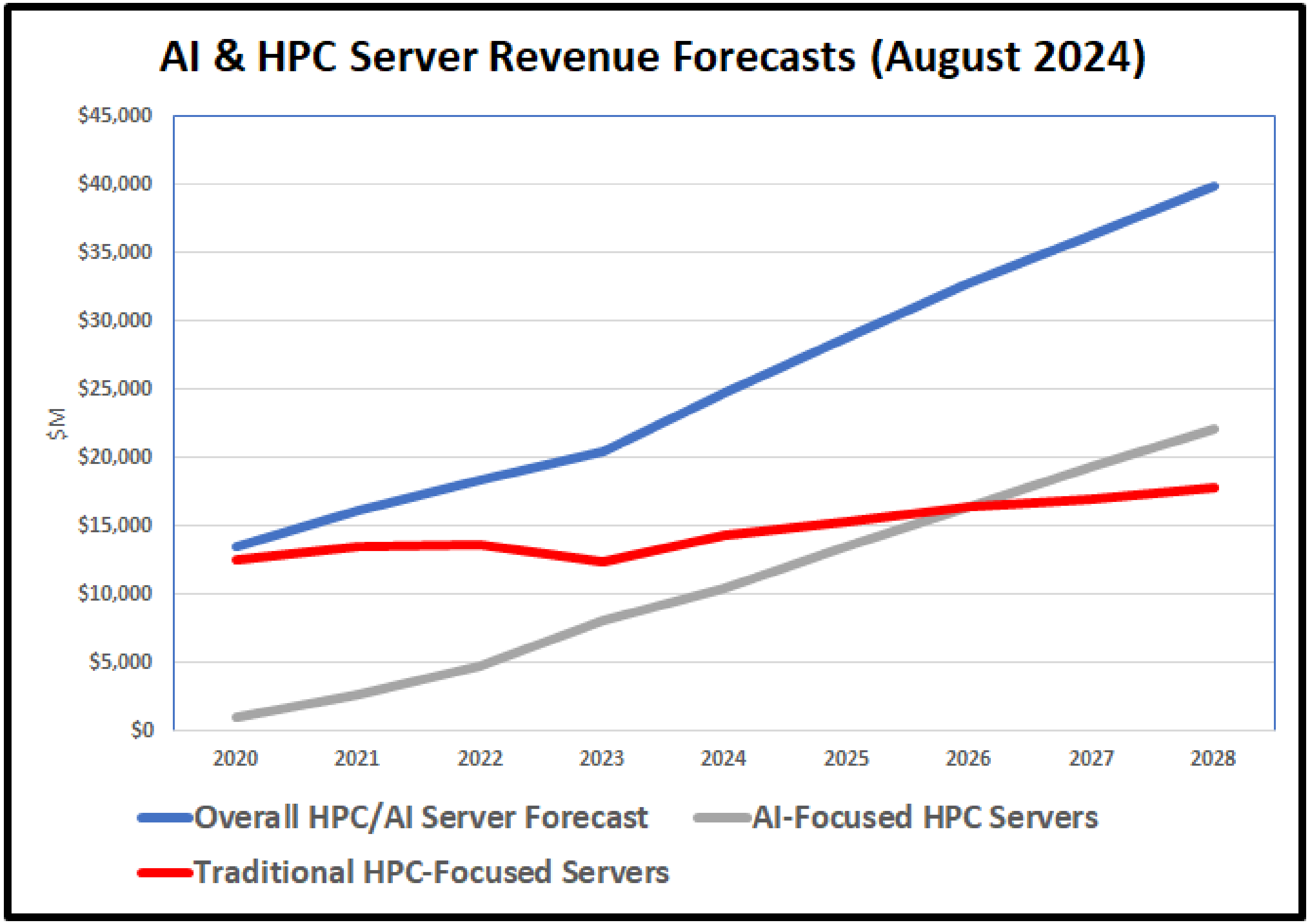
Market Segment Definition: *Non-Traditional Suppliers* *(new revenues added to the previous HPC market sizing)*

These are on-premises AI-centric HPC servers that are provided by non-traditional HPC suppliers like NVIDIA, Cerebras, SambaNova, SuperMicro, etc. These servers are designed primarily to run AI and AI-related workloads

- These servers are a subsegment of the overall HPC market but haven't historically been accounted for within prior HPC market numbers

