Hyperion Research ISC22 Market Update

May 2022

Earl Joseph, Bob Sorensen, Alex Norton,
Mark Nossokoff, Melissa Riddle,
Tom Sorensen and Jaclyn Ludema

www.HyperionResearch.com
www.hpcuserforum.com

The Hyperion Research Team

Analysts

- Earl Joseph, CEO
- Steve Conway, Senior Advisor
- Bob Sorensen, SVP Research
- Mark Nossokoff, Research Director
- Alex Norton, Research Director
- Jaclyn Ludema, Analyst
- Melissa Riddle, Associate Analyst
- Thomas Sorensen, Associate Analyst

Operations

- Jean Sorensen, COO

International Consultants

- Katsuya Nishi, Japan and Asia
- Jie Wu, China & Technology Trends
- Michael Feldman, Europe & Strategic Trends

Global Accounts

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

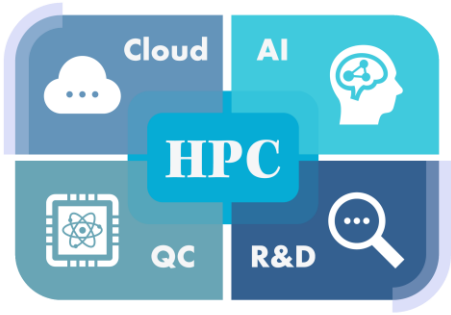
Data Collection

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

Example Research Areas

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

- Traditional HPC
- AI, ML, DL, Graph
- Cloud Computing
- Storage & Data
- Interconnects
- Software & Applications
- Power & Cooling
- Tracking all Processor Types & Growth rates
- Quantum Computing
- R&D and Engineering -- all types
- Edge Computing
- Supply Chain Issues



How We Do Our Research

100% focused on HPC & the results from applying HPC
(HPC includes AI, advanced technologies, and all types of R&D)

- **Worldwide coverage**
  - With a strong focus on international activities
- **Data is collected via direct interviews and surveys**
  - Many sites are surveyed directly by a team member
  - Research includes all types of R&D activities
  - Hyperion Research has deep relationships with the HPC suppliers and most of the larger HPC sites around the world
- **We offer both subscription services and custom research studies**
  - We publish over 90 documents a year

HYPERION RESEARCH

HPC Market Update

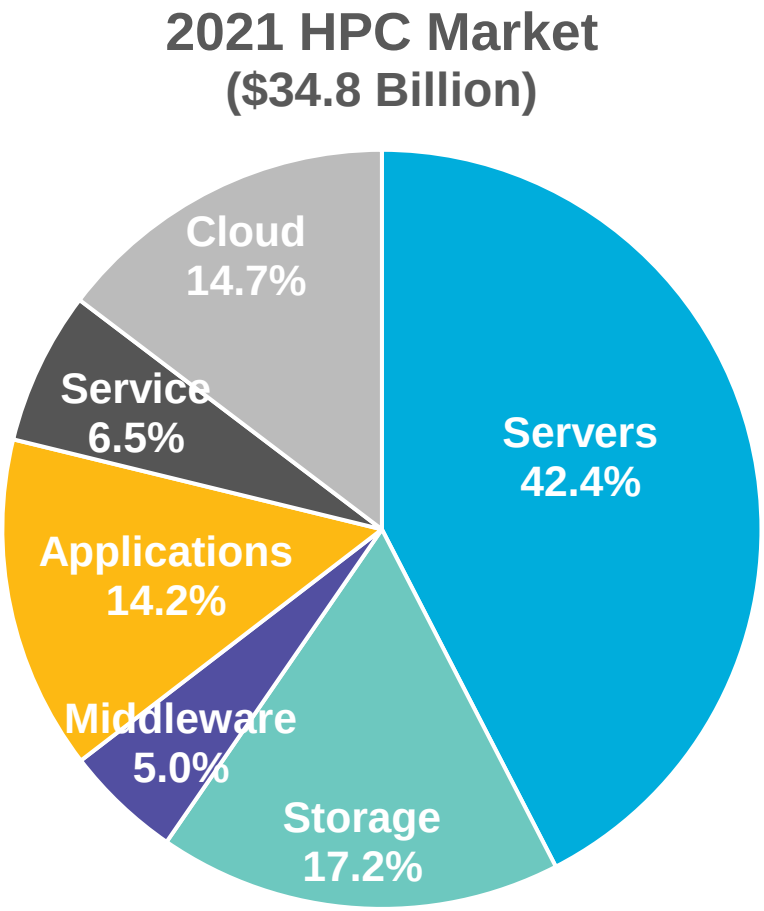
Our ISC22 Briefing Agenda

All registered attendees will receive the slide deck

- **Market Update and Forecasts**
- **Some Perspectives on European HPC**
- **Sustainability: Not Just a "Nice to Have"**
- **HPC and AI Talent Challenges**
- **Exascale Update**
- **Cloud Update**
- **Quantum Update**
- **AI Update**
- **Update on Storage & Interconnects**
- **Conclusions and Wrap-up**

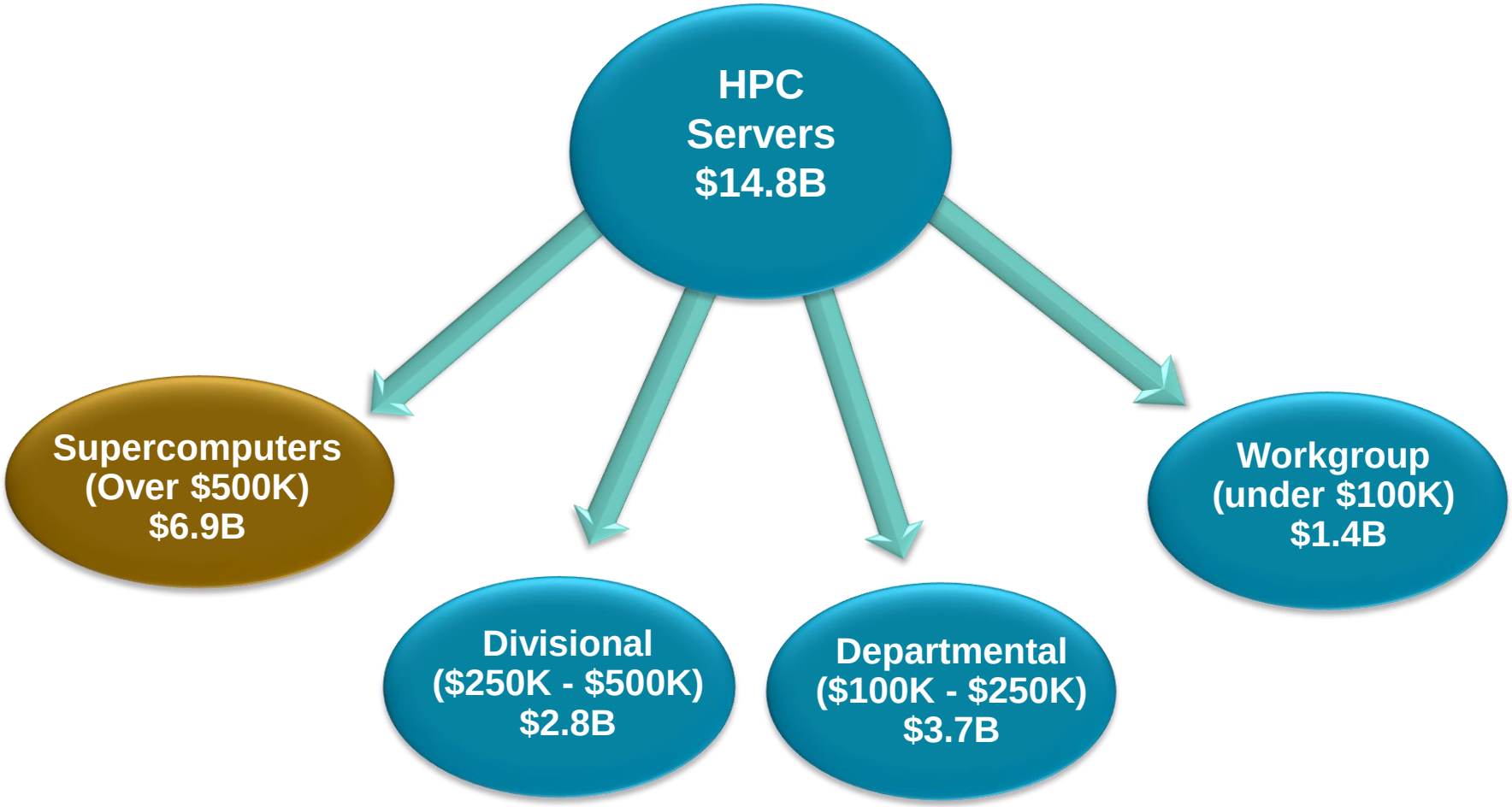
The Overall HPC Market in 2021

Looking at the overall HPC market, including servers, cloud usage, storage, software and repair services = $34.8 billion USD



The 2021 Worldwide On-Prem HPC Server Market: $14.8 Billion (up 9.1%)

2022 is projected to be around $17 Billion



WW HPC Market By Vertical ($ Millions)

2021 WW On-Prem High-Performance Systems Revenue by Applications ($M)	
	2021
Bio-Sciences	1,455
CAE	1,767
Chemical Engineering	177
DCC & Distribution	807
Economics/Financial	703
EDA / IT / ISV	849
Geosciences	1,010
Mechanical Design	59
Defense	1,552
Government Lab	2,866
University/Academic	2,637
Weather	681
Other	199
Total Revenue	14,763
Source: Hyperion Research, 2022	

WW HPC On-Prem Market By Vendor

($ Millions)

2021 On-Prem server Revenues By Vendor ($M)		
Vendor	**Server Revenues ($M)**	**Market Share**
HPE	5,050	34.2%
Dell Technologies	3,213	21.8%
Lenovo	1,174	8.0%
Inspur	993	6.7%
Atos	542	3.7%
Sugon	525	3.6%
IBM	463	3.1%
Penguin	378	2.6%
Fujitsu	176	1.2%
NEC	173	1.2%
Other	2,076	14.1%
Total On-Prem HPC	14,763	100.0%
Source: Hyperion Research, 2022		

The Broader On-premise Market Areas ($ Millions)

2021 total on-prem HPC spending reached $29.7B (adding cloud spending, brings it to $34.8B)

Revenues by the Broader HPC Market Areas			
	2019	2020	2021
Server	13,368	13,523	14,763
Storage	5,288	5,079	5,984
Middleware	1,572	1,491	1,731
Applications	4,569	4,315	4,952
Service	2,181	2,015	2,267
Total Revenue	26,979	26,423	29,697
Source: Hyperion Research, 2022			

HYPERION RESEARCH

HPC Market Update For EMEA

HPC Market By Regions

In 2020: very high growth in Japan, the rest of the market declined by over 7%
In 2021: healthy growth in all other regions

2021 HPC Server Sales By Region			
	2019	2020	2021
North America	6,119	5,424	6,235
EMEA	3,871	3,686	4,142
Asia/Pacific w/o Japan	2,416	2,492	3,230
Japan	746	1,699	898
Rest-of-World	217	222	257
Total	13,368	13,523	14,763
Source: Hyperion Research, 2022			

HPC Market In Europe By Verticals

In 2021: very healthy SERVER growth at 14.3%

HPC Industry/Application Segments for EMEA			
$US Millions	**2020**	**2021**	**Growth**
Bio-Sciences	365	422	15.8%
CAE	491	541	10.3%
Chemical Engineering	13	15	11.8%
DCC & Distribution	13	15	19.3%
Economics/Financial	82	94	14.0%
EDA	114	124	8.7%
Geosciences	350	385	10.1%
Mechanical Design	3	4	29.5%
Defense	330	357	8.1%
Government Lab	981	1,160	18.2%
University/Academic	634	727	14.6%
Weather	141	162	14.9%
Other	59	80	35.9%
EMEA Total	**3,577**	**4,087**	**14.3%**
Source: Hyperion Research, April 2022			

