Improving U.S. Competitiveness in the Global High Performance Computing and AI Markets:

Recent Changes, Concerns & Potential Impacts from The CHIPS and Science Act

www.HyperionResearch.com
www.hpcuserforum.com

December 13, 2022

About Hyperion Research

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

Hyperion Research mission:

- *Hyperion Research helps organizations make effective decisions and seize growth opportunities*
  - *By providing research and recommendations in high performance computing and emerging technology areas*

HPC User Forum mission:

- *To improve the health of the HPC/AI/QC industry*
  - *Through open discussions, information sharing and initiatives involving HPC users in industry, government and academia along with HPC vendors and other interested parties*

The Hyperion Research Team

Analysts

- Earl Joseph, CEO
- Bob Sorensen, SVP Research
- Mark Nossokoff, Research Director
- Alex Norton, Research Director
- Jaclyn Ludema, Analyst
- Melissa Riddle, Associate Analyst
- Thomas Sorensen, Associate Analyst
- Cary Sudan, Survey Specialist

Operations

- Jean Sorensen, COO

International Consultants

- Katsuya Nishi, Japan and Asia
- Jie Wu, China & Technology Trends

Global Accounts

- Rene Copeland, Dir. Business Development
- Mike Thorp, Sr. Global Sales Executive
- Kurt Gantrish, Sr. Account Executive

Data Collection

- Sue Sudan, Market Data Group
- Kirsten Chapman, KC Associates

Today's Agenda

9:00--9:15	Why HPC is Important to Science and Industry
9:15--9:50	An Overview of the HPC, AI and QC Markets
9:50--10:15	A Global Perspective on HPC Focusing on China
10:15--10:45	Part 1: U.S. Vendors Discuss the Impact of Act: HPE, AMD & IBM
10:45--11:00	Break
11:00--11:30	The role of the Department of Commerce/NIST in Implementing the CHIPS and Science Act
11:30--12:00	How NSF is Implementing The CHIPS Act & Why HPC Leading-edge HPC Tools are Important to Support & Advance Science
12:00--12:30	Ideas for Advancing Science via New Approaches to HPC/AI
12:30--1:15	Lunch
1:15--1:45	Compelling Science Examples from NASA
1:45--2:15	An Overview of the CHIPS and Science Act
2:15--3:15	Part 2: U.S. Vendors Discuss the Impact Act: Intel, Qualcomm, Microsoft, Goodyear, CATALOG, …
3:15--3:30	Break
3:30--4:00	Sustaining U.S. High-Performance Computing Leadership in an Exascale Era
4:00--4:25	Group Discussion: What Should We Track & How to Better Measure U.S. Competitiveness
4:25--4:30	Next Steps and Action Items

Some Background & Definitions

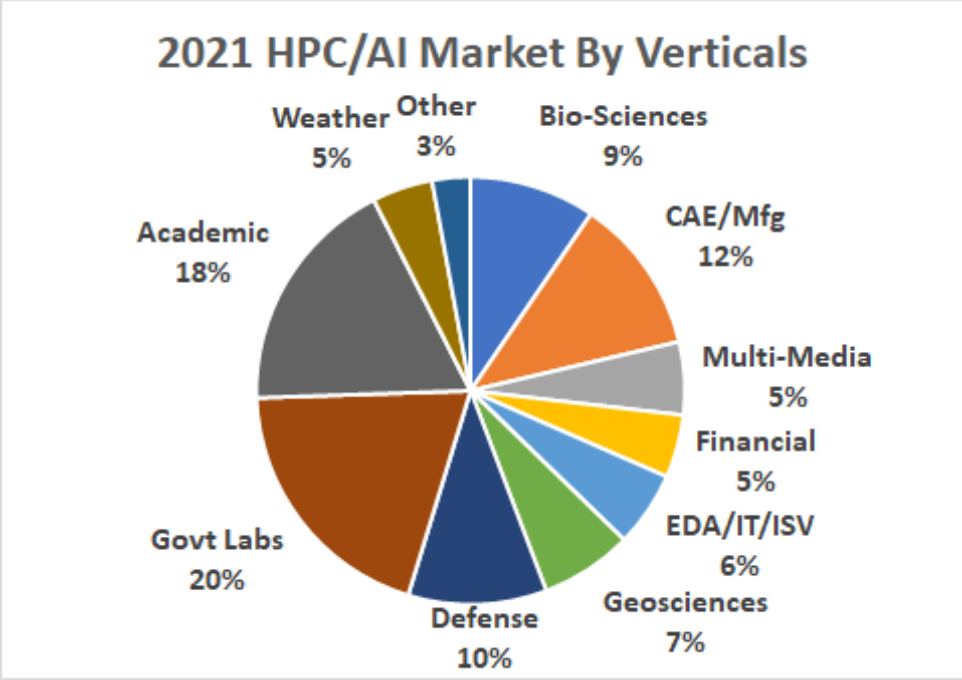
HPC (aka Technical Computing) is defined as:

The market for computer servers used by scientists, engineers, analysts, and other groups for computationally intensive and/or data-intensive modeling, simulation and analysis applications.

- It includes AI, Quantum and most emerging technologies
- Technical servers range from small servers costing less than $10,000 to the large-capability machines valued in hundreds of millions of dollars

Some Background & Definitions

- **Definition of US competitiveness for the day:**
  - Industry = growing revenues and jobs
  - Science = keeping the US ahead in science and new research
- **What is included in HPC/AI?**
  - Traditional HPC -- modeling, simulation and analysis
  - AI, ML, DL
  - Quantum computing and all types of advanced technology
- **Who uses HPC/AI?**



Why is HPC/AI important?

