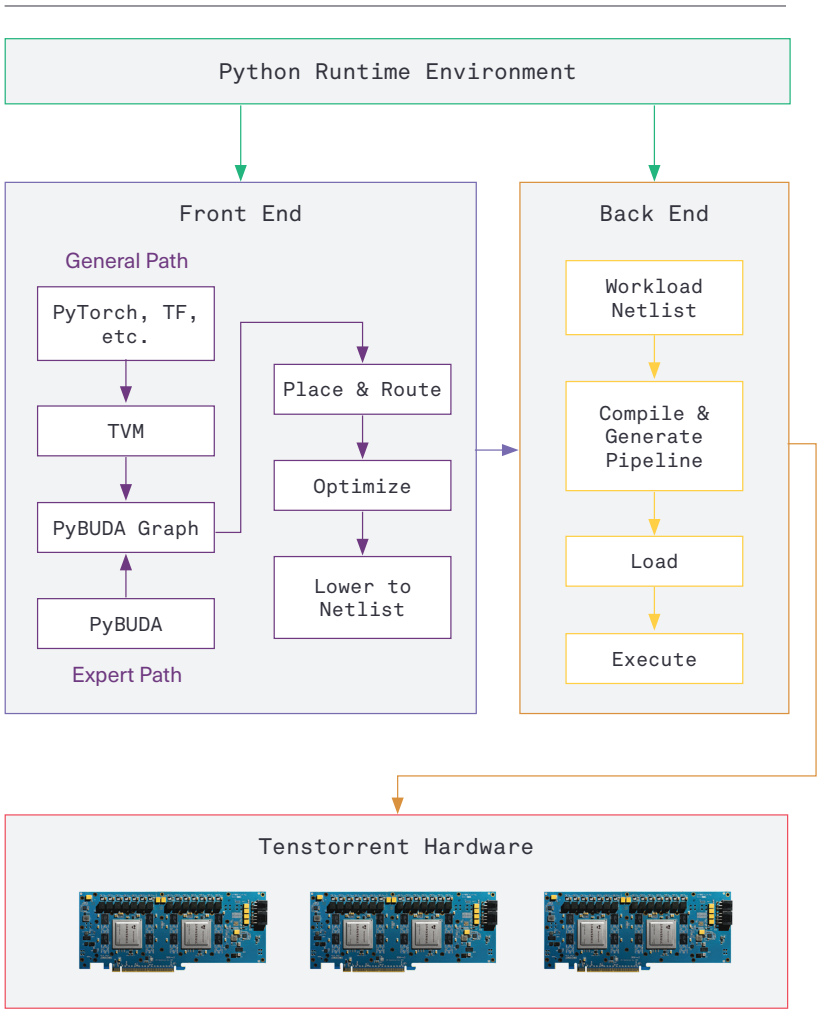
tenstorrent

Open-Source TT-Buda API

Tenstorrent uses a unique approach to machine learning by converting the model to a graph and pipelining it across the tensix cores vs. batch processing commonly found in GPUs. Maximizing efficiency with sparsity combined with a revolutionary approach to conditional compute and ultra scalable data movement, Tenstorrent offers an industry-leading solution combining both inference and training on a single high-performing, lower power chip. Our open source API, TT-Buda, allows unprecedented access to the hardware and programmable cores inside the architecture, enabling model-specific optimizations not possible on other architectures.

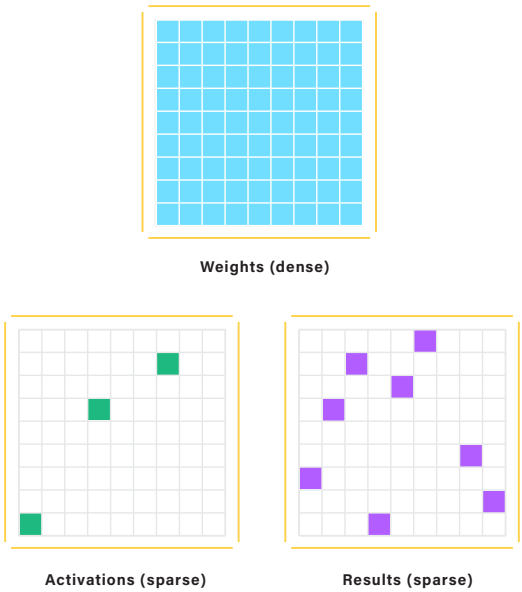
OUR GOAL

Our goal is to run the top 20-25 Hugging Face models “good enough” out of the box. That means exceptional performance with additional optimization available through our API.