HPC Broader Market In Europe

In 2021: very healthy overall growth at 13.9%

EMEA Revenues by the Broader HPC Market Areas			
$US Millions	**2020**	**2021**	**Growth**
Server	3,577	4,087	14.3%
Storage	1,355	1,561	15.2%
Middleware	405	466	15.0%
Applications	1,156	1,298	12.3%
Service	528	581	10.2%
Total Revenue	**7,020**	**7,993**	**13.9%**
Source: Hyperion Research, April 2022			

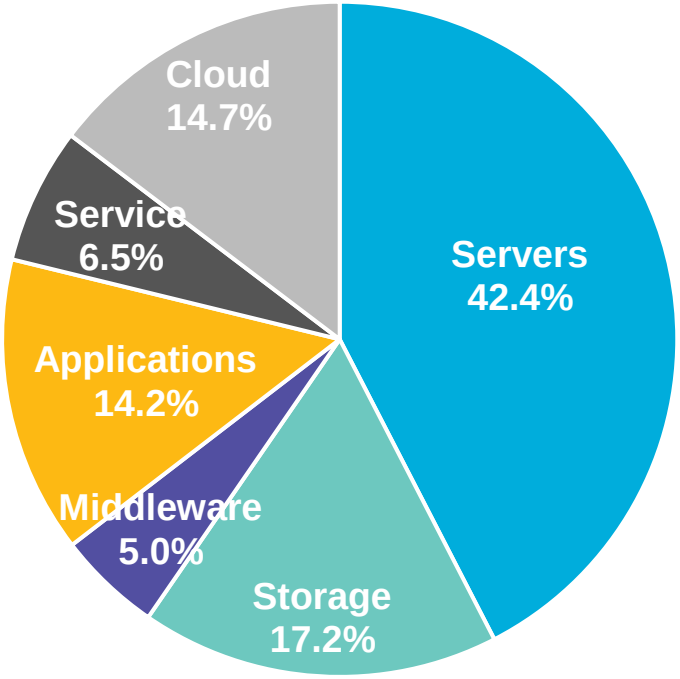
HYPERION RESEARCH

HPC Market Forecasts

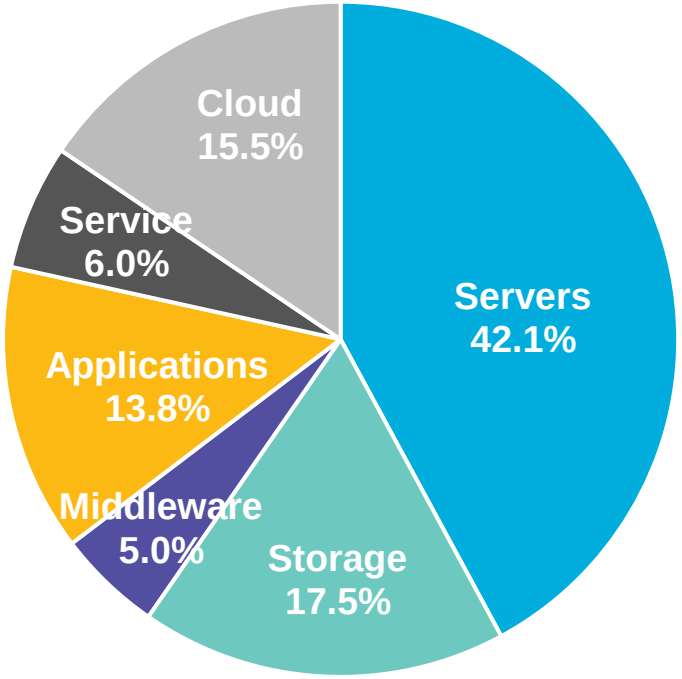
The HPC Market Will Grow in 2022

The strong market rebound in 2021 will carry into 2022 as more buyers look to HPC solutions to address new opportunities and compete more aggressively

2021 HPC Market
($34.8 Billion)



2022 HPC Market
($39.9 Billion)



The HPC Market Will Grow in 2022

<u>On-prem servers:</u> The strong market rebound in 2021 will carry into 2022 as more buyers look to HPC solutions to address new opportunities and compete more aggressively

- **Covid and the resulting supply chain issues have been a major concern for almost 2 years, but HPC vendors have found ways to keep producing products at record levels**
- **2022 is forecasted to reach an all-time high of around US $17 billion in on-prem HPC servers with US $35 billion in total on-premises HPC spending**
  - 2021 on-premises servers grew 9.1% over 2020
- **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**
- **New technological developments in big data, AI, processors, etc. are providing many new areas for users to advance their research and engineering**

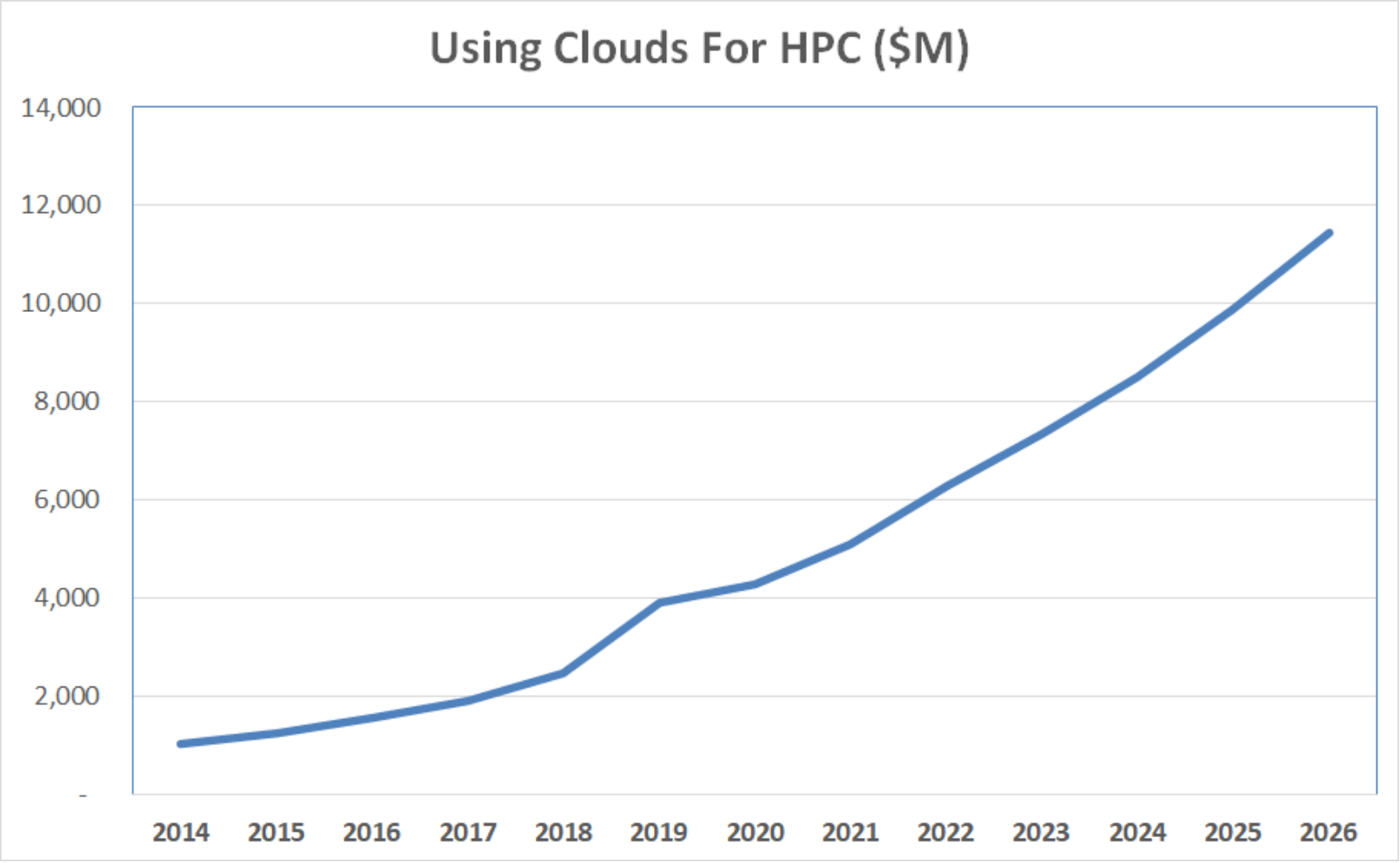
The HPC Cloud Market Will See Very Strong Growth in 2022

The growth will build on the fundamental changes in buying behavior seen in 2021

- **HPC & AI buyers around the world revealed for the first time that HPC buyers are planning to shift some of their on-premises budgets to spending in the cloud**
  - The shift is fundamental because up to 2021 very few sites were taking money from the on-premises budgets for cloud computing
- **End user spending on public cloud resources to run HPC workloads is projected to grow substantially in 2022, at a rate greater than 23%, to exceed US $6 billion**
  - This strong growth reflects the heavy work that the cloud service providers (CSPs) have done to make clouds more HPC friendly
  - Users have also gone through extensive work to profile and evaluate where clouds make the most sense
- **This major shift in buying behavior doesn't mean that on-premises HPC systems are going away**
  - The on-premises HPC server market is anticipated to exhibit healthy growth, 7%-8% a year, over the forecast period

