HPC Market Update During ISC21

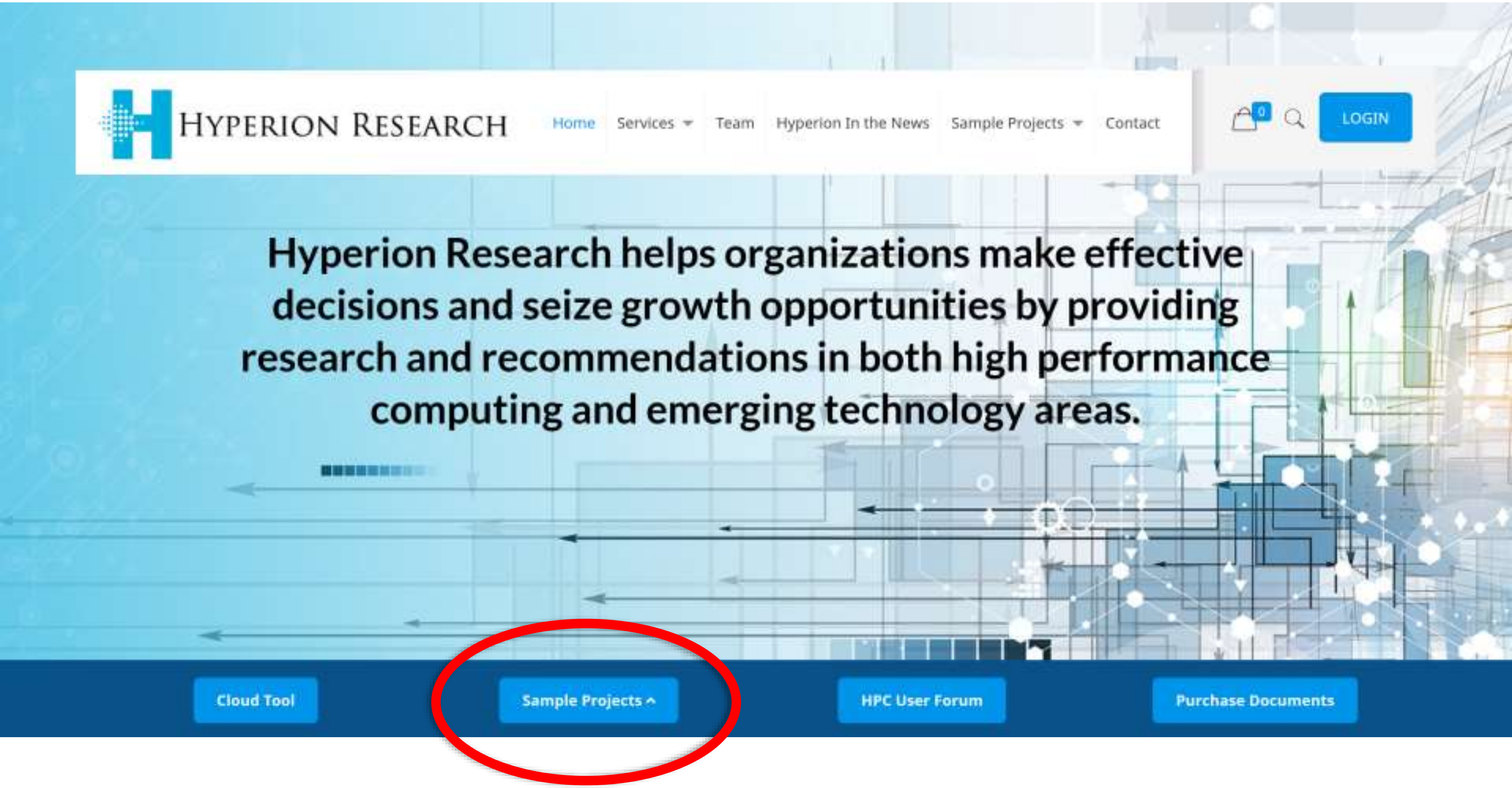
June 2021

Earl Joseph

www.HyperionResearch.com
www.hpcuserforum.com

Visit Our Website: www.HyperionResearch.com

Twitter: @HPC_Hyperion



Hyperion Research helps organizations make effective decisions and seize growth opportunities by providing research and recommendations in both high performance computing and emerging technology areas.

The Hyperion Research Team

Analysts

- Earl Joseph, CEO
- Steve Conway, Senior Advisor
- Bob Sorensen, SVP Research
- Mark Nossokoff, Senior Analyst
- Alex Norton, Principal Analyst & Data Manager
- Melissa Riddle, Associate Analyst
- Thomas Sorensen, Research Associate

International Consultants

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

Operations

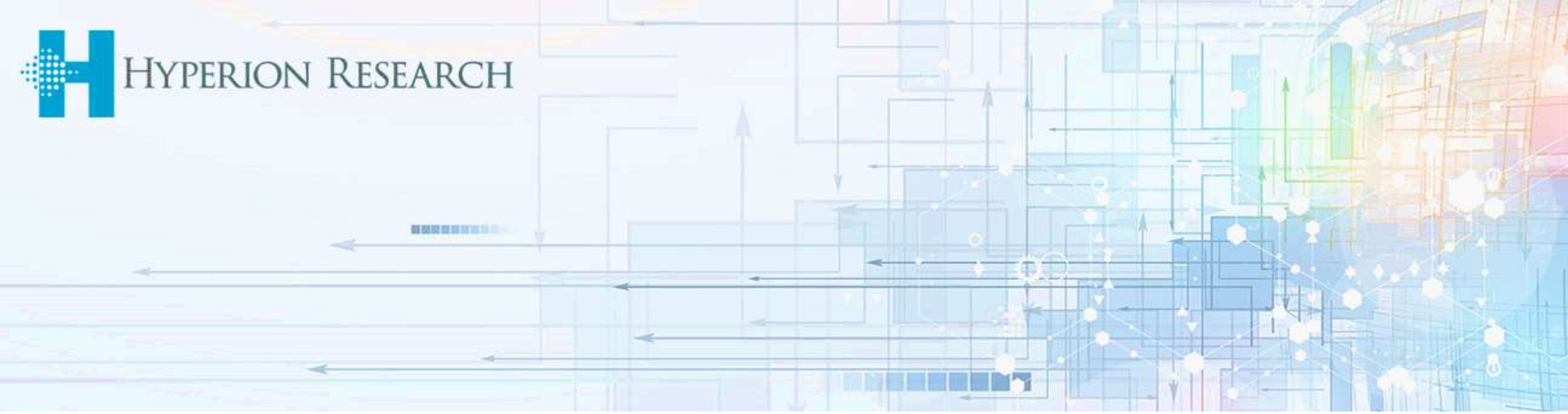
- Jean Sorensen, COO

Global Accounts

- Rene Copeland, Strategic Mktg & Sales Advisor
- Mike Thorp, Sr. Global Sales Executive
- Kurt Gantrish, Sr. Account Executive

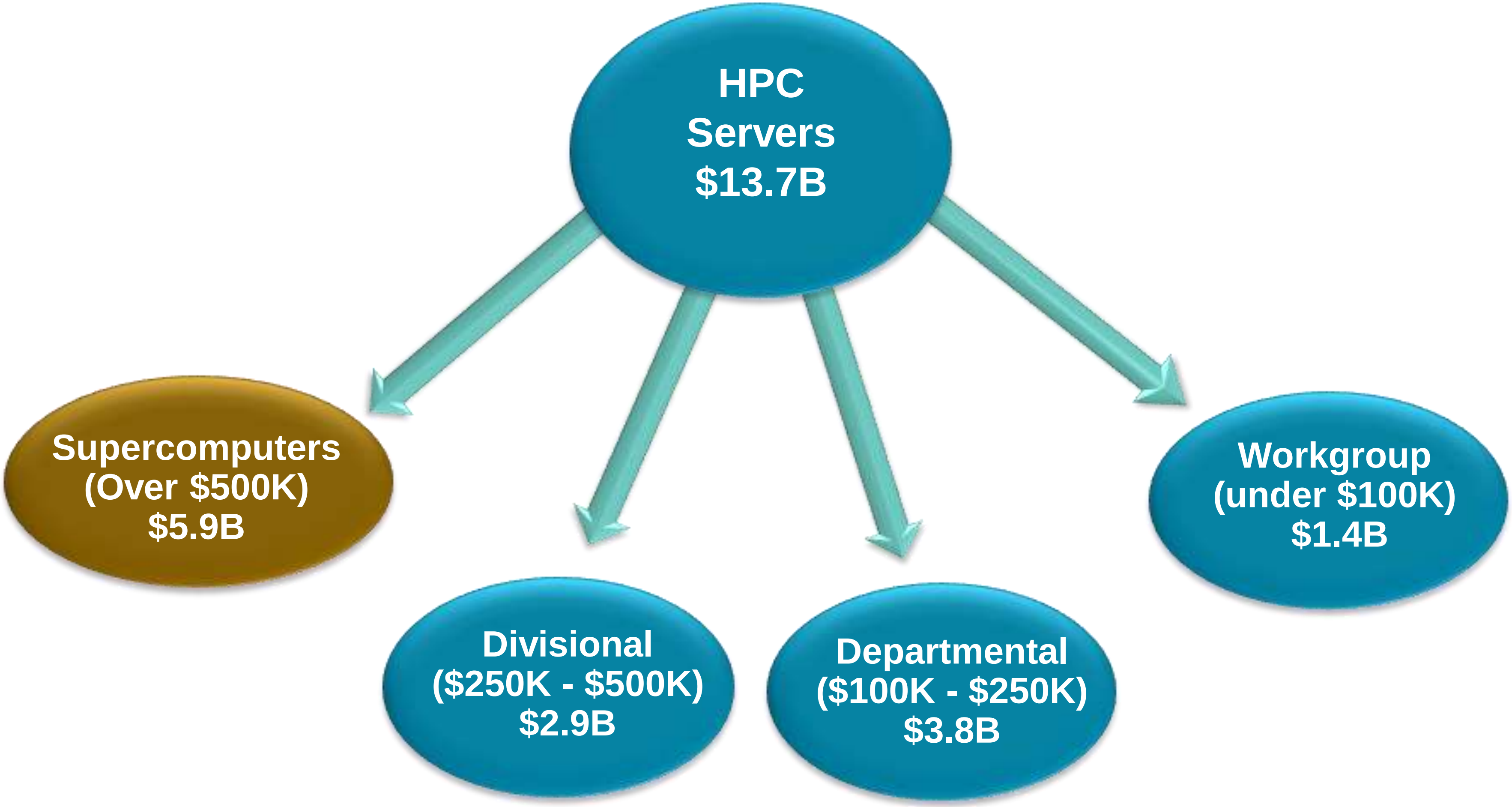
Data Collection

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



2020 Market Results

The 2020 Worldwide On-Prem HPC Server Market: $13.7 Billion (up 1.1%)



HPC Servers $13.7B

Supercomputers (Over $500K) $5.9B

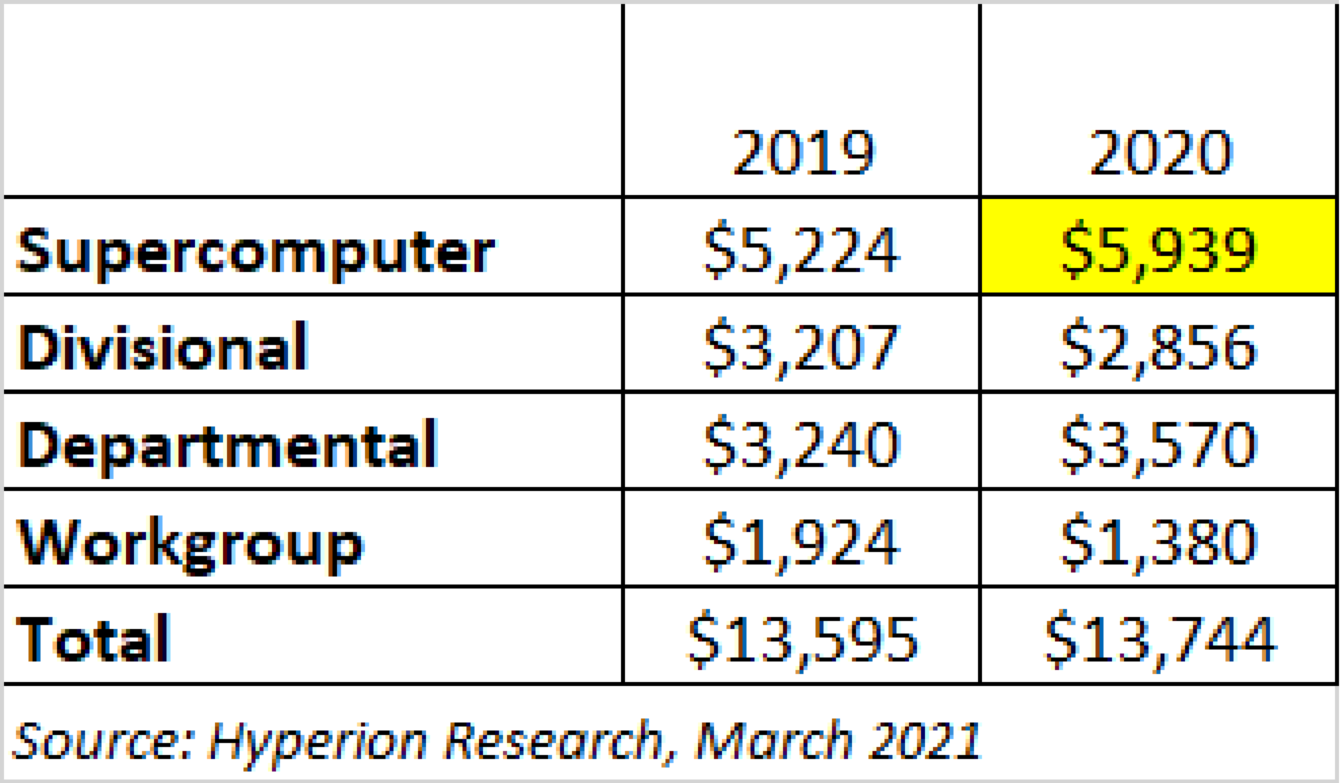
Divisional ($250K - $500K) $2.9B

Departmental ($100K - $250K) $3.8B

Workgroup (under $100K) $1.4B

WW HPC Market By Segments ($ Millions)

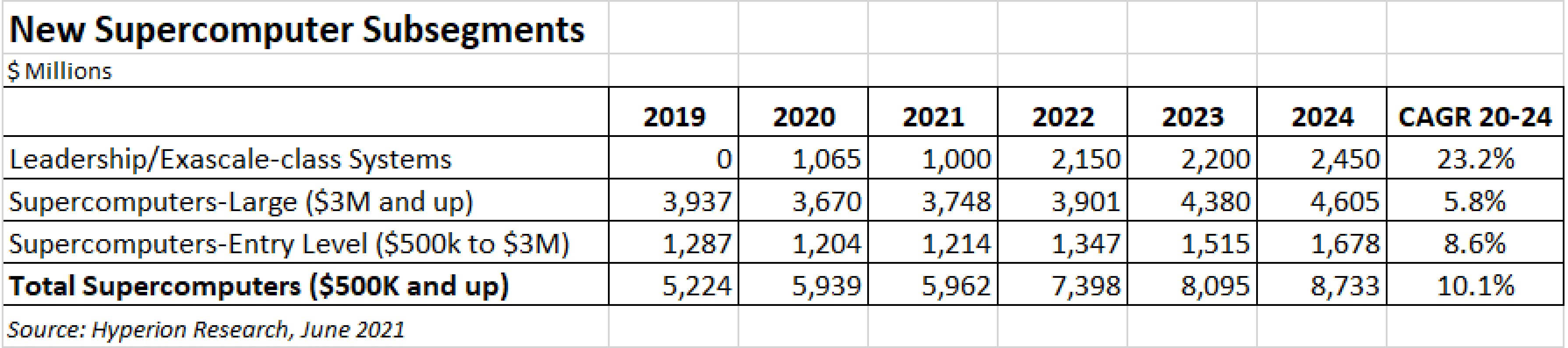
Fugaku made the supercomputer segment strong in 2020, while the workgroup declined the most



	2019	2020
Supercomputer	$5,224	$5,939
Divisional	$3,207	$2,856
Departmental	$3,240	$3,570
Workgroup	$1,924	$1,380
Total	$13,595	$13,744

Source: Hyperion Research, March 2021

New Supercomputer Subsegments

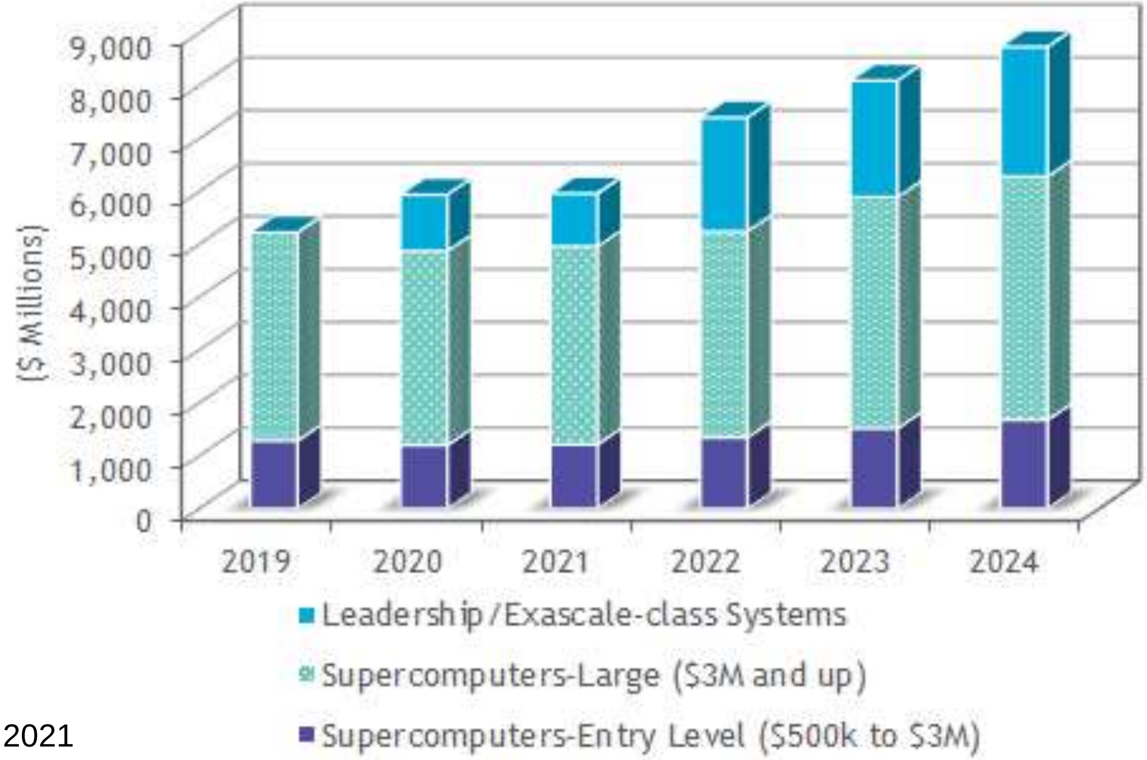


New Supercomputer Subsegments

$ Millions

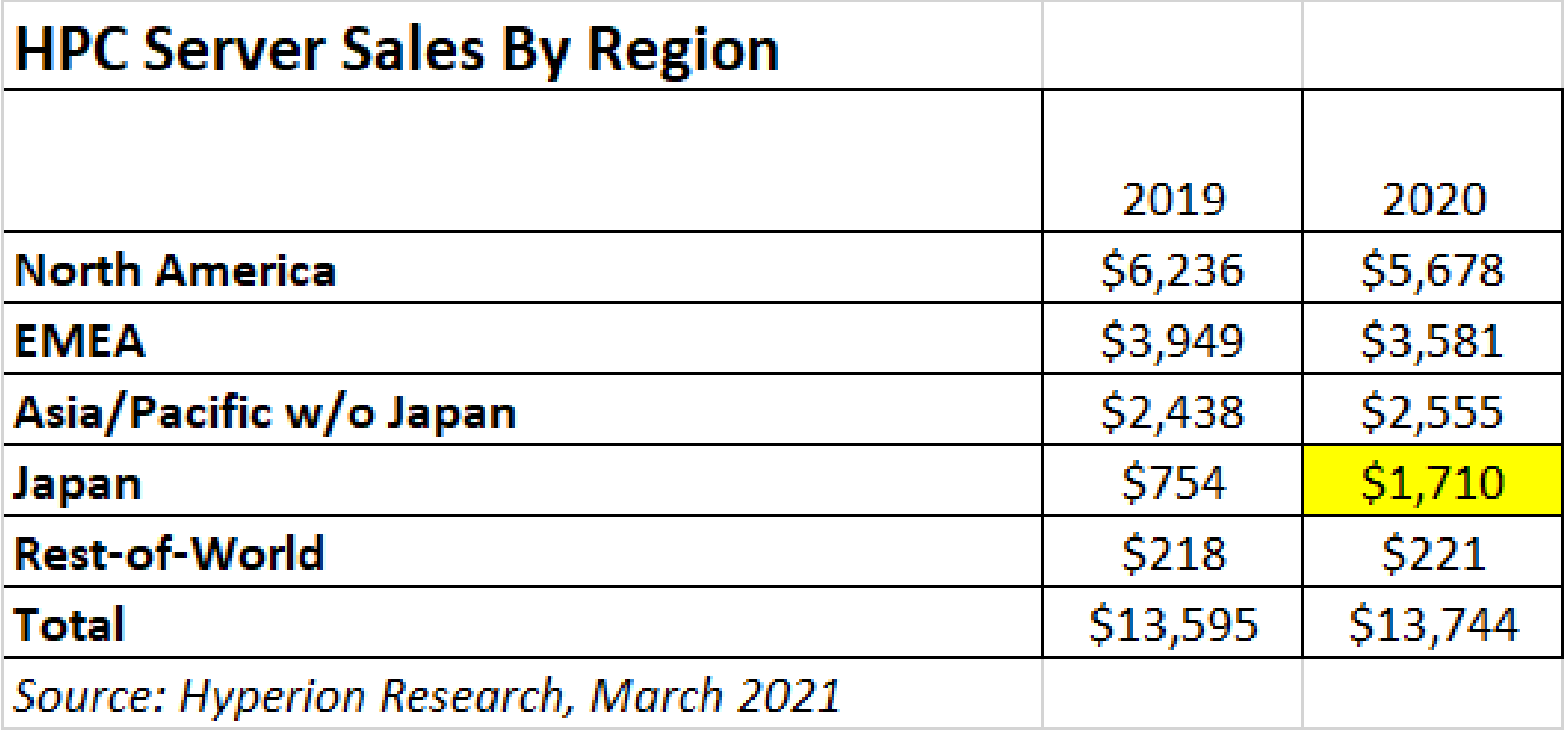
	2019	2020	2021	2022	2023	2024	CAGR 20-24
Leadership/Exascale-class Systems	0	1,065	1,000	2,150	2,200	2,450	23.2%
Supercomputers-Large ($3M and up)	3,937	3,670	3,748	3,901	4,380	4,605	5.8%
Supercomputers-Entry Level ($500k to $3M)	1,287	1,204	1,214	1,347	1,515	1,678	8.6%
Total Supercomputers ($500K and up)	5,224	5,939	5,962	7,398	8,095	8,733	10.1%