HPC Compared to AI-centric Servers

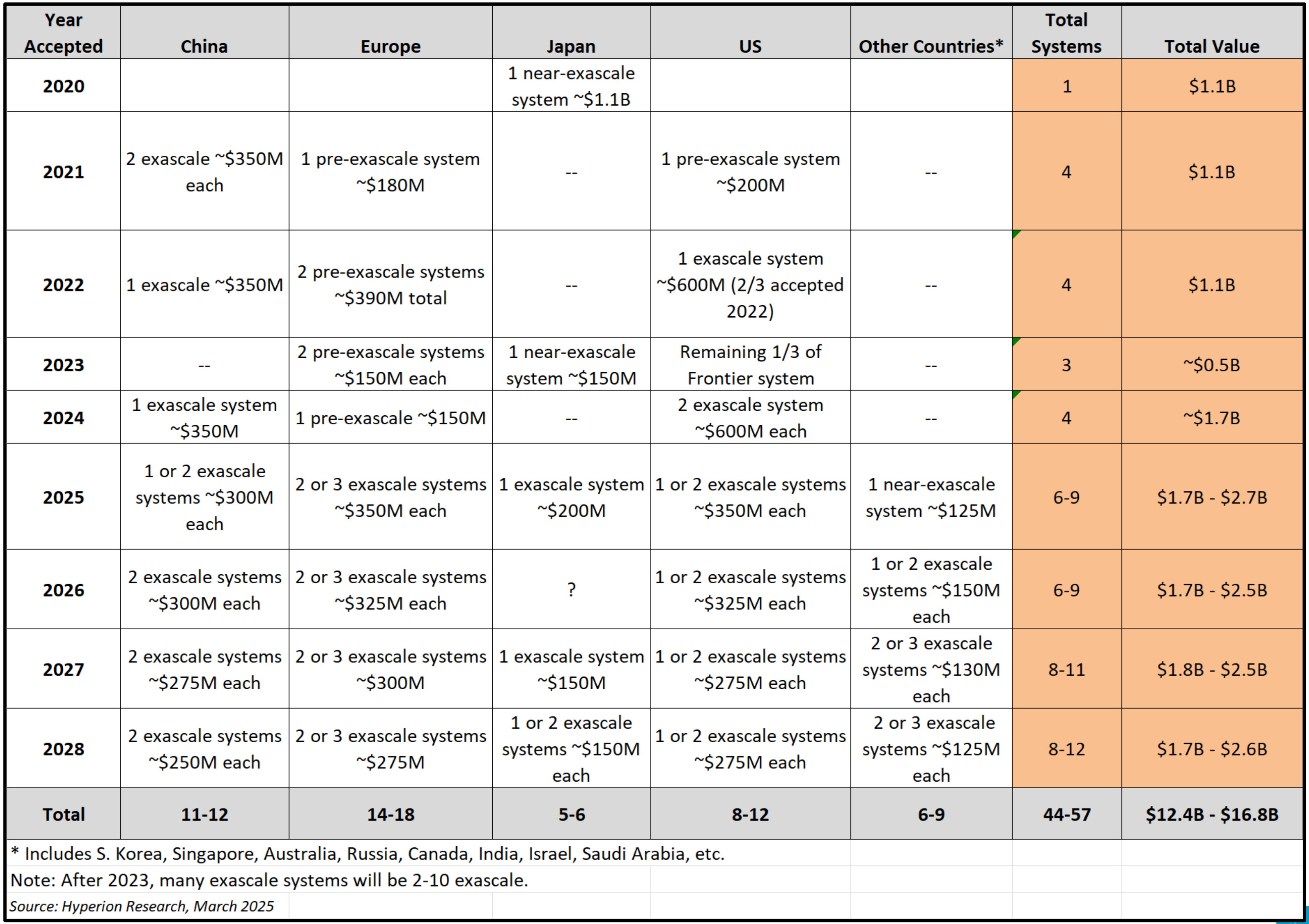
Many servers are running both traditional HPC and AI Workloads



Note: AI systems may still run some traditional HPC jobs (<50% of workload).
Likewise, traditional HPC systems often run some AI jobs (<50% of workload).

The Exascale Market (System Acceptances)