1. **To conduct R&D, engineering, financial analysis, make movies, forecast the weather, etc.**
   - To make scientists, engineers and analysts more productive
   - To develop better products and processes
2. **To research things that are physically impossible to do (or highly undesirable to do)**
   - **Physically Impossible:** landing a spacecraft on mars, researching the big bang, modeling the future climate, …
   - **Undesirable:** testing nuclear explosions, testing vaccines on people, testing cars with people inside, …
3. **To help address major issues facing society:**
   - Finding vaccines for new diseases
   - Forecasting climate disasters (hurricanes, flooding, etc.)
   - Understanding and developing solutions for climate change
   - Developing lower impact energy solutions
   - Detecting sophisticated cyber security breaches
   - Addressing economic challenges for everyone

Examples of How HPC/AI is Used

Switch to the videos

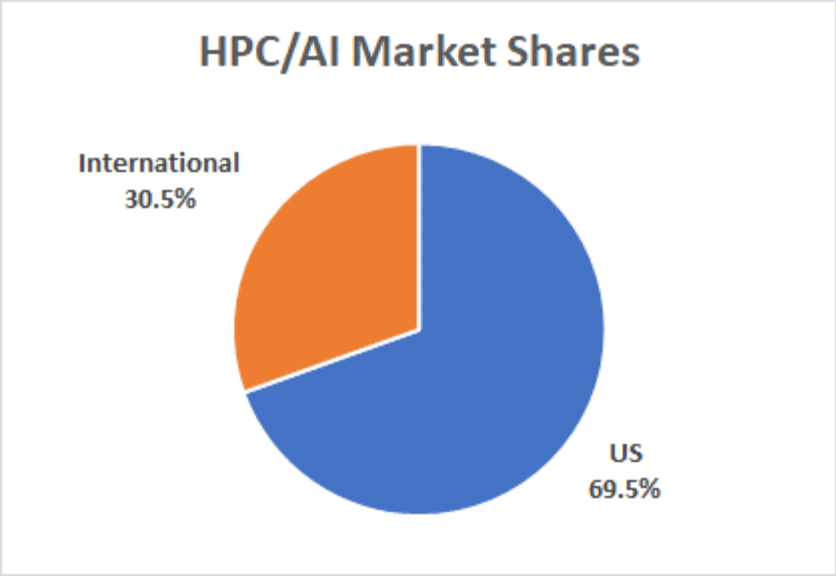
Today's Agenda

9:00--9:15	Why HPC is Important to Science and Industry
9:15--9:50	An Overview of the HPC, AI and QC Markets
9:50--10:15	A Global Perspective on HPC Focusing on China
10:15--10:45	Part 1: U.S. Vendors Discuss the Impact of Act: HPE, AMD & IBM
10:45--11:00	Break
11:00--11:30	The role of the Department of Commerce/NIST in Implementing the CHIPS and Science Act
11:30--12:00	How NSF is Implementing the CHIPS Act & Why HPC Leading-edge HPC Tools are Important to Support & Advance Science
12:00--12:30	Ideas for Advancing Science via New Approaches to HPC/AI
12:30--1:15	Lunch
1:15--1:45	Compelling Science Examples from NASA
1:45--2:15	An Overview of the CHIPS and Science Act
2:15--3:15	Part 2: U.S. Vendors Discuss the Impact Act: Intel, Qualcomm, Microsoft, Goodyear, CATALOG, …
3:15--3:30	Break
3:30--4:00	Sustaining U.S. High-Performance Computing Leadership in an Exascale Era
4:00--4:25	Group Discussion: What Should We Track & How to Better Measure U.S. Competitiveness
4:25--4:30	Next Steps and Action Items

On-Premise HPC Server Sales

U.S. vendors have 69.5% of servers sold

- **HPC servers generated over $14B in sales in 2021**
  - ~$30B when storage, software and services are added
- **The U.S. has been the market leader for decades**
- **Other nations are investing in domestic suppliers to gain economic value from this market**
  - China, Europe and Japan



2021 HPC/AI Server Sales ($M)		
Vendors	**US**	**International**
HPE	5,048	
Dell Technologies	3,196	
Lenovo		1,166
Inspur		993
Sugon		518
IBM	463	
Atos		520
Fujitsu		190
NEC		173
Penguin	378	
Others	1,149	927
Total	10,234	4,488
Market Share	**69.5%**	**30.5%**

On-Premise HPC Processor Market

U.S. vendors sell over 95% of the processors sold

- **Over 4 million base processors and over 330,000 GPUs/accelerators were sold in 2021**
- **The U.S. has led the processor components in HPC/AI for decades**
- **Other countries are investing to gain market share**
  - China, Europe, and Japan

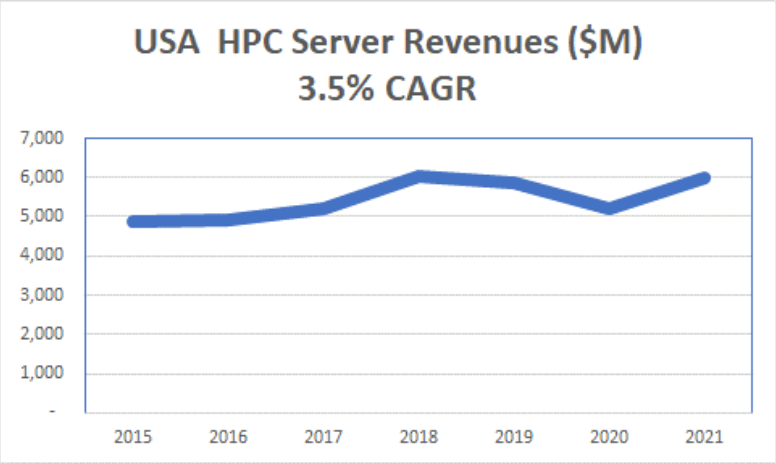
2021 HPC/AI Processors Sold		
Vendors	**US**	**International**
Intel & AMD	3,906,211	
IBM	100,500	
Custom	97,916	103,845
ARM	9,582	13,789
Total	4,114,209	117,634
Market Share	**97.2%**	**2.8%**

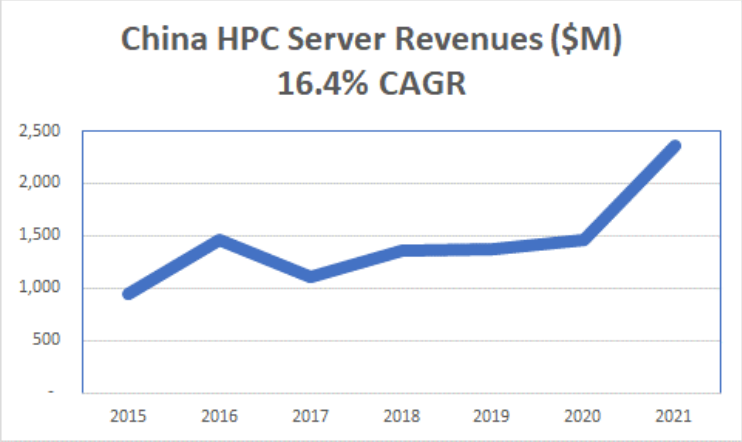
2021 GPUs & Accelerators Sold		
Vendors	**US**	**International**
Nvidia	325,108	
Others	4,714	4,215
Total	329,823	4,215
Market Share	**98.7%**	**1.3%**

