AV1000

3rd Gen. Intel Xeon Scalable ICE LAKE Liquid Cooled GPU Server

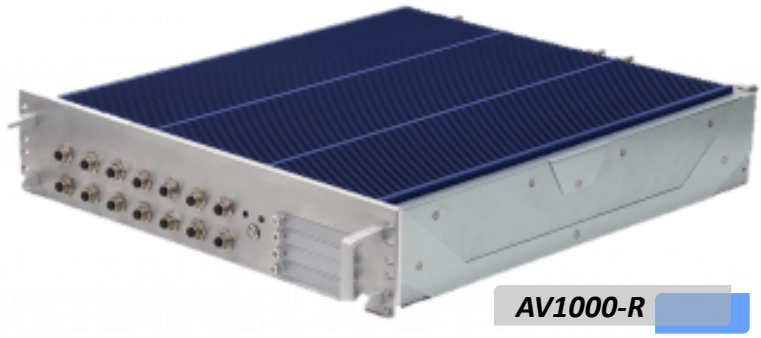
NVIDIA QUADRO RTX 6000

- **Thermal Dissipation Coefficient : 4kW**
- **Liquid Cooled Platform :**
  Gun Drilled -10 x 500 x 10mm Pass (⌀ x L x H)
- **CPU: 3rd Gen. Intel Xeon Scalable (ICE LAKE)**
  ® Platinum 8380 (40 x cores, 2.3 GHz, 270W)
- **GPU: 2 x NVIDIA RTX 6000 Passive (24 GB GDDR6, 4,608 CUDA Cores)**
- **8 x DDR4-3200MHz up to 2TB**
- **Storage(1): NVMe SSD (1 x PCIe Gen 4.0 x4, 1 x PCIe Gen 4.0 x8) up to 48TB**
- **Storage(2): 4 x SATA 6.0 SSD up to 16TB**
- **Dimension: 450 x 500 x 88 mm (WxDxH)**

INDEX

1. Introduction & Key Features

1-1 Overall Introduction



AV1000-R

The global transformation is rapidly scaling the demands for flexible computer, networking, and storage. Future workloads will necessitate infrastructures that can seamlessly scale to support immediate responsiveness and widely diverse performance requirements. The exponential growth of data generation and consumption, the rapid expansion of cloud-scale computing and 5G networks, and the convergence of high-performance computing (HPC) and artificial intelligence (AI) into new usages requires that today's data centers and networks evolve now— or be left behind in a highly competitive environment.

7Starlake's AV1000 AI Inference Rugged Server which are featuring **3rd Gen. Intel Dual Xeon Ice Lake Scalable Processors** (40 Cores, 2.3 GHz, 270W) with **2 x NVIDIA QUADRO RTX6000**, **8 x DDR4-3200MHz 2TB memory** and **2 x NVMe SSD up to 48TB**, to provide the seamless performance foundation for the data centric era from the multi-cloud to intelligent edge, and back. The Intel Xeon Scalable platform provides the foundation for an evolutionary leap forward in data center agility and scalability. Disruptive by design, this innovative processor sets a new level of platform convergence and capabilities across compute, storage, memory, network, and security.

AV1000 enables a new level of consistent, pervasive, and breakthrough performance in new AI inference to implement machine learning and deep learning. In addition to NVIDIA QUADRO RTX6000, AV1000 provides one M.2 NVMe slot for fast storage access. Combining stunning inference performance, powerful CPU and expansion capability, it is the perfect ruggedized platform for versatile edge AI applications.

With innovative liquid cooled design, AV1000 is built in most advanced " Gun Drilled ", which with 10 pipes (each pipe 5mm x 5mm x π x 500mm) to dissipate max 10KW heat.

AV1000

1-2 3rd Gen. Intel Xeon Ice Lake Scalable Platinum Processors

AV1000's 3rd Gen Xeon Scalable processor is based on a potent architecture that increases core performance, memory, and I/O bandwidth to accelerate diverse workloads from the data center to the edge. It is available with up to 40 powerful cores and built-in workload acceleration features include: Intel Deep Learning Boost, Intel Advanced Vector Extensions 512, and Intel Speed Select Technology.