Over 45 systems and over $12 billion in value



Year Accepted	China	Europe	Japan	US	Other Countries*	Total Systems	Total Value
2020			1 near-exascale system ~$1.1B			1	$1.1B
2021	2 exascale ~$350M each	1 pre-exascale system ~$180M	--	1 pre-exascale system ~$200M	--	4	$1.1B
2022	1 exascale ~$350M	2 pre-exascale systems ~$390M total	--	1 exascale system ~$600M (2/3 accepted 2022)	--	4	$1.1B
2023	--	2 pre-exascale systems ~$150M each	1 near-exascale system ~$150M	Remaining 1/3 of Frontier system	--	3	~$0.5B
2024	1 exascale system ~$350M	1 pre-exascale ~$150M	--	2 exascale system ~$600M each	--	4	~$1.7B
2025	1 or 2 exascale systems ~$300M each	2 or 3 exascale systems ~$350M each	1 exascale system ~$200M	1 or 2 exascale systems ~$350M each	1 near-exascale system ~$125M	6-9	$1.7B - $2.7B
2026	2 exascale systems ~$300M each	2 or 3 exascale systems ~$325M each	?	1 or 2 exascale systems ~$325M each	1 or 2 exascale systems ~$150M each	6-9	$1.7B - $2.5B
2027	2 exascale systems ~$275M each	2 or 3 exascale systems ~$300M	1 exascale system ~$150M	1 or 2 exascale systems ~$275M each	2 or 3 exascale systems ~$130M each	8-11	$1.8B - $2.5B
2028	2 exascale systems ~$250M each	2 or 3 exascale systems ~$275M	1 or 2 exascale systems ~$150M each	1 or 2 exascale systems ~$275M each	2 or 3 exascale systems ~$125M each	8-12	$1.7B - $2.6B
Total	**11-12**	**14-18**	**5-6**	**8-12**	**6-9**	**44-57**	**$12.4B - $16.8B**

* Includes S. Korea, Singapore, Australia, Russia, Canada, India, Israel, Saudi Arabia, etc.

Note: After 2023, many exascale systems will be 2-10 exascale.

Source: Hyperion Research, March 2025

Hyperion Research Predictions

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)

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

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

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

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

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

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 already on the books
    - IBM, QuEra, IQM, D-Wave
  - 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

