Quantum Information National Laboratory HUNGARY

# Ultra Low Latency AI support to drive detectors

**Peter Rakyta**, Gregory Morse
Eötvös Loránd University
HUN-REN Wigner Research Centre for Physics

Benjamin Mencer, Ryan Coffee
SLAC National Accelerator Laboratory

ELTE EÖTVÖS LORÁND UNIVERSITY

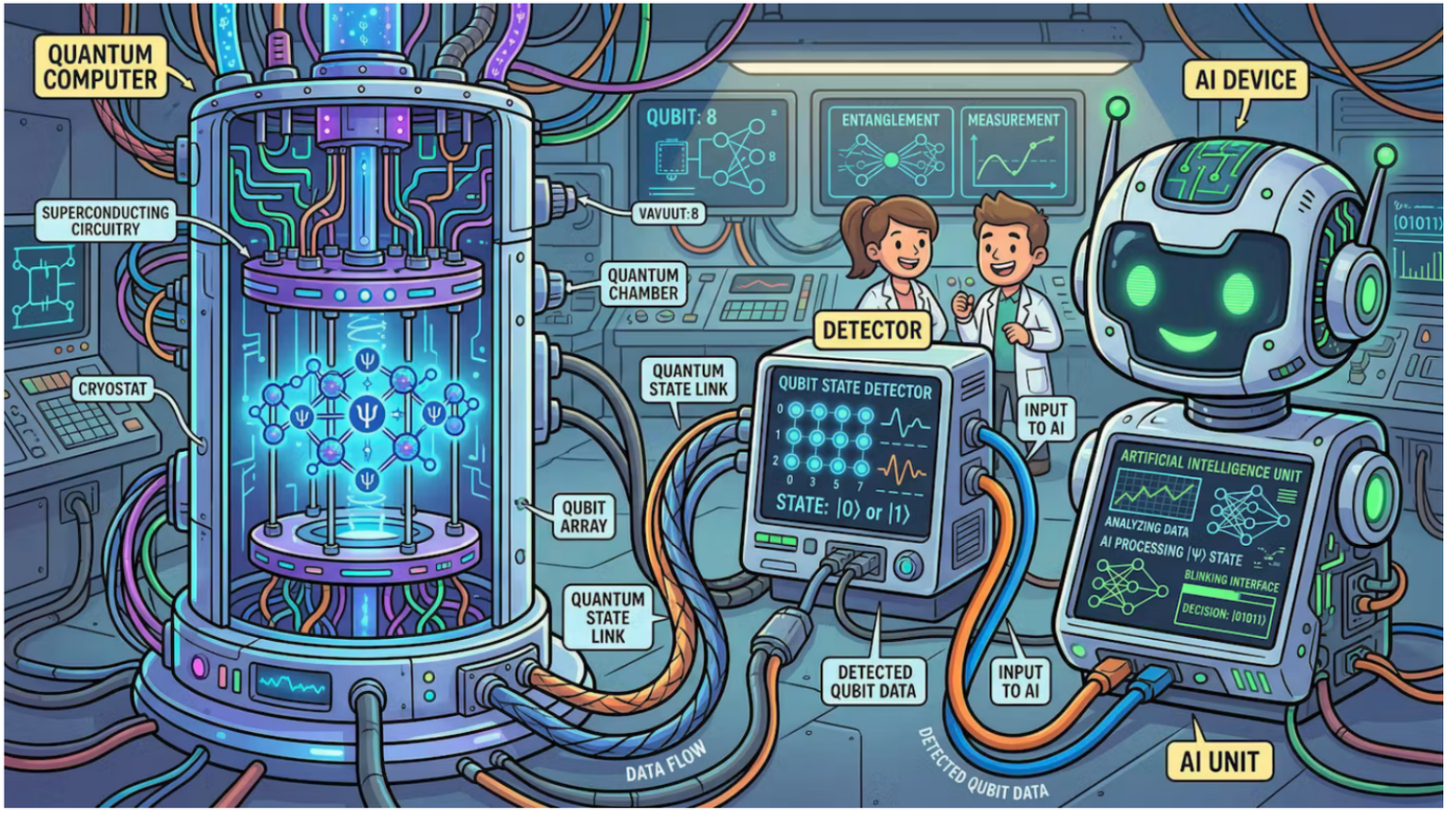
WIGNER

SLAC NATIONAL ACCELERATOR LABORATORY

# Ultra Low Latency AI support to drive (QC) detectors

- Error correction
- Error mitigation
- Pulse generation

- Requires AI feedback on a $\mu s$ time scale




ELTE EÖTVÖS LORÁND UNIVERSITY

Wigner

Quantum Information National Laboratory HUNGARY

# The Genesis Mission

A national initiative to build the world's most powerful scientific platform.

**The Genesis Mission unites DOE National Labs, industry, academia, and more to harness AI for breakthroughs in energy dominance, discovery science, and national security.**

