STAR LAKE

AV3000

AI Inference Rugged Server

Intel XEON CPU with TESLA T4 GPU

- Dual Xeon Scalable Platinum 8380 (40 cores)
- High Memory Capacity 768GB
- 4x NVIDIA TESLA T4 GPU
- MIL-461 18V~36V DC-DC 800W
- NVMe SSD (4GB/s) 80TB
- Liquid Cooling
- Intel Optane DC Persistent Memory support

Content

1. Introduction & Key Features

1-1 Overall Introduction

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.

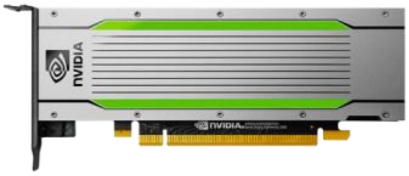
7Starlake's AV3000 AI Inference Rugged Server which are featuring Dual Xeon Scalable Platinum 8380 Processor (40 cores) with 4 x NVIDIA TESLA T4 , DDR4-768 GB memory and 80 TB NVMe, to provide the seamless performance foundation for the data centric era from the multi-cloud to intelligent edge, and back. 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.

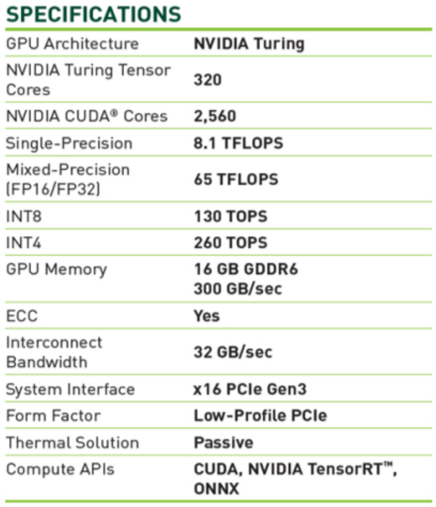
AV3000 enables a new level of consistent, pervasive, and breakthrough performance in new AI inference to implement machine learning and deep learning. In addition to Tesla T4, AV3000 provides one M.2 NVMe slot for fast storage access. Combining stunning inference performance, powerful CPU and expansion capability, it is the perfect ruggedized platform for versatile edge AI applications.

AV3000 ruggedized AI inference platforms designed for advanced inference acceleration applications such as voice, video, image and recommendation services. It supports NVIDIA® Tesla T4 GPU, featuring 8.1 TFLOPS in FP32 and 130 TOPs in INT8 for real-time inference based on trained neural network model.

1-2 NVIDIA TESLA T4

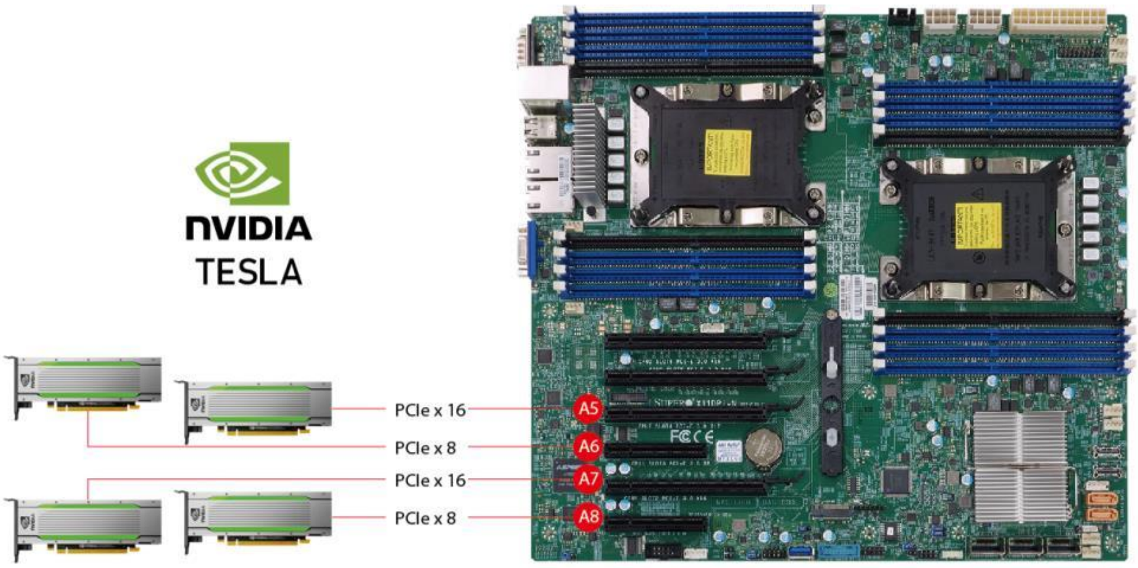
AV3000 Supports 4x NVIDIA Tesla T4 GPU Module; can power the planets most reliable mainstream workstations. Designed into a low-profile, 70-watt package, T4 is powered by NVIDIA Turing Tensor Cores, supplying innovative multi-precision performance to accelerate a vast range of modern applications. AV3000 w/ Tesla T4 GPUs accelerates diverse cloud workloads. These include high-performance computing, data analytics, deep learning training and inference, graphics and machine learning. T4 features multi-precision Turing Tensor Cores and new RT Cores. It is based on NVIDIA Turing architecture and comes in a very energy efficient small PCIe form factor. AV3000 delivers ground-breaking performance at scale.

