BMKT-102-1: High Performance Data Lakes and Lakehouses for AI

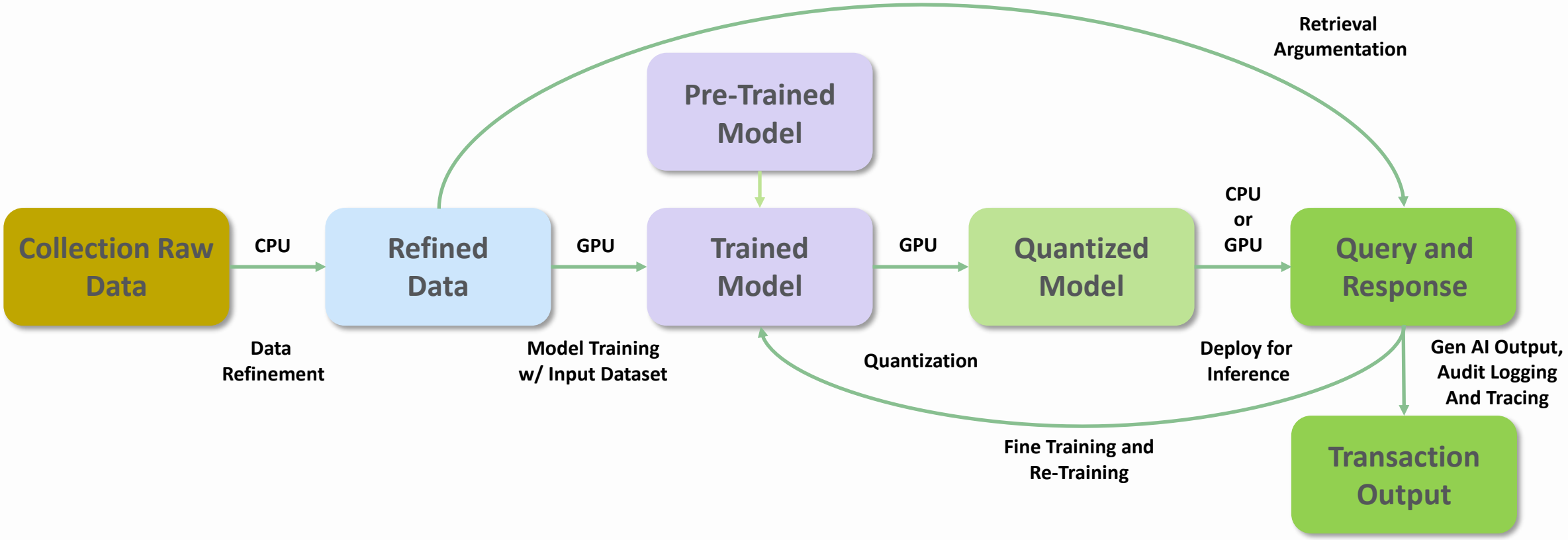
Wendell Wenjen

Accelerate Everything © 2025 Supermicro

1. AI Data Pipeline Overview
