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

ISC26 Market Update Briefing
June 2026

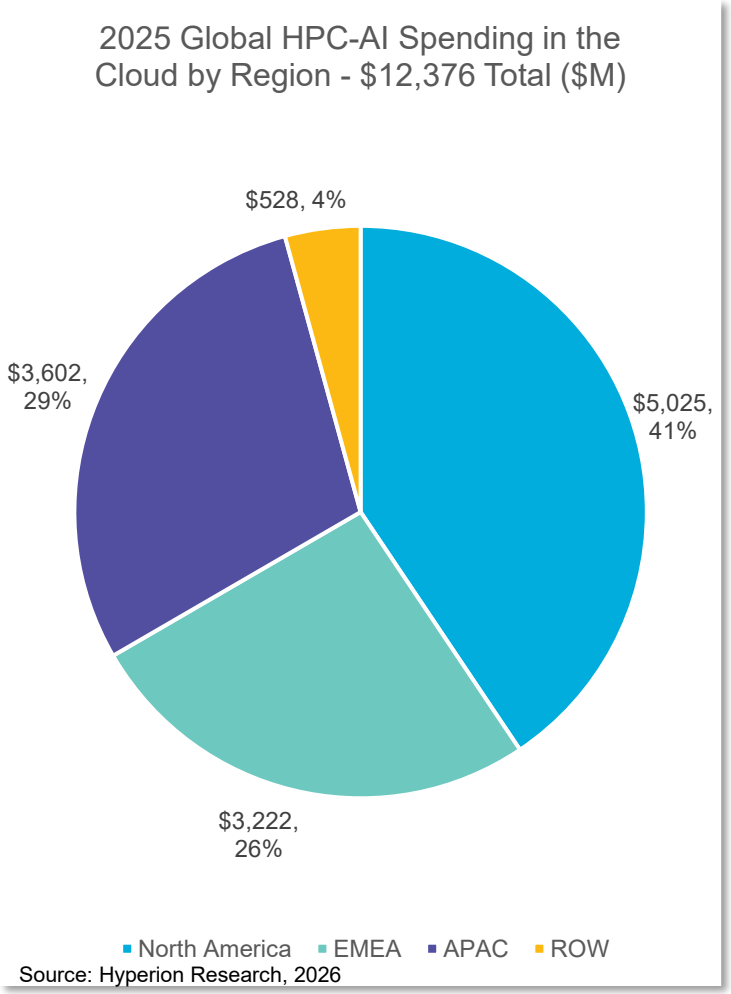
www.HyperionResearch.com
www.hpcuserforum.com