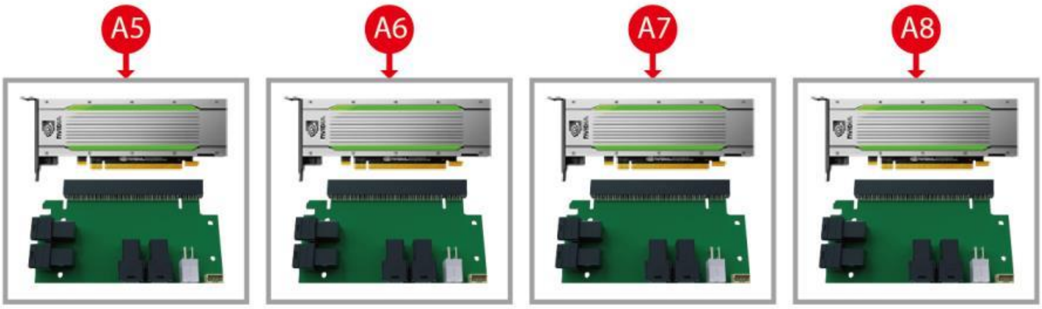




SPECIFICATIONS

GPU Architecture	NVIDIA Turing
NVIDIA Turing Tensor Cores	320
NVIDIA CUDA® Cores	2,560
Single-Precision	8.1 TFLOPS
Mixed-Precision (FP16/FP32)	65 TFLOPS
INT8	130 TOPS
INT4	260 TOPS
GPU Memory	16 GB GDDR6 300 GB/sec
ECC	Yes
Interconnect Bandwidth	32 GB/sec
System Interface	x16 PCIe Gen3
Form Factor	Low-Profile PCIe
Thermal Solution	Passive
Compute APIs	CUDA, NVIDIA TensorRT™, ONNX

