Hyperion Research:
ISC26 Market Update

June 2026

www.HyperionResearch.com
www.hpcuserforum.com

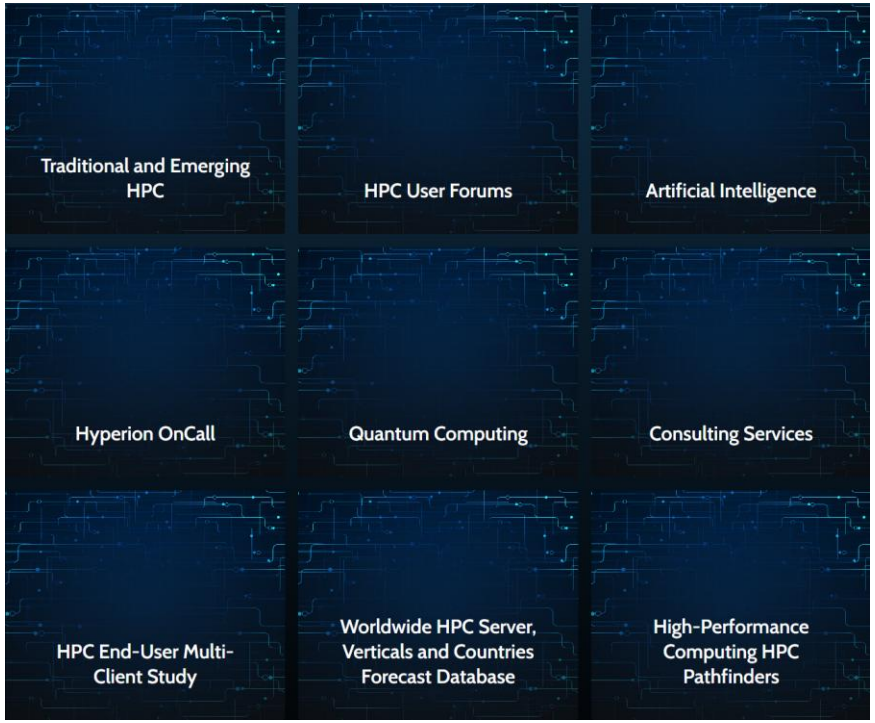
Earl Joseph, Bob Sorensen, Mark Nossokoff,
Tom Sorensen, and Jaclyn Ludema

Welcome

- Today's presentation will be sent to all registrants
- For follow-up details or discussions, please email mthorp@hyperionres.com
- The presentation will conclude at 8:30, but you are welcome to stay until 10:00 for networking and more coffee!

Who is Hyperion Research?

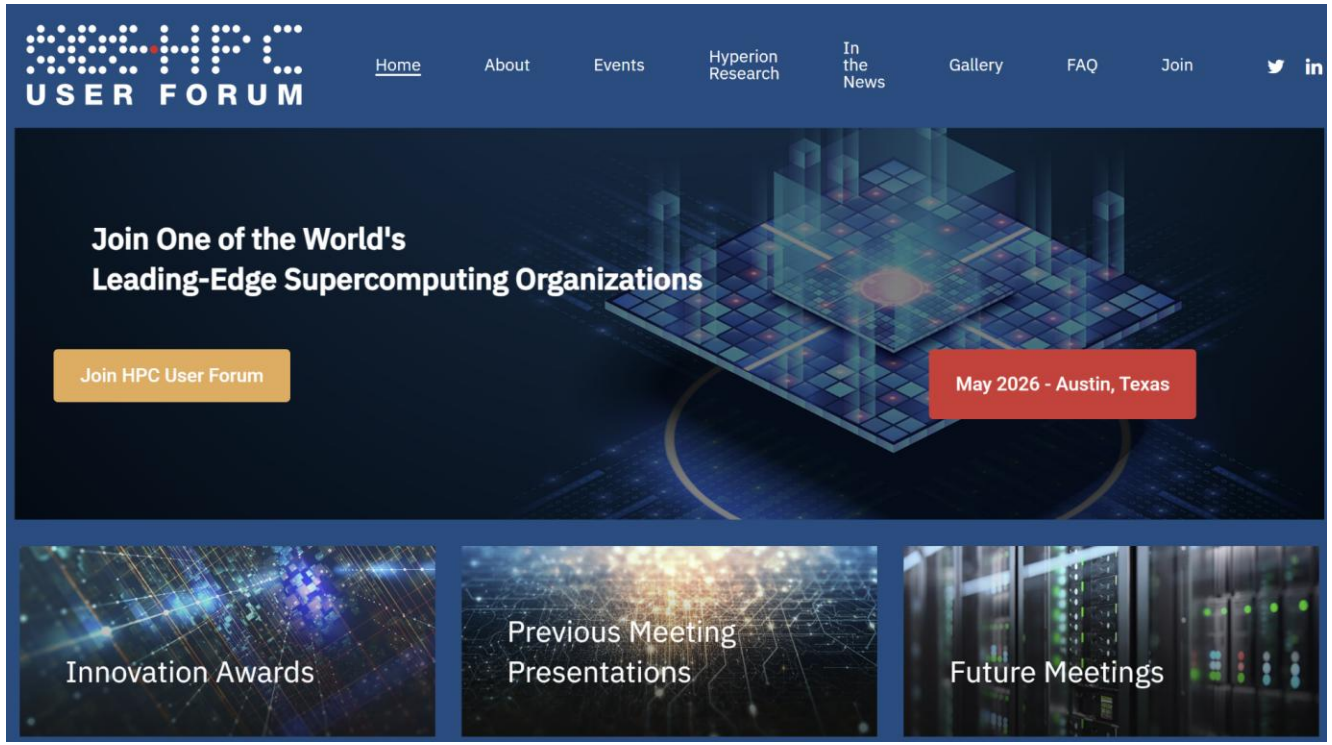
(www.HyperionResearch.com & www.HPCUserForum.com)

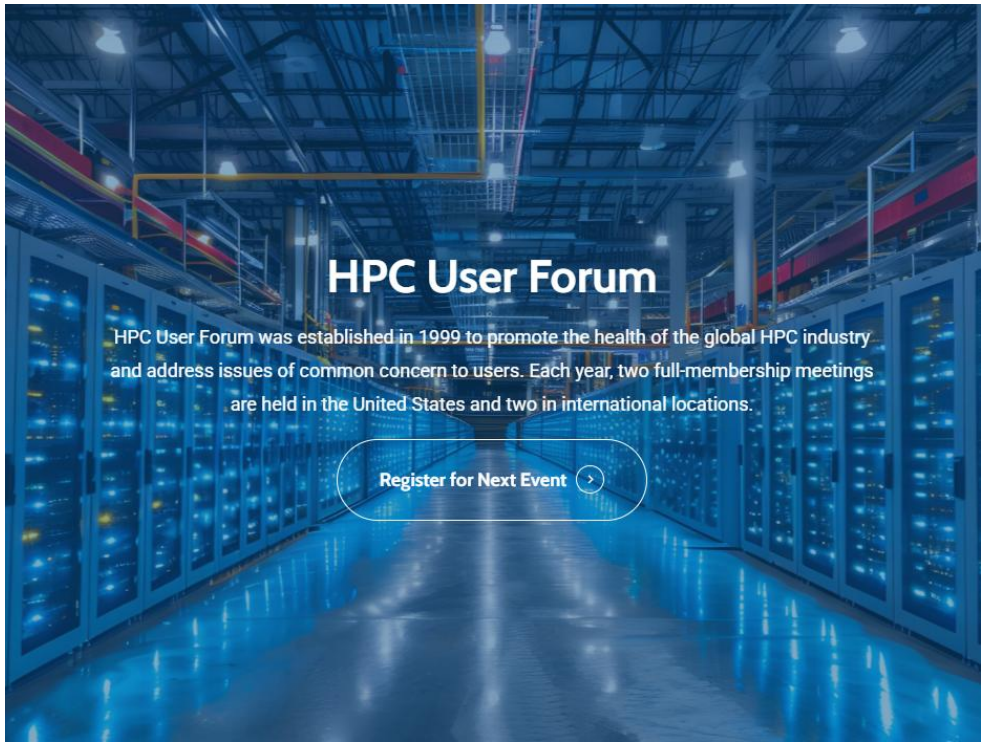


- **Provides timely, in-depth mission-critical insights across broad portfolio of relevant topics including HPC, AI, Quantum, storage and interconnects, data centers assessment, and ROI**
- **Delivered through variety of services consisting of custom projects, advisory, white papers, webinars, panel participation, and subscription**

Who is Hyperion Research?

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





- **Promote the health of the global HPC industry**
- **Address issues of common concern**
- **Hold multiple membership meetings throughout the year; typically, 2 in the U.S. and 2 globally**

Hyperion Research Team

Earl Joseph

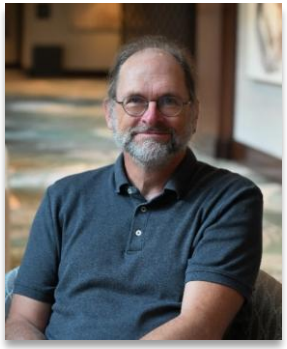
- CEO
- Executive Director HPC/AI User Forum

Jean Sorensen

- COO
- Human Resources

Bob Sorensen

- Sr. VP of Research
- Chief Quantum Analyst
- AI Analyst



Mark Nossokoff

- Research Director
- Chief Storage and Cloud Analyst

Jaclyn Ludema

- Sustainability Analyst
- Cloud Analyst

Tom Sorensen

- Principal AI/HPC Analyst
- Editor ACN Update

Mike Thorp

- Sr. Global Account Executive

Tyler Pla

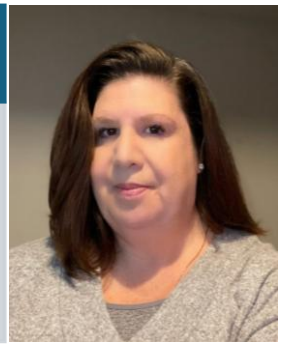
- Global Account Executive

Hyperion Research Partners

Katsuya Nishi

- Sr. Account Rep., Japan
- Chief Editor, HPCWire Japan

Mara Jacob



- HPC User Forum Event Director

Benjamin Portman

- President, Orpheus
- Web designer & developer

Andrew Rugg

- President, Certus Insights

Mike Heroux

- Consulting Analyst and Contributor

Kirsten Chapman

- Owner, K Chapman Consulting
- Data Collection Operations

KC Associates

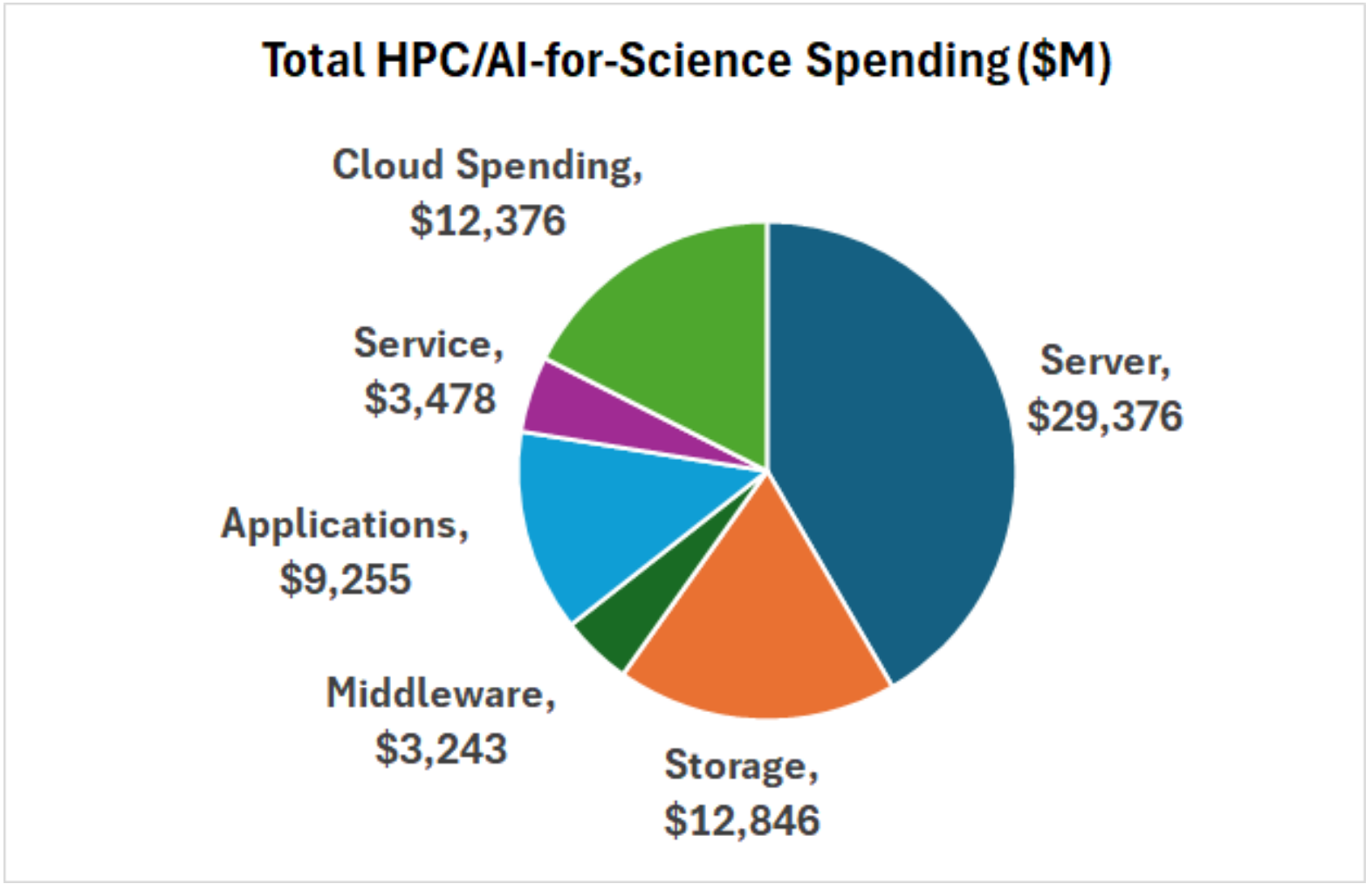
Today's Agenda

- **Mike Thorp, Senior Global Account Executive**
  - Introduction
- **Earl Joseph, CEO**
  - HPC and AI Market Update
- **Bob Sorensen, SVP, Chief QC & AI Analyst**
  - 6th Annual Global QC Market Survey
  - Recent Study Overviews: Gen AI ROI, AI in the Cloud
  - FP64 vs FP4: An Evolving Debate
- **Mark Nossokoff, Research Director, Chief Storage & Cloud Analyst**
  - Perspectives on HPC-AI Cloud, Storage, Interconnects, and Sustainability
  - Scientific Computing Cost, Value, and ROI Model
- **Conclusions**
- **Q&A**

2025 Was a Strong Growth Year

16.9% overall growth in spending!

- **15.0% growth in on-premises servers**
- **29.7% growth in the use of clouds**
- **Over $70 billion in total spending**



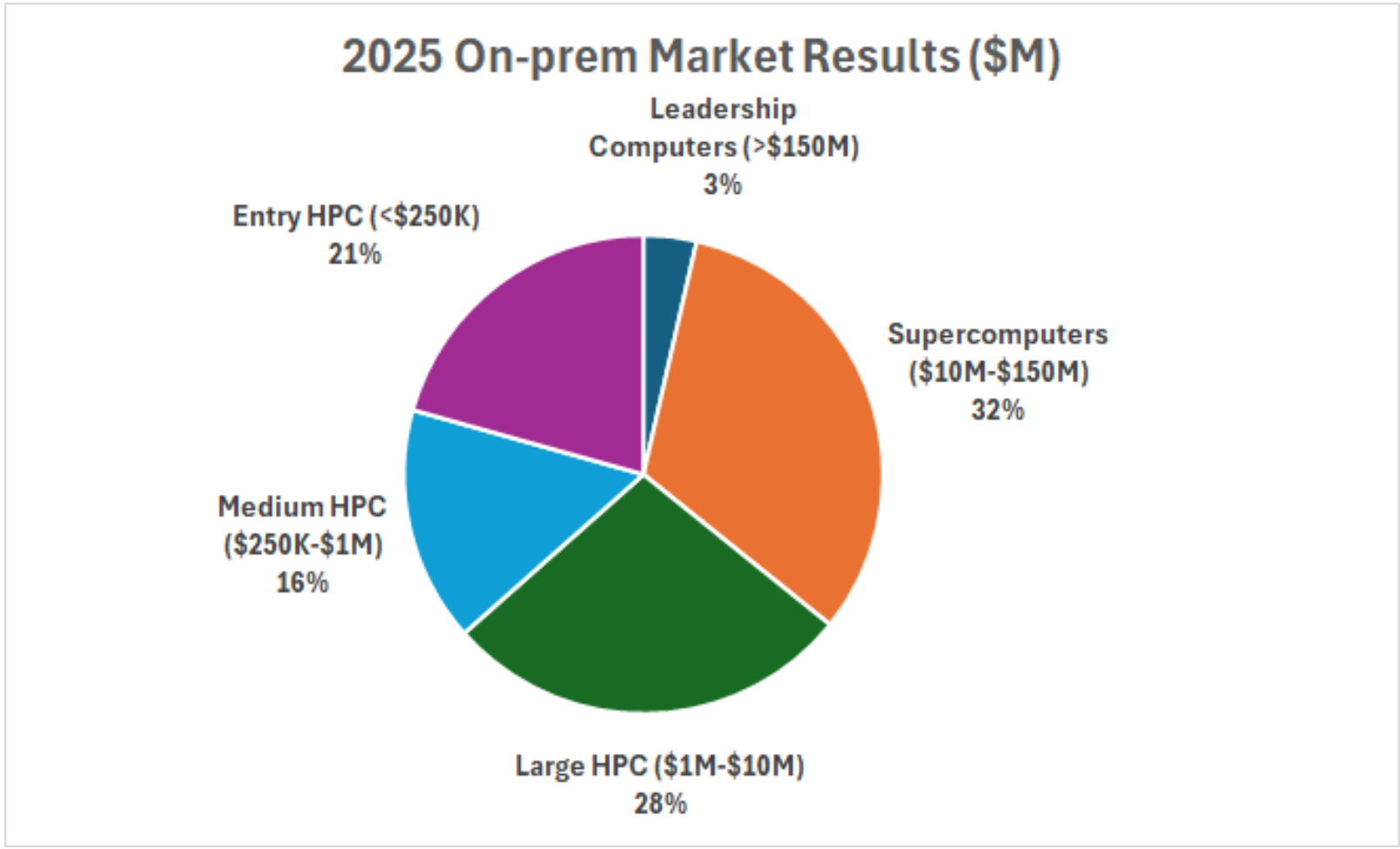
2025 HPC/AI Market By Vendor

On-prem sever sales were just under $30 billion in 2025

2025 HPC/AI On-Prem Sales		
Vendor	**2025 Sales**	**2025 Share**
HPE	$6,485	22.1%
Dell Technologies	$5,342	18.2%
Lenovo	$1,783	6.1%
Inspur	$1,236	4.2%
Sugon	$693	2.4%
Atos	$464	1.6%
IBM	$453	1.5%
Penguin	$508	1.7%
Fujitsu	$259	0.9%
NEC	$238	0.8%
Other HPC Suppliers	$2,735	9.3%
Non-Traditional Suppliers	$9,180	31.2%
Total	**$29,376**	**100.0%**
Source: Hyperion Research, April 2026		

2025 HPC/AI Market By Segment

Supercomputers ($10M-$150M) is the largest segment at $9.5 billion



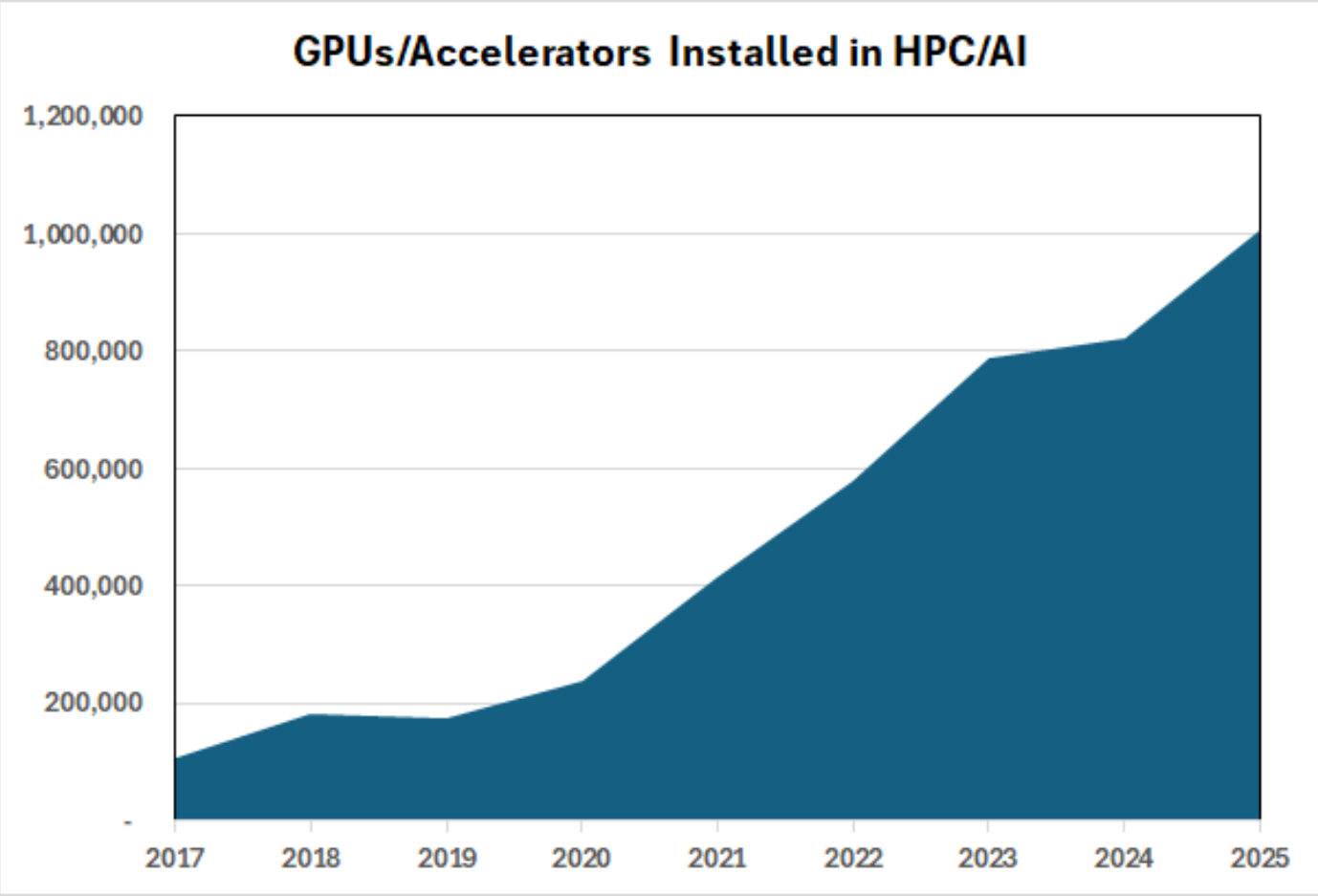
2025 HPC/AI Market By Vertical

Five verticals are now over $2 billion, and two are over $4 billion

HPC/AI On-prem System Installations ($M)		
	2024	2025
Bio-Sciences	2,302	2,622
CAE	2,751	3,134
Chemical Engineering	305	341
DCC & Distribution	1,384	1,610
Economics/Financial	1,336	1,526
EDA / IT / ISV	1,497	1,701
Geosciences	1,550	1,782
Mechanical Design	28	27
Defense	2,695	3,213
Government Lab	6,115	7,129
University/Academic	4,039	4,542
Weather	1,137	1,288
Other	415	462
Total Revenue	**25,554**	**29,376**
Source: Hyperion Research, May 2026		