Source: Hyperion Research, June 2021



WW HPC Market By Regions

In 2020: very high growth in Japan, the rest of the market declined by over 7%

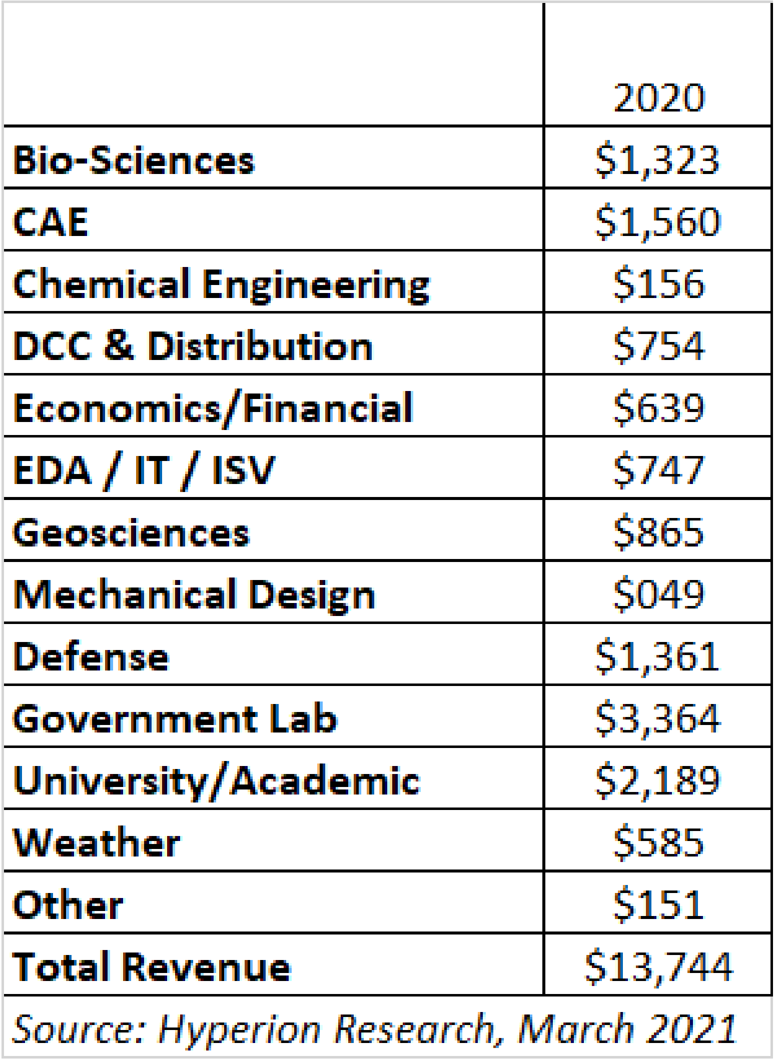


HPC Server Sales By Region		
	2019	2020
North America	$6,236	$5,678
EMEA	$3,949	$3,581
Asia/Pacific w/o Japan	$2,438	$2,555
Japan	$754	$1,710
Rest-of-World	$218	$221
Total	$13,595	$13,744

Source: Hyperion Research, March 2021

WW HPC Market By Verticals ($ Millions)

Five segments are over a $ billion a year

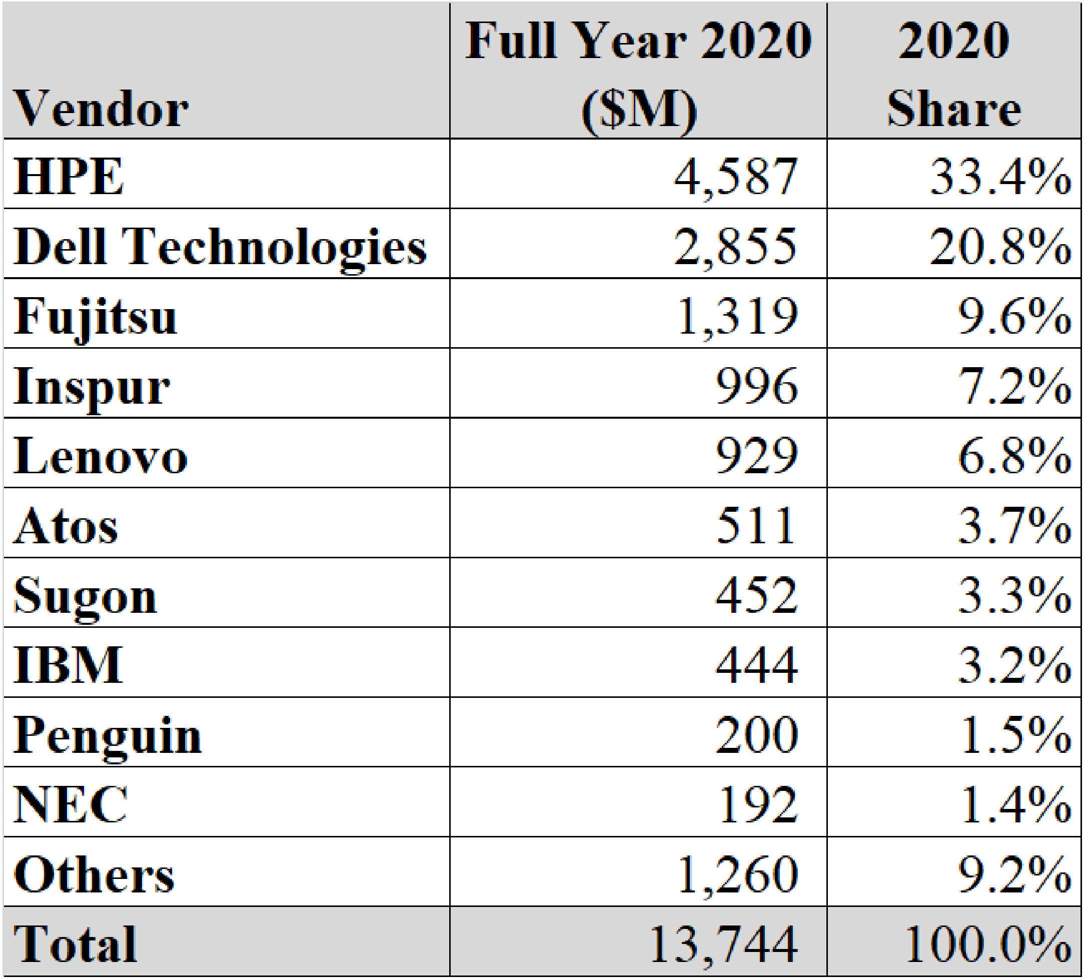


	2020
Bio-Sciences	$1,323
CAE	$1,560
Chemical Engineering	$156
DCC & Distribution	$754
Economics/Financial	$639
EDA / IT / ISV	$747
Geosciences	$865
Mechanical Design	$049
Defense	$1,361
Government Lab	$3,364
University/Academic	$2,189
Weather	$585
Other	$151
Total Revenue	$13,744

Source: Hyperion Research, March 2021

Worldwide HPC Vendor Market Shares

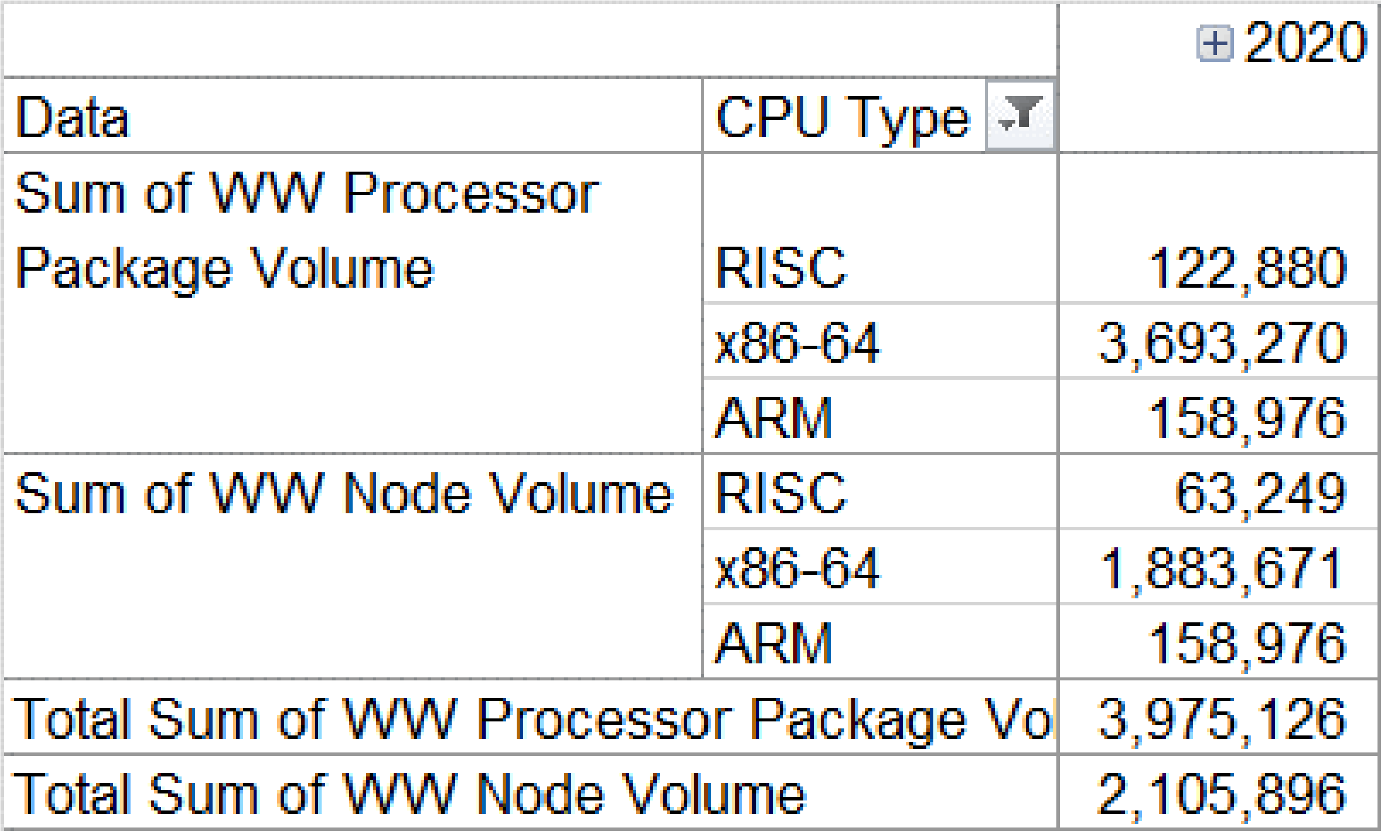
($ Millions)



Vendor	Full Year 2020 ($M)	2020 Share
HPE	4,587	33.4%
Dell Technologies	2,855	20.8%
Fujitsu	1,319	9.6%
Inspur	996	7.2%
Lenovo	929	6.8%
Atos	511	3.7%
Sugon	452	3.3%
IBM	444	3.2%
Penguin	200	1.5%
NEC	192	1.4%
Others	1,260	9.2%
Total	13,744	100.0%

WW HPC Market By Processors

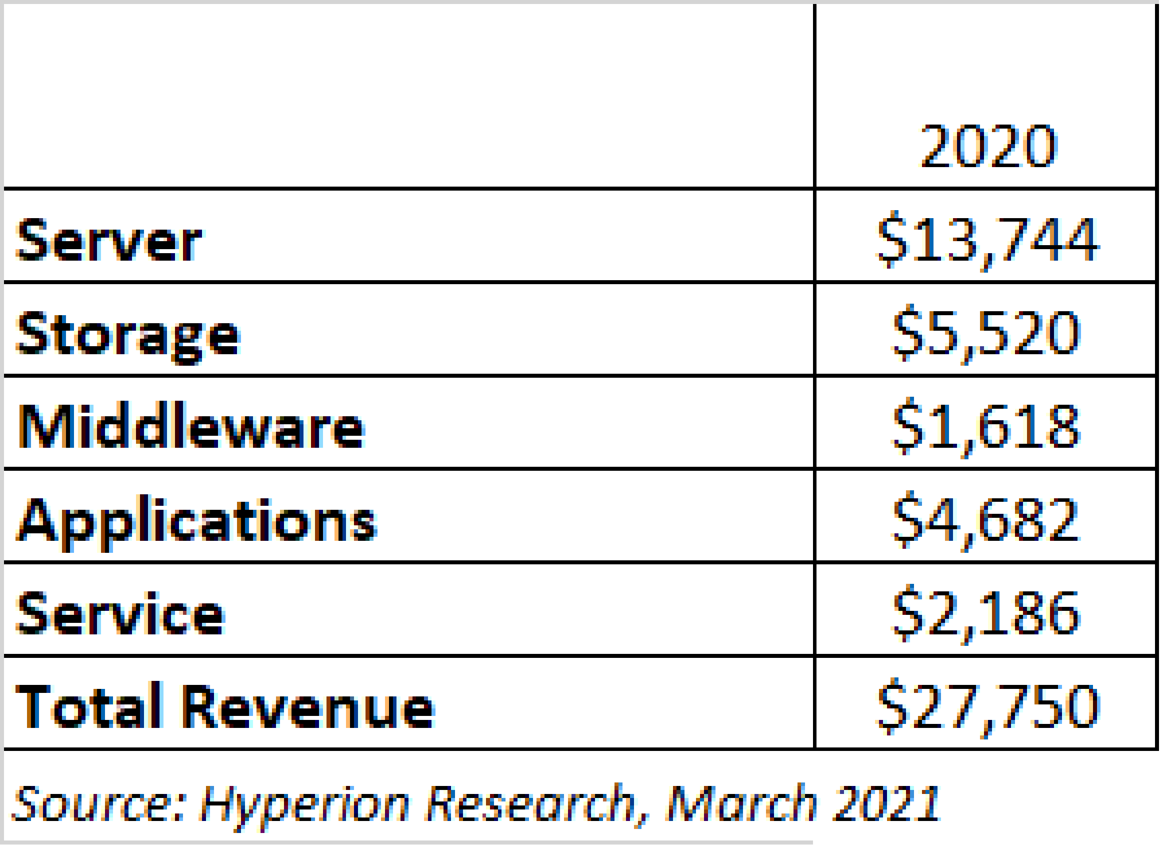
~4 million processors are sold each year



		⊞ 2020
Data	CPU Type	
Sum of WW Processor Package Volume	RISC	122,880
	x86-64	3,693,270
	ARM	158,976
Sum of WW Node Volume	RISC	63,249
	x86-64	1,883,671
	ARM	158,976
Total Sum of WW Processor Package Vo		3,975,126
Total Sum of WW Node Volume		2,105,896

The Broader On-premise Market Areas ($millions)

The 2020 total on-prem HPC spending exceeded $27 billion (excluding cloud spending)



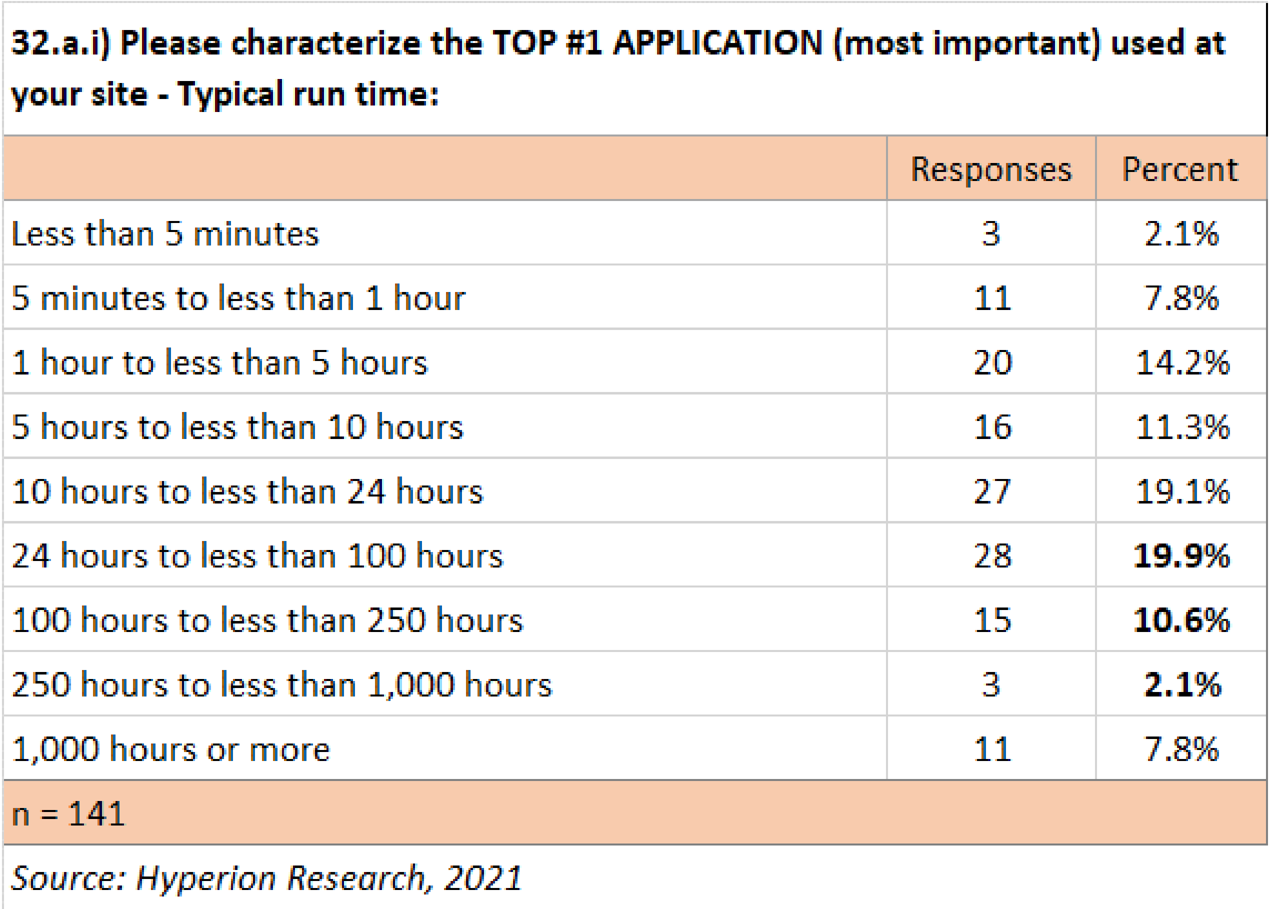
	2020
Server	$13,744
Storage	$5,520
Middleware	$1,618
Applications	$4,682
Service	$2,186
Total Revenue	$27,750

Source: Hyperion Research, March 2021

Some Results From Our New End-user Study

Largest Application Runtime

From our soon to be release end-user MCS study ~33% of the #1 applications run for over 24 hours