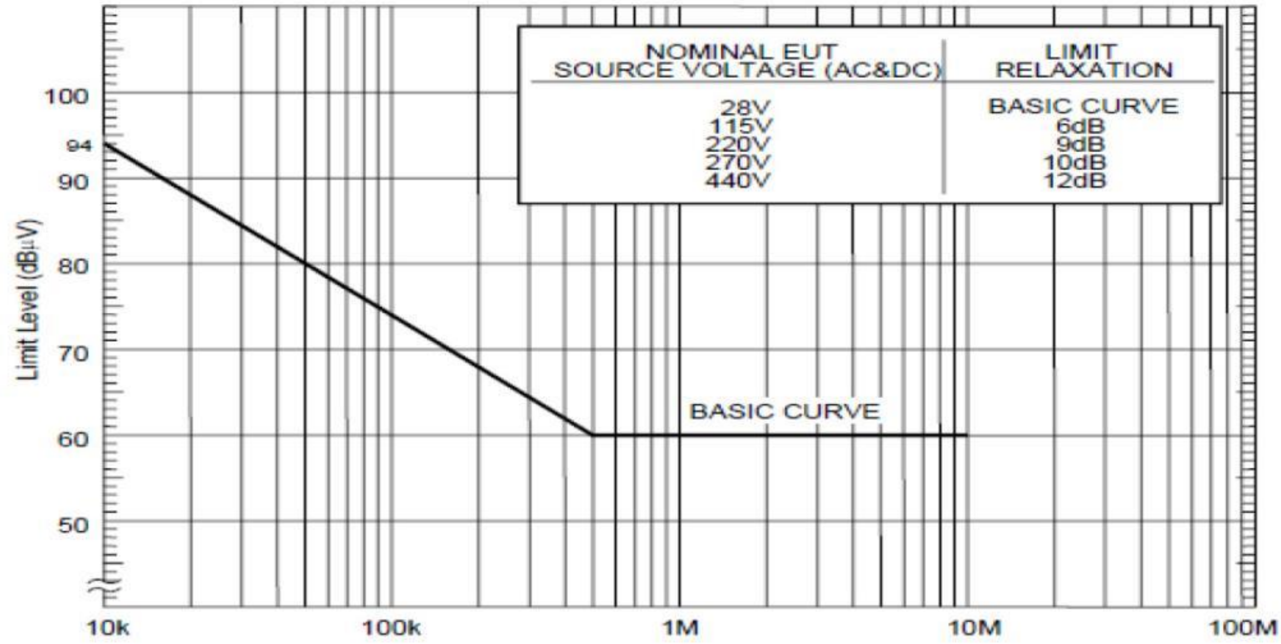




1-3 MIL-STD 461

AV3000 is designed to meet strict size, weight, and power (SWaP) requirements and to withstand harsh environments, including temperature extremes, shock/vibe, sand/dust, and salt/fog.

AV3000 is MIL-461 EMI/EMC compliant rugged Edge AI Inference server. It passes numerous environmental tests including Temperature, Altitude, Shock, Vibration, Voltage Spikes, Electrostatic Discharge and more. It support **18V-36V**DC input, 800W (+60 degree > 89%) DC-DC Efficiency, State-Of-The-Art External **EMI Filter** The sealed compact chassis shields circuit cards from external environmental conditions such as sand, dust, and humidity.



1-4 NVMe – The Foundation for Data - Centric Innovation

1-4-1: Faster Time to Value with 4X Speed

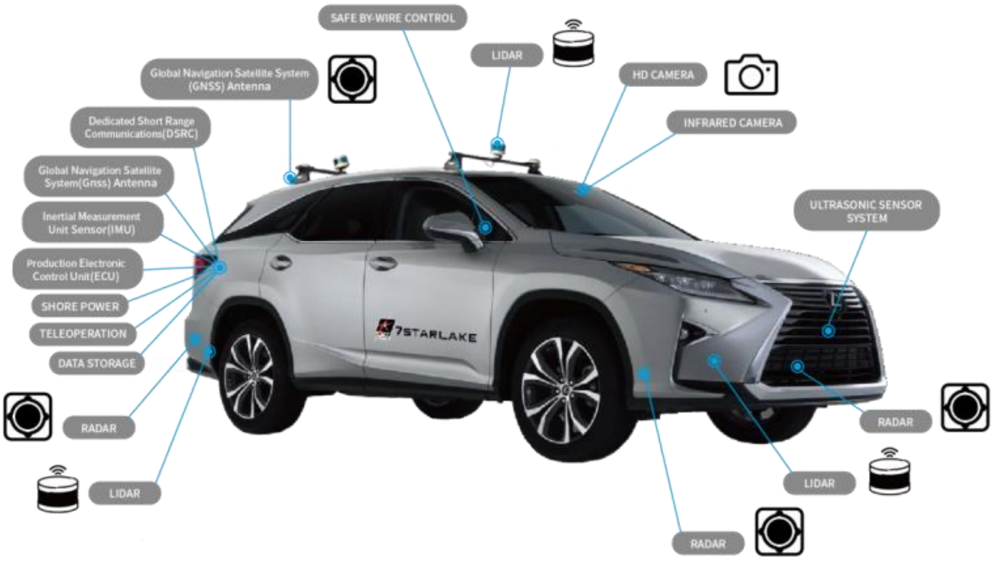
AI applications are inherently data-intensive, with multiple reads and writes to the file system. And, at the outset, the AI algorithm absorbs tremendous amounts of training data as it learns the parameters of its job. Compared to an old-school SATA SSD drive, an NVMe-based drive can write to disk up to **4x faster.** Also, seek times – the time it takes for a drive to locate the area in which a file is stored – are up to 10x faster. It is worth noting that NVMe is not merely fast because it connects via PCIe interfaces. There is also a lot of clever engineering on the drives themselves, particularly pertaining to how it organizes read/write requests.

