Highlights of Growth of AI and HPDA in HPC

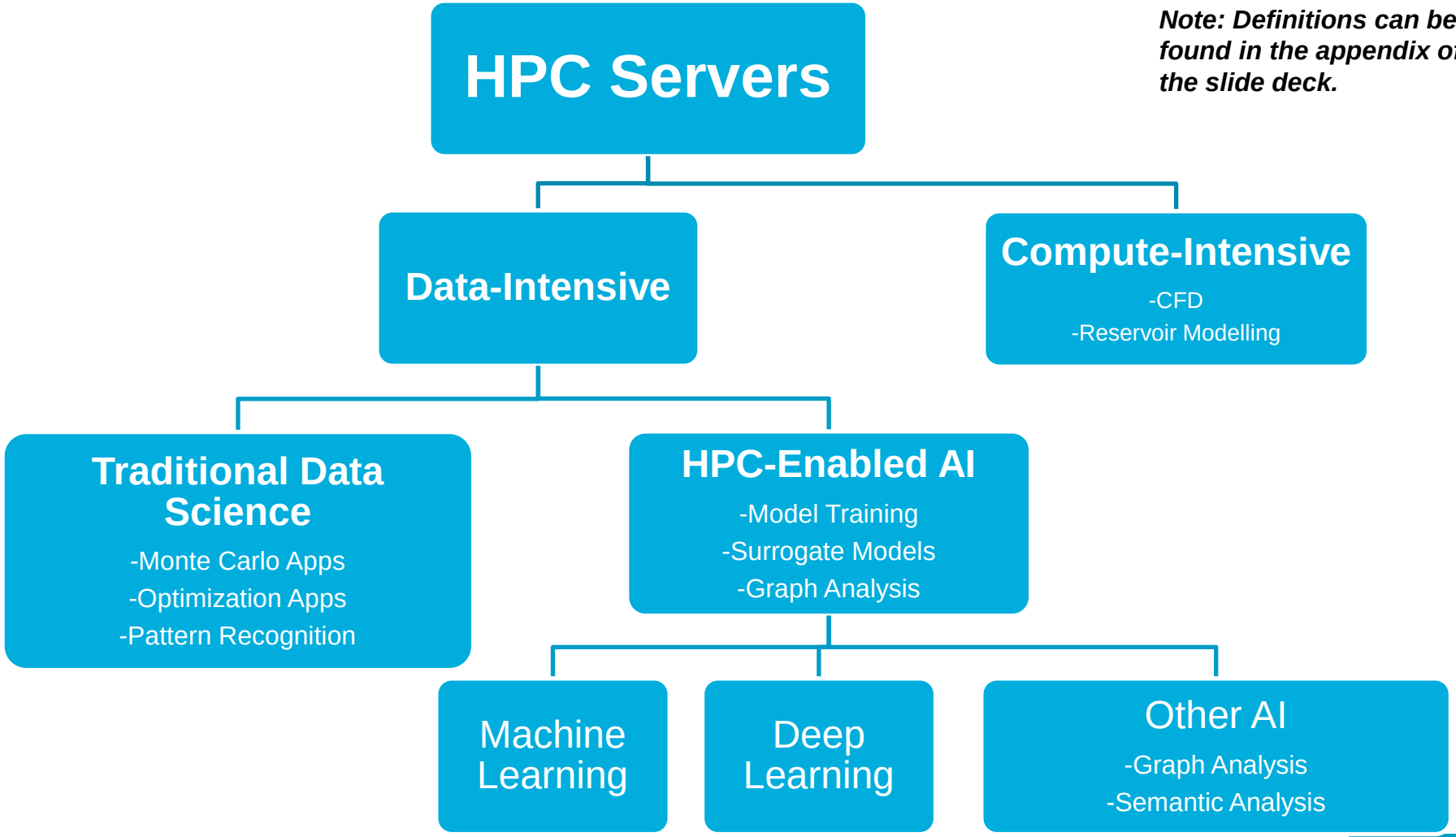
November 2021

www.HyperionResearch.com
www.hpcuserforum.com

Alex Norton

On-Prem HPC Market Segmentation

Server classification based on end-user application



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

The HPC Server Market Segmented

Data-Centric focused servers comprise ~25% today

Worldwide HPC Server Revenue Breakout by Compute-Intensive and Data-Intensive Focuses								
($M)	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**	**2025**	**CAGR '20-'25**
Compute-Centric Server Revenue	$9,771	$10,024	$10,900	$12,547	$13,637	$14,428	$13,719	6.5%
Data-Centric Server Revenue	$3,598	$3,499	$3,649	$4,400	$4,928	$5,519	$6,182	12.1%
HPC Server Revenue Forecast	$13,368	$13,523	$14,550	$16,947	$18,565	$19,947	$19,901	8.0%

- **Data-centric servers are growing nearly twice as fast as compute-centric servers, due to:**
  - A rise in data-intensive simulations
  - Aggressive growth of AI applications
  - The integration of AI techniques in traditional HPC apps