Mark Nossokoff and Jacklyn Ludema

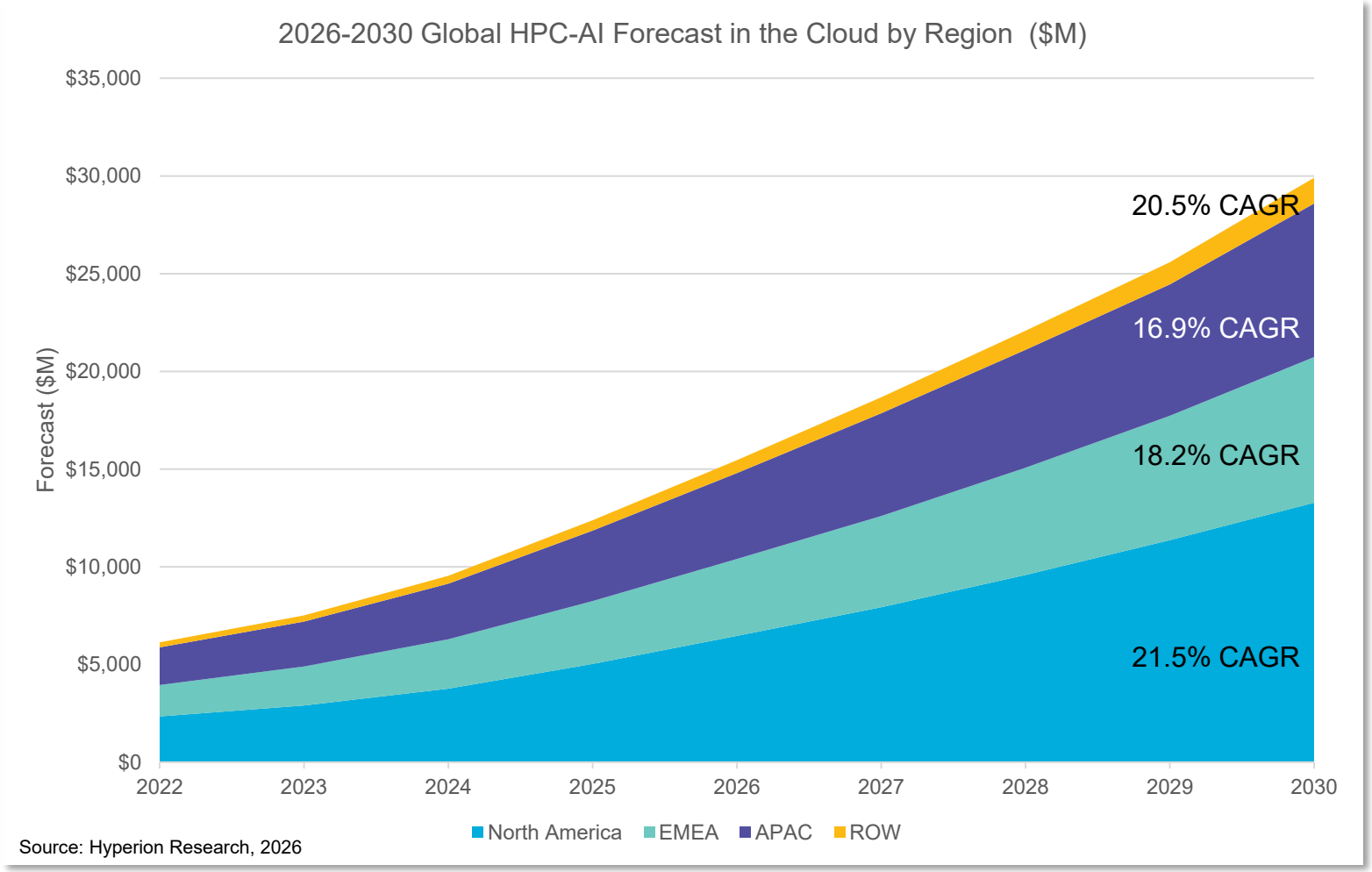
Cloud

2026-2030 HPC-AI Cloud Forecast

North America leads in cloud spending and projected growth



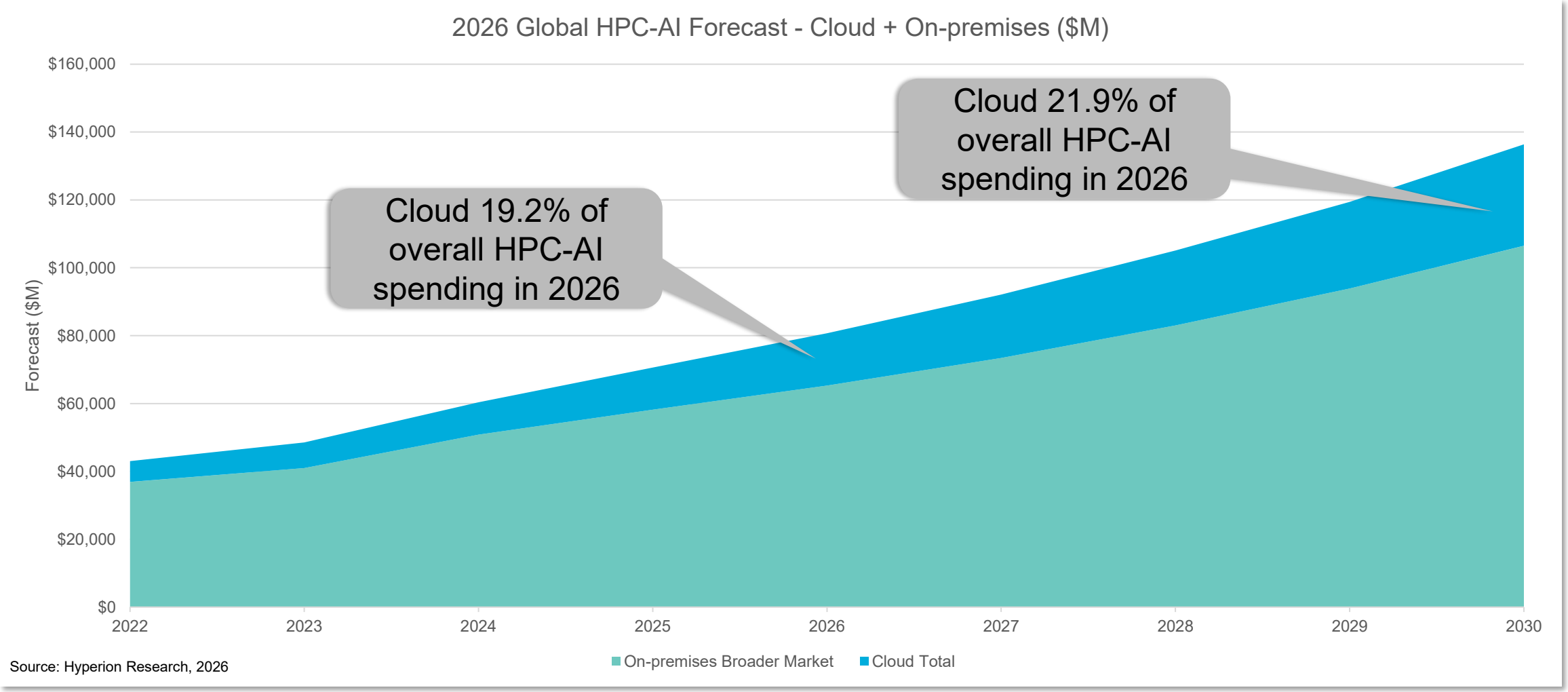
Source: Hyperion Research, 2026



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

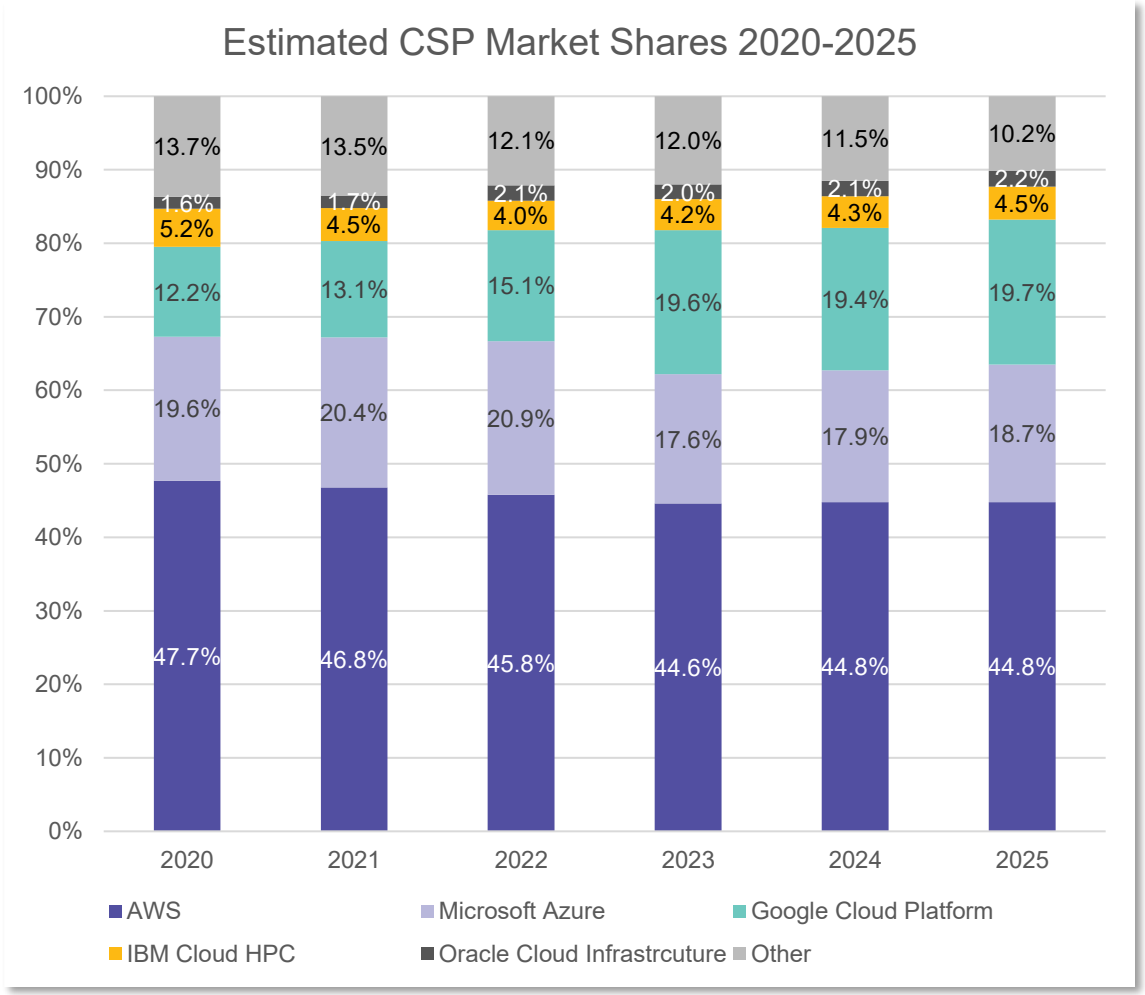


Source: Hyperion Research, 2026

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

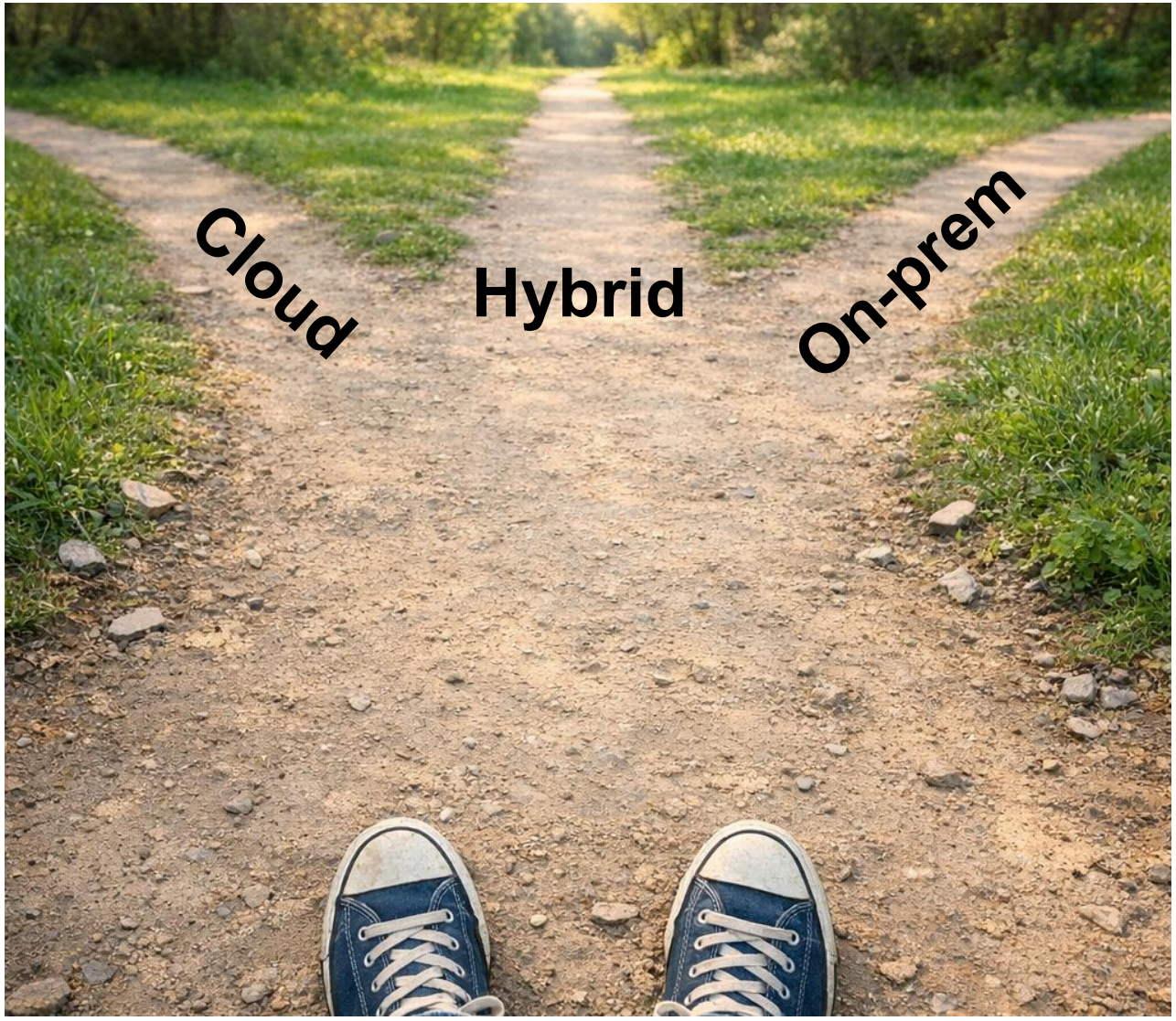
PNNL

May 2026