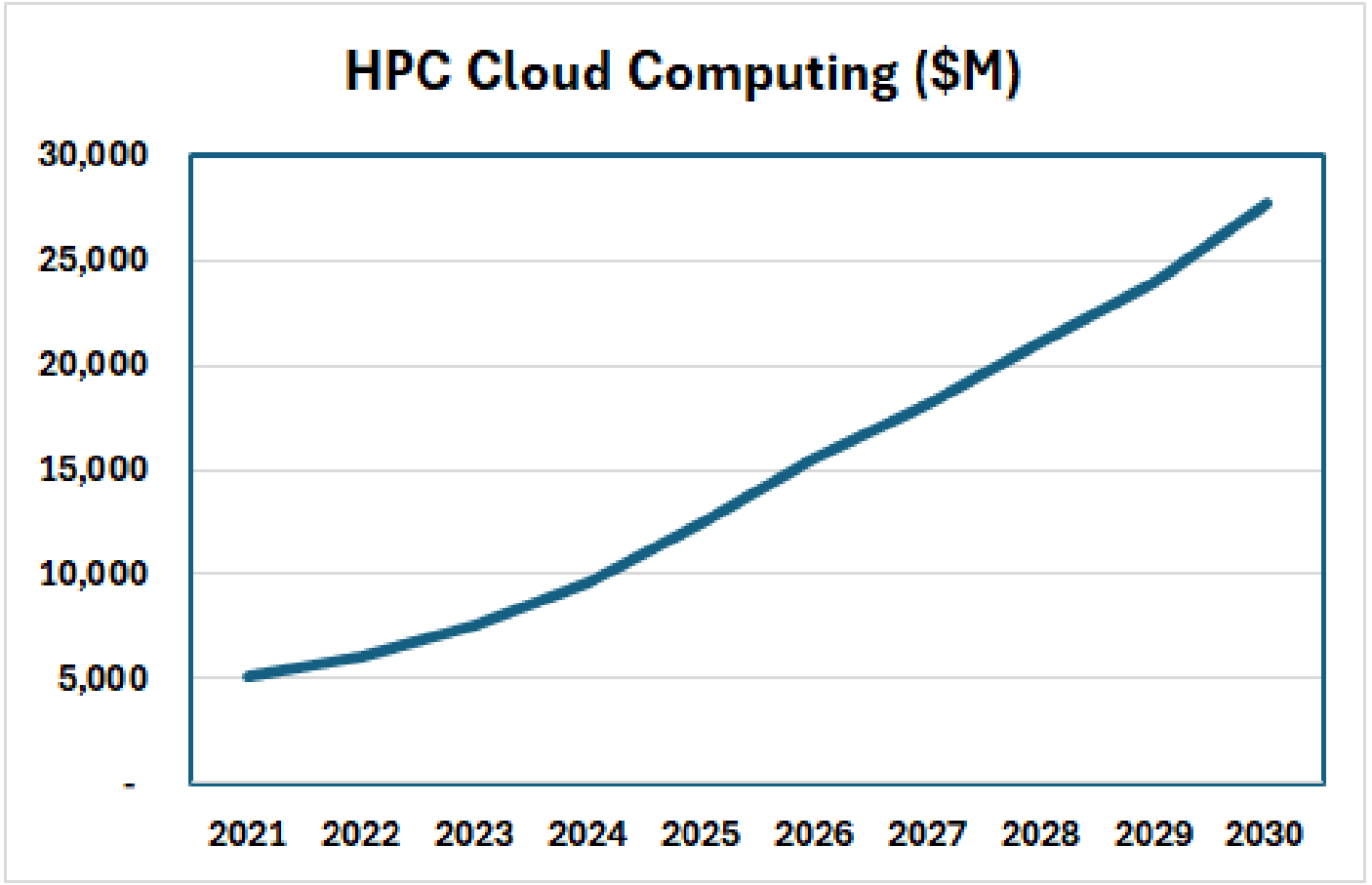
GPU Installations: Over 1 Million in 2025

In HPC & AI-for-Science = 33.3% CAGR over the last 5 years



Cloud Computing for HPC/AI-for-Science

Projected cloud HPC/AI spending to reach $30 billion by 2030
Grew 30% in 2025



HPC/AI-for-Science Market Should See Growth in 2026

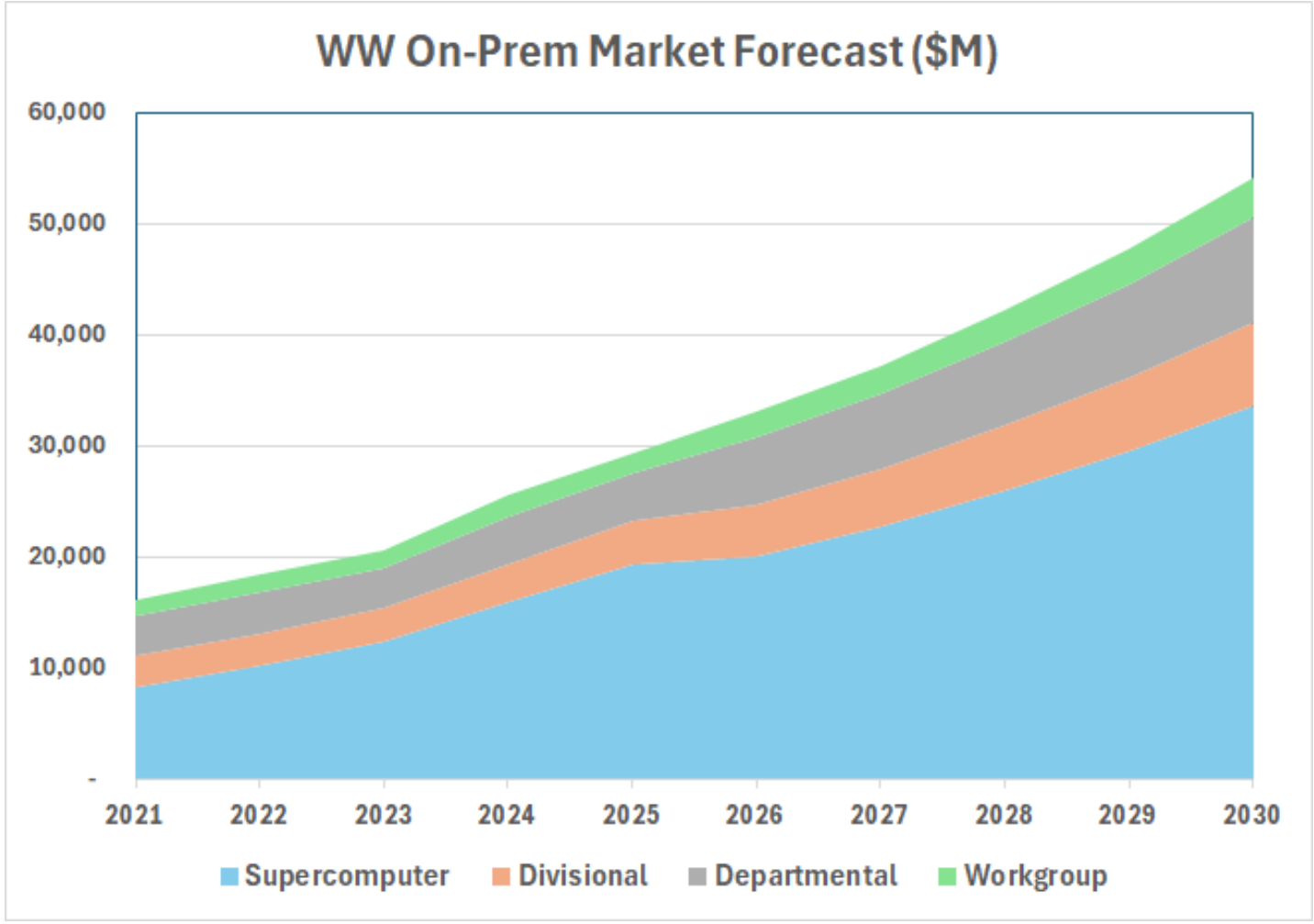
… but there are some major concerns

- **The global economic situation and changing trade rules could have a major impact to IT build outs in 2026**
- **Non-science AI systems are getting major attention and $$**
- **Supply chain issues are still impacting installations (e.g., GPUs, memory & SSDs)**
- **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 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

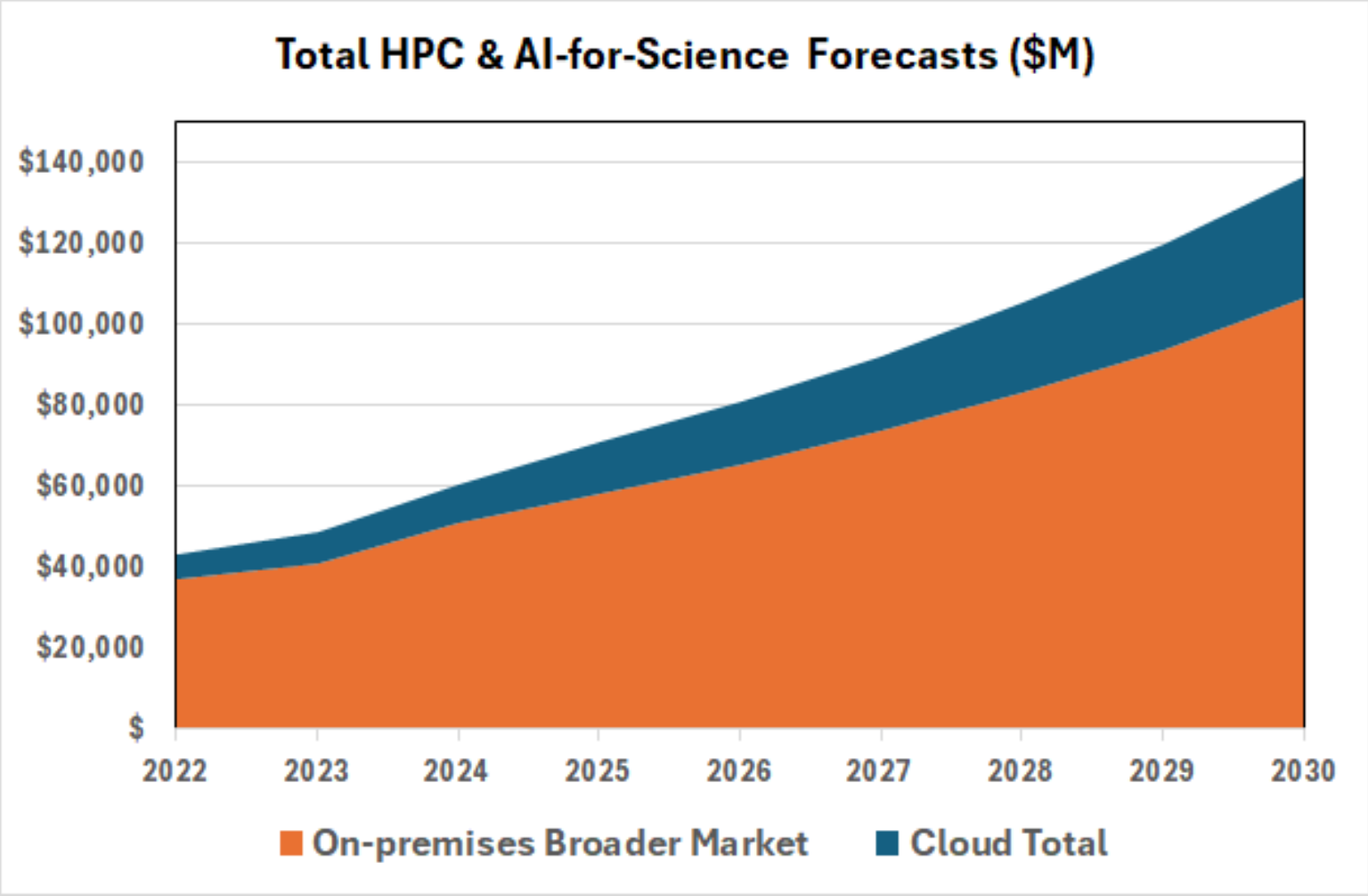
HPC/AI-for-Science On-Prem Server Forecast

On-prem servers are projected to reach $54 billion by 2030
Broader HPC/AI on-prem market is projected to exceed $100 billion



HPC & AI-for-Science On-Prem Plus Cloud Forecast

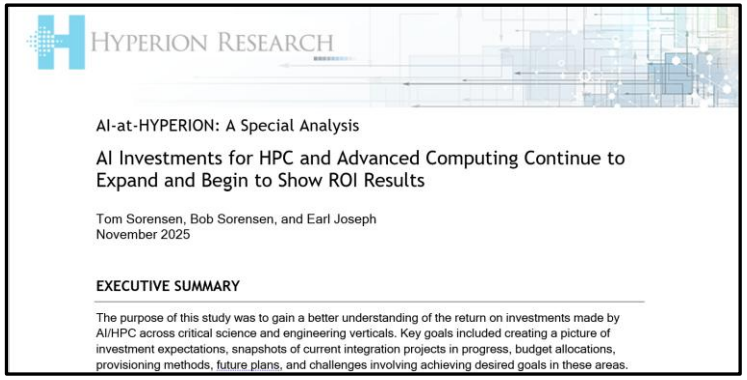
The overall market is projected to exceed $135 billion by 2030
Doubling over the next 5 years



Trends in AI-for-Science

Success in Using AI-for-Science

"AI Investments for HPC and Advanced Computing Continue to Expand and Begin to Show ROI Results"


HYPERION RESEARCH

AI-at-HYPERION: A Special Analysis

AI Investments for HPC and Advanced Computing Continue to Expand and Begin to Show ROI Results

Tom Sorensen, Bob Sorensen, and Earl Joseph
November 2025

EXECUTIVE SUMMARY

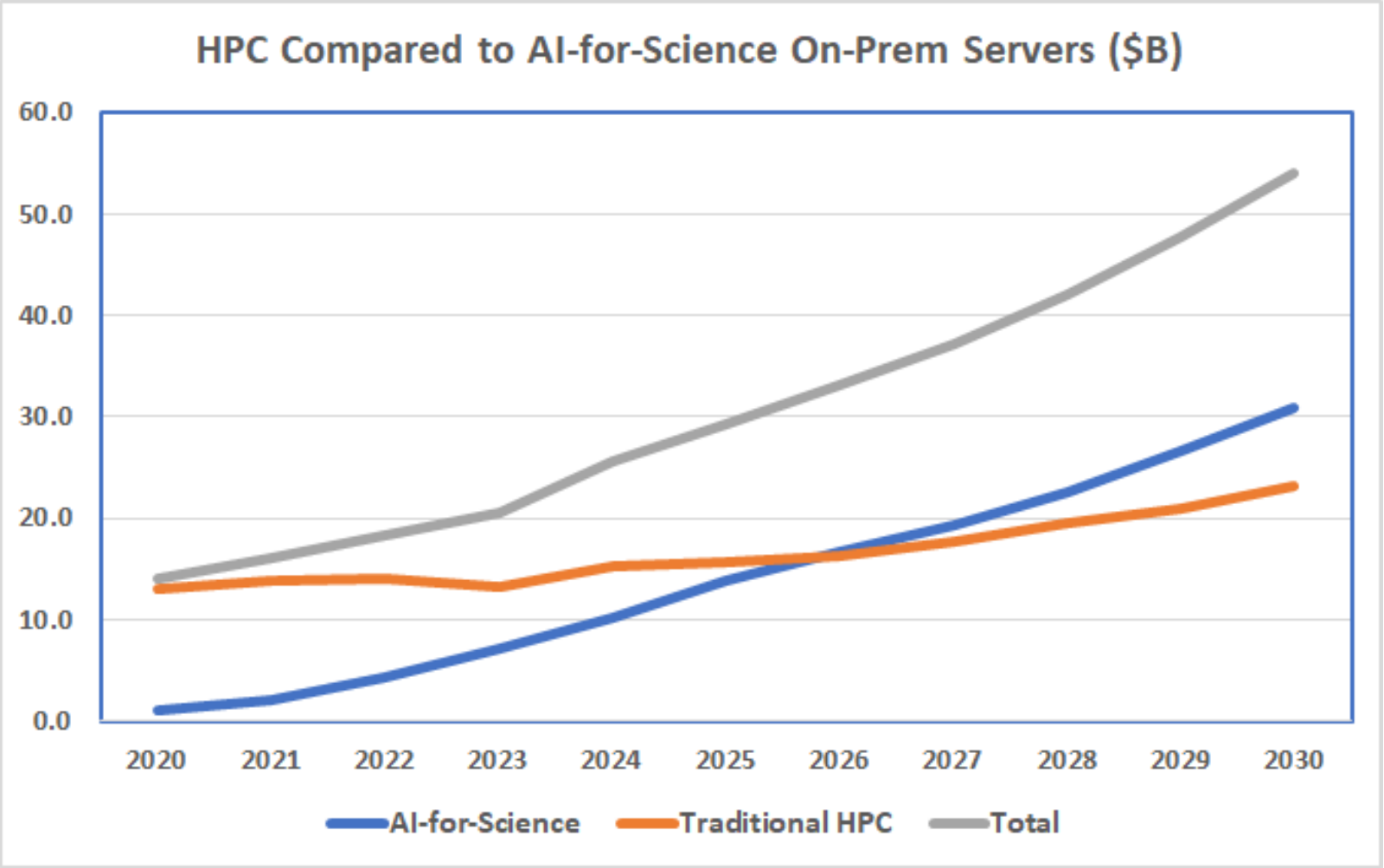
The purpose of this study was to gain a better understanding of the return on investments made by AI/HPC across critical science and engineering verticals. Key goals included creating a picture of investment expectations, snapshots of current integration projects in progress, budget allocations, provisioning methods, future plans, and challenges involving achieving desired goals in these areas.

Highlights from the study:

- **75.8% of the sites felt that their AI projects met or exceeded expectations**
- **74.9% of respondents indicated plans to moderately or significantly expand generative AI to support HPC workloads**
  - Less than 3% expect to contract their use of AI, none of which would characterize that contraction as significant
- **Roughly 40% of respondents are already using agentic AI models**
- **However, technical challenges continue to bring hesitation when it comes to broad adoption:**
  - Hallucinations, lack of explainability, and integration complexity are persistent concerns
  - This signals a transition from reactive adoption to more measured, application-specific onboarding

Traditional HPC & AI-for-Science

AI-for-Science servers are projected to reach $31 billion by 2030
Total HPC & AI-for-Science to reach $54 billion by 2030



AI-for-Science Servers by Region

North America is projected to grow at a high rate 18.5% CAGR
Followed by Europe at 17.4% CAGR

Hyperion Research AI-for-Science Server Forecast: By Region ($ Billions)								
	2024	2025	2026	2027	2028	2029	2030	CAGR 2025-30
North America	4.76	6.51	8.01	9.46	11.06	13.12	15.22	18.5%
EMEA	2.84	3.79	4.58	5.27	6.17	7.31	8.48	17.4%
Asia/Pacific w/o Japan	1.92	2.54	3.02	3.43	3.96	4.63	5.30	15.9%
Japan	0.56	0.74	0.88	1.00	1.14	1.32	1.49	15.2%
Rest-of-World	0.15	0.20	0.23	0.26	0.30	0.34	0.38	14.0%
Total Server Revenue	10.2	13.8	16.7	19.4	22.6	26.7	30.9	17.5%
Source: Hyperion Research, May 2026								

Note: these numbers do not yet include Genesis and European AI factories.

AI-for-Science Servers by Verticals

Government Lab is the leading sector: over $12 billion in 2030

Hyperion Research AI-for-Science Server Forecast: By Verticals ($ Billions)								
	2024	2025	2026	2027	2028	2029	2030	CAGR 2025-30
Bio-Sciences	1.30	1.76	2.10	2.36	2.67	3.06	3.43	14.2%
CAE	0.77	1.04	1.24	1.39	1.57	1.80	2.02	14.2%
Chemical Engineering	0.10	0.14	0.16	0.18	0.21	0.24	0.27	14.2%
DCC & Distribution	0.46	0.62	0.74	0.83	0.94	1.08	1.21	14.2%
Economics/Financial	0.36	0.52	0.62	0.69	0.79	0.90	1.01	14.2%
EDA	0.15	0.21	0.25	0.28	0.32	0.37	0.41	14.2%
Geosciences	0.18	0.24	0.28	0.32	0.36	0.41	0.46	14.2%
Defense	1.34	1.86	2.28	2.67	3.15	3.75	4.38	18.7%
Government Lab	3.66	4.79	5.96	7.19	8.69	10.60	12.63	21.4%
University/Academic	1.58	2.15	2.56	2.88	3.25	3.73	4.18	14.2%
Weather	0.23	0.31	0.36	0.41	0.46	0.53	0.60	14.2%
Other	0.10	0.14	0.17	0.19	0.21	0.25	0.28	14.2%
Total Server Revenue	**10.2**	**13.8**	**16.7**	**19.4**	**22.6**	**26.7**	**30.9**	**17.5%**

Source: Hyperion Research, May 2026

2026 MCS Study Findings: AI Strategies

94.8% are using or will soon be using AI/LLMs
But 8.2% plan to stop using them

Characterizing AI/LLM Strategy	
Q. Which of the following best describes your generative AI/LLM strategy?	
Strategy Characterization	**Percentage of Sites**
We use generative AI/LLMs today and plan to continue using it over the next 12-18 months	71.1%
We do not use generative AI/LLMs today but plan to start in the next 12-18 months	15.5%
We use generative AI/LLMs today but plan to stop using it within the next 12-18 months	8.2%
We do not use generative AI/LLMs today and do not plan to within the next 12-18 months	5.2%
n = 97	
Source: Hyperion Research, 2026	

Study Findings: Barriers to AI Growth

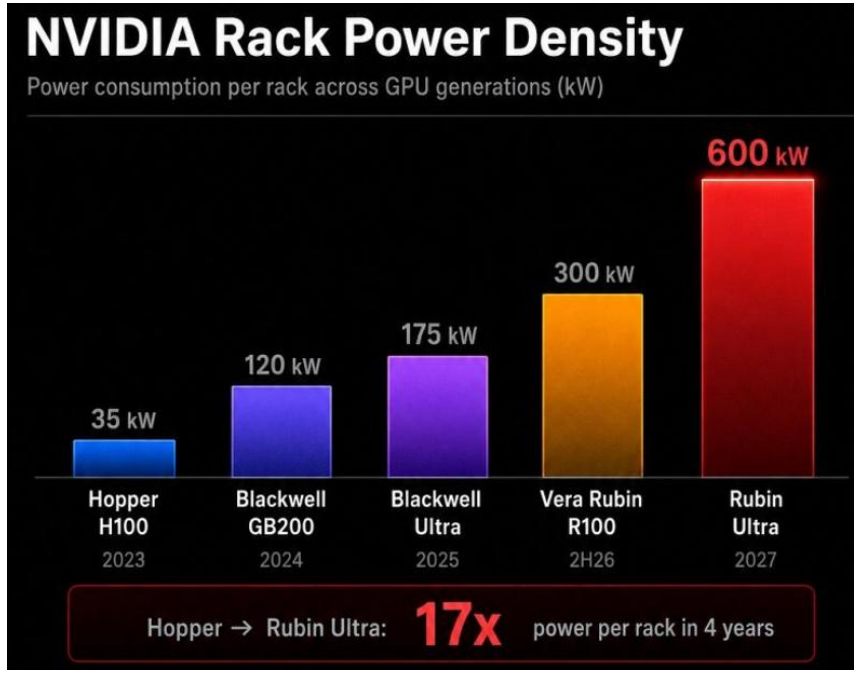
Expertise, quality of training data, dealing with complexity and high costs are the top barriers

Barriers to AI Capability Growth	
Q. Are any of the following barriers to furthering your AI capabilities? Please select all that apply:	
Barrier	**Percentage of Sites**
Level of in-house AI expertise	36.7%
Quality of available training data	35.0%
Complexity with integrating AI models into existing HPC workloads	30.0%
High/uncertain development costs	28.3%
Concerns with technical issues (i.e., scaling, hallucinations, ease of maintenance)	21.7%
High/uncertain operational costs	20.8%
Scale of available training data	19.2%
Access to specialized hardware	17.5%
The technology is moving too fast for credible assessment of value	17.5%
Lack of demonstrated return on investment	16.7%
Level of AI vendor or 3rd party expertise	10.8%
Access to specialized software	10.0%
Confusion/uncertainty with vendor selection	8.3%
Other (please specify)	3.3%
n = 120	
Notes: Respondents could select multiple options	
Source: Hyperion Research, 2026	

