

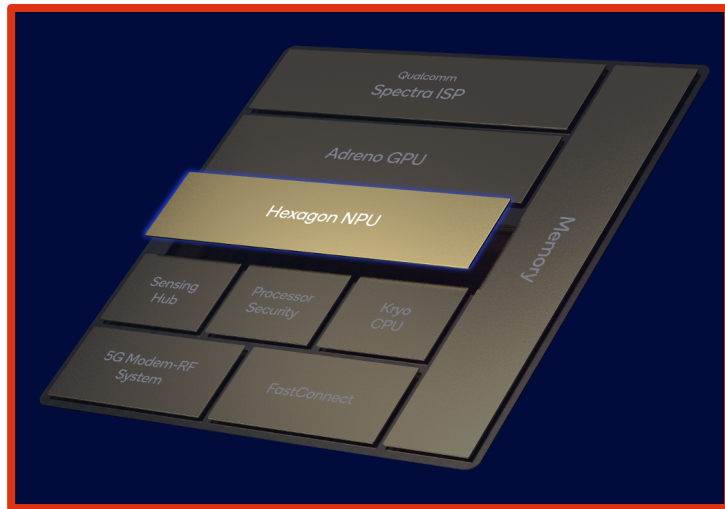
# Qualcomm® Hexagon™ NPU Backend for Triton



Enable mapping for Edge AI and Cloud AI

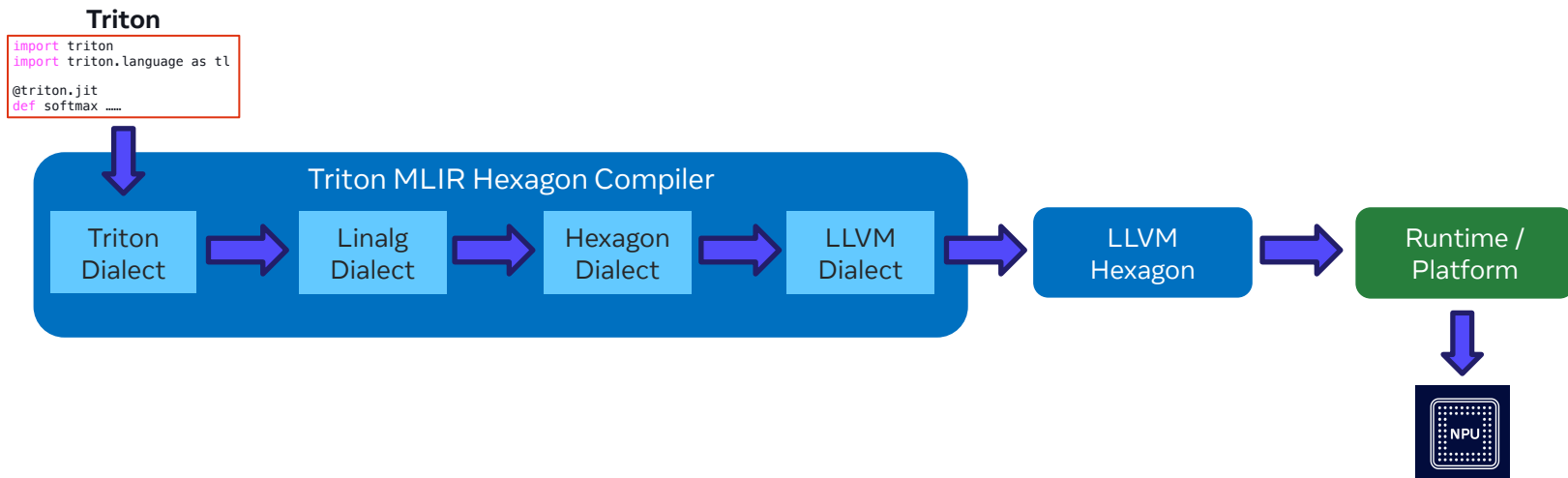
Qualcomm® Hexagon™ NPUs


- 4-way multi-threaded VLIW
- Hexagon Vector eXtensions (HVX)
- Vector registers and dedicated memory
- Tensor units