www.HyperionResearch.com
www.hpcuserforum.com

Jaclyn Ludema & Mark Nossokoff

The Project Compute Environment Decision



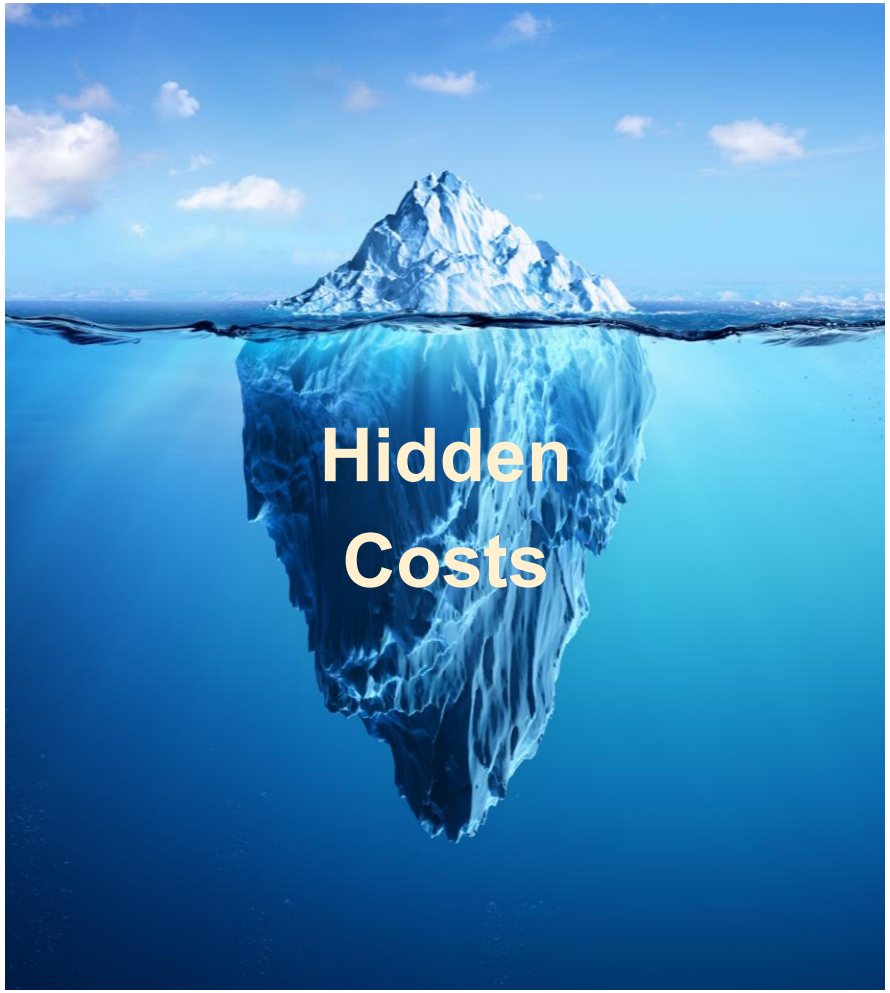
Project Planning Uncertainty

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

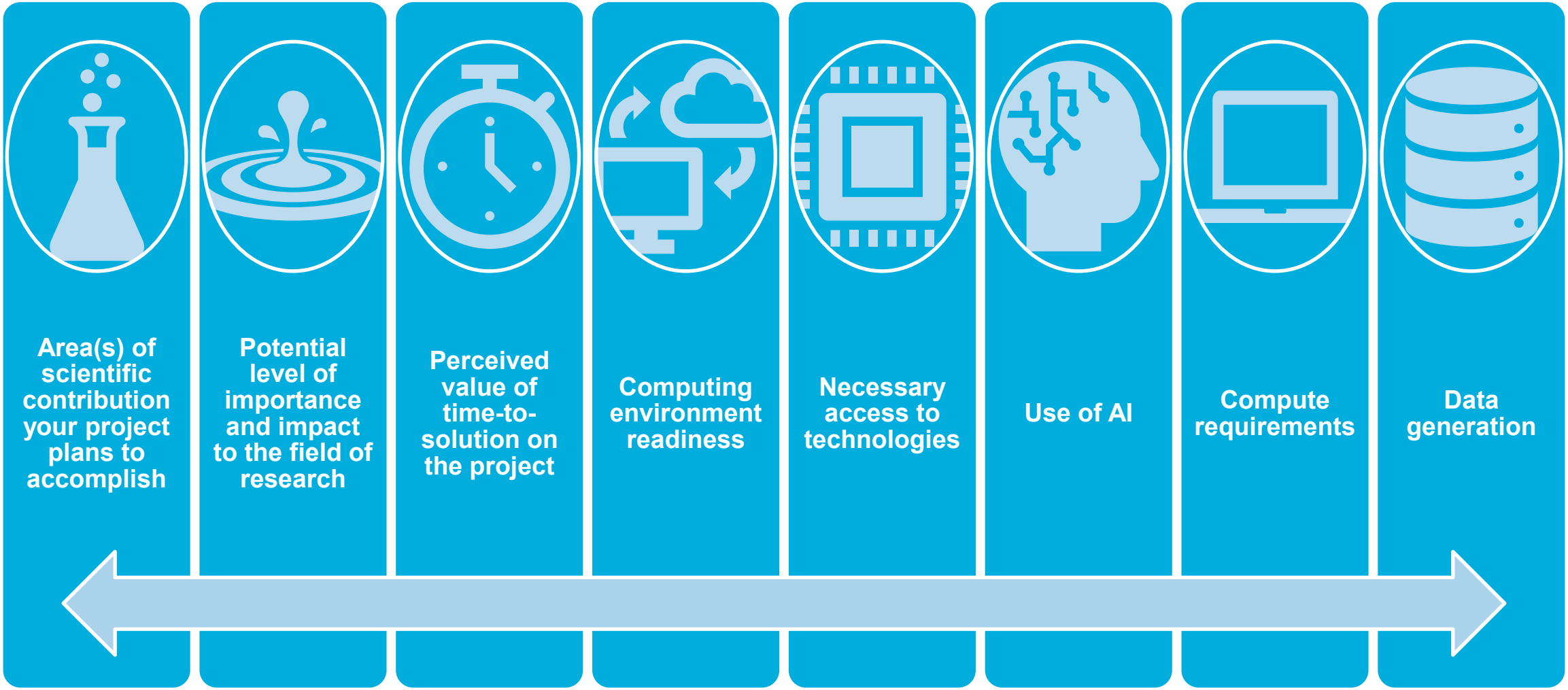
Technical, budget, and governance teams all look at TCO differently

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

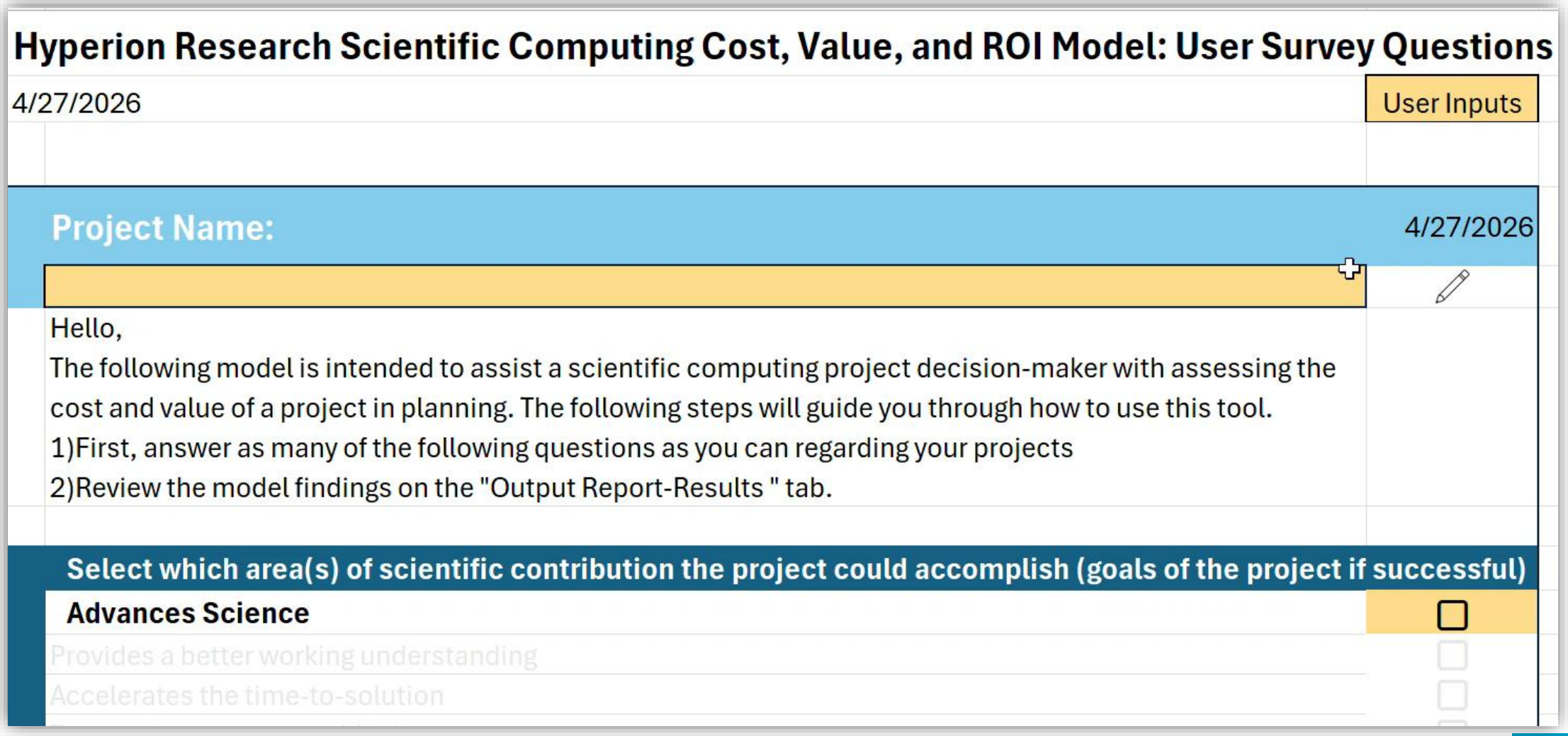
Bringing Together Project Goals and Needs