Market Changes:
Nodes Per System

Rack Power Requirements

Drives major increases in the price per rack



WHAT 600kW PER RACK MEANS FOR INFRASTRUCTURE

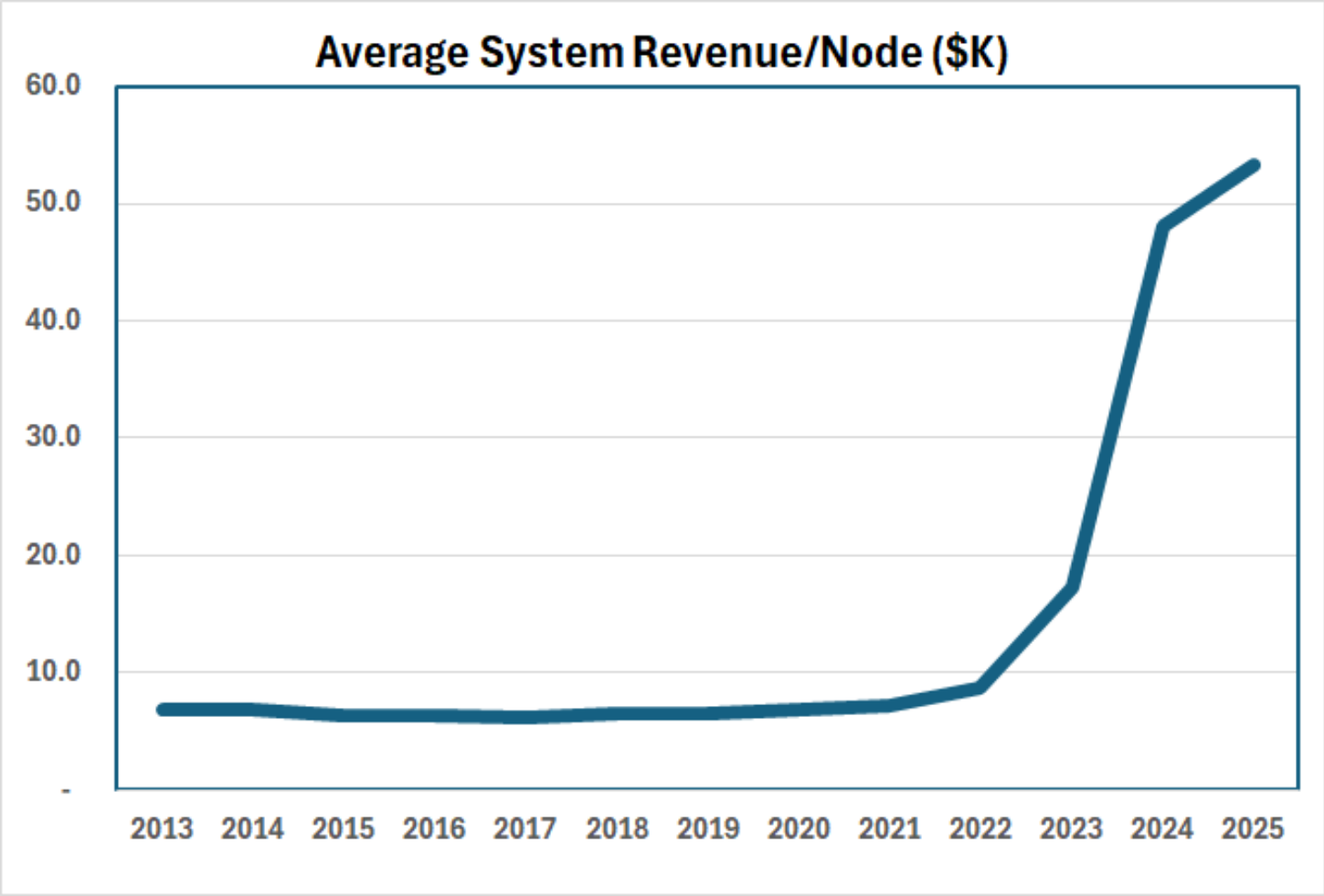
	CONSTRAINT	HOPPER ERA (35kW)	RUBIN ULTRA (600kW)
❄	Cooling	Air Cooled	Direct-to-Chip Liquid Only
⚡	Power per 10K racks	350 MW	6,000 MW (6 GW)
	Transformer demand	Baseline	~17x baseline

The bottleneck isn't chips. It's everything downstream: cooling, transformers, grid connections, and time. Jensen said NVIDIA will be "supply constrained through the entire life of Vera Rubin." The constraint he's talking about isn't silicon. It's infrastructure.

Prejean Consulting | May 2026

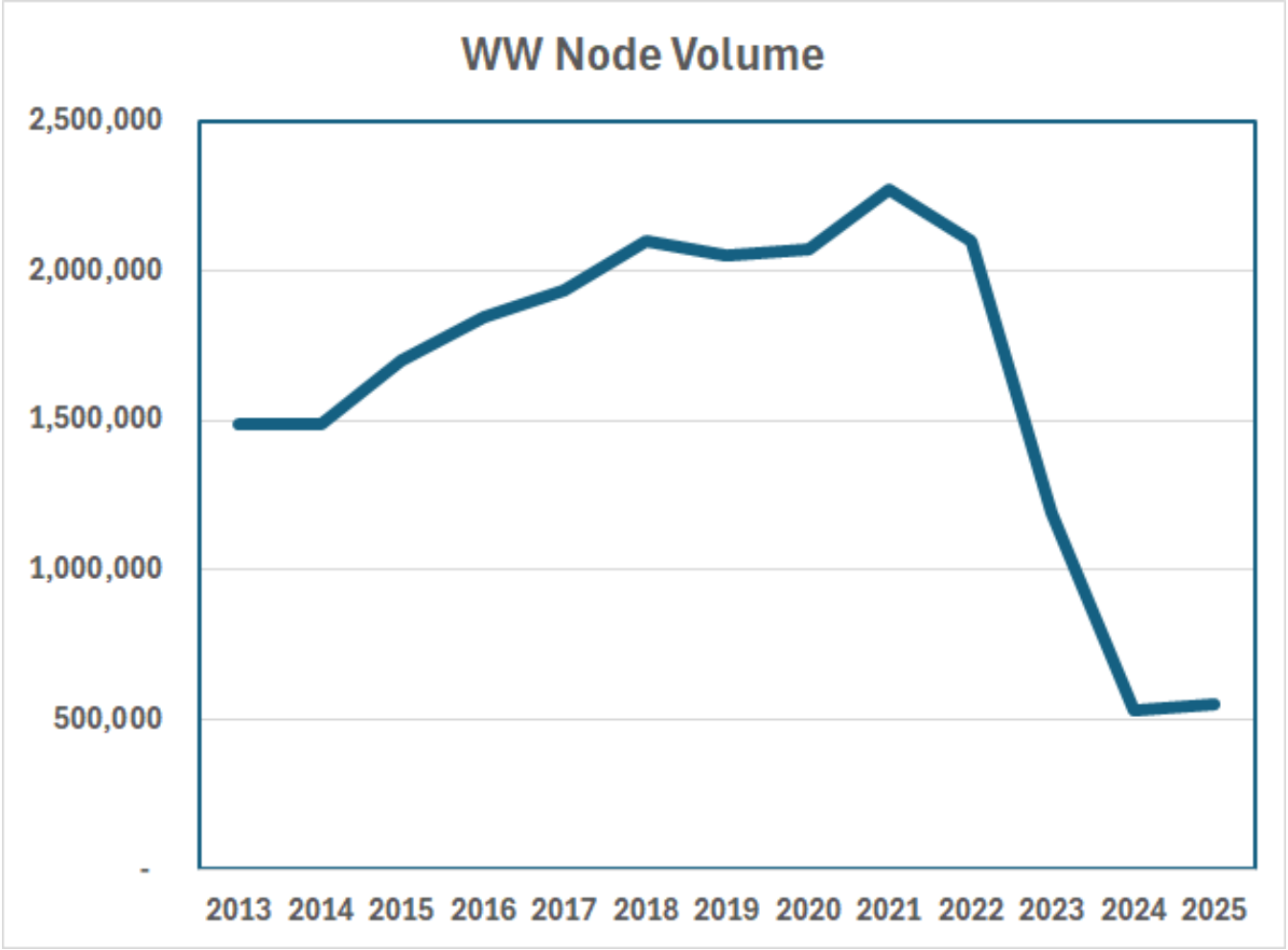
A Major Market Shift: High Node Costs

Driven by more systems having GPUs, higher GPU costs and higher node costs (memory, liquid cooling, higher power)



A Major Market Shift: Fewer Node Purchases

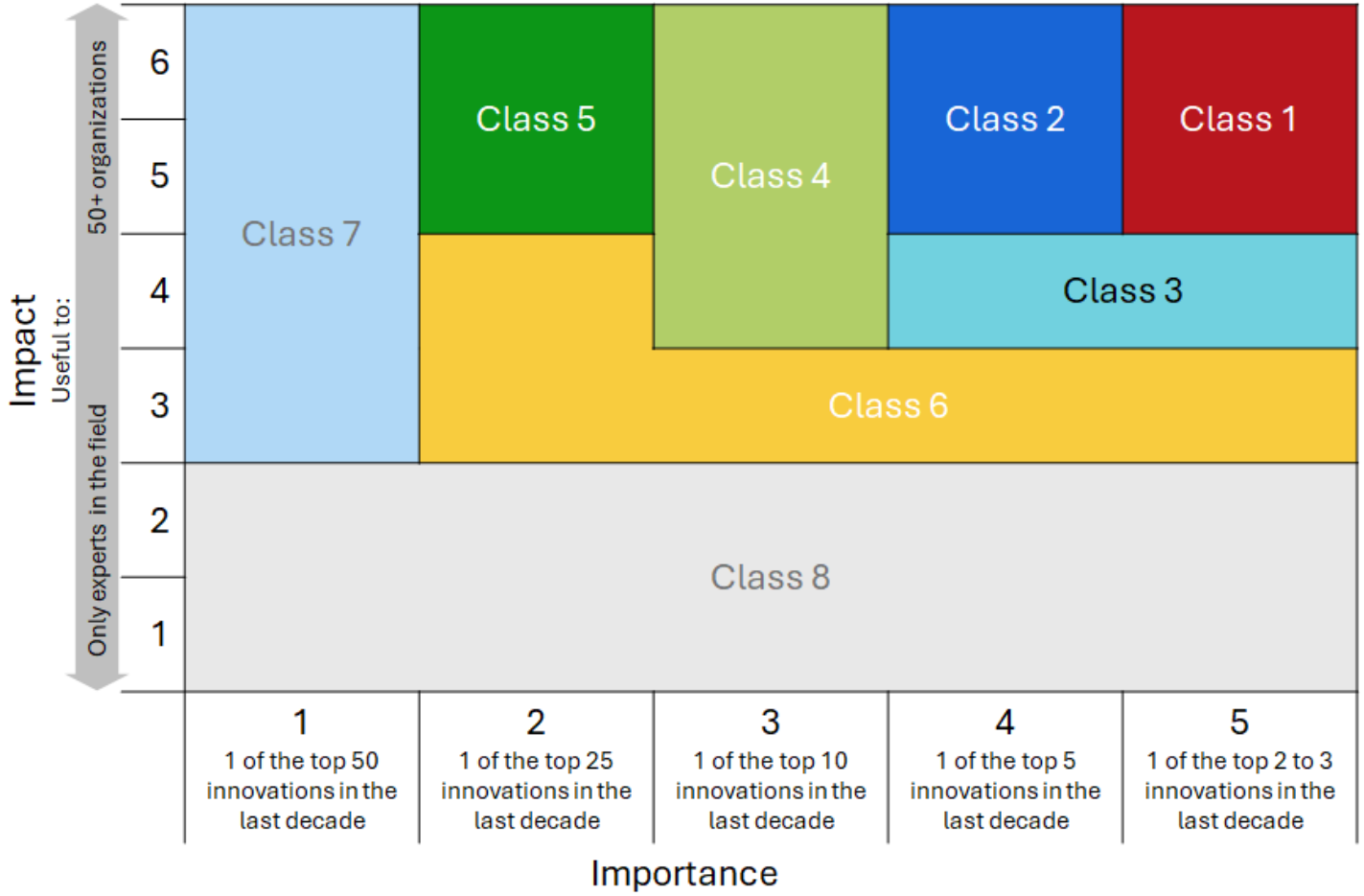
Driven by more systems having GPUs, higher GPU costs and higher node costs (memory, liquid cooling, higher power)



Measuring and Comparing Leadership Computing

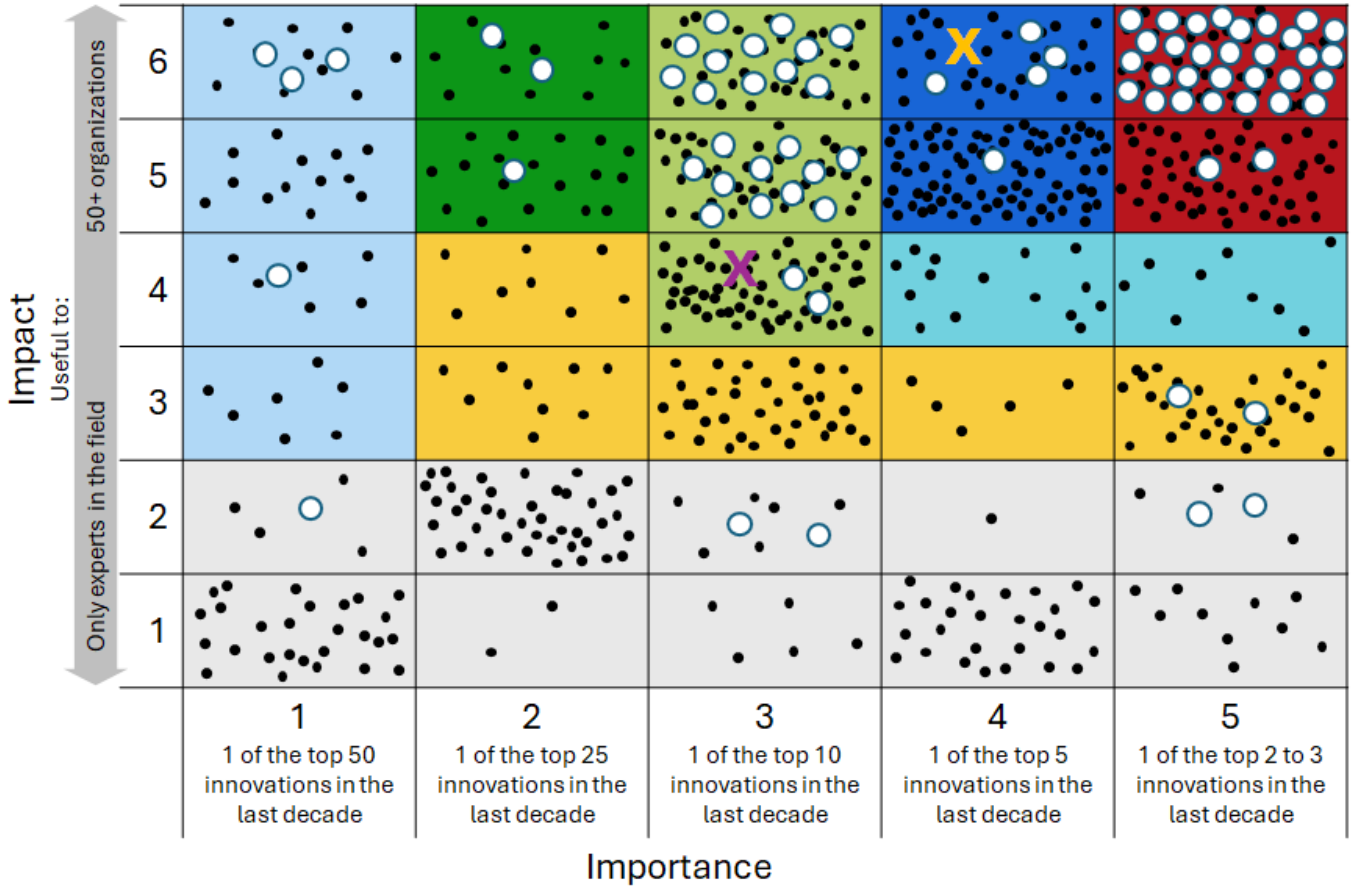
One Way to Show the Value of Leadership Computing

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



Showing the Value of Leadership Computing: RIKEN

An example from a 2024 study compared to 650 other projects
When applied to key societal goals, the increased value generated is impressive



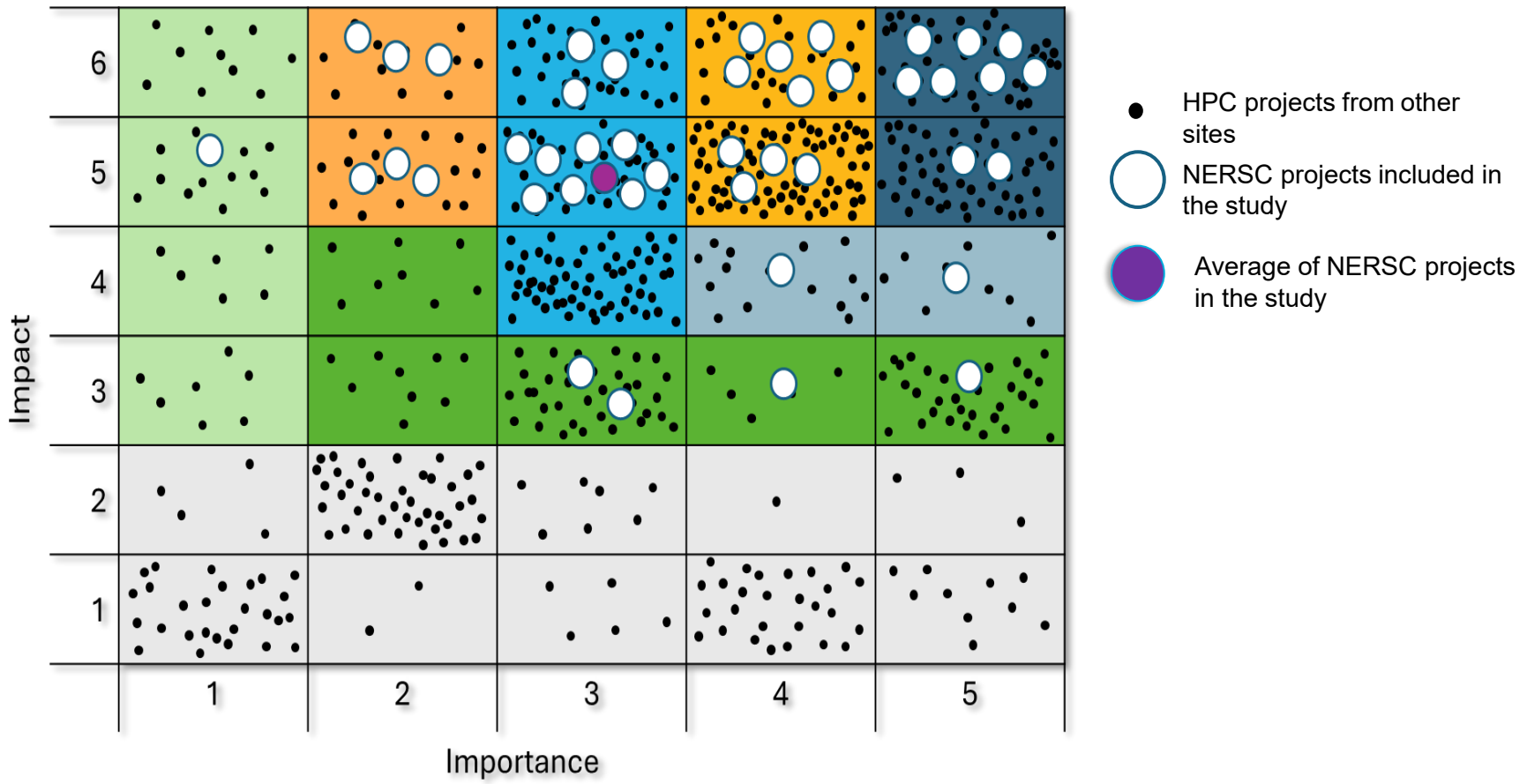
The K computer exceeded 9 times its investment.

If Fugaku delivers similar returns as the K computer, plus the added returns from addressing major societal issues, the returns could exceed 68 to 90 times the investment in the Fugaku system.

Showing the Value of Leadership Computing: NERSC

An example from a 2024 study compared to 650 other projects

Innovation Class Mapping: Showing Participating NERSC projects



Conclusions

Expecting strong growth, but there are some growing concerns…

- **2025 was a strong growth year**
  - AI-for-Science, GPUs, and cloud are high growth areas
  - New QC systems are being installed around the world
  - 2026 looks like a healthy growth year
- **New technologies are showing up large numbers:**
  - Agentic AI, 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 growing concerns around system costs, supply chains, power, AI impacts, data centers push-back, talent and political changes**

Next on the Agenda:

Elevator Speeches for QC, AI, and FP

Bob Sorensen

Elevator Speeches for QC, AI, and FP

ISC26 Market Update Briefing
June 2026

Bob Sorensen and Tom Sorensen

www.HyperionResearch.com
www.hpcuserforum.com

6th Annual Global QC Market Survey: Moving From Research Activity to Market Opportunity

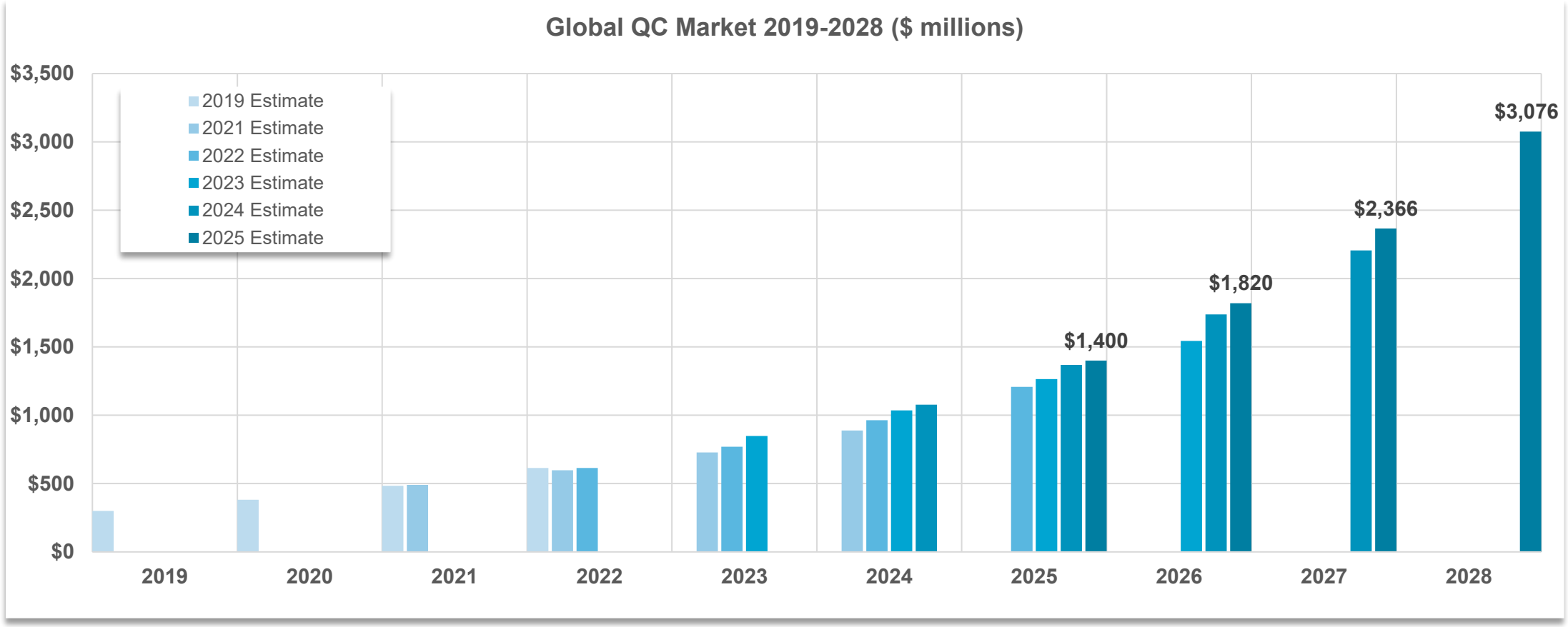
Bob Sorensen
Chief Analyst for Quantum Computing
Hyperion Research

QC Market Executive Summary/Highlights

- The global quantum computing (QC) market is estimated to have been worth $1.4 billion in 2025, with a projected average annual growth rate of 30%, driving the global market to $3 billion in 2028
  - Based on survey results from 116 QC experts, representing 99 different QC supplier companies
  - 53% of respondents had headquarters in North America, 29% in Europe, 17% in Asia/Pacific
- The QC hardware market will start a notable shift towards on-premises installations, reaching $1.2 billion in 2028 revenues
  - QC hardware for on-premises sites will be the largest segment of the market (29%) in 2028
  - QC hardware to support cloud access (10%) of the market in 2028
  - On-premises plus CSP software stack (25%) of the market in 2028
- Revenues aside, QC companies still rely on government-funded R&D and VC investment to support operations
  - 52% of companies received government funding in 2025, 34% from VCs
- Most promising applications center on quantum-level simulations
  - Lead by computational chemistry (26%) and material science (22%)
  - Crypto at 16%: An issue of mindshare or seen as needed to verify PQE schemes?
  - Optimization and logistics (11%)
  - Prospects for AI/ML continue to decline: 10% this survey, down from 23% in 2022 survey
  - Science and engineering applications at 5%

Global QC Market Estimate: $1.4 billion in 2025

With an 30% annual growth rate out to 2028 to reach $3 billion



- Continued revenue growth driven by increased emphasis of on-premises installations of larger systems
- Awaiting a hockey stick when quantum advantage is clearly demonstrated (2028, 2029, 2030?)

