



HYPERION RESEARCH

Global AI Update

SC22

November 2022

Alex Norton and Tom Sorensen

www.HyperionResearch.com
www.hpcuserforum.com

On-Prem HPC Market Segmentation

Server classification based on end-user application

HPC Servers

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

Data-Intensive

Compute-Intensive

- CFD
- Reservoir Modelling

Traditional Data Science

- Monte Carlo Apps
- Optimization Apps
- Pattern Recognition

HPC-Enabled AI

- Model Training
- Surrogate Models
- Graph Analysis

Machine Learning

Deep Learning

Other AI

- Graph Analysis
- Semantic Analysis

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%
								
<u>Compute-Intensive</u> Server Revenue	10,020	10,848	12,103	13,280	14,177	13,586	13,993	5.2%
<u>Data-Intensive</u> 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
 - Timesaving 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

What did we miss?



tsorensen@hyperionres.com

anorton@hyperionres.com