SCALE UP YOUR HPC, AI, AND ML WORKLOADS WITH SUPERMICRO SUPERBLADE

Put up to 1.68 petaFLOPs per rack of supercomputing capacity into your data center with our performance and density-optimized solution with a resource-saving architecture

Enterprise data centers have long considered on-premises supercomputing capabilities a dream that is out of reach. Today, dreams can become reality with the Supermicro® A+ SuperBlade® platform powered by AMD EPYC™ processors and AMD Instinct™ MI100 accelerators. Equipped with the highest-performing x86 server CPU¹ and the world’s fastest HPC GPU², you can run your AI/ML applications and reduce turnaround time with high 16-bit floating-point (FP16) matrix operations, speeding AI model training and inferencing. You can run high-performance computing (HPC) workloads and accelerate run times for processes such as EDA simulations with fast 64-bit floating-point (FP64) arithmetic. Even traditional enterprise workloads including database management systems and virtual desktop environments benefit from the computing density the platform delivers.

Supercomputing Within Reach

A single 48U rack of six SuperBlade platforms can support up to 120 SuperBlade nodes. When powered by 64-core AMD EPYC 7763 processors, you have up to 300 teraFLOPs of maximum theoretical computing power at your service. When you configure an AMD Instinct MI100 GPU in each node, you add 1380 teraFLOPs of maximum theoretical 64-bit floating-point performance per rack. The stunning total: 1.68 petaFLOPs per rack to speed your most compute-intensive AI, ML, and HPC workloads to completion. The SuperBlade Node prepares you for the future by being ready to accommodate the AMD Instinct MI210 GPU when available.

This density-optimized solution includes 200-Gbps InfiniBand for cluster connectivity and 25 Gigabit Ethernet switching for traditional IP traffic. The platform’s resource-saving architecture significantly reduces initial and operational expenses through shared power and cooling, network management, and unparalleled node density.
We deliver the broadest portfolio of the most high-performing, reliable servers with hyper-dense configurations that don’t sacrifice the versatility you need to get your work done.

### FOOTNOTES

1. SPECjbb®2015-Distributed critical-jOPS comparison based on highest system results published as of 01/01/2021. Configurations: 20 node, 1x AMD EPYC 7763 (2561044 SPECjbb2015-Distributed critical (OPS, 2919887 SPECjbb2015-Distributed max (OPS), https://spec.org/jbb2015/results/2919887 SPECjbb2015-20210226-00533.html) versus 2-node, 1x AMD EPYC 7702P (1877397 SPECjbb2015- Distributed critical) (OPS, 255878 SPECjbb-Distributed max (OPS), https://spec.org/jbb2015/results/255878 SPECjbb2015-20200402-00531.html) for ~1.36x the performance. MLNWR-115

2. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak double precision (FP64), 46.1 TFLOPS peak single precision matrix (FP32), 23.1 TFLOPS peak single precision (FP32), 184.6 TFLOPS peak half precision (FP16) peak theoretical, floating-point performance. Published results on the NVidia Ampere A100 (40GB) GPU accelerator resulted in 9.7 TFLOPS peak double precision (FP64), 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16) theoretical, floating-point performance. Server manufactures may vary configuration offerings yielding different results. MI100 03

For workloads including training, inferencing, and HPC, the key getting more work done in a day is the AMD Instinct MI100 GPU Accelerator. The AMD Instinct MI100 GPU is AMD’s most advanced accelerator with all-new AMD CDNA™ architecture with Matrix Core technology that offers superior performance for a full-range of mixed-precision operations. It’s the world’s fastest HPC accelerator with up to 11.5 TFLOPs of FP64 performance. Sixteen lanes of PCI-E 4.0 connectivity links the GPU with the CPU for maximum I/O bandwidth. Ultra-fast HBM2 GPU memory helps eliminate bottlenecks, with 1.2 TB/s of GPU memory bandwidth to support your largest data sets. The AMD ROCm™ open software platform supports the libraries your software needs, with optimized versions of PyTorch and TensorFlow. Update your software to use the ROCm platform once, and you can run it anywhere.