QC Partnerships: With Government Research Organizations

63% of respondent companies have/had QC-related government partnerships in past 3 years

Option	Percent Selected
Access to government funding	63%
Access to government-funded QC research activities	38%
Access to leading-edge QC hardware development	37%
Explore the co-design of QC systems	29%
Explore the hybrid quantum/classical QC systems	29%
Explore key government QC use cases	28%
Access to leading-edge QC research in algorithms	25%
Access to leading-edge QC research in applications	25%
Explore key government QC applications	25%
Access to key advanced quantum computing experts	24%
Foster public attention	22%
Help develop quantum/classical hybrid algorithms	21%
Access to leading-edge QC software development	16%
Support for publication of QC-related research in key journals	12%
Access to key advanced classical computing experts	11%
Access to key advanced classical computing hardware	8%
Access to key advanced classical computing software	7%
Other (Please specify)	5%
Don't know/Not sure	1%

N = 64, Select all that apply

- 63% of respondent companies used access to government funding
- 38% used access to government-funded QC research activities
- QC integration issues on the rise
  - Co-design and hybrid each mentioned by 30% of respondents
- Little interest in government-centric classical capabilities
- Others included
  - Access to quantum testbeds
  - Access to cryogenic cooling systems
  - Adoption of QC hardware and software
  - Development of QC/HPC middleware

QC Partnerships: With QC End Users

62% respondent companies have/had QC end user partnerships in past 3 years

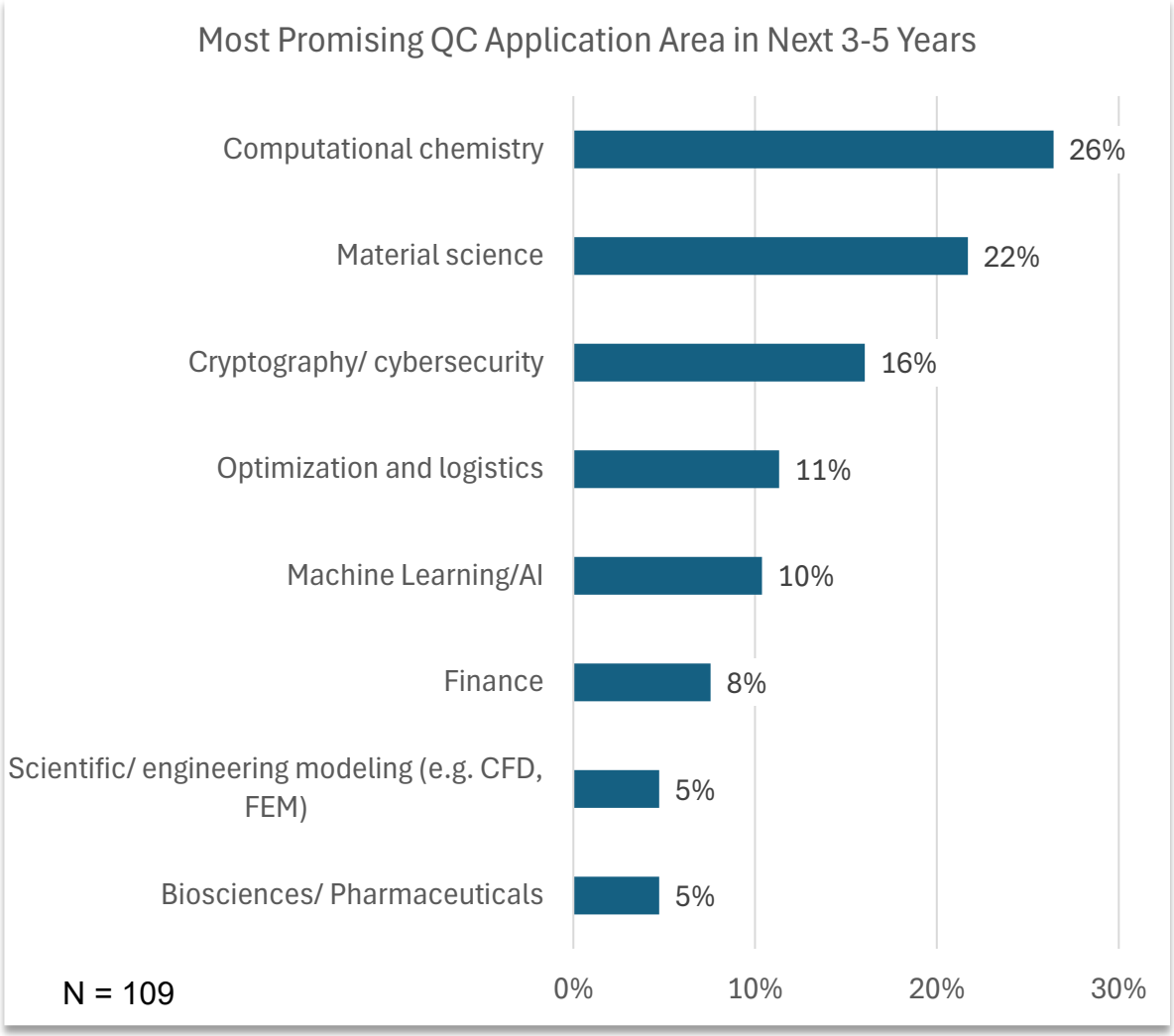
Option	% Selected
Explore new QC sector/vertical-specific QC-related opportunities	68%
Explore key performance gains over classical counterpart	42%
Explore QC sector/vertical-specific performance opportunities on existing classical workloads	39%
Explore QC/classical integration issues	36%
Field test/evaluate new QC hardware	32%
Field test/evaluate new QC software	30%
Access QC end-user QC expertise	29%
Foster public attention	29%
Establish sector-specific capabilities	22%
Encourage follow-on sales	20%
Access QC end-user classical IT expertise	8%
Other (Please specify)	5%
Don't know/Not sure	0%

N = 62, Select all that apply

- 68% engaged with end users to explore new sector/vertical specific QC-related opportunities
  - Plus, one in five seeking to establish sector-specific capabilities
- More than one third looking to explore QC/HPC integration issues with end users
- Others indicated efforts to co-develop a full stack quantum HW/SW solution, improve calibration routines, and build QEC workflows

Most Promising QC Applications in the Next 3-5 Years

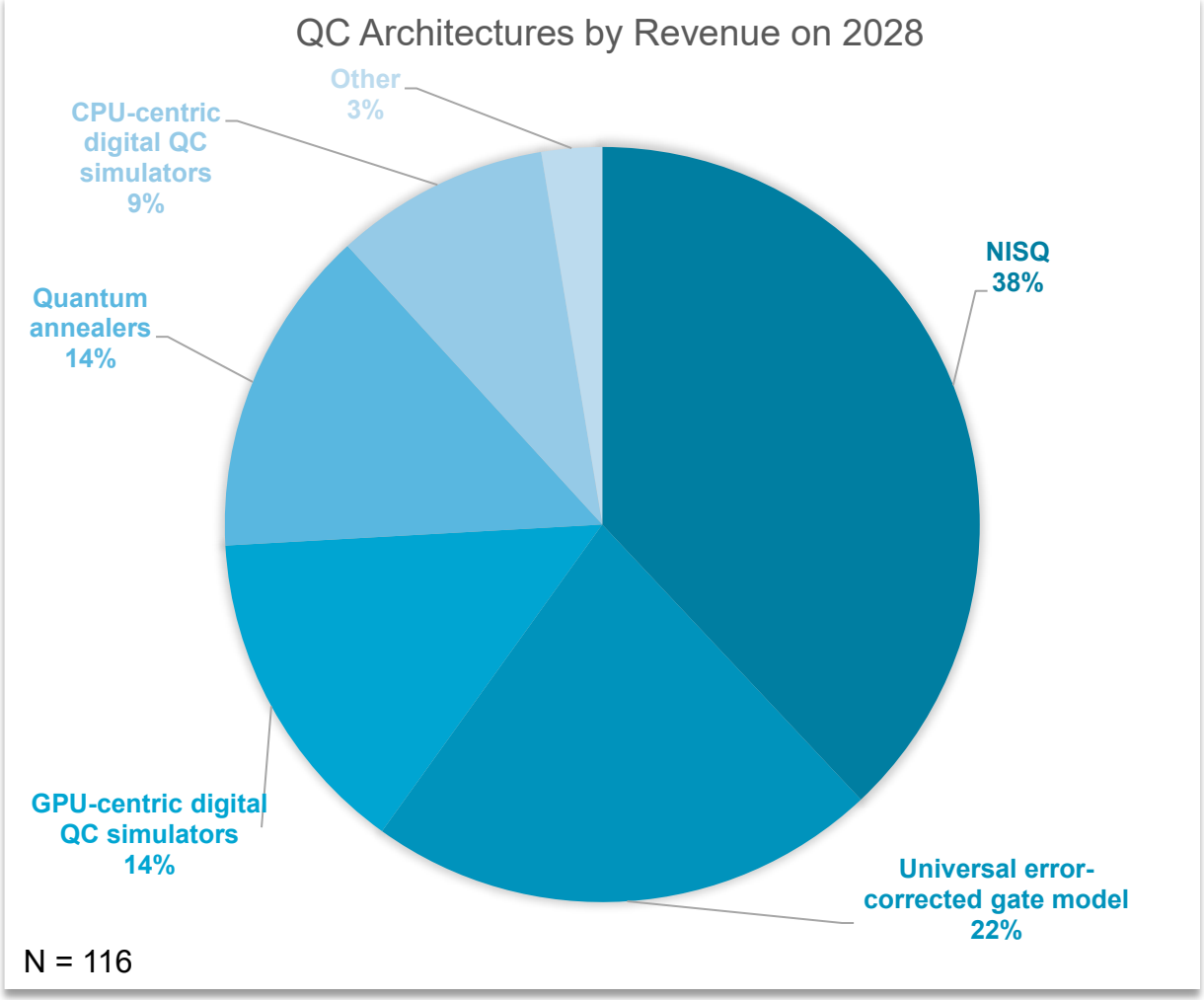
Quantum system simulations driving near-term QC applications



- Computational chemistry and material science top the list:
  - Combination represent nearly half of the market
  - Both use quantum systems to simulate quantum-level phenomena
- Crypto (16%): an issue of mindshare or seen as needed to verify post quantum encryption schemes?
- Prospects for AI/ML (10%) continue to decline
  - Was 23% in 2022 survey
- Scientific/engineering modeling at 5%
  - A dearth of algorithms or a perception that QC cannot handle traditional classical computational methods?
- Finance is currently an aggressive early adopter, but increasingly perceived as a niche market
  - Or a subset of optimization?

QC Market 2028: QC Architectures

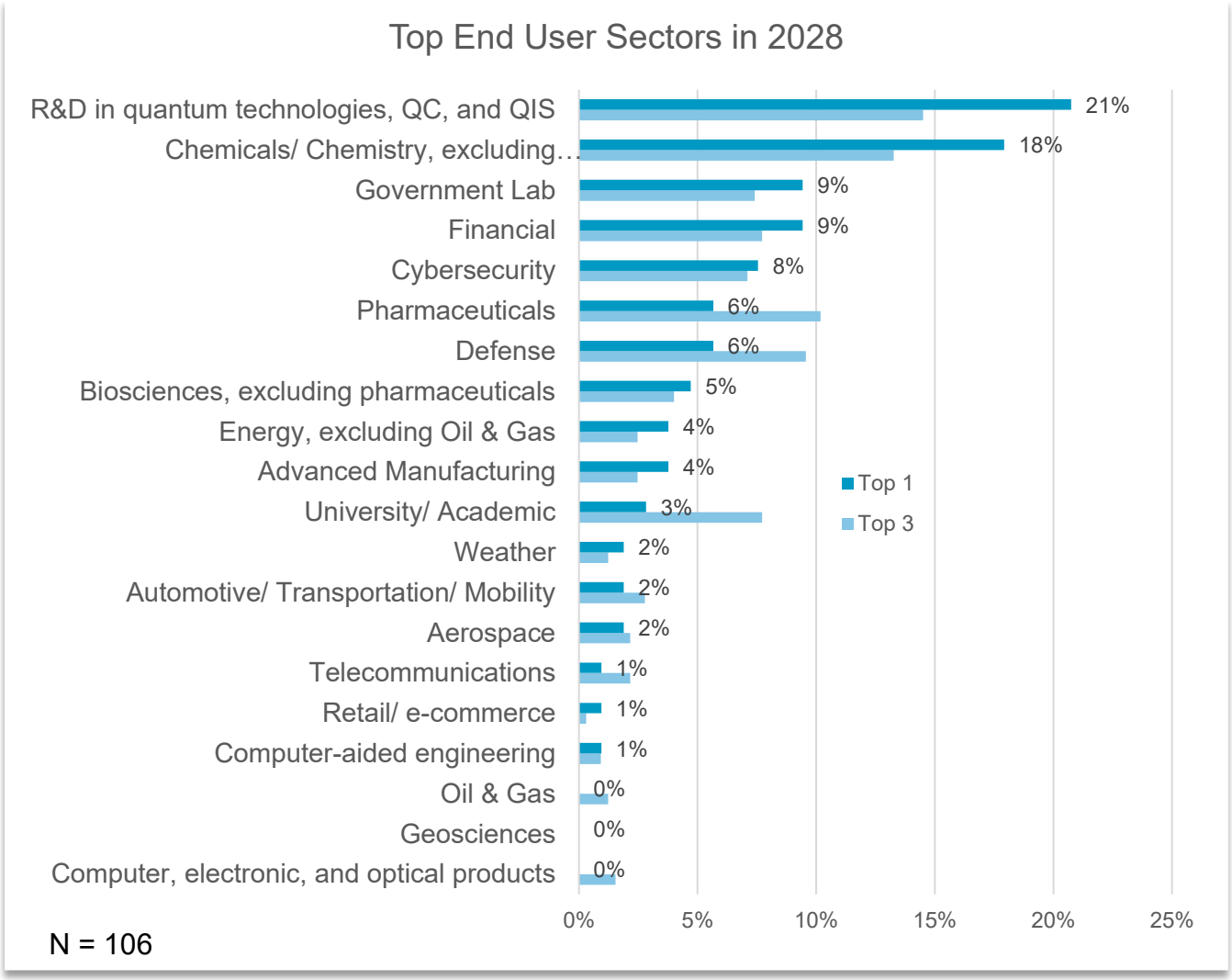
NISQ maintains lead, QC simulators still major element of QC architecture



- NISQ expected to dominate QC architecture in 2028 at 38% of total market revenues
  - Almost twice universal error corrected gate model alternative
- Digital simulators (CPU and GPU based) combine for almost one-quarter of hardware market
  - But GPUs are more preferred at twice CPU rate
  - Nearly the same market size as EC gate model
- Others included:
  - Digital annealers, quantum-inspired, classical, and analog

QC Market 2028: Top End User Sectors

QC R&D and chemicals on top, but broad applicability envisioned



N = 106

- Although nearly every sector choice deemed important by some, there are clear concentrations in key areas
- Most promising single sector is R&D in quantum technologies followed by chemicals/chemistry
  - Combined, selected by nearly 40% of respondents
- Government labs and defense combined selected by 15% of respondents
- Top 3 considerations broaden QC sector applicability
  - As a top three choice, academic sector goes from 3% to 8%, the greatest sectorial increase
  - Both defense and pharmaceuticals go from 6% to 10%

QC Market 2028: Primary End User Motivations

QC exploration and implementation lead drivers, classical concerns issues fading?

Option	% Selected
Explore relevant QC use case potential with no expectations of near-term advantage	47%
Develop in-house familiarization with QC skills with no expectations of near-term end use deployment	45%
Implement new algorithm(s) not possible on classical counterpart systems	44%
Engage with the QC vendor community for future activities	37%
Address concerns with future performance capabilities of classical computing systems	35%
Enable better real-time computational capabilities	22%
Realize faster turnaround time on existing classical counterpart systems	19%
Reduce overall computational power and cooling requirements	17%
Reduce overall computing systems costs	12%
Other	7%
Don't know/Not sure	9%

- QC exploration, QC familiarization, and implementation of new QC algorithms are key motivators
  - QC timeframe realities are well understood, driving QC end user interest
  - Likewise, QC vendor engagement on the rise
- QCs seen as addressing concerns with current classical performance falling from 51% (2025) to 35% (2028)
  - Reducing power, cooling, and cost remain minor considerations
- One in five looking at real-time QC compute opportunities

N = 116, Respondents were given the option of selecting all that apply.

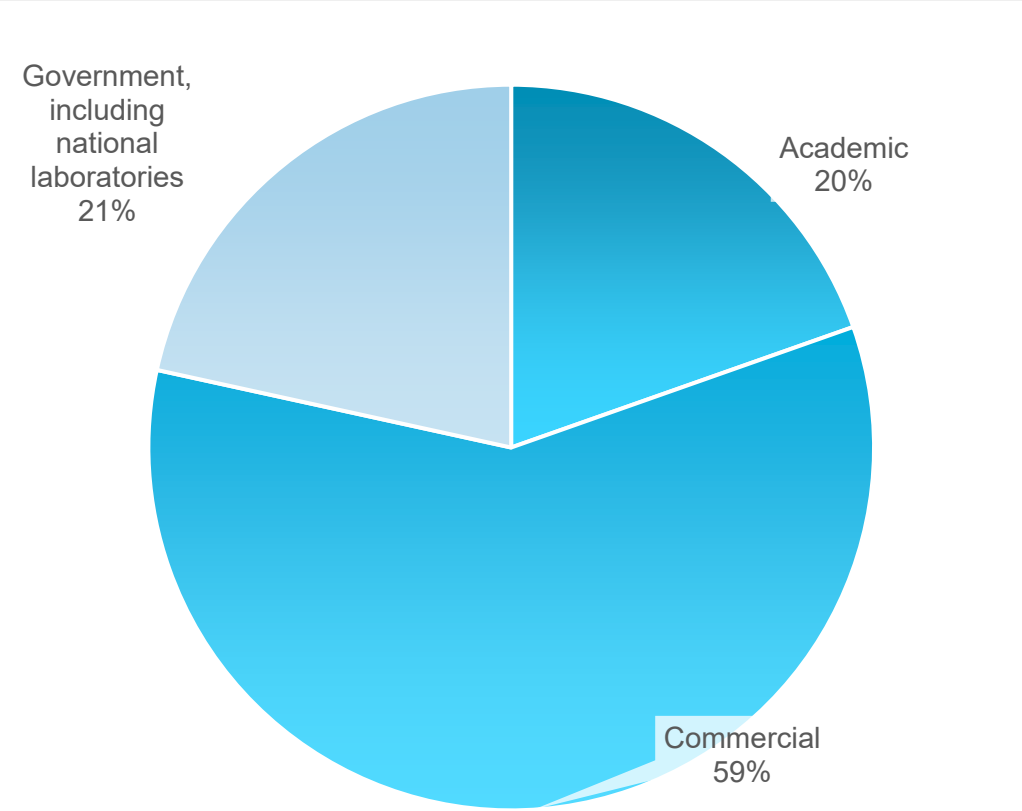
Recent Study Overviews: Gen AI ROI AI in the Cloud

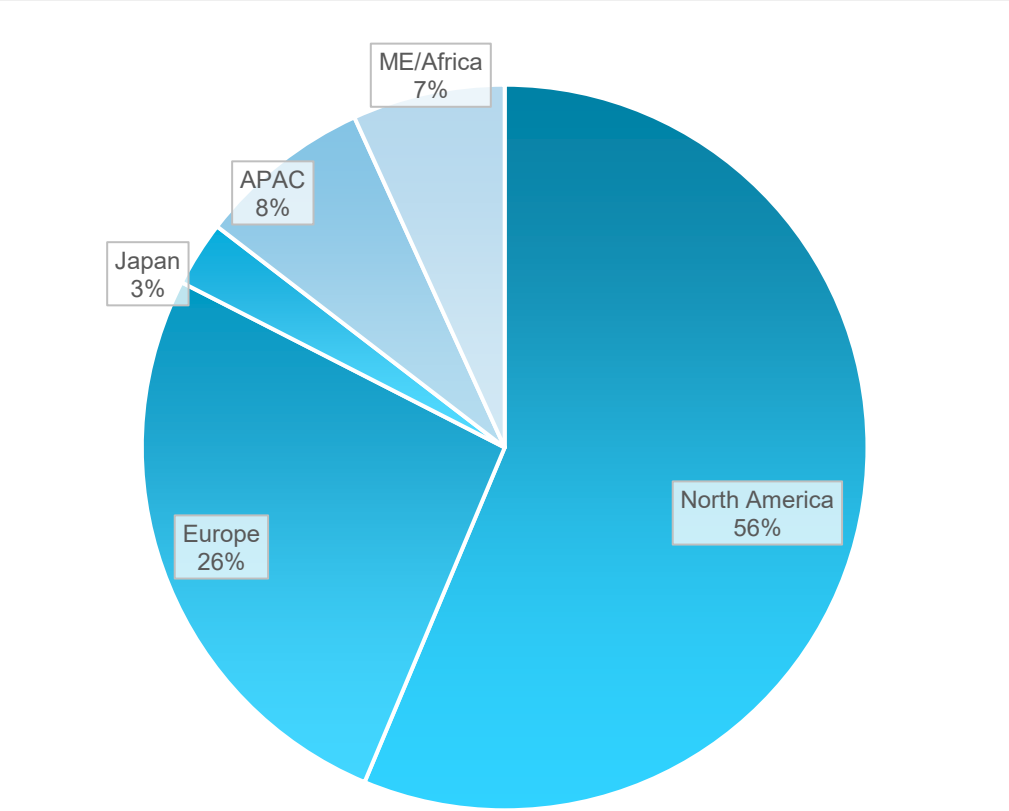
June 2026

www.HyperionResearch.com
www.hpcuserforum.com

Tom Sorensen

Study on Gen-AI Investments for HPC and Advanced Computing





Focus on commercial respondents, n=103

Key Findings

Generative AI models are most leveraged to support the front and back end of existing traditional HPC methods such as modeling and simulation

End Use	% Selected
Scientific data analysis	80.6%
Time series data analysis	61.2%
Text generation	60.2%
Code generation	56.3%
Synthetic data generation	54.4%
Image creation	43.7%
Audio or music generation	14.6%

Hyperion Research 2026, N=103

- The vast majority indicated the use of scientific data analysis applications
  - In line with the analytic and language generation strengths of current generative AI capabilities, data analysis and code/text generation were among the top 4 selected end uses
- Most respondents reported engaging in several activities linked to their generative AI models, demonstrating a high confidence and willingness to explore new options for their AI

Key Findings Continued

Roughly half of respondents expect a measurable monetary return on investment within 2 years with 30% expecting it in less than one year

