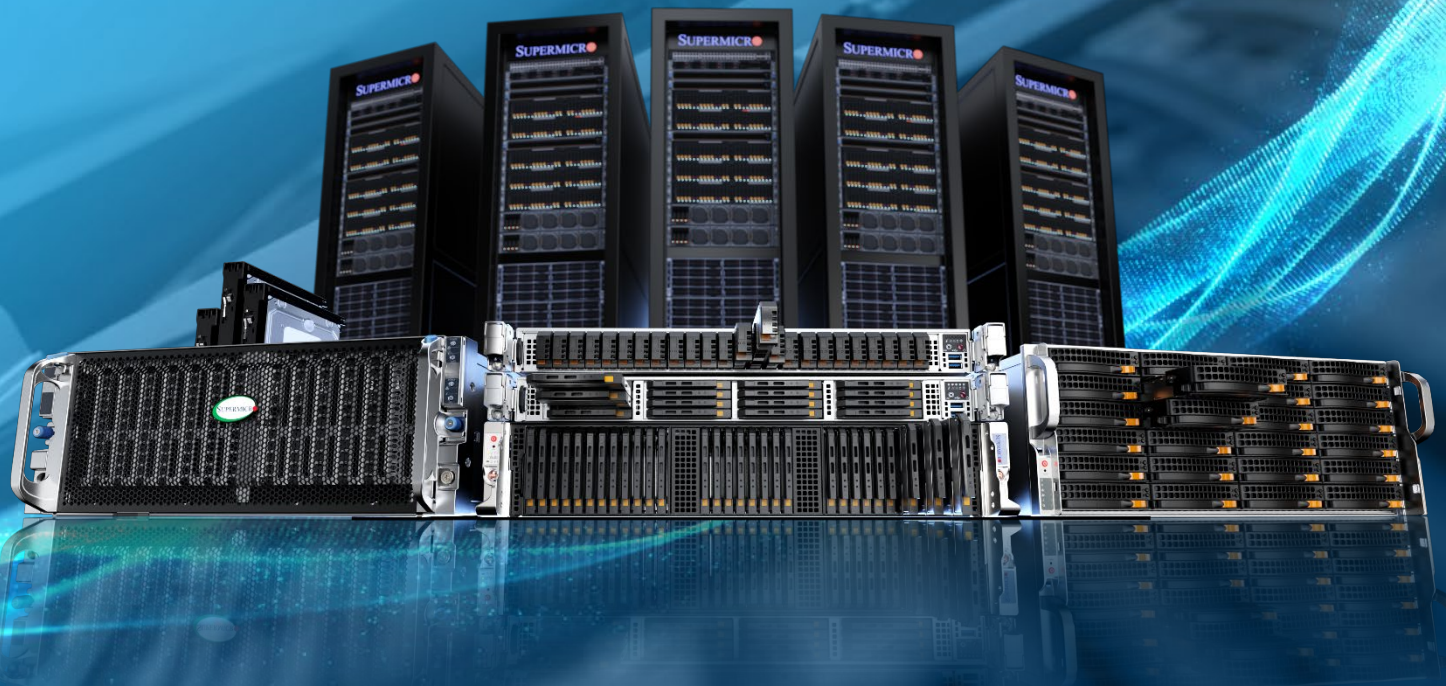
Accelerate AI Data Pipelines

With the Supermicro® Scale-Out Storage Architecture



Agenda

- The problems you face with AI
- Why you need multi-tier storage solutions
- Supermicro’s solution
- How you benefit from working with Supermicro
- How to get started

A Trend Toward Artificial Intelligence (AI)

- Big problems require big data analysis
- AI emerged as the technique of choice
- Model training requires lots of data and lots of GPU acceleration
  - More data means more accurate models
  - More acceleration means faster time to value
- Storing data and making it readily accessible can be expensive
  - Cloud storage costs can surprise you

