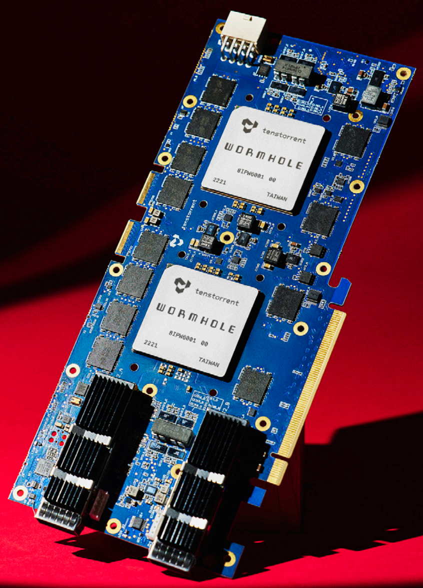
tenstorrent

AI Accelerators



The Tenstorrent e-series and n-series machine learning accelerators are powered by our Grayskull™ and Wormhole graph processors, architected from the ground up to meet the needs of current and future artificial intelligence models. These ASICs are engineered around a forward-thinking, open, and cost-effective approach to artificial intelligence processing, and the fundamental building block of each graph processor is the Tensix Core.

Each Tensix Core incorporates a cache of local scratch pad SRAM, five "baby RISC-V" microprocessors, Matrix and Vector Engines, and dedicated hardware streams built upon ethernet protocols that facilitate rapid core-to-core and chip-to-chip communication. The net result is a mesh of highly flexible machine learning cores supporting a broad range of precision formats, able to scale in concert with ever expanding models and evolve with the industry.

Key to maximizing performance, flexibility, and efficiency is Tenstorrent's open approach. The high-level TT-Buda SDK enables users and organizations to quickly implement their models on Tenstorrent hardware, while the low-level TT-Metalium SDK is open source and geared toward programming as close to the metal as possible.

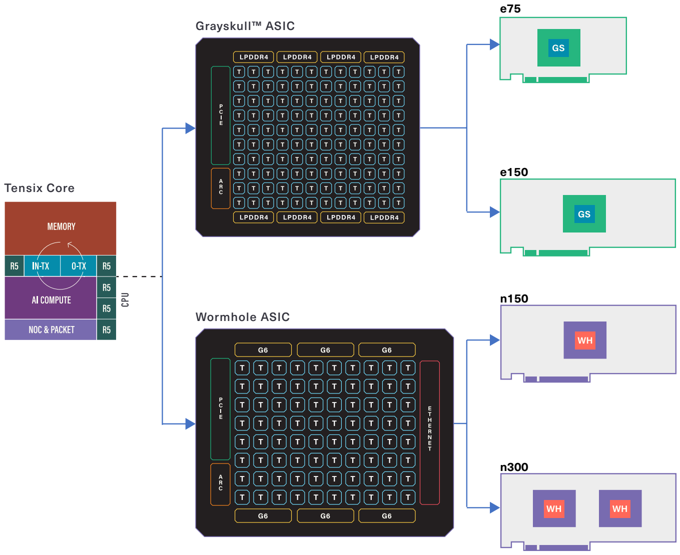
Tenstorrent's graph processors are designed to provide this scalable, flexible feature set in a cost-effective fashion by manufacturing on a mature, less costly process while employing a memory hierarchy able to take advantage of commodity memory technologies instead of expensive, exotic solutions. The e-series and n-series machine learning accelerators offer an entry point for organizations to familiarize themselves with Tenstorrent's open, novel architecture.

Comparison Chart

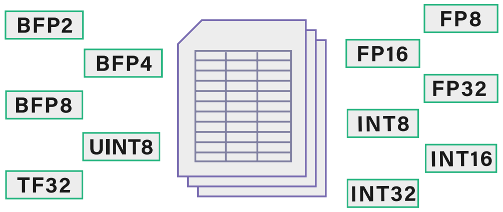
Card	e75	e150	n150	n300
ASIC	Grayskull™	Grayskull™	Wormhole	2x Wormhole
Tensix Cores	96	120	72	128
AI Clock	1 GHz	1.2 GHz	1 GHz	1 GHz
SRAM	96MB	120MB	108MB	192MB
Memory Capacity	8GB	8GB	12GB	24GB
Memory Type	LPDDR4	LPDDR4	GDDR6	GDDR6
Memory Bandwidth	102.4 GB/sec	118.4 GB/sec	288 GB/sec	288 GB/sec
TFLOPs (FP8)	221	332	262	466
Interface	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16
Total Board Power	75W	200W	160W	300W
Cooling	Active	Passive*	Passive*	Passive*
Form Factor	HHHL Single Slot	FHFL Dual Slot	FHFL Dual Slot	FHFL Dual Slot

* Active Cooling Kit available separately.