2. Hybrid and High-Performance Data Lakes
3. Data Lakehouses
4. Implementation

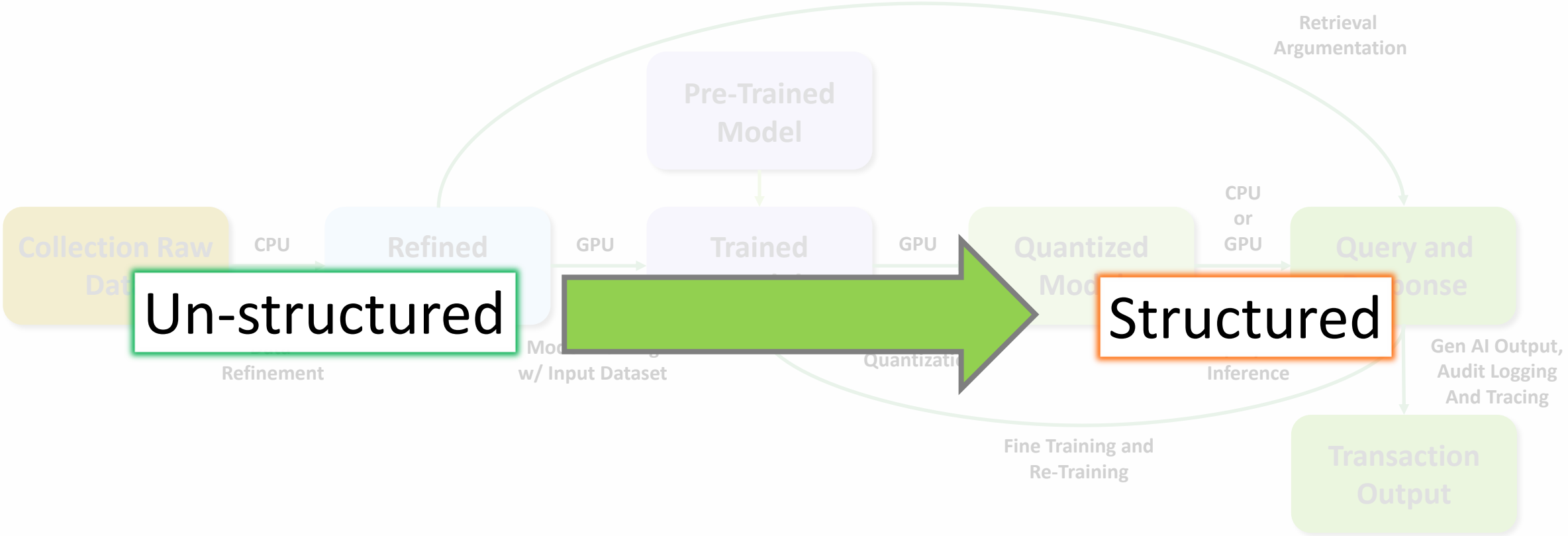
Better Faster Greener™ © 2024 Supermicro