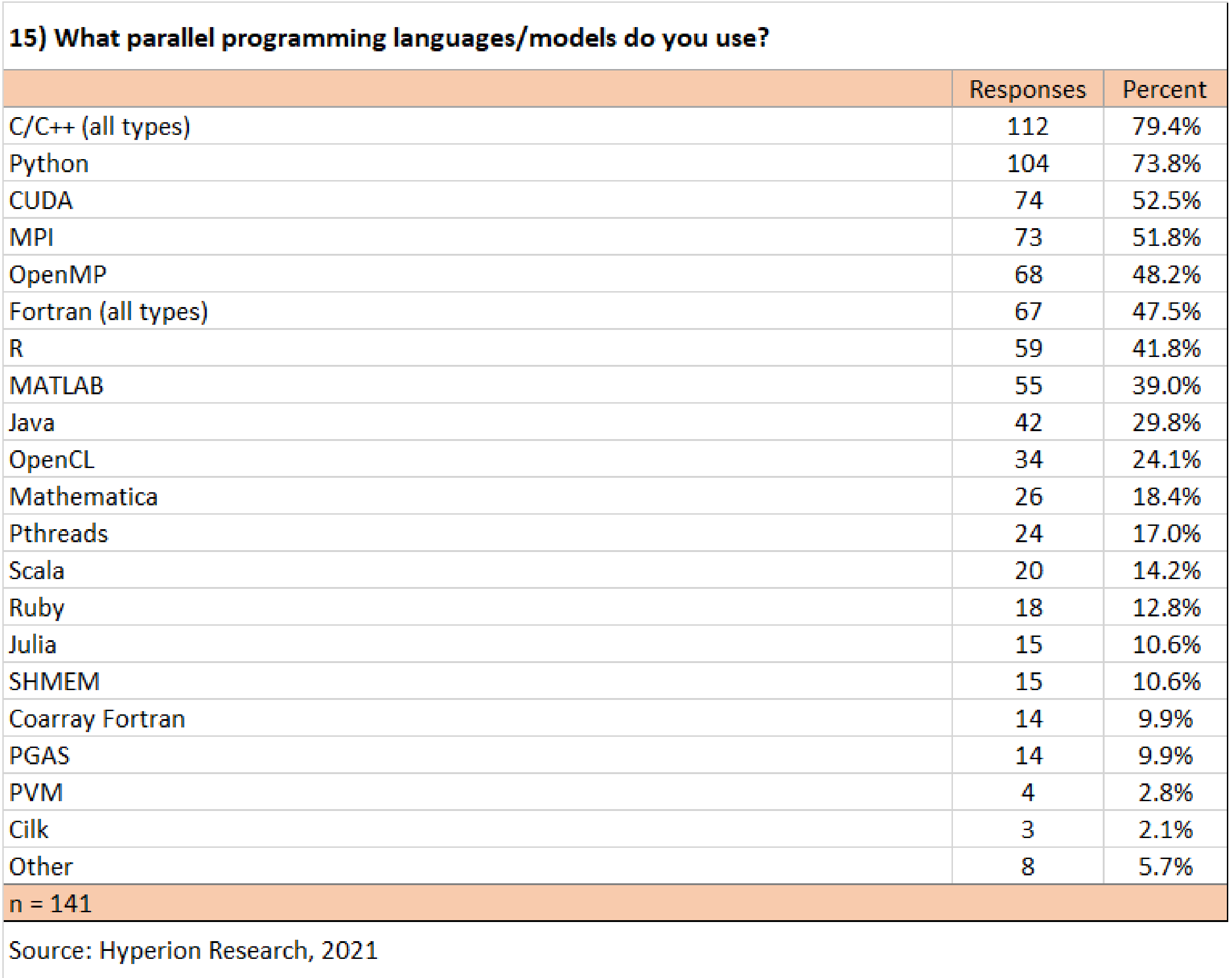
32.a.i) Please characterize the TOP #1 APPLICATION (most important) used at your site - Typical run time:

	Responses	Percent
Less than 5 minutes	3	2.1%
5 minutes to less than 1 hour	11	7.8%
1 hour to less than 5 hours	20	14.2%
5 hours to less than 10 hours	16	11.3%
10 hours to less than 24 hours	27	19.1%
24 hours to less than 100 hours	28	**19.9%**
100 hours to less than 250 hours	15	**10.6%**
250 hours to less than 1,000 hours	3	**2.1%**
1,000 hours or more	11	7.8%
n = 141		

Source: Hyperion Research, 2021

Programming Models Used Today

From our soon to be release end-user MCS study



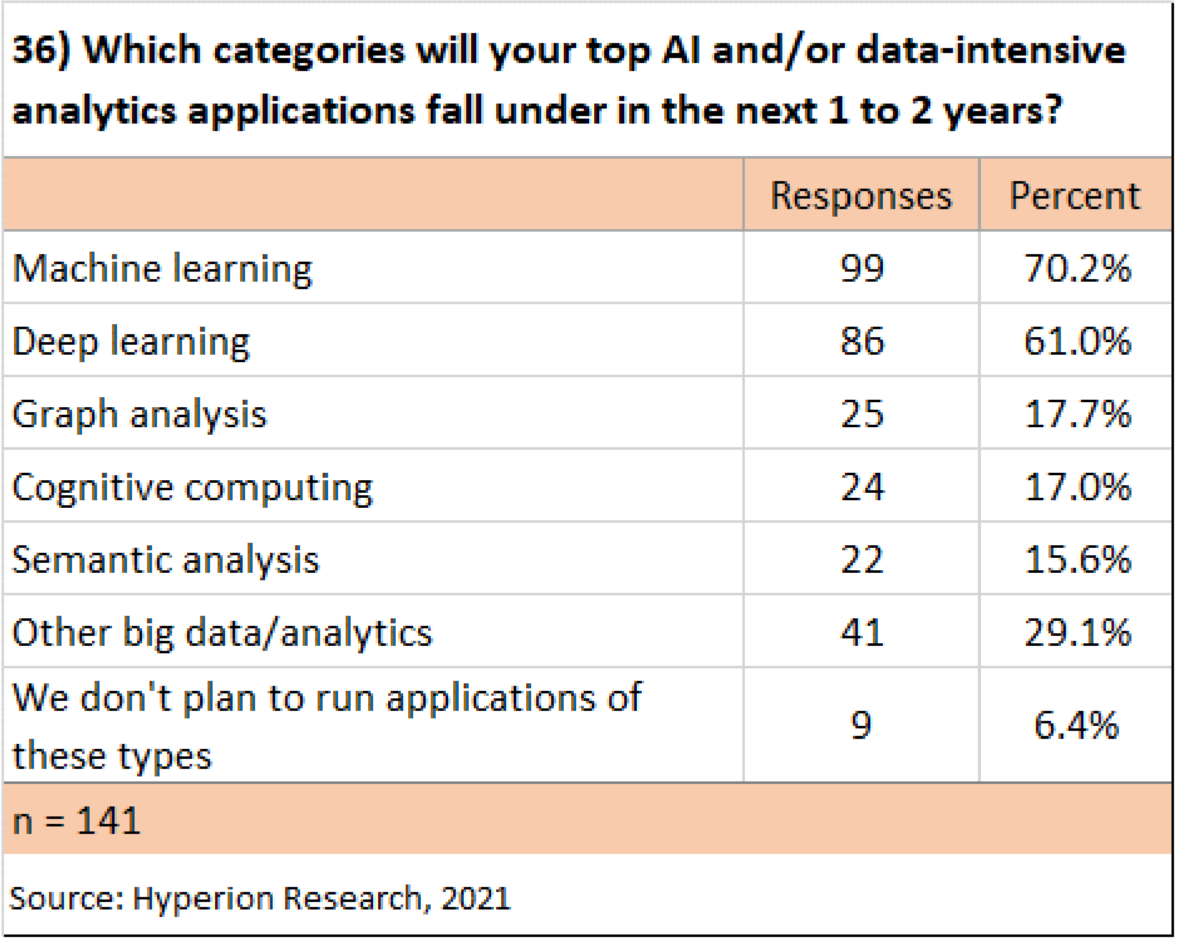
15) What parallel programming languages/models do you use?

	Responses	Percent
C/C++ (all types)	112	79.4%
Python	104	73.8%
CUDA	74	52.5%
MPI	73	51.8%
OpenMP	68	48.2%
Fortran (all types)	67	47.5%
R	59	41.8%
MATLAB	55	39.0%
Java	42	29.8%
OpenCL	34	24.1%
Mathematica	26	18.4%
Pthreads	24	17.0%
Scala	20	14.2%
Ruby	18	12.8%
Julia	15	10.6%
SHMEM	15	10.6%
Coarray Fortran	14	9.9%
PGAS	14	9.9%
PVM	4	2.8%
Cilk	3	2.1%
Other	8	5.7%
n = 141		

Source: Hyperion Research, 2021

Use Of Different AI/ML/DL Approaches

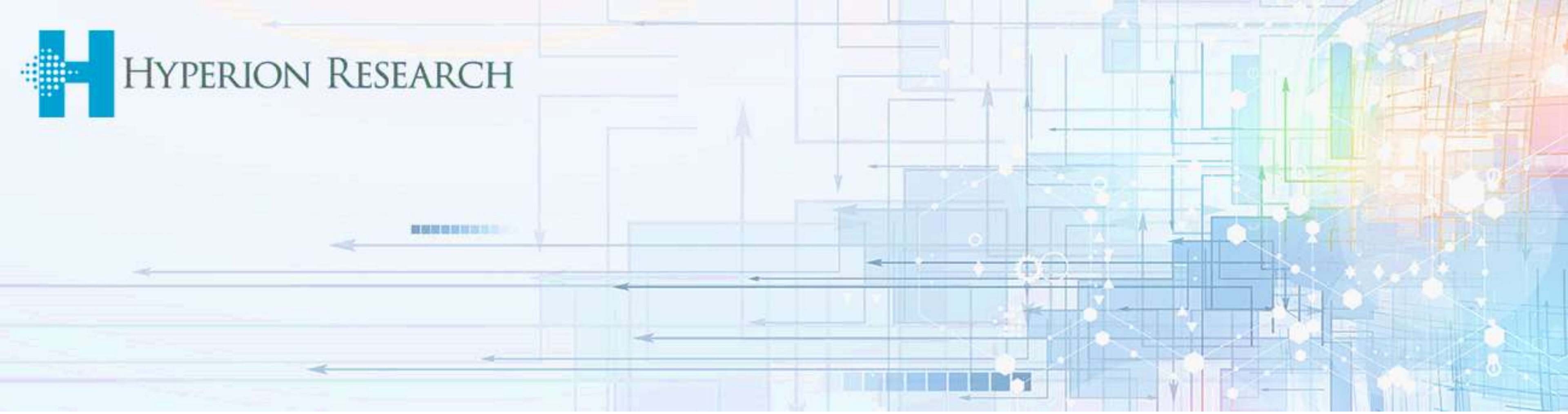
From our soon to be release end-user MCS study



36) Which categories will your top AI and/or data-intensive analytics applications fall under in the next 1 to 2 years?

	Responses	Percent
Machine learning	99	70.2%
Deep learning	86	61.0%
Graph analysis	25	17.7%
Cognitive computing	24	17.0%
Semantic analysis	22	15.6%
Other big data/analytics	41	29.1%
We don't plan to run applications of these types	9	6.4%
n = 141		

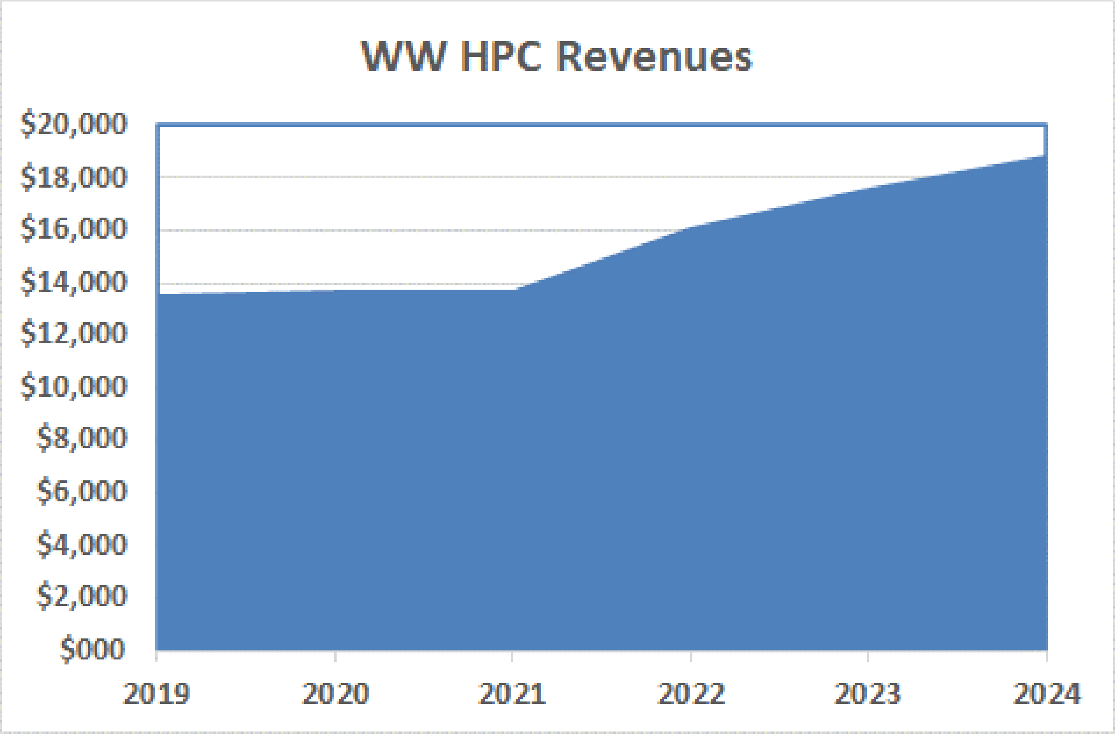
Source: Hyperion Research, 2021



HPC Forecasts

HPC On-Prem HPC Server Forecast ($millions)

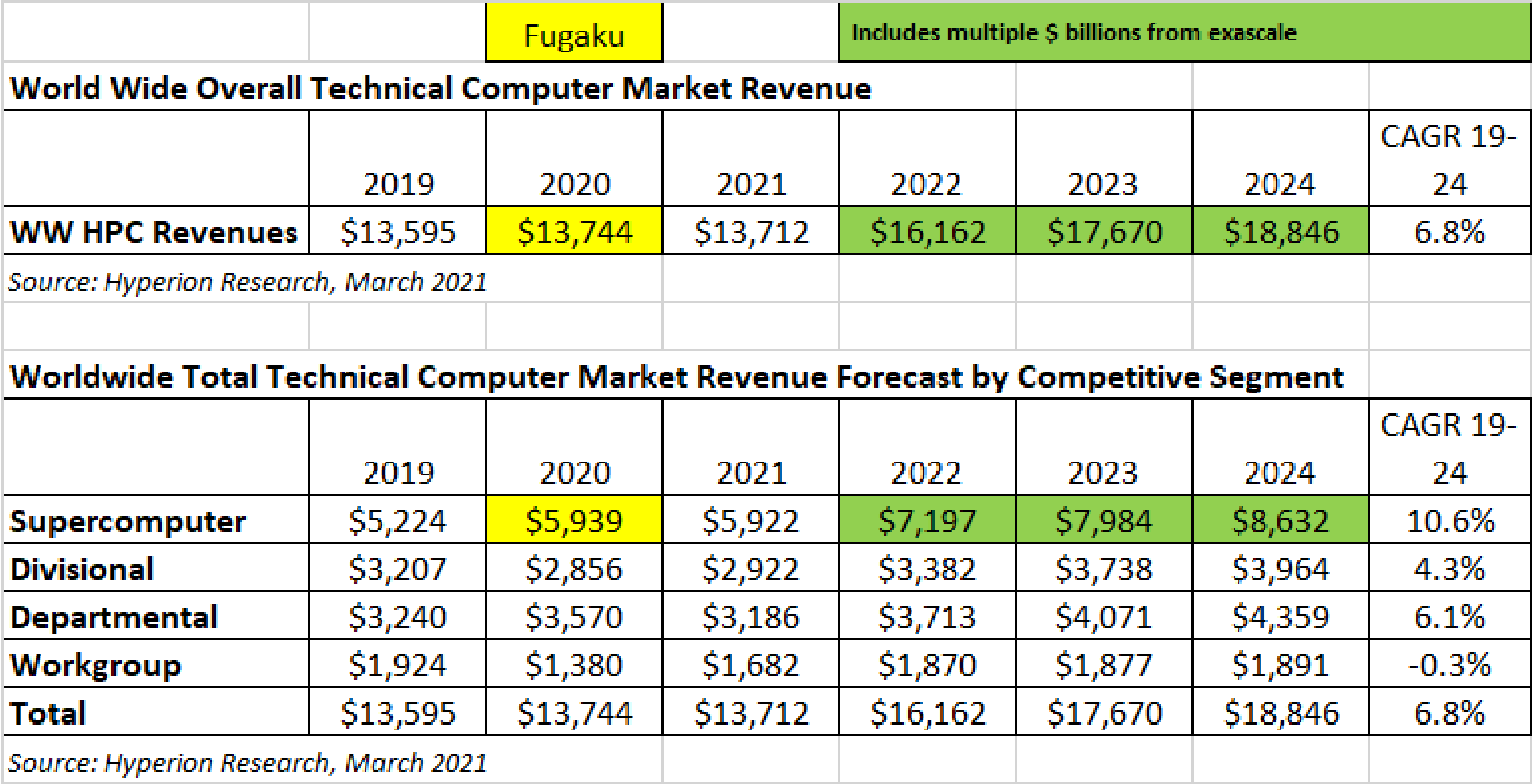
- **The five-year CAGR (2019 to 2024) is 6.8%**
  - Reaching close to $19 billion in 2024
  - Exascale is adding major growth
  - Covid is expected to still impact 2021



HPC On-Prem Server Forecast

($millions)

The overall CAGR is now 6.8%



		Fugaku		Includes multiple $ billions from exascale			
World Wide Overall Technical Computer Market Revenue							
	2019	2020	2021	2022	2023	2024	CAGR 19-24
WW HPC Revenues	$13,595	$13,744	$13,712	$16,162	$17,670	$18,846	6.8%

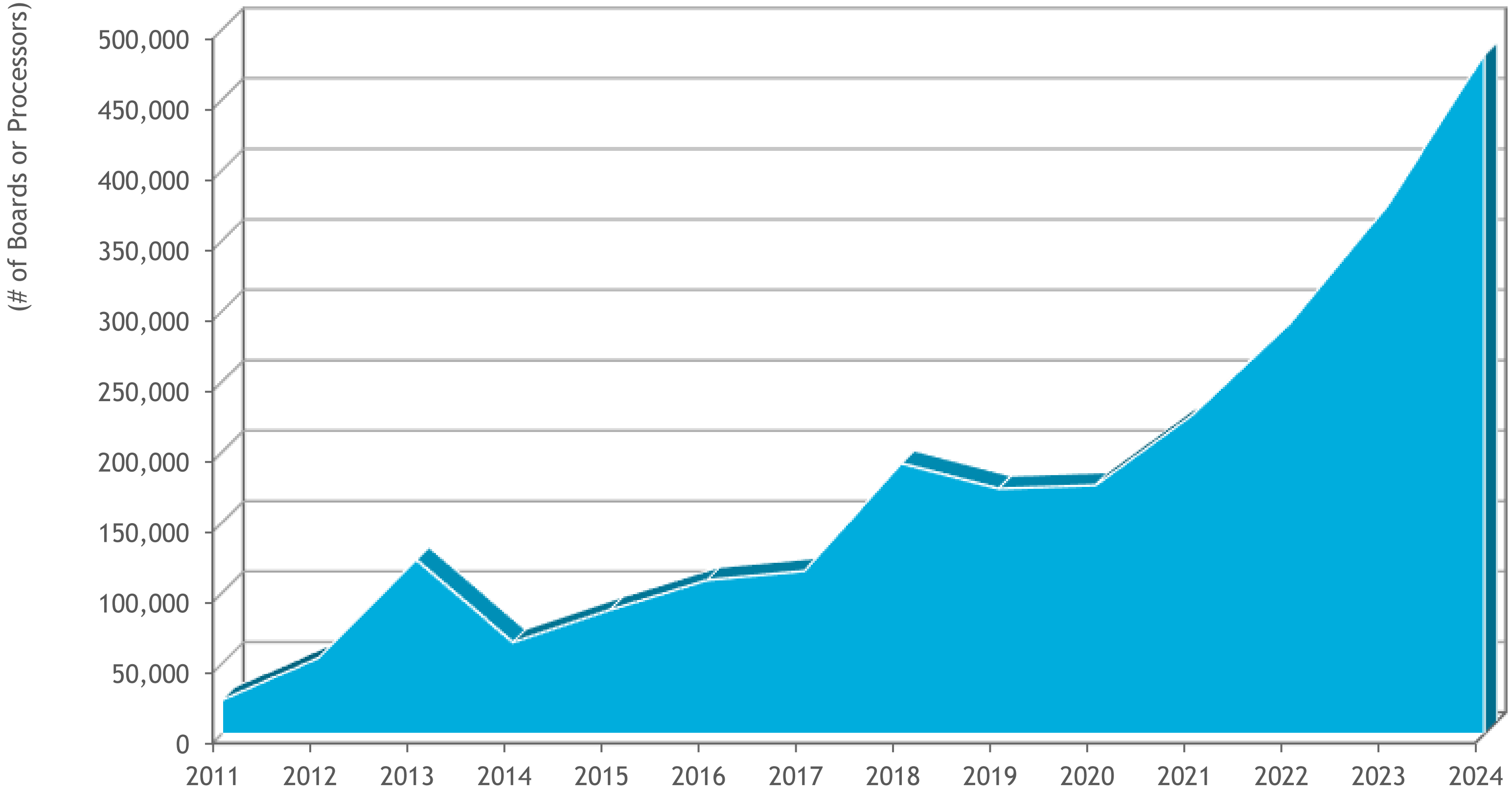
Source: Hyperion Research, March 2021

Worldwide Total Technical Computer Market Revenue Forecast by Competitive Segment							
	2019	2020	2021	2022	2023	2024	CAGR 19-24
Supercomputer	$5,224	$5,939	$5,922	$7,197	$7,984	$8,632	10.6%
Divisional	$3,207	$2,856	$2,922	$3,382	$3,738	$3,964	4.3%
Departmental	$3,240	$3,570	$3,186	$3,713	$4,071	$4,359	6.1%
Workgroup	$1,924	$1,380	$1,682	$1,870	$1,877	$1,891	-0.3%
Total	$13,595	$13,744	$13,712	$16,162	$17,670	$18,846	6.8%