Smart & Scalable From the Ground Up: The Tensix Core is the foundation of the Grayskull™ and Wormhole ASICs. It is designed specifically for AI/ML applications, incorporating spacious SRAM and a Network-on-Chip design to build out a mesh able to intelligently process and move data while leveraging commodity components, keeping build costs low.

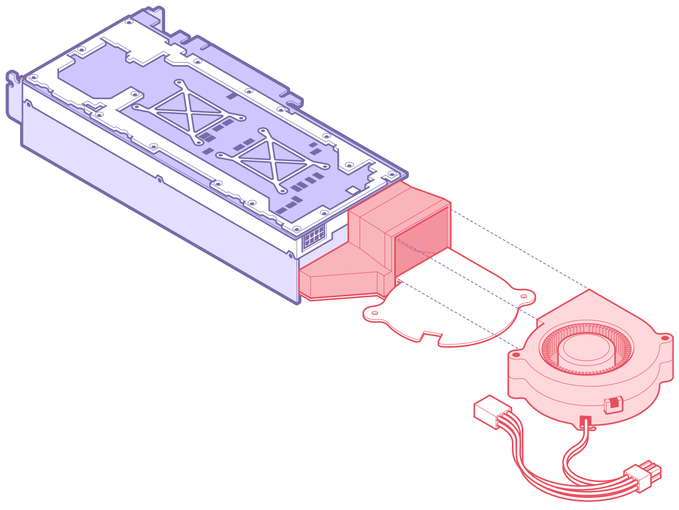


Flexible Precision Support: Tenstorrent's Tensix Cores support a broad range of data types, including highly efficient block floating point (BFP) precision. BFP offers most of the precision of conventional floating point formats while requiring just half the bandwidth and storage.

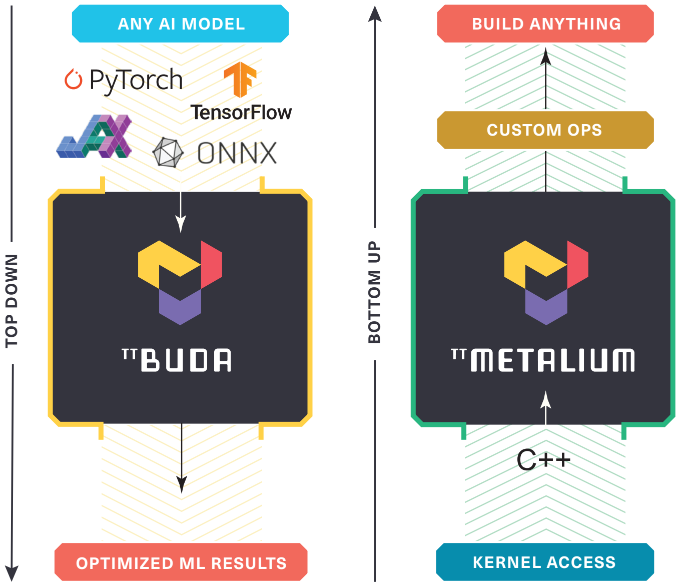


	Grayskull™	Wormhole
Floating point	FP8, FP16, BF16	FP8, FP16, BF16, FP32
Block Floating Point	BFP2, BFP4, BFP8	BFP2, BFP4, BFP8
Integer	-	INT8, INT16, INT32
Unsigned Integer	-	UINT8
TensorFloat	-	TF32

Flexible Cooling: The Tenstorrent e75 ships with a blower fan and is geared as a one-stop solution for getting started with Tensix Core architecture. For users who want to step up to the e150, n150, and n300, Active Cooling Kits are available for workstation use.



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-Metalium SDK provides low-level hardware access, enabling use of Python and C++ for both AI and non-AI workloads.



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