AI Data Pipelines



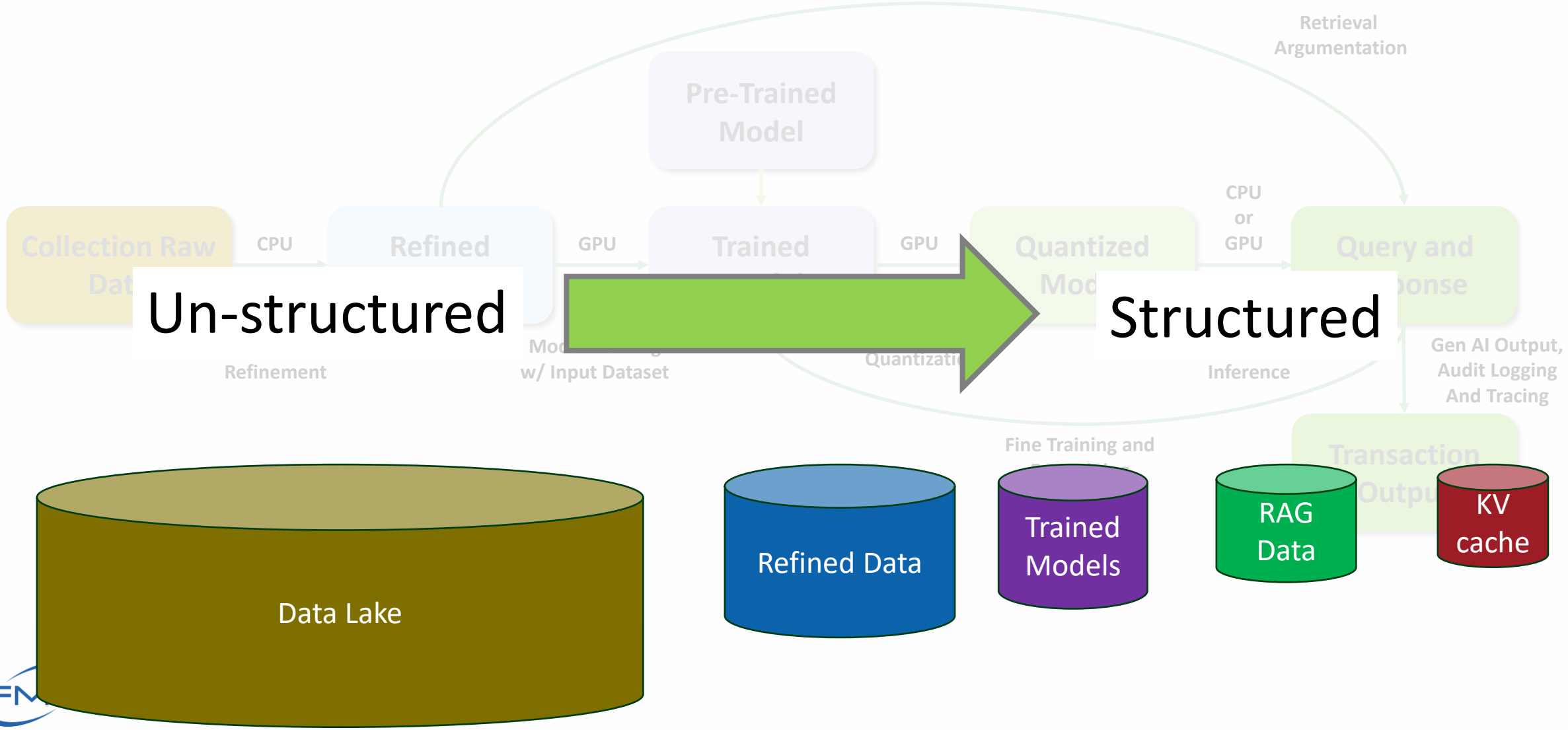
Accelerate Everything © 2025 Supermicro

AI Data Pipelines



Accelerate Everything © 2025 Supermicro

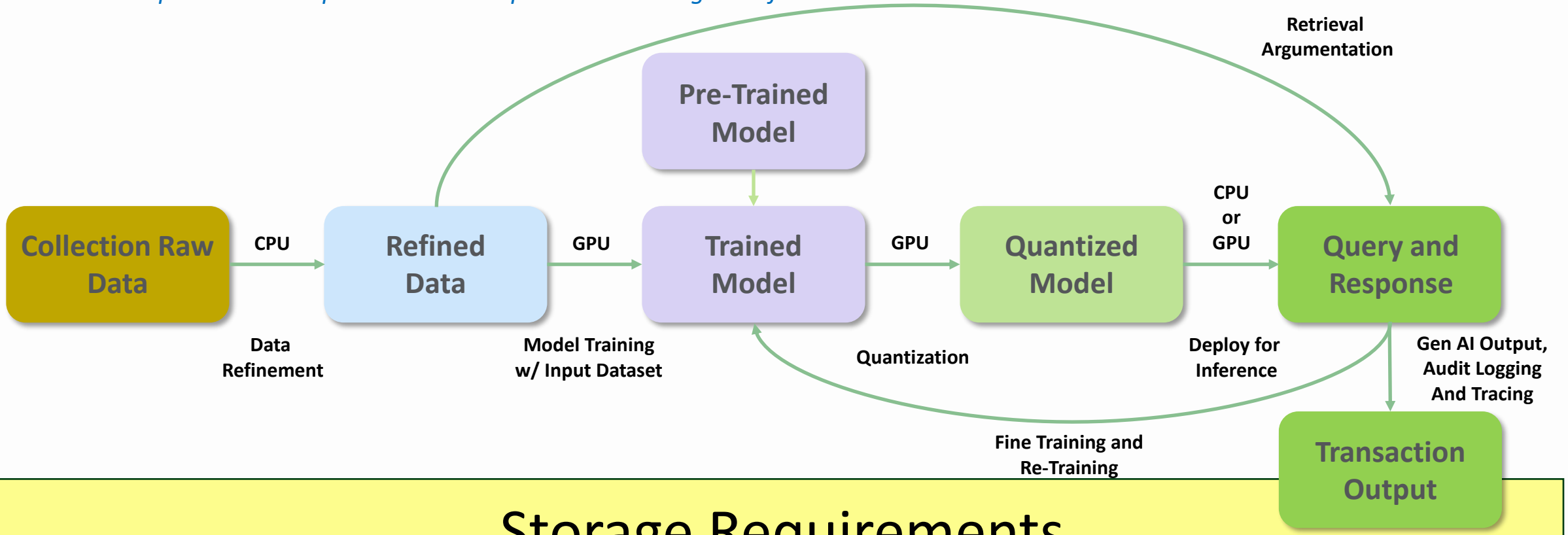
AI Data Pipelines