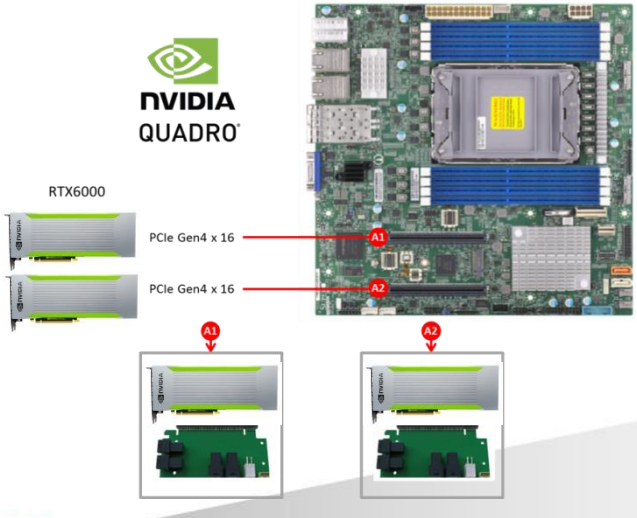
Compared to previous generation, 3rd Gen Xeon Scalable processor features:

- **1.46x** average gen-on-gen performance improvement
- **Up to 1.60x** higher memory bandwidth
- **Up to 2.66x** higher memory capacity
- **Up to 1.33x** more PCI Express lanes per processor

1-3 NVIDIA QUADRO RTX6000 Passive

AV1000 supports **2** **x NVIDIA Quadro RTX6000 Passive** PCIe 3.0 x16 Graphics card that can power the planets most reliable mainstream workstations. Meet your visual computing challenges with the power of NVIDIA® Quadro® RTX™ GPUs in the data center. Built on the NVIDIA Turing architecture and the NVIDIA RTX platorm, the NVIDIA Quadro RTX 6000 passively cooled graphics board features RT Cores and multi-precision Tensor Cores for real-time ray tracing, AI, and advanced graphics capabilities. Tackle graphics-intensive mixed workloads, complex design, photorealistic renders, and augmented and virtual environments at the edge with NVIDIA Quadro RTX, designed for enterprise data centers.

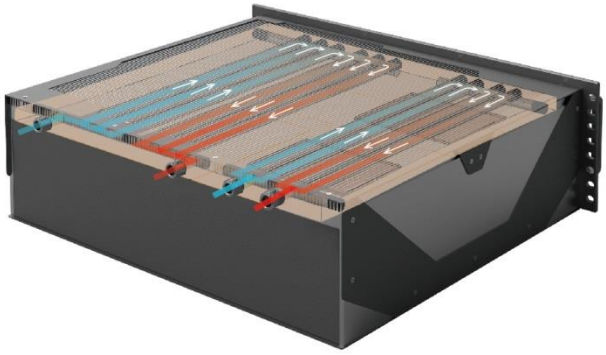
SPECIFICATIONS	
GPU Architecture	NVIDIA Turing
NVIDIA CUDA Cores	4,608
NVIDIA Tensor Cores	576
NVIDIA RT Cores	72
Single-Precision Perf.	14.38 TFLOPS
Half-Precision Perf.	28.75 TFLOPS
Tensor Perf.	119.4 TFLOPS
GPU Memory	24GB GDDR6
Memory Bandwidth	Up to 624 GB/sec
Power Consumption	260 W
Form Factor	PCIe 3.0 x16
Thermal Solution	Passive
Compute APIs	CUDA, DirectCompute, OpenCL™

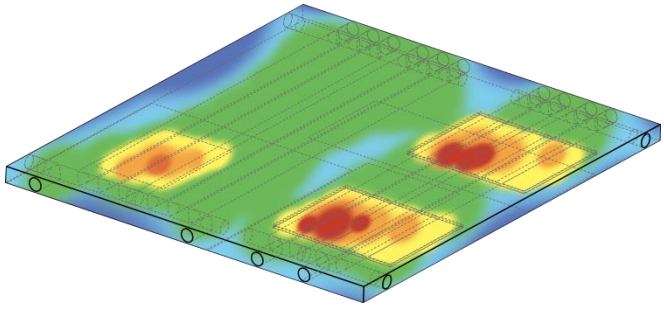


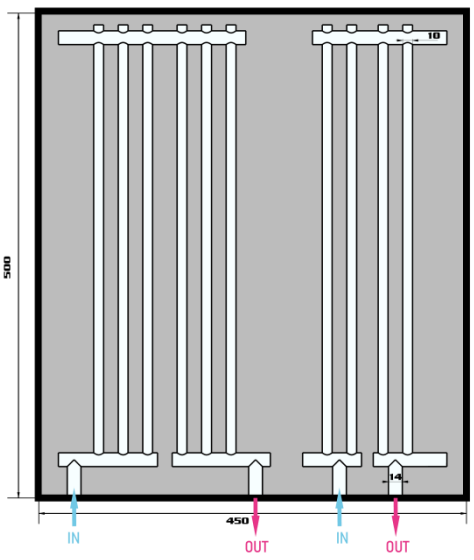
1-4 Conduction Liquid Cooling Plate (C.L.C.P.)

