

GPUltima-CI

FEATURES

- Single or multiple 19" racks
- Up to 128 PCle or SXM4 NVIDIA Ampere GPUs
- Up to 96 PCle or U.2 NVMe drives
- Up to 32 Dual Intel Xeon Scalable Processors nodes
- Up to 32 100Gb Infiniband or Ethernet NICs
- Up to 48-port Liqid Grid PCIe Fabric interconnect
- Power distribution and cooling up to 52kW per rack



GPUltima-CI is the ideal solution for datacenter resident AI training, complementing and supporting rugged edge deployed AI on the Fly platforms in the field. GPUltima-CI allows mixed use datacenters to increase GPU, networking and storage resource utilization by as much as 100% compared to similar hyperconverged server solutions. Unlike traditional architecture where applications must use the available datacenter hardware, OSS GPUltima-CI allows the high-performance application, via the Liqid Command Center API, to dictate the optimal bare-metal hardware configuration for each job to maximize efficiency. Large, flexible reservoirs of GPUs, NVMe storage and NICs are interconnected by a high-speed, low-latency PCIe switched fabric to banks of dual Intel Xeon Scalable Architecture server nodes in each rack. The Command Center Management Software then composes these resources into the optimal set of bare metal servers. This multi-petaflop compute accelerator system is perfect for AI training, deep learning,



SPECIFICATIONS

System		
Rack	42U tall 1200mm traditional rack or Scale Matrix DDC Also available in 24U, 44U and 48U tall versions Supports OSS GPU Accelerators, NVMe Flash Storage Arrays, PCIe fabric switches and quad-node servers	
Computer Accelerators	3U SCA8000 8-way SXM4 A100 expansion with up to four 128Gb PCIe fabric connections 4U EB3600 8-way PCIe A100 expansion with up to four 128Gb PCIe fabric connections Half 4U EB3450 4-way PCIe A100 expansion with up to two 128Gb PCIe fabric connections	
GPUs	SXM4 A100 with NVLink o 6,912 Cuda cores, 432 Tensor cores o 9.7 Tflops Double-Precision o 19.5 Tflops Single-Precision o 312 Tflops Tensor Performance o 64GB/s bi-directional interconnect band o 40GB or 80GB HBM2 memory o 400 watts	PCIe A100 o 6,912 Cuda cores, 432 Tensor cores o 9.7 Tflops Double-Precision o 19.5 Tflops Single-Precision o 312 Tflops Tensor Performance width o 64GB/s bi-directional interconnect bandwidth o 40GB or 80GB HBM2 memory o 250 watts
Flash Storage Arrays	2U FSAe-2 24-way U.2 NVMe JBOF with up to two 128Gb PCIe fabric connections 4U 4UV 16-way PCIe NVMe JBOF with up to two 128Gb PCIe fabric connections	
NVMe Drives	PCIe N1952 o 6.4TB, 3 DW/day o PCIe3.0 x8 64Gb/s o Max Read (128KB): 6.0GB/s o Max Write (128KB): 3.8GB/s o Random Read IOPS (4KB): 1,000,000 o Random Write IOPS (4KB): 300,000 o Write Latency (512B): 12µs	U.2 CM5, FIPS 140-2 o 6.4TB, 3 DW/day o PCle3.0 x4 32Gb/s o Max Read (128KB): 3.35GB/s o Max Write (128KB): 3.04GB/s o Random Read IOPS (4KB): 770,000 o Random Write IOPS (4KB): 165,000 o Write Latency (512B): 20μs



SPECIFICATIONS CONTINUED

2U, 4-node, dual Intel Xeon Scalable Processor server. Each node contains: Servers

Dual Socket P (LGA 3647) "SkyLake" CPUs up to 28 cores and 3.2GHz Up to 2TB ECC DDR4-2666MHz

Two Gen 3 x16 PCIe expansion slots Six 2.5" SATA3 SSDs

IPMI, dual USB 3.0 and Disk-on-module support

Infiniband Switch Mellanox 36 port Infiniband switch EDR 100Gb/s, QSFP connectors

1U form factor

Compostable Infrastructure

Management

Ligid Grid managed switch array up to 8U, 96-ports

128Gbps PCIe fabric per port

Fail-over and multi-topology support

1Gb Management port with Xeon D-1548 management CPU

Infiniband Interface Card Mellanox Connect-X5

EDR 100Gb/s, QSFP connectors Single or dual port available One card per server

Power Distribution Unit Tripp-Lite Monitored PDU

27.6kW power

Input: 380/400V 3 phase, 63A

Power monitoring via display and Ethernet 110kW total power ~ 97% over-provisioned

Cables Copper network and fabric cables inside each rack

Fiber Infiniband and PCIe fabric cables up to 100m available for multi-rack GPUltima-CI deployments

Fully integrated, cabled, racked and stacked datacenter solutions

Software OS, Frameworks

and Libraries

Operating Systems: CentOS, Ubuntu, Suse, Windows Optional Pre-installed deep learning frameworks:

> o Torch o Caffe2 o Theano

o TensorFlow

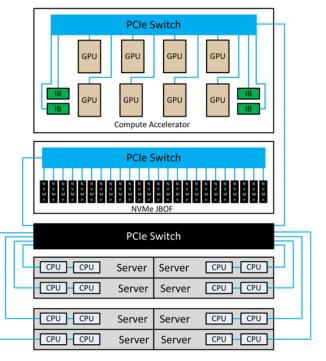
CUDA NVIDIA drivers

Optional Pre-installed deep learning libraries:

o MLPython o cuDNN o DIGITS

o Caffe on Spark

o NCCL



Architecture