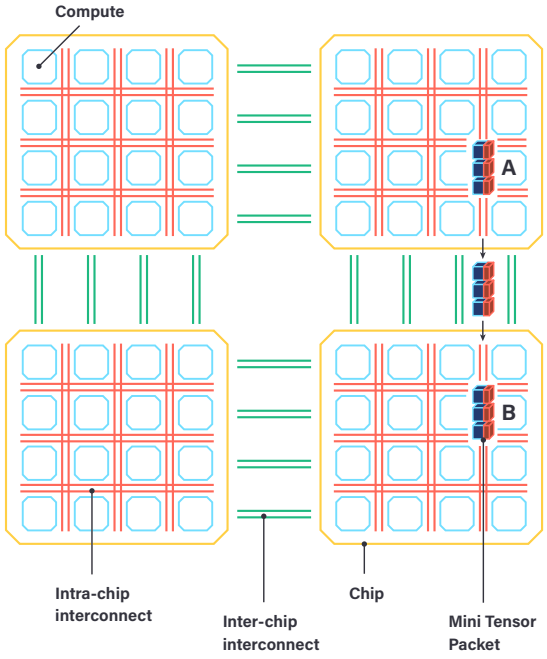


How we win at AI/MI

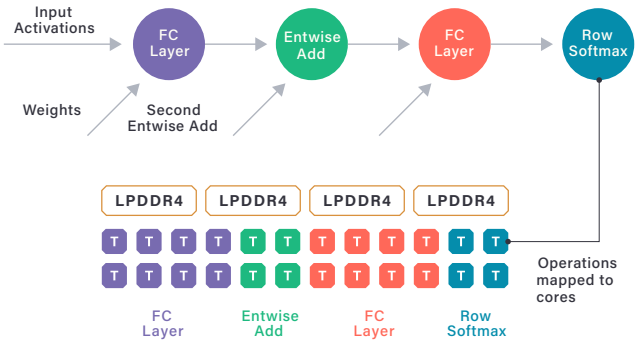
Conditional Execution: Don't waste your cycles. Process only what is needed for accuracy.



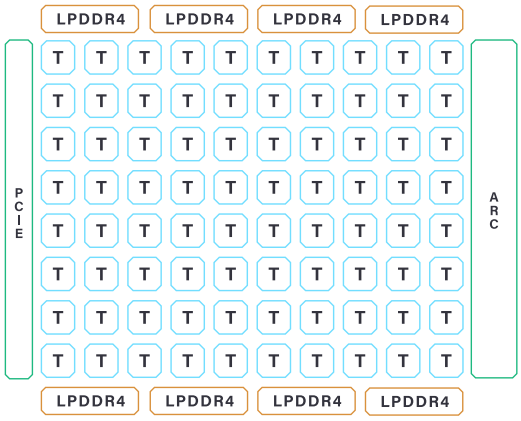
Native Scaling: Don't get hung up on interconnect. Build a computer as big or as small as you like, without the hassle.



On-Chip Dataflow & Near-Memory Compute: Spatial mapping of tensor operations in single batches enables full hardware utilization, no matter the model size.



Commodity Components: Spatial mapping eliminates the need for high memory bandwidth. Lower cost builds drive lower cost products, without sacrificing quality or performance.



Wide Set of Models: Tenstorrent makes the only AI/ML accelerator that has 70+ models up and running without the need for expensive and difficult to find ML experts.

TT-BUDA GPT-2 Hugging Face BERT

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