Source: Hyperion Research, March 2021

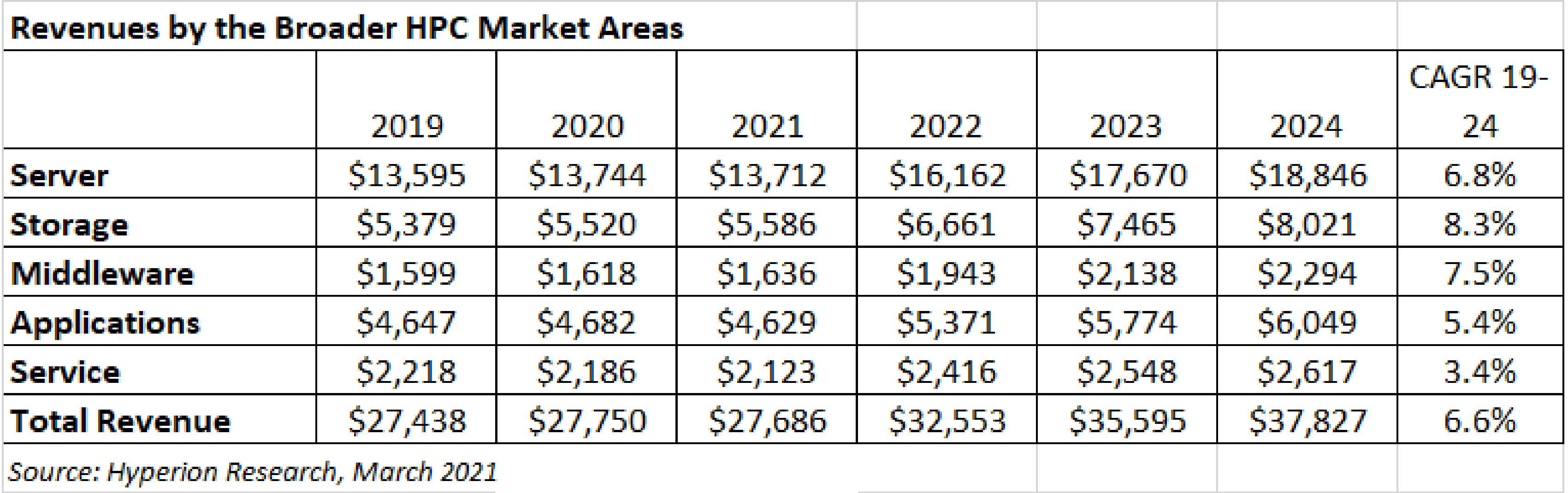
GPU/Accelerator Forecast

Anticipated high growth for accelerators over next 5 years



On-Prem Forecasts For The Broader Market Areas ($millions)

Storage is expected to grow the most at 8.3% CAGR



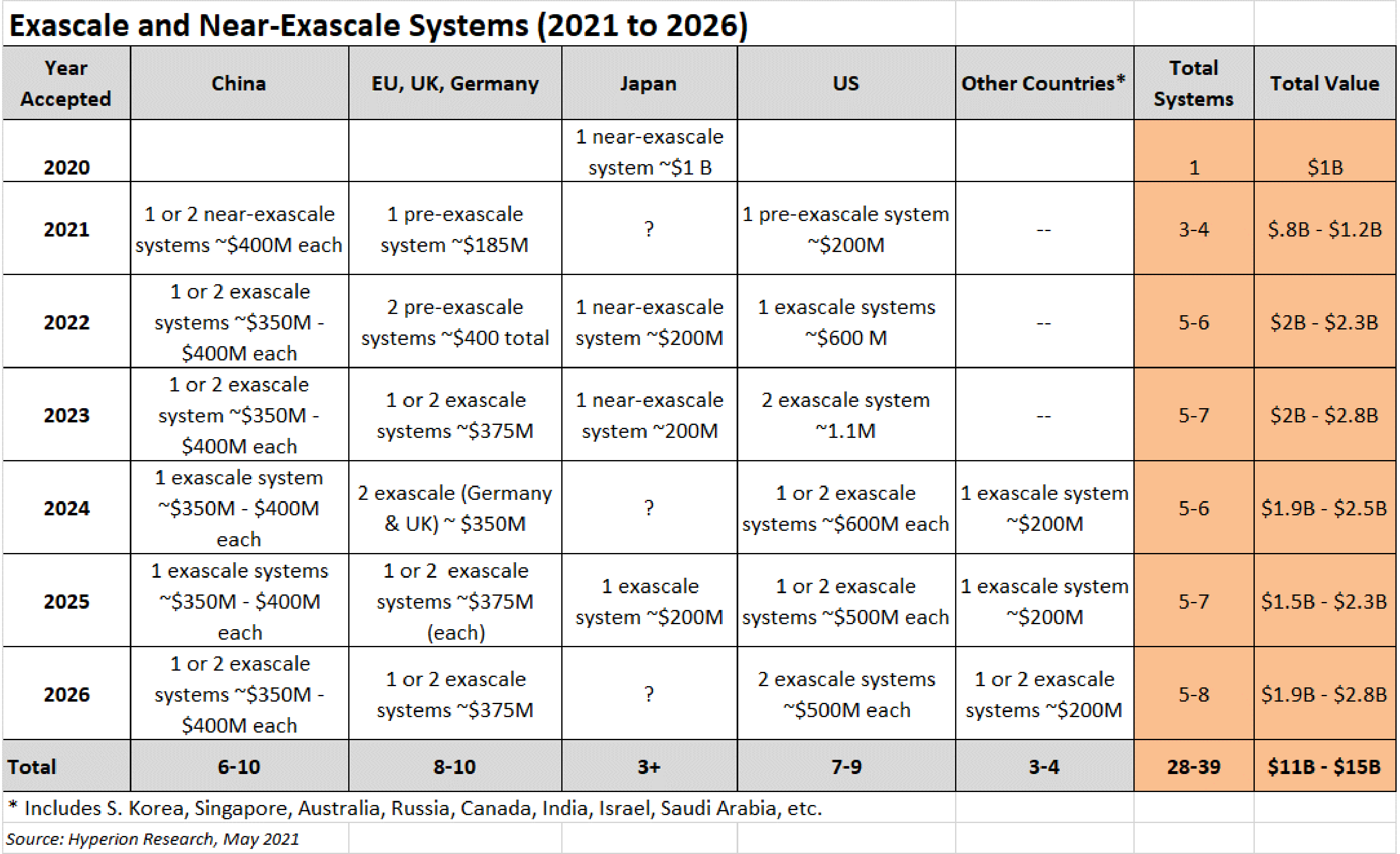
Revenues by the Broader HPC Market Areas

	2019	2020	2021	2022	2023	2024	CAGR 19-24
Server	$13,595	$13,744	$13,712	$16,162	$17,670	$18,846	6.8%
Storage	$5,379	$5,520	$5,586	$6,661	$7,465	$8,021	8.3%
Middleware	$1,599	$1,618	$1,636	$1,943	$2,138	$2,294	7.5%
Applications	$4,647	$4,682	$4,629	$5,371	$5,774	$6,049	5.4%
Service	$2,218	$2,186	$2,123	$2,416	$2,548	$2,617	3.4%
Total Revenue	$27,438	$27,750	$27,686	$32,553	$35,595	$37,827	6.6%

Source: Hyperion Research, March 2021

The Exascale Market (System Acceptances)

Over 30 systems and over $11 billion in value



Exascale and Near-Exascale Systems (2021 to 2026)

Year Accepted	China	EU, UK, Germany	Japan	US	Other Countries*	Total Systems	Total Value
2020			1 near-exascale system ~$1 B			1	$1B
2021	1 or 2 near-exascale systems ~$400M each	1 pre-exascale system ~$185M	?	1 pre-exascale system ~$200M	--	3-4	$.8B - $1.2B
2022	1 or 2 exascale systems ~$350M - $400M each	2 pre-exascale systems ~$400 total	1 near-exascale system ~$200M	1 exascale systems ~$600 M	--	5-6	$2B - $2.3B
2023	1 or 2 exascale system ~$350M - $400M each	1 or 2 exascale systems ~$375M	1 near-exascale system ~200M	2 exascale system ~1.1M	--	5-7	$2B - $2.8B
2024	1 exascale system ~$350M - $400M each	2 exascale (Germany & UK) ~ $350M	?	1 or 2 exascale systems ~$600M each	1 exascale system ~$200M	5-6	$1.9B - $2.5B
2025	1 exascale systems ~$350M - $400M each	1 or 2 exascale systems ~$375M (each)	1 exascale system ~$200M	1 or 2 exascale systems ~$500M 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 ~$375M	?	2 exascale systems ~$500M each	1 or 2 exascale systems ~$200M	5-8	$1.9B - $2.8B
Total	**6-10**	**8-10**	**3+**	**7-9**	**3-4**	**28-39**	**$11B - $15B**

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

Source: Hyperion Research, May 2021

HYPERION RESEARCH

New Trends on Using Cloud for HPC Workloads

June 2021

Alex Norton and Mark Nossokoff

www.HyperionResearch.com
www.hpcuserforum.com

HPC Cloud Forecast

HPC cloud market forecast to reach about $9 billion in 2024

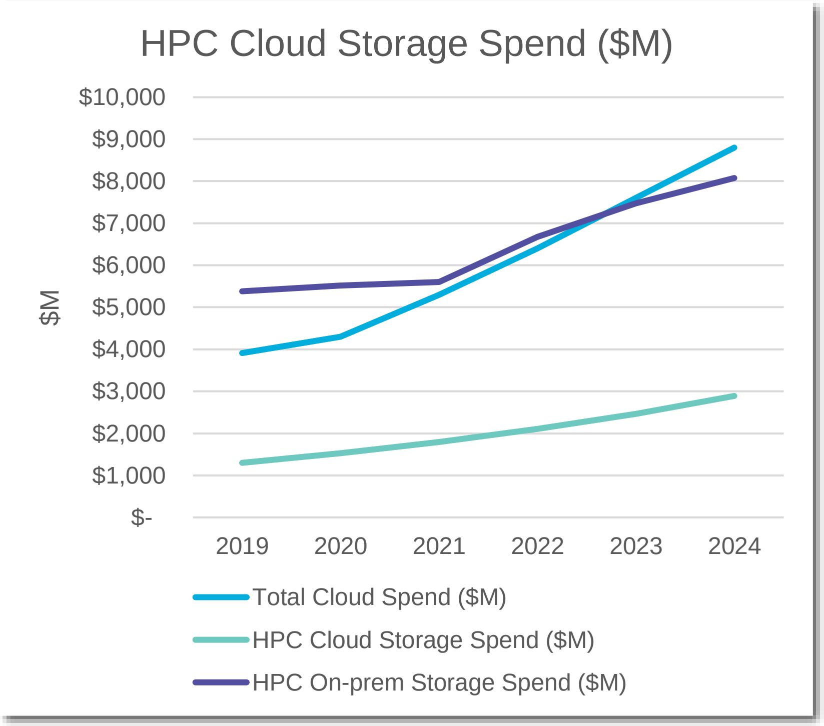
($M)	2019	2020	2021	2022	2023	2024	CAGR '19-'24
HPC Public Cloud Spend Forecast	3,910	4,300	5,300	6,400	7,600	8,800	17.6%
On-Premise HPC Server Forecast	13,595	13,744	13,741	16,197	17,708	18,977	6.9%

- **Public HPC cloud numbers are based on HPC user cloud spending, not CSP reported HPC revenues**
- **Cloud spending growth significantly outpaces on-premises**
- **Roughly 1/3 of cloud spend is for storage in the cloud, the rest for compute**

HPC Cloud Storage Forecast

Cloud Storage '19-'24 CAGR double equivalent on-prem storage CAGR

- **Total cloud spend of $3.9B in 2019**
- **Total cloud spend to surpasses on-prem storage spend in 2024**
- **Storage represented 1/3 of total 2019 cloud spend**
- **Cloud storage CAGR double on-prem storage CAGR**
  - Cloud: 17.3%
  - On-prem: 8.5%



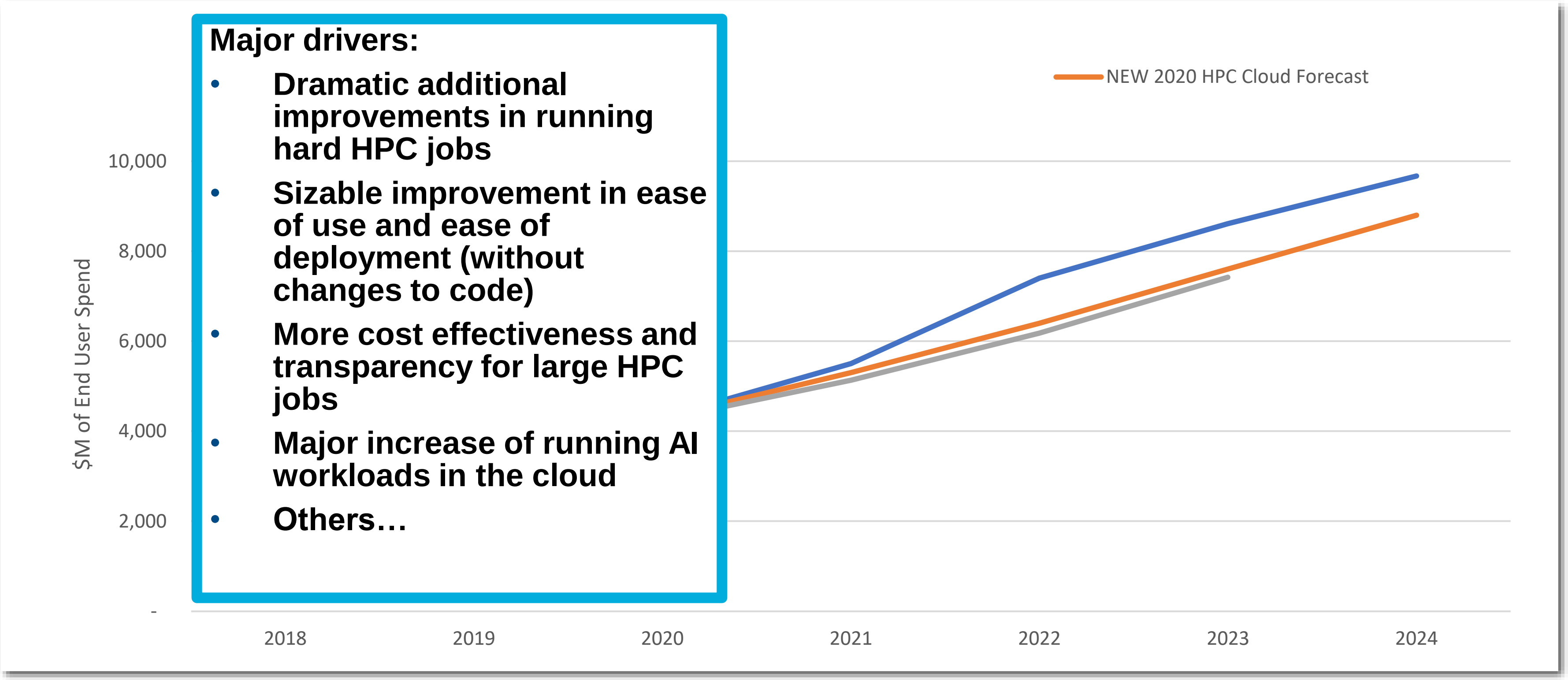
Source: Hyperion Research, 2021

Source: Hyperion Research, 2021

	2019	2020	2021	2022	2023	2024	'19 -'24 CAGR
Total Cloud Spend ($M)	$ 3,910	$ 4,300	$ 5,300	$ 6,400	$ 7,600	$ 8,800	17.6%
HPC Cloud Storage Spend ($M)	$ 1,303	$ 1,529	$ 1,793	$ 2,104	$ 2,467	$ 2,894	17.3%
HPC On-prem Storage Spend ($M)	$ 5,379	$ 5,520	$ 5,605	$ 6,675	$ 7,478	$ 8,075	8.5%

Cloud Market Possibilities

A hypothetical HPC cloud forecast incorporating potential market drivers



Major drivers:

- **Dramatic additional improvements in running hard HPC jobs**
- **Sizable improvement in ease of use and ease of deployment (without changes to code)**
- **More cost effectiveness and transparency for large HPC jobs**
- **Major increase of running AI workloads in the cloud**
- **Others…**

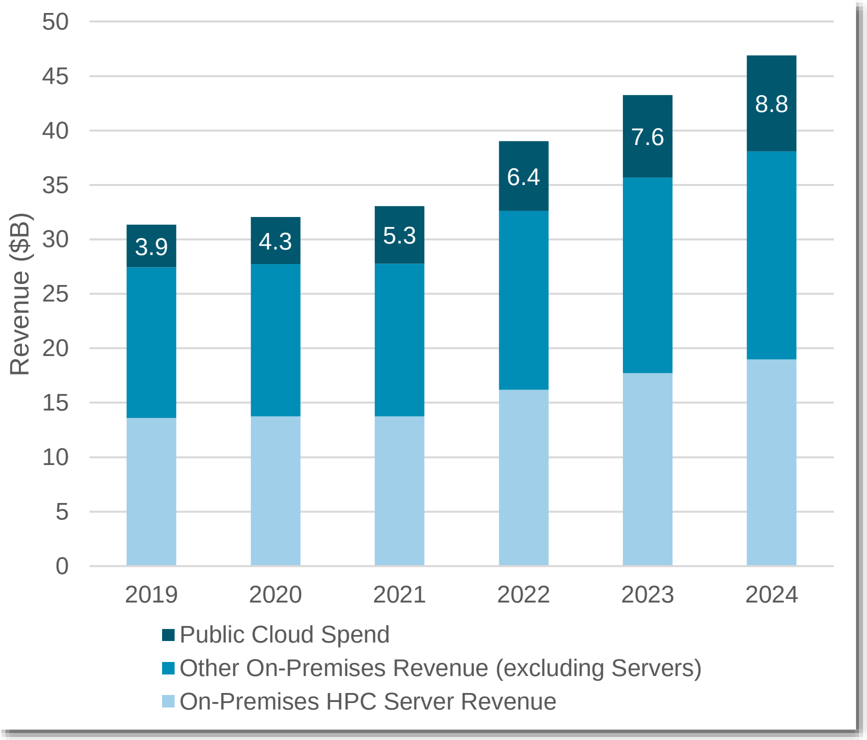
HPC Cloud Vertical Forecast

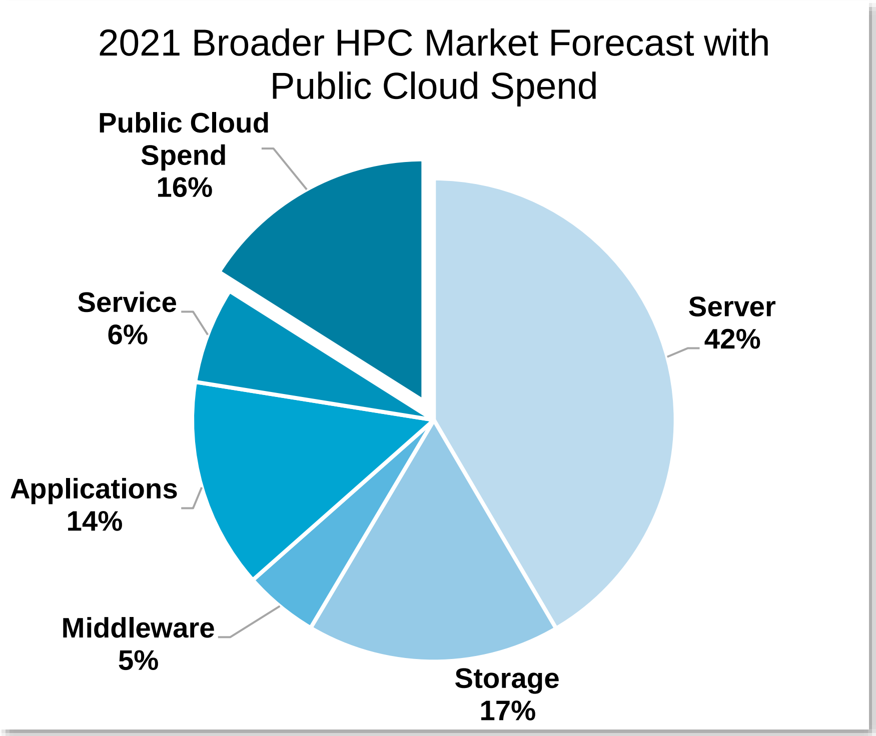
Bio-sciences and CAE the early adopting verticals; weather, geosciences and academia show highest growth over forecast period

($M)	2018	2019	2024	2019-2024 CAGR
Bio-Sciences	$778	$1,230	$2,453	14.8%
CAE	$469	$733	$1,540	16.0%
Chemical Engineering	$62	$98	$211	16.6%
DCC & Distribution	$141	$222	$519	18.5%
Economics/Financial	$123	$195	$430	17.2%
EDA	$178	$285	$677	18.9%
Geosciences	$148	$240	$660	22.4%
Mechanical Design	$12	$20	$44	17.5%
Defense	$185	$296	$705	18.9%
Government Lab	$173	$274	$625	17.9%
University/Academic	$123	$197	$528	21.8%
Weather	$26	$42	$220	39.0%
Other	$49	$79	$188	18.9%
Total	$2,466	$3,910	$8,800	17.6%