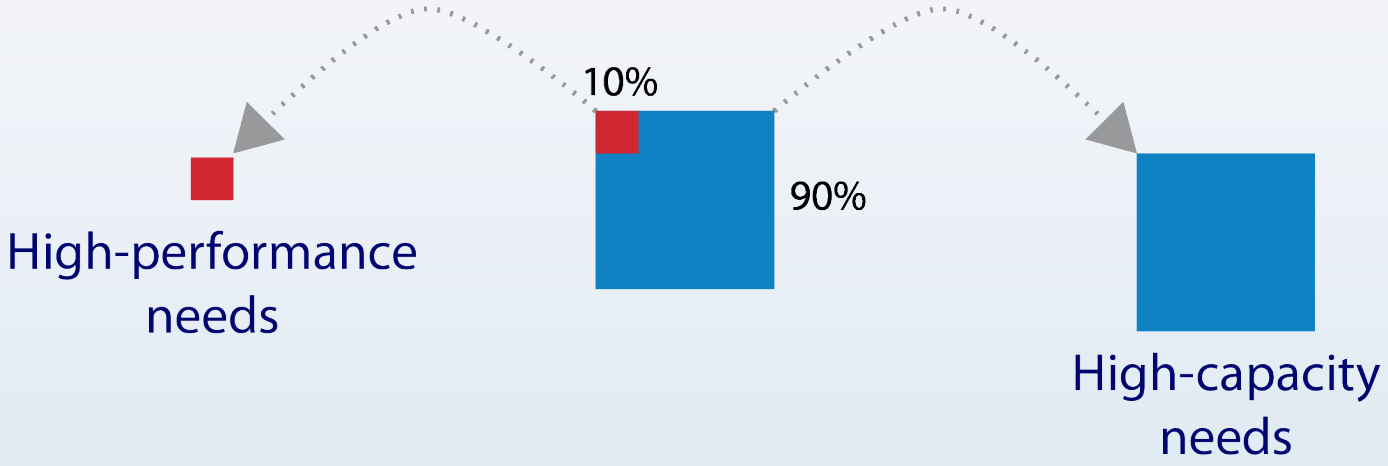
Lots of data and fast

- A customer using AI to automate defect analysis produced 25 PB of data in a week.
- GPU accelerators can consume 400 Gbps as they process image data.

You can’t afford to store all your data on the fastest storage available.

You Need a Multi-Tier Solution

- AI data pipeline
  - Collect and transform data
  - Ingest data into model training
- You don’t use all your data all the time
  - Don’t treat it the same
- Multi-tier storage balances cost and performance
  - **High-performance storage**—drive expensive GPU clusters to shorten the time-to-insight for AI insights
  - **Capacity-optimized storage**—a cost-effective, high-capacity storage a data lake for the rest of your data



A faster AI pipeline means faster time to insight

Our Products Help Tier Storage for Your Needs

From High-Capacity 90-Drive Systems to High-Performance All-Flash NVMe Systems

Capacity

Performance

Enterprise

Hybrid front/rear drive access for object storage and other needs.

High-Capacity

Up to 90 drive bays and dual CPUs for object and file storage.

Multi-Node

Multi-node scale-out storage for high-availability services, HCI, and vSAN.

Petascale All-Flash

Highest level of performance (IOPS, bandwidth) for fast application access.

Where AIOps and MLOps are performed

- GPU-dense servers accelerate AI training and inferencing
- GPUDirect storage transfers directly from/to GPU memory and eliminates CPU or main memory overhead
- Supermicro offers a range of GPU-dense servers to meet every level of application need