SIMPLER Grants.gov

U.S. DEPARTMENT of ENERGY

## The Genesis Mission: Transforming Science and Energy with AI

**Agency:** Office of Science

Assistance Listings: 81.049 -- Office of Science Financial Assistance Program

Last Updated: April 10, 2026

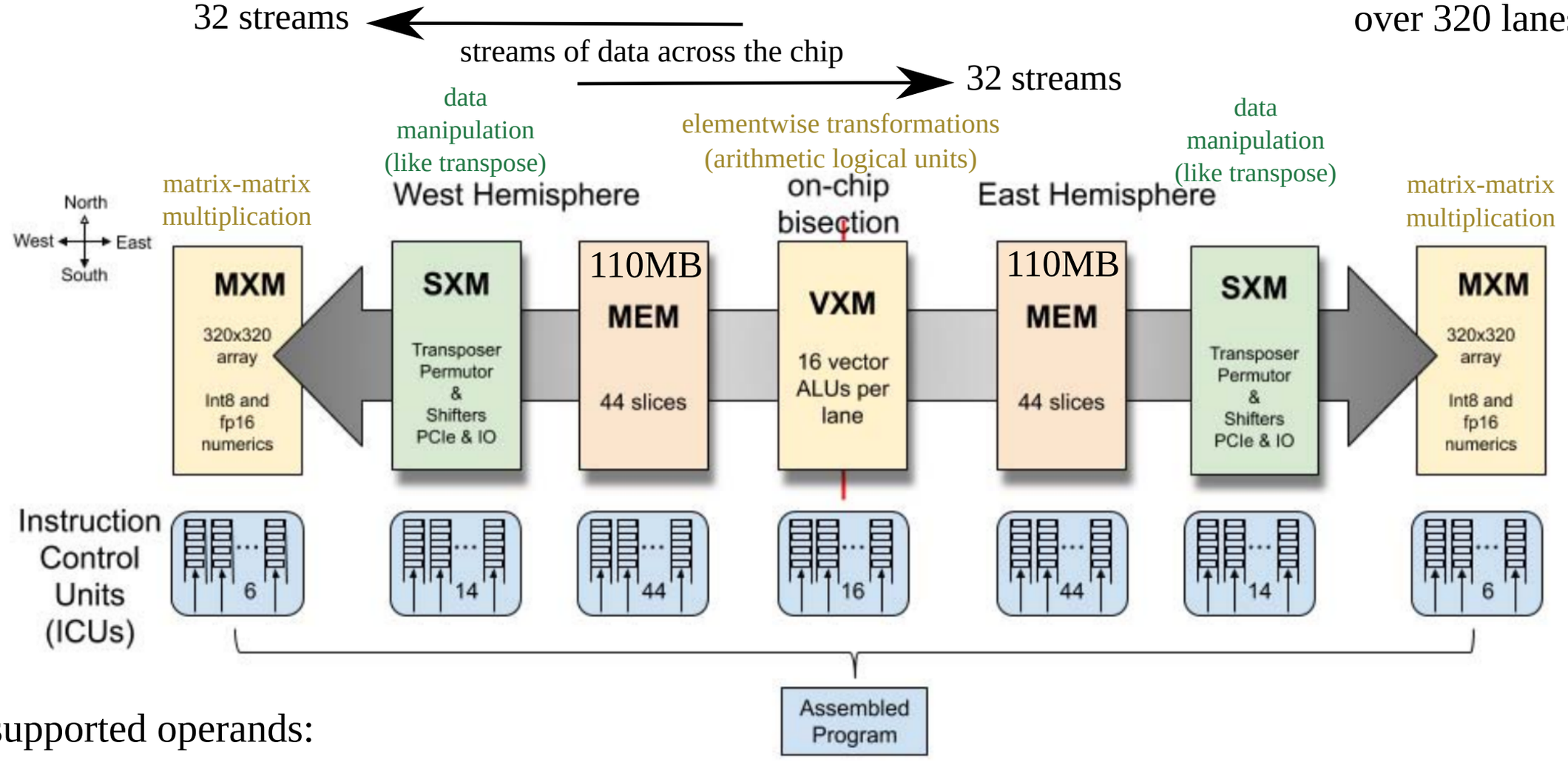
View version history on Grants.gov

# Think Fast: A Tensor Streaming Processor (TSP) for Accelerating Deep Learning Workloads

2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)

Quantum Information National Laboratory HUNGARY

each stream carries 320 vectorized bytes over 320 lanes




supported operands:

VXM: int8, int16, int32, uint8, uint16, uint32, float16, float32, bool8, bool16, bool32

