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

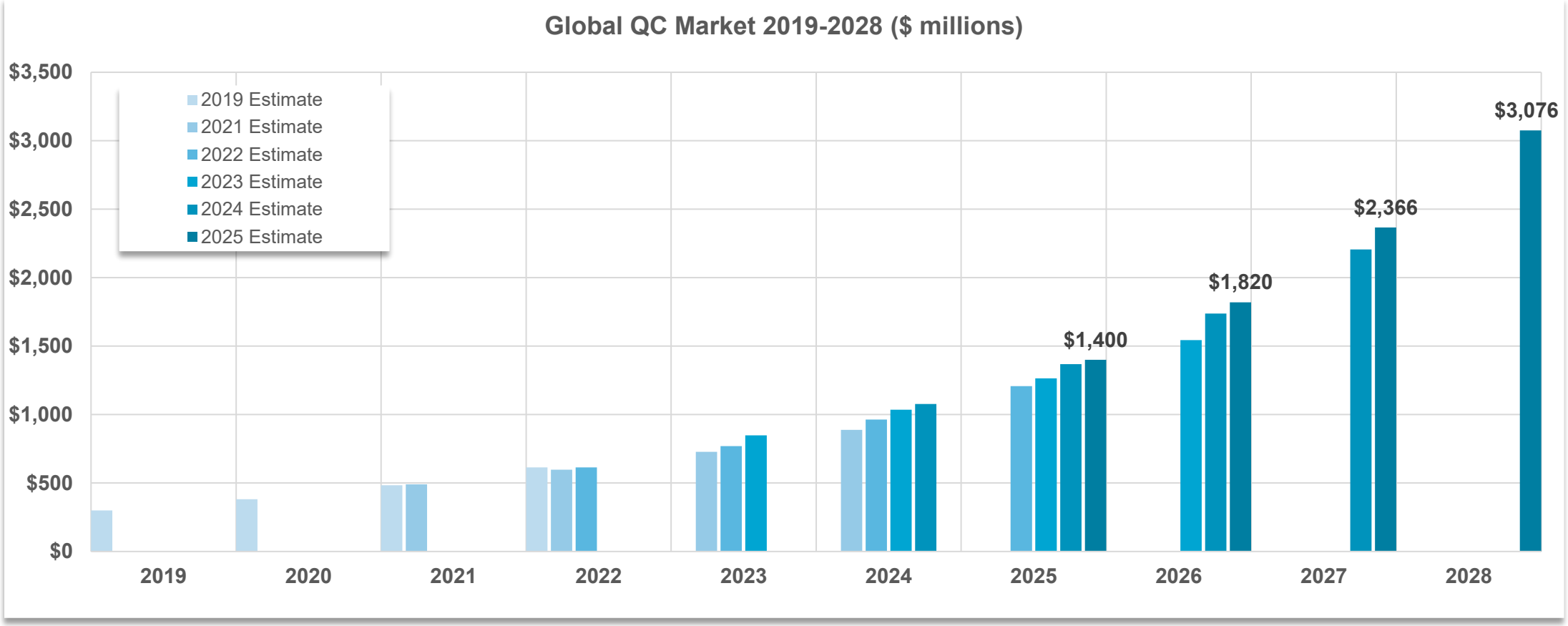
Bob Sorensen
Chief Analyst for Quantum Computing
Hyperion Research