NVMe-based storage, on the other hand, supports multiple I/O queues, with a theoretical maximum of 64,000 queues, each permitting 64,000 entries for a grand total of 4.096 billion entries. Also, the drive's controller software is designed to create and manage I/O queues. These are intelligently shaped by the system's characteristics and predicted workload, rather than some kind of hard-coded one-size-fits-all solution.

SSD	SATA SSD	M.2 SSD			
Bus Standard	**SATA Revision 3**	**SATA Revision 3**	**PCI-E 3.0x2**		**PCI-E 3.0x4**
Transfer Protocol	**AHCI**	**AHCI**	**AHCI**	**NVMe**	**NVMe**
Bandwidth	**6 Gbps**	**6 Gbps**	**6 Gbps**	**16 Gbps**	**32Gbps**
Transfer Rate	**600 MB/s**	**600 MB/s**	**600 MB/s**	**2 GB/s**	**4 GB/s**

1-4-2: 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.

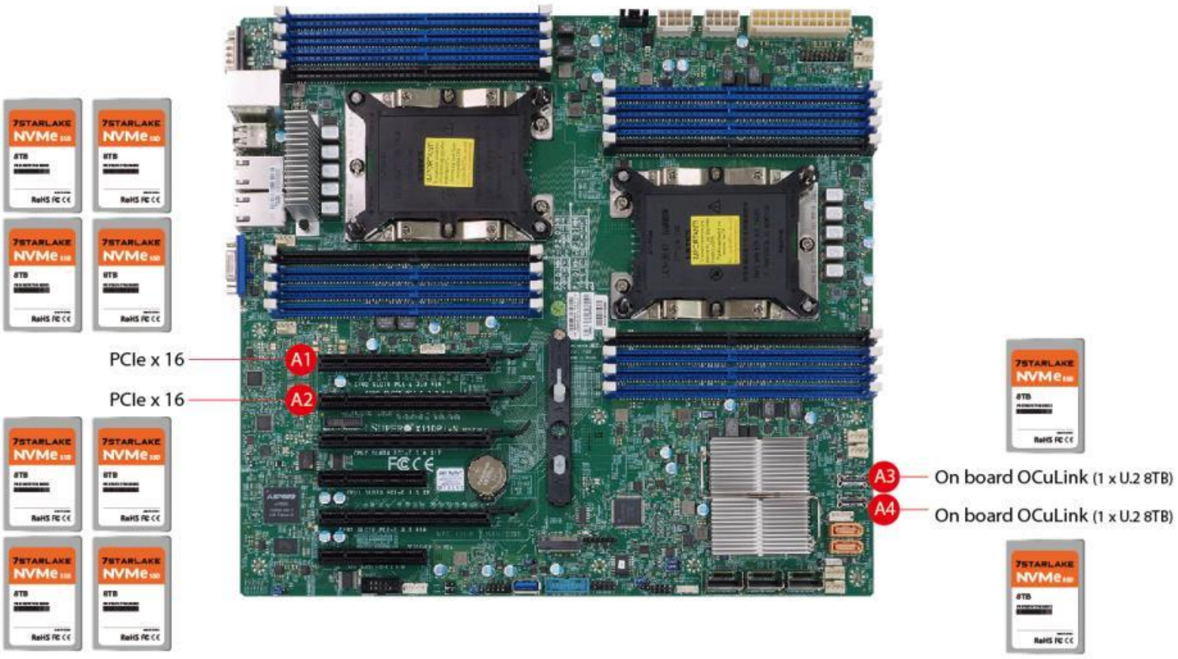

SAFE BY-WIRE CONTROL
LIDAR
Global Navigation Satellite System (GNSS) Antenna
HD CAMERA
Dedicated Short Range Communications(DSRC)