MXM: int8 x int8 → int32, float16 x float16 → float32

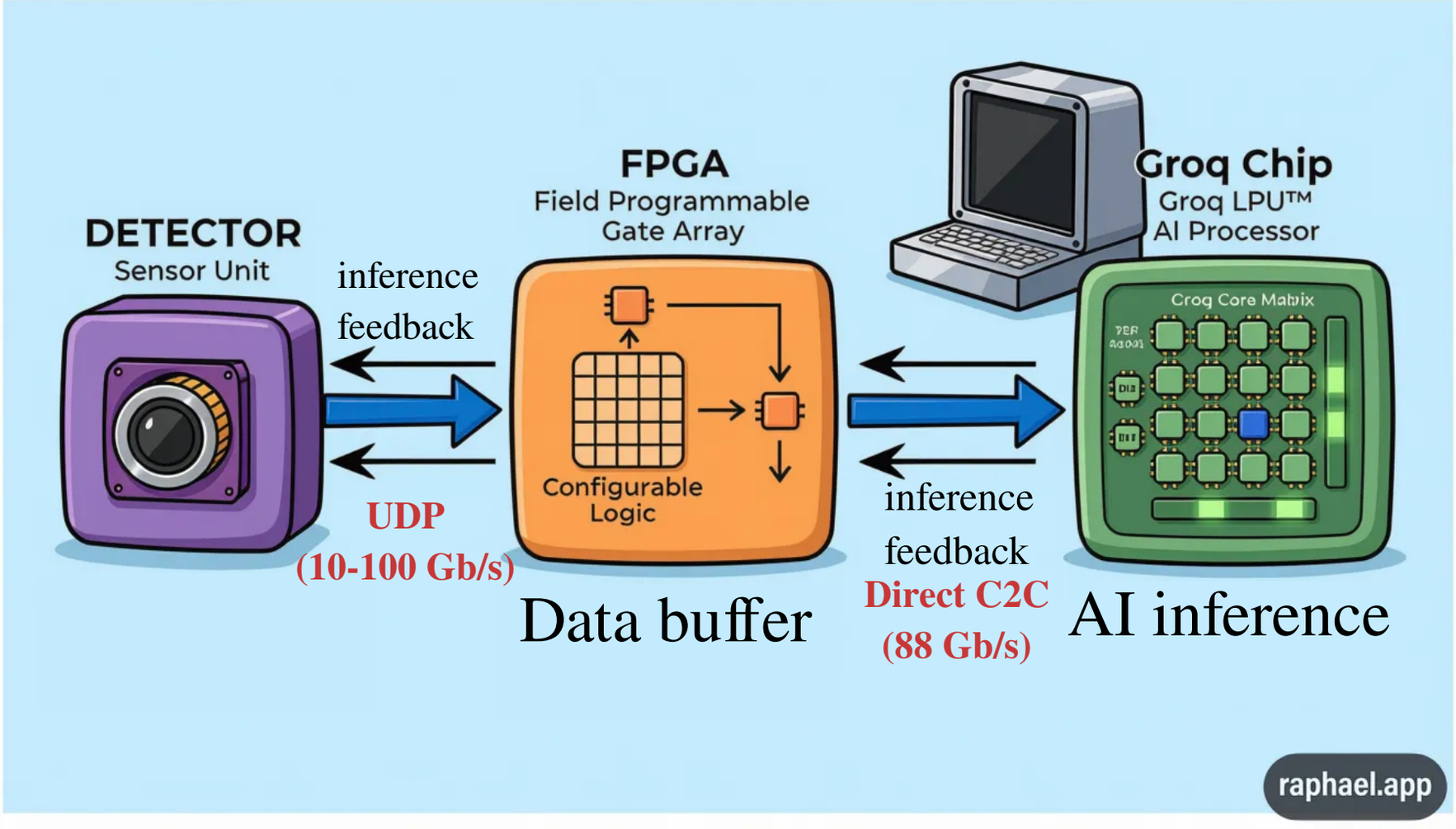
ELTE EÖTVÖS LORÁND UNIVERSITY

wigner

groq

MAXELER Technologies Maximum Performance Computing

# Inference loopback via data-flow deterministic architecture

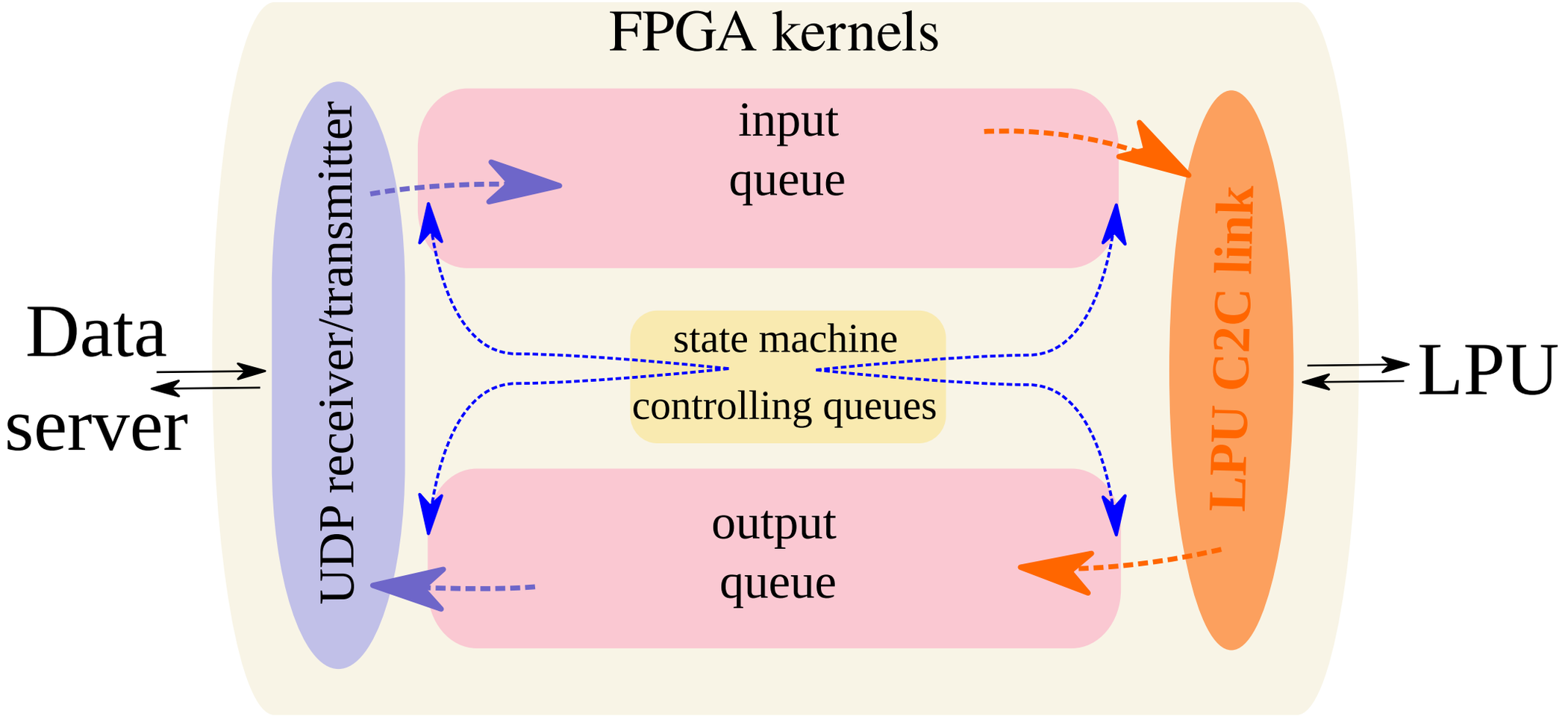



**Inference latency:**

X-ray imaging $\sim 45\ \mu s$ (22 kHz)

fusion plasma characterization $\sim 10\ \mu s$ (100 kHz)

# FPGA gate keeper



# Inference engine: workflow

- Hardware bring-up: Groq sends "**notify**" to the FPGA

  *How much input data is required for the inference?*

- FPGA collects input data incoming trough the UDP channel.
- FPGA triggers the Groq chip to run the inference program after the FPGA collected the input data
- The Groq chip sends "**read**" command to the FPGA as part of the inference program
- The FPGA starts streaming the inference input to Groq (92 Gbs)
- The Groq chip does the calculations
- The Groq chip sends "**write**" command and the data to the FPGA
- The FPGA starts streaming the received inference return via the UDP channel
- Groq sends "**notify**" to the FPGA

Quantum Information National Laboratory HUNGARY

SLAC NATIONAL ACCELERATOR LABORATORY

ELTE EÖTVÖS LORÁND UNIVERSITY

WIGNER

# I/O Data compression

- Transmitting data between the detector and the inference engine is a botleneck

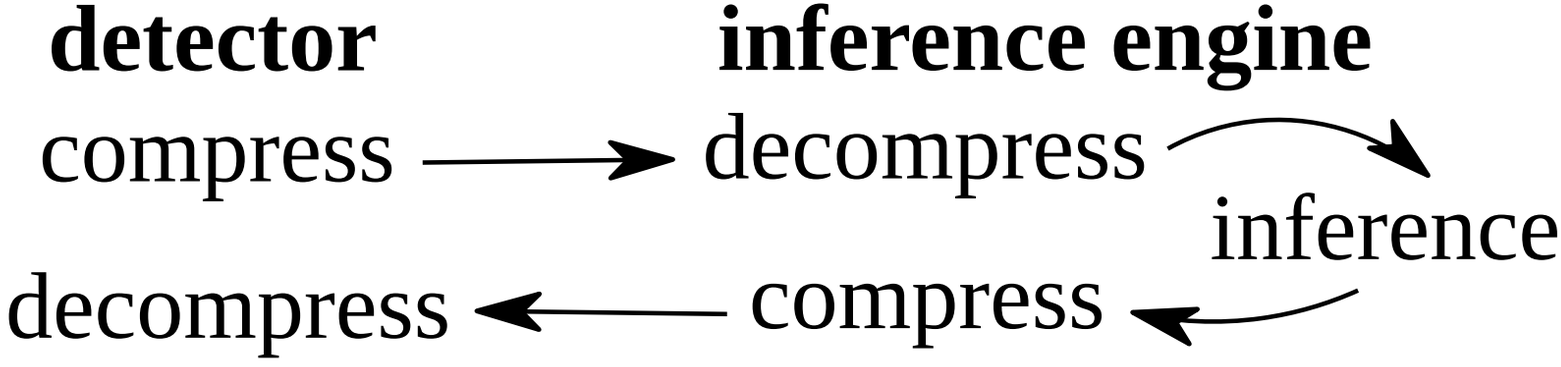
| Feature | FP4 (E2M1) | MXFP4 | NVFP4 |
|---|---|---|---|
| **Format Structure** | 4 bits (1 sign, 2 exponent, 1 mantissa) plus software scaling factor | 4 bits (1 sign, 2 exponent, 1 mantissa) plus 1 shared power-of-two scale per 32 value block | 4 bits (1 sign, 2 exponent, 1 mantissa) plus 1 shared FP8 scale per 16 value block |
| **Accelerated Hardware Scaling** | No | Yes | Yes |
| **Memory** | Up to 4x less memory than FP16 | | |
| **Accuracy** | Risk of noticeable accuracy drop compared to FP8 | Risk of noticeable accuracy drop compared to FP8 | Lower risk of noticeable accuracy drop particularly for larger models |

OCP Microscaling Formats (MX) Specification

concept of: FP8, FP4

definition of mantissa and exponent bits

- We apply the following compression approach



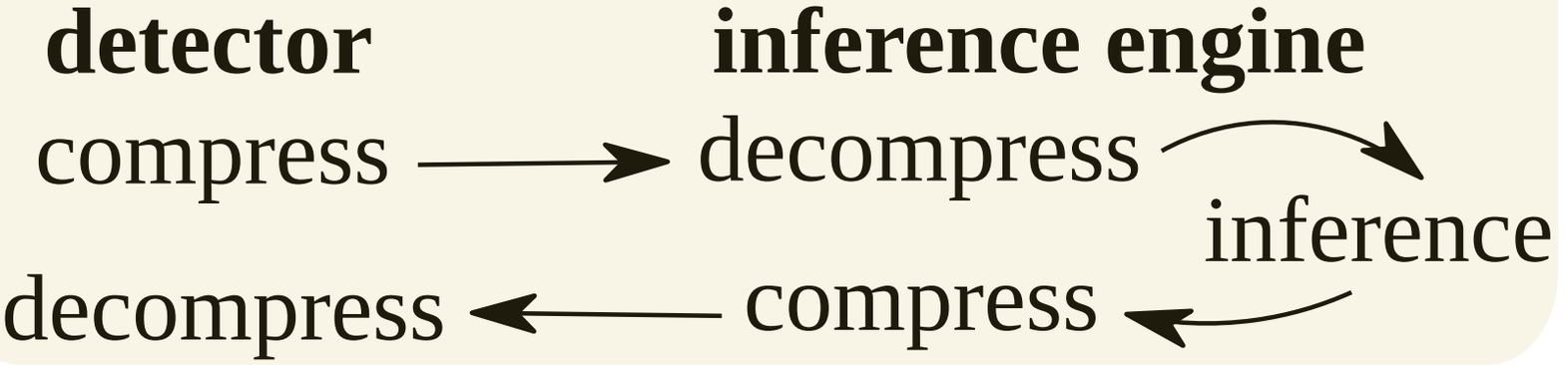
# I/O Data compression

- Transmitting data between the detector and the inference engine is a botleneck
- Our proof-of-concept solution is to use 0 exponent bits: fixed-point number representation

$$\text{FP32/FP16 data} = \text{scale}_{\text{FP32}} \times \text{uint8/uint4 data} + \text{offset}_{\text{FP32}}$$