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

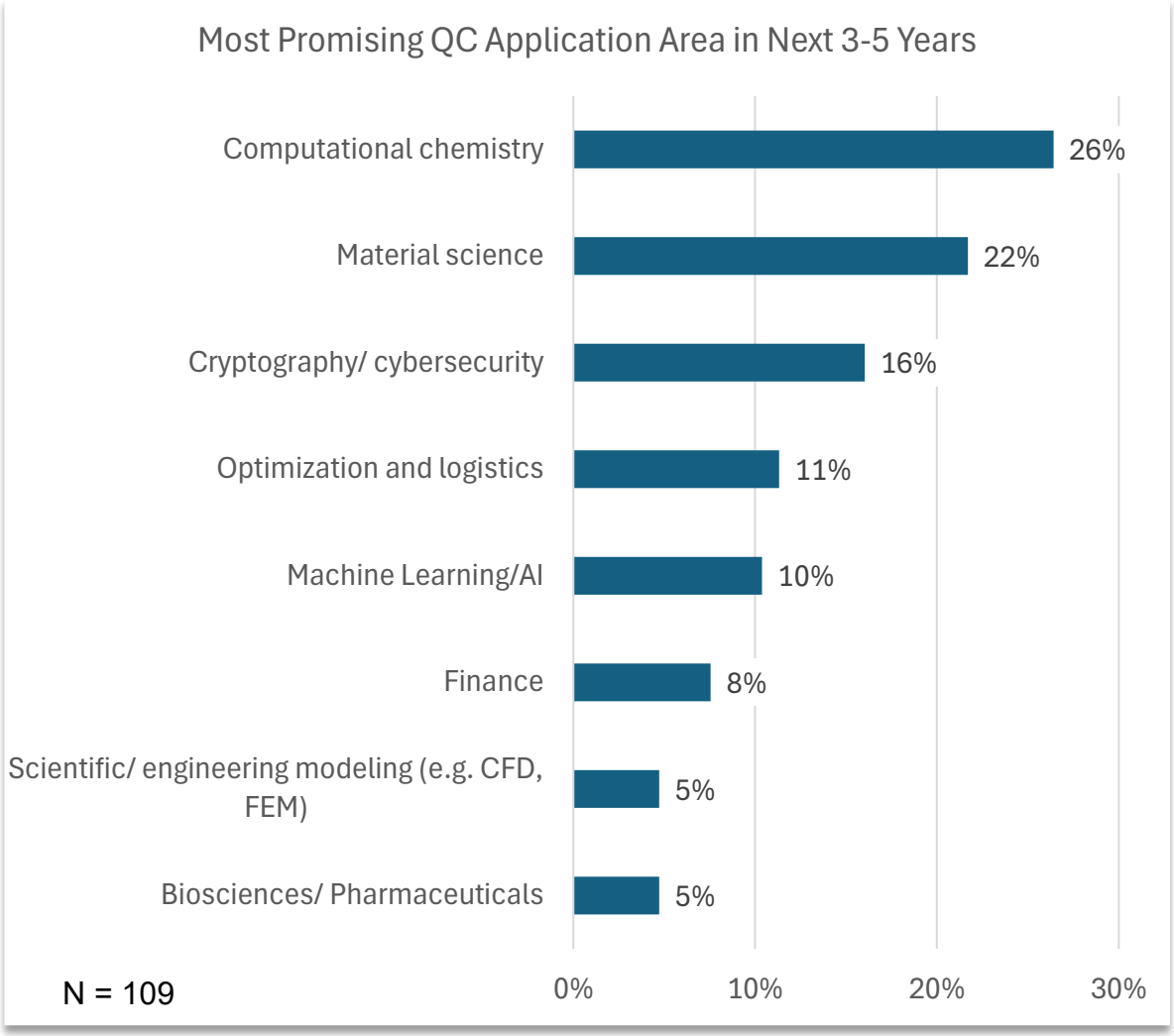
Option	% Selected
Explore new QC sector/vertical-specific QC-related opportunities	68%
Explore key performance gains over classical counterpart	42%
Explore QC sector/vertical-specific performance opportunities on existing classical workloads	39%
Explore QC/classical integration issues	36%
Field test/evaluate 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

