Navigating the New World of HPC...and AI...and Quantum...and Clouds...

Bob Sorensen
Senior Vice President of Research
Hyperion Research, LLC

The HPC Landscape is Undergoing Significant Changes (Again)

- Multiple developments are significantly altering the overall HPC landscape
  - Applications in machine learning and deep learning
    - Generative AI is a near-term game changer
  - Computing at the edge
    - Supporting real-time decision making
  - CSP's HPC-related impact
  - End user decisions for on-premises vs. cloud vs. hybrid
  - Dark HPCs on the horizon
  - Quantum computing making strides
  - Traditional modeling and simulation remain crucial

© Hyperion Research 2024

One Size No Longer Fits All

- HPC systems are becoming more complex due to the diverse workloads they are required to address
  - Variety in workloads demand variety in architecture
  - Accurately charactering workloads in an imperative
  - Simple benchmarking is no longer enough
    - Top 500: Perhaps even harmful to the sector?
- Happily, there are technical options to address this evolving environment
  - On-prem trends: moving away from monolithic architectures towards a more flexible, heterogenous design and multiple hardware partitions (or multiple systems)
  - Cloud-based options: more image types, bare metal options, CSP-HPCaaS, experimental/novel hardware access, containers, etc.
- Sampling of options for underlying computational engines
    - Processors: x86, ARM, RISC-V
    - Superchips: Nvidia Grace Hopper, AMD MI300
    - Accelerators: Nvidia Hopper, AMD MI300, and Gaudi, Graphcore, Cerebras, SambaNova, Cambracon, Groq