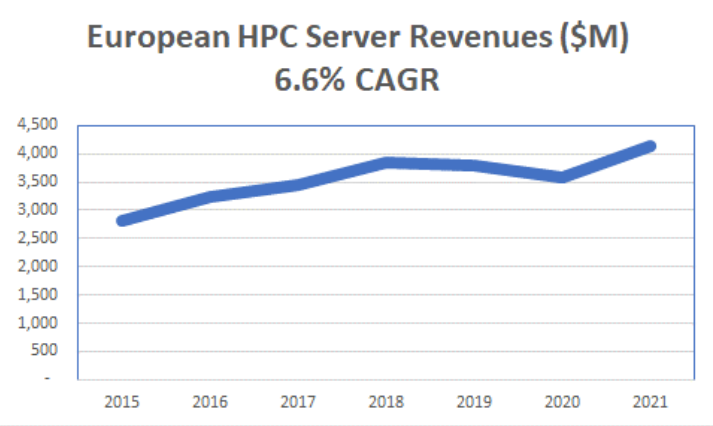
Historic Growth Rates In HPC

U.S. is the largest market, but hasn't been growing much

- **China has seen the largest growth at 16.4%**
- **Europe has been growing faster than the U.S.**







The Exascale Market: Highly Competitive

Over 30 systems and over $10 billion in value

Exascale and Near-Exascale Leadership Systems (2020 to 2027)

Year Accepted	China	Europe	Japan	US	Other Countries*	Total Systems	Total Value
2020			1 near-exascale system ~$1 B			1	$1.0B
2021	2 exascale ~$350M each	1 pre-exascale system ~$180M	?	1 pre-exascale system ~$200M	--	4	$1.1B
2022	1 exascale ~$350M each	2 pre-exascale systems ~$190 each	1 near-exascale system ~$150M	1 exascale systems ~$600M	--	5	$1.5B
2023	1 exascale system ~$350M	1 or 2 pre-exascale systems ~$150M each	1 near-exascale system ~$150M	1 or 2 exascale systems ~$600M each	--	4-5	$1.8B - $2.4B
2024	1 exascale system ~$350M	1 exascale ~$500M, plus 1 exascale (or pre) systems ~$200 M	?	1 or 2 exascale systems ~$400M each	1 exascale system ~$200M	4-6	$1.2B - $1.9B
2025	1 or 2 exascale system ~$300M each	1 or 2 exascale systems ~$350M each	1 exascale system ~$150M	1 or 2 exascale systems ~$350M each	1 exascale system ~$150M	5-8	$1.3B - $2.3B
2026	1 or 2 exascale system ~$300M each	1 or 2 exascale systems ~$325M each	?	1 or 2 exascale systems ~$350M each	1 or 2 exascale systems ~$150M each	4-8	$1.1B - $2.2B
2027	1 or 2 exascale systems ~$250M each	1 or 2 exascale systems ~$300M	?	1 or 2 exascale systems ~$300M each	1 or 2 exascale systems ~$150M each	4-8	$1.0B - $2.0B
Total	**8-11**	**8-12**	**4**	**7-12**	**4-6**	**31-45**	**$10B - $14B**

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

Source: Hyperion Research, July 2022

State of Play in AI

December 13, 2022

www.HyperionResearch.com
www.hpcuserforum.com

Alex Norton
anorton@hyperionres.com

What We Mean When We Say AI

Hyperion Research focuses on the pointy end of AI

- **What we track:**
  - As with HPC, track the highest end of AI
  - Focus primarily on the training of large models of all kinds
  - Track both stand alone AI applications and research as well as the application of AI to support traditional mod/sim
- **AI's impact on the technology ecosystem:**
  - Rise of GPUs and other accelerators
  - New system architectures
  - New deployment models and resource models
- **Growth of ML and promise of DL**
  - Emergence of other popular methodologies like reinforcement learning, transfer learning, foundation models, etc.
  - Three years ago, significant efforts and hype around deep learning, but that has been scaled back as transparency and reproducibility issues impact down-stream applications

AI's Importance

The promise of AI can be found in a myriad of places

- **In scientific domains, AI presents real opportunity to push beyond the boundaries of experimentation and physical observation to explore new areas of discovery**
- **Interesting work has been done in important areas of science, like:**
  - Drug discovery and virus response, especially Covid-19
  - Weather and climate forecasting
- **AI is also critical in furthering technological advancements in real-world applications**
  - Automated driving
  - Security applications
  - Battlefield operations (e.g., drones and satellite use in Ukraine now)

Impact of AI on HPC Datacenters

AI impacting the future roadmaps of most HPC datacenters

- **As AI rises to new levels of importance in HPC datacenters worldwide, technology roadmaps of future HPC datacenter deployments are evolving:**
  - Incorporation of GPUs for AI training
    - Most future systems planned to incorporate GPUs or other accelerators to run AI training workloads
  - Modernization of legacy HPC code
    - Exploit GPU capabilities when not being used for AI apps
  - Different system choices
    - Some choosing to build specialized AI systems for dedicated training capabilities
  - Use of cloud increasing to handle AI workloads
    - Access to flexible GPU resources
    - Tools and services available on the cloud
    - Access to public/shared data sets not available on-prem
    - Expertise of CSPs/hyperscalers

The Divergence of Types of AI Users

Two distinct types of users emerging in the AI ecosystem

- **Experimental users**
  - Just starting their AI journey
  - Working on small scale models
  - Can be in research or in industry
  - Not necessarily pushing scale of systems or dataset size

- **AI researchers pushing the boundaries of AI**
  - Billions to trillions of parameters in models
  - Models take larger teams to train over weeks to months
  - Handful of capable organizations at the top with:
    - Adequate funding for AI research
    - Infrastructure to handle scale of training
    - Amply sized team to pursue new AI capabilities
  - Can overlap with science, but recently are training more general models with extensive down-stream applications

AI Technology Trends

New accelerator market targeted at AI applications

- **A few years ago, a number of companies emerged to address the unmet needs of AI acceleration**
  - Few companies have become successful
  - Many were acquired or ceased to be a company altogether
  - According to the State of AI Report from 2022, the “Big 5 Startups” for AI processors are Cerebras Systems, SambaNova, Graphcore, Habana, and Cambricon
  - The combined private valuation of Cerebras, SambaNova, and Graphcore does not exceed the annualized datacenter revenue in 2021 by NVIDIA[1]
- **NVIDIA maintains solidified leadership in the GPU ecosystem today, despite challengers:**
  - AMD’s GPU line chosen for two US Exascale machines, and is gaining some traction in the market
  - Intel preparing to bring the Datacenter GPU Max to the market soon for general availability