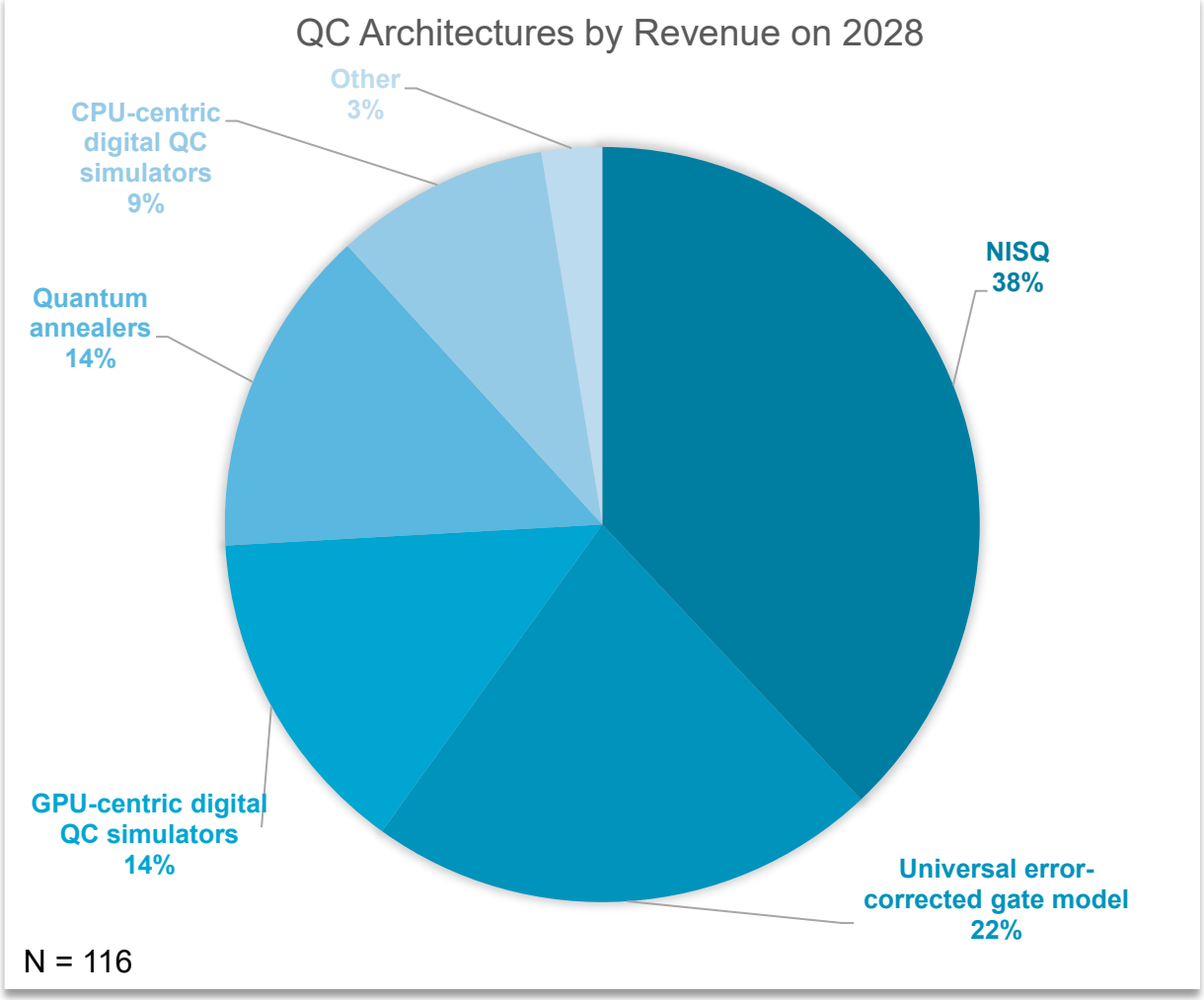
Quantum system simulations driving near-term QC applications