- The compressed data are always nonzero ($\text{offset}_{\text{FP32}}$)
- Scale data within (0.0, 1.0): $\text{scale}_{\text{FP32}}$
- Multiply with $2^4$ or $2^8$.
- Rounding to nearest even integer
- Boundle together compressed data and scaling factors

- We apply the following compression approach



arXiv > physics > arXiv:2502.16141

Physics > Applied Physics

[Submitted on 22 Feb 2025 (v1), last revised 14 Nov 2025 (this version, v2)]

# A Hybrid Neural Architecture: Online Attosecond X-ray Characterization

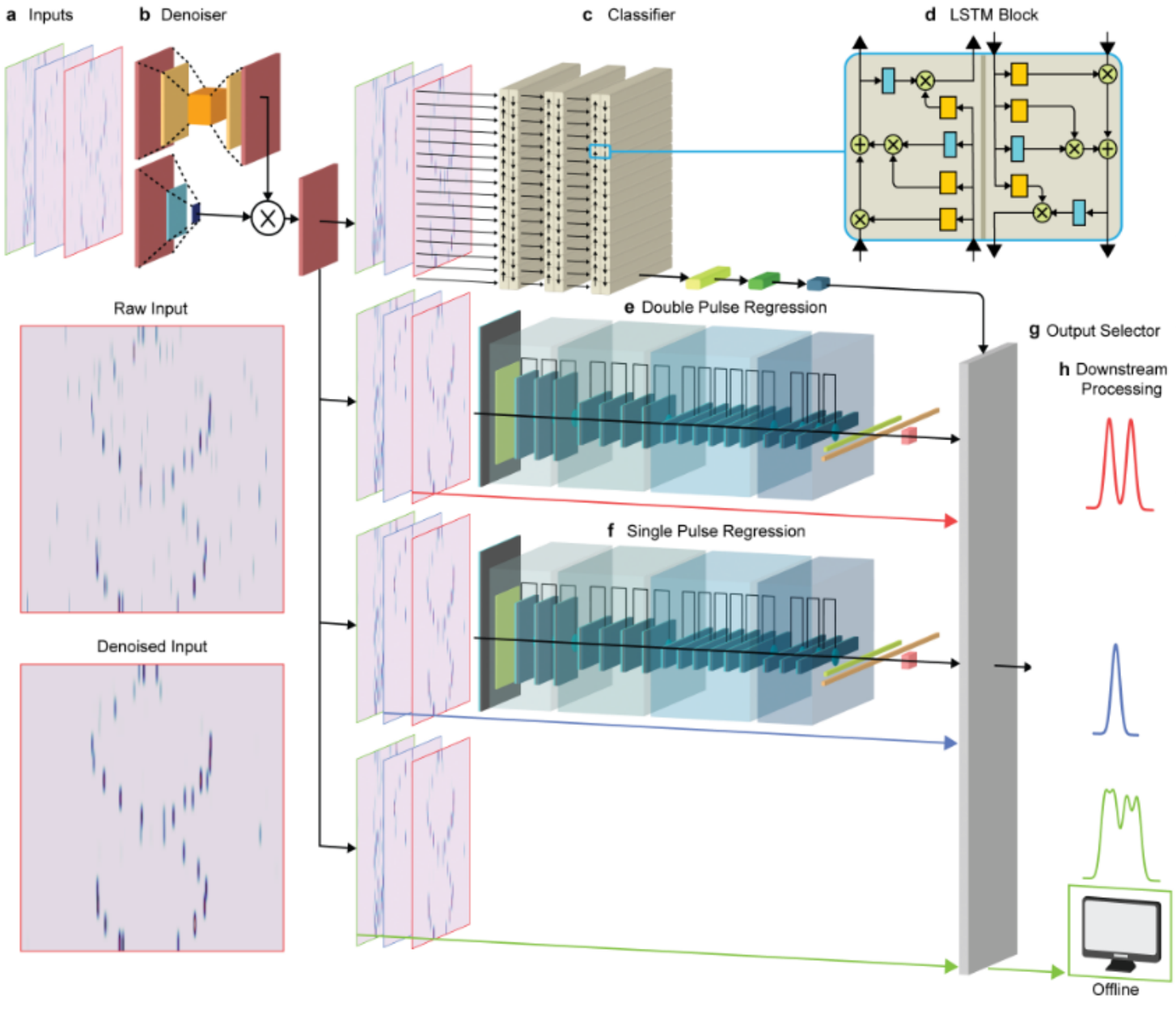
Jack Hirschman, Benjamin Mencer, Razib Obaid, Amanda Shackelford, Ryan Coffee

Quantum Information National Laboratory HUNGARY

ELTE EÖTVÖS LORÁND UNIVERSITY

Wigner

SLAC NATIONAL ACCELERATOR LABORATORY




- Denoiser
- Zero pulse classifier
- LSTM based pulse num classifier
- ResNet based feature regression

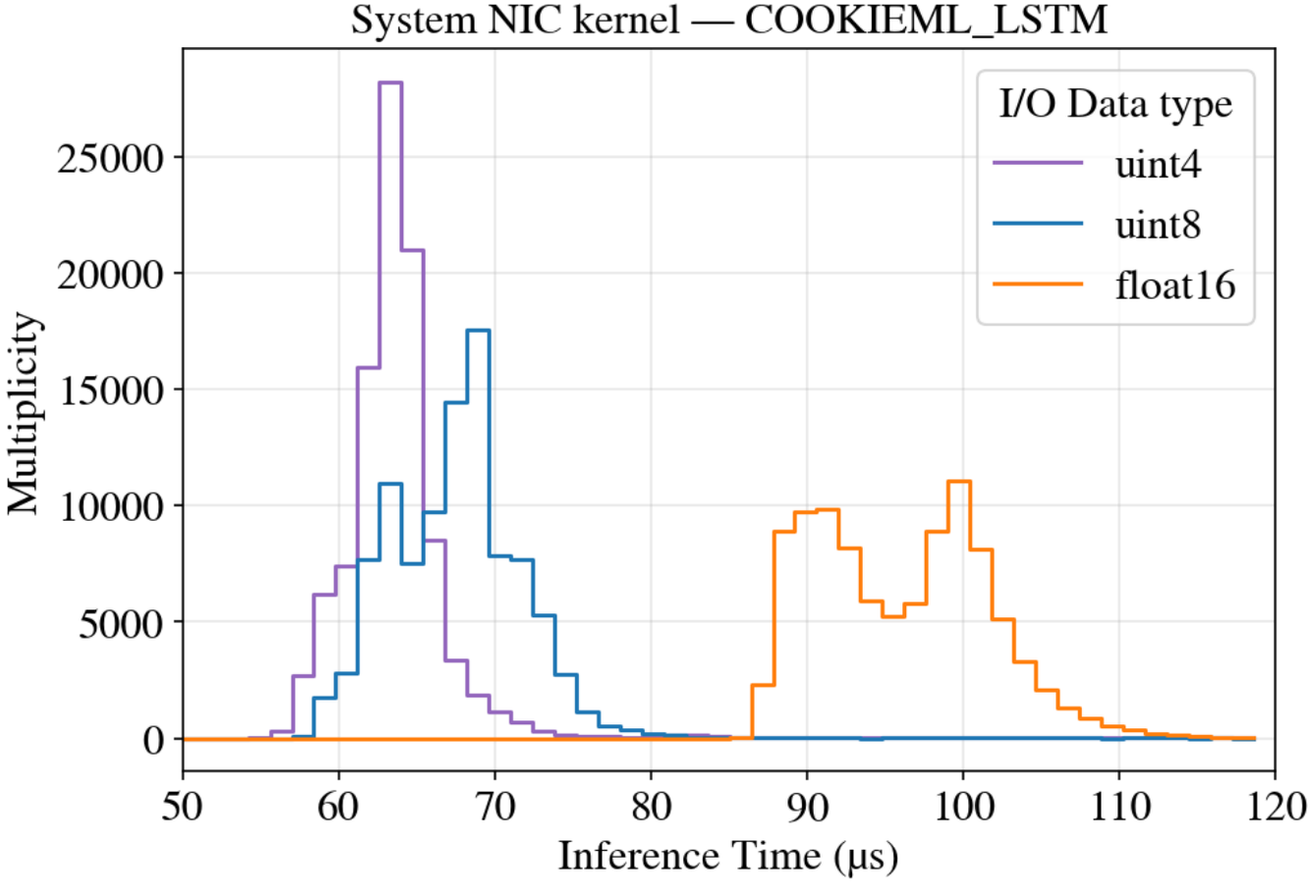
# Inference latency on the LSTM pipeline