## TRITON WORKFLOW FOR HEXAGON NPUS

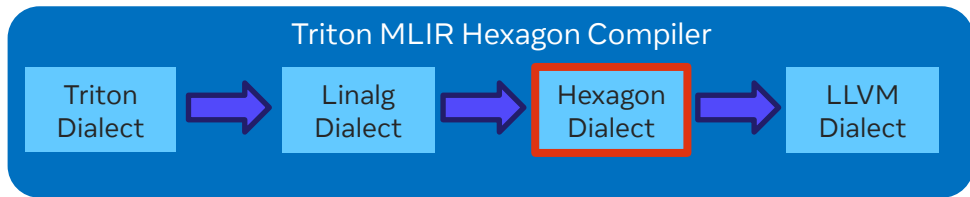
Approach: Leverage upstream Triton and MLIR developments in addition to building downstream target-specific optimizations



- Built by extending an MLIR-based compiler developed for Hexagon NPUs
- Uses 'Triton shared middle-layer' for doing Triton  Linalg conversion

**Hexagon Dialect** handles optimizations for Hexagon NPUs

- ❑ Multi-core multi-threaded parallelism
- ❑ Vectorization
- ❑ Mapping to tensor units
- ❑ Memory optimizations

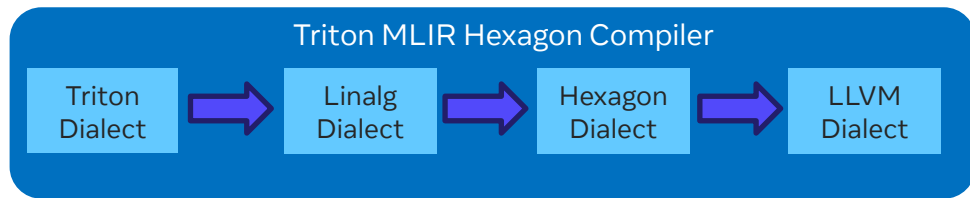


Successfully mapped key kernels such as matmul, flash attention, softmax, layer norm, and others, using this workflow

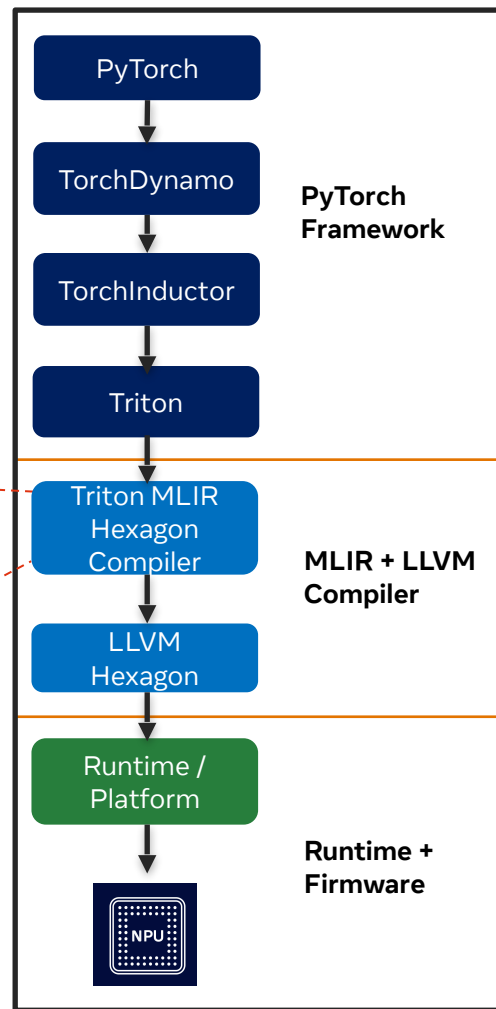
- ❑ Fine-tuning for performance
- ❑ Initial performance looking promising

## PYTORCH WORKFLOW VIA TRITON NPU BACKEND

Hexagon NPU backend for Triton enables PyTorch mapping



Developing tools to enable the workflow in a developer-friendly Python environment



Thank you!