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

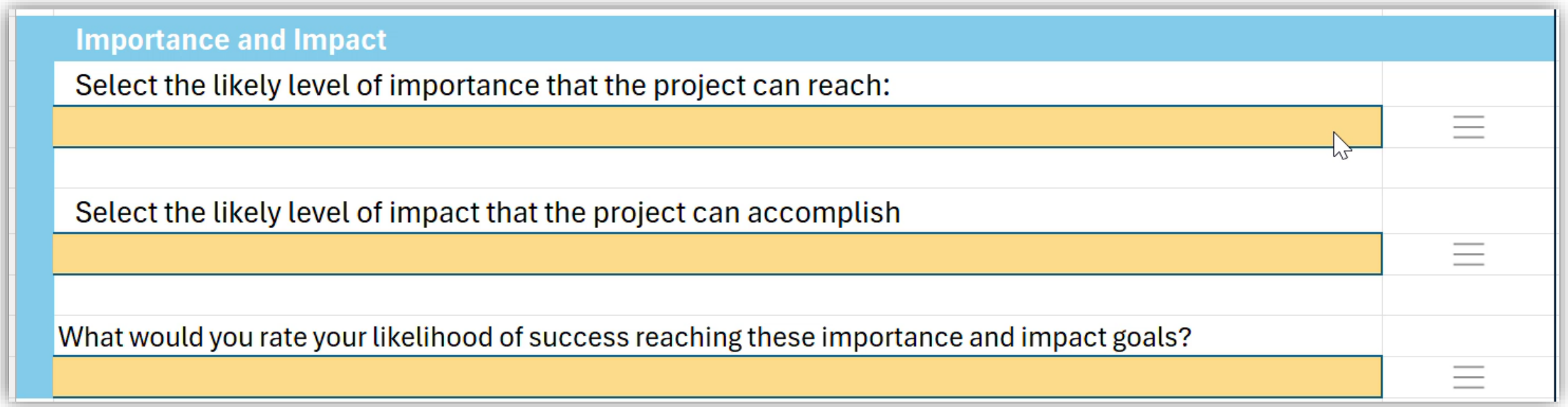
2)Review the model findings on the "Output Report-Results " tab.

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	☐

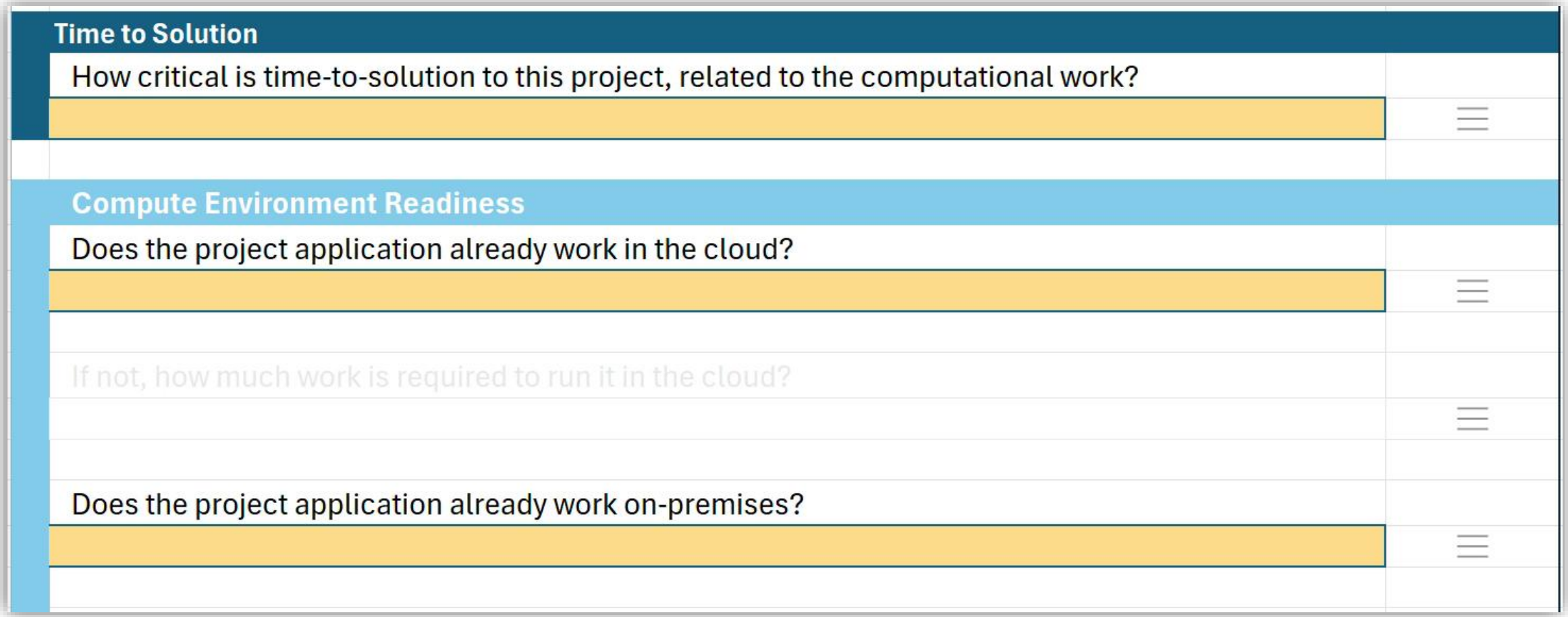
Using the Tool (Continued)

Importance and Impact



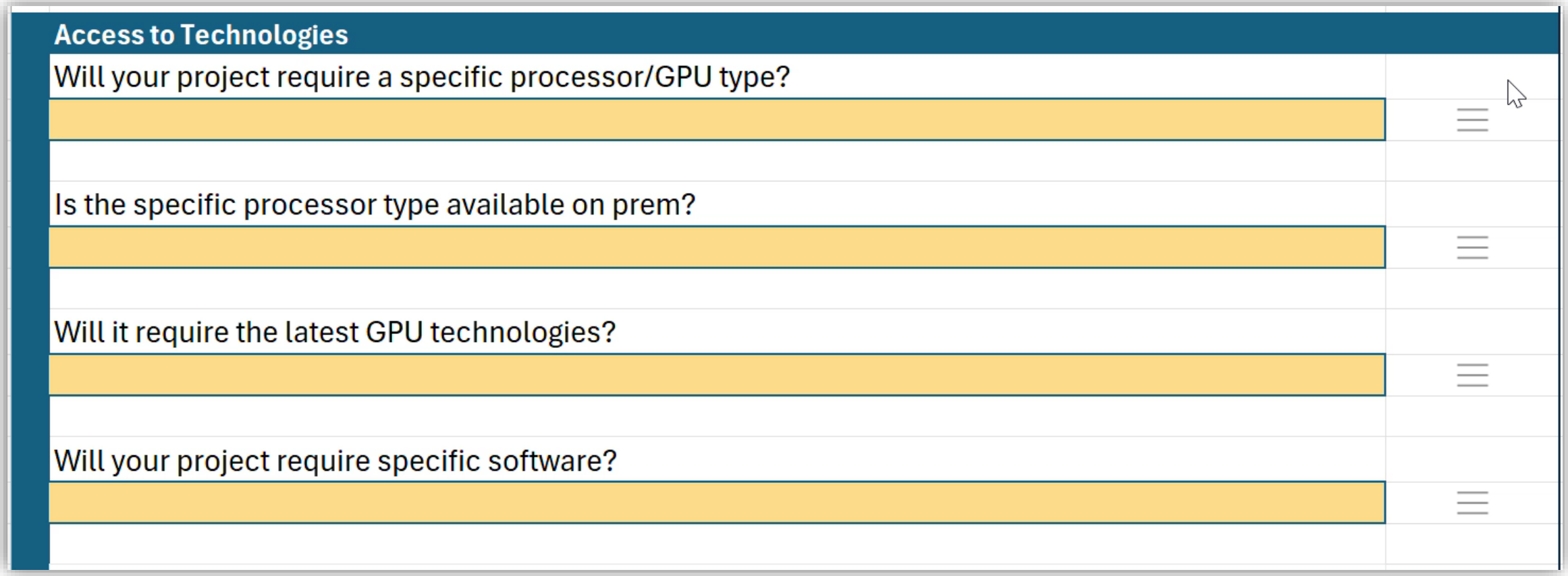
Using the Tool (Continued)