A Complete HPC Market Picture

Incorporating the cloud to the broader market forecast

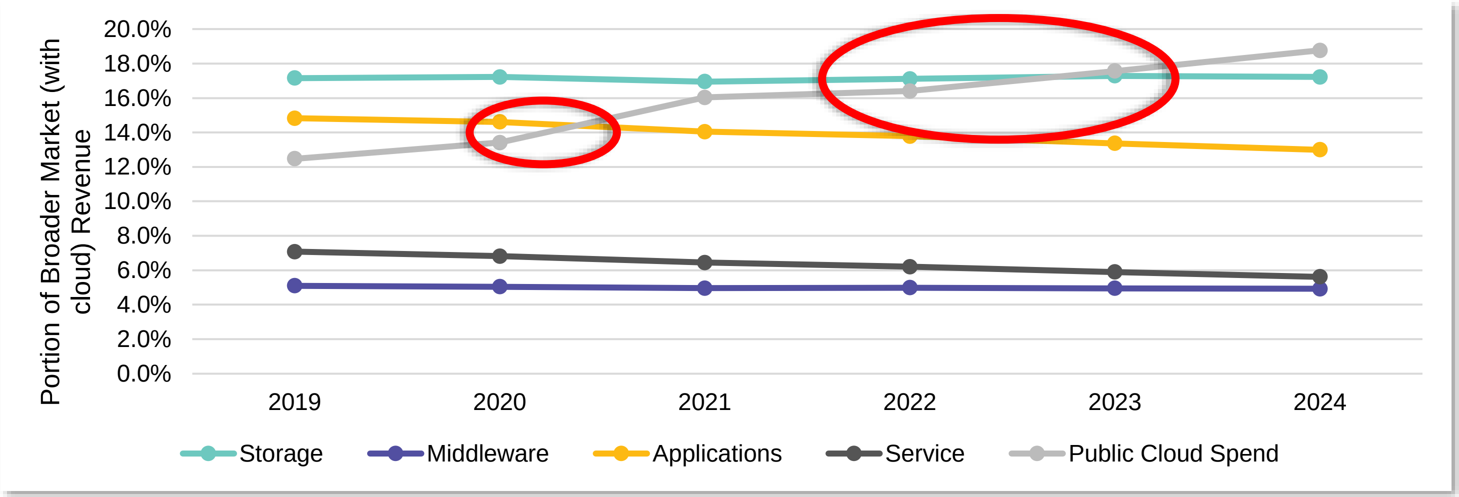




© Hyperion Research 2021

A Complete HPC Market Picture

HPC in the cloud projected to become larger part of HPC broader market



- **On-premises server revenues represent at least 40% of the HPC broader market (with cloud) over the forecast period**
- **Among other categories, public cloud spend will take over applications revenue for the 2nd largest segment, and by 2024 will be the second largest segment**

MAJOR RESEARCH FINDINGS: Impact of HPC Cloud on On-Premises

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

- **Today: Public cloud resources are complementary to on-premises deployments**
  - Many longitudinal studies show that cloud is used primarily for burst capabilities by many HPC users
- **The Next 12 Months: A recent study showed that almost 50% of the users are altering on-premises deployments due to cloud considerations**
  - Migrating HPC workloads to cloud platforms requires new skills for datacenter managers and researchers
    - Much of this education and training on using the cloud addresses 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

Barriers to HPC Cloud Adoption

Barriers to HPC cloud adoption have remained the same over the past few years

- **The most common barrier for users when evaluating cloud usage is cost**
  - Cloud computing can be very expensive for certain workloads, especially those that require long run times, large data transfers, or are “hard HPC jobs”
  - CSPs are working to make cost more transparent, as well as providing low-cost options, like spot instances
  - Users are evaluating which workloads are both cost effective and performant in the cloud
- **Data locality, data gravity, and data movement remain important issues for HPC users as well**
  - Upload and download speeds can be very long for larger data sets, and the cost can be high, resulting in some workloads remaining on-premises
  - Data movement within a cloud platform can also be costly
- **Security issues are usually in the top three most important barriers for HPC users as well, especially with new data privacy laws around the world**

Drivers of Adoption

Cloud usage continues to increase, driven by a few consistent factors

- **Burst capabilities are the most important driver for HPC cloud adoption**
- **Cost effectiveness for specific applications has grown in impact of HPC users' evaluation of cloud**
  - Applications that parallelize well and can take advantage of spot-pricing can be cost effective in the cloud
  - Applications that require expensive hardware not available on-prem (including accelerators, memory technologies, interconnects, etc.) can be more cost effective to run in the cloud rather than invest in the hardware on-prem (only if not needed full-time)
- **Users are experimenting with new hardware and software solutions in the cloud ahead of procuring for on-premises deployments**
  - On the supplier side, many emergent processor/accelerator companies view the cloud as their initial introduction to HPC users

Future Research Directions

A more in-depth focus on cloud storage options

- **Identify, measure, and forecast capacity and spending**
  - User spending
  - CSP consumption
- **Sector, vertical and segment differentiation**
- **Parallel file system adoption for HPC cloud computing**
  - Requirements for cloud high performance parallel file systems (price, performance, scale)
  - Sentiment and opportunity for current leading high performance parallel file systems
- **Cloud workload usage and requirements**
  - Establish which workloads/applications are best suited to drive cloud storage spending
  - Use cases – temporal/durable; file/block/object

Conclusions

HPC in the cloud continues to evolve, as well as augment the broader HPC market

- **HPC users continue to increase their cloud usage and cloud spend, resulting in an aggressive growth**
  - Users will increase their usage as they work to understand more completely which workloads make most sense to run in the cloud
  - Cloud is anticipated to be a critical component in future system designs for HPC, as well as overall resource capabilities for HPC applications
- **Many barriers to increased HPC cloud usage have remained consistent over a few years, despite CSPs and users, both, working to solve some of the continuing concerns of HPC cloud usage**
- **Given recent announcements, the weather sector is one of a few verticals to watch, not only for the potential of their increased usage, but also the added proof of performance and capabilities for harder HPC applications**

Hyperion Research

HPC-Driven AI Market Trends

June 2021

www.HyperionResearch.com
www.hpcuserforum.com

Steve Conway

AI is Still Near the Start

Today: Special (Weak) AI

- Many observations but few choices
- “One trick dogs”: 10 AI solutions in a box to solve 10 problems
- Rudimentary training/inferencing
- Short on real-world data
- Examples:
  - Image & voice recognition
  - Early automated driving
  - Reading an MRI

Future: General (Strong) AI

- Many observations, many choices
- Versatile decision-makers capable of serious experiential learning
- More intelligent training/inferencing
- High-volume synthetic data
- Examples:
  - Discerning human motivation
  - Mature automated driving
  - Diagnosing/”curing” a cancer

About AI

- **AI is about machines guessing (inferencing) much faster than humans**
  - Human intuition is far better at this but far slower
  - For the foreseeable future, machines will mainly carry out the tedious AI work and hand off the challenging work to humans
- **AI ethics and liability activities center around the position of humans vs. machines and the HMI**
- **Current issues (transparency, biased input) could slow but not stop AI momentum**
  - Our studies show almost all HPC sites are involved in AI

HPC and AI

- **HPC is crucial at the forefront of AI R&D**
  - HPC shows where the mainstream AI market is headed.
  - HPC market growth + AI potential is motivating vendors
- **HPC innovations heavily influence mainstream AI:**
  - Algorithmic sophistication
  - Parallelization
  - Clustered servers (“clusters”)
  - CPU-accelerator processing
  - Ultrafast system data rates
  - Capable memory subsystems
- **AI is exiting the peak of the hype cycle**
  - Vendors are less often setting unrealistic expectations
- **HPC data center & enterprise deployments are different**
  - HPC data center: monolithic, standalone upgrade
  - Enterprise data center: integrate into existing infrastructure and workflow

Addressing Obstacles to Progress Toward General AI

Obstacles

- Training data
  - Inadequate volume, bias
  - Dimensionality reduction
- Explainability (XAI)
  - DL trust issue worsening
  - Inferencing is explainable
- Very task-specific
- Learning models may ignore most “intelligence”

Potential Solutions

- Synthetic data
  - Potentially unlimited
  - Multiple efforts under way
- HPC community focusing on XAI
  - Boost inferencing ability
- Multimodal AI
- Neuro-symbolic AI to augment ML/DL

Important Commercial Use Cases

Most will take longer to mature than previously thought

Precision Medicine

Automated Driving Systems

Fraud and Anomaly Detection

Affinity Marketing

Business Intelligence

Cyber Security

IoT/Edge/Smart Cities

Edge Computing

A Newer Paradigm for Highly Distributed Computing

- **A relatively new approach where some or all of the necessary computation is done directly at or near data sources**
  - Vehicles and traffic sensors
  - Medical devices
  - Product manufacturing lines
  - Military sites
  - Other data-generating locations
- **Contrasts with norm of sending Big Data to data centers or cloud computing platforms**

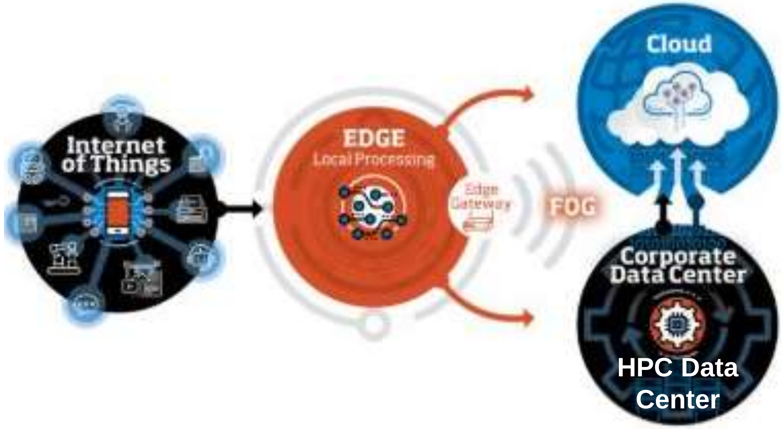
Benefits of Computing at the Edge

- **Faster responses to local issues (lower latency)**
  - Identifying lawbreakers in time for apprehension
  - Enabling vehicle-vehicle communication (e.g., truck fleets)
  - Monitoring patients in hospitals
  - Operating smart homes and factories
  - Reacting quickly to changing battlefield conditions
  - Contrasts with norm of sending Big Data to distant data centers or cloud computing platforms.
- **Lower costs**
  - Vs. moving and storing Big Data in distant facilities
- **Higher autonomy, reliability and (potentially) security**
  - No need to share resources (co-tenancy)
  - No network switches to fail and cause disruptions
  - Less data to transfer over vulnerable networks (but edge devices themselves need to be secure)
- **Scalability**
  - Adding low-cost edge devices can efficiently handle growth in source data

HPC’s Role in Edge Computing

When Wide-Area Situation Awareness/Control Are Needed

- **Examples:**
  - Smart city functions (traffic congestion, smart power grids)
  - Battlefield operations



- **Generally, only a small subset of data needs to be moved for deeper analysis from the edge to HPC systems in data centers or clouds**
  - A subset from millions of devices can still be large
  - “Fog computing” usually refers to small clusters close to the edge

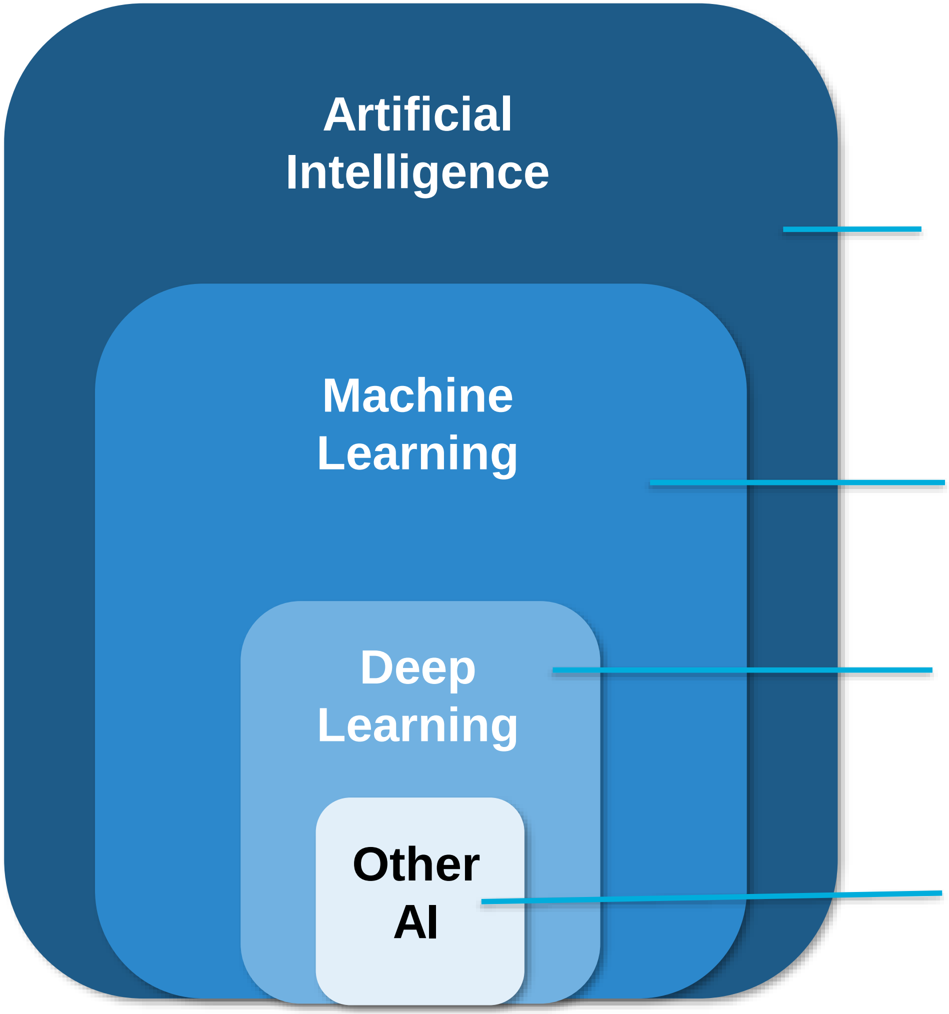
Highlights of the Continued Growth of HPDA/AI in HPC

June 2021

Alex Norton

www.HyperionResearch.com
www.hpcuserforum.com

Hyperion Research Definitions
AI: Machine Learning, Deep Learning



Artificial Intelligence (AI): a broad, general term for the ability of computers to do things human thinking does (but NOT to think in the same way humans think). AI includes machine learning, deep learning and other methodologies.

Machine Learning (ML): a process where examples are used to train computers to recognize specified patterns, such as human blue eyes or numerical patterns indicating fraud. The computers are unable to learn beyond their training and human oversight is needed in the recognition process. The computer follows the base rules given to it.

Deep Learning (DL): an advanced form of machine learning that uses digital neural networks to enable a computer to go beyond its training and learn on its own, without additional explicit programming or human oversight. The computer develops its own rules.

Other AI: AI methodologies not included in ML and DL. The main methodology in this segment today is graph analysis.

HPDA/AI Forecast

Dedicated AI servers growing more than 4X faster than overall on-prem servers

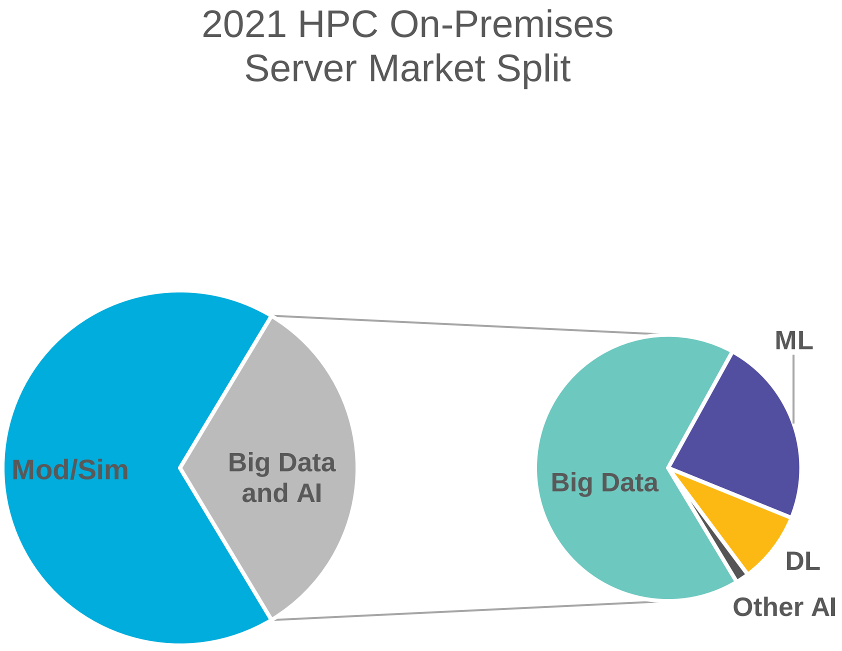
Forecast: Worldwide HPC-Based AI Revenues vs Total HPDA Revenues ($M)	2019	2020	2021	2022	2023	2024	CAGR '19-'24
WW HPC Server Revenue Forecast as of 05.02.2021	$13,595	$13,744	$13,741	$16,197	$17,708	$18,977	6.9%
WW HPDA Server Revenues [Includes Big Data and AI]	$3,598	$3,499	$4,500	$5,467	$6,650	$7,800	16.7%
WW HPC-Based AI (ML, DL & Other)	$918	$1,039	$1,500	$2,010	$2,745	$3,800	32.9%

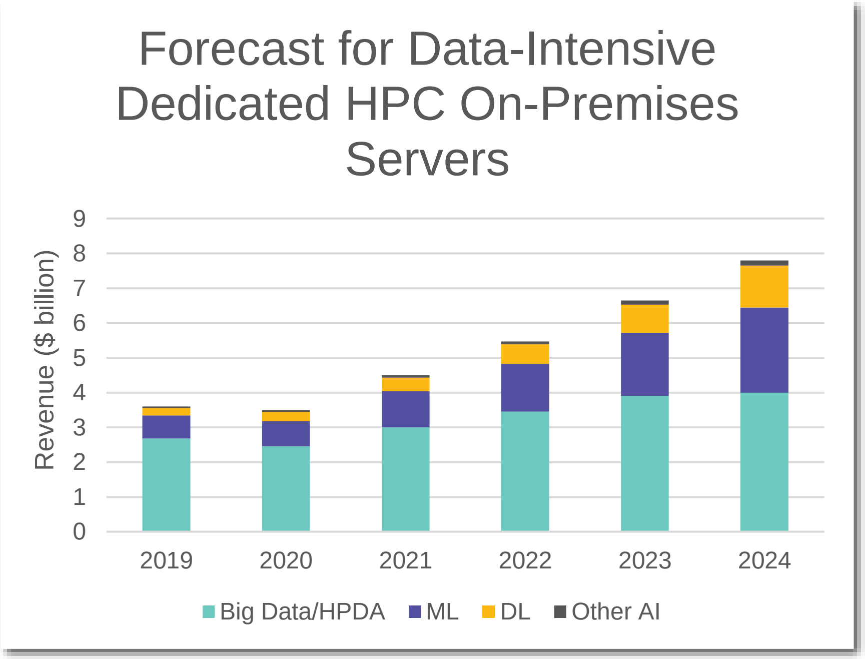
Forecast: Worldwide ML, DL & Other AI HPC-Based Revenues ($M)	2019	2020	2021	2022	2023	2024	CAGR '19-'24
ML in HPC	$667	$719	$1,039	$1,366	$1,816	$2,445	29.7%
DL in HPC	$209	$263	$390	$560	$804	$1,200	41.8%
Other AI in HPC	$42	$57	$71	$84	$125	$155	29.8%
Total AI Server Revenue	$918	$1,039	$1,500	$2,010	$2,745	$3,800	32.9%

- **Dedicated Deep Learning systems are growing at a faster rate than ML-dedicated systems**
  - ML-dedicated systems, however, make up the bulk of the AI server market revenue
  - Issues with the transparency in DL models had initially slowed DL growth, but as researchers work to solve the transparency issue, DL will exhibit higher growth

How AI Factors into Server Forecast

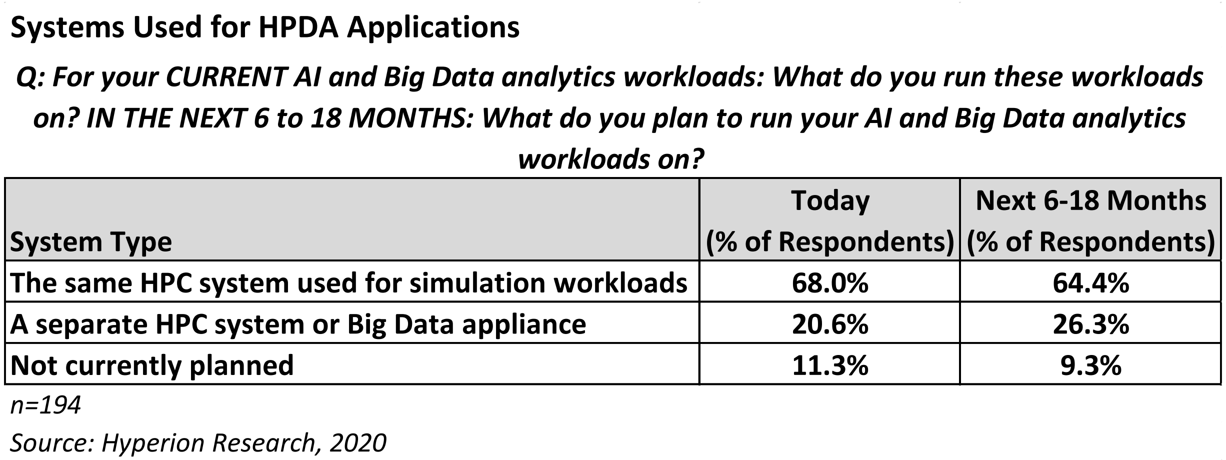
Data-intensive workloads, including HPDA and AI, comprise 1/3 of all on-prem HPC server revenue for 2021





Future HPC System Design

Users are building heterogeneous systems to handle HPDA/AI workloads



Systems Used for HPDA Applications

Q: For your CURRENT AI and Big Data analytics workloads: What do you run these workloads on? IN THE NEXT 6 to 18 MONTHS: What do you plan to run your AI and Big Data analytics workloads on?

