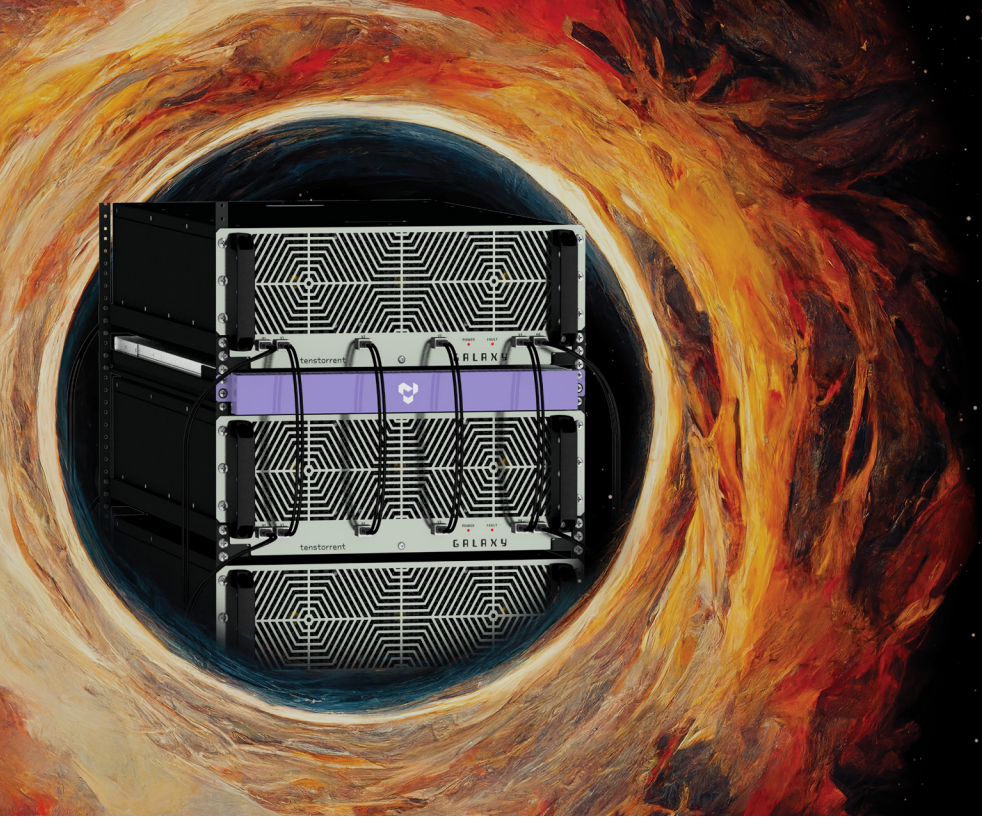


tenstorrent

Native AI scalability,
from the edge to

GALAXY

Tenstorrent's second-generation processor, Wormhole, builds off its high-performance predecessor with the addition of on-chip high-bandwidth Ethernet and switch. Using standard Ethernet, Tenstorrent's Wormhole products natively network together to enable the arbitrary scaling of computing resources without re-programming your model or infrastructure.

Each chip has sixteen 200Gb Ethernet ports around its edge (totaling 3.2Tb of chip-to-chip bandwidth), allowing for the extension of our Network-on-Chip to as many compute nodes as required. Tenstorrent's TT-Buda SDK automatically recognizes these additional devices to take full advantage of the available resources.

Wormhole chips are deployed on the Galaxy Modules, all part of the Tenstorrent pre-configured, rack-mounted Galaxy system. Galaxy systems are built with a pre-tested Ethernet backplane and deliver dense, high-performance AI compute with full binary equivalency to our smaller systems; this allows organizations to train models efficiently on a large-scale system and deploy fully-portable code to edge inference devices.

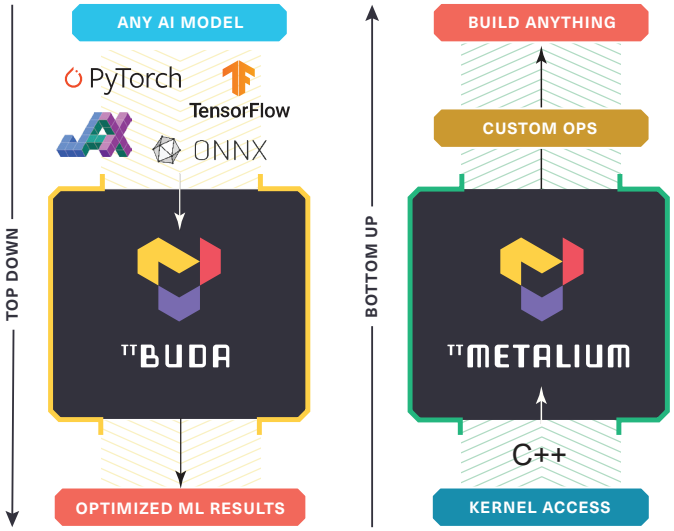
Scalability: Tenstorrent's hardware includes onboard Ethernet to enable high bandwidth chip-to-chip connectivity. With native support in our compiler, adding additional compute is as easy as installing another device. No special networking or configuration required.

