Rack-scale, application-centric, composable HPC infrastructure

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**

<table>
<thead>
<tr>
<th>System</th>
<th>42U tall 1200mm traditional rack or Scale Matrix DDC</th>
<th>Also available in 24U, 44U and 48U tall versions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Accelerators</td>
<td>3U SCA8000 8-way SMX4 A100 expansion with up to four 128Gb PCIe fabric connections</td>
<td>4U EB3600 8-way PCIe A100 expansion with up to four 128Gb PCIe fabric connections</td>
</tr>
<tr>
<td></td>
<td>Half 4U EB3450 4-way PCIe A100 expansion with up to two 128Gb PCIe fabric connections</td>
<td></td>
</tr>
<tr>
<td>GPUs</td>
<td>SXM4 A100 with NVLink</td>
<td>PCIe A100</td>
</tr>
<tr>
<td></td>
<td>6,912 Cuda cores, 432 Tensor cores</td>
<td>6,912 Cuda cores, 432 Tensor cores</td>
</tr>
<tr>
<td></td>
<td>9.7 Tflops Double-Precision</td>
<td>9.7 Tflops Double-Precision</td>
</tr>
<tr>
<td></td>
<td>19.5 Tflops Single-Precision</td>
<td>19.5 Tflops Single-Precision</td>
</tr>
<tr>
<td></td>
<td>312 Tflops Tensor Performance</td>
<td>312 Tflops Tensor Performance</td>
</tr>
<tr>
<td></td>
<td>64Gb/s bi-directional interconnect bandwidth</td>
<td>64Gb/s bi-directional interconnect bandwidth</td>
</tr>
<tr>
<td></td>
<td>40GB or 80GB HBM2 memory</td>
<td>40GB or 80GB HBM2 memory</td>
</tr>
<tr>
<td></td>
<td>400 watts</td>
<td>250 watts</td>
</tr>
<tr>
<td>Flash Storage Arrays</td>
<td>2U FSAe-2 24-way U.2 NVMe JBOF with up to two 128Gb PCIe fabric connections</td>
<td>4U 4UV 16-way PCIe NVMe JBOF with up to two 128Gb PCIe fabric connections</td>
</tr>
<tr>
<td>NVMe Drives</td>
<td>PCIe N1952</td>
<td>U.2 CMS, FIPS 140-2</td>
</tr>
<tr>
<td></td>
<td>6.4TB, 3 DW/day</td>
<td>6.4TB, 3 DW/day</td>
</tr>
<tr>
<td></td>
<td>PCIe3.0 x8 64Gb/s</td>
<td>PCIe3.0 x4 32Gb/s</td>
</tr>
<tr>
<td></td>
<td>Max Read (128KB): 6.0GB/s</td>
<td>Max Read (128KB): 3.35GB/s</td>
</tr>
<tr>
<td></td>
<td>Max Write (128KB): 3.8GB/s</td>
<td>Max Write (128KB): 3.04GB/s</td>
</tr>
<tr>
<td></td>
<td>Random Read IOPS (4KB): 1,000,000</td>
<td>Random Read IOPS (4KB): 770,000</td>
</tr>
<tr>
<td></td>
<td>Random Write IOPS (4KB): 300,000</td>
<td>Random Write IOPS (4KB): 165,000</td>
</tr>
<tr>
<td></td>
<td>Write Latency (S12B): 12µs</td>
<td>Write Latency (S12B): 20µs</td>
</tr>
</tbody>
</table>

sales@onestopsystems.com • +1 877-438-2774 • www.onestopsystems.com
SPECIFICATIONS CONTINUED

Servers
2U, 4-node, dual Intel Xeon Scalable Processor server. Each node contains:
- 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
Liqid 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:
- Torch
- Caffe2
- Theano
- TensorFlow

CUDA NVIDIA drivers
Optional Pre-installed deep learning libraries:
- MLPython
- cuDNN
- DIGITS
- Caffe on Spark
- NCCL

Architecture