INFRARED CAMERA
Global Navigation Satellite System(Gnss) Antenna
Inertial Measurement Unit Sensor(IMU)
ULTRASONIC SENSOR SYSTEM
Production Electronic Control Unit(ECU)
SHORE POWER
TELEOPERATION
DATA STORAGE
7STARLAKE
RADAR
RADAR
LIDAR
LIDAR
RADAR

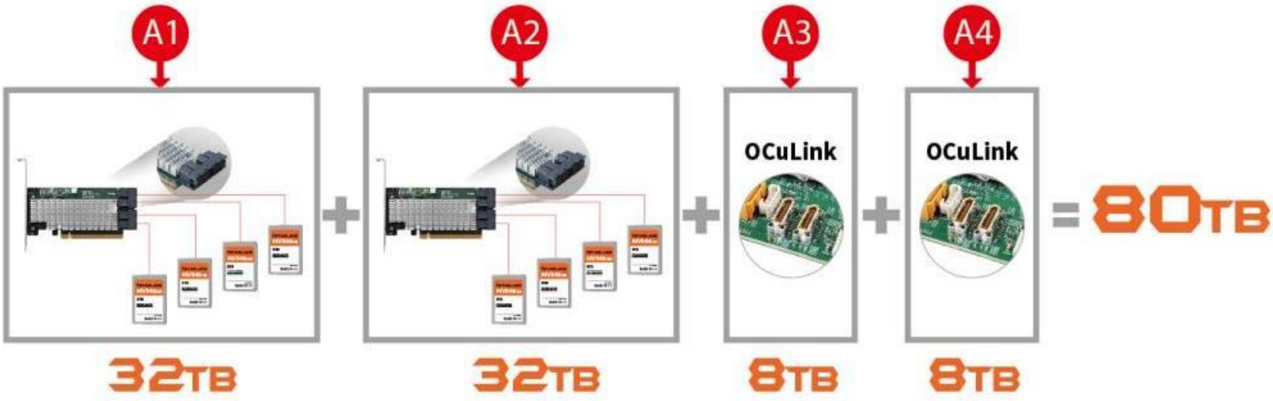
The modern datasets used for model training can be up to terabytes each. Even if the training itself runs from RAM, the memory should be fed from non-volatile storage, which has to support very high bandwidth. In addition, paging out the old training data and bringing in new data should be done rapidly 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.

1-4-3: How AV3000 Support 80TB NVMe

By applying NVMe-based data storage on solid-stage drives, can provide excellent performance such as create and manage large datasets, Overcome Capacity Limitations of Local SSDs, Accelerate epoch time of Machine Learning, Improve Utilization of GPUs.