- Denoiser
- Zero pulse classifier
- LSTM based pulse num classifier
- Groq compute time:

  int4: 19 µs

  int8: 20 µs

  float16: 20 µs




- detector ←10 Gbs→ FPGA Groq LPU

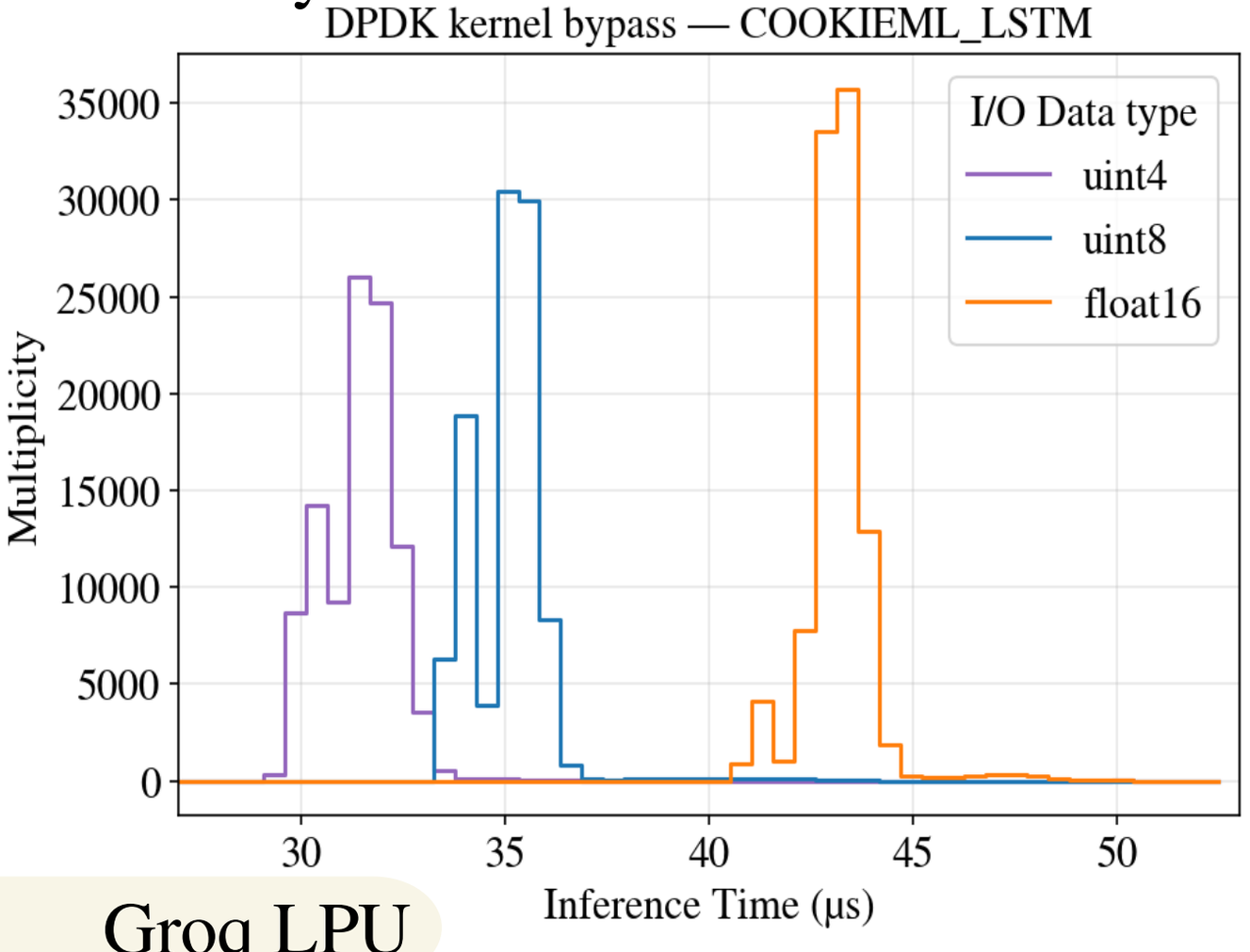
# Inference latency on the LSTM pipeline: DPDK kernel bypass

- bypasses the traditional Linux kernel networking stack to reduce latency
- DPDK is widely used in telecom, cloud networking, firewalls, routers, etc
- Groq compute time:

  int4: 19 µs

  int8: 20 µs

  float16: 20 µs




- detector ←10 Gbs→ FPGA Groq LPU

# RTCAKENN

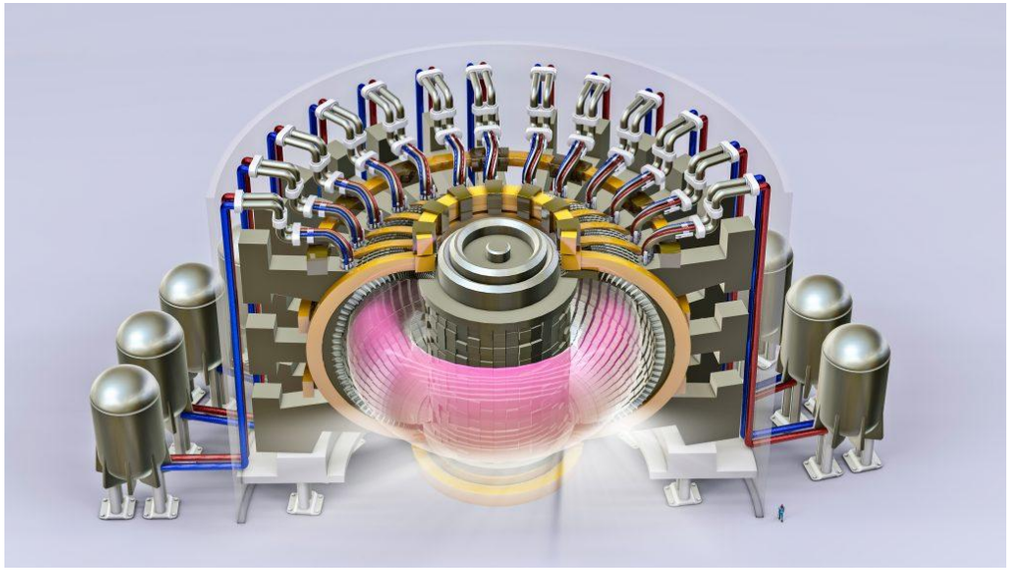
## (Real-Time Convolutional Autoencoder Kernel-based Embedded Neural Network)

Quantum Information National Laboratory HUNGARY

ELTE EÖTVÖS LORÁND UNIVERSITY

wigner

SLAC NATIONAL ACCELERATOR LABORATORY



- machine-learning model developed for real-time kinetic profile reconstruction in tokamak fusion plasmas
- developed by DIII-D National Fusion Facility