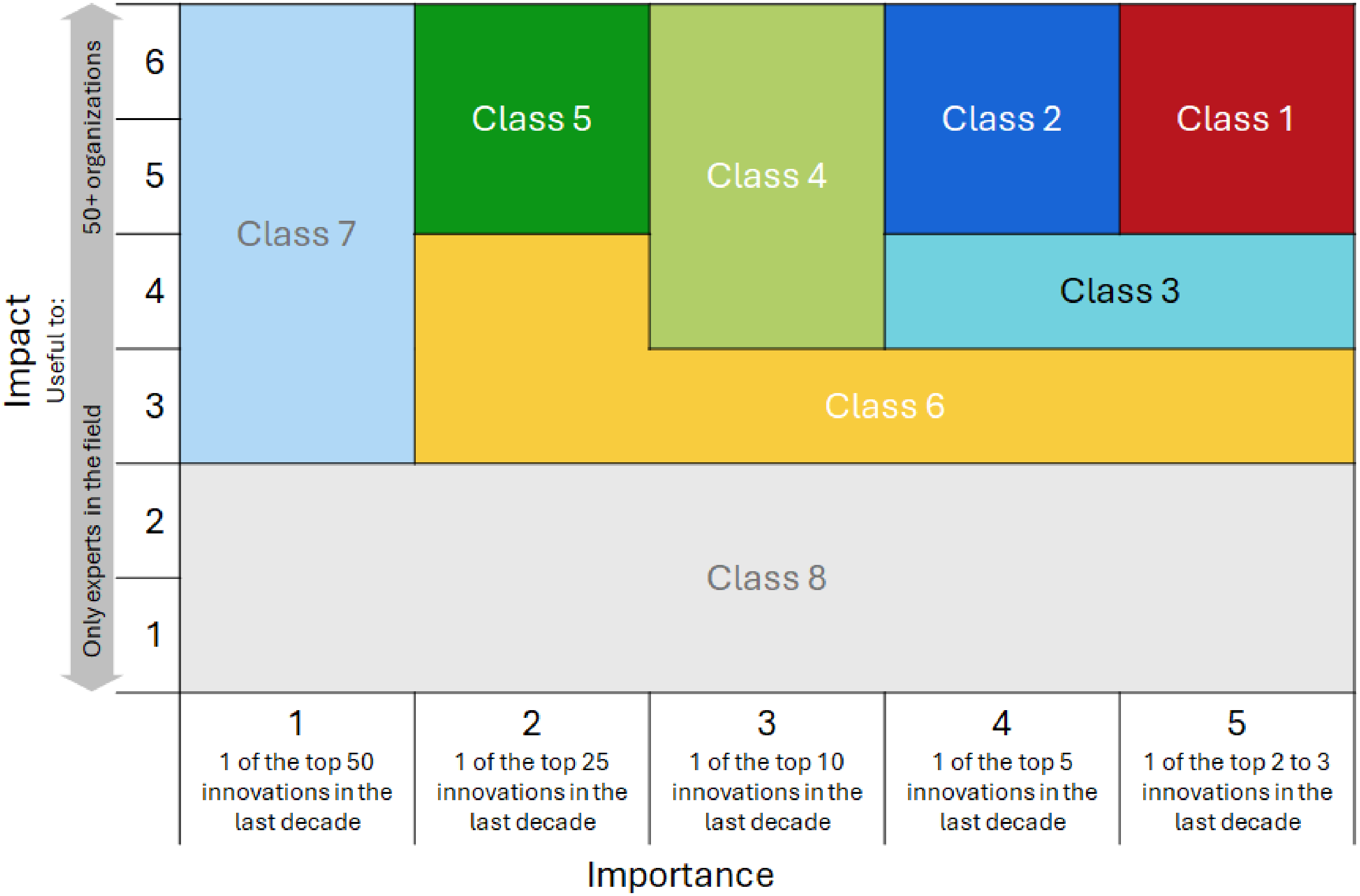
In Summary

Conclusions

- **2024 was a strong growth year**
  - GPUs, cloud, AI/ML/DL/LLM are high growth areas
  - QC systems are being installed around the world
- **New technologies are showing up large numbers:**
  - Generative AI, smarter AI, LLMs and SLLs are fueling a new level of growth
  - Processors, AI hardware & software, memories, new storage approaches, etc.
  - The cloud has become a viable option for many HPC workloads
- **Storage will likely see major growth driven by AI, big data and the need for much larger data sets**
- **There are still growing concerns around power & talent**

A New Way to Show the Value of Leadership Computing

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



Impact

Useful to:

50+ organizations

Only experts in the field

6

5

4

3

2

1

Class 5

Class 4

Class 2

Class 1

Class 7

Class 3

Class 6

Class 8

1
1 of the top 50 innovations in the last decade

2
1 of the top 25 innovations in the last decade

3
1 of the top 10 innovations in the last decade

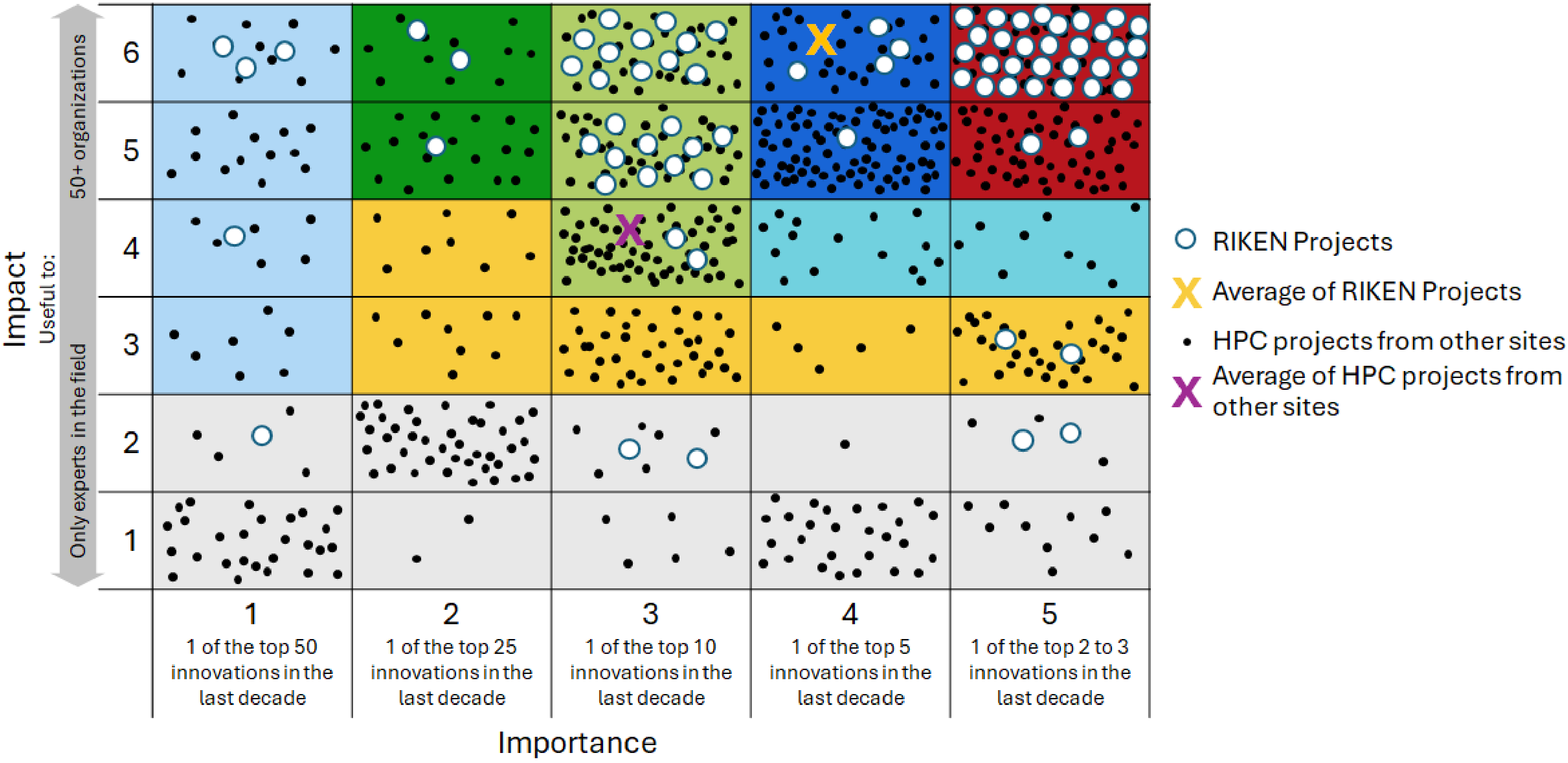
4
1 of the top 5 innovations in the last decade