PCIe x 16
A1
PCIe x 16
A2
A3
On board OCuLink (1 x U.2 8TB)
A4
On board OCuLink (1 x U.2 8TB)
7STARLAKE
NVMe



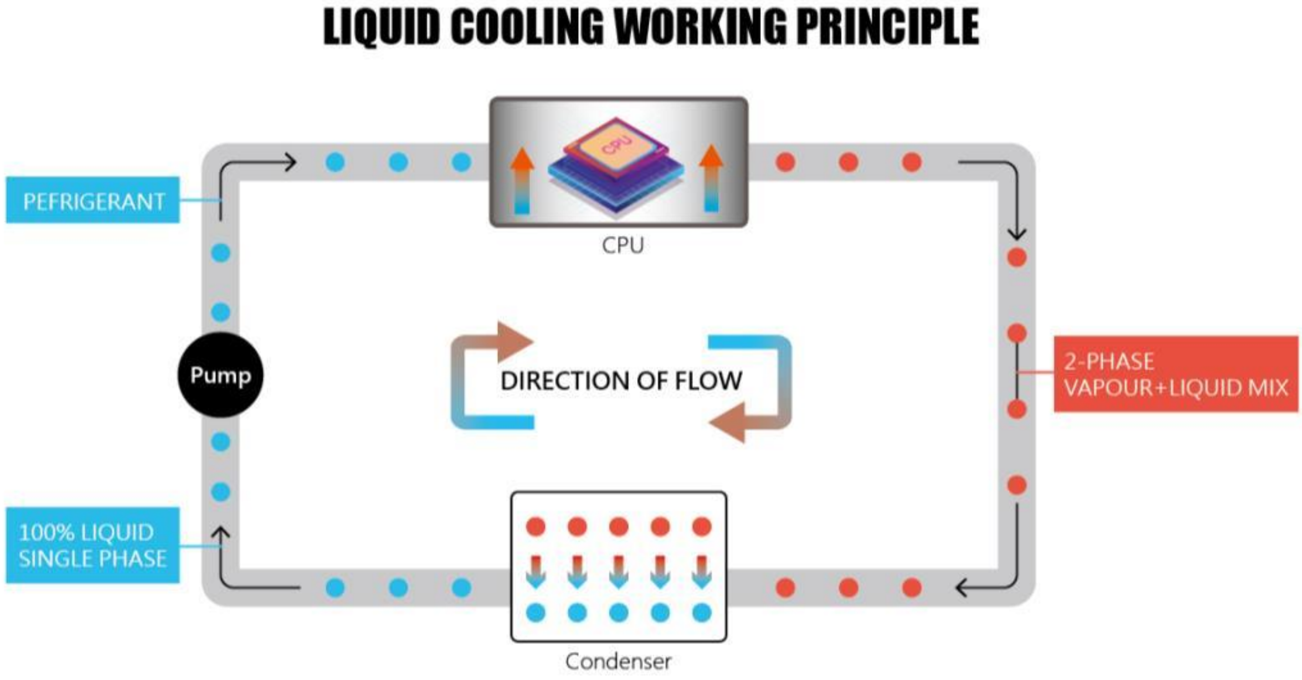
7Starlake is devoted to uncovering the unique way of optimising storage. AV3000 supports 6-slot PCIex16 expansions that provide extreme PCIe Gen 3x4 speed for GPUs at an exceptional value. AV3000 can make the best use of two PCIe x16 host connections that associate with eight PCIe gen 3x4 U.2 NVMe. Two extended card are individually installed on A1 and A2 slot (Please refer to the demonstration above); each of them has four ports that can total support 32TB NVMe;(8TB U.2 SSD x 4 Ports x 2 PCIe16 Slot). In other words, one port is able to back 8TB NVMe. In addition to the 8 x PCIe U.2 NVMe, there are other two on board OCuLink(labelled as A7 and A8) that are capable of offering 8TB each. As a result, AV3000 is able to surmount the limitations of traditional SSDs and supply large data storage.

1-5 Liquid Cooling

To enhance way of cooling CPU and GPU, 7Starlake 's newest rugged platform is designed to accommodate boards requiring liquid cooling. AV3000 uses the exceptional thermal conductivity of liquid to provide dense, concentrated cooling to targeted areas. By using LC, the dependence on fans and expensive air handling systems is drastically reduced. These results in much higher rack density, overall reduced power use, enhance higher level of efficiency, cools off high-performance GPU and the sound of silence and improve overclocking potential.