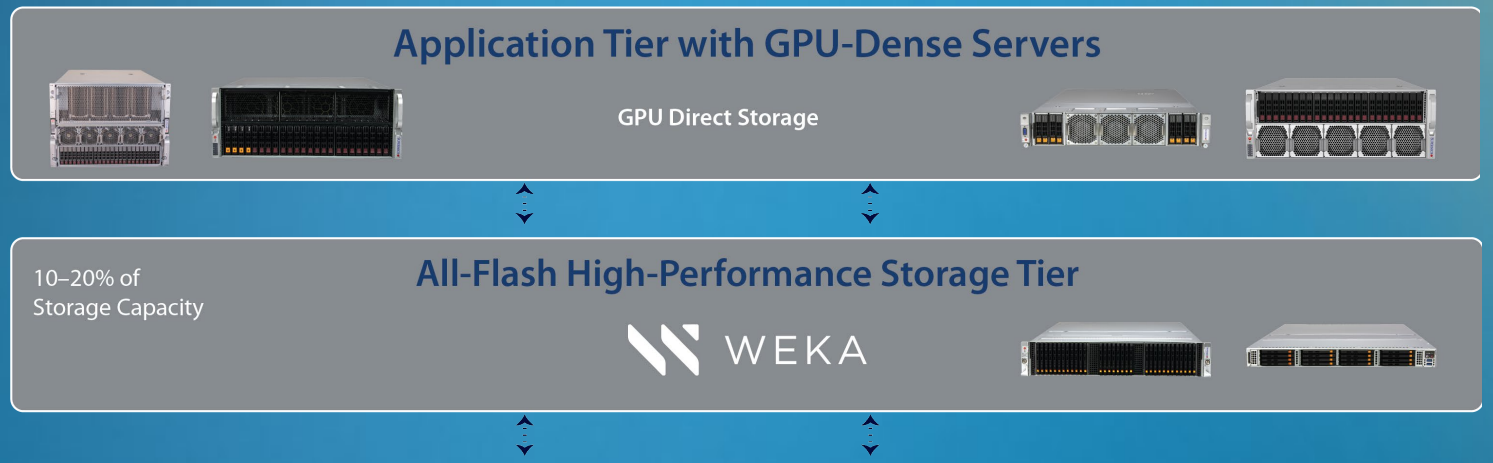
Supermicro-Powered Application Tier

No vendor offers more choices for GPU-accelerated computing

High-Performance All-Flash Tier

Active data is stored here

- Optimized for performance delivers data to applications as fast as they can consume it
  - More utilization -> Better ROI
- Supermicro Peta-scale storage systems
  - 1U and 2U servers using the latest E3.S NVMe storage
- Weka Data Platform
  - Scale-out, hierarchical storage solution
  - Clustered storage solution
  - Data protection and performance