HPC Cloud Usage Forecast

17.6% growth over the next 5 years



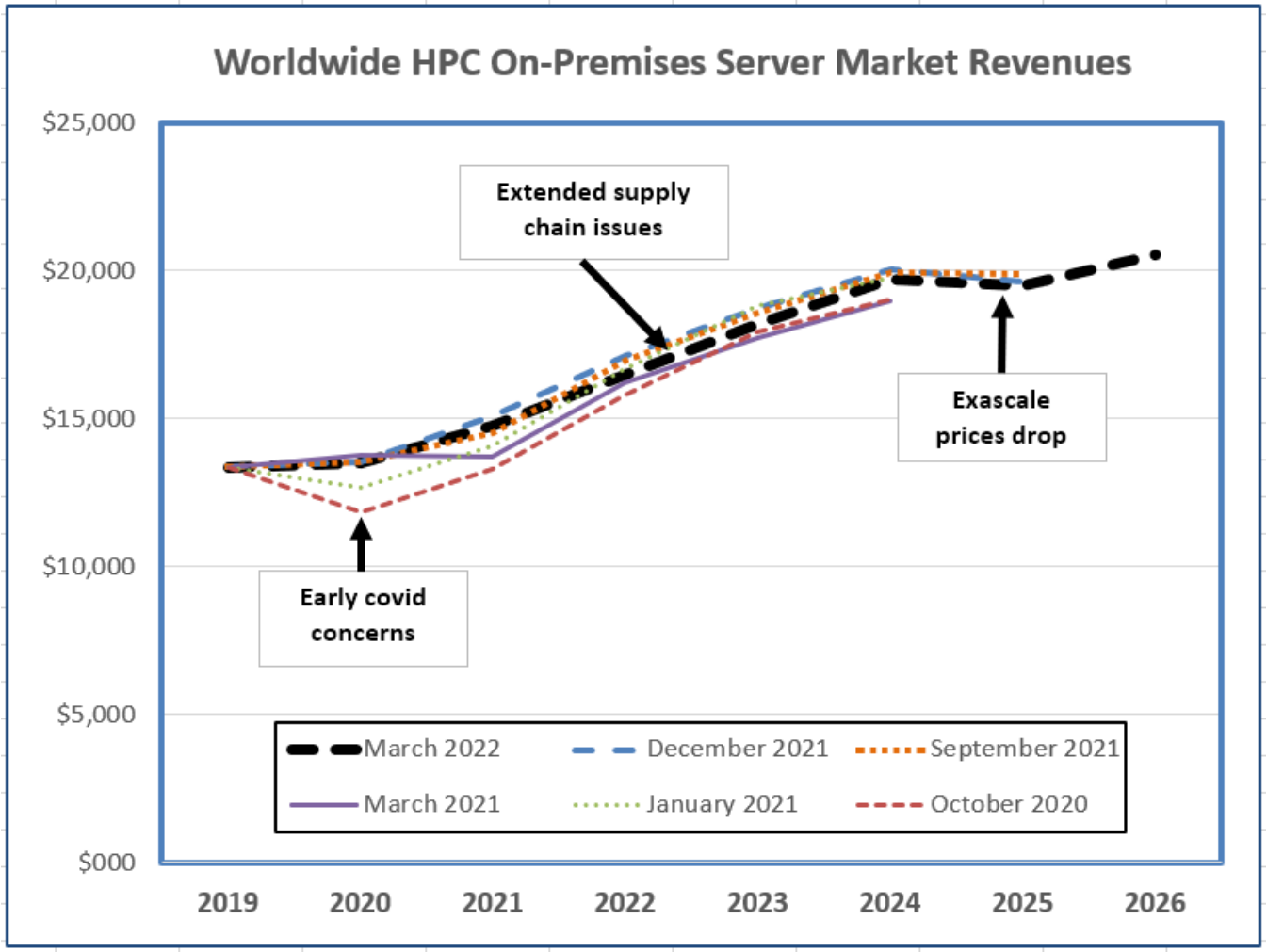
5-Year On-Prem HPC Server Forecast

6.9% growth over the next 5 years

On-Prem HPC Market Forecast by Competitive Segment							
$US Millions	**2021**	**2022**	**2023**	**2024**	**2025**	**2026**	**CAGR 21-26**
Supercomputer	6,926	8,125	9,023	9,846	9,661	9,543	6.6%
Divisional	2,803	3,077	3,397	3,681	3,677	4,397	9.4%
Departmental	3,648	3,909	4,316	4,645	4,704	5,137	7.1%
Workgroup	1,373	1,392	1,472	1,525	1,450	1,471	1.4%
Total	**14,763**	**16,503**	**18,208**	**19,697**	**19,492**	**20,549**	**6.9%**
Source: Hyperion Research, April 2022							

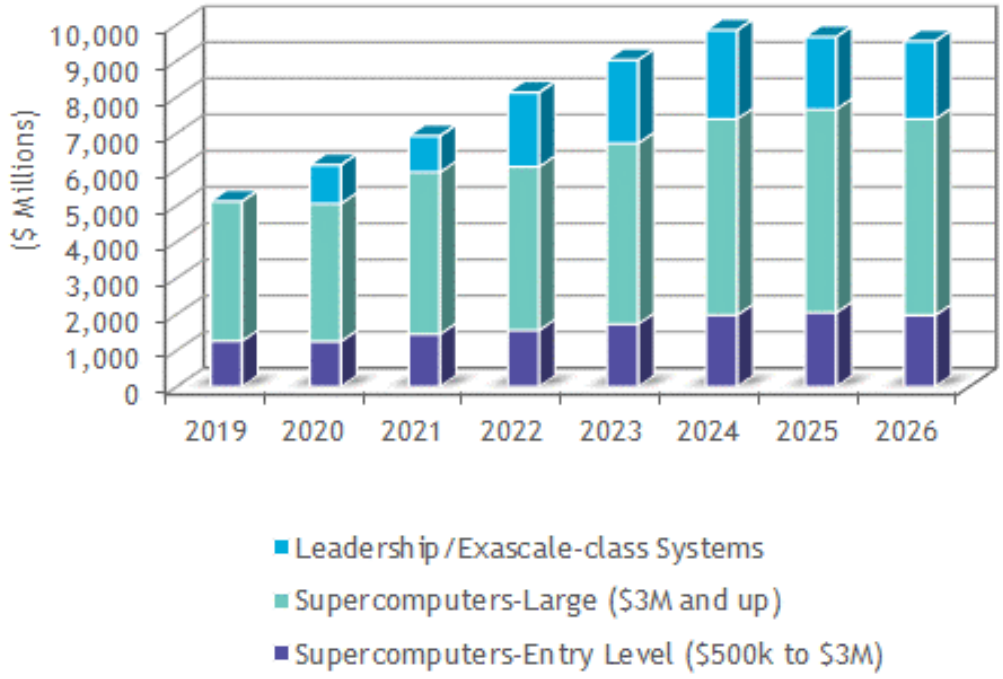
5-year HPC Server Forecast Changes

Quarterly adjustments due to covid



New Supercomputer Subsegments

New Supercomputer Subsegments									
$ Millions									
	2019	2020	2021	2022	2023	2024	2025	2026	CAGR 21-26
Leadership/Exascale-class Systems	0	1,065	1,000	2,050	2,300	2,450	2,000	2,150	16.5%
Supercomputers-Large ($3M and up)	3,858	3,808	4,477	4,516	4,995	5,420	5,615	5,419	3.9%
Supercomputers-Entry Level ($500k to $3M)	1,260	1,250	1,450	1,559	1,728	1,975	2,046	1,975	6.4%
Total Supercomputers ($500K and up)	**5,118**	**6,123**	**6,926**	**8,125**	**9,023**	**9,846**	**9,661**	**9,543**	**6.6%**
Source: Hyperion Research, May 2022									



The Exascale Market (System Acceptances)

Over 30 systems and over $11 billion in value

Exascale and Near-Exascale Systems (2020 to 2026)

Year Accepted	China	Europe	Japan	US	Other Countries*	Total Systems	Total Value
2020			1 near-exascale system ~$1 B			1	$1B
2021	1 or 2 near-exascale systems ~$350M to $400M each	1 pre-exascale system ~$185M	?	1 pre-exascale system ~$200M	--	3-4	$.8B - $1.1B
2022	1 or 2 exascale systems ~$350M - $400M each	2 pre-exascale systems ~$400 total	1 near-exascale system ~$150M	1 exascale systems ~$600M	--	5-6	$2B - $2.3B
2023	1 or 2 exascale system ~$350M - $400M each	1 or 2 exascale systems ~$350M each	1 near-exascale system ~$150M	1 or 2 exascale systems ~$600M each	--	5-7	$2B - $2.8B
2024	1 exascale system ~$350M - $400M	2 exascale ~$350M each	?	1 or 2 exascale systems ~$500M each	1 exascale system ~$200M	5-6	$1.9B - $2.5B
2025	1 exascale systems ~$350M - $400M	1 or 2 exascale systems ~$350M each	1 exascale system ~$200M	1 or 2 exascale systems ~$400M each	1 exascale system ~$200M	5-7	$1.5B - $2.3B
2026	1 or 2 exascale systems ~$350M - $400M each	1 or 2 exascale systems ~$350M each	?	1 or 2 exascale systems ~$400M each	1 or 2 exascale systems ~$200M	5-8	$1.9B - $2.8B
Total	**6-10**	**8-10**	**3+**	**6-9**	**3-4**	**28-39**	**$11B - $15B**

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

Source: Hyperion Research, May 2022

Our ISC22 Briefing Agenda

All registered attendees will receive the slide deck

- **Market Update and Forecasts**
- **Some Perspectives on European HPC**
- **Sustainability: Not Just a "Nice to Have"**
- **HPC and AI Talent Challenges**
- **Exascale Update**
- **Cloud Update**
- **Quantum Update**
- **AI Update**
- **Update on Storage & Interconnects**
- **Conclusions and Wrap-up**

Some Perspectives on HPC in Europe

ISC22
June 2022

www.HyperionResearch.com
www.hpcuserforum.com

Mark Nossokoff

European HPC Perspectives

Total EU HPC revenue growing ahead of market at 7.9%

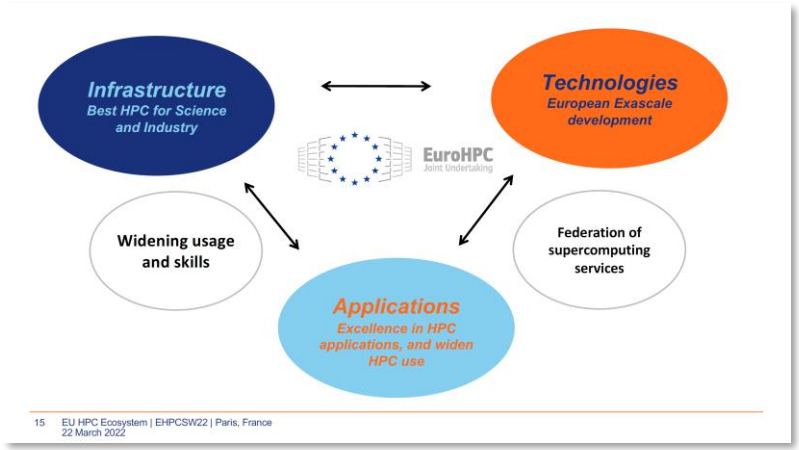
HPC Server (€M)	2019	2020	2021
North America	5,324	4,719	5,424
EMEA	3,368	3,207	3,604
APAC w/o Japan	2,102	2,168	2,810
Japan	649	1,478	781
ROW	189	193	224
Total	**11,631**	**11,765**	**12,843**

- **HPC Server 2021 demand/consumption**
  - EMEA ~ 28.1% of the global server market
  - EU ~ 80% of European consumption
- **EU growth in the worldwide HPC market projected to grow at a CAGR of 7.9% between 2021-2026, ahead of the overall market CAGR of 6.9%**

European HPC Perspectives

Broad-based progress with EuroHPC JU leadership

- **Public/private funding up to €7B through 2027**
- **Strategic Pillar strategy**
  - Technologies
  - Infrastructure
  - Applications
  - Federation, Usage, Skillsets



- **Top HPC installations:**
  - Five petascale
  - Three pre-exascale
- **Multiple calls open to support 4-pillar strategy**
- **Develop new quantum computing ecosystem**
- **EPI and SiPearl progress**
- **SAGE 2 complete and transitioning to IO-Sea**
- **Pan-European EUMasters4HPC program**

European HPC Perspectives

Other activity bodes well for HPC in Europe

- **UK HPC**
  - Archer 2
    - #22 on Nov 2021 Top500
    - UKRI, EPCC, the University of Edinburgh
  - Cambridge-1
    - Business and academic research focused on health and life sciences
    - NVIDIA datacenter
- **Intel investment in fab capacity**
  - Ireland
  - Germany

HYPERION RESEARCH

PLEASE SHARE YOU THOUGHTS!

mnossokoff@hyperionres.com

Our ISC22 Briefing Agenda

All registered attendees will receive the slide deck

- **Market Update and Forecasts**
- **Some Perspectives on European HPC**
- **Sustainability: No Longer a "Nice to Have"**
- **HPC and AI Talent Challenges**
- **Exascale Update**
- **Cloud Update**
- **Quantum Update**
- **AI Update**
- **Update on Storage & Interconnects**
- **Conclusions and Wrap-up**

Sustainability: No Longer a "Nice-to-Have"

ISC

June 2022

Jaclyn Ludema and Alex Norton

www.HyperionResearch.com
www.hpcuserforum.com

Upcoming and Future Systems

Power considerations of future systems are becoming more important to user sites

- **As the high end of the HPC space continues to plan and deploy larger systems, the power envelopes of those systems increases significantly**
  - GPUs draw more power than CPUs
  - CPUs and GPUs are increasing in power requirements
  - More racks are being installed
- **Frontier, as an example, anticipated to be a 20MW system, while Aurora has been discussed approaching 60MW**
- **Countries with tighter restrictions on power are looking at different approaches to leadership-class HPC systems**
  - LUMI – feeding back power to the local city
  - More power-efficient processors

Rising Power Envelopes of Chips

Chips getting hotter and drawing more power

- **New NVIDIA Hopper architecture GPUs have increased power envelopes north of 500 watts, the highest end topping out at 750 watts**
- **CPUs from Intel and AMD both top 200 watts at the high end of their respective CPU lines**
- **What does this boil down to?**
  - A rising need for liquid cooling in dense systems
  - A growth of power-efficient processor alternatives
    - Arm, RISC-V, ASICs, etc.
    - The European Processor Initiative highlights potential power efficiency as key pillar of development
  - Geographic-centric considerations for future systems
    - Cost of power
    - Access to power
    - Power consumption impacting local communities

The Carbon Offset of AI

Is the juice worth the squeeze?