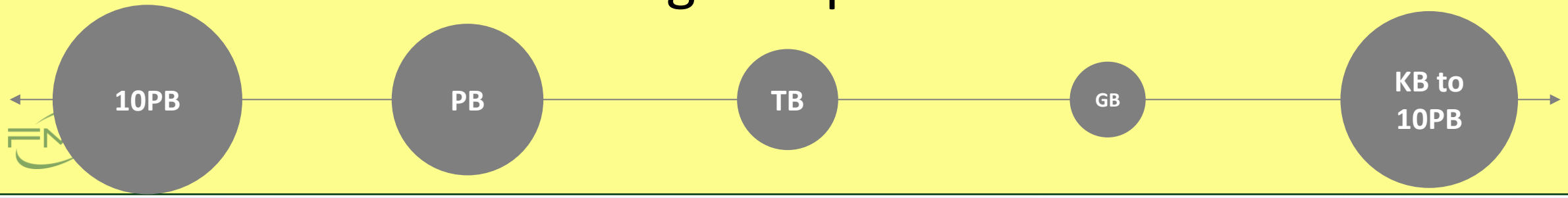
Accelerate Everything © 2025 Supermicro

Storage for AI Pipelines

New Complex AI Data Pipeline Needs Purpose-Built Storage Platforms



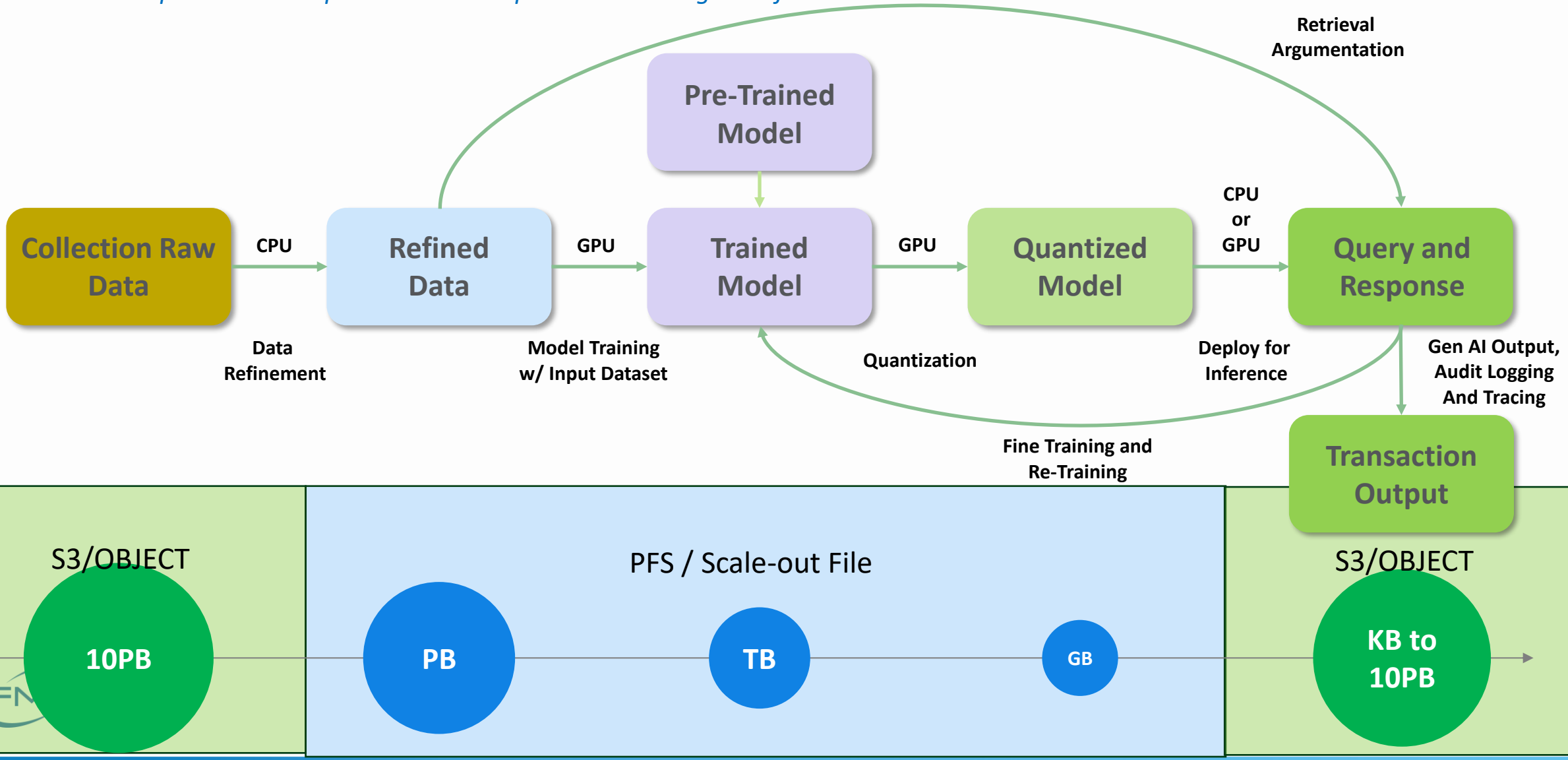
Storage Requirements



