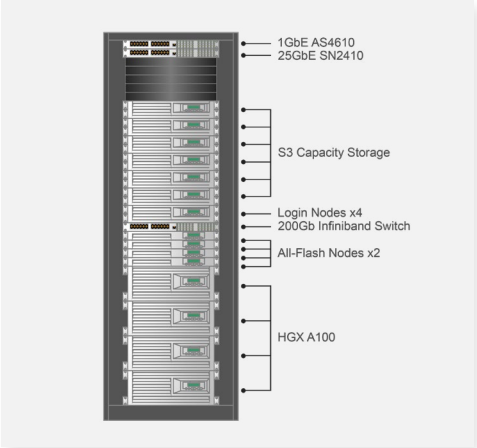
THINKMATE

# AI CLUSTER ARCHITECTURE

## THE NEXT GENERATION IN AI SUPERCOMPUTING INFRASTRUCTURE




Everyone's AI workload and organizational needs are unique. But the desire for faster time-to-result and strong ROI isn't.

The Silicon Mechanics Atlas AI Cluster was specifically designed to achieve this combination with powerful building blocks, resulting in a low TCO system that can be tailored to the needs of your organization.

Without any vendor lock-in. Without worries about scaling. And without concerns over how easily you can upgrade.

Welcome to the next generation of AI hardware. Welcome to Silicon Mechanics Atlas.

## BENEFITS

- Pre-configured for AI, reducing time-to-result even on the largest workloads
- Low total cost of ownership compared to traditional supercomputers
- Seamless, linear scaling
- Includes 3rd Gen AMD EPYC™, the world's highest-performing x86 server CPUi
- Lower initial cost and no platform lock-in compared to the NVIDIA® DGX A100
- Includes NVIDIA A100 GPUs, providing world's fastest memory bandwidth (over 2 TB/s) to run the largest models and datasetsii, and GPU partitioning

## GPU-ACCELERATED COMPUTE NODE COMPONENTS

- 2x AMD EPYC 7742 64-core CPUs (128 cores total), with support for PCIe 4.0
- 9x NVIDIA Mellanox® ConnectX-6 VPI HDR IB 200Gb/s, supports optional GPUDirect RDMA
- NVIDIA Mellanox® Spectrum® SN2000 Gigabit Ethernet HDR switches
- 2x 1.92TB M.2 NVMe Storage (OS) (RAID-1)
- 4x 2200W Power Supplies (3+1)
- 4U Form Factor 8x NVIDIA HGX A100 GPUs with 640GB GPU Memory)
- 2TB DDR4 System Memory
- 30TB NVMe Storage (7.68TB U.2 SSD in RAID-0)
- Includes Silicon Mechanics' AI Stack, Silicon Mechanics' Scientific Computing Stack, and support for popular frameworks

## STORAGE COMPONENTS

- 8 Weka.IO storage nodes + 1 additional Weka.IO node per additional GPU node
- S3 object storage capacity tier. Node count and drive density based on capacity requirements
- Additional optional storage nodes available as needed

## SUPPORTED SOFTWARE & AI FRAMEWORKS

Includes the Silicon Mechanics AI Stack, Silicon Mechanics' Scientific Computing Stack, and support for popular frameworks.

ubuntu

NVIDIA CUDA

PyTorch

R

Keras

TensorFlow

cuDNN

NVIDIA GPU CLOUD

## QUALITY ASSURANCE FOR ZERO DEFECTS

We build each of our systems to “zero defect” standards in our U.S.-based manufacturing facilities. Then we hand-inspect every order, testing them to ensure they are 100% operational and optimized to support rapid deployment.

## STANDARD 3-YEAR WARRANTY

We offer a comprehensive 3-year warranty standard, with every system purchased. But you can add extended or custom warranties if your situation calls for it.

## IN-HOUSE SUPPORT

We offer customer support at different levels, each based on a detailed service level agreement (SLA) that fits your needs.

# ABOUT THINKMATE

All our systems are proudly built in our state-of-the-art, ISO-Certified facility in Norwood, MA. Here, our systems are tested thoroughly before being shipped to you. We keep detailed reliability records of everything, so you can build your own system knowing that every component has been carefully considered by our engineers before being made available to you. But our commitment to quality goes beyond that - Every system is covered under our unified 3-Year Warranty with advance parts replacement. This commitment to superior quality and customer service has kept us the number one server solutions provider for over twenty years.

Learn more at www.thinkmate.com

[i] MLN-016: Results as of 01/28/2021 using SPECrate®2017_int_base. The AMD EPYC 7763 measured estimated score of 798 is higher than the current highest 2P server with an AMD EPYC 7H12 and a score of 717, https://spec.org/cpu2017/results/res2020q2/cpu2017-20200525-22554.pdf. OEM published score(s) for EPYC may vary. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

[ii] https://www.nvidia.com/en-us/data-center/a100/