1: https://www.stateof.ai/

Highlights from Applied AI Conference

ChatGPT to one million users in five days because of approachability and power of application

- **Two major themes persisted through the talks attended at the conference:**
  - Ethical and responsible AI in end use applications
  - Approachability of AI and integration into business operations at scale
- **Days before the conference, announcement and release of both GPT3.5 and the ChatGPT module**
  - ChatGPT reached one million users in five days
  - Fastest of any platform on the web
  - ChatGPT is an approachable representation of the capability of AI
- **With approaching EU regulations on AI and data, many businesses looking to solidify data governance and security protocols, as well as ethical use of data for AI applications**

AI in the US Versus China

China's growth outpacing the world in publishing AI papers

- **China continues to publish more work on AI than the United States**
  - Investing heavily in areas such as research in image and video recognition, automated driving, and smart cities
  - According to State of AI Report 2022, Tsinghua University published the most AI documents of any organization worldwide[1]
  - US published more papers than China in 2022, but China's growth much higher than the rest of the world
- **In wake of recent US restrictions on high-tech to Chinese organizations, new processor developments have emerged**
  - Biren Technologies is bringing a Chinese-developed GPU to the market after high end GPUs from US vendors banned from Chinese market
  - Other smaller organizations developing Chinese-developed processors to enhance an indigenous technology base in China

1: https://www.stateof.ai/

HYPERION RESEARCH

QUESTIONS?

anorton@hyperionres.com

A Whirlwind Tour of the Quantum Computing Sector

Dec 13, 2022

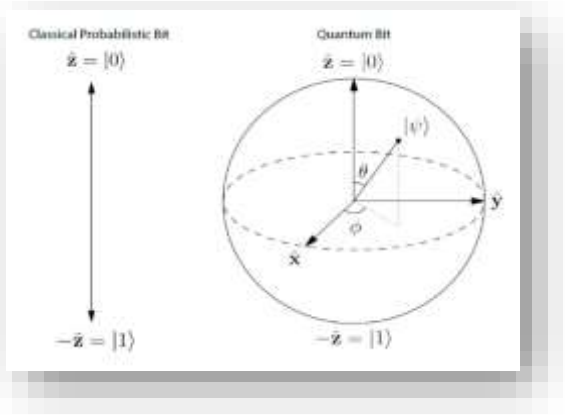
www.HyperionResearch.com
www.hpcuserforum.com

Bob Sorensen
Chief Analyst for Quantum Computing

Currently, the Promise of QC is Substantial

- **QC systems have the potential to exceed the performance of conventional computers for problems of importance to humankind and businesses alike in areas such as:**
  - Physical Simulation
  - Machine Learning
  - Optimization
- **And the list grows longer each day**
- **Initial Buzz: Demonstrate so-called quantum supremacy using a programmable quantum device to solve a problem that no classical computer can solve in any feasible amount of time**
  - For example: Factoring a large number into its two prime integers
- **But this is not the main goal today**
  - Seeking quantum advantage
  - Or even quantum competitive advantage

A Flavor of the Complexity



Strings of 1s and 0s

[0,0,0]

Input (Classical)

Site of Quantum Goodness

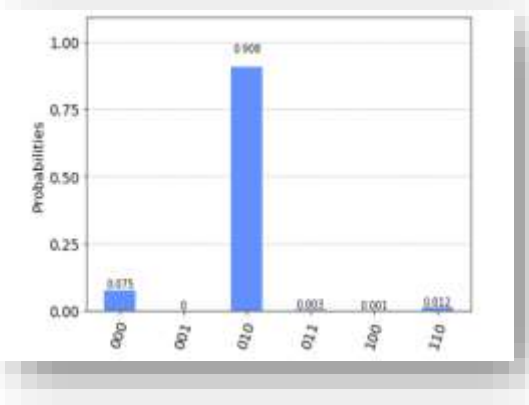
- Superposition
- Entanglement
- Quantum Logic Gates

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$$

$$|\alpha|^2 + |\beta|^2 = 1$$

Processing (Quantum)

Probability of Strings of 1s and 0s



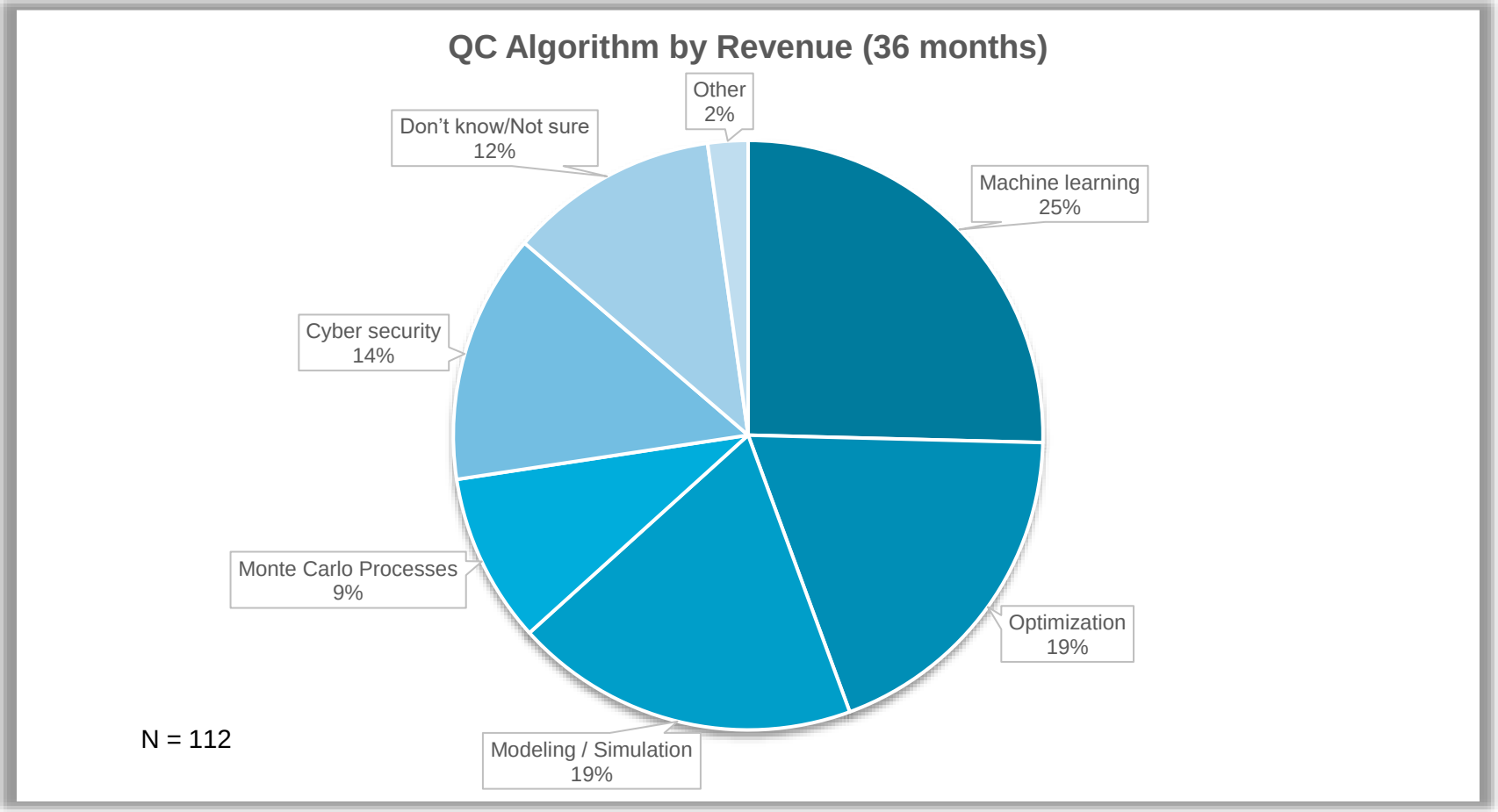
Output (Classical)