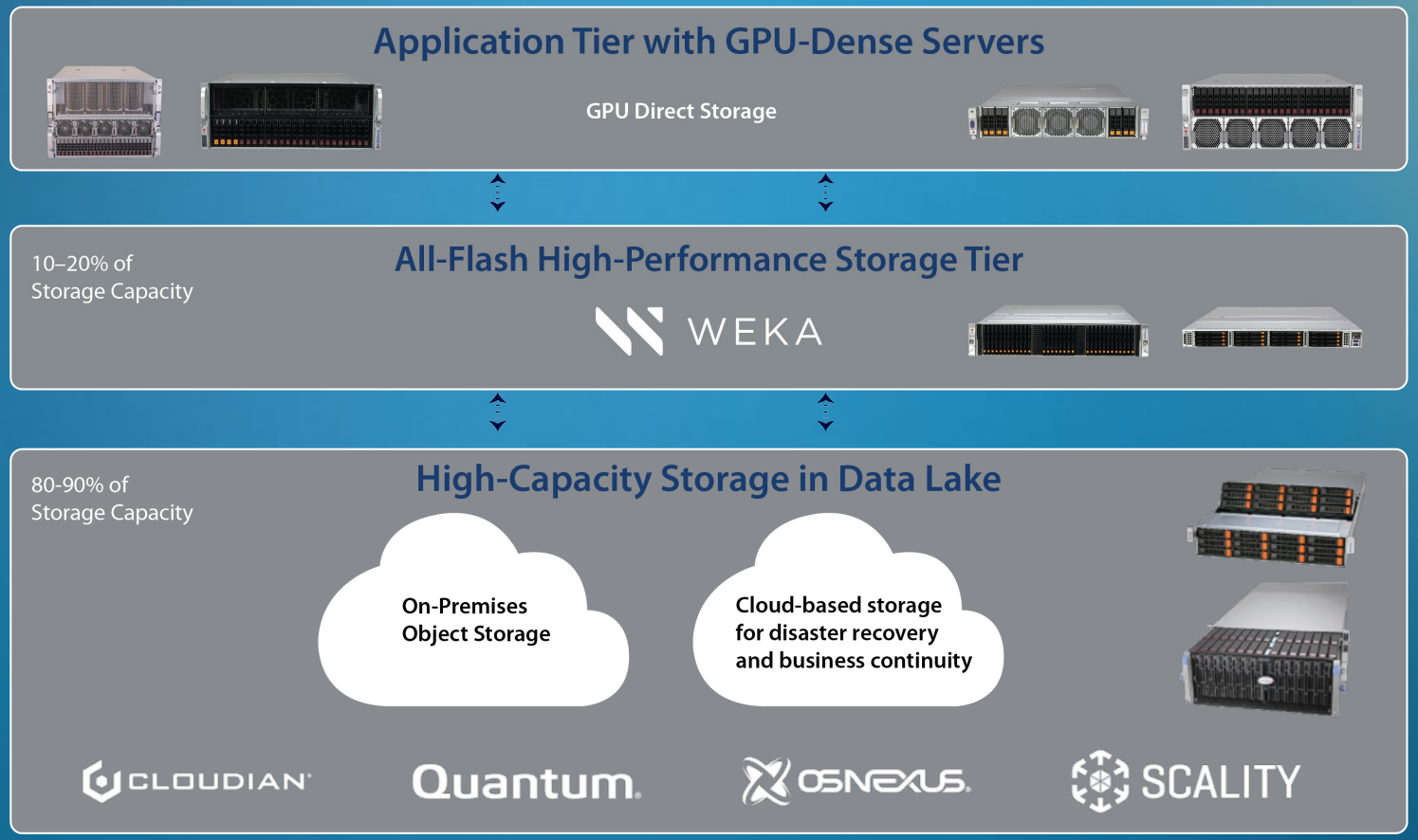


We worked with Weka engineers to optimize for Supermicro storage servers

High-Capacity Data Lake

All training data sets and models stored on-premises

- Data lake uses capacity-optimized storage
- High-capacity spinning-disk storage at lower cost per TB
- Capable back up & Tiering to the private cloud
- Supermicro servers join a scale-out cluster supported by 3rd party software partners