System Type	Today (% of Respondents)	Next 6-18 Months (% of Respondents)
The same HPC system used for simulation workloads	68.0%	64.4%
A separate HPC system or Big Data appliance	20.6%	26.3%
Not currently planned	11.3%	9.3%

n=194

Source: Hyperion Research, 2020

- **As workloads become more diverse, system designs have shifted to incorporate technologies for varied workloads**
  - The need for different processor technology, different memory and storage solutions, and interconnects have shifted HPC system designs
  - Some users are procuring separate systems dedicated to HPDA/AI applications

A Variety of AI-Specific Hardware

New processors and accelerators emerging to handle AI workloads

- **AI workloads require different hardware capabilities than traditional HPC**
- **GPUs are currently the accelerator market leaders for AI workloads, but…**
  - Diverse AI workloads are driving new, varied hardware requirements
  - Emerging technologies from both startups and large vendors are targeted at specific classes of AI applications
  - Many new processors are diverging from the current standard architectures
- **While it is still early in this new wave of processors, optimism is high, and funding is abundant for alternative processor options**

Public Cloud and AI

Recent studies show higher utilization of public cloud for HPC-enabled AI

- **Based on data collected in late 2019, users anticipate running about 20% of their HPC-enabled AI workloads in public clouds**
  - Growth is expected to continue over the next few years
  - AI workloads are expected to exhibit high growth in the overall HPC ecosystem over the next few years as well
- **HPC users looking to the cloud for AI due to:**
  - Access to diverse hardware and software solutions
  - Access to public data sets
  - CSPs' AI expertise
  - Ability to aggregate data in cloud for storage

AI Concerns

Current issues with AI concern explainability, transparency, and reproducibility

- **In talking with AI experts, there are a few key concerns that arise around the future of AI**
  - The transparency and explainability of trained models is crucial in understanding the process by which a model came to a decision
  - The reproducibility to the trained model is critical in verifying the output of a model
  - Working on these two issues will not only increase the accuracy of the models, but will also build public trust in real world applications
  - Understanding and mitigating data bias is valuable for producing applications that have wider usage opportunities
- **AI experts agree that data quality and data diversity are inadequate in many cases**

Environmental Concerns

AI training models consume massive amounts of energy

- **There are environmental concerns about AI as well, given the required energy for the compute power can be very high**
  - According to an article in Nature from 2020, the cost of training large NLP models can produce a similar CO2 emission as 125 round-trip flights between New York and Beijing (Source: https://www.nature.com/articles/s42256-020-0219-9?proof=t)
  - One possible solution is the power-efficiency of newer processor technologies in development
- **Members of the AI community are starting to talk about the carbon footprint of training models as a measure akin to the time to complete training**

Future Research Directions

Topics of interest for the next year of research

- **System design**
  - Processor and accelerator decisions
  - Interconnect schemes for heterogeneous system design
  - Storage solutions to handle the ever-growing size of data sets
  - Memory considerations for AI applications
- **Application considerations**
  - Changes in data privacy and sharing laws
  - Explainability, transparency, and reproducibility concerns
- **Continued intersection of AI and HPC**
  - Use of AI methodologies to augment modelling and simulation applications
  - Use of HPC simulations to generate synthetic data for training
  - Application of HPC techniques to enhance AI capabilities
  - Growing adoption of HPC-enabled AI techniques by traditional IT enterprises

HPC Storage Market Update

June 2021

Mark Nossokoff

www.HyperionResearch.com
www.hpcuserforum.com

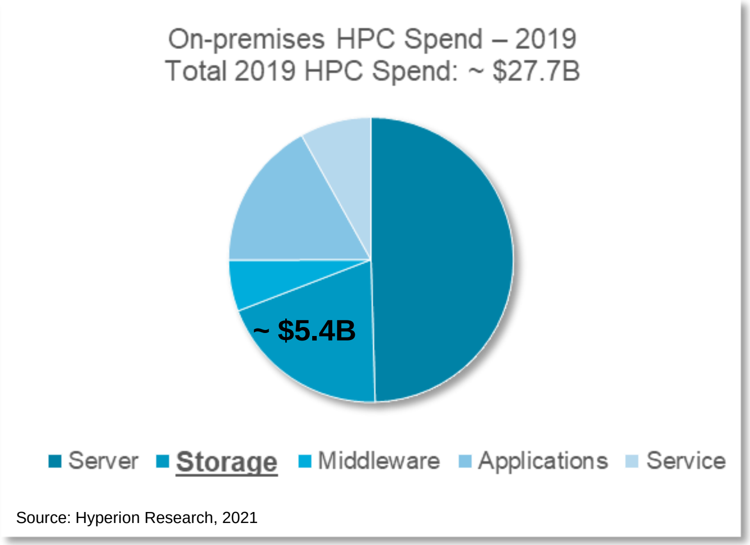
What’s happened in HPC storage since our SC20 update?

*“Gradually, then suddenly…”**

- **Brief look at the numbers**
- **Global deals and plans**
- **DoE storage highlights**
- **Business models**
- **Cloud storage**

* *The Sun Also Rises*, Ernest Hemmingway

HPC Storage is an Attractive Market



Source: Hyperion Research, 2021

- **Storage historically the highest growth HPC element**
- **Storage represents ~ 20% of HPC spending**
- **For every $1 spent on compute, ~ $0.40 is spent on storage**

Area ($M)	2019	2020	2021	2022	2023	2024	CAGR '19-'24
Server	$13,595	$13,744	$13,741	$16,197	$17,708	$18,977	6.9%
Add-on Storage	**$5,379**	**$5,520**	**$5,605**	**$6,675**	**$7,478**	**$8,075**	**8.5%**
Middleware	$1,599	$1,618	$1,640	$1,946	$2,142	$2,310	7.6%
Applications	$4,647	$4,682	$4,643	$5,380	$5,783	$6,092	5.6%
Service	$2,218	$2,186	$2,131	$2,421	$2,552	$2,636	3.5%
Total Revenue	$27,438	$27,750	$27,761	$32,619	$35,662	$38,090	6.8%

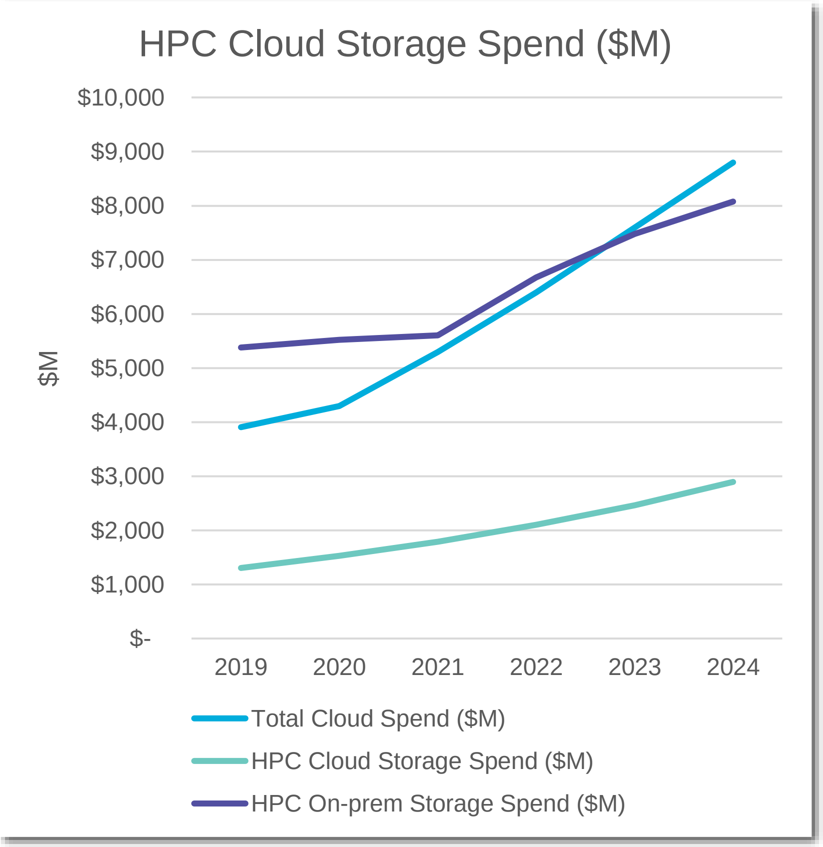
HPC Cloud Storage Forecast

Cloud Storage '19-'24 CAGR 2x on-prem storage CAGR

- **$3.9B total cloud spend in 2019**
- **Total cloud spend to surpass on-prem storage spend in 2024**
- **Storage represented 1/3 of total 2019 cloud spend**
- **Cloud storage CAGR double on-prem storage CAGR**
  - Cloud: 17.3%
  - On-prem: 8.5%

($M)	2019	2020	2021	2022	2023	2024	'19 -'24 CAGR
Total Cloud Spend	$3,910	$4,300	$5,300	$6,400	$7,600	$8,800	17.6%
HPC Cloud Storage Spend	$1,303	$1,529	$1,793	$2,104	$2,467	$2,894	17.3%
HPC On-prem Storage Spend	$5,379	$5,520	$5,605	$6,675	$7,478	$8,075	8.5%

Source: Hyperion Research, 2021



Source: Hyperion Research, 2021

Global Deals and Plans

- **“Lumi” from EuroHPC JU in Finland starting in 2021**
  - 7 PB all flash
  - 80 PB scratch
  - 30 PB object
- **Singapore NSCC in early 2022**
  - ASPIRE 1 successor
  - 20 PB @ 300GB/s read/write performance
- **“Alps” @ CSCS in 2023**
  - Successor to Top500 #12 (Nov 2020), “Piz Daint” (> 8 PB)
  - New architecture
- **Meteo France: 1.3 EB by 2025**
  - Numerical modeling and climate predictions
  - Ingest, render up to 2 and 1.3 PB per day, respectively
- **Exascale in South Korea by 2030**
  - Current 5th generation (“Nurion”) is #21 on Top500 (Nov 2020)
  - 6th generation by 2023
  - 7th generation by 2028

DoE Storage Highlights

- **NERSC "Perlmutter"**
  - Single-tier all-flash 35PB Lustre
  - Optimizations for bandwidth and throughput
- **LLNL "El Capitan"**
  - Rabbit storage subsystem
  - Flux resource manager
- **OLCF "Frontier"**
  - Orion file system (open-source Lustre and ZFS technologies)
    - 40 Lustre MSS; 450 Lustre OSS
    - 11.5 PBs NVMe flash performance tier @ 10 TBps peak read/write; random read > 2M IOPs
    - 679 PB HDD capacity tier (47,700 HDD spindles) @ Peak Read/Write 5.5/4.6 TBps and > 2M peak random read IOPs
    - 10 PB NVMe metadata flash tier with 480 NVMe devices
  - In-system storage layer
    - Peak Read/Write 75/35 TBps, respectively
    - 15B random Read/Write IOPs

Business Models

- **CentOS stewardship evolution**
  - Red Hat re-direction resources to upstream RHEL
  - Rocky Linux, AlmaLinux potential heir apparent
- **Big Memory market**
  - Samsung exiting NVMe Optane to focus on CXL
  - Emerging technologies
  - In-memory and near-memory computing
- **Uptake in vendor-offered consumption models**
  - HPE Greenlake
  - VAST SW-only
- **File system intrigue**
  - Lustre – open-source and custom enhancements
  - Seagate CORTX object archive file system
  - HPE File System Storage embracing Spectrum Scale

HPC Cloud Storage

- **UK Met and Microsoft**
  - Weather and climate simulations produce massive amounts of data
  - Active data archive system expected to support ~ 4 EBs of storage, query, and retrieval
- **RIKEN/Oracle collaboration**
  - Leveraging cloud more for storage applications
  - Not about compute and capabilities

Look for us…

- **Coming to a city near you as the world opens back up**
  - United States (e.g., Austin, Bay area, D.C. area, Houston, New York, Seattle)
  - International
  - "…we'll be at [suggest a location] if you'd like to talk…"
- **Virtual HPC User Forum in September**
- **Meet us in St. Louis – SC21**

HYPERION RESEARCH

Quantum Computing: Finding Its Place in the Advanced Computing Sector

June 2021

Bob Sorensen

www.HyperionResearch.com

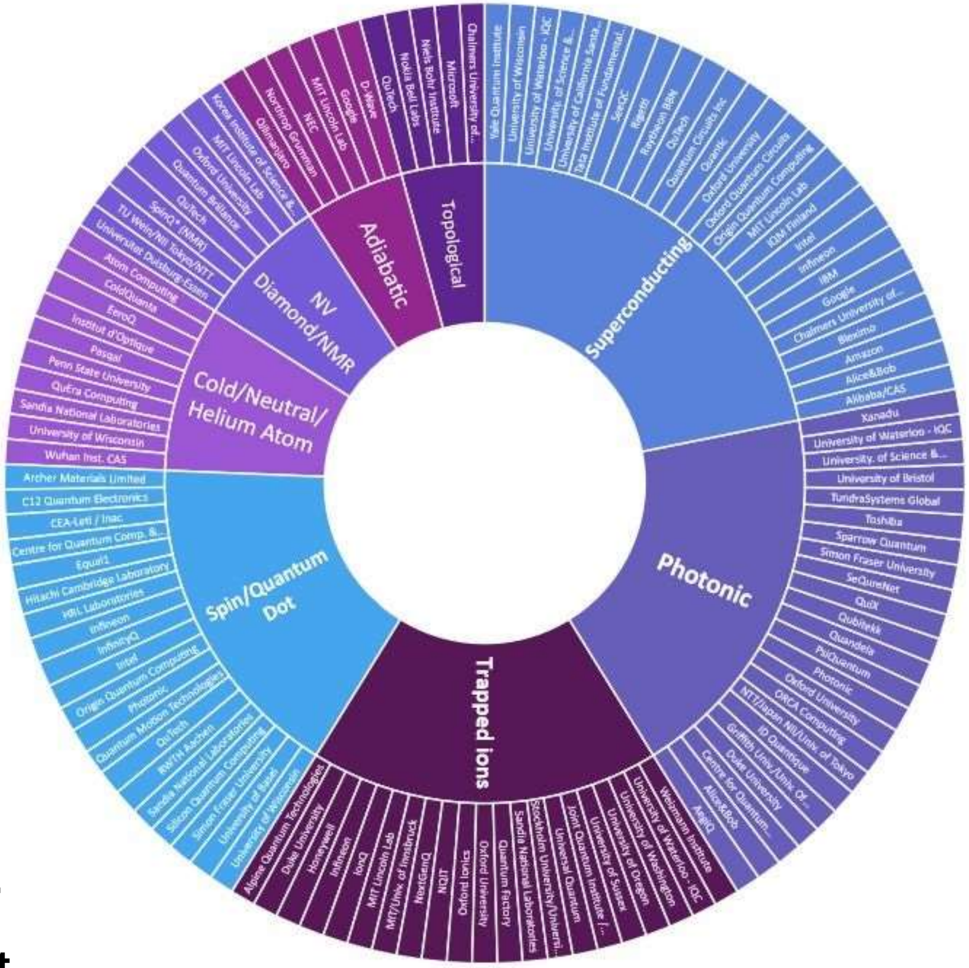
Currently, the Promise of QC is Substantial

- **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
    - Materials science
    - Chemistry
    - Pharmaceuticals
    - Oil and gas
  - Machine learning
  - Optimization
- **And the list grows longer each day**

- **Goal: Demonstrate so-called <u>quantum supremacy</u> using a programmable quantum device to solve a problem that no classical computer can solve in any feasible amount of time**
  - Shor's Algorithm: Factoring a large number into its two prime integers
  - This goal may not be the same for everyone

Broad Range of Contenders

- **There are several competing quantum modalities currently under development**
  - Superconducting
  - Photonic
  - Trapped Ions
  - Spin/Quantum Dot
  - Cold/Neutral/Helium Atom
  - NV/Diamond/NMR
  - Adiabatic
  - Topological
- **Each offers their own unique strengths and weaknesses**
- **There may not be a clear winner**
- **And the ultimate winner may not be here**



Source : Michel Kurek https://www.linkedin.com/in/michelkurek/

Substantial Challenges Ahead

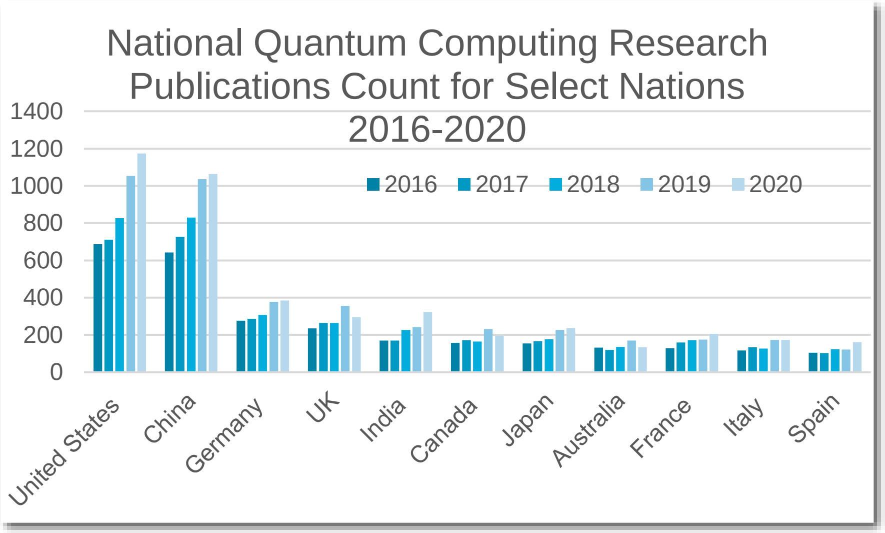
Formidable technical issues in QC hardware and software

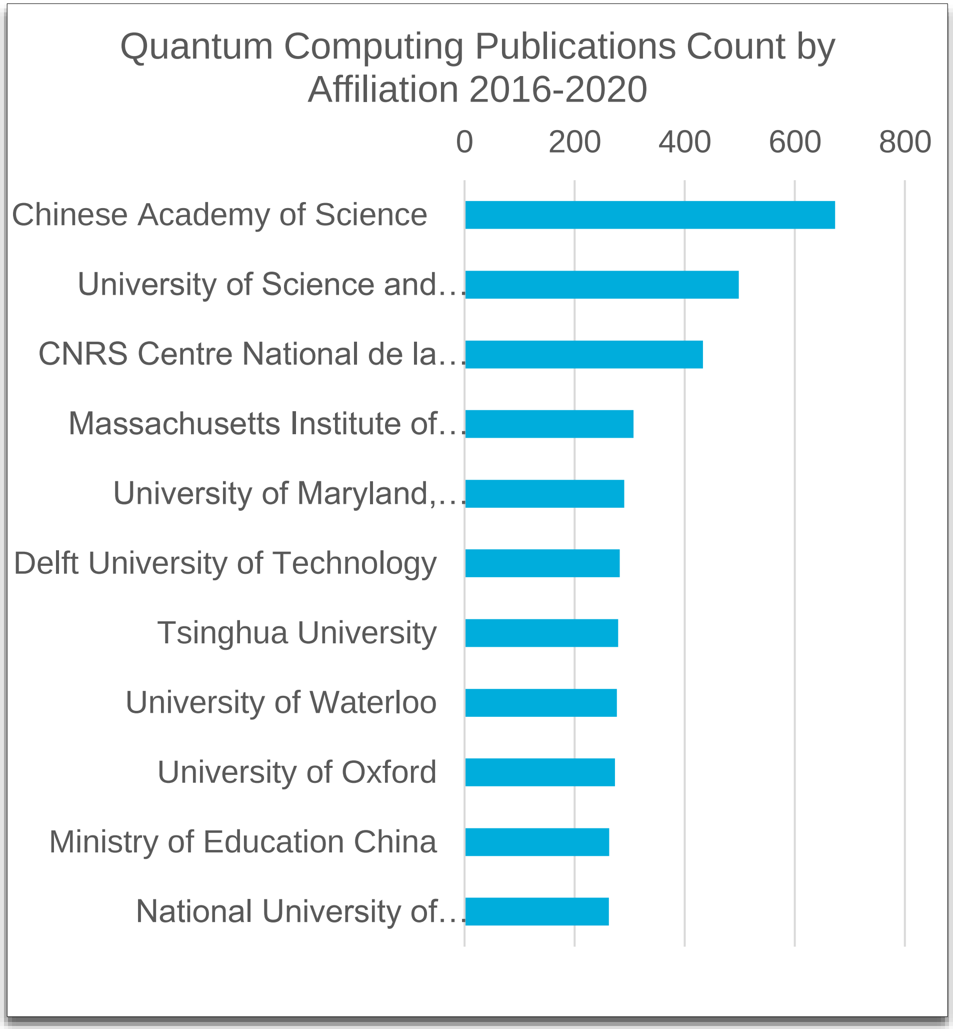
- Uncertain performance gains
- Unclear time frames
- Disorganized progress in algorithm/application development
- Looming workforce issues

All these factors complicates treating QC as a stable market alongside more traditional IT sectors

- Making a business case is tough...but it needs to be done
  - Use cases, revenue gains, customer acquisition

Comparisons of National-Level Quantum Computing Research Publications 2016-2020





- Summary of leading R&D publications between 2016 and 2020, using Scopus with the following query to search a publication's title, abstract, and keywords
- Query: "quantum comput*" OR "qubit" OR "quantum simulat*“ where * represents wildcard letters
- The searches, conducted on May 25, 2021, yielded a total data set of 17,534 documents