Substantial Challenges Ahead

- **Challenges include:**
  - Formidable technical issues in QC hardware and software
  - Uncertain performance gains
  - Unclear time frames
  - Disorganized progress in algorithm/application development
  - Funding continuity
  - Looming workforce issues
- **All these factors complicate treating QC as a stable market alongside more traditional IT sectors**
  - Making a business case is tough...but it needs to be done

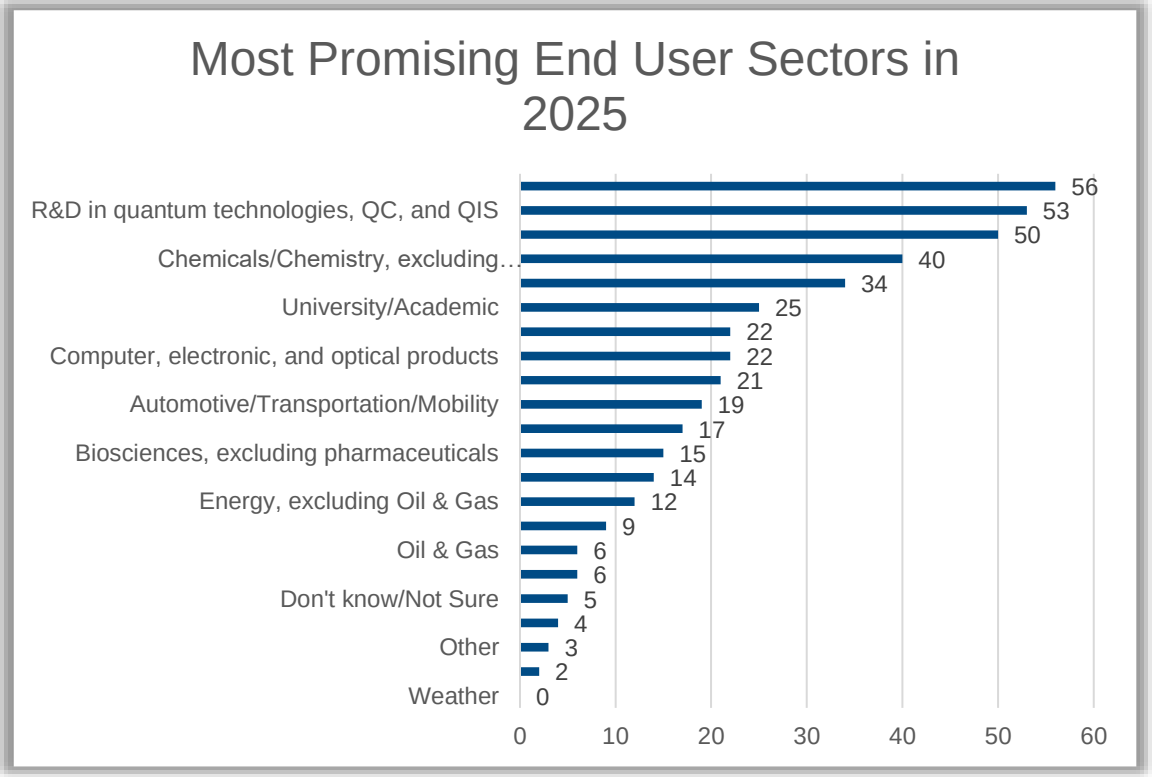
QC Algorithm by Revenues

ML & optimization, M/S and cybersecurity close behind



QC Market 2025: Top Three End User Sectors

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

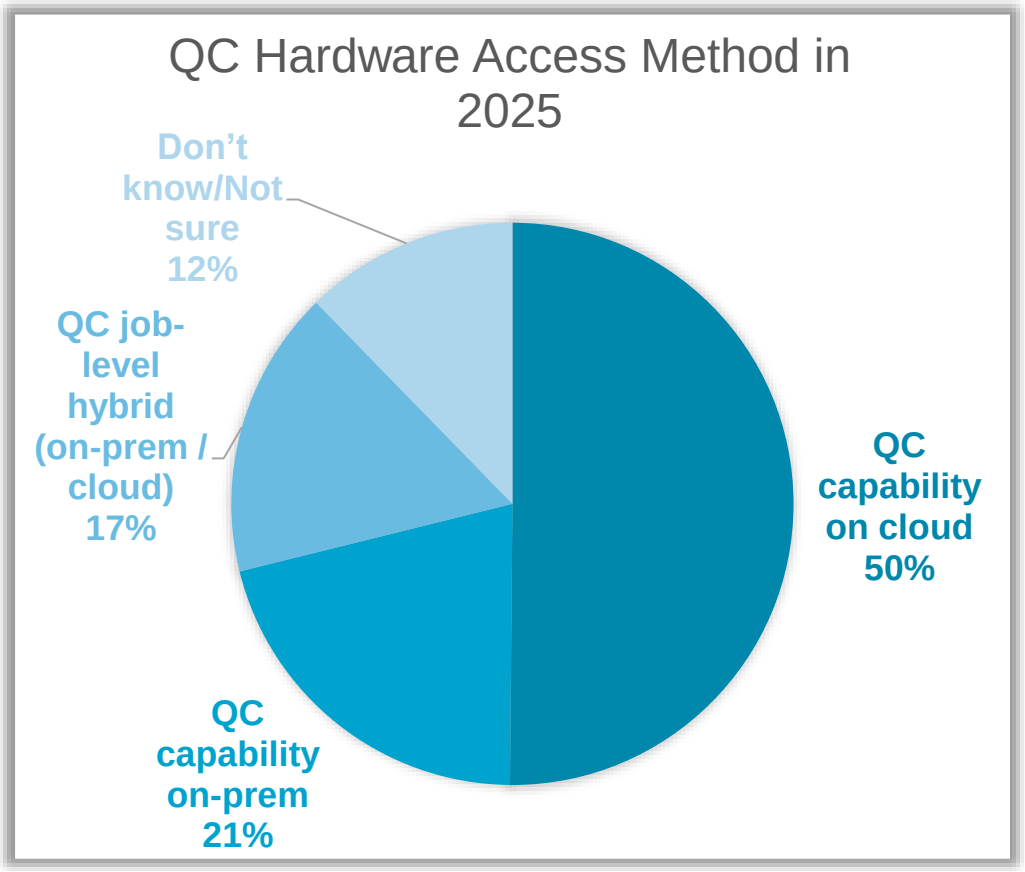


N = 145, Select top three

- **Financial chosen by one in three respondents as a top three most promising QC end user sector**
  - **But only narrowly ahead of QC R&D and cybersecurity**
- **Nearly every sector choice deemed important by some**
  - **Crossing academic, commercial, and government spaces**

QC Market 2025: QC Access Method

Cloud continues to dominate as preferred QC hardware access method



