BIOS ANNA A100

Featuring the latest generation
NVIDIA A100™ Tensor Core GPUs

BIOS IT 's latest iteration of its Artificial Neural Network Accelerator (ANNA) series is built on Supermicro Hardware and features the latest NVIDIA® A100™ Tensor Core GPUs.

The ANNA A100 is designed for the most demanding AI workloads and is optimised for the new HGX™ A100 4-GPU baseboard. With the newest version of NVIDIA® NVLink™ and NVIDIA NVSwitch™ technologies, these servers can deliver up to 5 PetaFLOPS of AI performance in a single system. The system supports PCI-E Gen 4 for fast CPU-GPU connection and high-speed networking expansion cards.

NVIDIA A100 TENSOR CORE GPU

The NVIDIA A100 Tensor Core GPU delivers unprecedented acceleration at every scale for AI, data analytics, and high-performance computing (HPC) to tackle the world's toughest computing challenges. As the engine of the NVIDIA data center platform, A100 can efficiently scale to thousands of GPUs or, with NVIDIA Multi-Instance GPU (MIG) technology, be partitioned into seven GPU instances to accelerate workloads of all sizes. And third-generation Tensor Cores accelerate every precision for diverse workloads, speeding time to insight and time to market.

NVIDIA HGX A100

As a balanced data center platform for HPC and AI applications, the new ANNA A100 system leverages the NVIDIA HGX A100 4 GPU board with four direct-attached NVIDIA A100 Tensor Core GPUs using PCI-E 4.0 for maximum performance and NVIDIA NVLink for high-speed GPU-to-GPU interconnects. This advanced GPU system accelerates compute, networking and storage performance with support for one PCI-E 4.0 x8 and up to four PCI-E 4.0 x16 expansion slots for GPUDirect RDMA high-speed network cards and storage such as InfiniBand HDR, which supports up to 200Gb per second bandwidth.

GET IN TOUCH

AMERICAS
1-800-654-BIOS

123 10th St.
San Francisco, CA 94103.

EMEA
+44 (0) 203 178 6467

Salisbury House, 29 Finsbury Circus,
London, EC2M 5QQ

APAC
+61 (0)2 8866 3343

Suite 701,275 Alfred St.,
North Sydney, NSW 2060 Australia

www.bios-it.com | sales@bios-it.com

DEEP LEARNING TRAINING

NVIDIA A100's third-generation Tensor Cores with Tensor Float (TF32) precision provide up to 20X higher performance over the prior generation with zero code changes and an additional 2X boost with automatic mixed precision and FP16.

DEEP LEARNING INFERENCE

A100 introduces groundbreaking new features to optimize inference workloads. It brings unprecedented versatility by accelerating a full range of precisions, from FP32 to FP16 to INT8 and all the way down to INT4.

HIGH PERFORMANCE COMPUTING

A100 introduces double-precision Tensor Cores, providing the biggest milestone since the introduction of double-precision computing in GPUs for HPC. This enables researchers to reduce a 10-hour, double-precision simulation running on NVIDIA V100 Tensor Core GPUs to just four hours on A100

HIGH PERFORMANCE DATA ANALYTICS

Accelerated servers with A100 deliver the needed compute power—along with 1.6 terabytes per second (TB/sec) of memory bandwidth and scalability with third-generation NVLink and NVSwitch—to tackle these massive workloads.

ENTERPRISE READY UTLISATION

A100 with MIG maximizes the utilization of GPU-accelerated infrastructure like never before. MIG allows an A100 GPU to be partitioned into as many as seven independent instances, giving multiple users access to GPU acceleration for their applications and development projects.

BIOS ANNA A100	
CPU	Dual AMD EPYC™ 7002 Series Processors
FORM FACTOR	2U
MEMORY	8TB Registered ECC DDR4 3200MHz SDRAM in 32 DIMMs
GPU	Supports 4 NVIDIA A100 GPUs
STORAGE	4 Hot-swap 2.5" drive bays (SAS/SATA/NVMe Hybrid)
EXPANSION SLOTS	4 PCI-E Gen 4 x16 (LP), 1 PCI-E Gen 4 x8 (LP)
PSU	2x 2200W Redundant Power Supplies, Titanium Level + 4 Hot-swap heavy duty fans

All products and companies referred to herein are trademarks or registered trademarks of their respective companies or mark holders.