The backplane is designed to handle total 300 Watts caused from Dual XEON Processor per plate. AV3000 establishes the mechanical design interface, outline and mounting requirements. Liquid through cooled plug-in pipe within associated sub-racks. While the connector layout remains common with AV3000.

AV3000 are cooled by liquid flowing through an integral pipe, boards and electronic components are cooled more effectively as compared to other more traditional cooling methods such as air or conduction cooling alone. Quick disconnect coupling assemblies allow fluidic coupling to the chassis manifold.



2. AV3000 Specifications

System

Processor	2 x 3rd Gen. Intel® Xeon® Scalable Platinum 8380 CPU(2.3 GHz, 40Cores, TDP 270W)
LAN	Dual LAN with 1GbE LAN with Intel® X722
Memory	768 GB DDR4-RDIMM
Storage	80 TB NVMe

4 x TESLA GPU Card

NVIDIA	TESLA T4
Tensor Core	320
CUDA Cores	2560
Memory	16GB GDDR6, 300 GB/Sec.
GPU 1	PCIex16 – A3 Slot
GPU 2	PCIex 8 – A4 Slot
GPU 3	PCIex16 – A5 Slot
GPU 4	PCIex 8 – A6 Slot

Storage Configurations

4x 8TB U.2 = 32TB	PCIe x 16 – A1 (4xPorts PCIe U.2 Riser Card)
4x 8TB U.2 = 32TB	PCIe x 16 – A2 (4xPorts PCIe U.2 Riser Card)
1x 8TB U.2 = 8TB	On Board OCuLink
1x 8TB U.2 = 8TB	On Board OCuLink

PSU

MIL-461 800W DC-DC 18V~36V Power module

Physical

Dimension	416.2 x 206.4 x 506.4mm (W x D x H)		
Weight	15KG	Finish	Anodic aluminum oxide
Chassis	Aluminum Alloy, Corrosion Resistant.		

REAR I/O

LAN	2 x RJ45 Gigabit Etherne
USB	4 x USB 2.0 ports (2 rear + 2 via headers) 5 x USB 3.2 Gen1 ports (2 rear + 2 via headers + 1 Type A)
VGA	1 x VGA
COM	2x COM(RS232)

EXPANSION SLOT (BY REQUEST)

SATA connector x2 available

iPass cable connector x3

NvME slot x1

MIL COMPLIANCE

MIL-STD-810G (Operation Test)

Low Temp.	Method 502.5 Procedure 2	Exposure(24H x 3 cycle) at -10℃ min.
High Temp.	Method 501.5 Procedure 2	60ºC for 2 Hrs after temperature stabilization.
Humidity	Method 507.5 Procedure 2	RH -95%. Test cycles: ten 24-hrs , functional test after 5th and 10th cycles
Vibration	Method 514.6 Category 20	10-500Hz 1.04Grms Test duration: 1 Hr x 3 axis (total 3 Hrs)
Shock	Method 516.6 Procedure 1	20G, 11mSec, 3 per axis

MIL-STD-810G (Non-Operating Tests)

Low Temp.	Method 502.5	Exposure(24H x 7 cycle) at -20℃ min.
High Temp.	Method 501.5 Procedure 1	71ºC for 2 Hrs after temperature stabilization.
Vibration	Method 514.6 Category 24	200 to 2000Hz Test duration: 1hr per axis; rms = 7.7 Gs
Shock	Method 516.6 Procedure 1	20G, 11mSec, 3 per axis