Within the Data-Centric Market

HPC-Enabled AI systems growing to ~50% of data-intensive market by 2025

($M)	2019	2020	2021	2022	2023	2024	2025	CAGR '20-'25
HPC-Enabled AI (ML, DL & Other) Server Revenue	$918	$1,039	$1,216	$1,618	$2,034	$2,429	$2,905	22.8%
Traditional Data Science (non-AI HPDA) Focused Server Revenue	$2,680	$2,460	$2,433	$2,782	$2,894	$3,091	$3,276	5.9%
Total Data-Centric HPC Server Revenue	$3,598	$3,499	$3,649	$4,400	$4,928	$5,519	$6,182	12.1%

- **HPC-enabled AI servers today account for less than $^1/_3$ of the data-intensive servers**
  - That portion is expected to grow to nearly ½ by 2025
  - AI-focused servers growing more than 3x faster than traditional data science servers

HPC-Enabled AI Market

ML-focused machines the bulk of the AI market today

($M)	2019	2020	2021	2022	2023	2024	2025	CAGR '20-'25
Machine Learning	$667	$719	$806	$1,018	$1,213	$1,368	$1,569	16.9%
Deep Learning	$209	$263	$341	$501	$692	$899	$1,133	33.9%
Other AI	$42	$57	$70	$98	$129	$162	$204	29.0%
Total HPC-Enabled AI Server Revenue	$918	$1,039	$1,216	$1,618	$2,034	$2,429	$2,905	22.8%

- **Within HPC-enabled AI servers, DL-focused servers growing aggressively**
- **Machine-learning-focused servers comprise majority of revenue today**

Future HPC System Designs

Users are building heterogeneous systems to handle AI & HPDA workloads

- **As workloads become more diverse, system designs have shifted:**
  - Some sites are building heterogeneous systems
  - Some sites are building out multiple systems to handle different workloads specifically
  - Some sites are looking to the cloud to address specific subsets of their workload portfolio
- **Technology options have diversified as well**
  - New accelerator options
  - ASICs designed for specific AI applications
  - Various memory, interconnect, and storage products
- **Compute resource allocation should be treated as an optimization problem:**
  - Find a balance among diverse technology options
  - Optimize for key workloads

Use of Surrogates in HPC

The injection of AI models to support large simulations

- **Using deep learning models to approximate simulations and speed up time to solution**
  - While these models lack the accuracy of true simulation, they allow researchers to search the solution space and complete large simulations faster
- **Surrogate models are a prime example of the convergence of HPC and AI**
- **Examples of the use of surrogates have emerged in:**
  - Earth system modelling
  - Genetics applications
  - Drug design
  - High-energy physics

For more information about surrogates: https://bit.ly/3o0L2uL

Future Research Directions

Topics of interest for the next year of research

- **New Components**
  - Emergent processors and accelerators
  - New memory technologies
  - Interconnect solutions
- **Application considerations**
  - Changes in data privacy and sharing laws
  - Data storage and sharing considerations
- **Continued intersection of AI and HPC**
  - Use of AI methodologies to augment traditional HPC apps
  - Use of large-scale simulations to generate synthetic data
  - Growing adoption of HPC-enabled AI techniques by traditional IT enterprises
- **Ethics of AI**
  - Explainability, transparency, and reproducibility concerns
  - Societal impacts of AI applications

HYPERION RESEARCH

Want to continue the conversation?

info@hyperionres.com
anorton@hyperionres.com

HYPERION RESEARCH

Appendix

Definitions

- **Compute-Centric HPC Application**: HPC applications that focus primarily on being able to scale across multiple processors for the most computation power possible. Examples of compute-centric HPC applications include crash simulation, reservoir modelling, genome sequencing, and many others. This category is usually referred to as HPC Modelling and Simulation workloads.
- **Data-Centric HPC Application**: HPC applications that rely primarily on the use of large data sets and involve high data transfer between nodes during the execution of the application. Examples of data-centric HPC applications include big data, HPDA, AI, ML, DL, and other data-driven simulation workloads. Many data intensive applications require larger memory profiles on the systems they are run on.

Definitions (cont.)

- **High Performance Data Analysis (HPDA):** refers to data-intensive computing that exploits HPC resources. HPDA includes long-standing, data-intensive modelling and simulation (M&S) methods in the HPC industry/application segments, and newer high-performance analytics methods that are used in these segments, as well as by commercial organizations that are adopting HPC for the first time. HPDA may employ either long-standing numerical modelling and simulation methods, newer methods such as for big data and large-scale graph analytics, semantic technologies, and knowledge discovery algorithms, or some combination of long-standing and newer methods.
- **Artificial Intelligence (AI):** a broad, general term for the ability of computers to do things human thinking does, but in different ways. AI includes machine learning, deep learning (a.k.a. cognitive computing) and other methodologies.

Definitions (cont.)

- **HPC-Enabled AI:** a term to highlight the specific AI applications tracked by Hyperion Research and classified in this forecast. Similar to the definition of HPC applications overall, HPC-enabled AI applications are those that are run by scientists, researchers, engineers, etc. Social media-type AI applications, mainly simple image recognition or tagging, are not included in the forecast. Hyperion Research focuses on the cutting edge, scientific or highly computational AI applications.
- **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.
- **Deep Learning (DL):** an advanced form of machine learning that often uses digital neural networks to enable a computer to go beyond its training and learn on its own, without explicit programming or human oversight.