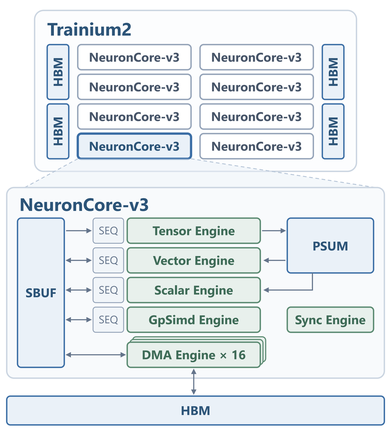
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Tensor Contractions on AWS Trainium2

Automatically compiling tensor contractions to high-performance NKI kernels

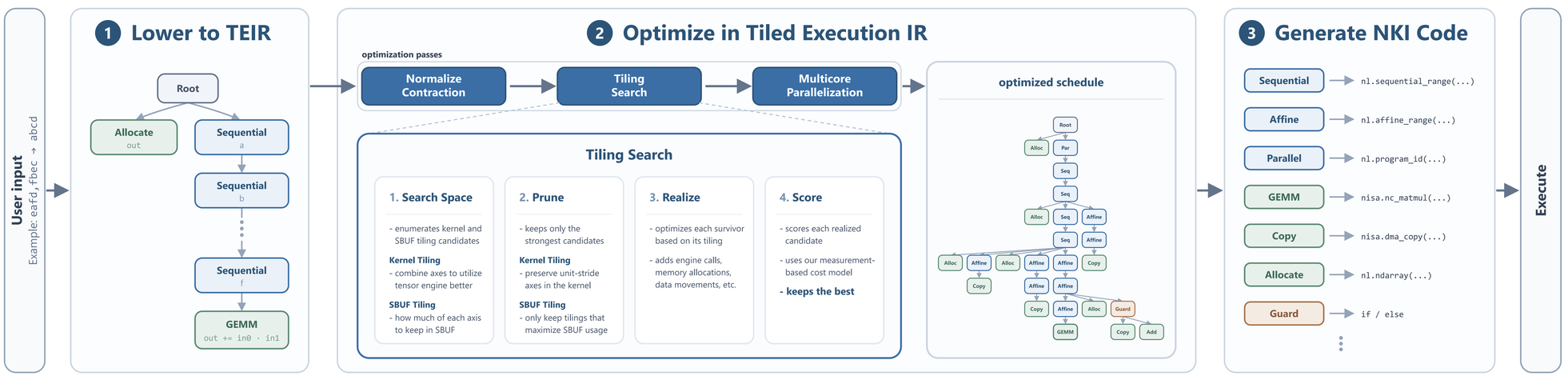
Max Engel, Alex Breuer - Friedrich Schiller University Jena

THE TRAINIUM2 CHIP



Trainium2 architecture, redrawn after the AWS Neuron SDK documentation [1].

THE PIPELINE - ONE TENSOR CONTRACTION TO A HIGH-PERFORMANCE NKI KERNEL



UTILIZING THE TENSOR ENGINE

Axis sizes rarely match the tensor engine

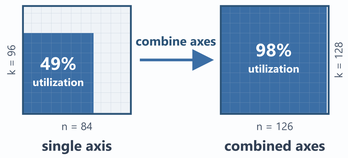
The tensor engine is a **systolic array** of 128×128 **Processing Elements** (PEs), but axis sizes are **rarely multiples of 128**.

Example: `eafd,fbec → abcd`

n-axes: b = 84, c = 84 **k**-axes: e = 96, f = 96

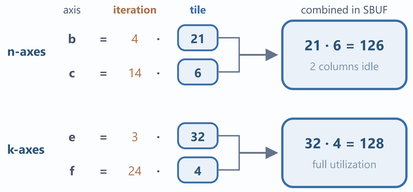
Combining axes increases utilization

■ = second input tile loaded into the PEs (held **stationary**).