Galaxy Module	164 TOPS at BFP8
	3.2Tbps Ethernet (16 x 200Gbps)
	12GB GDDR6
	200W per module
Galaxy Server	4U form factor
	32 Galaxy Modules
	5.2 PetaOPS at BFP8
	41.6 Tbps internal connectivity
	384GB of globally accessible GDDR6 memory
	3.8GB SRAM
	7.5 kW per server
Galaxy Rack	48U form factor
	8 Galaxy 4U enclosures
	42 PetaOPs at BFP8
	I/O up to 76.8 Tbps
	More than 3TB of GDDR6 memory

Our Galaxy's Edge:

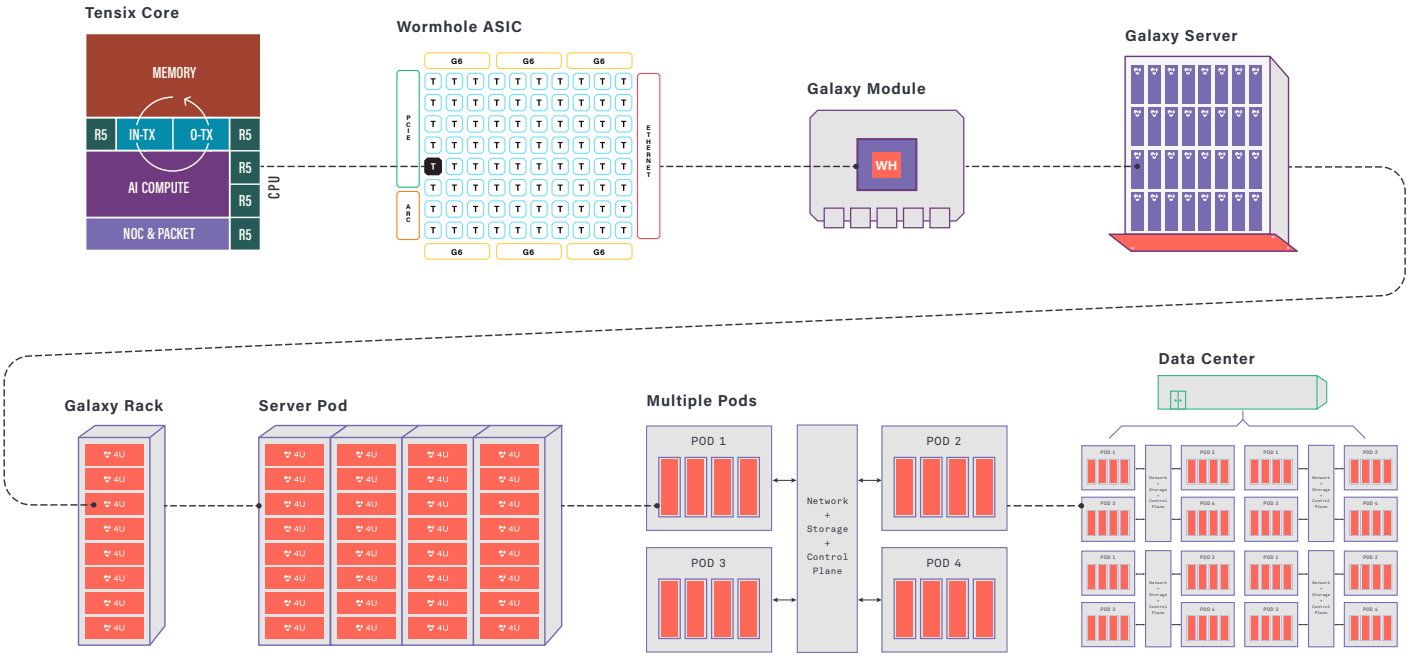
Power: Our devices deliver high performance at relatively low power, starting at 150W, while high-density configurations can maximize performance per-square-foot.

Ease of Code/Application Portability: Tenstorrent's TT-Buda SDK allows users to compile code from common ML frameworks like PyTorch or TensorFlow directly and abstracts the underlying hardware, while the TT-Metallium SDK provides low-level hardware access, enabling use of Python and C++ for both AI and non-AI workloads.



Cost vs Alternative(s): Tenstorrent is driving down the cost of performance per dollar by designing hardware specifically for AI/ML applications, while leveraging commodity components, which keeps the build costs as low as possible.

Faster Compute = Accelerated Science: Support for a range of data types from FP32 to BFP4 allows us to deliver high performance on a range of precisions.



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