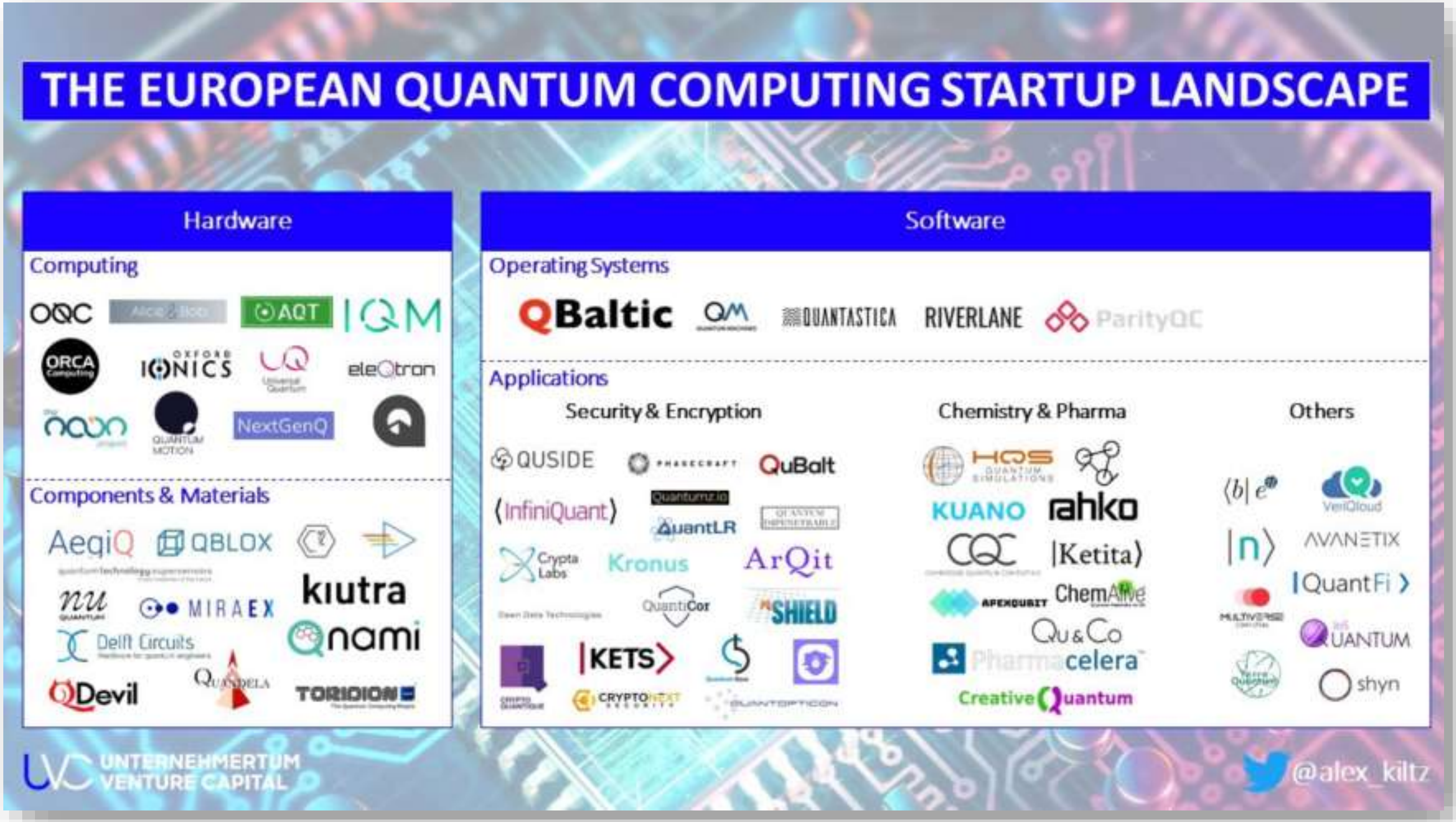
Select National/Regional Government Quantum Programs

- 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

A Growing Collection of Quantum Computing Hardware Suppliers…

- **A wide and diverse range of QC hardware suppliers have emerged to populate a growing QC ecosystem**
  - Legacy players (Fujitsu, IBM, Atos)
  - Integrated player (Honeywell)
  - New entrants:
    - Pure play: IonQ, Rigetti, ColdQuanta, Quantum Circuits Inc, Xanadu, IQM, etc.
    - Component players: Intel
  - Non-traditional players (Alibaba, AWS, Baidu, Google, Microsoft)
  - Myriad stealth players

European Commercial Presence is Substantial



THE EUROPEAN QUANTUM COMPUTING STARTUP LANDSCAPE

Hardware

Computing

OQC Alice & Bob AQT IQM
ORCA Computing OXFORD IONICS UQ Universal Quantum eleQtron
the nAon project QUANTUM MOTION NextGenQ

Components & Materials

AegiQ quantum technology ingenerators QBLOX
nu QUANTUM MIRAEX kiutra
Delft Circuits nami
QDevil QUANDELA TORIDION

Software

Operating Systems

QBaltic QM QUANTASTICA RIVERLANE ParityQC

Applications

Security & Encryption

QUSIDE PHASECRAFT QuBalt
InfiniQuant Quantumz.io QuantLR QUANTUM IMPENETRABLE
Crypta Labs Kronus ArQit
Deen Data Technologies QuantiCor SHIELD
KETS CRYPTO QUANTIQUE CRYPTONEXT SECURITY QUANTOPTICON

Chemistry & Pharma

HQS QUANTUM SIMULATIONS
KUANO rahko
CQC |Ketita⟩
APEXQUBIT ChemAlive
Qu & Co
Pharmacelera
Creative Quantum

Others

⟨b| e VeriQloud
|n⟩ AVANETIX
|QuantFi⟩
MULTIVERSE COMPUTING
QUANTUM
Terra Quantum shyn

UNTERNEHMERTUM VENTURE CAPITAL

@alex_kiltz

© Hyperion Research 2021

…and QC Software Suppliers Abound

	Amazon	Atos	D-Wave	Google/ Cirq	Honey- well	IBM/ Qiskit	Microsoft	Rigetti
1QBit	X		X		X	X	X	X
Accenture					X	X		
A*Quantum						X		
Agnostiq						X		
AIQTech						X		
Aliro Technologies					X	X		
Apply Science						X		
Beit					X	X		
Boxcat						X		
blueqat (formerly MDR)	X		X			X		
Cambridge Quantum Computing				X	X	X	X	
ColdQuanta						X		
Entropica Labs					X	X	X	X
equal1.labs						X		
GTN							X	
Horizon Quantum Computing						X		X
HQS Quantum Simulations				X			X	X
Jij							X	
JoS Quantum						X		
Max Kelsen						X		
Miraex						X		
Multiverse	X					X	X	
Netramark						X		
Nordic Quantum Computing Group (NQCG)						X		
Opacity						X		
OTI Lumionics			X				X	X

	Amazon	Atos	D-Wave	Google/ Cirq	Honey- well	IBM/ Qiskit	Microsoft	Rigetti
Phasecraft						X		
ProteinQure			X			X	X	X
Q-CTRL						X		X
QC Ware	X		X	X	X	X	X	X
Qedma						X		
QSimulate	X						X	
Qu & Co						X	X	
Quantastica								X
QuantFi						X		
Quantum Benchmark				X		X		
Quantum Machines						X		
Quantum-South					X			
Qubit Engineering							X	
QunaSys	X				X	X	X	
Rahko	X				X	X	X	
Rigetti	X							
Riverlane Research							X	X
softwareQ						X		
Solid State AI						X	X	
Strangeworks					X	X	X	X
Super.tech						X		
Xanadu	X					X	X	
Zapata Computing	X	X		X	X	X	X	X
Zurich Instruments						X		

Source: Quantum Computing Report, Updated April 21, 2021

QC Market Forecast Summary

- The global QC market was worth about $320 million (+/- $30 million) in 2020
- Based on an anticipated CAGR of 27% between 2020 and 2024, the global QC market will grow from approximately $320 million in 2020 to $830 million in 2024
- On-prem and cloud access QC hardware will comprise about 50% of the global QC market for the next three years
- Optimization, physical simulation, and machine learning will near equally divide the algorithm space
- NISQ will be the near-term architecture of choice, followed by quantum annealers and digital simulators
- User access to QC will be primarily through the cloud, at three times the rate of an on-premise option

QC Buyer/User Performance Considerations

QC Performance Expectations: Modest Gains Prevail

What is the minimum application performance gain you would require to justify using quantum computing for your existing and planned workloads?

	# of Responses	Percent
2-5X	9	7.8%
5-10X	14	12.2%
10-50X	26	22.6%
50-100X	30	26.1%
100-250X	11	9.6%
250-500X	3	2.6%
Greater than 500X	4	3.5%
Application or required performance not possible on classical systems	18	15.7%
n = 115		
Source: Hyperion Research, 2020		

- **Expectations for the minimum QC performance gains for both existing and planned workloads were relatively modest**
  - 78% of respondents would see a performance boost of less than 250X as justification for using QC
  - 42% would only need 50X or below
  - 20% would need less than 10X
- **A 50X performance improvement translates into a roughly 4-5-year lead over counterpart classical computing performance**
- **Only 18 respondents (16%) would require true quantum supremacy or quantum-only applications to justify using a QC for their existing or planned workloads**

How This Plays Out Near-Term

- **Near-term QC Ramp Up**
  - Many exploring applications/use cases and not just for traditional HPC but for enterprise IT computing environments
  - QC/SME application development could be next crucial milestone
  - That the market will be growing – at least for the next few years - is demonstrated
- **Quantum computing is not a replacement for classical computing, but a companion technology**
- **The sector is not at the Moore’s Law stage**
  - Development is happening in many dimensions and in parallel:
    - Hardware (qubit, QC-Lan, architecture)
    - Software (middleware, applications, use cases)
    - Algorithms
    - Hybrid classical/quantum systems
    - Quantum inspired hardware and software

ISV and Open-Source Application Growth

June 2021

Melissa Riddle

www.HyperionResearch.com
www.hpcuserforum.com

ISV and Open-Source Applications are Growing

Expanding workloads are driving application use

- **ISV software license spending is growing along with HPC workloads**
  - Average number of systems per site is growing to meet demands for differentiated HPC resources
  - Widespread cloud growth is also increasing average workload size
- **Open-source application use is growing in HPC, especially with the growth of cloud and GPUs**
  - ISV licenses can be challenging with the cloud
  - Many ISVs have not ported their code to run on accelerators yet

Top ISV Applications by Vertical

Updated list from existing database

- **Respondents were asked to list their top 3 applications only**
  - This is not an exhaustive list of all applications used
- **Large breadth of ISV applications in Government Lab, Academic/University, and Other verticals**
  - Government Lab
    - NAMD, VASP, Ansys, GAMESS, STAR-P
  - Academic/University
    - MATLAB, VASP, Gaussian, NAMD, Ansys
  - Other
    - SAP, Spark, GAMESS, Ansys, Gaussian, VASP

Top ISV Applications by Vertical (cont'd)

Updated list from existing database

- **Notably, some ISVs were represented across multiple verticals**
  - Bio-Sciences
    - Blast, MATLAB, NAMD, Schrodinger
  - CAE
    - ABAQUS, LS-DYNA, CONVERGE, SAP, Ansys
  - Chemical Engineering
    - LS-DYNA, MATLAB, RADIOSS
  - EDA
    - MATLAB, Ansys, COMSOL, Hadoop, SAP

Top ISV Applications by Vertical (cont'd)

Updated list from existing database

- **Verticals such as DCC, Economics/Financial, and Weather had fewer ISV applications, instead favoring in-house and open-source applications**
  - DCC & Distribution
    - Spark
  - Economics/Financial
    - Hadoop, Spark, IRIS
  - Weather
    - COSMO, FLUENT

ISV Software Across Verticals

Certain ISV applications prominent in multiple verticals

- **Ansys (FEA)**
  - Academic, CAE, EDA/IT/ISV, Government Lab
- **FLUENT (CFD)**
  - Academic, CAE, EDA/IT/ISV, Weather
- **GAMESS (Chemistry)**
  - Academic, Government Lab, Other
- **MATLAB (Calculator)**
  - Academic, Bio-Sciences, CAE, Chemical Engineering, EDA/IT/ISV
- **NAMD (Molecular dynamics)**
  - Academic, Bio-Sciences, Government Lab
- **Spark (Big Data)**
  - DCC, Economics/Financial, EDA/IT/ISV, Gov Lab
- **VASP (Quantum mechanics)**
  - Academic, Government Lab, Other

ISV Software in HPC Workloads

Most HPC users rely on ISVs for part of their workload

- **Most MCS respondents (77%) use ISV software for at least part of their HPC workload**
- **On average, an HPC workload consists of:**
  - 22% ISV software applications
  - 39% in-house applications
  - 39% open-source (free) applications
- **Over time, ISV software has been decreasing as percent of workload while open-source and in-house applications have been increasing**

Motivators for ISV Application Spending Growth

Expanding HPC resources are increasing ISV licenses

- **ISV spending is expected to increase 3.9% in 2021**
- **Demand for more differentiated resources (e.g., accelerators or specialized AI processors) is driving up average number of HPC systems per site**
- **HPC sites are using the cloud to address expanding workloads**
  - Many HPC users value continuity between their on-prem and cloud applications but licenses in the cloud (when available) come with additional premiums
- **Many HPC users perceive ISV software as producing faster, more accurate results when compared to open-source applications**

Open-Source Application Popularity

Open-source applications have wide appeal across many verticals

- Nearly all HPC users (92%) reported using open-source software at least some of the time
- In all verticals, some respondents report one of the top 3 applications at their site is open-source
- Academic, Bio-Sciences, Economics/Financial, EDA/IT/ISV, and Weather more likely to report one of their top 3 applications is open-source
- ISVs tend to get more expensive as applications are scaled up, so more users are supplementing with open-source to offset these costs
- 19% of HPC users report their cloud use is limited by the availability or cost of ISV codes in the cloud
- Open-source is integral to HPC sites worldwide

Open-Source Application Popularity (cont'd)

Open-source applications have wide appeal across many verticals

- **Many of the top open-source applications originated from publicly funded universities and are now supported by a larger community**
- **Some open-source applications/codes are cited among the top 3 applications per site across multiple verticals**
  - Gromacs: Academic, Bio, EDA/IT/ISV, Gov, Weather
  - LAMMPS: Academic, CAE, EDA/IT/ISV, Government
  - TensorFlow: Academic, Bio, DCC, EDA/IT/ISV, Other
  - OpenFOAM: Academic, CAE, EDA/IT/ISV
  - Quantum ESPRESSO: Academic, EDA, Gov, Other
  - WRF: Academic, EDA/IT/ISV, Weather

Importance of Programming Languages

Programming languages rank as top HPC applications

- **7% of users consider a programing language to be one of their top 3 HPC applications**
- **MATLAB was most popular as a top 3 application, followed by R and Python**
- **Programming languages were more likely to be ranked in a site's top 3 within certain verticals**
  - Academic, Bio-Sciences, Econ/Financial, EDA/IT/ISV
- **Most sites use multiple parallel programming languages/libraries/APIs**
  - Most common are Python (78% of sites), C/C++ (74%), MPI (61%), OpenMP (50%), Fortran (46%), CUDA (45%), and R (32%)

Conclusion

ISV and open-source software are both integral parts of most HPC sites

- **Patterns of increasing cloud use and differentiated hardware are driving growth in both paid ISV software licenses and open-source software**
  - New hardware and cloud are both challenging for HPC sites
  - When possible, HPC sites often choose to offset the challenges of cloud or unfamiliar hardware by using an application they are already familiar with on-prem
  - When ISV software itself is the challenge, HPC users adapt to open-source software

A Quick Global Tour of Exascale Systems

June 2021

Bob Sorensen

www.HyperionResearch.com
www.hpcuserforum.com

A Whirlwind of Exascale Development

- US: DoE-centric development
- Japan: Fugaku for the long-term
- China: Three diverse rollouts soon -- ISC Next Week?
- EU and individual European nations: A plethora of machines
- UK: Going it Alone
- Concerns with Future HPC Development

Near-term US Exascale Plans

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

- **Aurora: DOE Office of Science, Argonne National Laboratory**
  - Intel Prime/HPE/Cray Sub
  - Delivery in late 2022, acceptance in 2023 (12 month-ish late)
  - Cray Shasta architecture with Intel Xeons and Intel Xe GPU

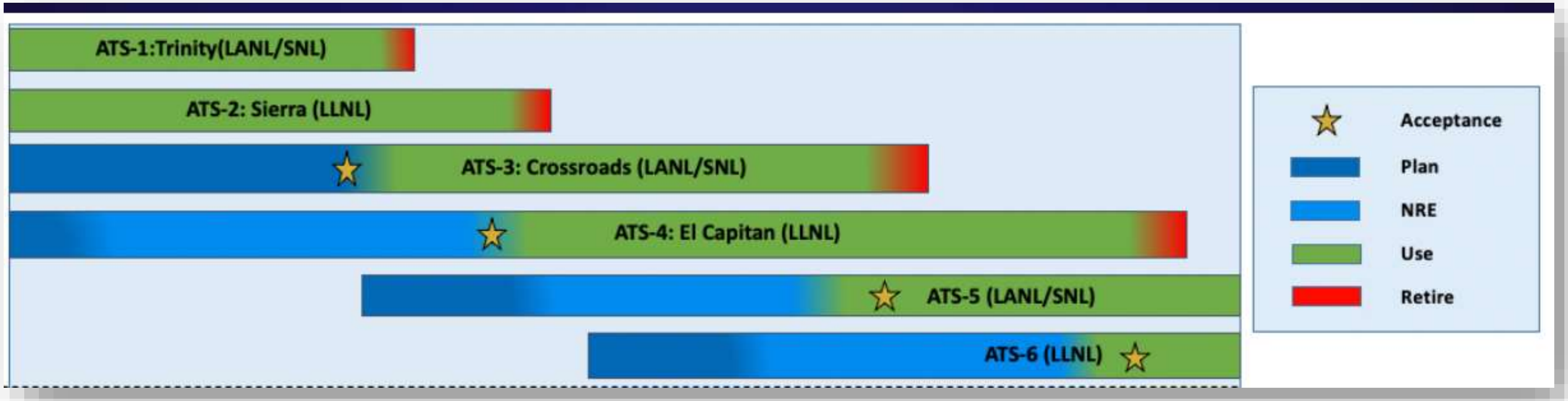
- **Frontier: DOE Office of Science: Oak Ridge National Laboratory**
  - HPE/Cray Prime
  - Delivery in late 2021 and acceptance in 2022
  - Cray Shasta with AMD EPYC CPU and future Radeon GPUs

- **El Capitan: DOE NNSA’s LLNL**
  - HPE/Cray Prime
  - Delivery in late 2022, with full production targeted for late 2023
  - Cray Shasta architecture AMD EPYC processors, next generation Radeon Instinct GPUs

Mid-Size US Exascale Plan

Crossroads on the Horizon: LANL and SNL

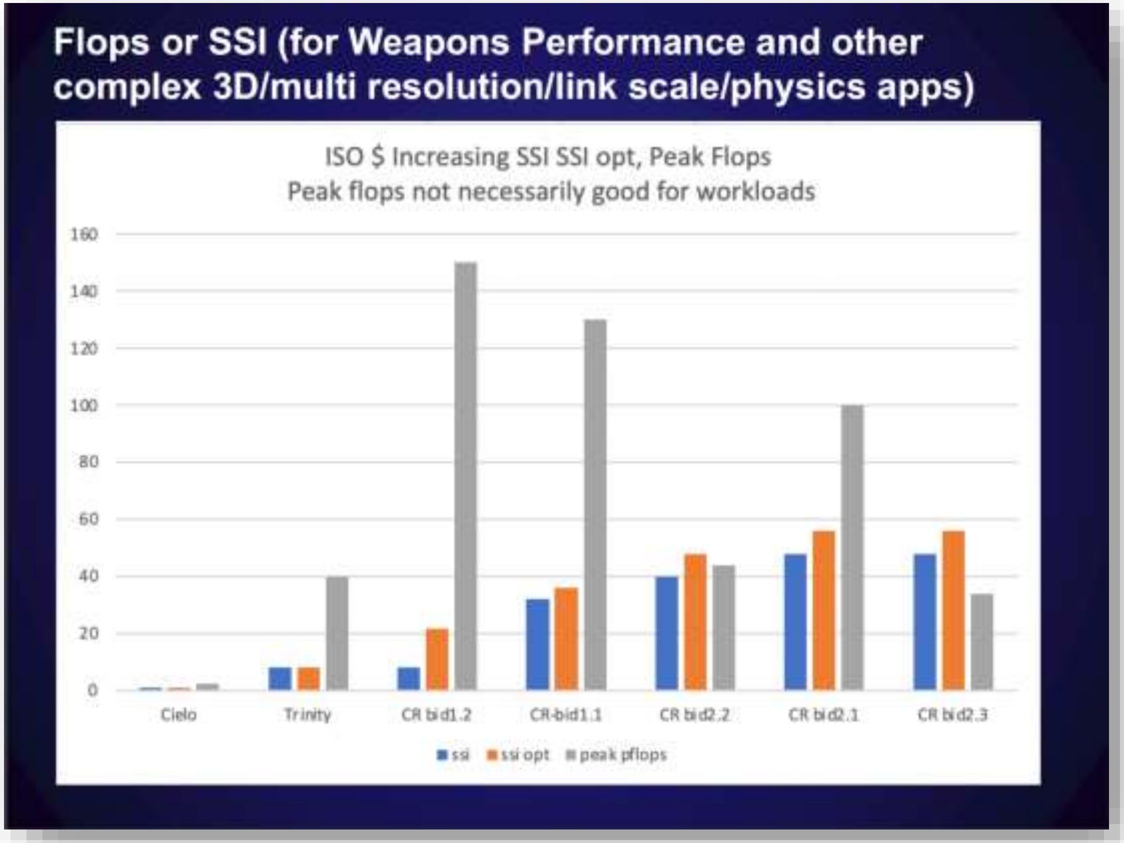
- **Procurements by the Alliance for Computing at Extreme Scale (ACES) partnership between Los Alamos National Laboratory and Sandia National Laboratories**
- **$105 million contract awarded to HPE to deliver Crossroads, a next-generation HPE Cray EX supercomputer to be sited at Los Alamos for 2022 operation**
- **Future Intel Xeon processor Sapphire Rapids with next-generation Intel Deep Learning Boost (with Advanced Matrix Extensions)**



Source: Gary Grider LANL, DoE, LA-UR-21-22315

Mid-Size US Exascale Plan (cont.)