Time to Solution & Compute Environment Readiness



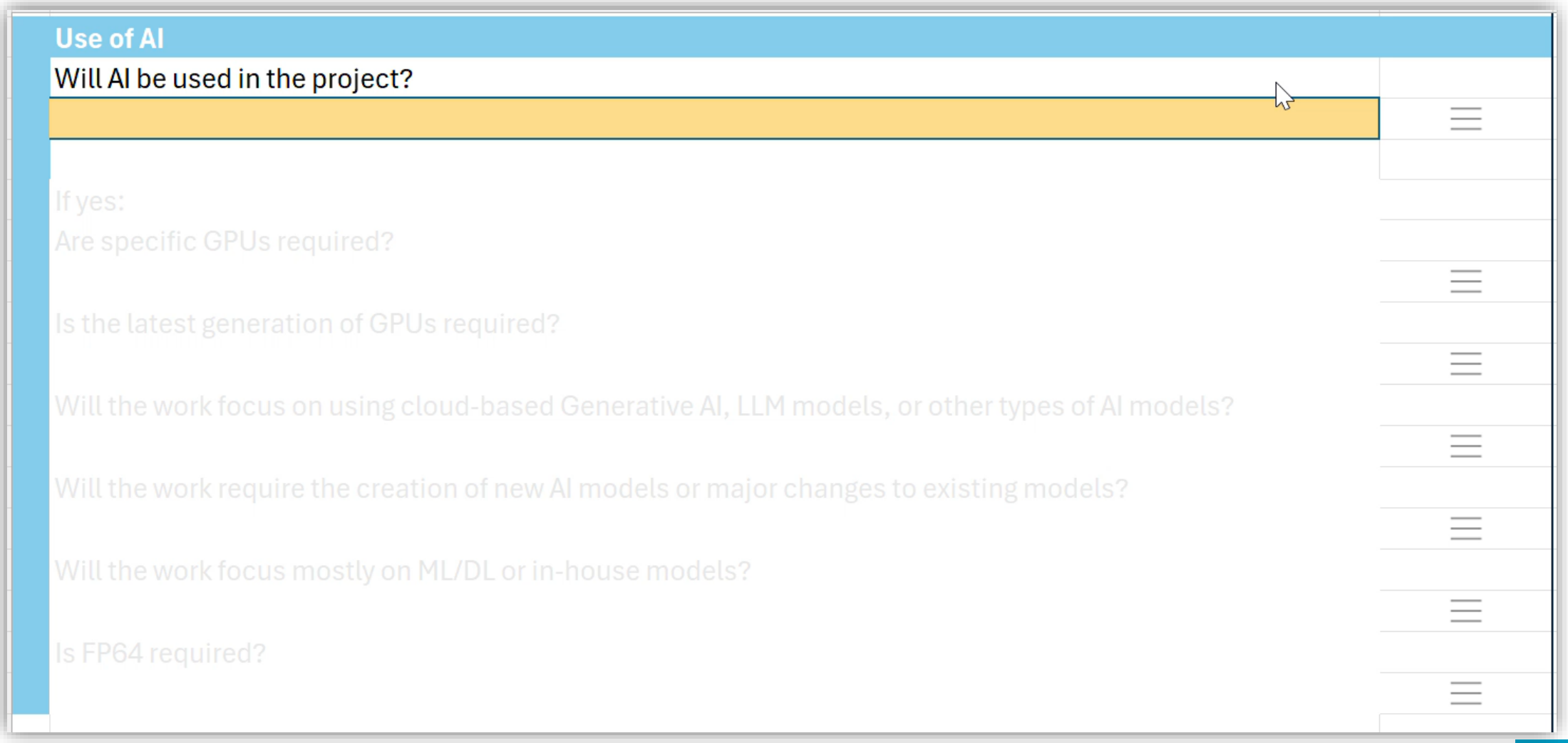
Using the Tool (Continued)

Access to Technology



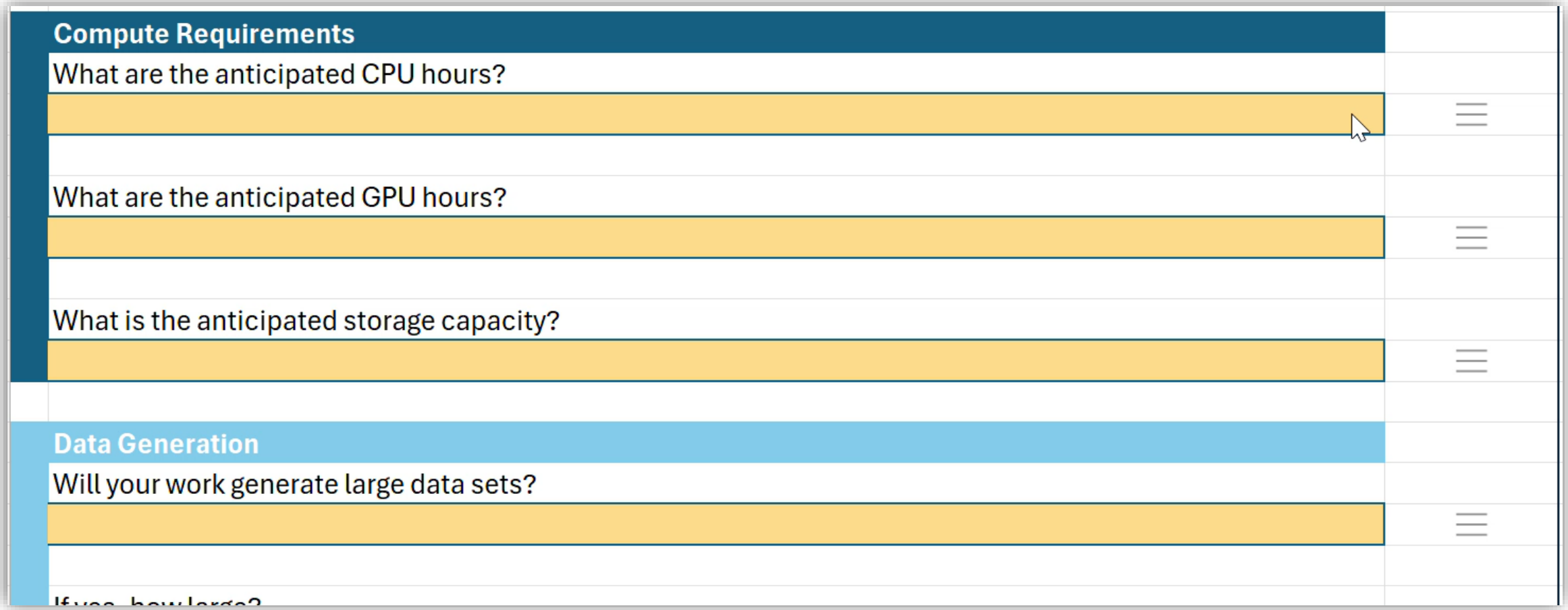
Using the Tool (Continued)

Use of AI



Using the Tool (Continued)