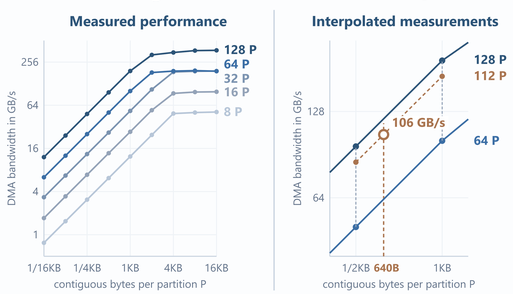
Splitting axes into iteration and tile

Every axis splits into an **iteration** factor and a **tile** factor. The tile factors of several axes are combined in SBUF into one engine dimension.



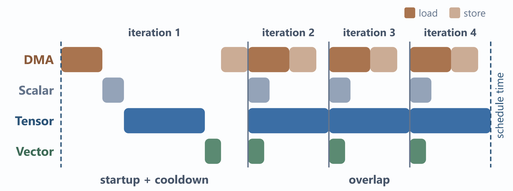
MODELING TRAINIUM2 PERFORMANCE

Per-engine cost model: shown for the DMA engines



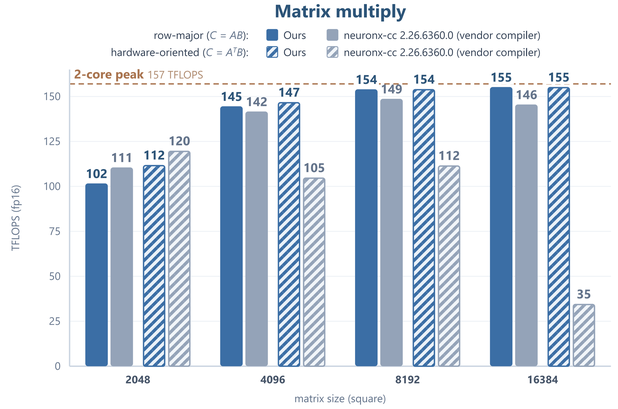
Schedule cost model

- the first iteration is priced **sequentially**, covering startup and cooldown of the execution
- **double buffering** enables overlapping execution of consecutive iterations
- later iterations are priced with **overlap**, the busiest engine determines the iteration time



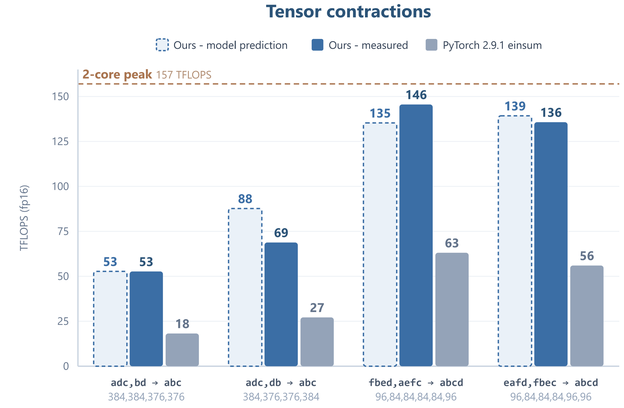
EVALUATION

Matrix multiply



Square matrix multiplication in two data layouts.

Tensor contractions



Four tensor contractions from the TCCG benchmark [2] in row-major data layout. Axis sizes are in alphabetical index order.

CONCLUSIONS

- **Automatically compiles contractions to NKI code.** Automatic search over candidate schedules, ranked by the cost model.
- **High tensor engine utilization takes several axes.** Combining axes increases tensor engine utilization from 49% to 98%.
- **Matrix multiply reaches 155 of 157 TFLOPS.** High performance in both data layouts and across all matrix sizes.
- **Contractions outperform PyTorch einsum by 2.3× to 2.9×.** Complex contractions reach up to 146 TFLOPS.

Interested in this work? Reach out!

Max Engel
max.engel@uni-jena.de

Alex Breuer
alex.breuer@uni-jena.de

References

[1] Amazon Web Services. Trainium2 Architecture Guide for NKI, 2026. awsdocs-neuron.readthedocs-hosted.com/en/v2.31.0/nki/guides/architecture/trainium2_arch.html

[2] P. Springer, P. Bientinesi. Design of a High-Performance GEMM-like Tensor–Tensor Multiplication. ACM TOMS 44(3), 2018.