- **Despite the growing power envelopes of major HPC systems, they continue to be built**
  - Not only is it about the power in, but also the carbon out
  - How can renewable energy sources help this?
- **The focus becomes an analysis of the trade off between running the system (and the power associated with that) and the impact of the work**
  - How can AI models be trained more efficiently to reduce their carbon impact on the globe?
  - Is the Green500 (or benchmarks like it) going to become a new critical component to system ranking and design?
- **There is a growing difficulty of monitoring the issue of carbon footprint with more systems moving fully or partially to cloud**

Addressing Carbon Outputs

US DOE announces MPC4Mfg to help optimize energy output of manufacturing organizations

- **Earlier this month, the US Department of Energy released a solicitation, HPC4Mfg for projects in the manufacturing sector:**
  - The goal: to use HPC to reduce the carbon footprint and improve efficiency of operations
  - The DOE allocated $3 million USD towards these projects
  - Funds awarded will support compute cycles on national lab systems
- **"The advanced modeling, simulation, and data analysis capabilities of these supercomputers will help manufacturers identify the most effective ways to improve performance, increase energy efficiency, and cut emissions across the industrial sector"[1]**

For more information about the HPC4Mfg: https://hpc4energyinnovation.llnl.gov/

Who Can Improve Sustainability?

Efficient business practices = less power waste

- **Business Analysts**
  - What is the optimized frequency of running applications to provide business value and save cost/power?
- **HPC Architects**
  - Resource optimization
  - Containerization
- **Site Planners**
  - Thorough evaluation of the impacts to the power supply infrastructure and local communities
  - Proper assessment of impact to adjacent ecosystems of the excess heat and noise produced by HPCs
- **Executives**
  - Plan to recycle components of HPCs at the end of life
  - Sensible identification of workloads that require HPCs

HYPERION RESEARCH

Continue the conversation

jludema@hyperionres.com

anorton@hyperionres.com

Our ISC22 Briefing Agenda

All registered attendees will receive the slide deck

- **Market Update and Forecasts**
- **Some Perspectives on European HPC**
- **Sustainability: No Longer a "Nice to Have"**
- **HPC and AI Talent Challenges**
- **Exascale Update**
- **Cloud Update**
- **Quantum Update**
- **AI Update**
- **Update on Storage & Interconnects**
- **Conclusions and Wrap-up**

HPC and AI Talent Challenges

ISC22
June 2022

Melissa Riddle

www.HyperionResearch.com
www.hpcuserforum.com

HPC/AI Expertise Shortage

Growing scarcity of HPC experts to implement new technologies will be a roadblock for many sites

- **HPC workforce is shrinking as outflow of retirees exceeds pipeline of new HPC staff**
- **New enterprise IT entrants are increasing the number of HPC sites worldwide**
- **Existing HPC sites are becoming more complex: emerging technologies such as AI, cloud, and GPUs require different skillsets and are driving up the number of systems per site**
- **Competition for HPC staff will intensify**
- **HPC users need major improvements in ease-of-use, ease-of-selection, & ease-of-optimization**

On-Premises HPC Talent

68% of respondents reported on-prem talent concerns

Q: What are your top 3 most significant barriers to expanding use of HPC/technical computing on-premises?

		SECTOR			REGION		
	Overall	Industry	Government	Academia	EMEA	NA	APAC
Lack of knowledge, or skilled HPC/ technical computing support staff	31.9%	40.2%	21.7%	19.4%	31.4%	32.9%	31.3%
Difficulties in scaling/moving our work up to an HPC technical server	27.7%	30.5%	21.7%	25.0%	17.6%	30.1%	50.0%
Programming hurdles with hybrid environments	24.1%	18.3%	60.9%	13.9%	23.5%	24.7%	25.0%
Ease-of-use issues	16.3%	20.7%	8.7%	11.1%	13.7%	20.5%	6.3%

n = 141; 82; 23; 36; 51; 73; 16 (respectively) with 3 responses per respondent. Additional choices not shown.

Source: Hyperion Research, 2021

- **Industry sector is most concerned about overall lack of skilled staff while Government is most concerned about programming for hybrid environments**
- **Compared to WW trend, EMEA is less concerned about scaling HPC workloads while APAC is more concerned**

Cloud HPC Talent

55% of respondents reported cloud talent concerns

Q: What are your top 3 most significant barriers to expanding use of HPC/technical computing in the cloud?

		SECTOR			REGION		
	Overall	Industry	Government	Academia	EMEA	NA	APAC
Lack of knowledge, or skilled cloud computing support staff	28.4%	32.9%	8.7%	30.6%	25.5%	28.8%	37.5%
Difficulties in scaling/moving our work up to an HPC technical server	22.0%	24.4%	34.8%	8.3%	29.4%	16.4%	25.0%
Ease-of-use issues	17.0%	22.0%	4.3%	13.9%	9.8%	23.3%	12.5%

n = 141; 82; 23; 36; 51; 73; 16 (respectively) with 3 responses per respondent. Additional choices not shown.

Source: Hyperion Research, 2021

- **Industry sector is most concerned about overall lack of skilled HPC cloud staff while Government sector is most concerned about scaling cloud workloads**
- **Compared with the WW trend, EMEA is more concerned about scaling HPC cloud workloads and less concerned about ease-of-use**

AI HPC Talent

72% of respondents reported AI talent concerns

Q: Which of the following is a barrier to furthering your AI capabilities?

		SECTOR			REGION		
	Overall	Industry	Government	Academia	EMEA	NA	APAC
Access to AI expertise	51.1%	52.4%	56.5%	44.4%	39.2%	60.3%	50.0%
Skills in AI programming	46.8%	51.2%	26.1%	50.0%	41.2%	50.7%	50.0%
Skills in AI model development	44.7%	48.8%	34.8%	41.7%	35.3%	52.1%	43.8%

n = 141; 82; 23; 36; 51; 73; 16 (respectively) with 2.8 responses per respondent. Additional choices not shown.

Source: Hyperion Research, 2021

- **Half of all respondents reported concern about general AI expertise access**
- **Among these regions, North America is the most concerned about each aspect of AI talent while EMEA is the least concerned**
- **Overall, respondents were most concerned about talent in AI, followed by on-prem HPC and cloud**

HYPERION RESEARCH

Questions?

mriddle@hyperionres.com

Our ISC22 Briefing Agenda

All registered attendees will receive the slide deck

- **Market Update and Forecasts**
- **Some Perspectives on European HPC**
- **Sustainability: No Longer a "Nice to Have"**
- **HPC and AI Talent Challenges**
- **Exascale Update**
- **Cloud Update**
- **Quantum Update**
- **AI Update**
- **Update on Storage & Interconnects**
- **Conclusions and Wrap-up**

A Quick Update on Exascale Systems

ISC 22

Bob Sorensen

www.HyperionResearch.com
www.hpcuserforum.com

Near-Term US Exascale Plans

Three systems over two years with budget of ~ $1.8 billion

- **Frontier: DOE Office of Science: Oak Ridge National Laboratory**
  - Slated to be first US exascale system in US
  - ~30MW, >1.5EF, 100 cabinets
  - *Crusher*, partial install (1.5/100 cabinets) currently running code
  - Cray Shasta with AMD EPYC CPU and AMD Radeon Instinct GPUs
  - Full user operations January 2023 (some delay)
- **Aurora: DOE Office of Science, Argonne National Laboratory**
  - **60MW**, ~ 1EF DP sustained
  - 08/21: Polaris testbed system (44PF DP and 1.5EF AI)
  - Cray Shasta architecture with Intel Xeons and Intel Xe GPU
  - Delivery in late 2022, acceptance in 2023 (delayed at least 12 months)
- **El Capitan: DOE NNSA's LLNL**
  - ~ 2EF
  - Cray Shasta architecture AMD EPYC processors, next generation Radeon Instinct GPUs
  - Fully deployed in 2023

China Exascale Status

The official situation

- **Sunway Pro OceanLight**
  - ~1.3 EFlops Rpeak, ~1.05 EFlops Rmax
  - 35 MW, 38 million cores
  - ShenWei post-Alpha CPU
  - National Supercomputing Center-Wuxi
- **Tianhe-3**
  - Dual-chip FeiTeng ARM and matrix accelerator nodes
  - ~ 1.7 EFlops Rpeak, 1.3 EFlops Rmax
  - NSCC-Tianjin
- **Sugon**
  - Hygon processors (low confidence), may go AMD Zen4
  - NSCC-Shenzhen

China Exascale Status

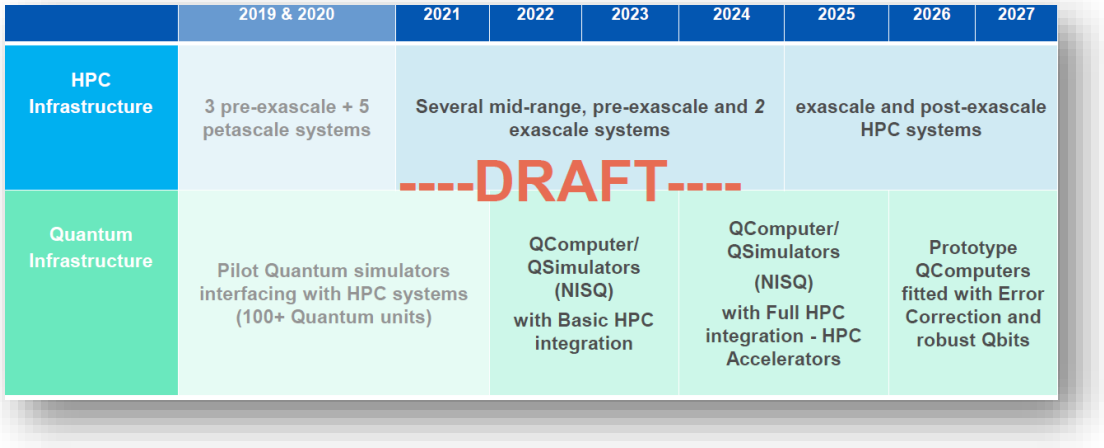
The unofficial reality?