Compute Requirements & Data Generation



Calculations Overview

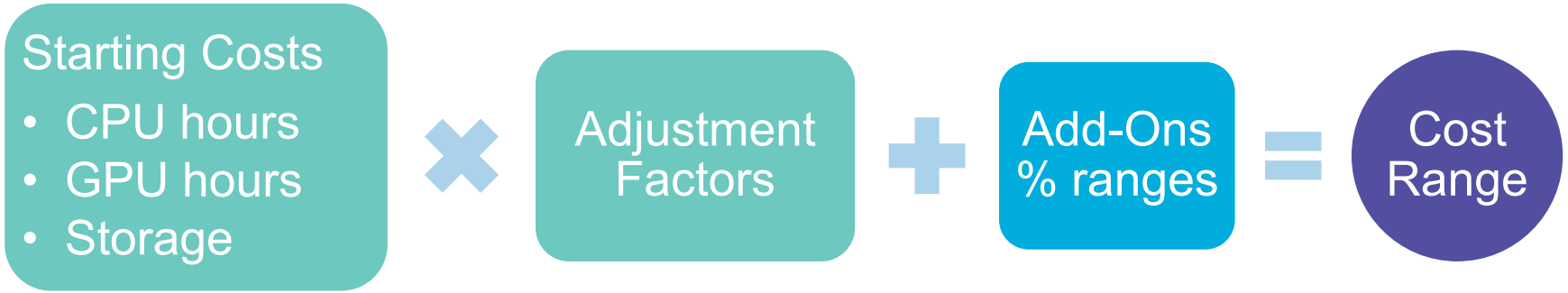
User inputs

Hyperion Research data

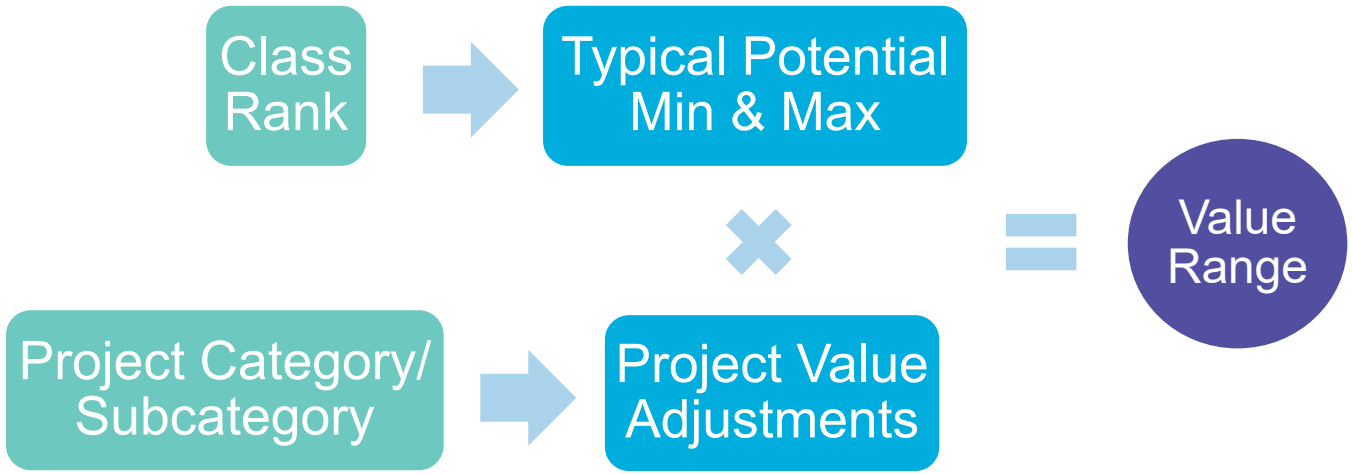
Model Outputs

Final Output

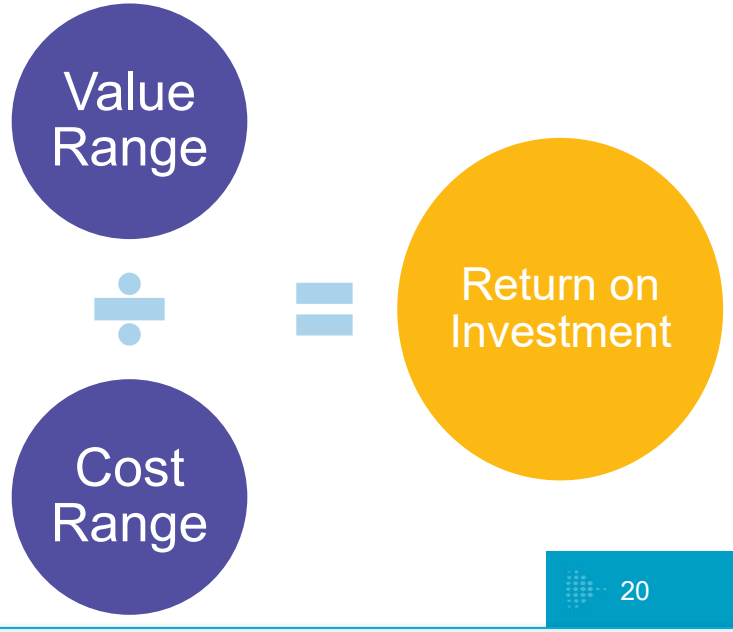
TCO Model



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

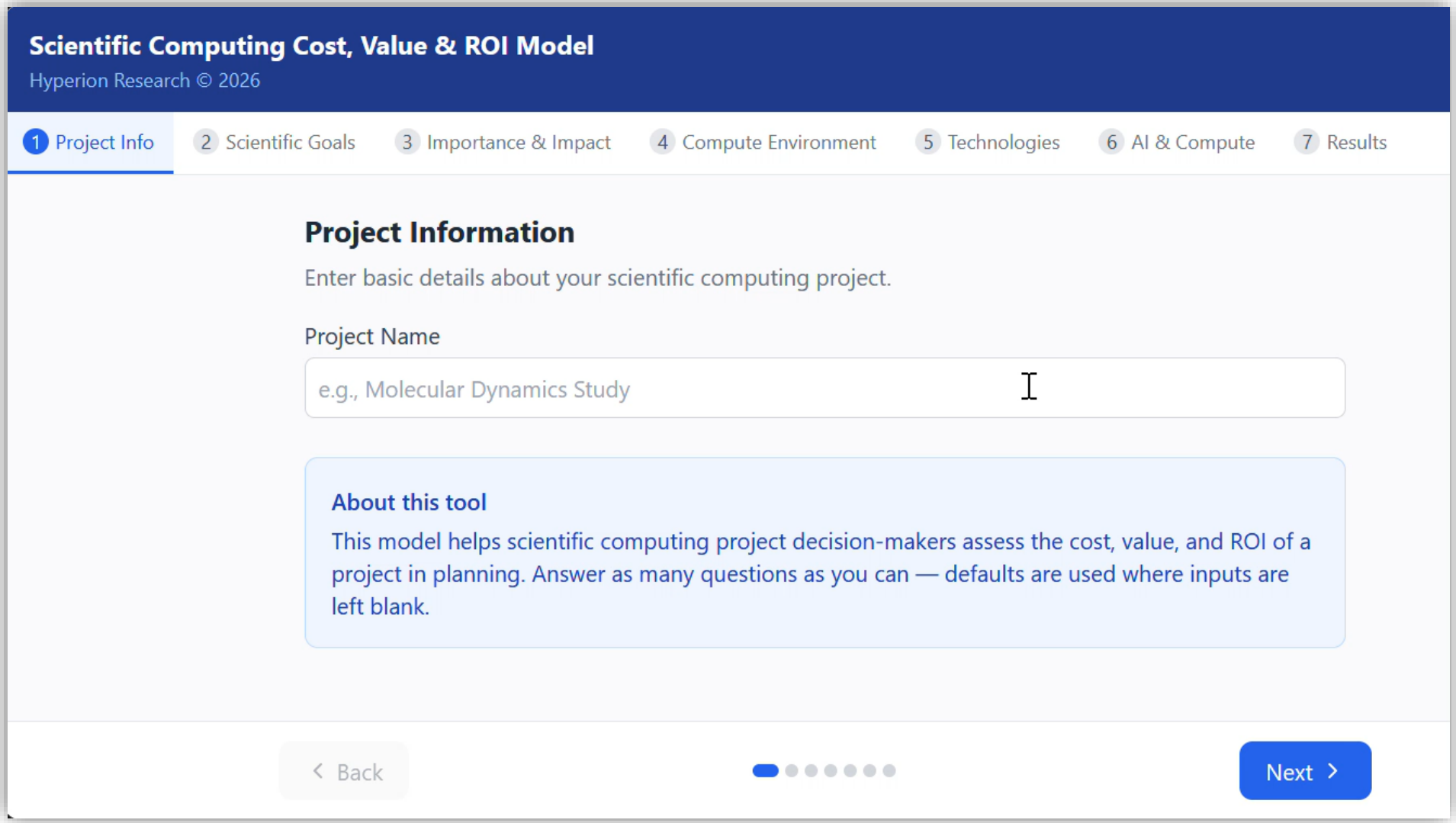
Model Output and Adjusted Outputs

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

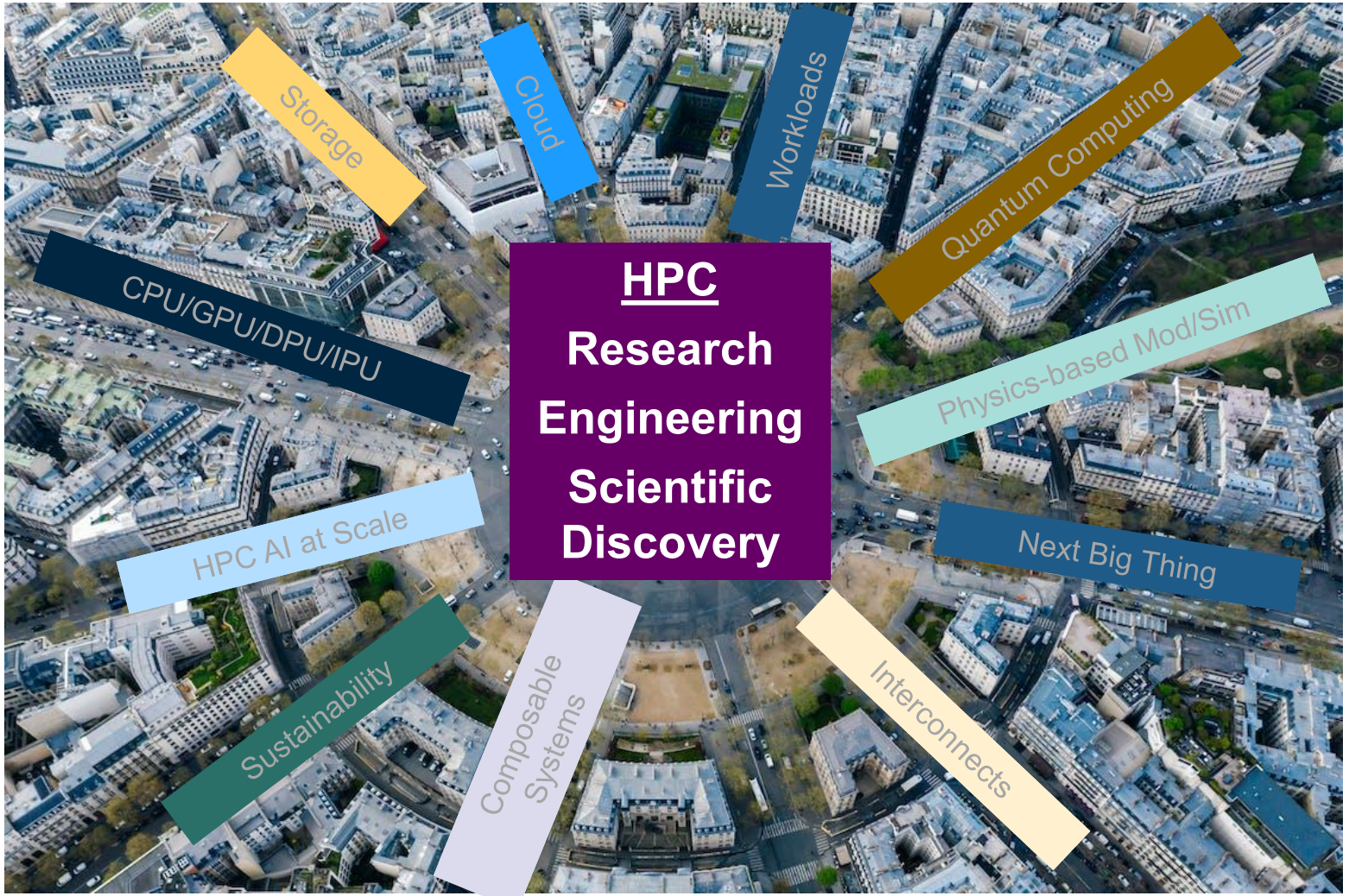