N = 145

- **Cloud supports half of total QC hardware access**
  - **Combined with hybrid, cloud involved in 67% of total QC hardware market access**
- **No major changes from last year's study**
  - **On-prem holds steady at ~20%**
  - **Slight increase in pure cloud-based option over hybrid alternative**

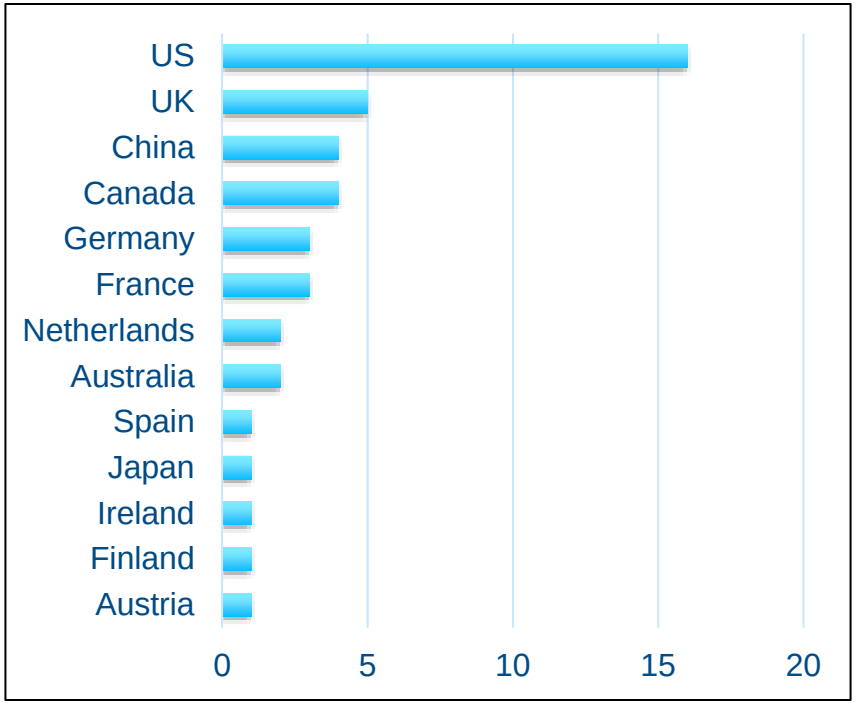
A Growing Collection of Quantum Computing HW and SW Suppliers

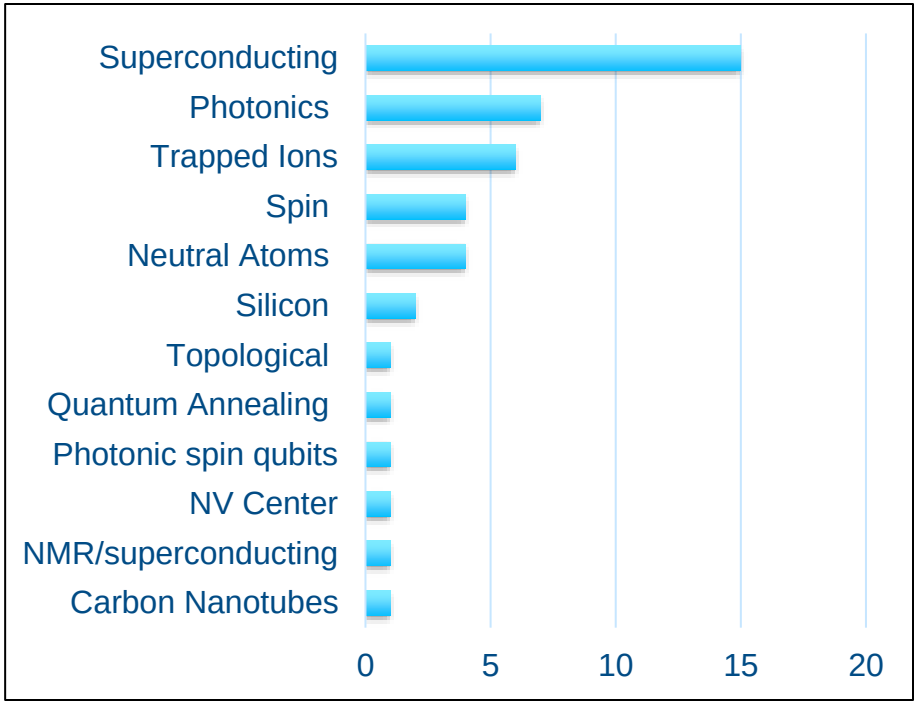
- **A wide and diverse range of QC hardware suppliers have emerged to populate a growing QC ecosystem**
  - Legacy players (IBM, Atos, NEC, Fujitsu)
  - Integrated player (Quantinuum)
  - New entrants:
    - Pureplay: D-Wave, IonQ, Rigetti, Infleqtion, Quantum Circuits Inc., Xanadu, IQM, Pasqal, etc.
    - Component players: Intel
  - Non-traditional players (Alibaba, AWS, Baidu, Google, Microsoft)
  - Myriad stealth players
- **The software stack is moving to specialize, e.g.:**
  - Strangeworks: turn-key software integration services
  - Classiq: quantum gate complier/optimizer
  - 1Qbit: QC-based AI/optimization software