Most liquid cooling solutions are using close loop design — Direct to Chip (D2C) integrated pump & cold plate in the system. Users may worry about potential risk of liquid leakage.

7Starlake highly values system reliability. In the pursuit of stability and power, our experienced team has successfully optimized the thermal design, bringing out an unprecedented model AV1000. Instead of normal D2C design, 7Starlake innovated an unique heat exchanger integrating **Conduction Liquid Cold Plate (CLCP)** on the computing system.







CLCP includes multi-channel cold water inlet/outlet owning high flexibility in adjusting numbers of inlet/outlet (up to 4in 4out) by request. When coolant flows through top sink, liquid can absorb the heat and take it away from the heat sources quickly to the heat exchanger. When heated liquid flushes into the heat exchanger, it will be cooled by 9 units of 12 x 12 cm active fan which can run at 2K~3K RPM.

AV1000	
Item	**Description**
CPU	1 x Intel Xeon Platinum 8380
GPU	2 x NVIDIA QUADRO RTX 6000
System Size	480 x 500 x 88mm (W x L x H)
CLCP Size	450 x 500 x 20mm (W x L x H)
Material of CLCP	AL6063
Gun Drilled HoleØ	10mm

Leveraging both liquid-cooling and air-cooling's strong points; these features accomplish higher rack density and efficiency, comprehensive reduction in power use, and increase of overclocking potential.

1-5 How to Leverage NVMe for AI & Machine Learning Workloads

The massive amount of data which is collected from approaches mentioned above is requested for further purpose. Driverless vehicles training can employ the sorted data for future improvement and progressively enhance the road safety and further development, and this occupies the most data. An instrumented vehicle can consume over 30TB of data per day while a fleet of 10 vehicles can generate 78PB of raw data. Normally, the data rate of a camera is approximately 120MB/s while that of radar is close to 220 MB/s. To sum up, if a vehicle has a combination of 6 cameras and 6 radars, the complete vehicle RAW data will roughly be 2.040 KB/s, which is around 58TB in an 8 hour test drive shift.

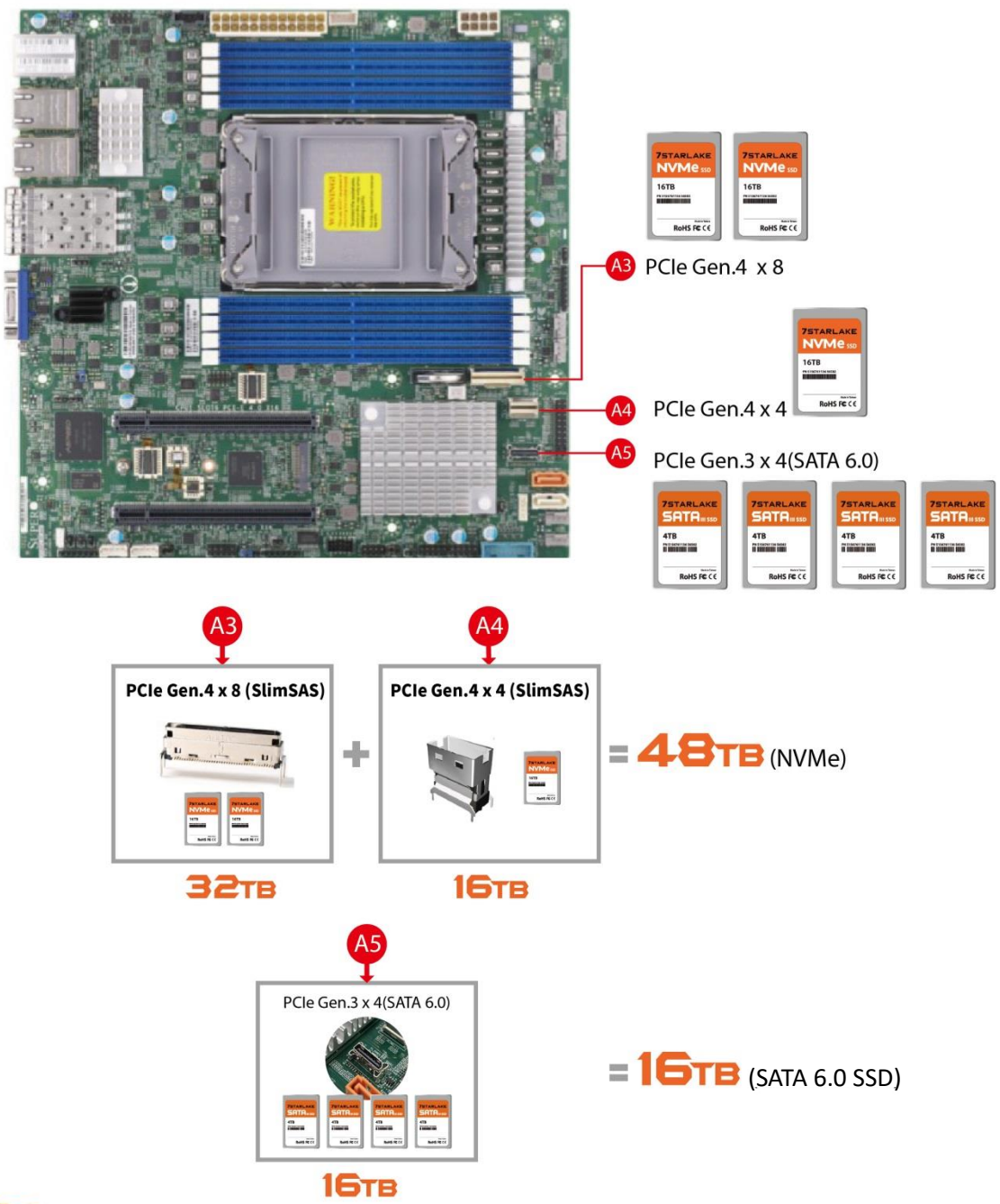
The modern datasets used for model training can be up to terabytes each. Even if the training itself runs from RAM, the memory should be fed from non-volatile storage, which has to support very high bandwidth. In addition, paging out the old training data and bringing in new data should be done rapidly to keep the GPUs from being idle. This necessitates low latency, and the only protocol allowing for both high bandwidth and low latency like this is NVMe.