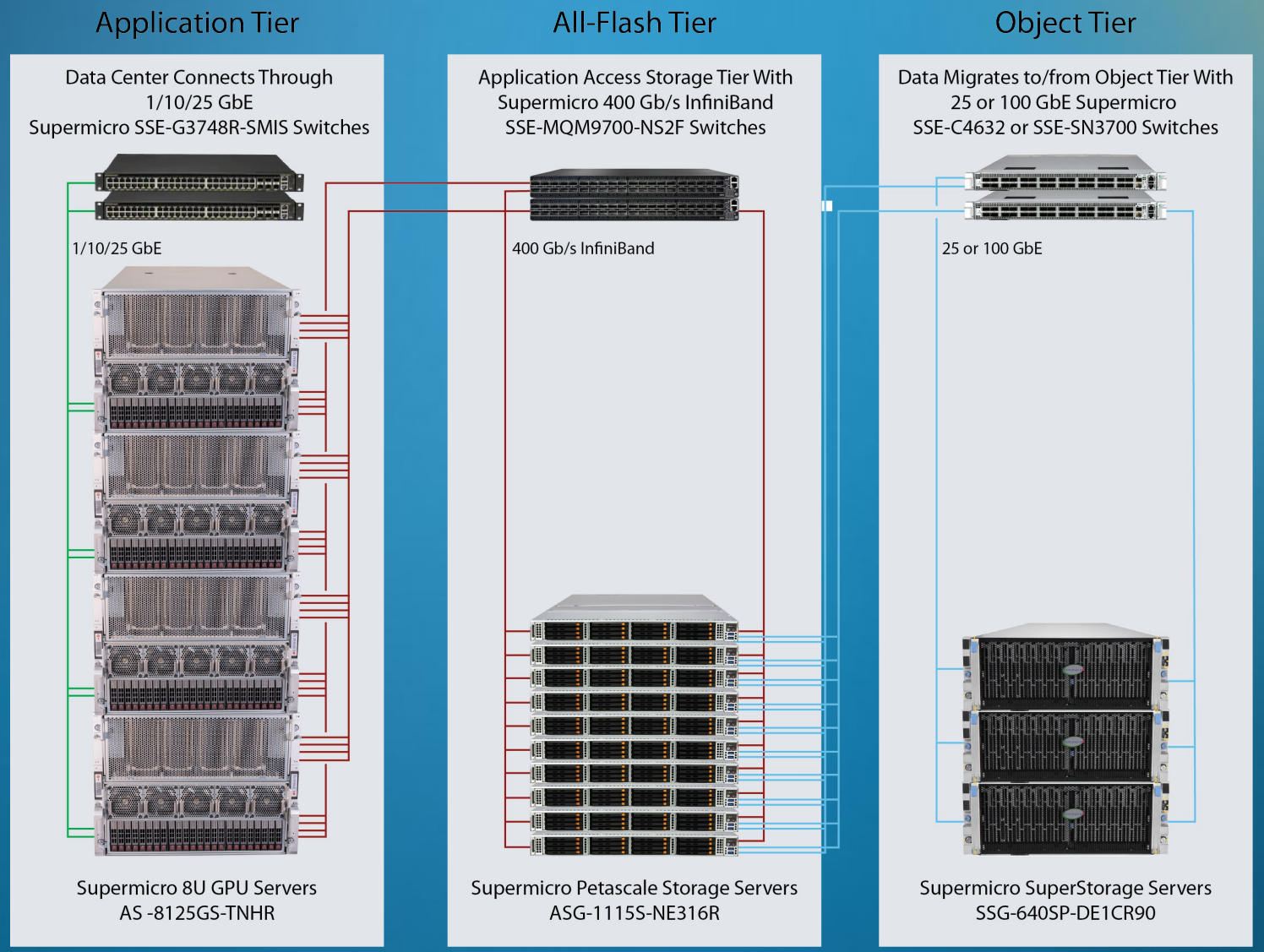


The key to cost-effectively storing all your data, safely, on premises

- Flexible application tier
  - Range of GPU server choices
- Flexible all-flash tier
  - Petascale storage servers
- Capacity-optimized object tier
  - SuperStorage servers
- Interconnections
  - 100–400-Gbps InfiniBand or Ethernet to application tier
  - 25 or 100 GbE for object tier
- Designed for Ease-of-Use for AIOps / MIo|Ops
- Tailored design, integration, testing and deployment to meet exact customer requirements.

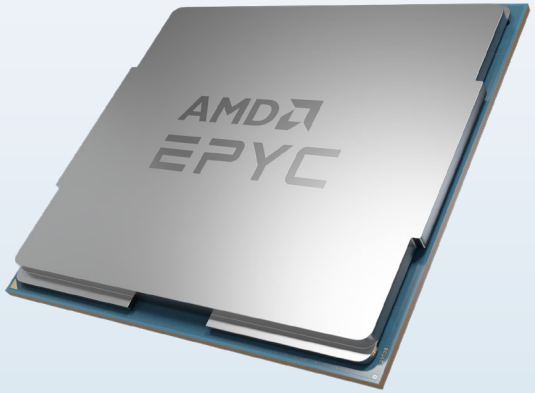
Engineered for AI-powered manufacturing defect analysis

Tested and Validated Solution



Maximize Your Investment

- Keep your training and inferencing pipelines saturated
- Avoid idle GPUs that waste investments and delay time to value
- Our solution:
  - Engineered with industry-leading capacity and performance
  - AMD EPYC™ processors with high core counts
  - PCIe Gen 5 interfaces, up to 160 lanes
  - High-density, high-performance E3.S drives