Time	% Selected
Less than one year	29.1%
One year to less than two years	21.4%
Two years to less than three years	20.8%
Three years to less than four years	7.6%
Four years to less than five years	4.2%
More than five years	4.1%
Never	7.8%
Don’t know/ Not sure	5.0%

Hyperion Research 2026, N=103

- Nearly 30% expect a measurable monetary return within the year
- Over 70% of respondents expect this monetary return within a 3-year window

Key Findings Continued

Most users reported their AI integration exceeded cost expectations at least moderately, but they intended to continue investing in the technology

Cost Expectation	% Selected
Significantly more cost than expected	11.7%
Moderately more cost than expected	40.8%
Met expectations	34.0%
Somewhat less costly than expected	4.9%
Significantly less costly than expected	1.9%
Have not yet integrated gen-AI into our HPC workloads	3.9%
Don’t know/ Not sure	2.9%

Hyperion Research 2026, N=103

- Over half (52.4%) of respondents indicated integration being more costly than expected
- 34.0% felt that the costs met expectations
- 6.8% found that costs were lower than expected
- Demonstrates continued trust in ROI and efficacy of integration

Study Cloud/On-Premises AI Activities

AI activity in the cloud more prevalent than on-premises in every major category

On premises: exploring the range of potential AI performance enhancements	50%
Cloud: exploring the range of potential performance enhancements	64%
On premises: reaching out to AI hardware and software suppliers for information	30%
Cloud: Reaching out to cloud service providers for hardware and software information	35%
On-premises hardware procurement for AI activities	25%
Cloud-based hardware procurement for AI activities	30%
On-premises software procurement for AI activities	20%
Cloud-based software procurement for AI activities	34%
On-premises: standing up limited AI-integrated pilot programs	22%
Cloud: standing up limited cloud-based AI-integrated pilot programs	31%
On premises: testing/assessing AI-integrated workload performance	25%
Cloud: testing/assessing cloud-based AI-integrated workload performance	39%
On premise: running production level AI-enabled workloads on-premises	30%
Cloud: running production level AI-enabled workloads in the cloud	43%

N= 103, Respondents could select all options that apply

Source: Hyperion Research 2026

Highlights of Recent Studies

- LLM Study: C*urrently available*
  - Conducted across industry verticals, governmental organizations, and academic institutions to capture the current LLM activity and applications
- AI in the Cloud Study: *Currently available*
  - Captures insights at the intersection of AI usage and cloud resources
- HPC End-User Multi-Client Study 2025: *Currently available*
  - The seventh edition of a comprehensive study that surveys many HPC customer sites worldwide to create a detailed profile of HPC activities
- End User Inferencing: *Currently available*
  - Targeted towards the inferencing side of production and near-production integration of advanced AI/LLM
- AI Investments and ROI: *Currently available*
  - Explore investment expectations, current integration progress, bullet allocations, and current/expected return on investment for AI integration
- AI/HPC Metrics and Adoption Standards: Coming Summer 2026

Sneak Peak at Upcoming Survey

- What metrics are used within your organization for measuring the success of gen-AI integration into a key workload?
- How would you describe your organization's level of satisfaction with its current ability to meet desired gen-AI metrics standards?
- What are the main targets for agentic AI in your organization?
- What gen-AI capabilities, currently or in the near future, are considered least viable for your current compute workloads? (Select all that apply)
- What compute resources are used for supporting your gen-AI inferencing requirements? (Select all that apply)

FP64 vs FP4
An Evolving Debate

ISC26

www.HyperionResearch.com
www.hpcuserforum.com

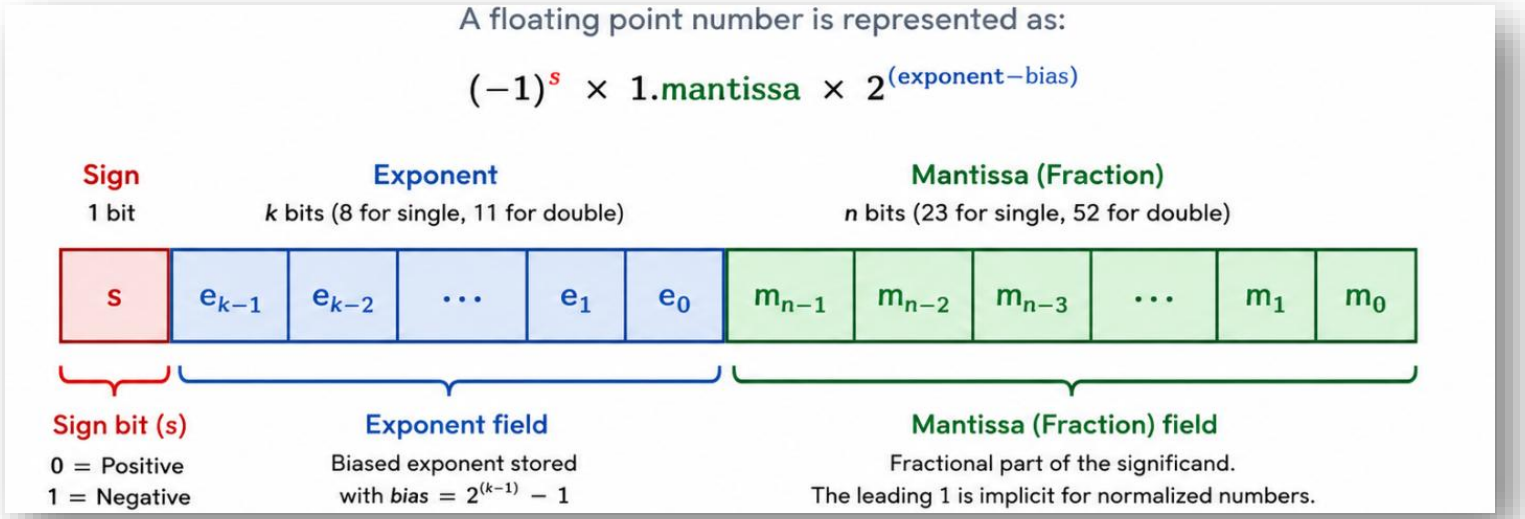
Bob Sorensen

In the Beginning: Chaos

- Burrough B5500 (1961)
  - 47 bit words
  - Bit 0 - always 1, Bit 1 – mantissa sign, Bit 2 – exponent sign (radix = 8), Bit 3-8 – unsigned exponent, Bit 9-47 - unsigned mantissa (integer)
- CDC 6000 (1964)
  - 59 bit words
  - Bit 0 – mantissa sign, Bit 1-11 – exponent, Bit 12-69 unsigned mantissa (integer)
  - Two's compliment format for exponent and mantissa
- IBM System 360/370 (1964)
  - 64 bit words (double precision)
  - Bit 0 - mantissa sign, Bit 1-7 – (exponent in excess – 64) (radix = 16), Bit 8-31, unsigned mantissa
  - Bits 32-63, optional add on mantissa
- Vax PDP-11 (1977)
  - 32 bit words
  - Bit 0 - mantissa, Bits 1-8 (exponent in excess – 128) (radix =2), Bits 9-15, 16-31 unsigned mantissa fraction

Let There Be Light:

- In I985, along came IEEE 754, a technical standard developed by the IEEE
- Specified interchange and arithmetic formats, along with methods for binary and decimal floating-point arithmetic in computer programming environments, including handling of exception conditions
- Addressed the need for portability and consistency in floating-point computations by defining precise representations and operations that ensure predictable results
- But just as important, defined binary and decimal formats, arithmetic operations (such as addition, subtraction, multiplication, division, and square root), rounding modes, and exception handling for overflow, underflow, and invalid operations



- E.g. FP64 accommodates values roughly from 2.2×10^{-308} to 1.8×10^{308} with about 15 decimal digits of precision
- **It took over a decade for this standard to move from concept to reality**

A Quick Example of Thoughtful IE^3 754 Standards: Rounding

Or: Why it took a decade

- Round To Nearest (Ties to Even) — *Default*
  - Rounds to the nearest representable value
  - If the exact value falls exactly halfway between two representable values (a tie), it rounds to the value with an even least significant bit (i.e., ending in 0)
- Round To Nearest (Ties Away from Zero)
  - Also rounds to the nearest representable value
  - However, if the value falls exactly halfway, it rounds to the value further from zero (i.e., rounds positive numbers up and negative numbers down)
- Round Toward Zero (Truncation):
  - Drops all extra bits and chooses the representable value closest to, but not greater in magnitude than, the exact value (towards 0)
- Round Toward Positive Infinity (Round Up):
  - Rounds up to the closest representable value that is strictly greater than the exact value (towards +∞)
- Round Toward Negative Infinity (Round Down):
  - Rounds down to the closest representable value that is strictly less than the exact value (towards –∞)

Along Comes FP4

Two main and evolving standards for FP4 targeted for current AI-centric hardware

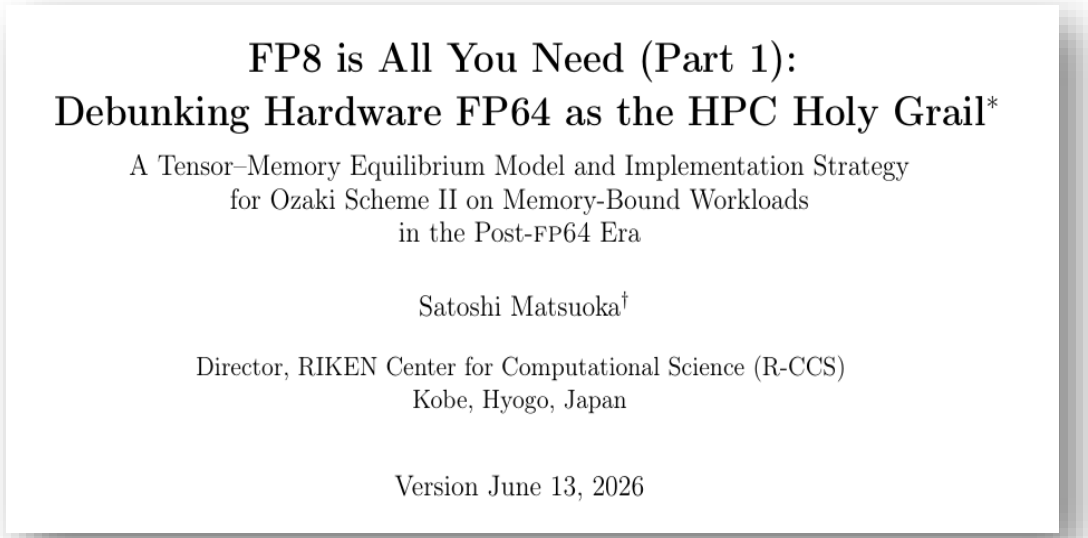
- NVIDIA FP4: Introduced with NVIDIA Blackwell GPUs
  - Uses an E2M1 layout (1 sign bit, 2 exponent bits, 1 mantissa bit) in 16-element blocks
  - Two-Level Scaling: Employs fine-grained E4M3 (FP8) scaling factors per 16-value block and a second-level FP32 scalar for the entire tensor
- Microscaling FP4: Part of the Open Compute Project standard, supported by AMD (CDNA) and NVIDIA, which uses an E2M1 layout in 32-element blocks
  - Uses Shared Scaling: Groups of 32 elements share a common E8M0 (8-bit exponent, no mantissa) scaling factor
- Sample FP4 Rounding Modes:
  - For inferencing often uses round-to-nearest (RTN)
  - For training: stochastic rounding (SR)

    For example, 2.4: 40% chance of rounding to 3 and a 60% chance of rounding to 2

FP64 in an FP4 World

- NVIDIA GPUs use a specialized emulation approach to handle FP64 demands by leveraging high-throughput, low-precision tensor cores
- However, there are concerns with FP64 emulation that include IEEE 754 non-compliance
  - Data-dependent accuracy
  - Potential numerical instability in complex simulations
  - Failure to properly account for *Not a Number* errors, infinite numbers, or specific signed zero scenarios
- Although DGEMM (double-precision matrix multiplication) emulation is deemed effective, there are few production-ready solutions for more complex transcendental function (exponential, logarithmic, etc.) and trigonometric math functions (sin(x), cos(x), etc.)
- For its part, AMD primarily focuses on native FP64 in its MI430X
  - Consumes much more chip real estate and power than emulated counterpart (~16-64X)
  - AMD is ‘studying’ FP64 emulation but favors delivering the highest native FP64 GPU on the market

FP64 in an FP4 World or Vice Versa



FP8 is All You Need (Part 1):
Debunking Hardware FP64 as the HPC Holy Grail*

A Tensor–Memory Equilibrium Model and Implementation Strategy
for Ozaki Scheme II on Memory-Bound Workloads
in the Post-FP64 Era

Satoshi Matsuoka†

Director, RIKEN Center for Computational Science (R-CCS)
Kobe, Hyogo, Japan

Version June 13, 2026

https://arxiv.org/pdf/2606.06510

From NVIDIA Technical Blog*

- *At the same time, dedicated FP64 vector performance remains critical for scientific applications that are not dominated by matrix kernels*
  - In these cases, performance is constrained by data movement through registers, caches, and high-bandwidth memory (HBM) rather than raw compute.
  - A balanced GPU design therefore provisions sufficient FP64 resources to saturate available memory bandwidth, avoiding over-allocation of compute capacity that cannot be effectively utilized

** https://developer.nvidia.com/blog/inside-the-nvidia-rubin-platform-six-new-chips-one-ai-supercomputer/*

- The IEEE 754 FP standard is not just about matrix multiplies that gives high fidelity results
  - It is a way to ensure portability and consistency in calculations across a wide range of numerical rules and practices
- **How long to resolve these issues with FP4 to the satisfaction of the scientific and engineering community? (2036?)**

Questions?

bsorensen@hyperionres.com

tsorensen@hyperionres.com

Perspectives on HPC-AI Cloud, Storage, Interconnects, and Sustainability

ISC26 Market Update Briefing
June 2026

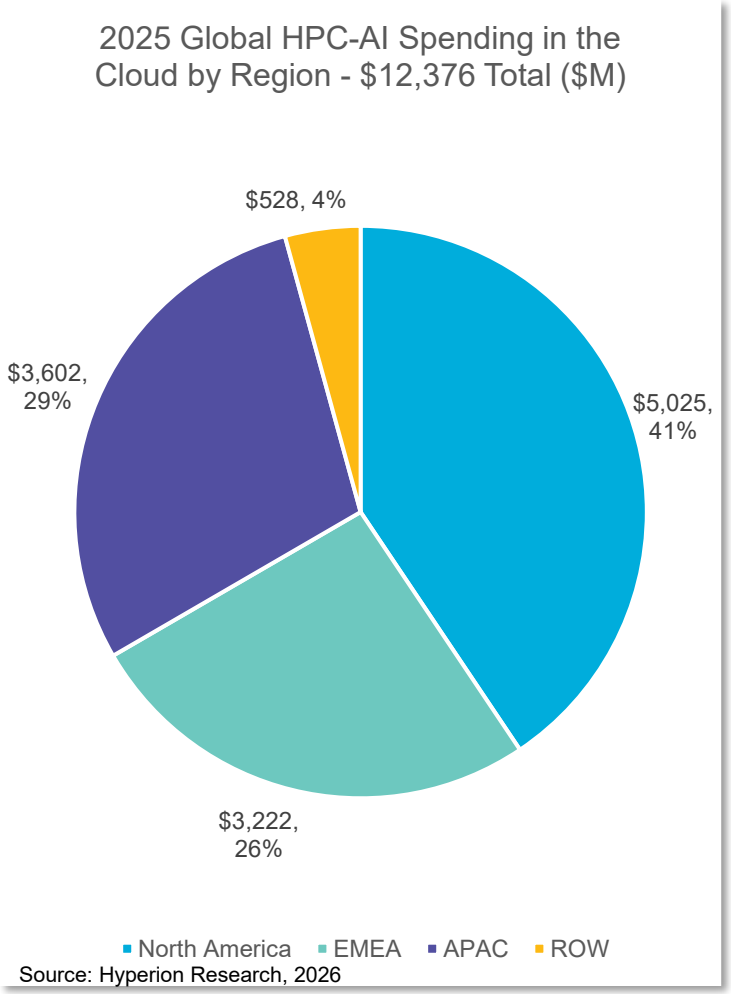
www.HyperionResearch.com
www.hpcuserforum.com

Mark Nossokoff and Jacklyn Ludema

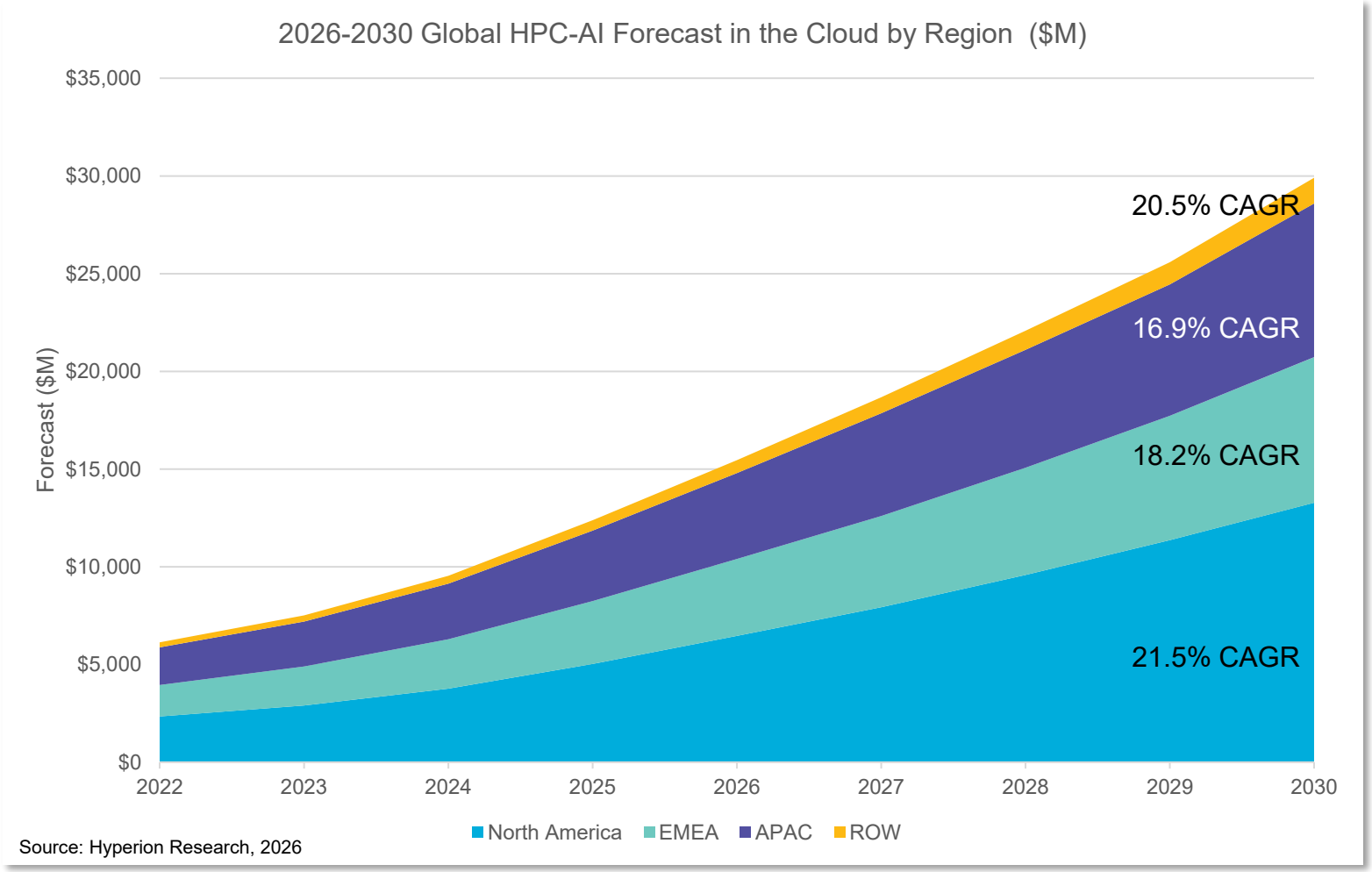
Cloud

2026-2030 HPC-AI Cloud Forecast

North America leads in cloud spending and projected growth



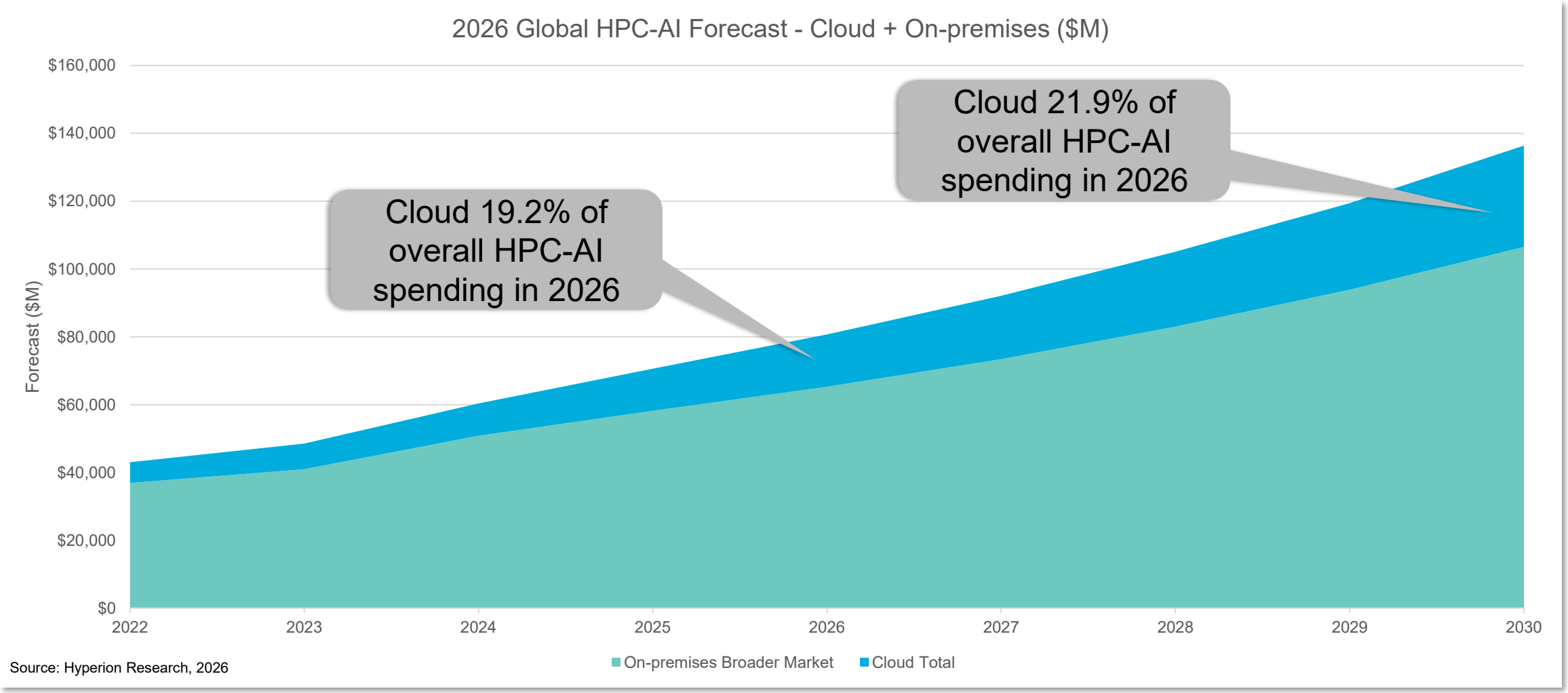
Source: Hyperion Research, 2026



Source: Hyperion Research, 2026

2026-2030 HPC-AI Cloud Forecast

Cloud growth projected to be 19.3% 5-year CAGR to $29.9B in 2030

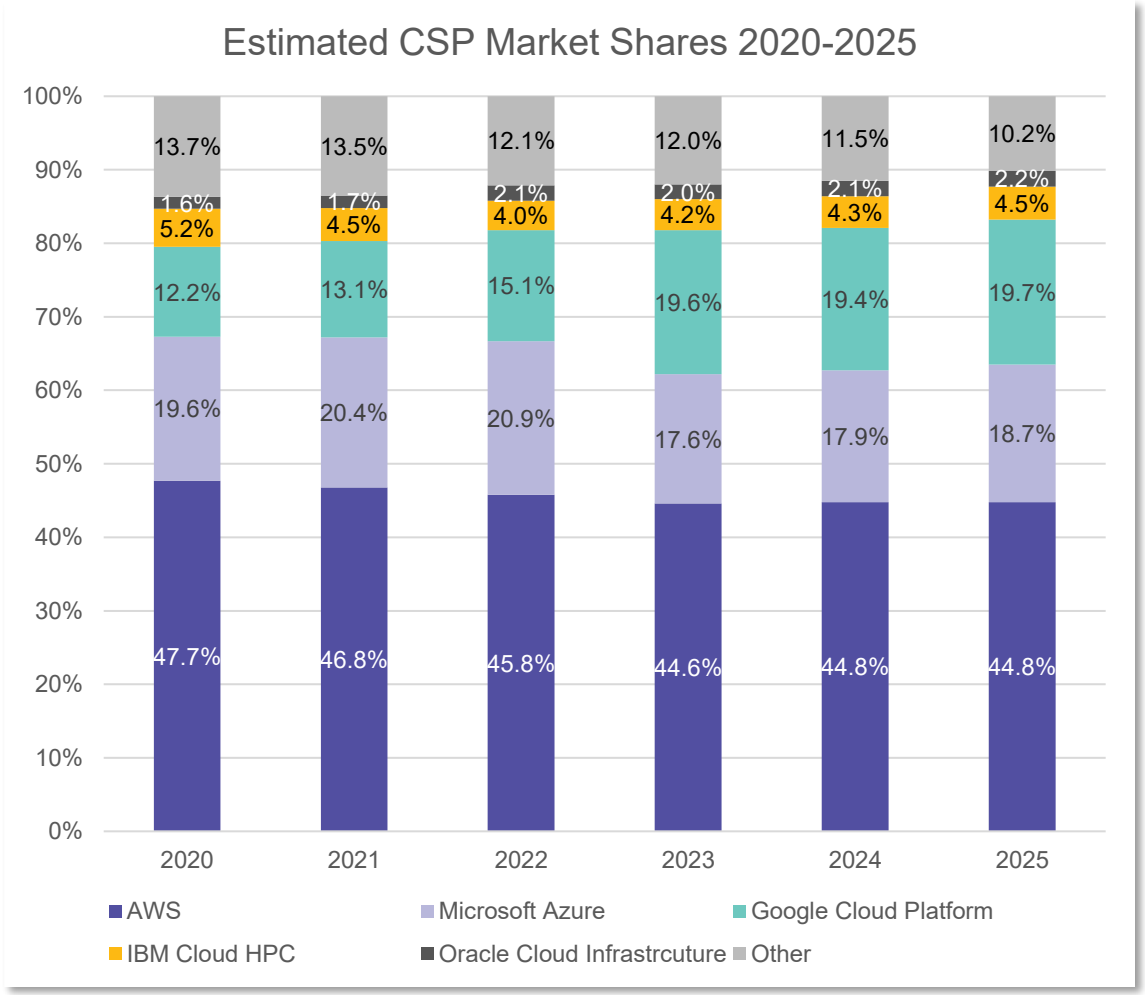


Source: Hyperion Research, 2026

Estimated CSP HPC-AI Market Shares

Market share order remains the same while all CSPs increased revenues

- **Leading CSPs retained or grew share**
- **“Other” (includes neoclouds) grew revenue, but at a lesser pace than the leading CSPs, decreasing its overall market share**