Fortunately, GPU servers have massive network connectivity. They can ingest as much as 48GB/s of bandwidth via 4-8 x 100Gb ports – playing a key part in one of the ways to solve this challenge.

NVMesh enables the chipset, core count, and power to be customized to match varying data workload and performance requirements. Combined with the Ultrastar Serv24-4N, NVMesh allows GPU optimized servers to access scalable, high performance without sacrificing performance and practicality. NVMe flash storage pools as if they were local flash. This technique ensures efficient use of both the GPUs themselves and the associated NVMe flash. The end result is higher ROI, easier workflow management and faster time to results.

1-6 How AV1000 Support 48TB NVMe

PCIe Gen 4 is the upper version of PCIe Gen 3, which surpassed PCIe Gen 2 and Gen 1. The bandwidth provided by PCIe Gen 4 is double as compared to PCIe Gen 3. It is backward compatible with prior generations of PCIe. PCIe Gen 4 is expected to satisfy, to a large extent, the requirements of high speed servers, gaming, graphics and data centers, where number of servers and solid-state devices are increasing as per market demand. PCIe Gen 4 is the new evolution over the PCIe Gen 3. The PCIe Gen 4 provides the data rate of 16 G/Ts as compared to 8G/Ts provided by PCIe Gen 3. It's architecture is fully compatible with all the previous generations of PCIe.



2. Specifications

System	
CPU	3rd Gen. Intel Xeon Scalable (ICE LAKE) ® Platinum 8380 (40 x cores, 2.3 GHz, 270W)
Memory type	8 x DDR4-3200MHz up to 2TB
GPU	2 x NVIDIA RTX 6000 (24 GB GDDR6, 4,608 CUDA Cores)
Storage	
Storage(1)	2 x NVMe SSD (1 x PCIe Gen 4.0 x4, 1 x PCIe Gen 4.0 x8 up to 48TB)
Storage(2)	4 x SATA 6.0 SSD up to 16TB
Ethernet	
Ethernet	2 x 10G Base-T
Rear I/O	
VGA	1 (M12)
Ethernet	2 x 10GbE (M12)
USB	2 x USB 3.0 (M20)
IPMI	1 x IPMI 2.0
Power Requirement	
Power	16V~32V DC-IN DC-DC
Mechanical	
Dimension	450 x 500 x 88 mm (W x D x H)
HeatExchanger (HE4K)	
Cooling Capacity	4KW
Server Managed	4
Facility Liquid Integration	No
Dimension	450 x 450 x 176 mm (W x D x H)

3. System Configuration