- 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

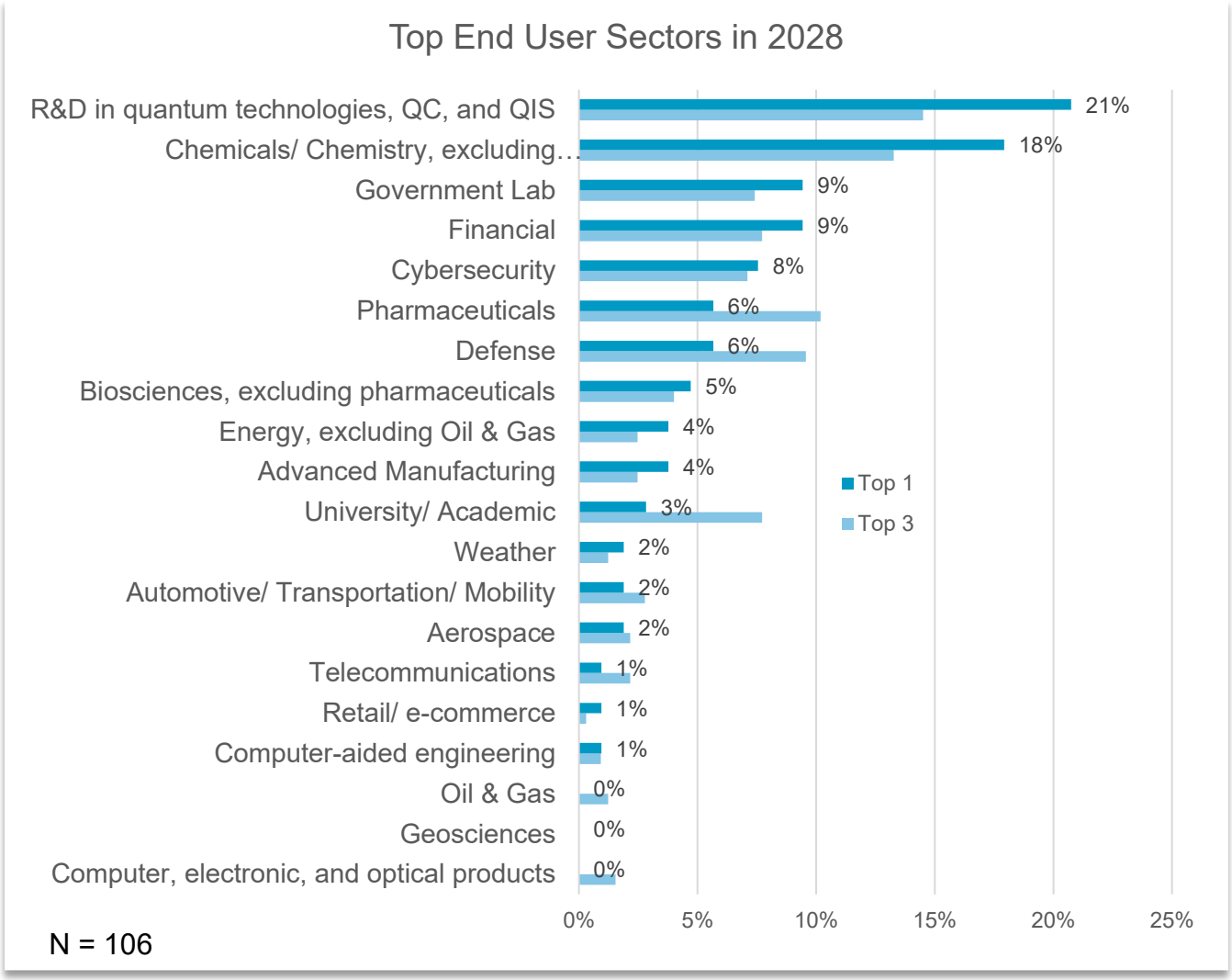
NISQ maintains lead, QC simulators still major element of QC architecture



- 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.

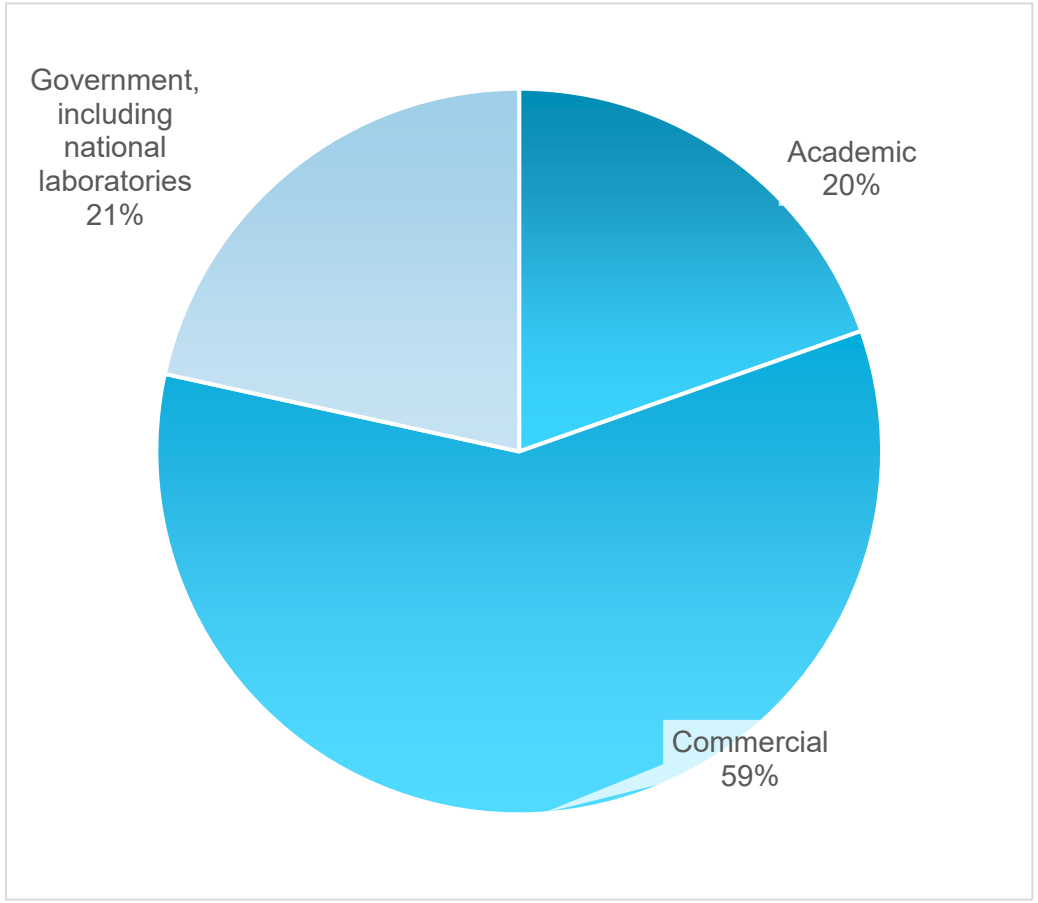
Recent Study Overviews: Gen AI ROI AI in the Cloud