- **Sunway Pro OceanLight**
  - Up and running since March 2021
- **Tianhe-3**
  - Up and running in last six (?) months
- **Sugon**
  - Potentially delayed
- **No official announcements**
- **No entries for June 2021, November 2021 Top 500 list**
  - Maybe this time around…will find out soon
  - Some hesitancy by Chinese leadership
- **Strong evidence of at least five other Chinese systems that could make top 10 list today**

EU HPC Plans

Exascale plans going forward

- **EU plan calls for acquisition of two exascale systems in the 2021-2024 timeframe**
  - At least one to use European technology: specifically using an EPI-developed processor
  - Additional procurements in Germany in 2024, 2025
  - EU may include 2 additional ES systems in 2023-2026
- **Post Exascale System around 2027**
  - Plans call for integration and deployment of the first hybrid HPC/quantum infrastructure in Europe



Source: Leonardo Flores Añover, Senior Expert

DG CNECT, HPC & Quantum Technology Unit - European Commission 2021

UK EU Plans

- **UK plans to roll out its first exascale system in 2025.**
- **System will be hosted at Advanced Computing Facility of EPCC, formerly the Edinburgh Parallel Computing Centre, a supercomputing centre based at the University of Edinburgh**
- **Build on current work within Exascale Computing ALgorithms and Infrastructures Benefiting UK Research (ExCALIBUR)**
  - Aims to deliver the next generation of high-performance simulation software for the highest priority fields in UK research
  - Stand-alone UK programme with a £46 million investment exclusively focused on exascale software and algorithms
- **Complements on-going UK HPC efforts**
  - Recently stood up ARCHER2, a £79 million effort, (28 Pflop/s HPE Cray EX HPC) resource for UK academic and commercial scientific research
  - On-going DiRAC 3, a distributed HPC ecosystem spanning extreme scaling, data intensive applications, and memory intensive applications across four UK sites

HYPERION RESEARCH

QUESTIONS?

bsorensen@hyperionres.com

Float to the top or sink to the bottom. Everything in the middle is the churn. -Amos Burton, The Expanse

HPC in the Cloud Update

ISC

June 2022

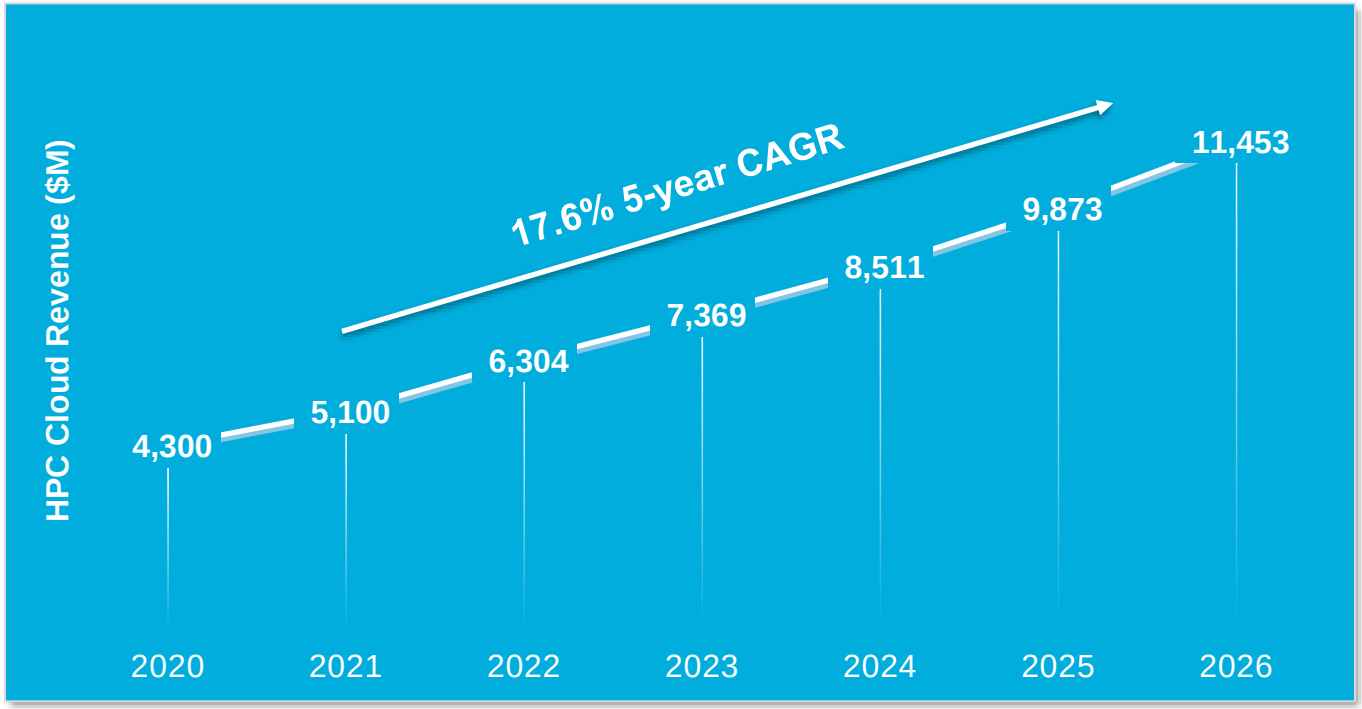
Alex Norton and Mark Nossokoff

www.HyperionResearch.com
www.hpcuserforum.com

HPC Cloud Forecast

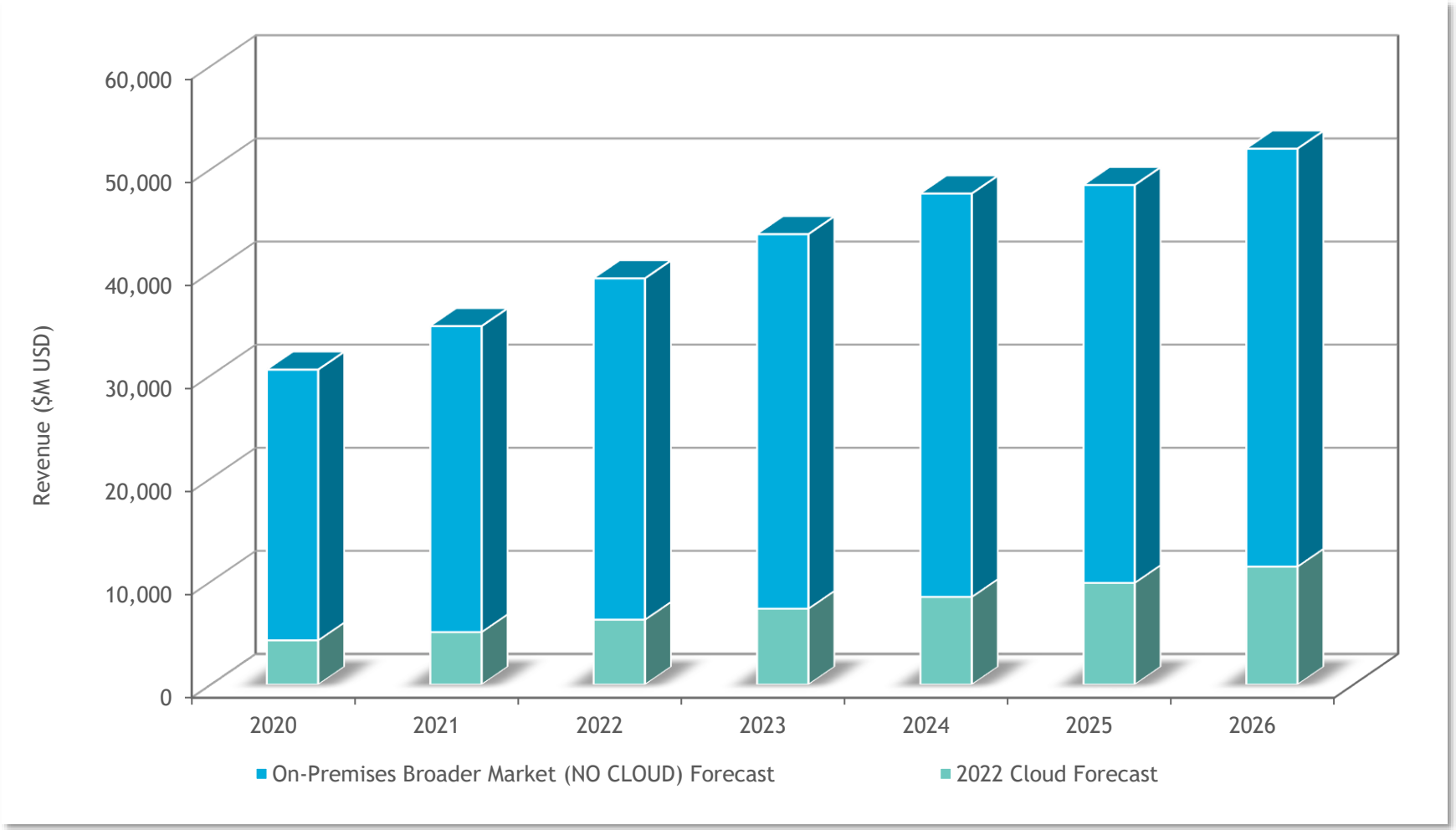
HPC cloud revenue is expected to exceed $11 billion by 2026

- **Storage-specific components comprise roughly 1/3 of cloud revenue for HPC**
- **AI and other data-intensive applications are a high growth segment for cloud adoption in HPC**



The Total HPC Market: On-prem And Cloud Computing

The cloud market is smaller, but growing faster



Impact of HPC Cloud

Organizations are increasingly factoring cloud into future on-premises deployment plans

- **Public cloud resources have historically been seen as complementary to on-premises**
  - Many longitudinal studies show that cloud is used primarily for burst capabilities by many HPC users'
  - This perception is shifting:
    - A recent study showed that almost 50% of the users are altering on-premises deployments due to cloud
- **Migrating HPC workloads to cloud platforms requires new skills for center managers & researchers**
  - Much of the education and training on using the cloud focuses on which workloads can and should be run in the cloud versus remain on-premises
  - IT departments are factoring in data movement and security as they expand their resource pools to consist of cloud resources

No One-Size-Fits-All

Cloud and on-prem decisions depend on variety of factors

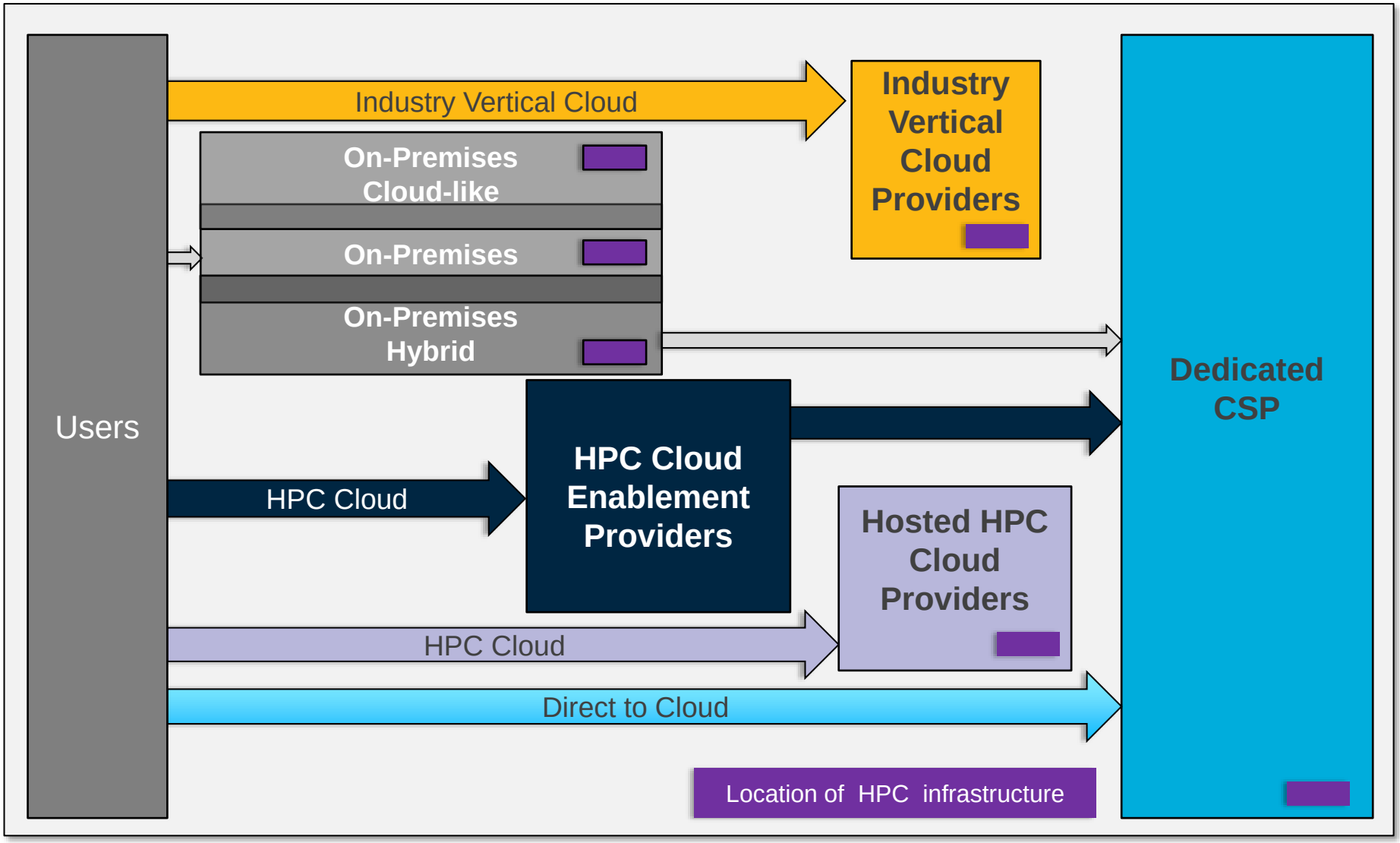
- **Choosing where to run HPC applications depends on a wide set of factors:**
  - Internal skillsets and knowledge
  - Budget and cost
  - Performance characteristics
  - Hardware availability
  - Time constraints
  - Other
- **Understanding which applications can be run cost effectively and performant in the cloud can aid in on-premises system design and optimization**
  - What technologies to invest in?
  - What scale to deploy on-premises?
  - Which middleware and software tools are necessary to optimally run HPC applications across platforms?
- **There is no single solution for all HPC user sites**