Current Field of Commercial QC Hardware Suppliers

Tracking the visible players from a global perspective

- **44 identified QC hardware developers**
- **12 quantum modalities under consideration**



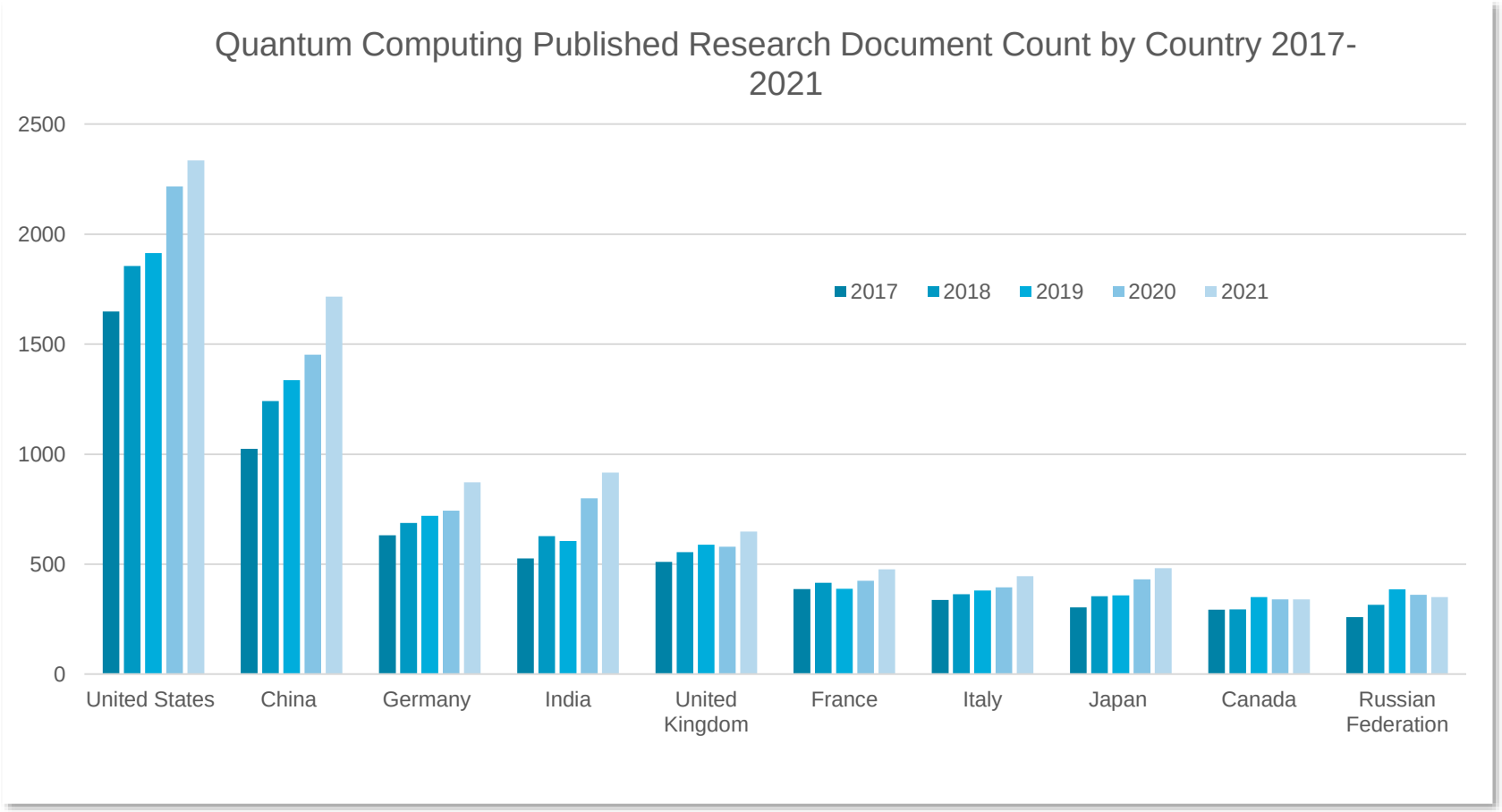


Select National/Regional Government Quantum Programs

- **Canada: Quantum Science Funding Framework**
- **China: Key National R&D project, Quantum Control and Quantum Information**
- **EU: The Quantum Flagship**
- **France: Quantum: the technological shift that France will not miss**
- **Germany: Government Framework Programme for Quantum Technologies**
- **Japan: Q-LEAP**
- **Russia: Digital Economy National Program**
- **UK: National Quantum Technologies (UKNQT) Programme**
- **US: National Quantum Initiative Act**