Accelerate Everything © 2025 Supermicro

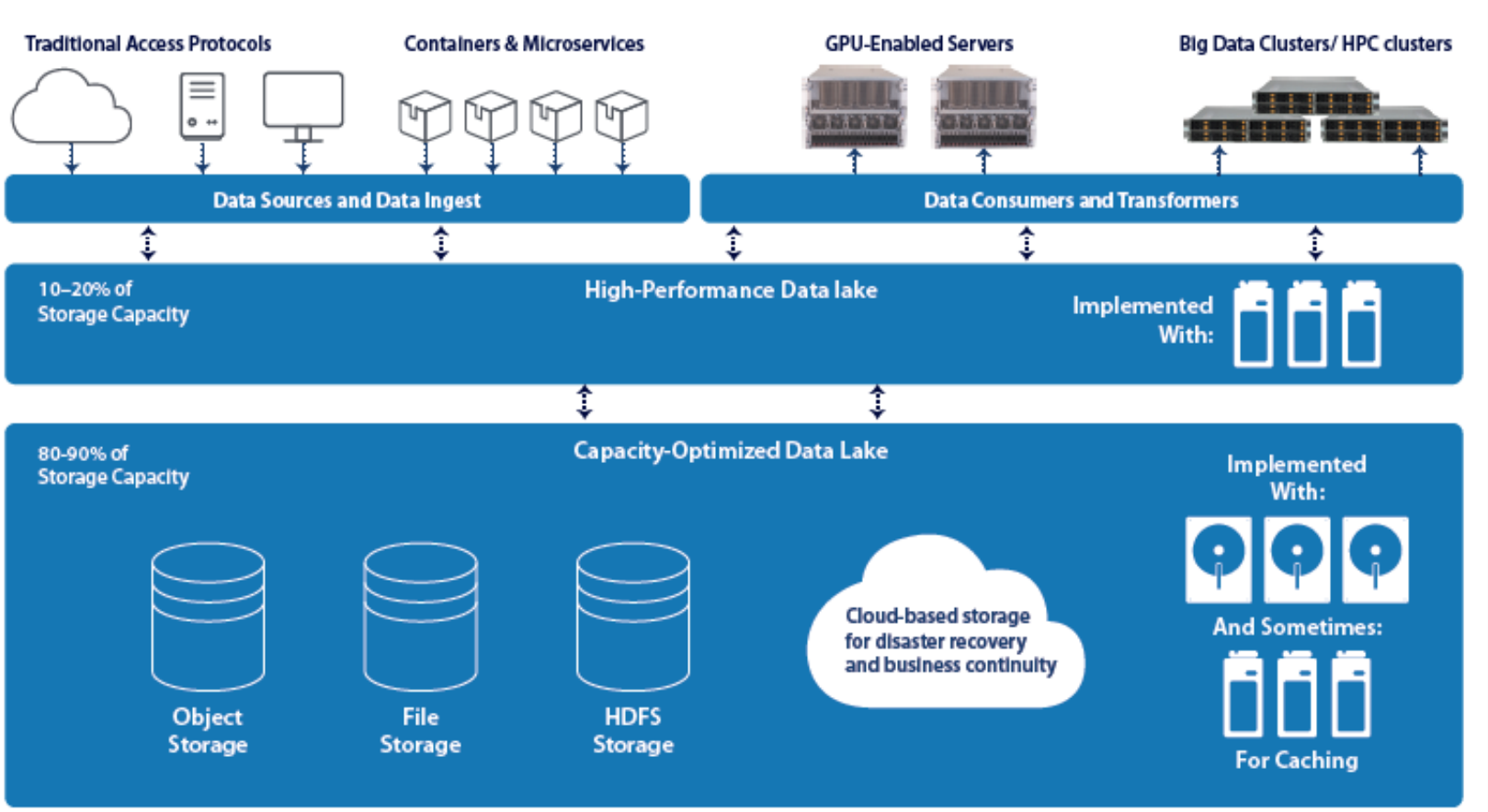
Storage for AI Pipelines

New Complex AI Data Pipeline Needs Purpose-Built Storage Platforms



Accelerate Everything © 2025 Supermicro

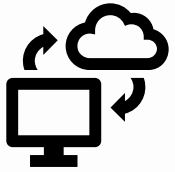
What is a Data Lake?



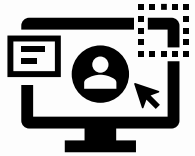
- Data Lake aggregates large amounts of structured data and unstructured data in the original format for further processing
- Can be implemented with HDD storage or Flash