HPC Cloud Consumption Approaches

HPCaaS permeates all avenues to access HPC resources

- **On-prem and hybrid clouds**
  - Cloud-like OPEX model for on-prem utilization
  - Incorporates utilization of third party CSPs
- **Industry vertical clouds**
  - Resources housed at third party site
  - Focused on specific vertical use cases
- **Cloud Enablement Providers**
  - Leverages best-of-breed CSPs
  - Black-box provisioning of applications and instances
- **Hosted HPC Clouds**
  - Manages HPC infrastructure
  - Focus exclusively on diverse set of HPC use cases
- **Direct to Cloud**
  - Primary business is to provide cloud resources
  - HPC use case is only one of many supported workloads

Multiple Models for Accessing HPC Cloud Services



Open Research Questions

Facets that we are paying attention to over the next year

- **What is the right balance of on-prem and cloud?**
  - Is it workflow driven?
  - Is it skillset driven?
  - What factors should be considered in finding the balance?
- **How can TCO analyses be improved to create more apples-to-apples comparison of cloud and on-prem?**
- **Where are the concentrations of cloud-born users?**
- **Can the impact of cloud on the on-prem market be quantified in terms of revenue?**

HYPERION RESEARCH

Continue the conversation

anorton@hyperionres.com

mnossokoff@hyperionres.com

Our ISC22 Briefing Agenda

All registered attendees will receive the slide deck

- **Market Update and Forecasts**
- **Some Perspectives on European HPC**
- **Sustainability: No Longer a "Nice to Have"**
- **HPC and AI Talent Challenges**
- **Exascale Update**
- **Cloud Update**
- **Quantum Update**
- **AI Update**
- **Update on Storage & Interconnects**
- **Conclusions and Wrap-up**

A Whirlwind Tour of the Quantum Computing Market

ISC 22

www.HyperionResearch.com
www.hpcuserforum.com

Bob Sorensen
Chief Analyst for Quantum Computing

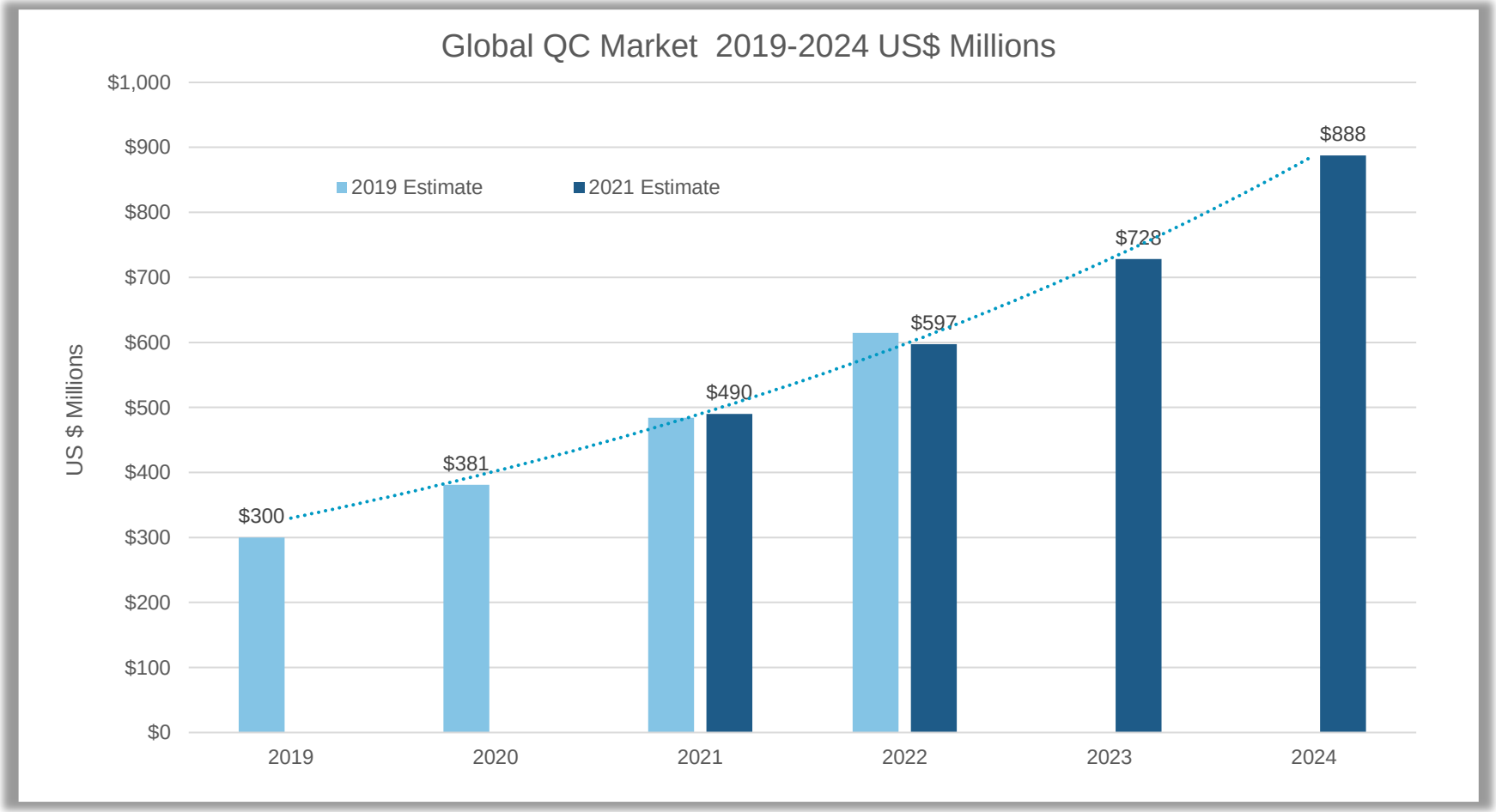
QC Market Study Summary

Continued strong growth seen for global QC sector

- **Based on a study of 112 QC suppliers from around the world, the estimated global QC market was worth about US $490 Million in 2021**
  - The anticipated CAGR for the sector between 2021 and 2024 is 21.9%
  - The global QC market in 2024 is estimated to be approximately US $900 Million
- **Current QC supplier base dominated by a few players**
  - 49% of companies < US $500K, 7% > US $10 Million
- **Cloud access model dominates for next three years**
  - All cloud (43%) + hybrid (21%) = 64%
- **Most Promising Market Segments**
  - QC, Cybersecurity, Financial, Academic, and Chemical/Chemistry

Global QC Market 2019-2024

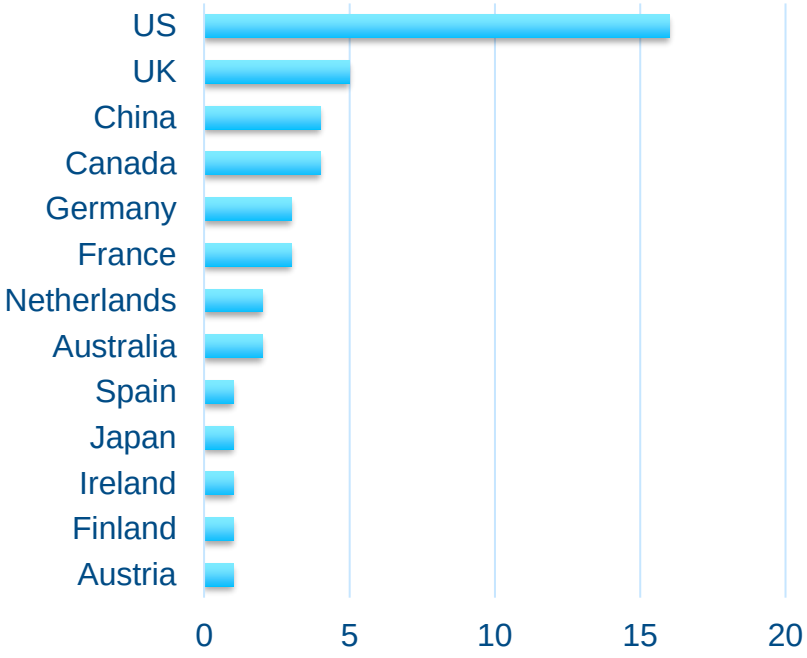
Global QC market reaches ~US $900 Million in 2024

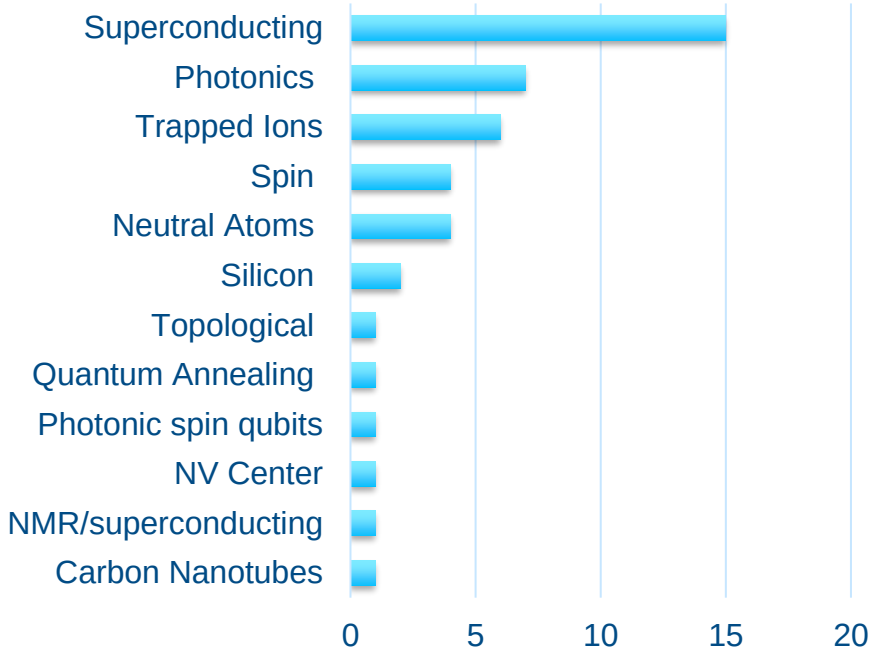


Current Field of Commercial QC Hardware Suppliers

Tracking the visible players from a global perspective

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





QC End User Study Key Takeaways

Commercial end users' interest in QC is high

- **Almost 70% of 415 companies surveyed worldwide have some in-house QC program**
  - Additional 20% plan to do so in the next few years
- **QC technology is seen as offering a range of corporate-level benefits including improved research capabilities and increased revenue**
- **End users are looking for solutions in optimization, factory processes, scheduling, etc.**
- **Every vertical surveyed had a significant number of organizations currently involved in some level of QC activity**
- **Most interested companies already involved in data analysis, ML/DL, optimization, mod/sim, material science—some of the most promising areas of QC today**

US Congress: H. R. 8132 Section 4

Study to Explore Quantum Computing Vulnerabilities

SEC. 4. STUDY TO ADVANCE QUANTUM COMPUTING.