Source: Hyperion Research, 2026

Scientific Computing Cost, Value, and ROI Model

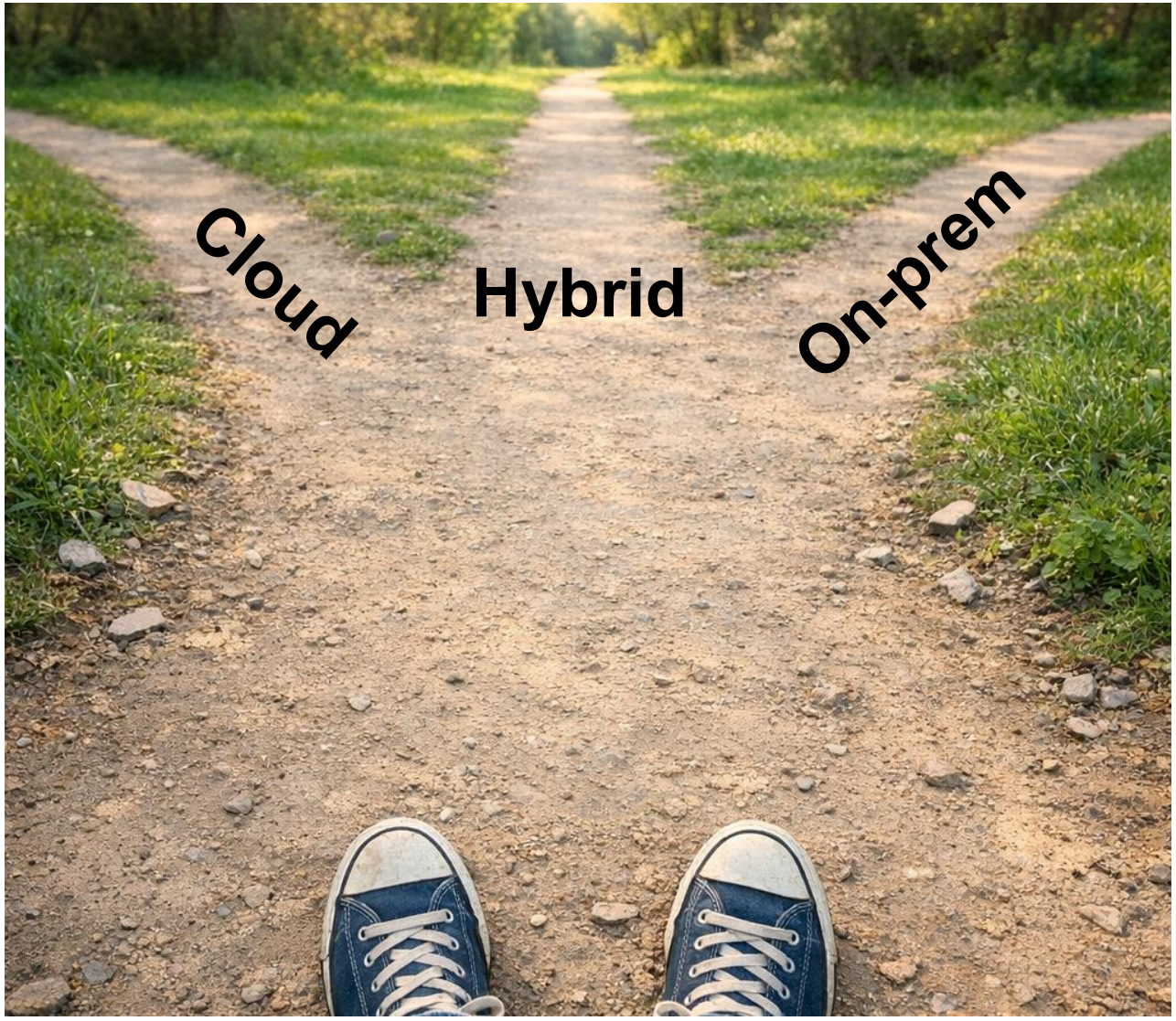
PNNL

May 2026

www.HyperionResearch.com
www.hpcuserforum.com

Jaclyn Ludema & Mark Nossokoff

The Project Compute Environment Decision



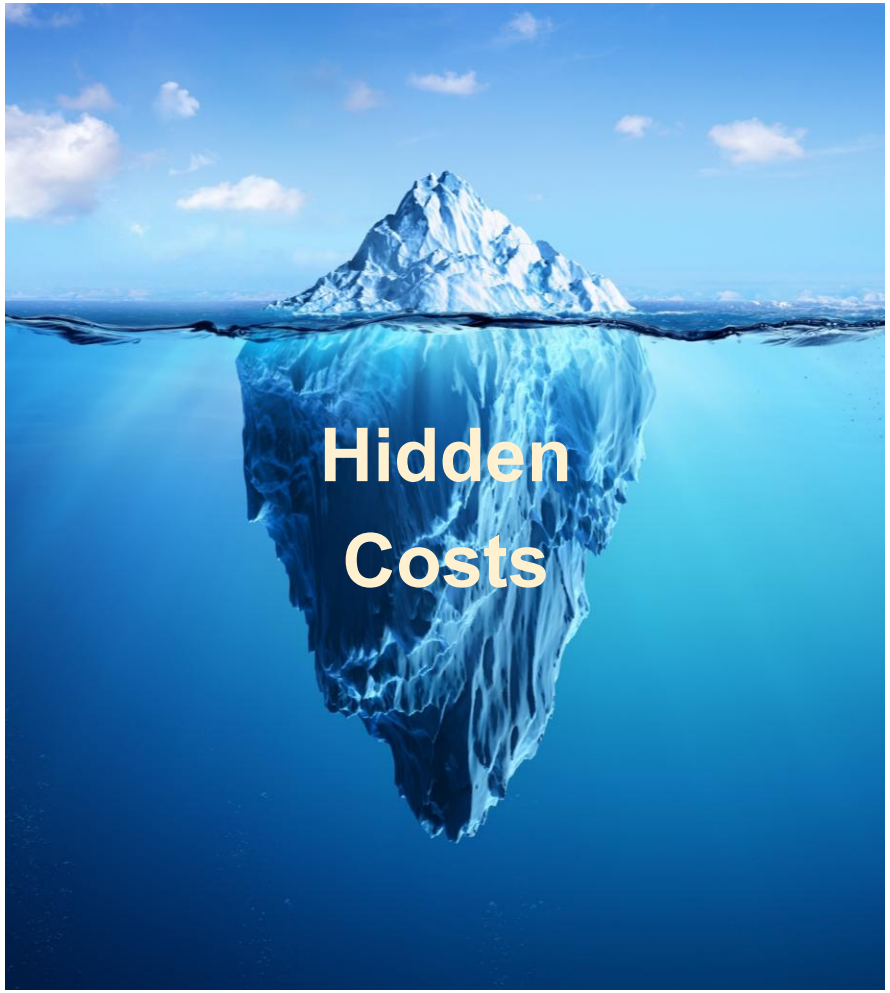
Project Planning Uncertainty

Early-stage projects lack precise workload or value figures

- **Workload estimation is inherently speculative**
- **Requirements evolve as understanding deepens**
- **Optimal infrastructure depends on:**
  - Compute requirements
  - Multiple workload characteristics
  - Software and data access requirements
  - Access to technology
  - Project goals
- **Proper decision-making requires consideration of all of the above**

Some Costs are Fragmented and Hidden

Costs are budgeted and tracked in various places, not always seen together by project planners



- **Starting Costs are easy enough to estimate:**
  - # of CPU hours
  - # of GPU hours
  - Necessary storage
- **Hidden costs vary between organizations, projects, and computing environments**
  - Software & Licensing
  - Maintenance & Support/Managed Services
  - Utilities (Power & Cooling)
  - Networking
  - Data Egress/Transfer
  - Staff Labor: Help desk and project salaries
  - Operational
  - Vendor maintenance
  - Training
  - Building and floor space
  - Indirect / Overhead

Quantifying Value of Scientific Research

Difficulty in assigning a monetary value to the societal impact of fundamental research

= $?

Fragmented Dialogue

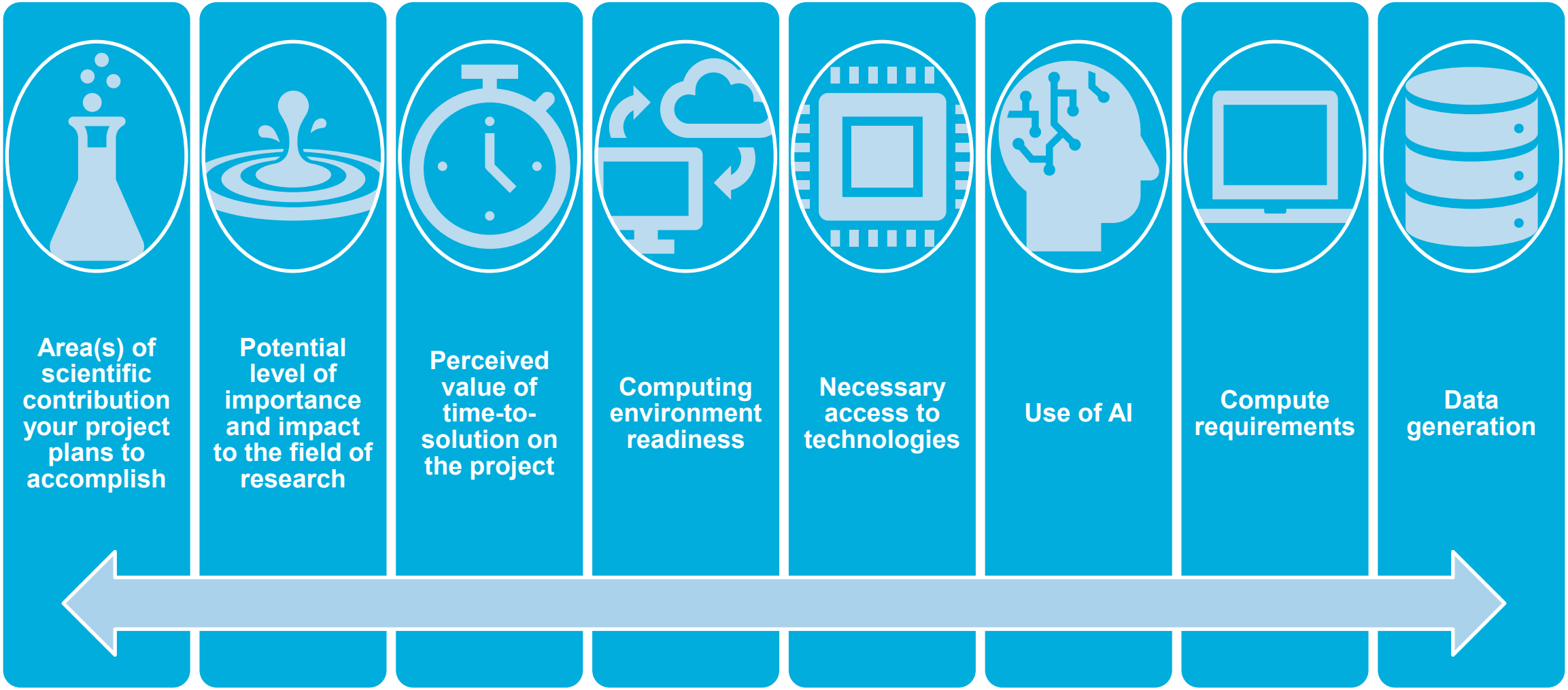
Technical, budget, and governance teams all look at TCO differently

The Solution: Scientific Computing Cost, Value, and ROI Model

- **A Decision-Support Tool:** An Excel-based model developed to guide project planning
- **Holistic Comparison:** Provide a transparent, data-driven comparison of on-premises, cloud and hybrid environments
- **Comprehensive Metrics:** Calculates and compares TCO, societal value, and ROI
- **Fosters Consensus:** Creates a common factual basis to streamline conversations and justify compute environment decisions

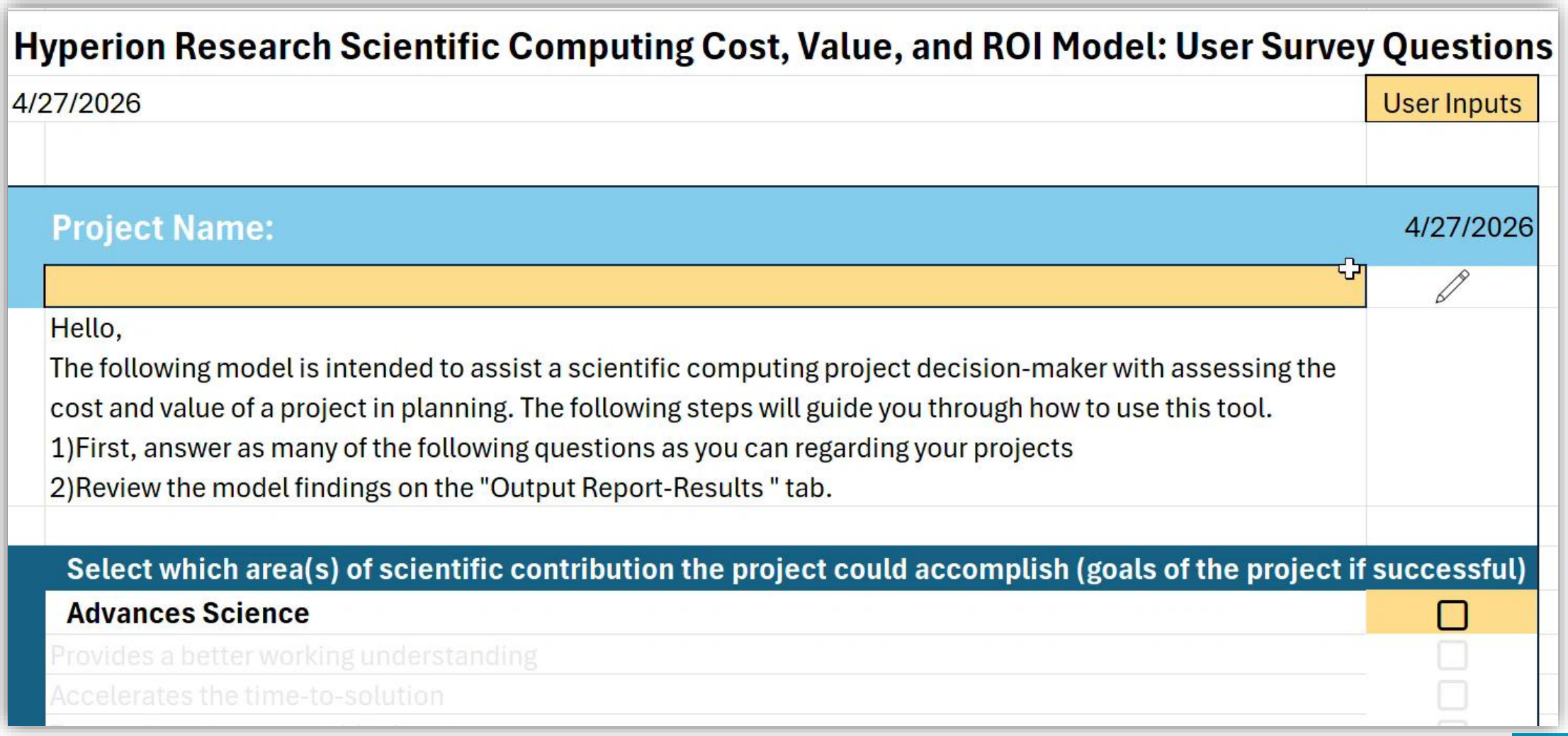
Bringing Together Project Goals and Needs

Researchers & project managers identify project-specific characteristics



Using the Tool

Area of Scientific Contribution



Hyperion Research Scientific Computing Cost, Value, and ROI Model: User Survey Questions

4/27/2026

User Inputs

Project Name: 4/27/2026

Hello,

The following model is intended to assist a scientific computing project decision-maker with assessing the cost and value of a project in planning. The following steps will guide you through how to use this tool.

1)First, answer as many of the following questions as you can regarding your projects

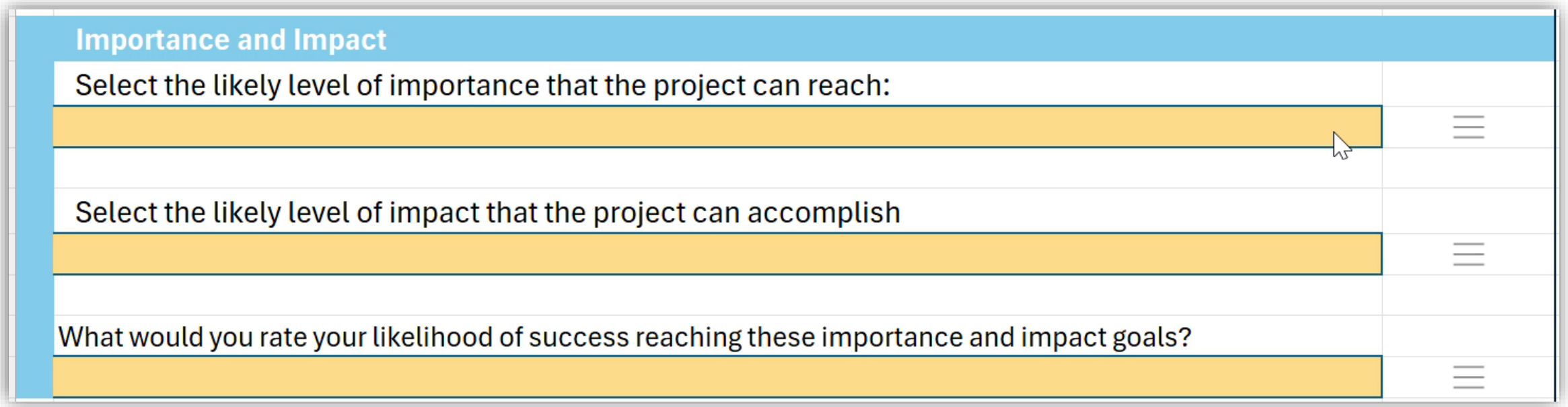
2)Review the model findings on the "Output Report-Results " tab.