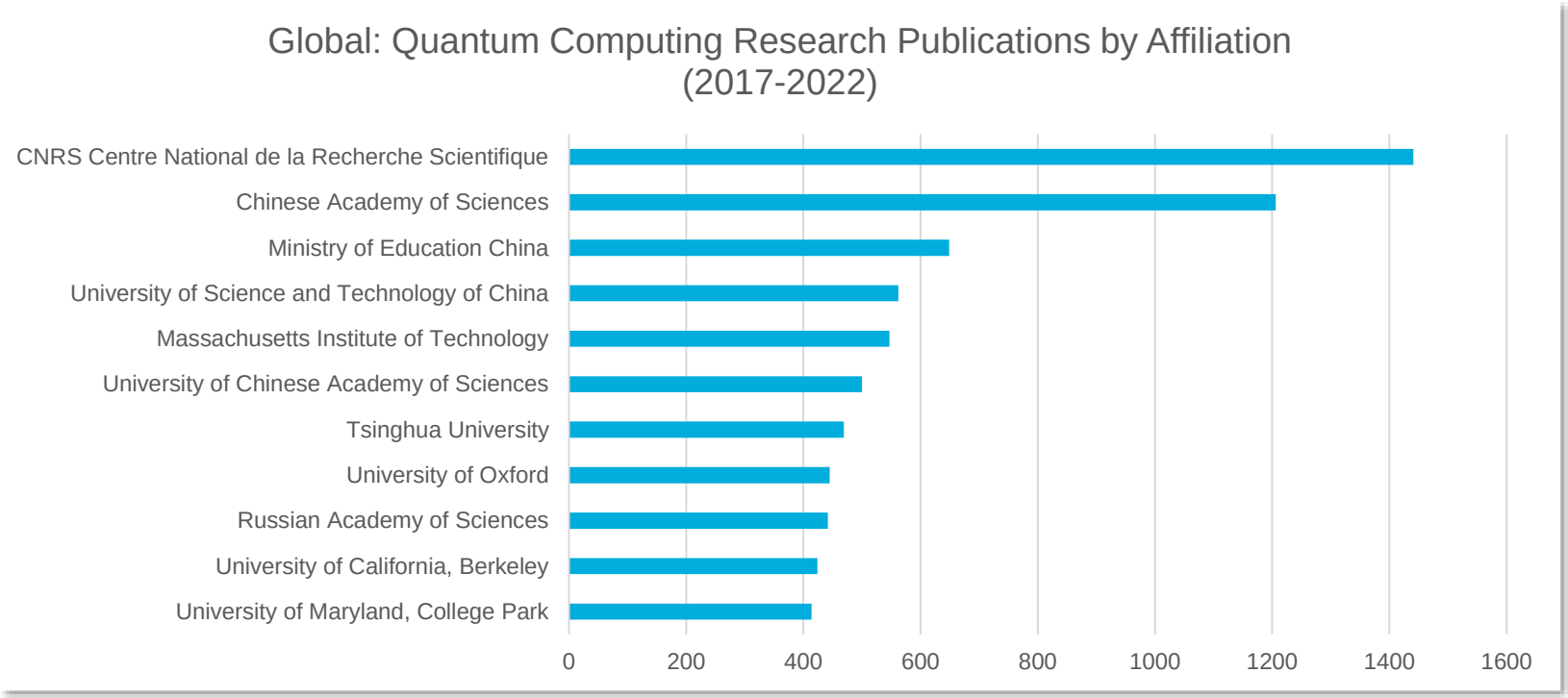
QC Research Activity From a Global Perspective

Wide range of efforts on the rise



Hyperion Research 2022

Leading QC Research Organizations



- **National Affiliation**
  - China: 5 – CAS, n.b. composed of 114 research entities
  - Russia: 1, RAS composed of 400 research entities
  - US: 3, France: 1, UK: 1

Hyperion Research 2022

Some Closing Thoughts

Time permitting

- **Non-US QC participation in the global QC supply and market ecosystem appears to be on the rise**
  - Activity mainly remains local -- for now
- **QC professional services critical contributor**
  - Mix of QC and non-QC-specific consulting entities
- **As in HPC, large government procurements could soon alter the trajectory of the sector**
  - A double-edged sword in the making?
- **Questions about continued VC investment (currently at an impressive magnitude and rate of growth) is causing no small degree of angst**
  - Many private firms will continue to rely on private funding instruments
  - Government role needs to be thoughtful, stable, and long-term

HYPERION RESEARCH

QUESTIONS?

bsorensen@hyperionres.com

Insufficient facts always invite danger.

- Spock, *Stardate: 3141.9.*

HYPERION RESEARCH

What U.S. Industry Could Use To Be More Competitive

From a Previous Hyperion Research Study: On What U.S. Industry Could Use

If you could create a "Wish List" for U.S. Government investments in new HPC technologies, what would your organization like to see improved the most to become more competitive?

1. Lead in developing lower power computers and components
2. Help improve security via standards and R&D to improve security at all levels
3. Lead co-design and application design/re-design to make HPC systems more powerful and easier to use
4. Help improve the scalability of systems & making them more robust at scale
5. Improve big data & storage capabilities, costs and power consumption
6. Invest in creating better performing designs & more cost-effective systems
7. Provide funding and technical support for training (and for developing more technical experts in all areas)
8. Create better tax codes, standards and regulations (that help US companies grow faster and more easily)
9. Improve cloud capabilities to make them more useful and compatible (perhaps via standards, technical support, and developing tools to make them easier to use)

A Major Concern About HPC/AI

The shortage of talent and the need for diversity

- **The shortage of technical talent is growing**
  - More are retiring then entering the labor pool
  - Large companies are hiring away many experts
  - There continues to be a lack of diversity, limiting the talent pool
- **The need for technical talent is increasing**
  - AI with machine and deep learning is complex and is in its infancy
  - Systems are becoming more complex with multiple processor types, larger scales, different types of hardware and software, …
  - Adding cloud resources is complex
  - Quantum computing has great promise, but is complex

The new technologies present a great opportunity, but require a larger workforce to take advantage of them

Today's Agenda

9:00--9:15	Why HPC is Important to Science and Industry
9:15--9:50	An Overview of the HPC, AI and QC Markets
9:50--10:15	A Global Perspective on HPC Focusing on China
10:15--10:45	Part 1: U.S. Vendors Discuss the Impact of Act: HPE, AMD & IBM
10:45--11:00	Break
11:00--11:30	The role of the Department of Commerce/NIST in Implementing the CHIPS and Science Act
11:30--12:00	How NSF is Implementing the CHIPS Act & Why HPC Leading-edge HPC Tools are Important to Support & Advance Science
12:00--12:30	Ideas for Advancing Science via New Approaches to HPC/AI
12:30--1:15	Lunch
1:15--1:45	Compelling Science Examples from NASA
1:45--2:15	An Overview of the CHIPS and Science Act
2:15--3:15	Part 2: U.S. Vendors Discuss the Impact Act: Intel, Qualcomm, Microsoft, Goodyear, CATALOG, …
3:15--3:30	Break
3:30--4:00	Sustaining U.S. High-Performance Computing Leadership in an Exascale Era
4:00--4:25	Group Discussion: What Should We Track & How to Better Measure U.S. Competitiveness
4:25--4:30	Next Steps and Action Items

HYPERION RESEARCH

Thank You
For Joining Us Today!

Questions?

We welcome questions, comments and suggestions

Please contact us at:

info@hyperionres.com