5
1 of the top 2 to 3 innovations in the last decade

Importance

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

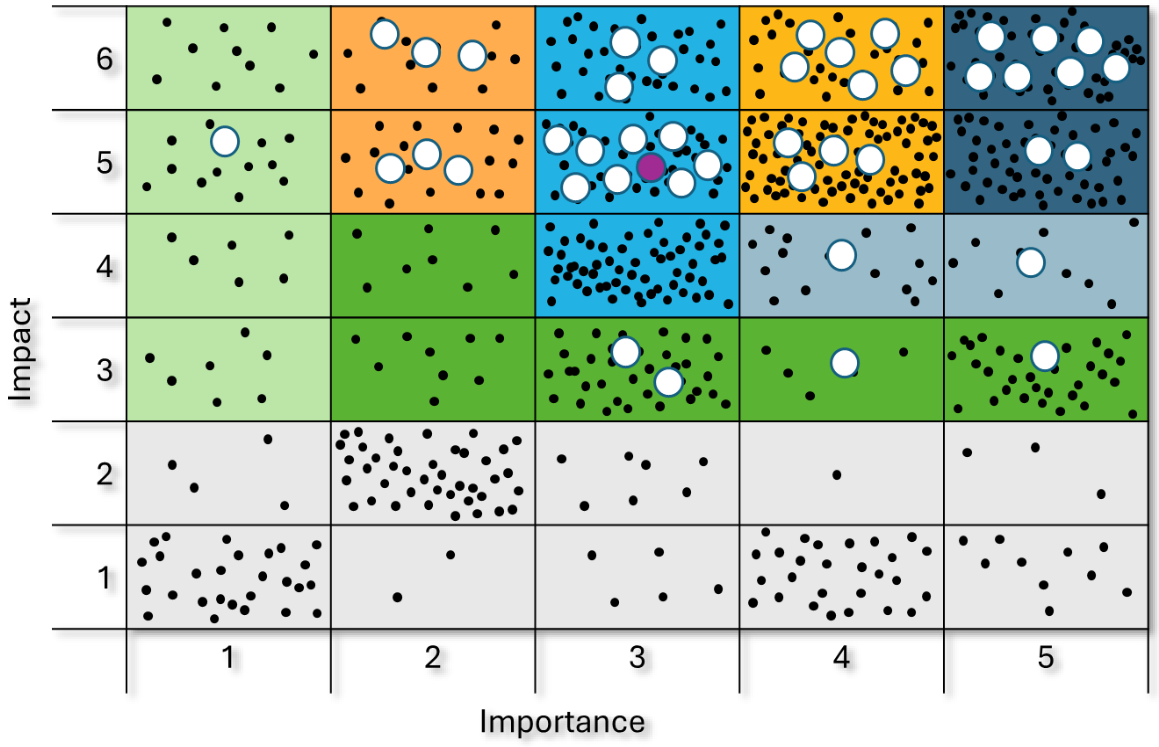
An example from a 2024 study compared to 650 other projects



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

An example from a 2024 study compared to 650 other projects

Innovation Class Mapping: Showing Participating NERSC projects



● HPC projects from other sites

○ NERSC projects included in the study

● Average of NERSC projects in the study

We Welcome Questions, Comments and Suggestions

Please contact us at:
info@hyperionres.com