Web-based version of this Tool



Storage

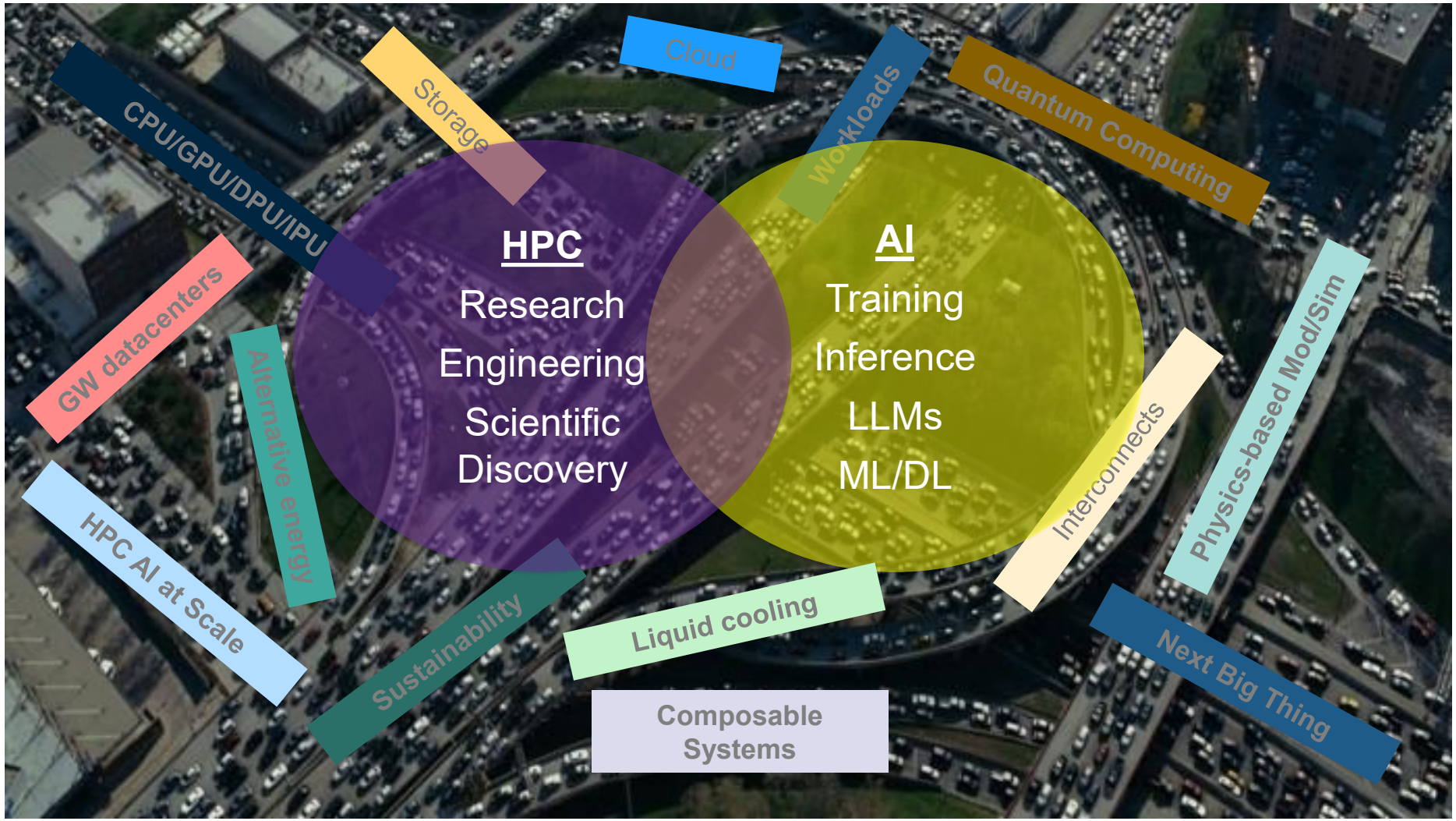
Not Your Father's HPC

A Busy Intersection of Complex Challenges



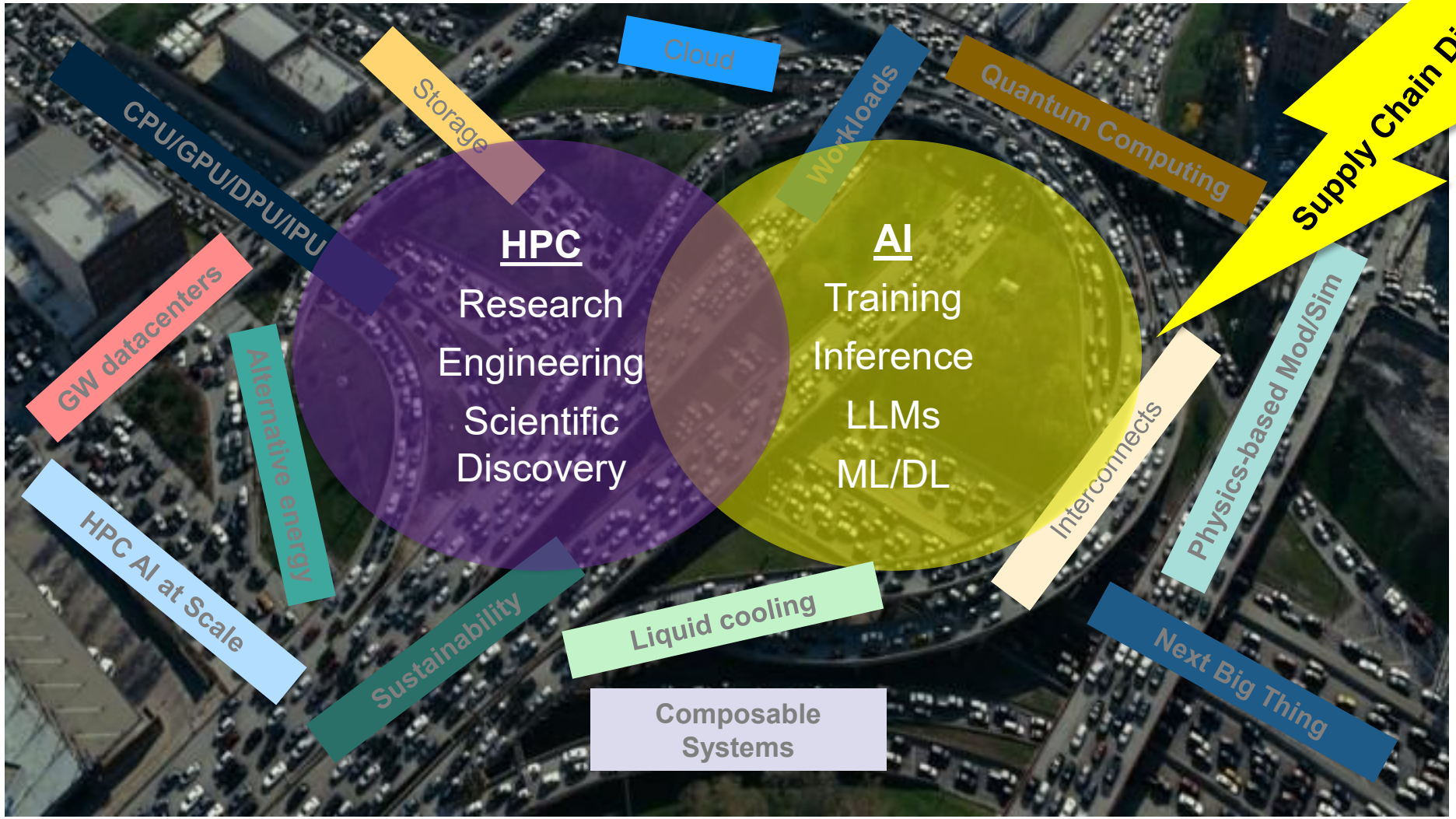
Not Your Father's HPC... ...now add AI

A Busy Intersection of Complex Challenges



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

A Busy Intersection of Complex Challenges

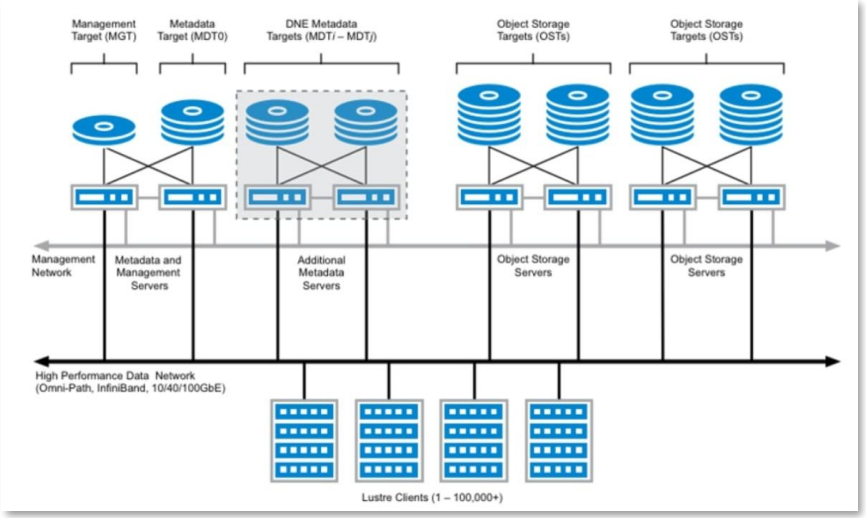


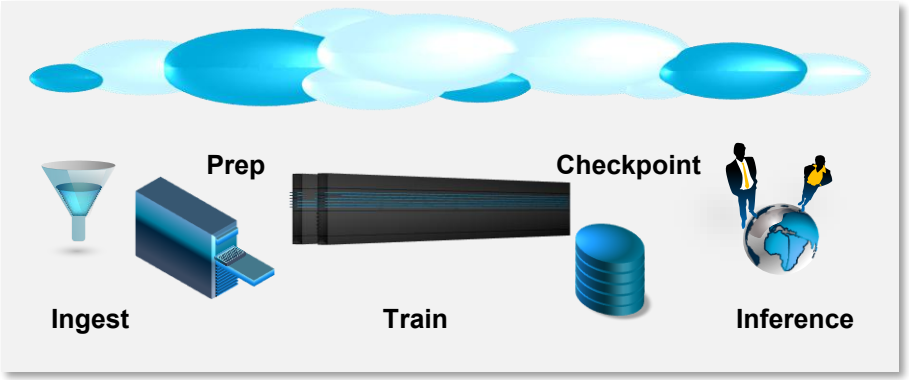