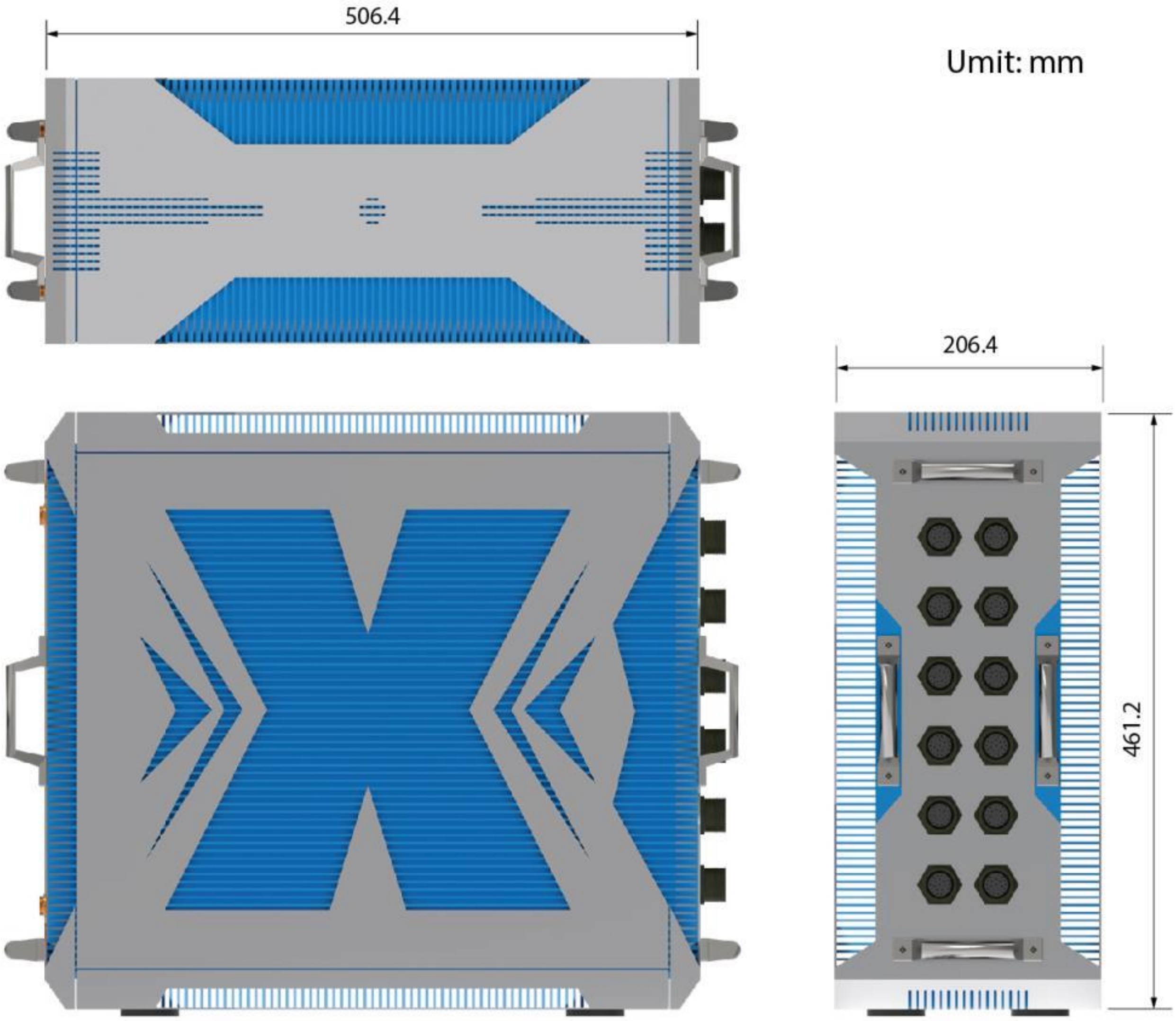
MIL-STD-461E

CE102	Basic curve, 10kHz - 30 MHz
RE102-4, (1.5 MHz)	(1.5 MHz) -30 MHz - 5 GHz
RS103	1.5 MHz - 5 GHz, 50 V/m equal for all frequencies EN 61000-4-2: Air discharge: 8 kV

ENVIRONMENTAL QUALIFICATIONS

Regulatory	CE ,FCC Compliance
Operation Temp.	-20~+50°C
Storage Temp.	-40~+85 °C
Green Product	RoHS, WEEE compliance

3. Dimension


506.4
Umit: mm
206.4
461.2