The model predicts

- pressure
- safety factor
- toroidal current density
- electron temperature
- electron density
- ion temperature
- plasma rotation profiles

Machine learning-based real-time kinetic profile reconstruction in DIII-D

JOURNAL ARTICLE · 22 December 2023 · Nuclear Fusion

DOI: https://doi.org/10.1088/1741-4326/ad142f · OSTI ID: 2251556

Shousha, Ricardo ; Seo, Jaemin ; Erickson, Keith; Xing, Zichuan ; Kim, SangKyeun ; Abbate, Joseph ; Kolemen, Egemen

Inference runs in under about 8 ms (10 runs to get averaged output)

# Inference latency on the cakeNN pipeline

Quantum Information National Laboratory HUNGARY

ELTE EÖTVÖS LORÁND UNIVERSITY

Wigner

SLAC NATIONAL ACCELERATOR LABORATORY

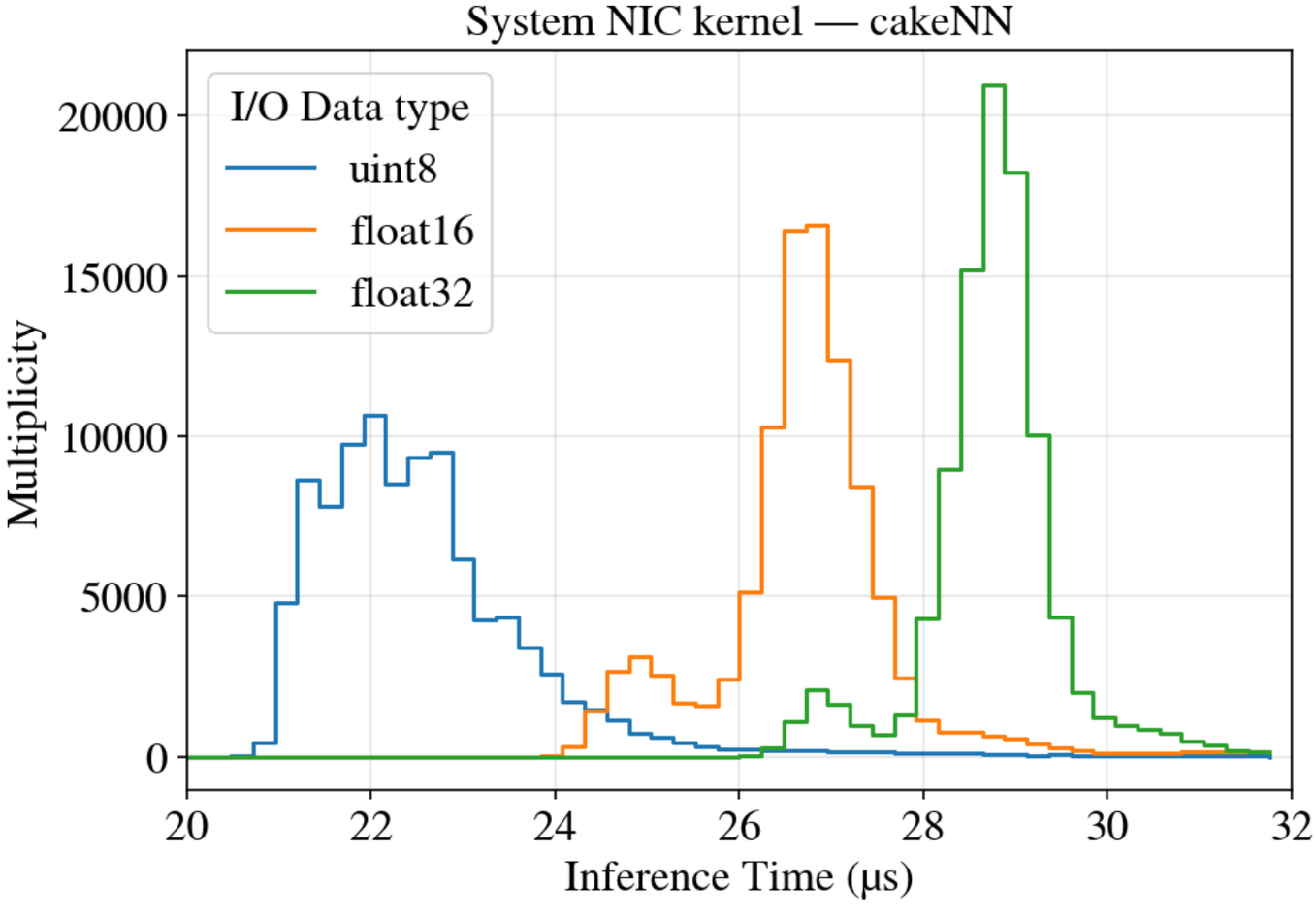
- pressure
- safety factor
- toroidal current density
- electron temperature
- electron density
- ion temperature
- plasma rotation profiles

- Groq compute time:

  int8: 4 µs

  float16: 4 µs

  float32: 4 µs




- detector ←10 Gbs→ FPGA Groq LPU

previously: ms time scale

# Inference latency on the cakeNN model: DPDK kernel bypass

Quantum Information National Laboratory HUNGARY

ELTE EÖTVÖS LORÁND UNIVERSITY

Wigner

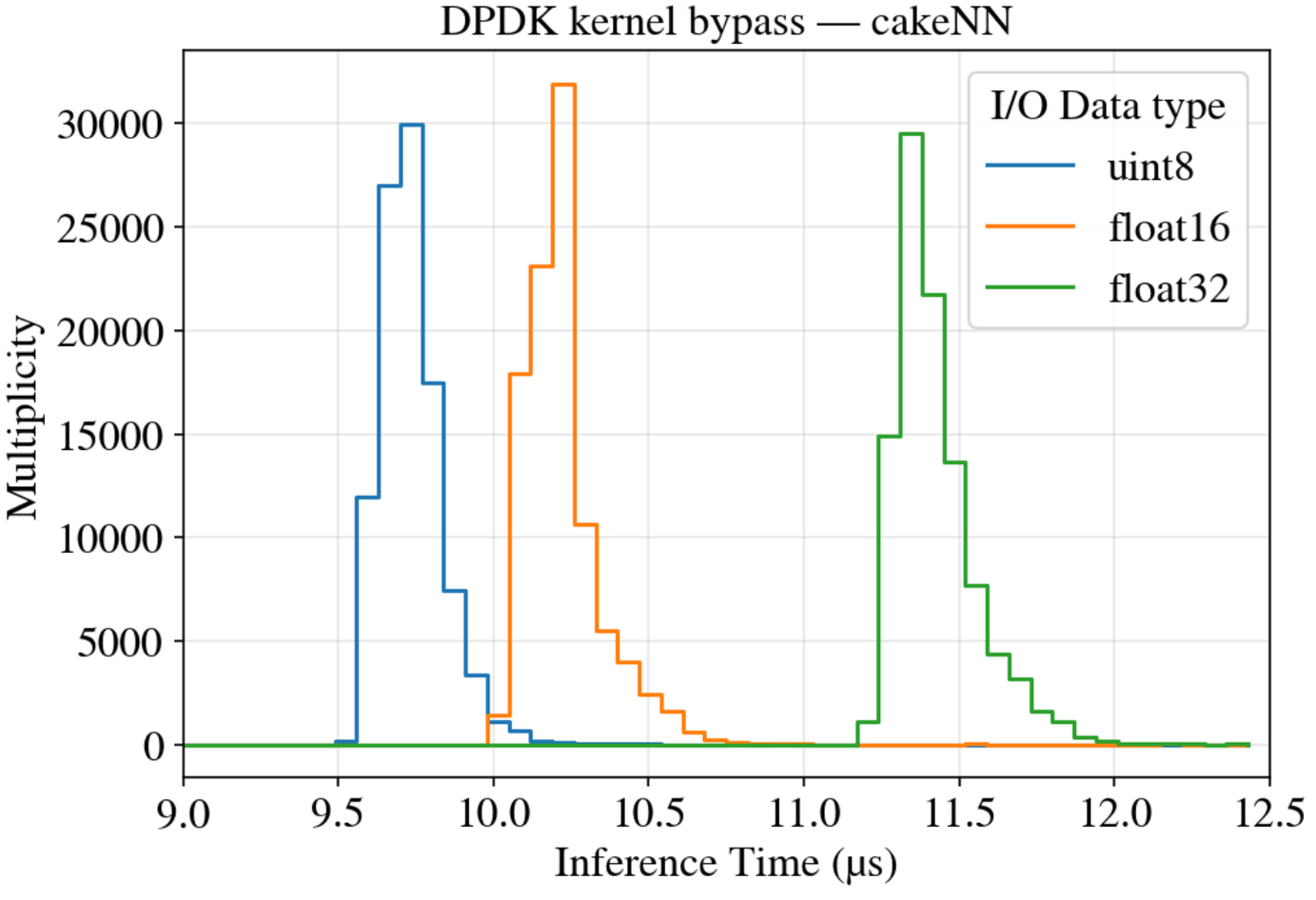
SLAC NATIONAL ACCELERATOR LABORATORY

- The I/O overhead is still too big
- further improvements:
  - FPGA ↔ FPGA
  - reimplement the FPGA gate keeper
- Groq compute time:

  int8: 4 µs

  float16: 4 µs

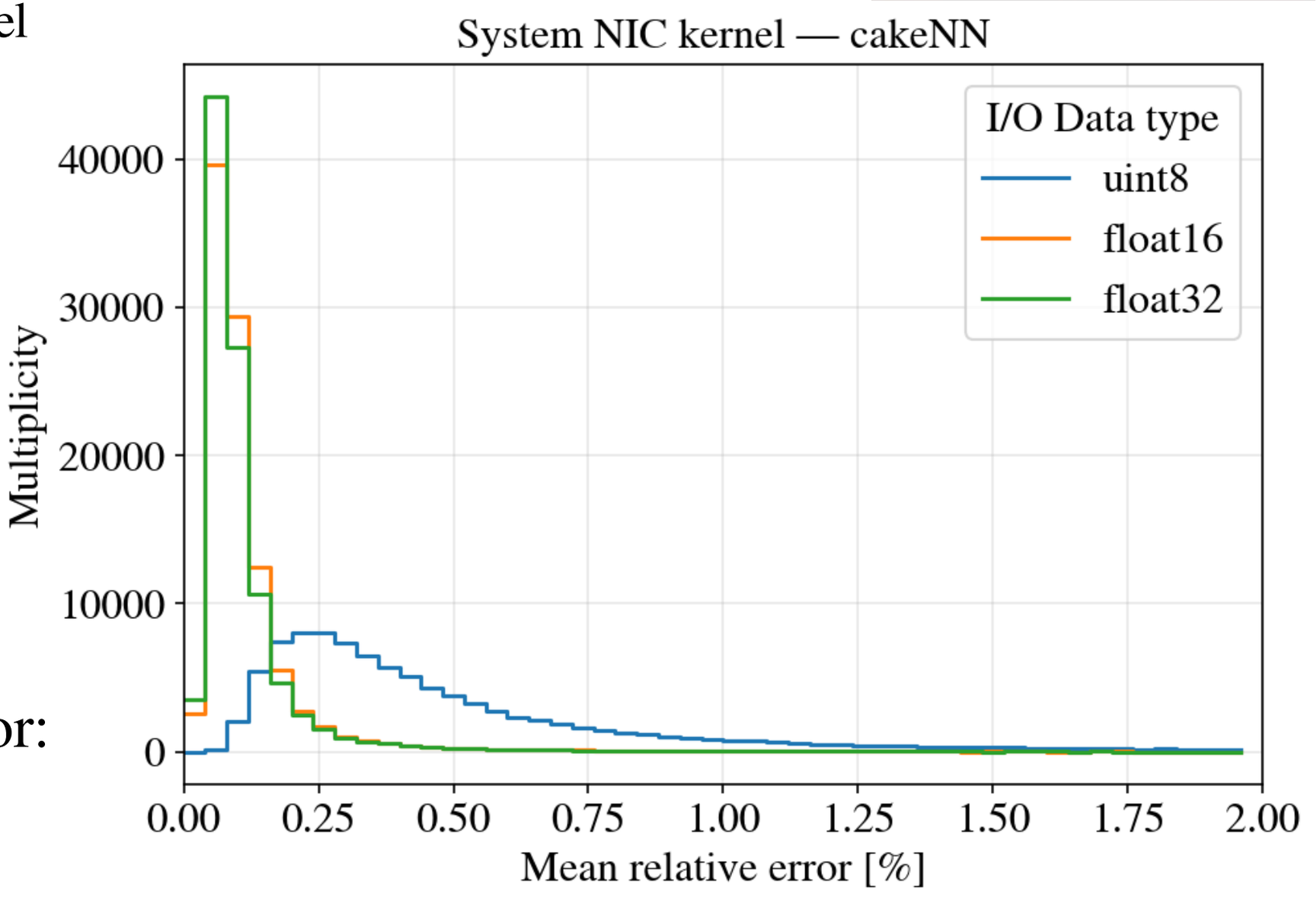
  float32: 4 µs




- detector ↔ (10 Gbs) FPGA Groq LPU

# Numerical accuracy: cakeNN model

Element-wise comparision with float32 torch model evaluated on CPU

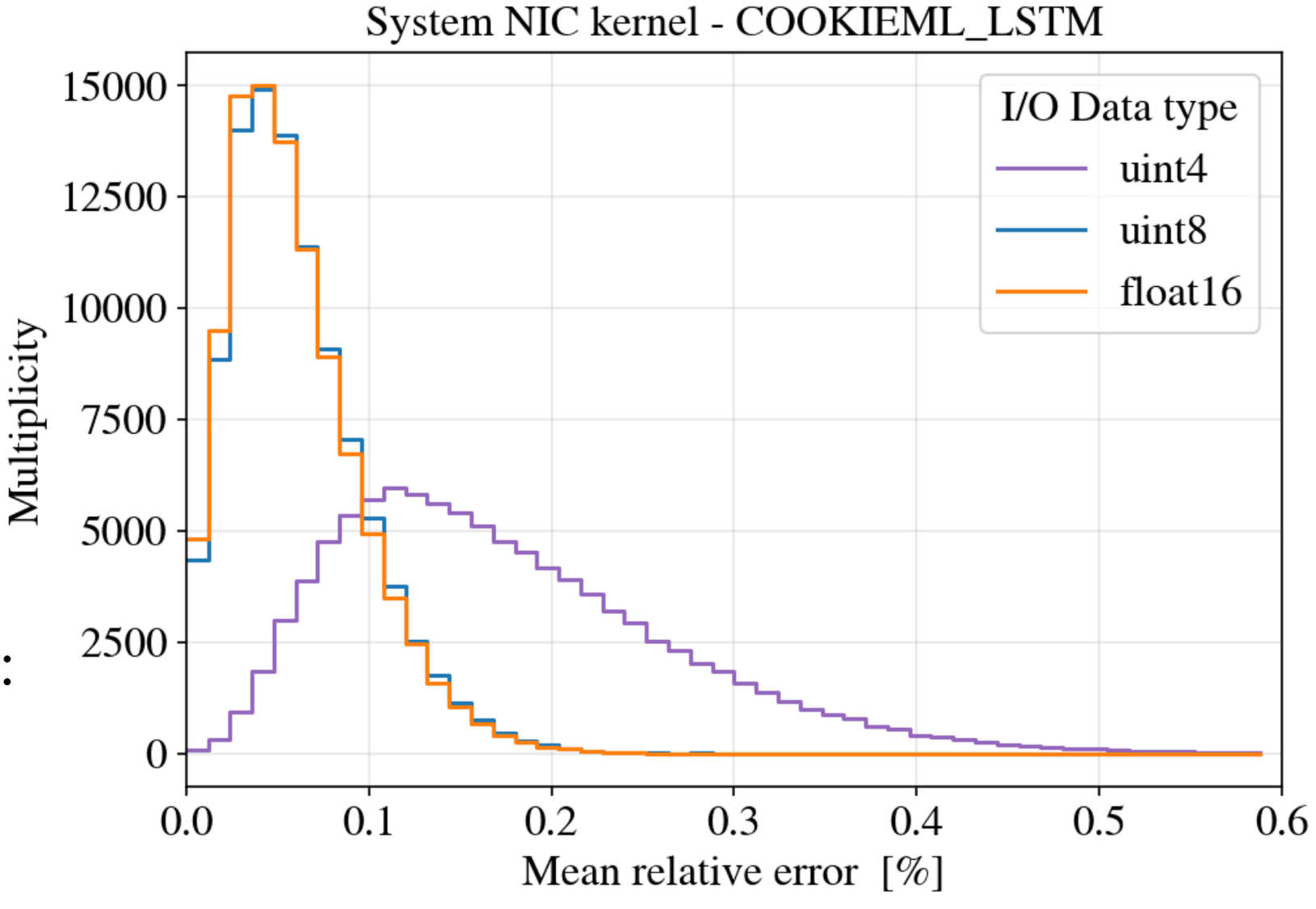



max mean error:

less than 4-6%

# Numerical accuracy: CookieML LSTM model

Element-wise comparision with float16 torch model evaluated on CPU

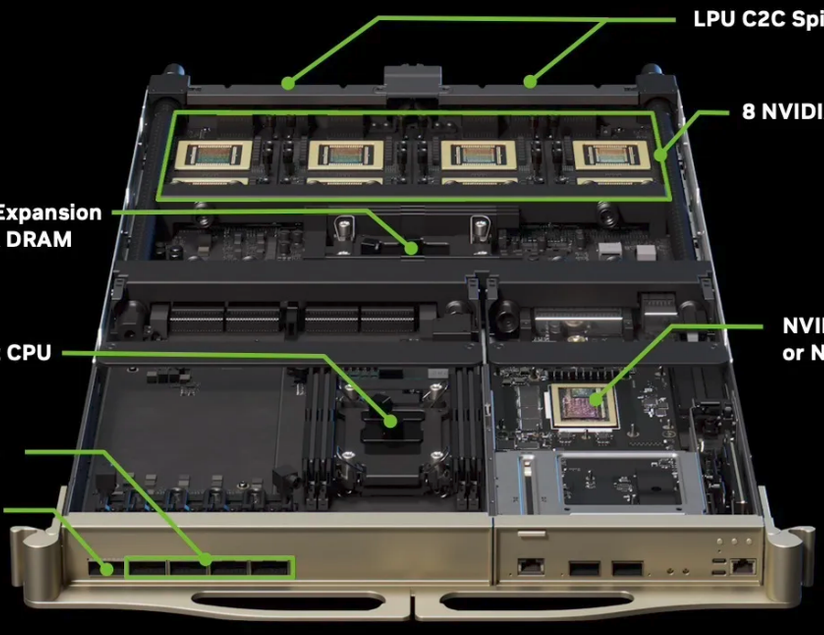



max mean error:

less than 1-2%

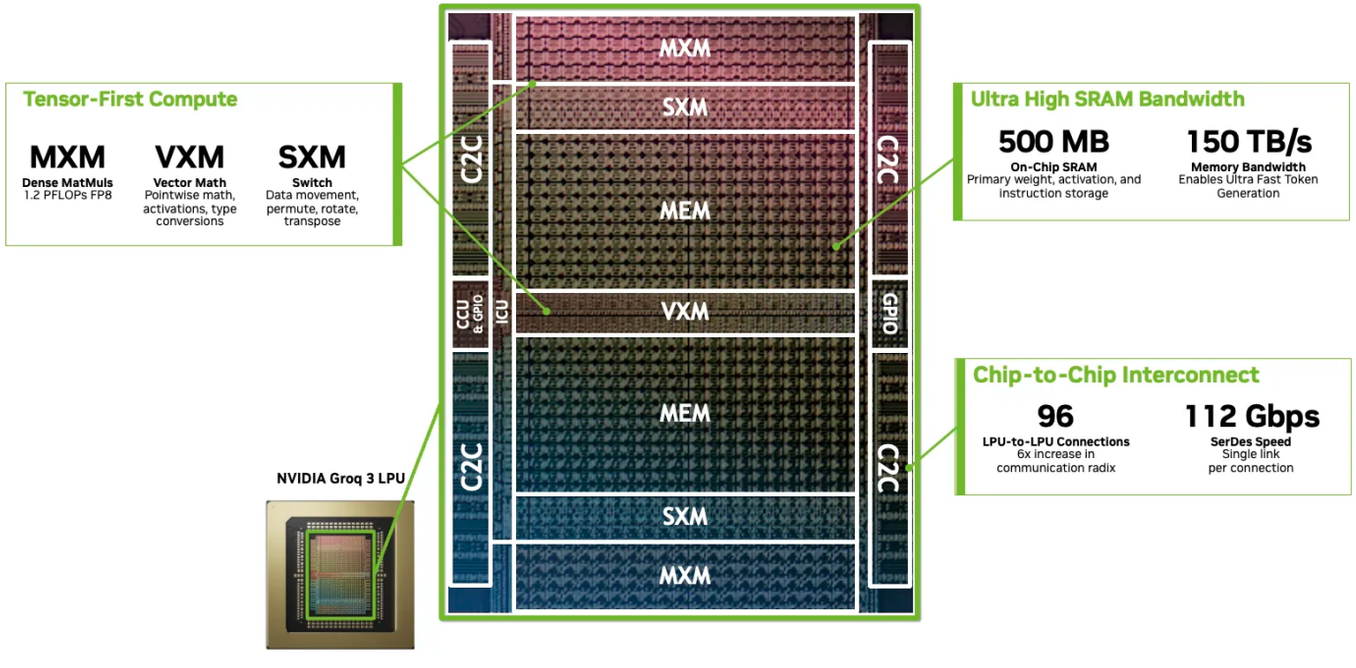
# NVIDIA Groq 3 LPX Compute Tray

Rack-ready, liquid cooled, 1U compute engine for low-latency inference




LPX compute tray specifications and configuration are preliminary and subject to change;

## NVIDIA Groq 3 LPU Chip Architecture




| Specification | NVIDIA Groq 3 LPX |
|---|---|
| AI inference compute | 315 PFLOPS |
| Total SRAM capacity | 128 GB |
| On-chip SRAM bandwidth | 40 PB/s |
| Scale-up density | 256 chips |
| Scale-up bandwidth | 640 TB/s |

# Aknowledgement

This research was supported by the Ministry of Innovation and Technology and the National Research, Development and Innovation Office within the Quantum Information National Laboratory of Hungary and Grants No. 2022-2.1.1-NL-2022-00004,
by the ÚNKP-24-5 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund,
by the Hungarian Scientific Research Fund (OTKA) Grant No. K134437
and by the Hungarian Academy of Sciences through the Bolyai János Stipendium (BO/00571/22/11).

We acknowledge the computational resources provided by the Wigner Scientific Computational Laboratory (WSCLAB) (the formerWigner GPU Laboratory)

contact: Peter Rakyta, rakyta.peter@wigner.hu