tenstorrent

Galaxy

Scalability • Performance • Density

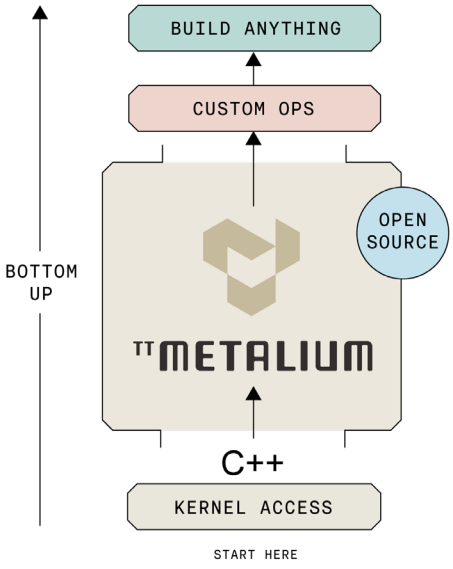
Built for scale-out and maximum performance at low total cost of ownership, Tenstorrent's Galaxy Wormhole Server utilizes Tenstorrent's novel processor architecture comprised of 32 Wormhole™ Tensix Processors. The Wormhole™ ASIC is designed to scale, featuring a Network-on-Chip (NoC) that offers 3.2 Tbps of Ethernet connectivity to surrounding chips, configurable to balance between a massive multi-chip mesh and flexible multi-tenant clusters. Each Galaxy features a cumulative 9.32 PetaFLOPs (FP8) of compute performance, 3.8GB of pooled on-die SRAM, and 384GB of globally accessible GDDR6 memory, enabling performant AI inferencing on demanding LLMs or CV models. Tenstorrent's Ethernet-based interconnect enables a mesh, so you can combine as many Galaxy Servers as desired to power your universe of computing needs.

Specifications

	Wormhole ASIC	Galaxy Module	Galaxy Wormhole Server
Units per Server	32	4	1
Tensix Cores	80 @ 1GHz	640 @ 1GHz	2,560 @ 1GHz
TFLOPs (FP8)	292 TFLOPs	2.3 PetaFLOPs (2336 TFLOPs)	9.32 PetaFLOPs (9322 TFLOPs)
TFLOPs (FP16)	82 TFLOPs	656 TFLOPs	2.62 PetaFLOPs (2621 TFLOPs)
TFLOPs (BFP8)	164 TFLOPs	1.3 PetaFLOPs (1312 TFLOPs)	5.24 PetaFLOPs (5243 TFLOPs)
SRAM	120MB	960MB	3.8GB
GDDR6 Memory	12GB @ 12 GT/sec	96GB @ 12 GT/sec	384GB
Power	200W	1.6 kW	7.5 kW

Application Portability

Tenstorrent's Tensix Processors are supported by the TT-Metalium™ open source SDK, providing direct access to the metal. Code written on a single or smaller mesh of Wormhole™ processors can be easily ported to the larger pool of resources in Galaxy.



Flexible Precision Support

Wormhole Tensix Processors at the heart of the Galaxy Wormhole Server support a broad range of data formats, including highly efficient block floating point (BFP) precision. BFP offers most of the precision of conventional floating point formats while requiring just half the bandwidth and storage ensuring you can run your models with the accuracy and throughput you demand with a minimum of compute.

Floating point	FP8, FP16, BF16, FP32*
Block floating point	BFP2, BFP4, BFP8
Integer	INT8, INT32*
Unsigned integer	UINT8
TensorFloat	TF32
Vector	VFP32, VTF19

*Output only.

For additional specifications, hardware and software compatibility, and volume pricing, contact Tenstorrent at sales@tenstorrent.com