Better Faster Greener™ © 2024 Supermicro

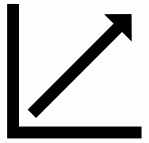
Key Benefits of Data Lakes



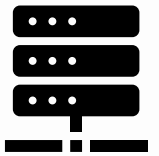
Native Data Storage Flexibility



Enhanced Security Features



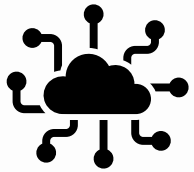
Scalability and Elastic Growth



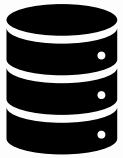
Cost-Effective On-Prem Infrastructure

Better Faster Greener™ © 2024 Supermicro

Applications for Data Lakes



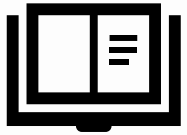
AI/ML Training and Inference



Data Warehousing and Business Intelligence

IoT Data Storage

Healthcare Records

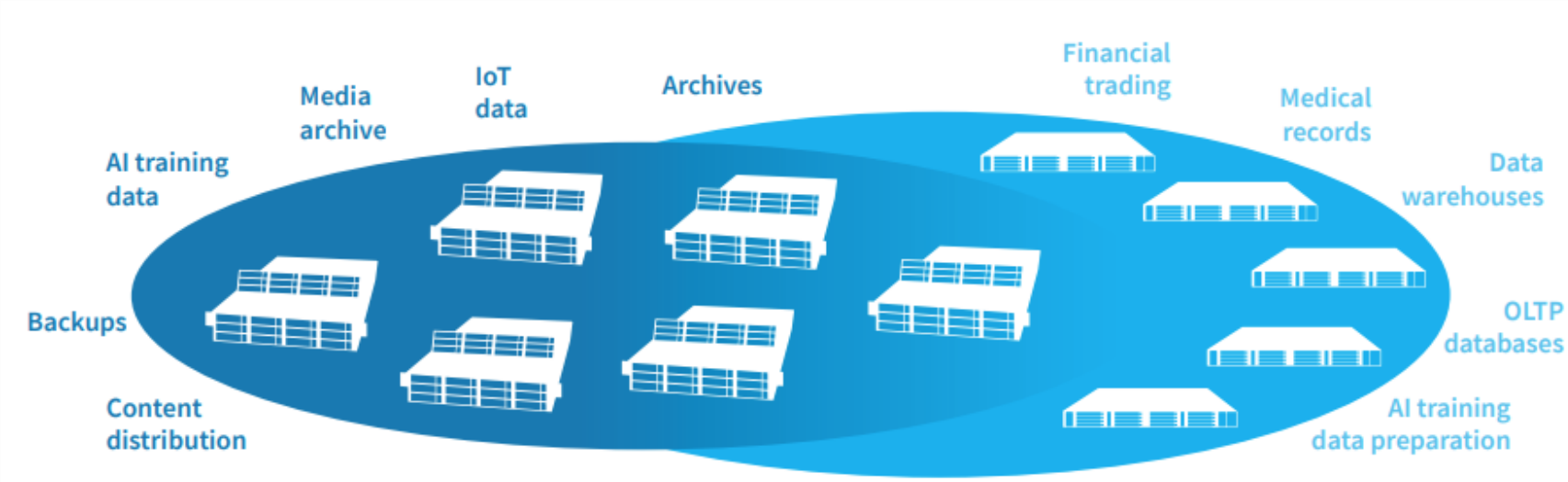


Content Asset Management

Better Faster Greener™ © 2024 Supermicro

Legacy vs. Modern Data Lake Architectures

- Legacy Data Lake Architectures
  - Data tightly coupled to applications, leading to isolated data silos
  - Hadoop data lakes have high complexity in accessing and managing data
  - Data warehouses have fixed schema, optimized for analytics
- Modern Data Lake Architectures
  - Utilize object storage for scalability and access
  - Seamless integration across systems



Better Faster Greener™ © 2024 Supermicro

Data Lakehouses

- Merges the scalability of data lakes with transactional integrity of data warehouses
- Decouple storage from compute, so compute resources can be scaled on demand
- Optimizes for AI/ML workloads requiring high-speed writes

Better Faster Greener™ © 2024 Supermicro

SuperStorage Family

ENTERPRISE	TOP LOADING	SIMPLY DOUBLE	HIGH AVAILABILITY DUAL-PORT	PETASCALE ALL-FLASH
				
Capacity Optimized - Cloud Density 3.5" Storage			**TCO Optimized All-Flash Storage**	
• Optimized for enterprise data centers and edge deployments • Front access, double-sided, product families • 2U/3U/4U form factors • 16-36 3.5" LFF Bays	• Maximum capacity for cloud-scale storage in 4U FF • Highest storage capacity and $/TB value • Single node and dual-server systems • 60, 90 drives in 4U • Dual Intel CPUs	• High density LFF storage in 2U FF • 2U Form Factor • Up to 24x 3.5" bays and up to 4x Gen5 NVMe SSDs for caching • Single Intel (X13) or AMD (H13) CPU	• High Availability Dual Port (Active-Active or Active-Passive) • Shared Everything Architecture (Scale-up Building Block) in 2U • All-flash 24x U.2 NVMe drives • Single Intel Xeon 5th Gen CPU, dual server configuration	• Revolutionary storage performance and capacity • Highest IOPS and BW • All-flash EDSFF NVMe in 1U and 2U • Up to 32x E3.S NVMe SSDs • Dual Intel and Single AMD CPUs
3.5” HDD Storage Servers			**All-Flash Storage Servers**	