Select which area(s) of scientific contribution the project could accomplish (goals of the project if successful)

Advances Science	☐
Provides a better working understanding	☐
Accelerates the time-to-solution	☐

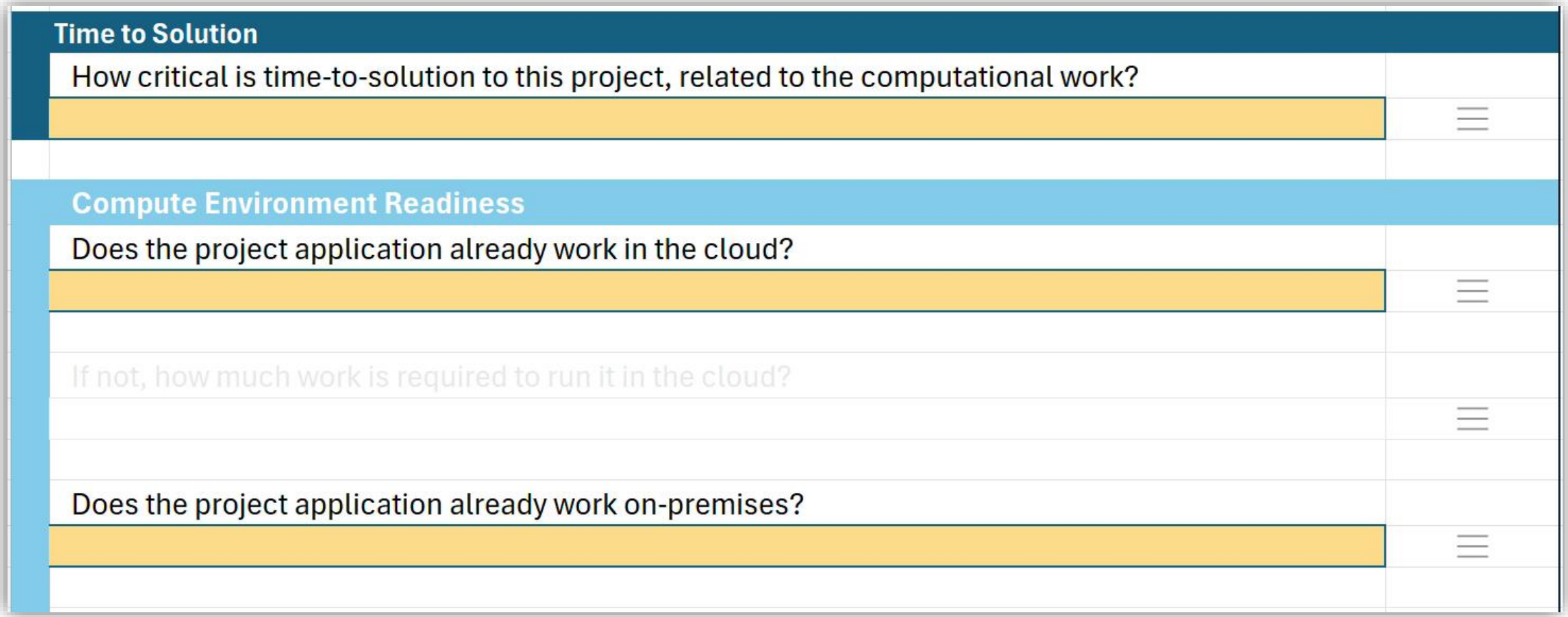
Using the Tool (Continued)

Importance and Impact



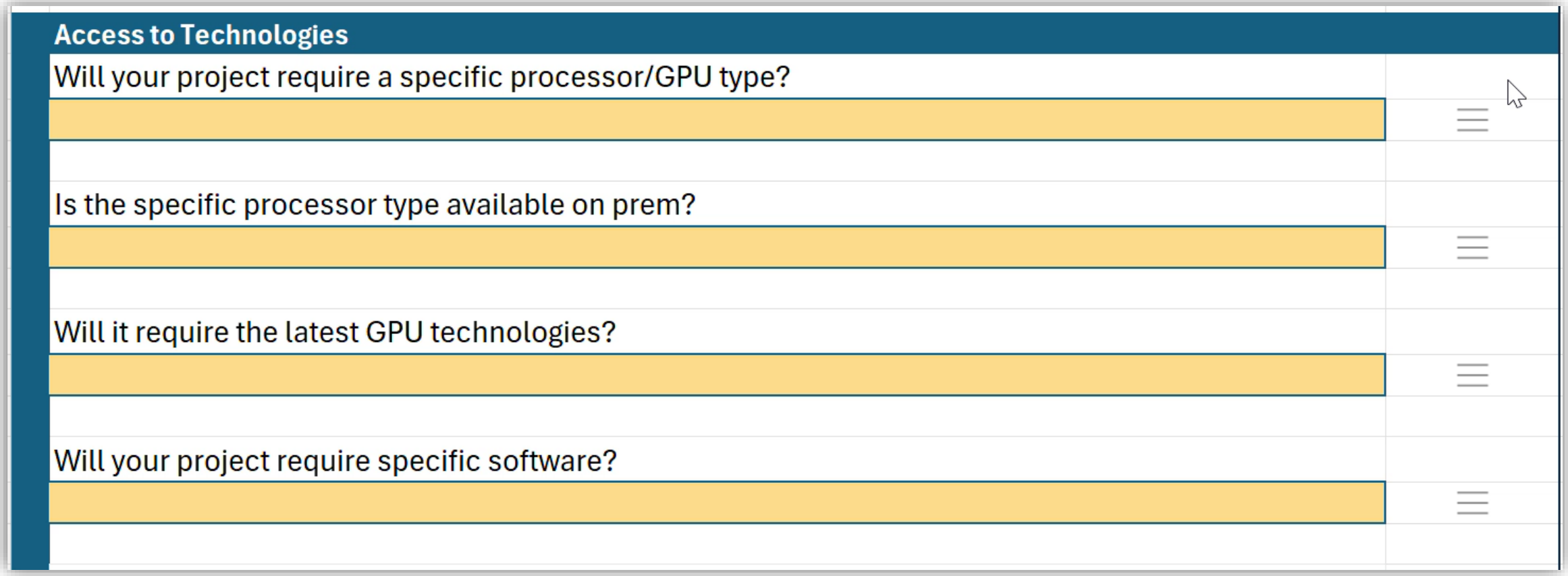
Using the Tool (Continued)

Time to Solution & Compute Environment Readiness



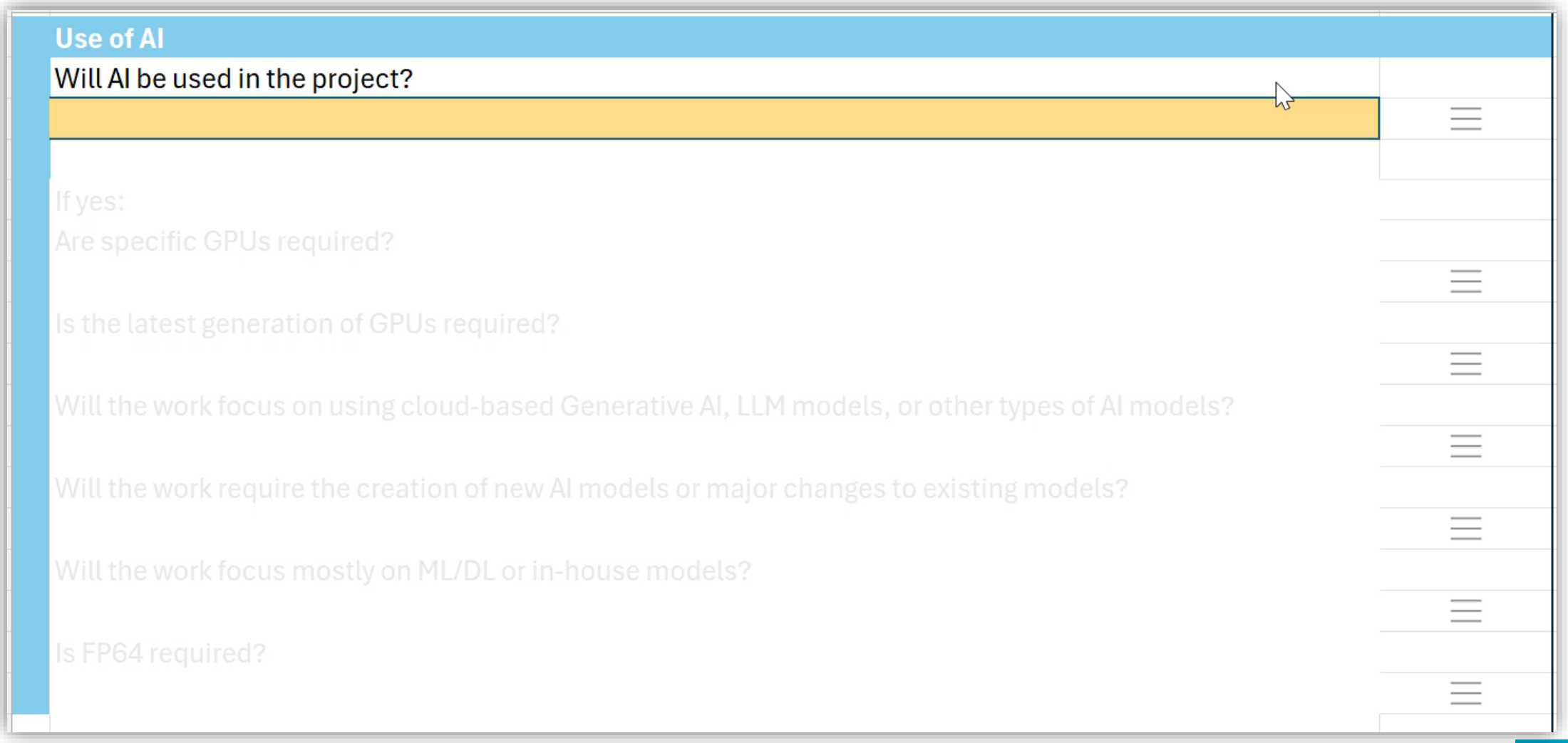
Using the Tool (Continued)

Access to Technology



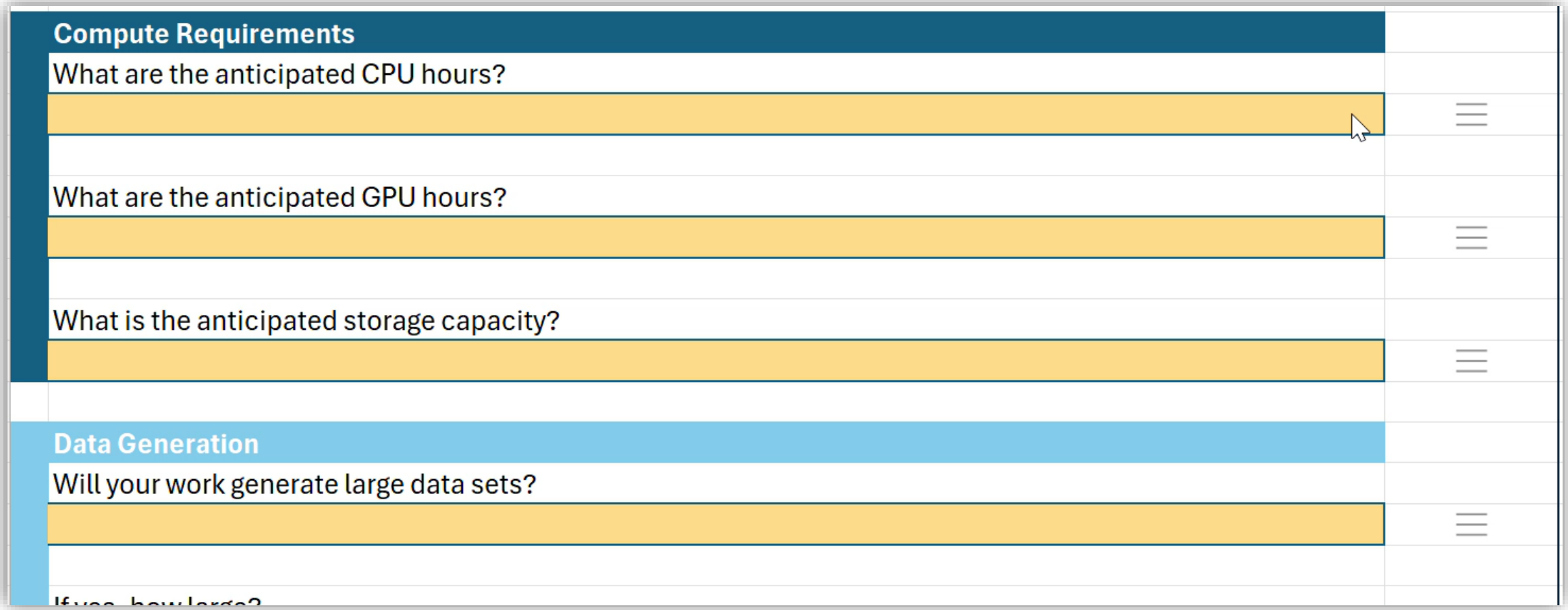
Using the Tool (Continued)

Use of AI



Using the Tool (Continued)

Compute Requirements & Data Generation



Calculations Overview

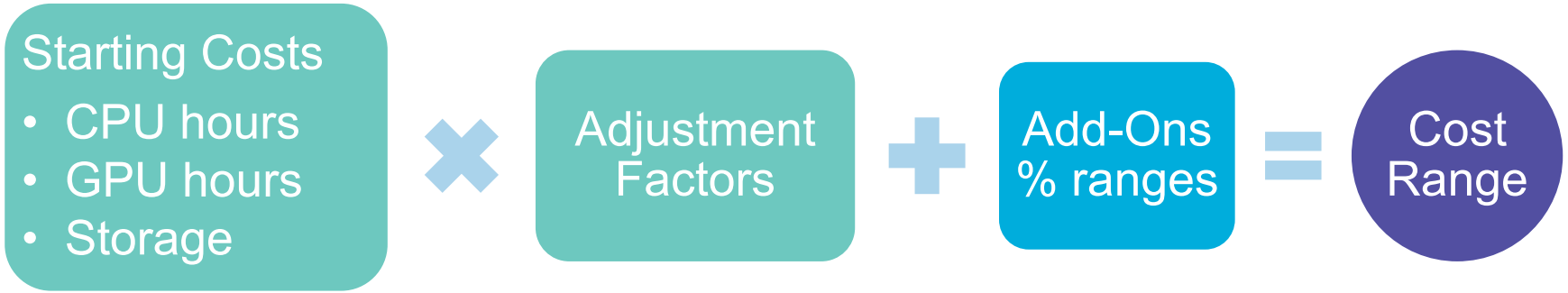
User inputs

Hyperion Research data

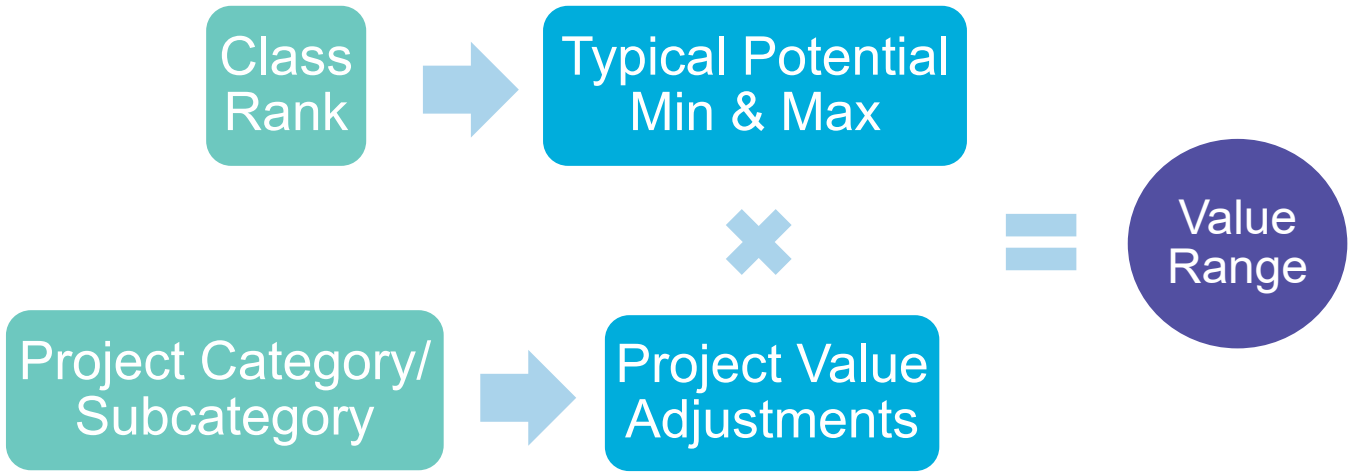
Model Outputs

Final Output

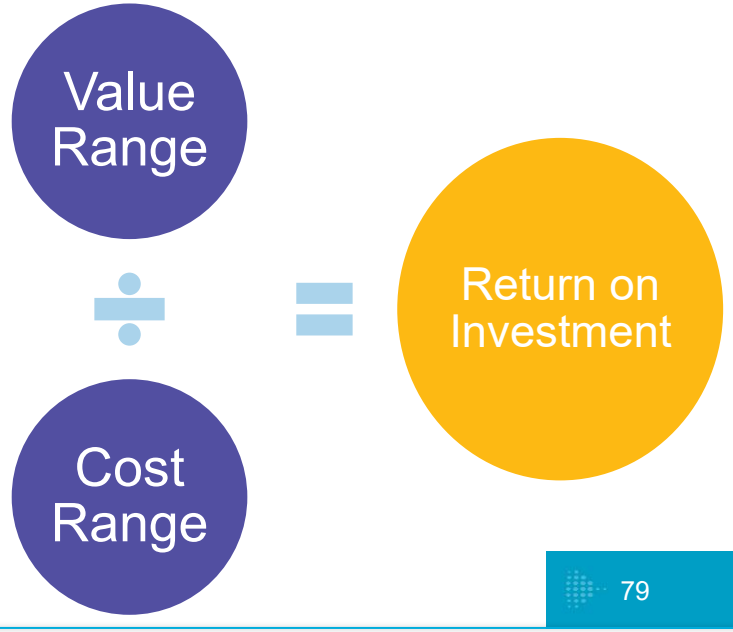
TCO Model



Value Model



ROI Model



Adjustment Factors

How project requirements and characteristics influence TCO

- Answers to the questions regarding time-to-solution, compute environment readiness, access to technology, use of AI, and data generation are tied to percentage adjustments that are applied to the starting cost TCO.
- There are several answers to adjustment questions that negate the use of a certain computing environment.
- Example:

	Adjustments to the Overall Budget					
Adjustment Questions	**Hybrid**		**On-premises**		**Cloud**	
	Lower Range	**Higher Range**	**Lower Range**	**Higher Range**	**Lower Range**	**Higher Range**
How critical is time-to-solution to this project, related to the computational work?						
Very important	8%	10%	15%	20%	0%	0%
Moderately important	5%	8%	10%	15%	0%	0%
Slightly important	3%	5%	5%	10%	0%	0%
Not important	0%	0%	0%	0%	0%	0%
Don't know	0%	0%	0%	0%	0%	0%

Add-Ons to TCO

Ensuring all elements of TCO are considered

- Add-ons are elements of TCO beyond the cost of CPU hours, GPU hours, and storage, and can be hidden or obscured in some TCO discussions.
- Based on Hyperion Research TCO data, this tool provides a lower- and higher-percent range of how these Add-ons change TCO.

Add-Ons to TCO

	Hybrid		On-premises		Cloud	
	Lower Range	Higher Range	Lower Range	Higher Range	Lower Range	Higher Range
Software & Licensing	3%	7%	3%	8%	3%	6%
Maintenance & Support/Managed Services	5%	9%	10%	18%	0%	0%
Utilities (Power & Cooling)	4%	8%	8%	15%	0%	0%
Networking	4%	9%	5%	10%	3%	8%
Data Egress/Transfer	3%	7%	3%	5%	3%	8%
Staff Labor: Help desk and project salaries	10%	18%	15%	25%	5%	10%
Operational**	5%	8%	10%	15%	0%	0%
Vendor maintenance	4%	6%	8%	12%	0%	0%
Training	1%	3%	1%	3%	1%	3%
Building and floor space	3%	5%	5%	10%	0%	0%
Indirect / Overhead	8%	13%	10%	15%	5%	10%
Total Add on budget	**49%**	**91%**	**78%**	**136%**	**20%**	**45%**

**Operational: bidding, selecting, purchasing, installation, operating, and upgrading

Results: Model Output

Users are first presented with the results of model

Project Name:	Example Project					4/28/2026
Model TCO Summary	Cost Range ($K)		Value Range ($K)		ROI	
Compute Options	Typical Min	Typical Max	Typical Min	Typical Max	Maximum	Minimum
Hybrid	$593	$3,078	$1,310	$1,938	2.2 X	0.6 X
Primarily On-premises	$796	$3,349	$1,310	$1,938	1.6 X	0.6 X
Primarily Cloud	$390	$2,647	$1,310	$1,938	3.4 X	0.7 X

- **Cost ranges, value ranges, and ROIs for 3 compute environment options**
- **But wait, there's more!**

Adjusting the TCO Values

Users may know more accurate starting costs or add-on costs, so the tool gives them an opportunity to adjust the numbers and document changes

So here is what we think your total cost is going to be:						Please adjust where you see fit:			
Hybrid	**Lower Range**		**Higher Range**			**Include**	**Adjust Lower Range**	**Adjust Higher Range**	**A**
	%	**$**	**%**	**$**			**$**	**$**	
Starting costs*	67%	$ 398,166	52%	$ 1,615,875					St
Software & Licensing	2%	$ 11,945	4%	$ 113,111		☑			Sc
Maintenance & Support/Managed Services	3%	$ 19,908	5%	$ 145,429		☑			M Sc
Utilities (Power & Cooling)	3%	$ 15,927	4%	$ 121,191		☑			U
Networking	3%	$ 15,927	5%	$ 145,429		☑			N
Data Egress/Transfer	2%	$ 11,945	3%	$ 105,032		☑			D
Staff Labor: Help desk and project salaries	7%	$ 39,817	9%	$ 282,778		☑			St sa
Operational**	3%	$ 19,908	4%	$ 121,191		☑			O
Vendor maintenance	3%	$ 15,927	3%	$ 96,953		☑			V
Training	1%	$ 3,982	2%	$ 48,476		☑			Tr
Building and floor space	2%	$ 9,954	3%	$ 80,794		☑			B
Indirect / Overhead	5%	$ 29,862	7%	$ 201,984		☑			In
Totals		$ 593,267		$ 3,078,242					To

The Final Results

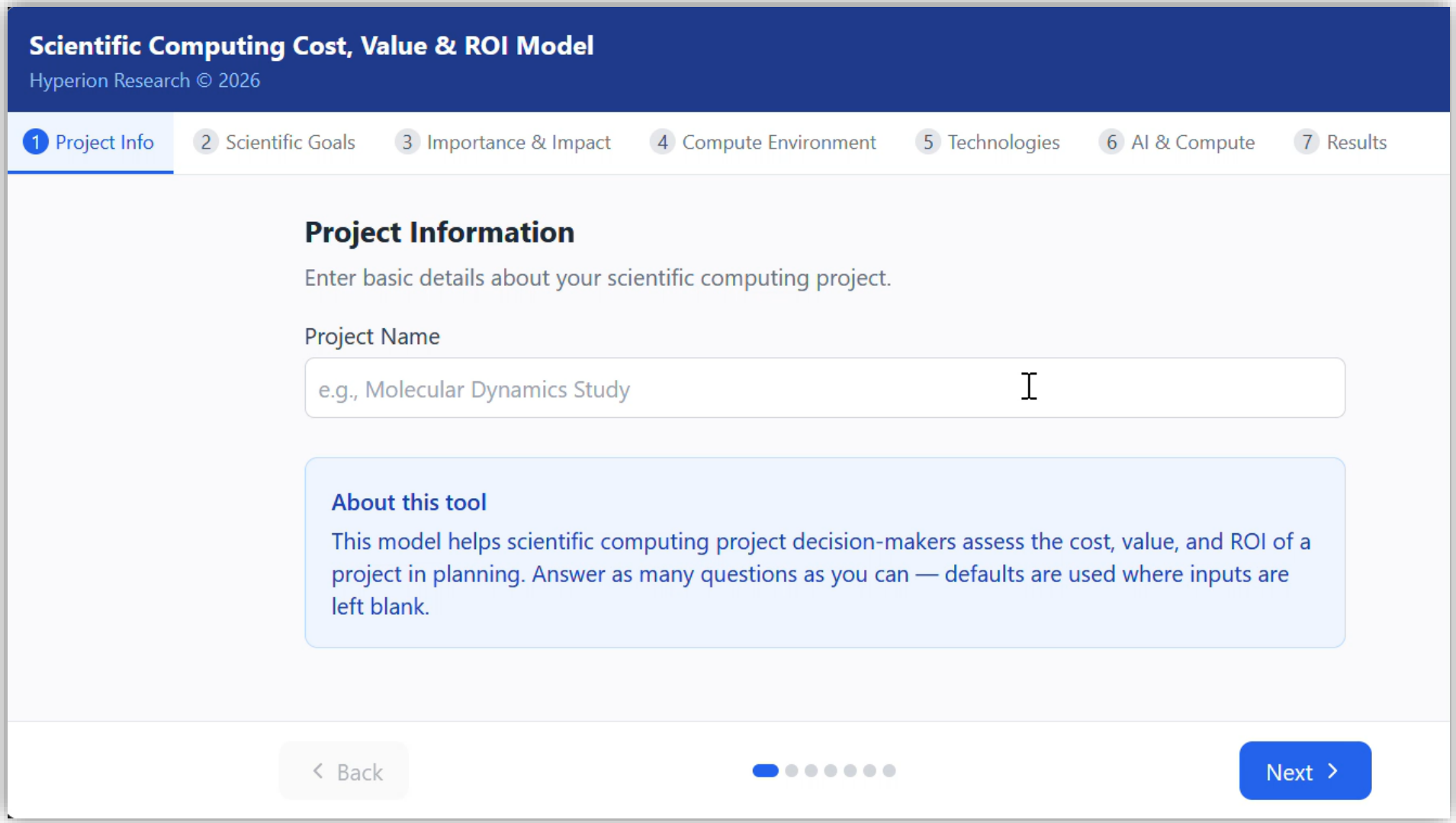
Model Output and Adjusted Outputs

Project Name:	Example Project					4/28/2026
Model TCO Summary	Cost Range ($K)		Value Range ($K)		ROI	
Compute Options	Typical Min	Typical Max	Typical Min	Typical Max	Maximum	Minimum
Hybrid	$593	$3,078	$1,310	$1,938	2.2 X	0.6 X
Primarily On-premises	$796	$3,349	$1,310	$1,938	1.6 X	0.6 X
Primarily Cloud	$390	$2,647	$1,310	$1,938	3.4 X	0.7 X
Adjusted TCO Summary	Cost Range ($K)		Value Range ($K)		ROI	
Compute Options	Typical Min	Typical Max	Typical Min	Typical Max	Maximum	Minimum
Hybrid	$695	$2,462	$1,310	$1,938	1.9 X	0.8 X
Primarily On-premises	$738	$2,980	$1,310	$1,938	1.8 X	0.7 X
Primarily Cloud	$395	$2,531	$1,310	$1,938	3.3 X	0.8 X