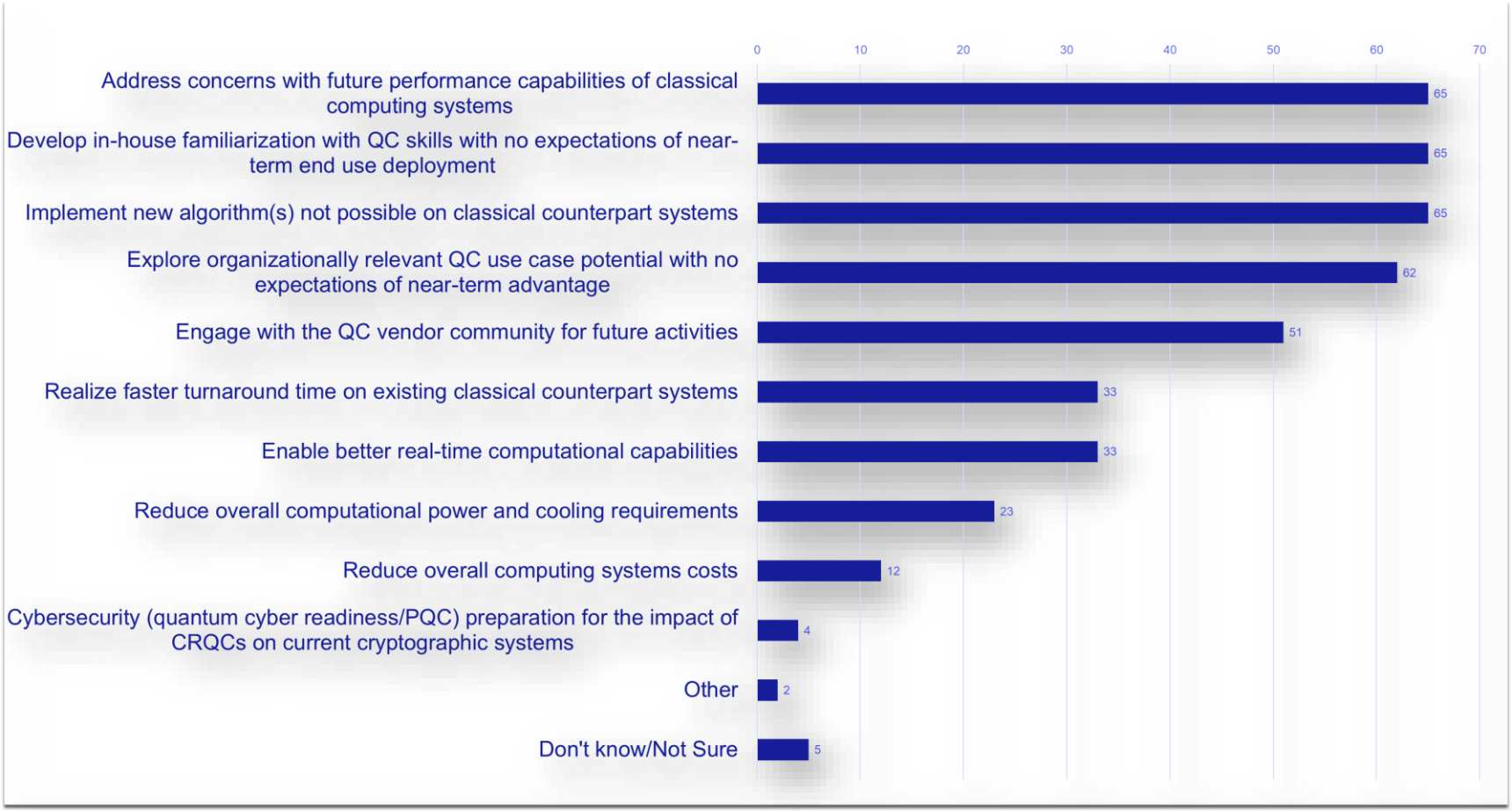
© Hyperion Research 2024

Different Workloads/Different Solutions

HPC Use Case Characteristics/Use Cases	Modeling/Simulation	Big Data/Data Science	AI: Large Language Models	Cloud-based HPC
Data Format	64-bit floating point numerical formats	64-bit floating point or integer data format	Low, mixed, or AI-specific precision formats	Variable formatting
General Code Characterization	Mix of both parallel and serial codes	Primarily parallel codes suited to cluster architectures	Distributed parallel codes, tightly coupled compute engines	Favors either small serial or large task parallel codes, loosely clustered systems
Processor and Accelerator Configuration	High core count CPU-based, GPUs support/augment CPU computation	High GPU counts, CPU managed data flows	Emphasizes GPU or related AI-specific accelerators, strong GPU-GPU interactions	Flexible node configurations of CPU/GPU/Accelerators, both virtual and bare metal options
Data Storage	Consistent, uniform storage formats, typically with large file size and consistent storage access patterns	Varying data storage formats: text, semi-structured data, structured data, binary, random data access patterns	Large numbers of small read-only files, multiple re-read iterations during training, high rate of data reuse	Supports multiple instances of varying storage configurations. Can be geographically distributed
Job Characterization	Small data input, compute intensive, large data output	Large data input, processes data at high velocity, small data output	Significant aggregate Flops count, large data input, matrix operation intensive	Wide job specifics addressable by both virtual and bare metal options
Exemplar Software	C++, Fortran, MPI	Python, Java, R, Scala, MATLAB	BERT, GPT, Megatron-Turing	Docker, Fargate, Kubernetes
Data Characterization	Program accuracy dependent on empirical validation/verification process	Verifiable data with strong statistical underpinnings	Dependent on existing training data availability/validity	Data physical location affects data size, access, performance, and price

© Hyperion Research 2024

QC 2026: QC End User Perceptions



© Hyperion Research 2024

Summary of HPC-related LLM-related Activities

	Currently	Next 12-18 months	Change Over Time
Exploring the range of potential performance enhancements by integrating LLMs into existing HPC-based workloads	58%	48%	-10%
Exploring in-house requirements for integrating LLMs into HPC-based workloads	55%	51%	-4%
Testing/assessing LLM-integrated workload performance	34%	45%	11%
Procuring access to necessary LLM software	31%	31%	0%
Reaching out to LLM hardware and software suppliers for information	30%	35%	5%
Passively monitoring LLM technology developments	27%	14%	-13%
Procuring access to necessary LLM hardware	26%	28%	2%
Standing up limited LLM-integrated pilot programs	26%	36%	10%
Porting LLM capability into existing workloads	25%	34%	9%
Running production level LLM-enabled workloads	22%	50%	28%
Standing up a fully funded LLM research efforts	17%	27%	10%
No current activity	1%	0%	-1%
Other	1%	0%	-1%

© Hyperion Research 2024

Thanks

- Questions and comments are most welcome and can be sent to bsorensen@hyperionres.com

© Hyperion Research 2024