(a) IN GENERAL.—

(1) STUDY REQUIRED.—Not later than 1 year after the date of enactment of this Act, the Secretary of Commerce and the Federal Trade Commission shall complete a study on the state of the quantum computing industry and the impact of such industry on the United States economy.

(2) REQUIREMENTS FOR STUDY.—In conducting the study, the Secretary and the Commission shall—

(A) develop and conduct a survey of the quantum computing industry through outreach to participating entities as appropriate to—

(i) establish a list of industry sectors that implement and promote the use of quantum computing;

(ii) establish a l
developed, or are de

(iii) the status o

(iv) provide a d

(B) develop a comprehen

(C) identify which Federa

(D) identify all interagenc

(E) develop a brief descri

(F) identify all regulations, guidelines, mandatory standards, voluntary standards, and other policies implemented by each of the Federal agencies identified under subparagraph (B), as well as all guidelines, mandatory standards, voluntary standards, and other policies implemented by industry-based bodies; and

(G) identify Federal Government resources that exist for consumers and small businesses to evaluate the use of quantum computing.

(b) MARKETPLACE AND SUPPLY CHAIN SURVEY.—The Secretary and Commission shall conduct a survey of the marketplace and supply chain of quantum computing to—

(1) assess the severity of risks posed to such marketplace and supply chain;

(2) review the ability of foreign governments or third parties to exploit the supply chain in a manner that raises risks to the economic and national security of the United States; and

(3) identify emerging risks and long-term trends in such marketplace and supply chain.

(b) MARKETPLACE AND SUPPLY CHAIN SURVEY.—The Secretary and Commission shall

(1) assess the severity of risks posed to such marketplace and supply chain;

(2) review the ability of foreign governments or third parties to exploit the supply chain i

(3) identify emerging risks and long-term trends in such marketplace and supply chain.

HYPERION RESEARCH

QUESTIONS?

bsorensen@hyperionres.com

Insufficient facts always invite danger.

- Spock, *Stardate: 3141.9.*

Global AI Update

ISC

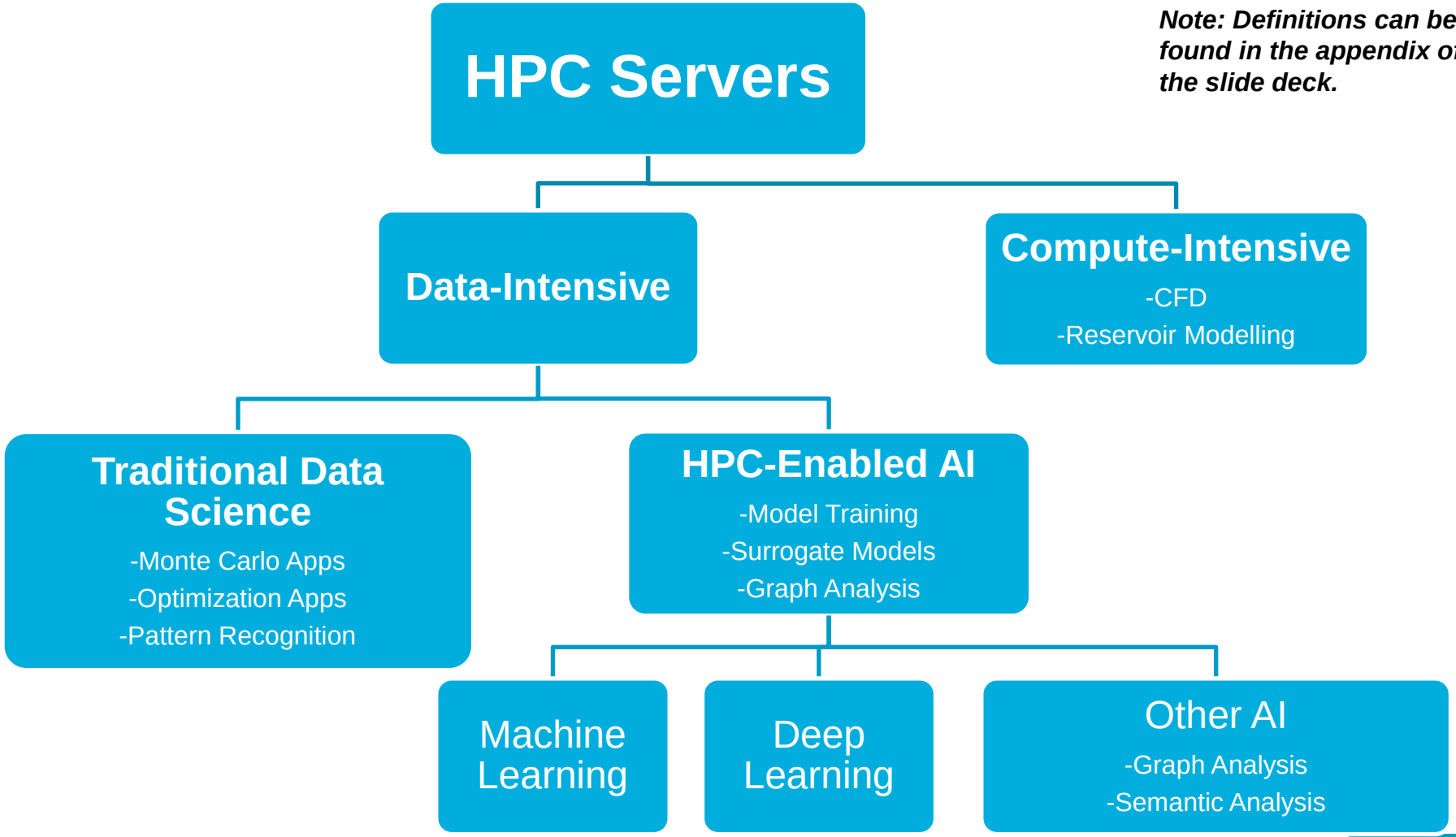
June 2022

Alex Norton and Tom Sorensen

www.HyperionResearch.com
www.hpcuserforum.com

On-Prem HPC Market Segmentation

Server classification based on end-user application



Note: Definitions can be found in the appendix of the slide deck.

HPC-enabled AI Forecast

5 year CAGR expected to reach over 22% growth

Forecast: Worldwide HPC server revenue breakout by compute-intensive and data-intensive focuses ($M)	2020	2021	2022	2023	2024	2025	2026	CAGR 2021-2026
Worldwide HPC Server Revenue Forecast	13,519	14,750	16,503	18,208	19,697	19,492	20,549	6.9%
Compute-Intensive Server Revenue	10,020	10,848	12,103	13,280	14,177	13,586	13,993	5.2%
Data-Intensive Server Revenue	3,499	3,901	4,400	4,928	5,519	5,906	6,555	10.9%
HPC-enabled AI (ML, DL & Other) Server Revenue	1,039	1,300	1,718	2,083	2,484	2,941	3,619	22.7%
Traditional Data Science (non-AI HPDA) Focused Server Revenue	2,460	2,601	2,682	2,845	3,036	2,965	2,937	2.5%

Future HPC System Design

AI and HPDA workloads pushing sites to consider new system architectures

- **As workloads become more diverse, system designs have shifted:**
  - Some sites are building single, large, heterogeneous systems to address a wide variety of applications
  - Some sites are building out multiple, smaller systems to handle different workloads specifically
  - Cloud resources are growing in utilization to address data-intensive workloads
- **Technology options have diversified as well**
  - New accelerator options, including AI-specific ASICs
  - Various memory, interconnect, and storage solutions
- **Compute resource allocation should be treated as an optimization problem:**
  - Find a balance among diverse technology options
  - Optimize for key workloads

Intersection of HPC and AI

Modeling and simulation workloads working in harmony with AI techniques

- **AI applications growing in the HPC space:**
  - Stand-alone AI models
  - AI incorporated into traditional simulation workloads:
    - Surrogate models
    - Data preparation and cleansing
    - Simulation steering with trained AI models
- **Mod/sim workloads benefiting from AI**
  - Acceleration of time to solution
  - Exploring new solution spaces
  - Parsing sparse matrices of data
- **AI benefiting from mod/sim workloads**
  - Generation of large synthetic datasets for training
  - Verification and testing of trained models in simulation

The Role of Explainability

For optimization, engagement, and compliance

- **Reproducibility and transparency as optimizers**
  - Automated monitoring hastens and ameliorates training
  - Time-saving capabilities (automated reporting, bias or drift detection) ease developer load and free valuable time
    - More models are developed overall
    - More models make it to production
- **Explainability drives engagement**
  - Many application spaces highly value auditability
  - Reassurance for previously hesitant domains
  - Contributes positively to development of AI workforce
- **Growing efforts to regulate and standardize**
  - Bolstered public knowledge and trust
  - Auditability as a legal obligation
  - Explainability tools mitigate regulatory fines

AI and HPC in the news

Interesting recent highlights of AI and HPC in the news

- **DALL-E from OpenAI**
  - Generation of images from text input
  - Model trained with 12 billion parameters
- **Meta AI system deployment**
  - Meta acquires large system from Penguin Computing
  - System consists of 760 DGX boxes, 1 exabyte of storage
  - "…Meta believes it will make it the largest AI supercomputer in the world."[1]
- **AI test bed at ANL[2]**
  - Multiple test systems from emergent AI processor and system vendors
  - Open calls for researchers to propose work to test and run on novel architectures
- **AI is one of three pillars for CORAL-2 procurements**
  - Frontier and Aurora both expected to be up and running this year
  - Systems include GPU technology and other architecture factors to address data-intensive workloads

1: https://tinyurl.com/yutnese7
2: https://tinyurl.com/mssbzast

HYPERION RESEARCH

What did we miss?

tsorensen@hyperionres.com

anorton@hyperionres.com

Our ISC22 Briefing Agenda

All registered attendees will receive the slide deck

- **Market Update and Forecasts**
- **Some Perspectives on European HPC**
- **Sustainability: No Longer a "Nice to Have"**
- **HPC and AI Talent Challenges**
- **Exascale Update**
- **Cloud Update**
- **Quantum Update**
- **AI Update**
- **Update on Storage & Interconnects**
- **Conclusions and Wrap-up**