Source: Hyperion Research, 2025

Plan for Next Steps

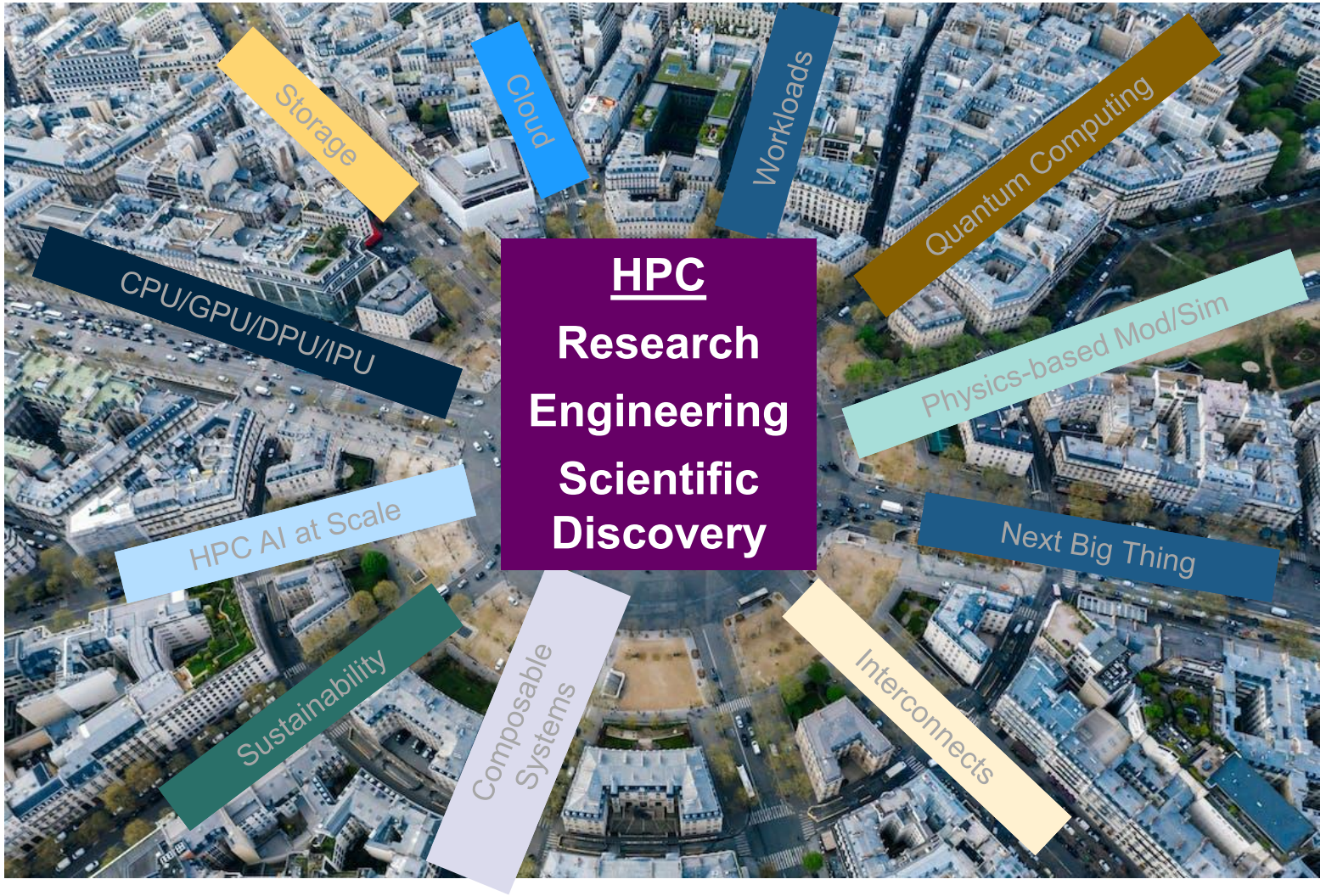
Web-based version of this Tool



Storage

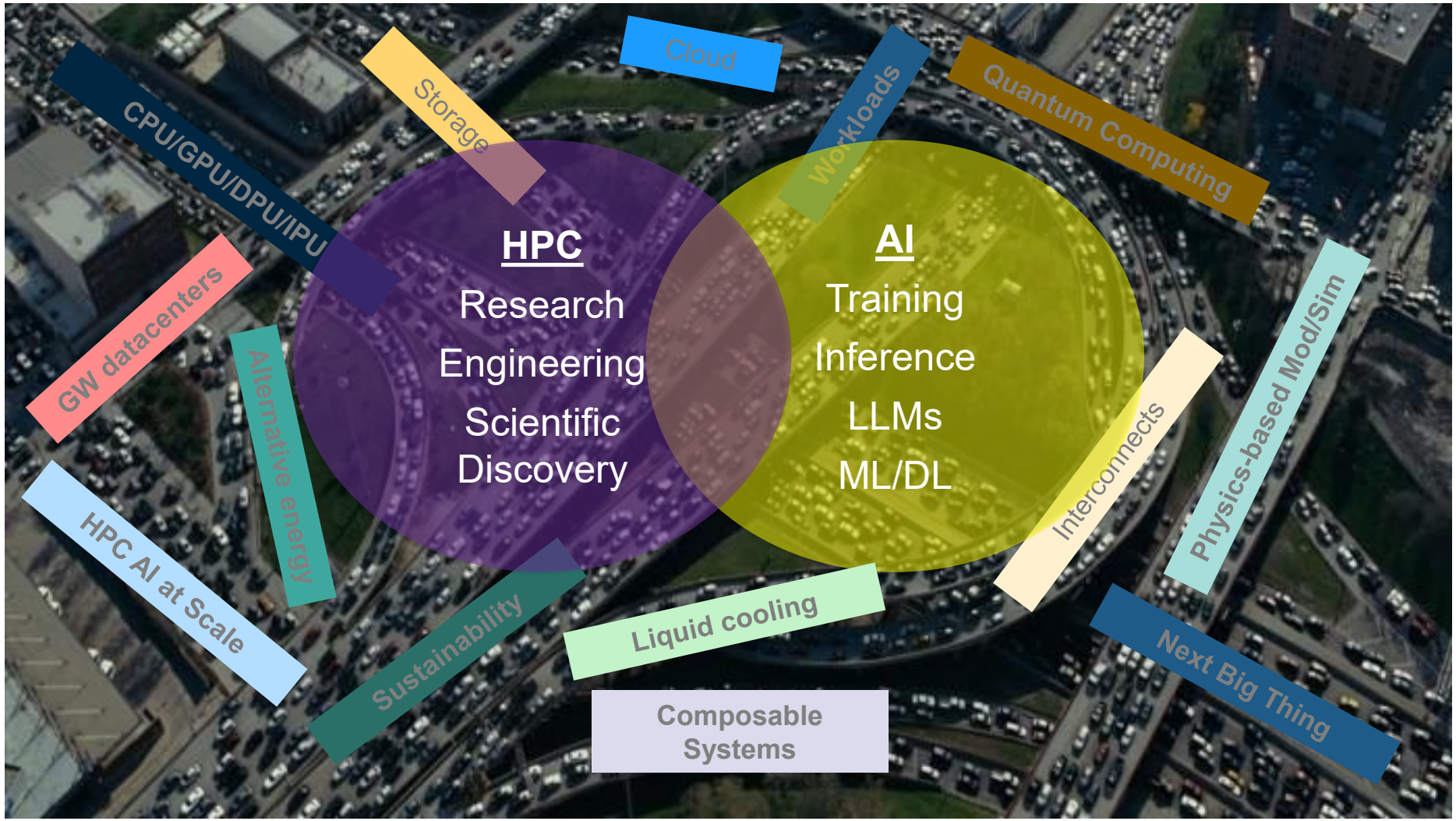
Not Your Father's HPC

A Busy Intersection of Complex Challenges



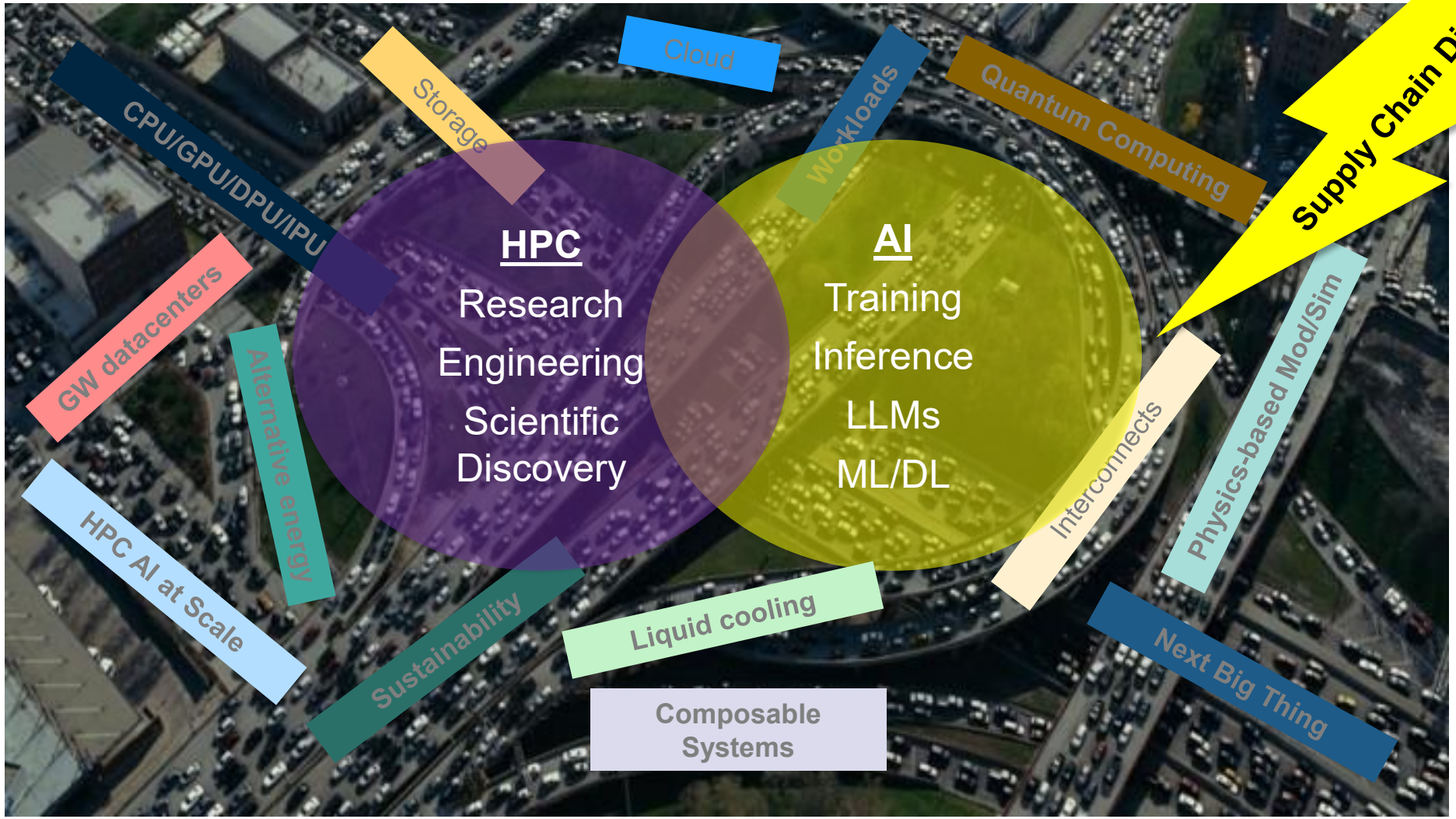
Not Your Father's HPC... ...now add AI

A Busy Intersection of Complex Challenges



Not Your Father's HPC... ...now add AI

A Busy Intersection of Complex Challenges

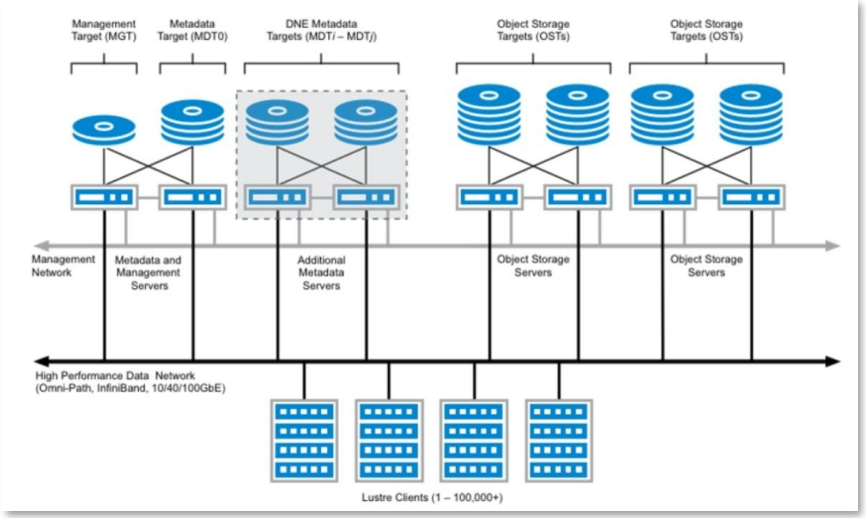


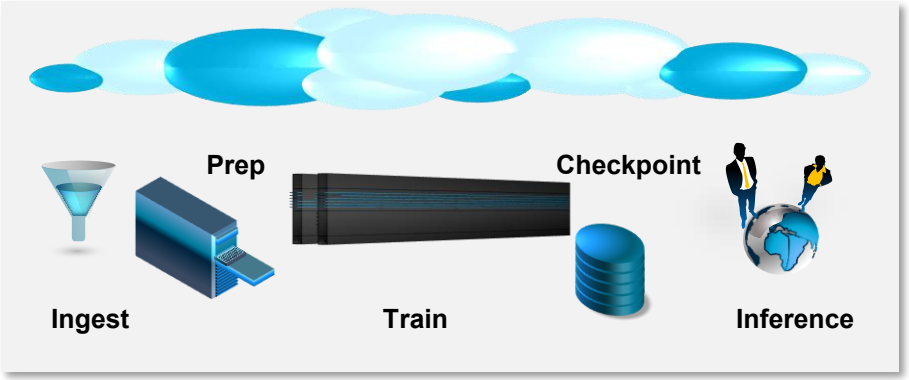
Storage: Data Platforms Take Hold

In: Accommodation of the heterogeneous workloads of traditional HPC and modern AI workloads

Out: Storage systems tailored primarily for homogenous traditional HPC workloads

- **Storage evolving into content-aware data platforms for heterogeneous workloads**
- **AI workloads driving need for high-performance, flexible I/O**
- **Storage market providing innovation and business growth opportunities**
- **Market growth likely to spur acquisitions, funding rounds, and IPOs**
- **Providers are increasingly integrating AI into their data platforms, in addition to optimizing them for AI data pipeline**
- **CSPs are also addressing data platform needs**





Supply Chain Challenges and Opportunities

Pricing, allocations, and lead times leading to creative responses

- **Memory manufacturers reallocating NAND capacity to HBM capacity**
- **Cluster memory capacities being driven higher by AI inferencing and requirements for long context windows**
- **Pricing rising in response 2x-3x over the past couple of years...**
- **...driving storage to become a large % of a balanced system architecture's budget**
- **New manufacturing capacity projected to not come on-line until mid-late 2027**
- **Maximize effective flash capacity from existing assets in AI and HPC workloads**

Data Platforms

Growing number of vendors adopting the market segment

- DDN Infinia
- Everpure
- Hammerspace
- HPE
- Huawei
- IBM
- NetApp data platform
- Oracle data platform
- VAST AI Operating System
- VDURA
- Weka

Interconnects

Emerging (Proliferation?) Standards

- **Ultra Ethernet Consortium (UEC)**
  - Scale-out
  - Contributions to Linux kernel
  - Released Rev 1.0
  - NVIDIA joined
  - Much of UEC is Slingshot
  - Now under the Linux Foundation
- **Ethernet Scale-Up Networking (ESUN)**
  - Scale-up
  - Announced at 2026 OCP Summit
- **UltraAccelerator Link (UALink) Consortium**
  - Released version 1.0 of spec
  - NVIDIA absent
- **InfiniBand**
  - Incumbent but adoption may have peaked
  - Quasi-standard; sole-sourced by NVIDIA
- **NVLink Fusion for 3rd party integration**
  - More than serdes
  - NVIDIA's response to UALink
  - Intel added in conjunction with corporate investment from NVIDIA
- **MRC on Ethernet**
  - Multiple Reliable Connections
  - Aggregate multiple serdes lanes into a single non-blocking system/cluster fabric
  - Topologically equivalent to NVLink GPU connectivity

Other Interconnect Considerations

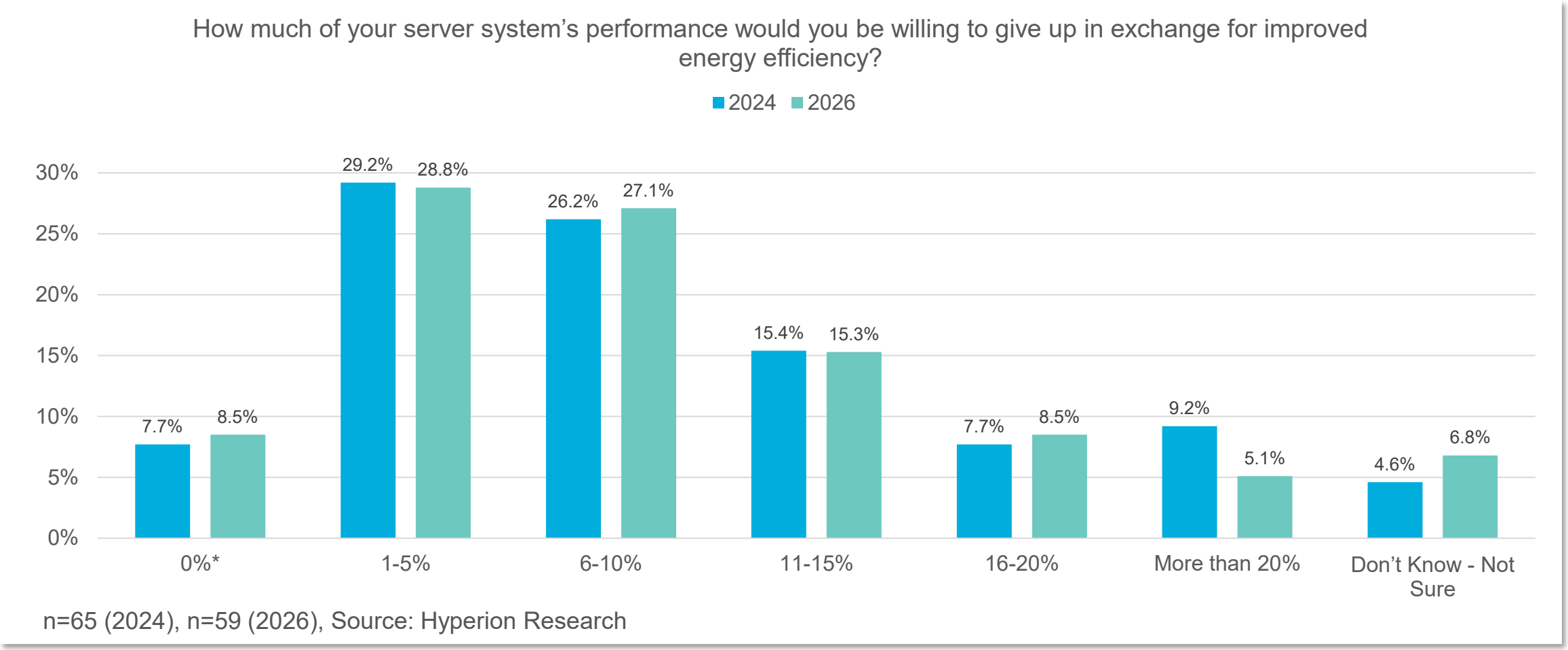
- **Other interconnects**
  - HPE Slingshot
    - Increasing line rates
    - Heavy contributions to UEC
    - Increasing promotion and visibility within the market (e.g., Slingshot workshop at SC25)
  - Eviden Bxi
    - Increasing line rates
    - Roadmap to intercept UEC
  - Cornelis OmniPath
    - Increasing line rates
    - Roadmap to intercept UEC
  - Huawei
    - UB-Mesh
    - Challenging NVLink
    - Open source the spec
- **CSPs**
  - Oracle Zettascale10 Acceleron RoCE networking
  - AWS EFA sidecar
  - Google
    - Falcon
    - optical switching
    - Virgo (TPU8i)
    - Boardfly (TPU8t)
- **Technology**
  - Optical adoption

Sustainability

HPC/AI Sustainability

- **Grid Stability**
  - May 4, 2026 – North American Electric Reliability Corporation (NERC) issued a rare Level 3 “Essential Action” Alert after repeated events of more than 1,000 MW of data center load abruptly tripped offline, highlighting that large datacenter load drops are now a documented grid stability threat
- **Creative Behind-the-Meter Solutions**
  - On-site generation capacity, powered by natural gas (primarily US facilities)
  - Small modular reactors (SMRs)
  - Banking off-peak energy via facility-sized battery energy storage systems
  - Interest in hydrogen-based backup generators
    - Groningen facility and Microsoft Lathan, NY, US (2022)
    - Bloom Energy: Calistoga, CA, US (Sept 2025)

Willingness to Exchange Performance for Energy Efficiency



n=65 (2024), n=59 (2026), Source: Hyperion Research

* I am not willing to give up any performance for energy efficiency

Call to Action and Coming Attractions

- **New website:** High Performance Computing (HPC) Research | Hyperion Research
  - Check it out!
- **New global site survey**
  - Look for results to start rolling out in 3Q26
- **Continuum Computing TCO/Value/ROI Model & Tool**
  - Sponsored by PNNL
  - Assist in providing project-based guidance on TCO and ROI analysis between cloud and on-premises infrastructure
  - Based on direct research on TCO

HYPERION RESEARCH

Questions?

mnossokoff@hyperionres.com
jludema@hyperionres.com

Hyperion Research ISC26 Update: Conclusions

June 2026

www.HyperionResearch.com
www.hpcuserforum.com

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

We Invite You to Apply for the HPC Innovation Excellence Awards:

https://www.hpcuserforum.com/innovationaward/

Examples Of Previous Winners

The Trophy and Certificate For Winners

We invite you to apply for the next round of awards: https://www.hpcuserforum.com/innovationaward/

The Innovation Award Program Goals

We invite you to apply for the next round of awards: https://www.hpcuserforum.com/innovationaward/

The HPC Innovation Excellence Awards recognize noteworthy achievements by users of HPC and AI-for-Science

The program's main goals are to:

- Recognize HPC & AI enabled innovations in science, engineering, and data analytics, including both public sector advances in science and public or private sector returns on investment (ROI)
- Showcase HPC accomplishments in various environments such as traditional HPC centers, enterprise data centers, and cloud computing platforms as well as quantum computing
- Help convey the broad benefits of adopting HPC
- Demonstrate the value of HPC to funding organizations, elected officials, and investors.
- Expand public understanding of and support for HPC

2026 Events

Event	Date	Location
~~SCA/HPCAsia~~	~~January 26 - 29~~	~~Osaka, Japan~~
~~HPC User Forum - Spring~~	~~May 5 - May 6~~	~~Austin, TX (UT-Austin)~~
ISC26 Breakfast Briefing	June 23	Hamburg, Germany
HPC User Forum - Fall - United States	September 8 - September 9	Chicago, IL (Argonne National Lab)
HPC User Forum - Fall - Europe	October 6 - October 7	Stuttgart, Germany (HLRS)
SC26 Breakfast Briefing	November 17	Chicago, IL

HYPERION RESEARCH

We Welcome Questions, Comments and Suggestions

Please contact us at:

info@hyperionres.com