Accelerate Everything © 2025 Supermicro

Data Lake Solution Partners

CLOUDIAN	ddn	EDB
Cloudian Hyperstore is an on-premises, S3-compatible data lake platform, particularly suited for AI/ML applications. HyperStore can scale to exabyte levels without disruption to operations.	*DDN* offers the EXAScaler parallel file system to support a wide range of data-intensive workflows in a hyperconverged model. It supports POSIX, NFS, SMB, HDFS, and S3 access to storage. Its Infinia product supports file and object stores.	*EDB Postgres AI* solves common enterprise challenges by extending enterprise-grade Postgres with native AI vector processing, an analytics lakehouse, and a unified platform for observability and hybrid data management.
HAMMERSPACE DATA IN MOTION	IBM	MINIO
Hammerspace Global Data Platform unifies unstructured data across edge, data centers, and clouds. It provides extreme parallel performance for AI, GPUs, and high-speed data analytics, and orchestrates data to any location and supports a wide range of protocols including NFS, pNFS, SMB, S3 and CLI.	*IBM Storage Ceph* is a software-defined scale-out enterprise storage platform that provides block, file, and object capabilities. Its best in class S3 capabilities support the largest and most intensive data applications.	*MinIO Enterprise Object Store* is an ultra high-performance object store that is used to deliver against AI/ML, analytics and archival workloads - all from a single platform. Software-defined, MinIO is ideal for a broad class of SuperMicro servers.
OSNEXUS STORAGE MADE SIMPLE	Quantum.	Qumulo
OSNexus QuantaStor is a scale-out shared-storage solution that delivers multi-tenant, S3-compatible, object storage with unique dynamic tiering features to optimally place objects for maximum performance and cost efficiency.	*Quantum ActiveScale* merges flash, disk, and optionally, tape libraries, to build data lakes, storage clouds, and computing clusters at any scale with outstanding performance, efficiency, availability, and durability at up to 80% lower cost than alternative solutions.	*Qumulo* is a single platform for all unstructured data, no matter where that data resides. Qumulo is built for managing geographically distributed file and object data on-premises, at the edge, in the core, and in the cloud.
SCALITY	VAST	WEKA
Scality solves organizations' biggest data storage challenges — security, performance, and cost. The world's most discerning companies trust Scality so they can grow faster and execute AI data-driven ideas quicker — while increasing efficiency and avoiding lock-in.	*VAST Data* enables data-intensive enterprises to effortlessly capture, catalog, refine, and enrich data through an API-driven data pipeline, providing real-time insights for applications like agentic AI. VAST empowers organizations to challenge conventional thinking and unlock transformative possibilities through its innovative data platform.	*WEKA* delivers a software-defined data platform that helps get your data out of silos and islands and into streaming data pipelines to fuel information workloads like AI and HPC.

Object Storage

- Cloudian
- DDN (Infinia)
- IBM Ceph
- Minio
- OSNexus
- Quantum ActiveScale
- Scality

File with Object Tier

- Hammerspace
- Qumulo
- WEKA
- VAST

Data Lakehouse:

- EDB
- VAST

Better Faster Greener™ © 2024 Supermicro

DISCLAIMER

Super Micro Computer, Inc. may make changes to specifications and product descriptions at any time, without notice. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of Super Micro Computer, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and Super Micro Computer, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. Super Micro Computer, Inc. assumes no obligation to update or otherwise correct or revise this information.

SUPER MICRO COMPUTER, INC. MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

SUPER MICRO COMPUTER, INC. SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL SUPER MICRO COMPUTER, INC. BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF SUPER MICRO COMPUTER, Inc. IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

© 2025 Super Micro Computer, Inc. All rights reserved.

THANK YOU

www.supermicro.com