Update on Storage and Interconnects

ISC22
June 2022

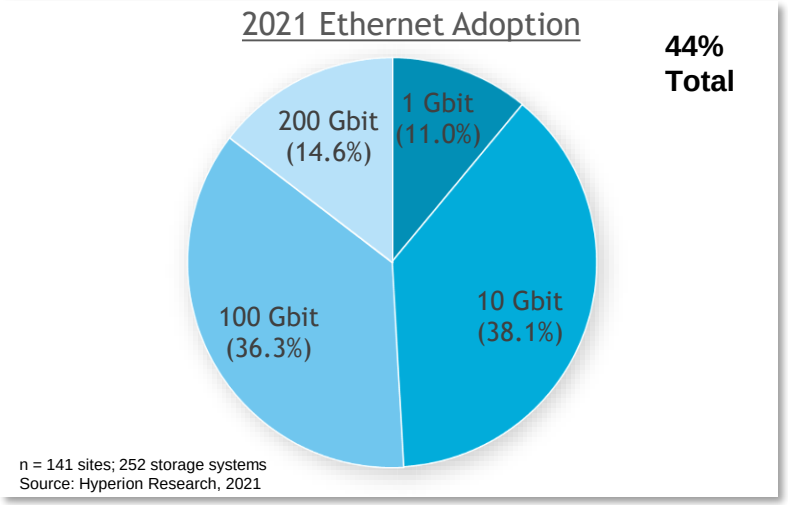
www.HyperionResearch.com
www.hpcuserforum.com

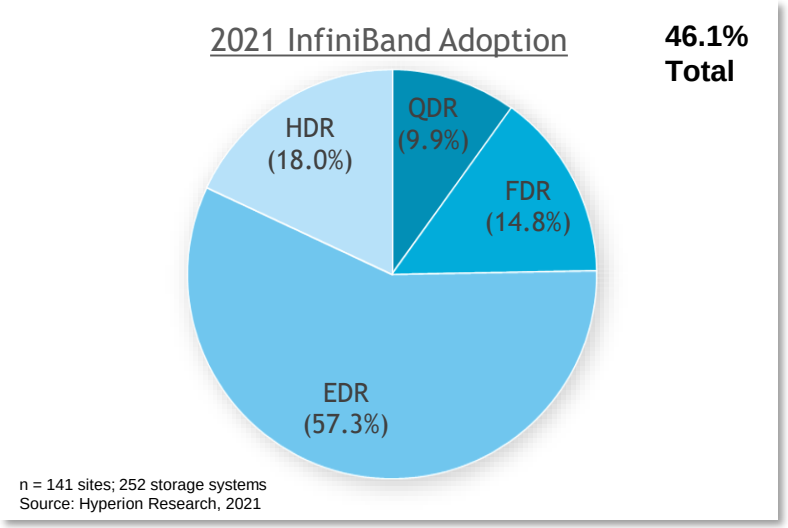
Mark Nossokoff

System Interconnect Preferences

InfiniBand edges ahead of Ethernet in aggregate adoption

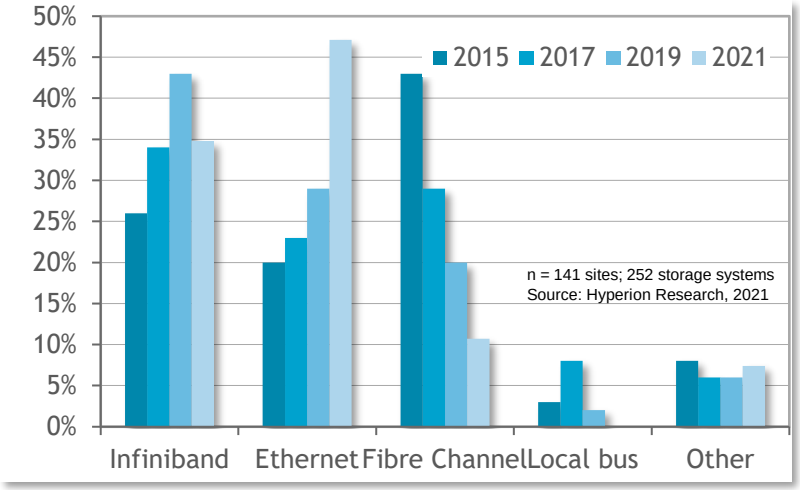
- **Initiated distinction between storage and system interconnects**
- **Ethernet**
  - Adopted at 44% of sites surveyed
  - 10 Gbit most widely adopted
- **InfiniBand**
  - Deployed at 46.1% of sites surveyed
  - EDR 100 Gbit most widely deployed
- **Omni-Path**
  - Some adoption across all sectors
  - 2nd largest deployed in academic sites (13.9%)



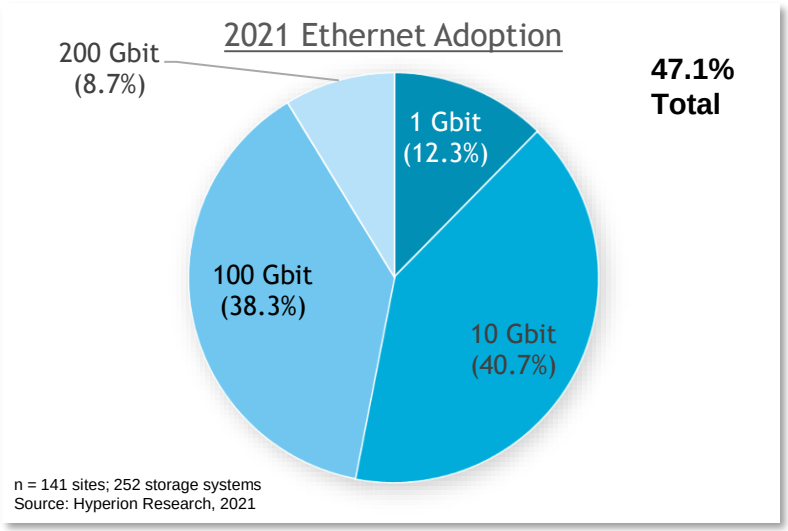


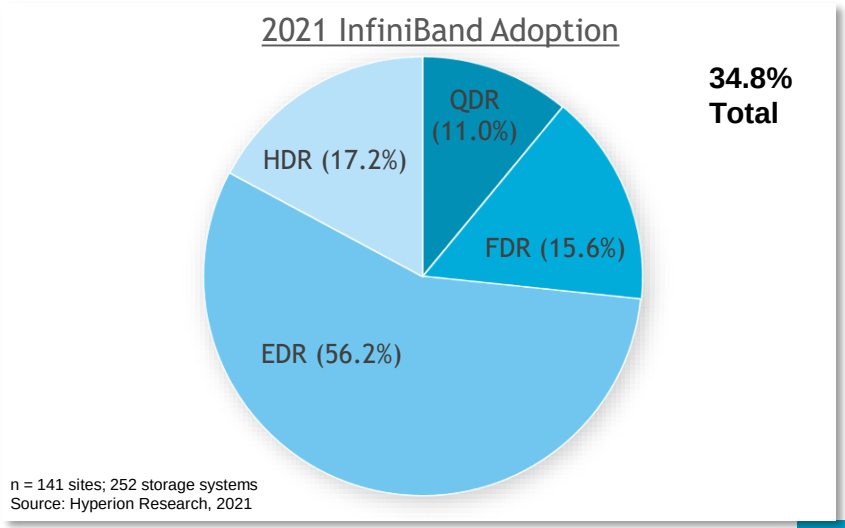
Storage Interconnect Preferences

Ethernet emerged as preferred storage interconnect



- **Sector preference differences**
  - Industry leans ethernet
  - Government split between ethernet and InfiniBand
  - Academia leans InfiniBand
- **Omni-Path is large part of other**





Interconnect Topics

System scale and heterogenous workloads are pushing existing system interconnects to their limits

- **Ethernet and ethernet-based**
  - Additional HPE Slingshot deployments
  - Rockport Networks emerging from stealth
- **Infiniband**
  - Continued investment in new features
  - Focus on datacenter scale and operating environment
- **OPA (OmniPath Architecture)**
  - DOE NNSA adoption of and investments in Cornelis Networks
    - Networking partner for Tri-labs CTS-2
    - $18M for NG-HPCN

Interconnect Topics

Composability, disaggregation driving interest in new interconnects and protocols for dramatic latency improvements

- **Emergence of DPUs**
  - Offload compute complex of data processing functionality
  - Mellanox, Fungible, and Pensando (soon-to-be-AMD) notable developers of DPUs
- **Industry-wide support of CXL**
  - Composable systems to dynamically pool resources depending on specific application and workload needs
  - Marvell acquires Tanzanite
  - Rambus acquires PLDA (Jun 2021) & Hardent (May 2022)
- **Google Aquila**
  - Convergence of ultra-low latency disaggregation and HPC/ML workload adoption by traditional datacenters
    - Cost-effective topology: direct-connected Dragon Fly
    - Ultra-low latency: shallow buffer cell-based GNet L2 protocol
    - Unified fabric: integrated RMA primitives with ToR-in NIC (TIN)
    - Co-existing in larger ethernet network: ethernet at the edges of Aquila-connected cliques

Interconnect Topics

Growing instantiations of chiplets are exposing the need for inter-die, intra-packaging interconnects

- **Collection of dies using different geometry nodes from different vendors**
- **Portability and re-use of older node-based functions**
- **Standardization efforts**
  - UCIe – Universal Chiplet Interconnect express
  - OCP ODSA – Open Compute Project Open Domain-Specific Architecture

HYPERION RESEARCH

PLEASE SHARE YOUR THOUGHTS!

mnossokoff@hyperionres.com

Our ISC22 Briefing Agenda

All registered attendees will receive the slide deck

- **Market Update and Forecasts**
- **Some Perspectives on European HPC**
- **Sustainability: No Longer a "Nice to Have"**
- **HPC and AI Talent Challenges**
- **Exascale Update**
- **Cloud Update**
- **Quantum Update**
- **AI Update**
- **Update on Storage & Interconnects**
- **Conclusions and Wrap-up**

Hyperion Research ISC22 Conclusions

May 2022

Earl Joseph, Bob Sorensen, Alex Norton,
Mark Nossokoff, Melissa Riddle,
Tom Sorensen and Jaclyn Ludema

www.HyperionResearch.com
www.hpcuserforum.com

Conclusions

- **2021 was a strong year with a 9% increase**
  - 2022 is also expected to be a strong year
    - Exascale systems will drive growth in 2022 to 2024
    - AI, HPDA/big data are hot areas
- **New technologies are showing up large numbers:**
  - Processors, AI hardware & software, memories, new storage approaches, etc.
  - Quantum
- **The cloud has become a viable option for many HPC workloads**
  - HPC in the cloud is lifting the sector writ large
- **Storage will likely see major growth driven by AI, big data and the need for much larger data sets**
- **There are growing concerns about the supply chain and HPC talent availability**

A Concern: HPC Expertise Shortage

The growing scarcity of HPC experts to implement new technologies is the number one roadblock for many HPC sites

- **Two major trends:**
  1) A shrinking HPC workforce
  2) A massive increase in system complexity

- **HPC experts are an aging workforce**
  - The pipeline of new HPC staff entering the workforce does not adequately match the outflow of retirees
  - Competition for HPC staff will intensify
- **Increasingly complex workloads are more difficult to manage**
  - Increasing HPC systems per site
  - Augmenting traditional modeling/simulation with AI and big data
  - Incorporating multiple processor types, co-processors, accelerators, and other specialized hardware
  - Balancing on-prem and cloud
  - And Enterprise IT users are entering HPC space, and need HPC expertise
- **HPC users will need major improvements in ease-of-use, ease-of-selection, & ease-of-optimization**

2022 HPC User Forum Meetings

Please join in!

- **June 21-22, 2022**
  - In-person at Oak Ridge National Laboratory
- **September 6-8, 2022**
  - In-person in Tucson, Arizona
- **October 3-6, 2022**
  - In-person in Paris (October 3-4)
  - In-person at HLRS, Stuttgart, Germany (October 5-6)

HYPERION RESEARCH

Thanks for joining us today!

We welcome questions, comments and suggestions

Please contact us at:
info@hyperionres.com