Storage: Data Platforms Take Hold

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

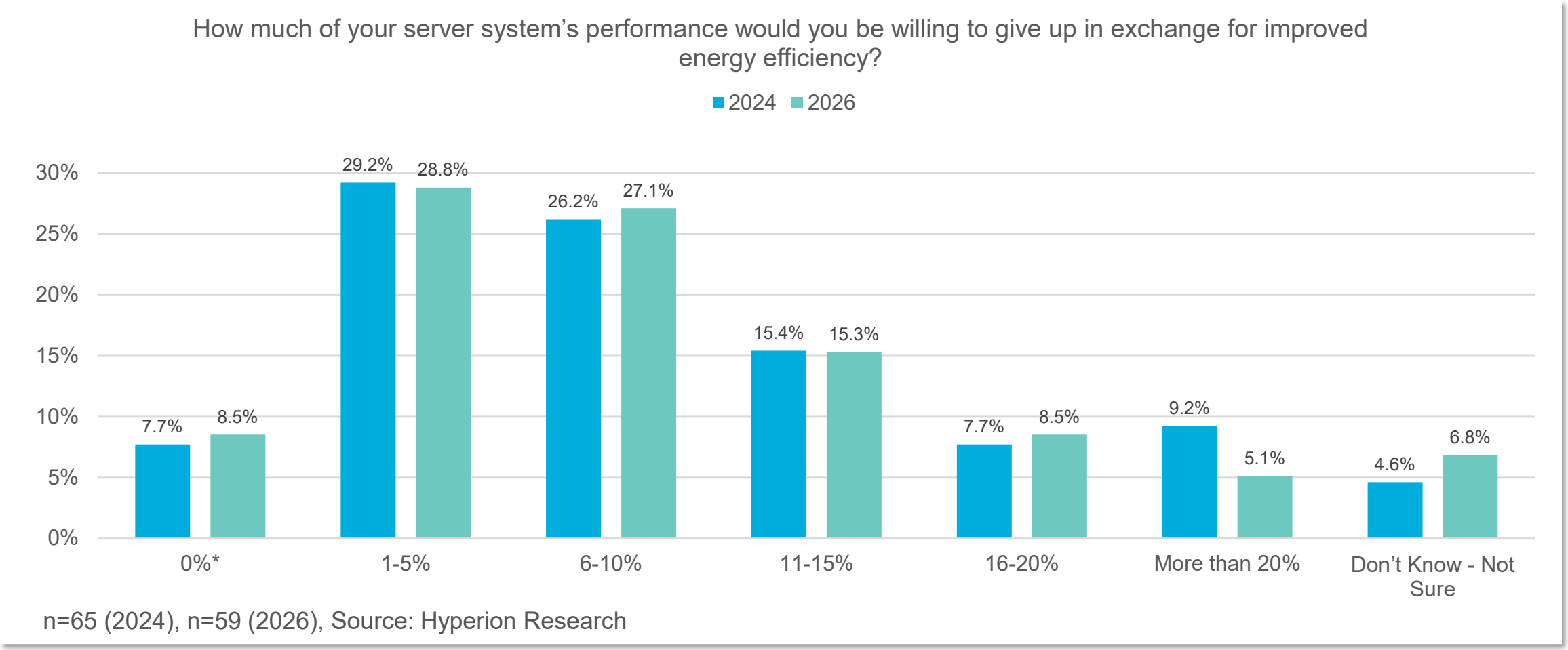
- **Other interconnects**
  - HPE Slingshot
    - Increasing line rates
    - Heavy contributions to UEC
    - Increasing promotion and visibility within the market (e.g., Slingshot workshop at SC25)
  - Eviden Bxi
    - Increasing line rates
    - Roadmap to intercept UEC
  - Cornelis OmniPath
    - 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