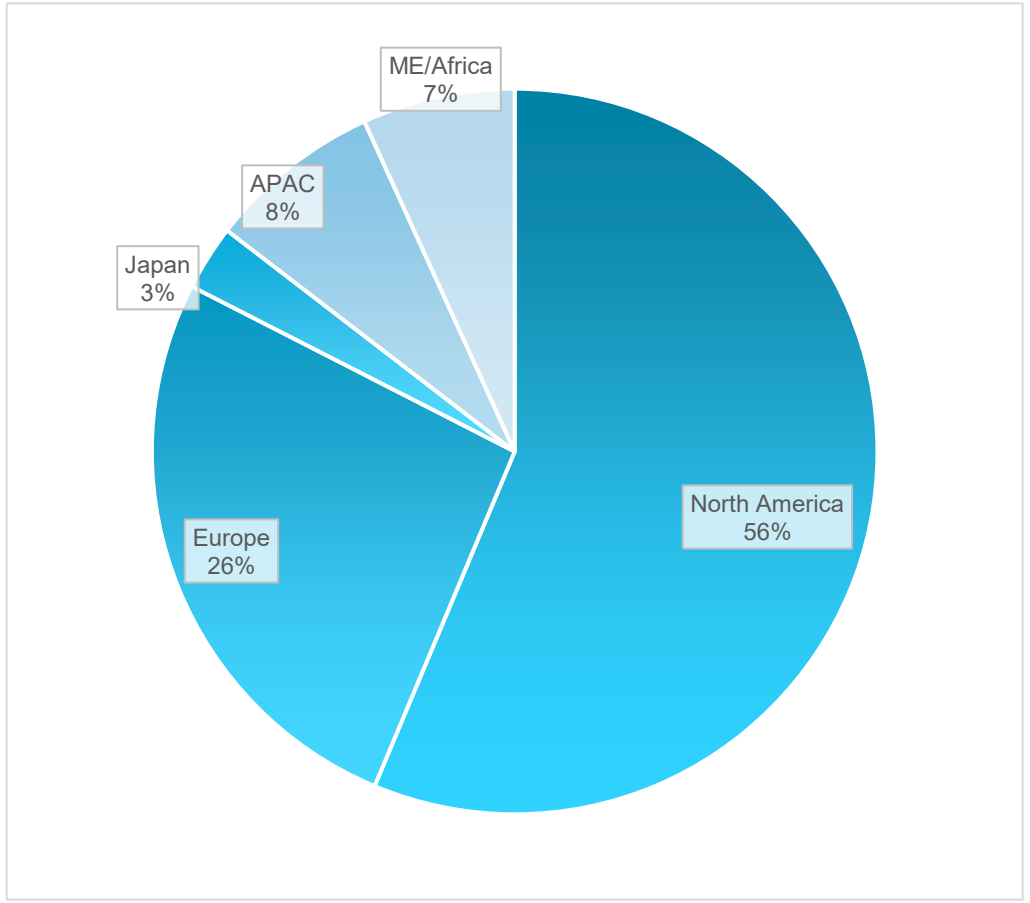
June 2026

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

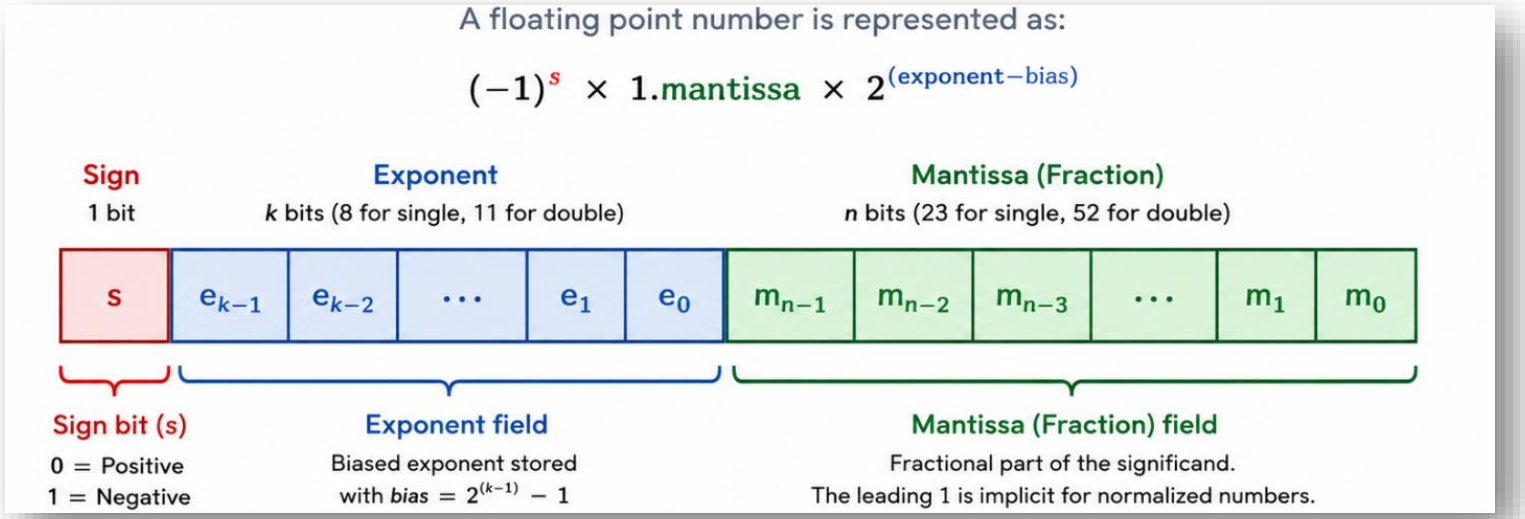
Bob Sorensen

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

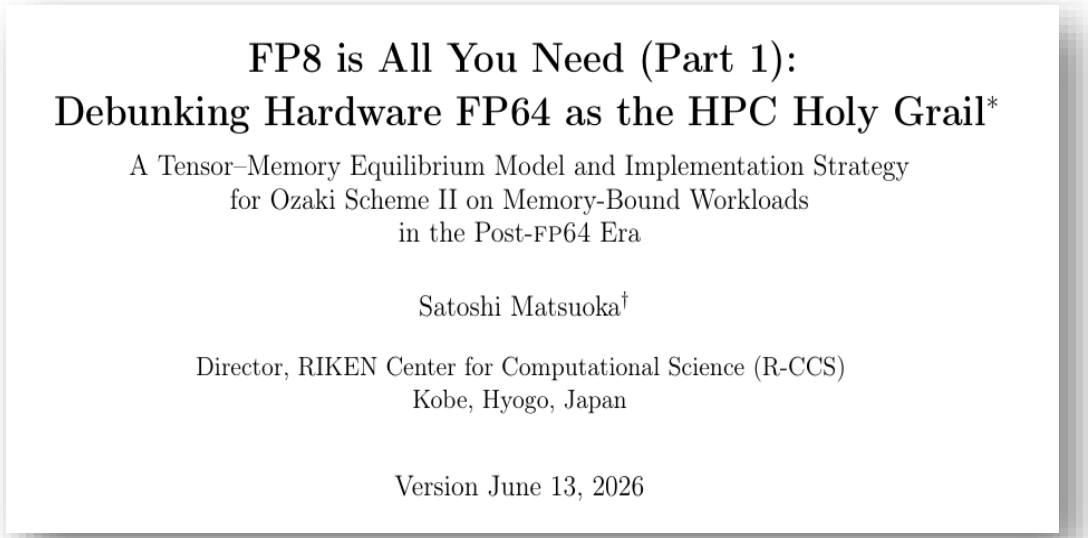
- NVIDIA FP4: Introduced with NVIDIA Blackwell GPUs
  - Uses an E2M1 layout (1 sign bit, 2 exponent bits, 1 mantissa bit) in 16-element blocks
  - Two-Level Scaling: Employs fine-grained E4M3 (FP8) scaling factors per 16-value block and a second-level FP32 scalar for the entire tensor
- Microscaling 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