Procurement Insights: May the “Best” Machine Win



Flops or SSI (for Weapons Performance and other complex 3D/multi resolution/link scale/physics apps)

Source: Gary Grider LANL, DoE, LA-UR-21-22315

- **No Peak Performance (in Flops) requirement specified in RFP**
- **Specific workload speed up specifications increasingly common**
  - Based on SSI mini apps representing workload and workflow information about apps
- **For Crossroads, lowest TPP system bid performed best on representative workloads**

The Cycle Continues

Looking Ahead Starts Now

US Rep. Jay Obernolte, R-Calif., recently introduced *The Next Generation Computing Research and Development Act*

- Targeted to create the Beyond Exascale Computing Program for developing systems with capabilities that exceed those of the fastest supercomputers in the U.S.
- Tasked with maintaining foundational research programs:
  - Mathematical
  - Computational science
  - Computer sciences
- Focusing on new and emerging computing:
  - Post-Moore’s law computing architectures
  - Novel approaches to modeling and simulation
  - Artificial intelligence and scientific machine learning
  - Quantum computing
  - Extreme heterogeneity

Japan’s Exascale System

Riken’s Fugaku #1 on June 2020 Top 500 list

- **High Performance Linpack (HPL) result of 415.5 petaflops**
- **Uses Fujitsu A64 ARMv8.2-A processor**
  - 48/52 compute cores with GPU-like vector extensions
  - 4x 8 GB HBM with 1024 GB/s, on-die Tofu-D network BW (~400 Gbps)
  - High SVE FLOP/s (3.072 TFLOP/s)
- **No GPUs**
- **158,976 single socket nodes**
- **Tofu-D bandwidth 10X total global CSP traffic**
- **Peak DP > 400 Pflops, Peak SP > 800 Pflops, Peak HP > 1600 Pflops**
- **Typically, 37X faster that predecessor K system on target co-design applications**
- ***Eight Year Lifetime***

China Exascale Plans

One or More May Make the June 2021 Top 500 List?

Three prototypes under development – one or more prototypes may be selected for full-up production

- **NUDT (Tianhe)**
  - Indigenous CPU, possibly Arm-based Phytium Xiaomi or (less likely) Fujitsu A64FX
  - MT-2000+ NUDT accelerator (or follow-on)
  - 400 Gbps homegrown network
- **Sugon**
  - Heterogeneous architecture (2 CPU/2 DCU accelerators per node)
  - Hygon processor is licensed clone of AMD Gen 1 EPYC processor
  - Hygon-developed accelerator
  - Six-dimensional torus network for ~10,000 nodes
  - Board-level liquid immersion cooling
- **Sunway (prototype specifications)**
  - CPU-only SW26010 chip follow-on (260 cores @ 3Tflops per chip)

EU HPC Plans

EuroHPC program stood up in 2018

- **Chartered to develop EU-wide HPC development program**
  - 33 participating States + EU
    - (Malta recently joined)
  - Operational duration: November 2018-2026
- **Three sites recently selected for 150-200 Pflops systems**
  - Kajaani Finland, Barcelona Spain, and Bologna Italy
  - Total Investment: 650 million Euros
    - 50% EU
    - 50% Consortium
- **Five sites selected for medium range HPCs (~4Pflops)**
  - Investment ~180 million Euros
- **Systems are owned by EuroHPC Joint Undertaking**
- **Installations started 4Q2020**

EU HPC Plans (continued)

Exascale Plans Going Forward

- **EU Plan calls for acquisition of two exascale systems in the 2022-2023 timeframe**
  - At least one to use European technology: specifically using an EPI-developed processor
  - EU plans may include 2 ES systems in 2023-2026
- **Additional procurements in Germany, France, others in 2024, 2025**
- **Post Exascale System around 2027**
  - Plans call for integration and deployment of the first hybrid HPC/quantum infrastructure in Europe

UK Plans

UK Looking at 2024 for First Post-Brexit Exascale System

- The UK, which will not likely be eligible to fully take part in EuroHPC projects or access calls when Horizon 2020 ends, has plans for a domestic exascale system
- Exascale project requirements include support for both traditional modeling and simulation as well as AI/Deep Learning
- System targeted for both scientific community and industrial users
  - Typical European emphasis on industrial competitiveness
- Exascale rollout schedule
  - Procurement during 2022
  - Assembly and installation 2023
  - Final changes to hosting environment 2023
  - Planned service opening April 2024
- 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

Risks in HPC

Factors that Could Endanger the Future Growth of HPC Performance

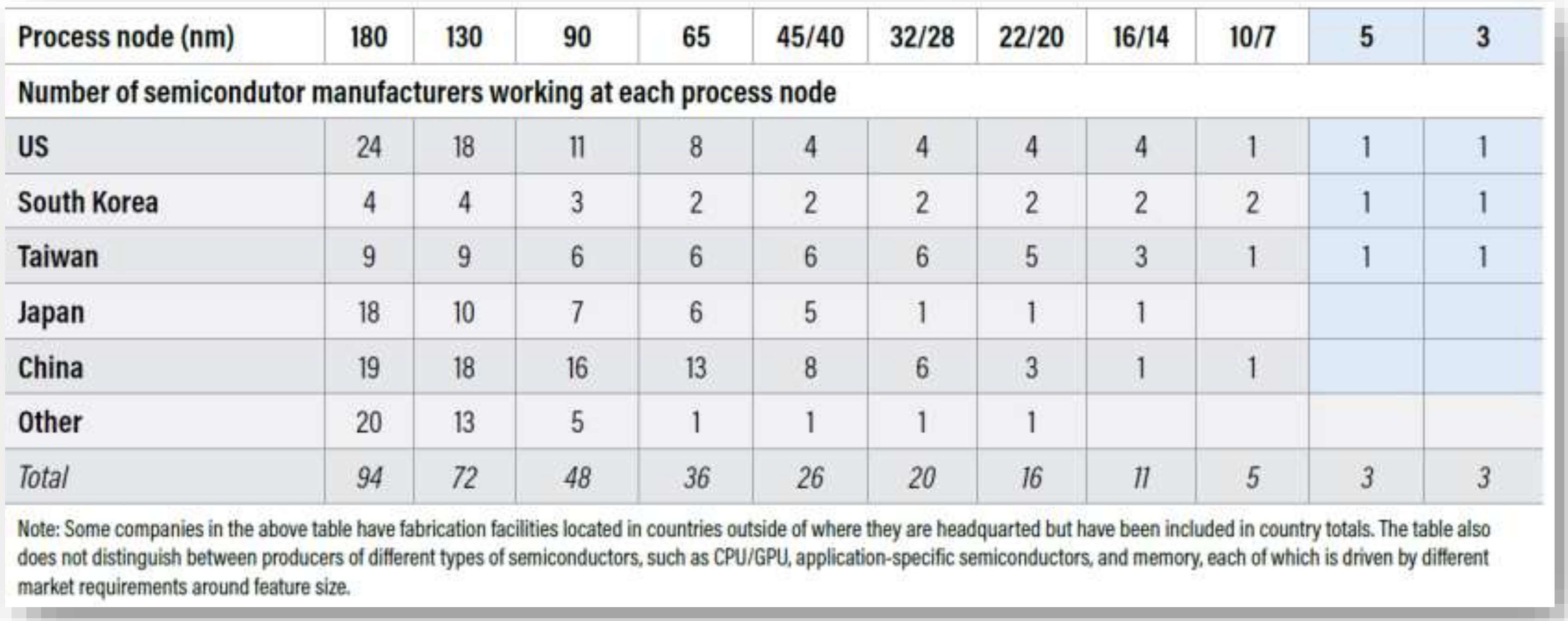
- **Near-term Concerns (now to 5 years out)**
  - Rising cost - and applicability - of high-end system development
    - Driving new procurement models at the high end?
  - Country/regional emphasis on indigenous HPC food chain development
    - Red, blue, purple etc. supply chains
    - Rising invasive technology policy
    - Reduced international collaboration
    - Programmability concerns
  - No next big thing
    - (GPU/AI now, quantum computing not exactly next)
  - CSPs: A double-edged sword

- **Some other specifics:**

Risks in HPC

Shrinking Number of Semiconductor Makers

- **And then there were three: TSMC (Taiwan), Samsung (S Korea) and Intel (US)**



Process node (nm)	180	130	90	65	45/40	32/28	22/20	16/14	10/7	5	3
Number of semicondutor manufacturers working at each process node											
US	24	18	11	8	4	4	4	4	1	1	1
South Korea	4	4	3	2	2	2	2	2	2	1	1
Taiwan	9	9	6	6	6	6	5	3	1	1	1
Japan	18	10	7	6	5	1	1	1			
China	19	18	16	13	8	6	3	1	1		
Other	20	13	5	1	1	1	1				
Total	*94*	*72*	*48*	*36*	*26*	*20*	*16*	*11*	*5*	*3*	*3*

Note: Some companies in the above table have fabrication facilities located in countries outside of where they are headquarted but have been included in country totals. The table also does not distinguish between producers of different types of semiconductors, such as CPU/GPU, application-specific semiconductors, and memory, each of which is driven by different market requirements around feature size.

Source: The Geopolitics of Semiconductors, Prepared by the Eurasia Group September 2020

N.B. Apple's M1 chip locks up significant portion of TSMC's 5-nm production capacity

Risks in HPC

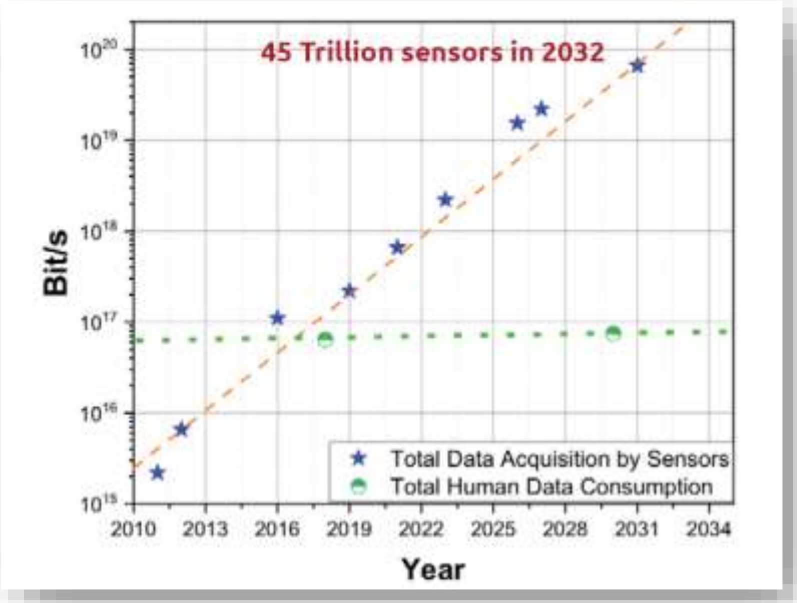
Narrowing of Advanced Lithographic Options

- **Samsung and TSMC require EUV systems for their 5nm lines**
- **There is only one supplier of EUV steppers: ASML of the Netherlands**
  - An EUV machine is made of more than 100,000 parts and costs approximately $120 million
  - Several dozen exist, with approximately two-year back-order
  - In 2023, ASML plans for high-NA EUV capability for 3 (or 2) nm process node
  - That may be the limit of NA-EUV capability
  - Little work on next generation technology
- **Alternate immersion systems are 7nm capable**
  - ASML and Nikon of Japan are the only two immersion vendors currently

Risks in HPC

Data Deluge from IoT, Edge, and Sensors

- Fundamental breakthroughs in analog hardware are required to generate smarter world-machine interfaces to facilitate storage, processing, and analysis



- Sensor technologies are experiencing exponential growth with forecasts of ~45 trillion sensors in 2032
  - Generating >1 million zettabytes (10^{27} bytes) of data per year
  - This is equivalent to ~10^{20} bit/s
- Realistic synthetic data is needed to feed DL's data appetite
- Mitigating factor: only a small subset of edge/IoT data needs HPC in a cloud or data center

In Summary

The ROI From HPC Is High

The average ROI is $507 for revenue, and on average $47 for profits/cost savings

Updated results continue to indicate substantial returns for investments in HPC:

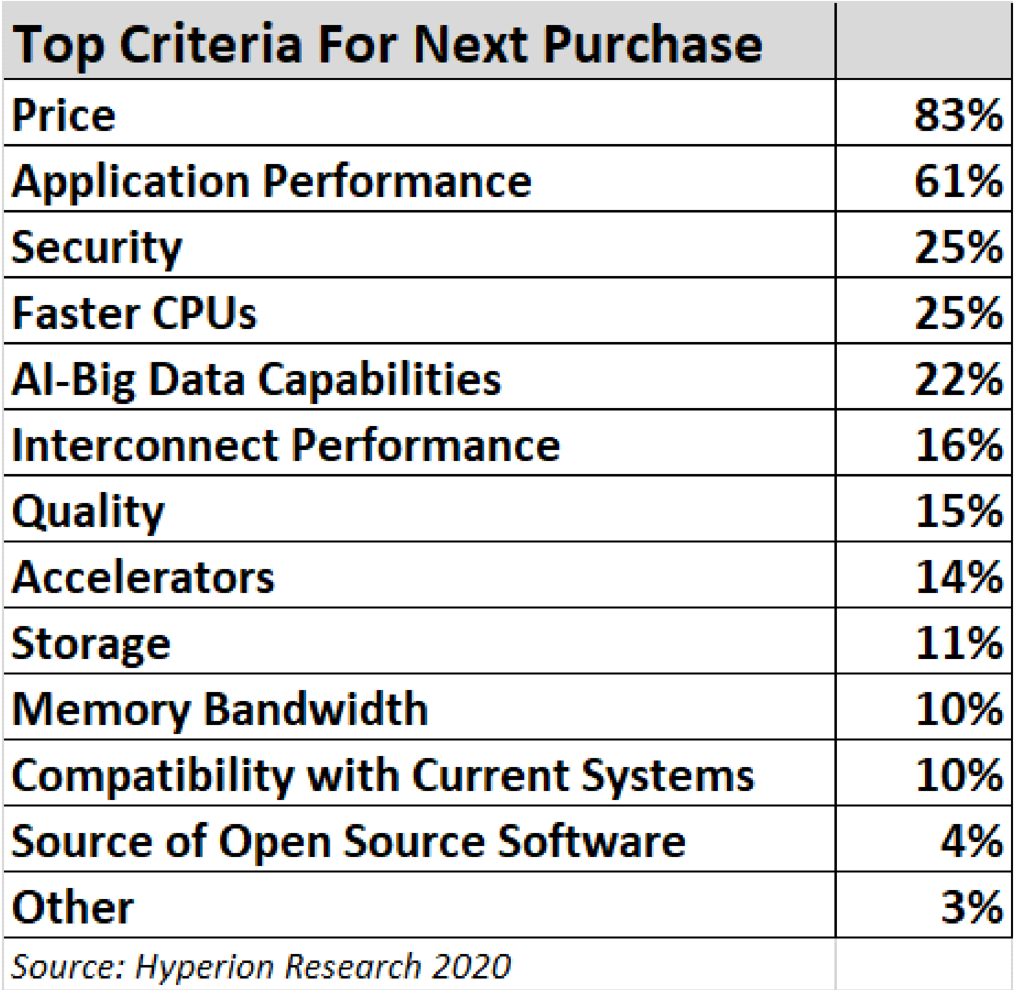
- The data now covers 763 successful HPC projects
- On average <u>$507 dollars in revenue per dollar of HPC invested</u> was generated (excluding outliers)
- On average <u>$47 dollars of profit (or cost savings) per dollar of HPC invested</u> was generated (excluding outliers)
- The average HPC investment per innovation was $2.6 million

Note that this research is looking at the economic impacts based on the HPC investment compared with the output of revenue/sales and/or profits and cost savings. It excludes the additional costs of production, sales etc. that are also required for each project.

The full data and results of this research are available at: www.hpcuserforum.com/ROI/

Key Buying Requirements For HPC

#1 = price/performance (for running their specific applications) and performance on their specific applications

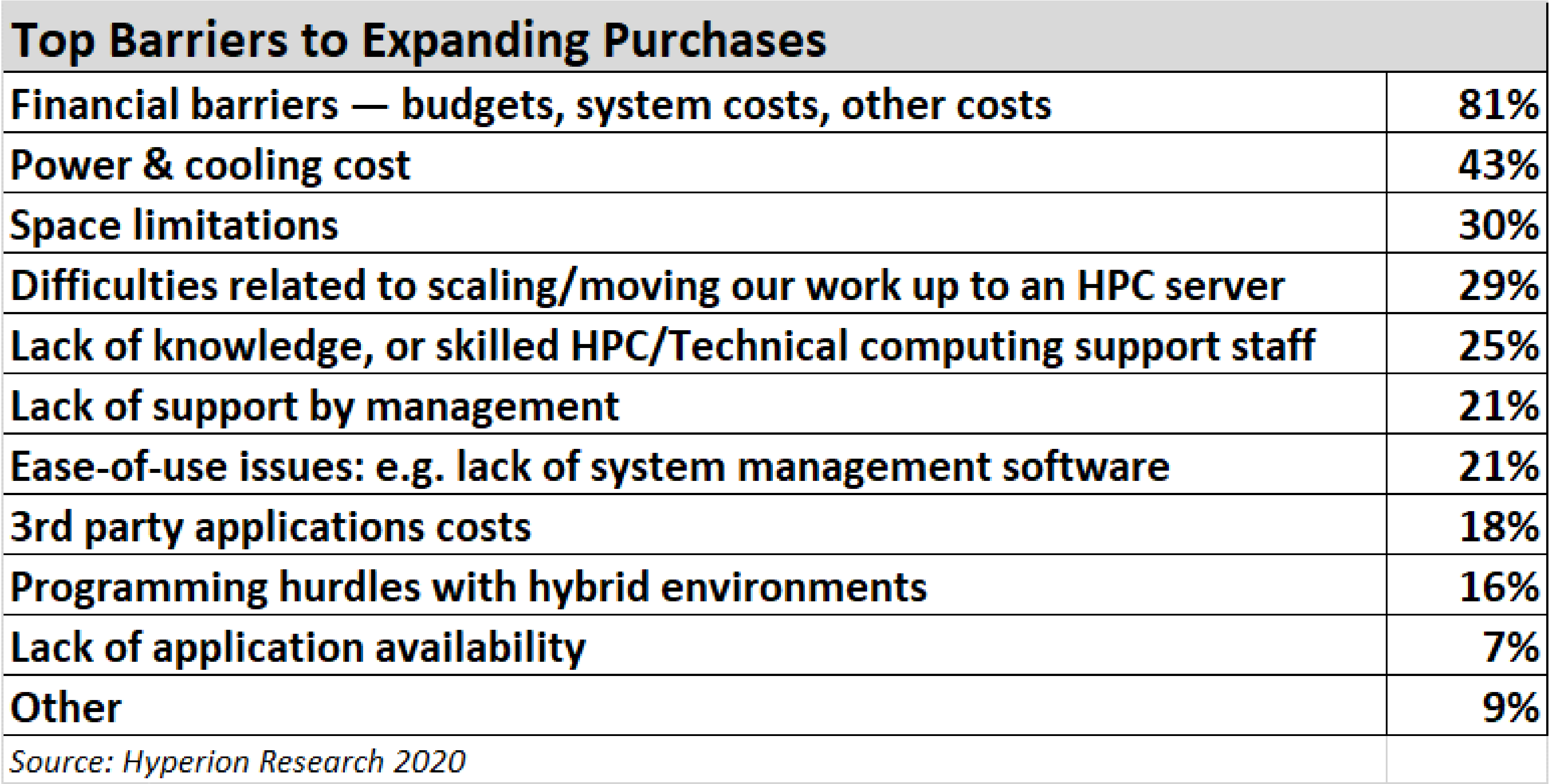


Top Criteria For Next Purchase	
Price	83%
Application Performance	61%
Security	25%
Faster CPUs	25%
AI-Big Data Capabilities	22%
Interconnect Performance	16%
Quality	15%
Accelerators	14%
Storage	11%
Memory Bandwidth	10%
Compatibility with Current Systems	10%
Source of Open Source Software	4%
Other	3%

Source: Hyperion Research 2020

Barriers For Buying More On-prem

#1 = budgets



Top Barriers to Expanding Purchases	
Financial barriers — budgets, system costs, other costs	81%
Power & cooling cost	43%
Space limitations	30%
Difficulties related to scaling/moving our work up to an HPC server	29%
Lack of knowledge, or skilled HPC/Technical computing support staff	25%
Lack of support by management	21%
Ease-of-use issues: e.g. lack of system management software	21%
3rd party applications costs	18%
Programming hurdles with hybrid environments	16%
Lack of application availability	7%
Other	9%

Source: Hyperion Research 2020

Conclusions

- **The pandemic was expected to impact 2020 by ~8% decline, but Fugaku made 2020 a growth year!**
  - 2022 to 2024 are expected to be strong growth years
    - Exascale systems will drive growth in 2022 to 2024
    - AI, HPDA, big data are hot growth areas
    - HPC in the cloud will lift the sector writ large
- **New technologies are showing up in larger numbers:**
  - Processors, AI hardware & software, memories, etc.
- **The cloud has become a viable option for many HPC workloads**
- **Storage will likely see major growth driven by AI, big data and the need for much larger data sets**

QUESTIONS?

Questions or comments are welcome.

Please contact us at:

info@hyperionres.com

Please take a minute to do a short survey to help us improve our update briefings