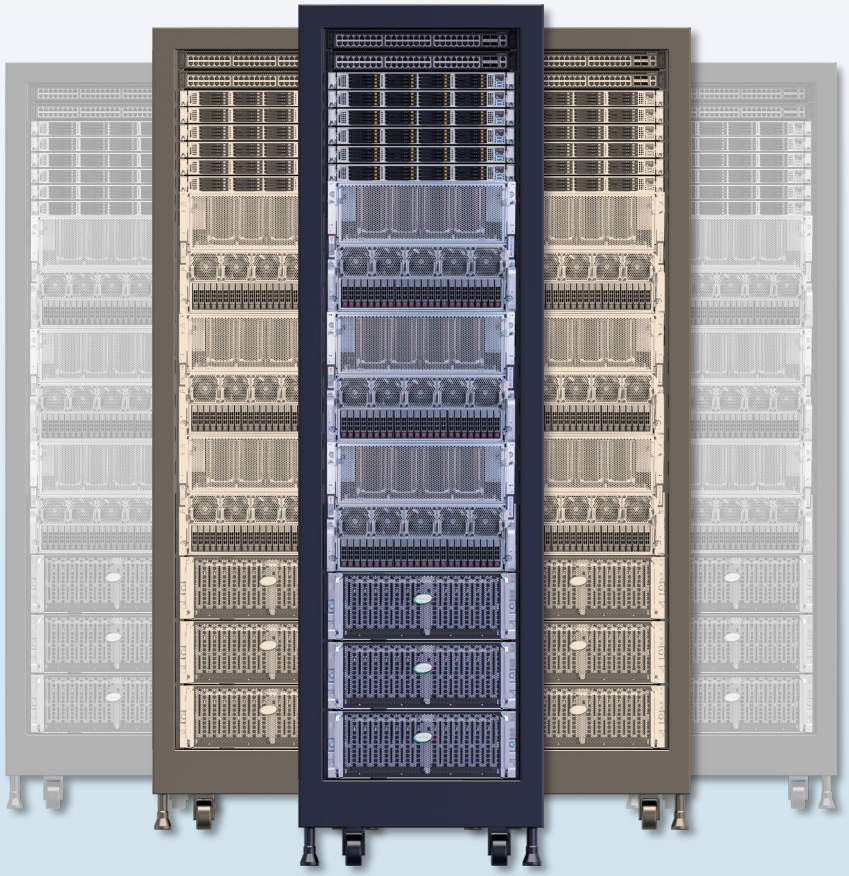


The performance you need to scale your AI workloads.

Complete Rack-Level Integration

Shorter time to value means better return on investment.

- Turn-key solution that is factory tested and customer validated
- Completely configured solutions at the rack level



https://www.supermicro.com/en/solutions/rack-plug-and-play

Choose Turn-Key Data Storage from Supermicro

- Collect, transform, and load massive amounts of raw training data
- Use a proven, multi-tier reference architecture at multi-petabyte scale
- Deploy a diverse set of servers for every workload
- Scale to meet any level of demand
- Flex to change and scale the deployment as needs change
- Tap into an experienced team of storage specialists led by Supermicro
- Help your organization stay on the cutting edge

Leverage a validated blueprint that de-risks deployment.

Elevator Pitch

Updated January 10, 2024

Confidential

Supermicro offers a complete, **proven** reference storage architecture **optimized for performance and capacity** to support AI and machine learning **data workloads**. This multi-tier architecture balances high performance storage requirements needed to keep **AI data pipelines** saturated with large capacity requirements needed for multi-Petabyte data lakes. Supermicro collaborates with leading storage software partners including parallel file provider Weka, object storage partner Quantum ActiveScale and many other object storage partners to **validate, integrate and deliver** the optimized software-defined storage with Supermicro's rack integration process to shorten time-to-deployment. These solutions are deployed on a variety of Supermicro's AMD-powered storage servers including **Petascale servers for the flash tier** which have a massive 32 E3.S drives enabling up to 491 TB of all-flash storage per system and **SimplyDouble capacity-optimized** storage servers for the object tier.

