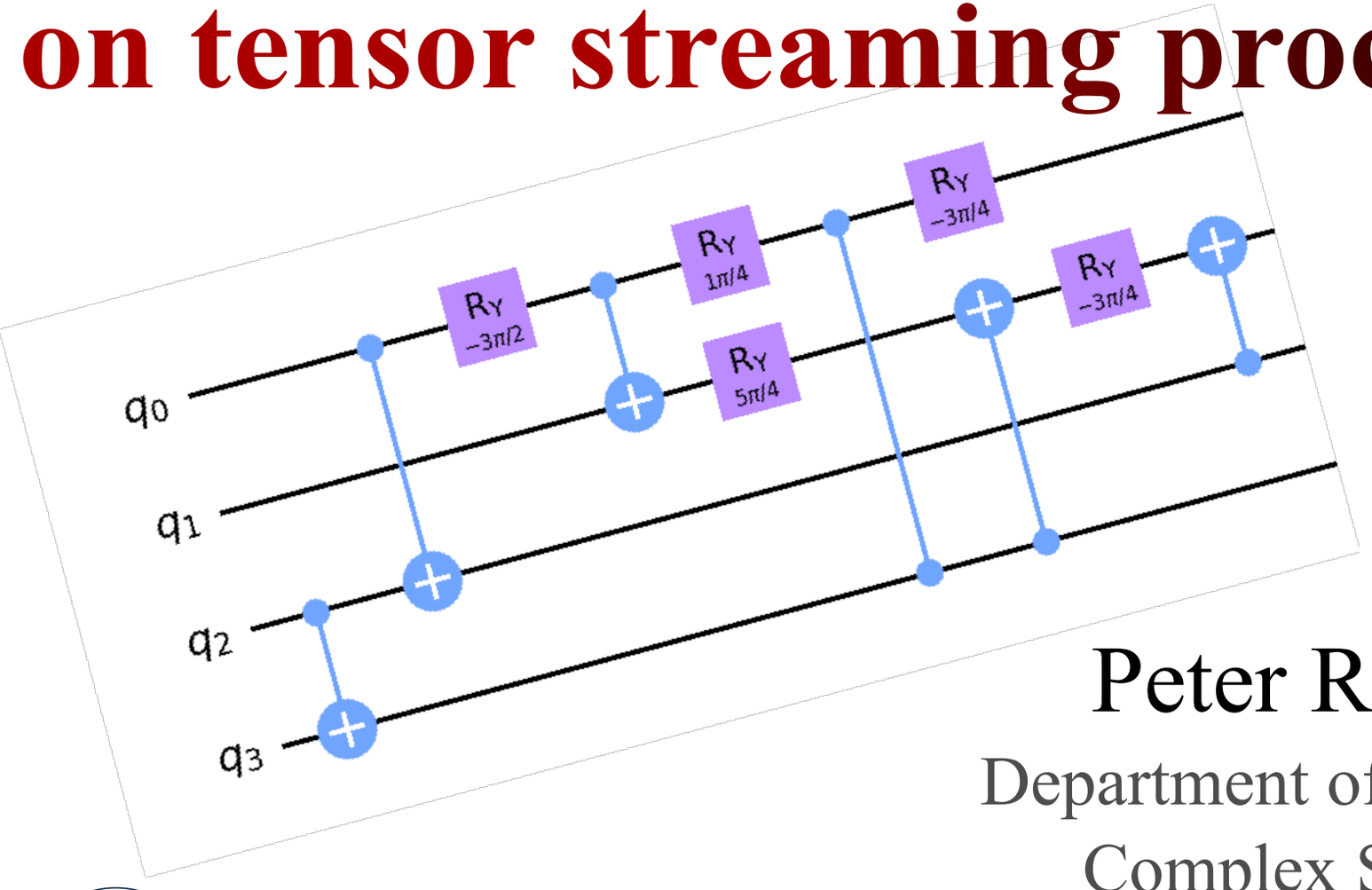
Quantum Information
National Laboratory
HUNGARY

Simulation of quantum computers on tensor streaming processors



Peter Rakyta
Department of Physics of Complex Systems

ELTE
EÖTVÖS LORÁND
UNIVERSITY

Data-flow implementation of a quantum computer simulator

Quantum Information National Laboratory HUNGARY

Organize data into streams flowing through the chip

Computations: operations on the elements of a data stream

Data-flow hardware + data-flow programming model =
Data-flow engine (DFE)

ELTE EÖTVÖS LORÁND UNIVERSITY

wigner

groq

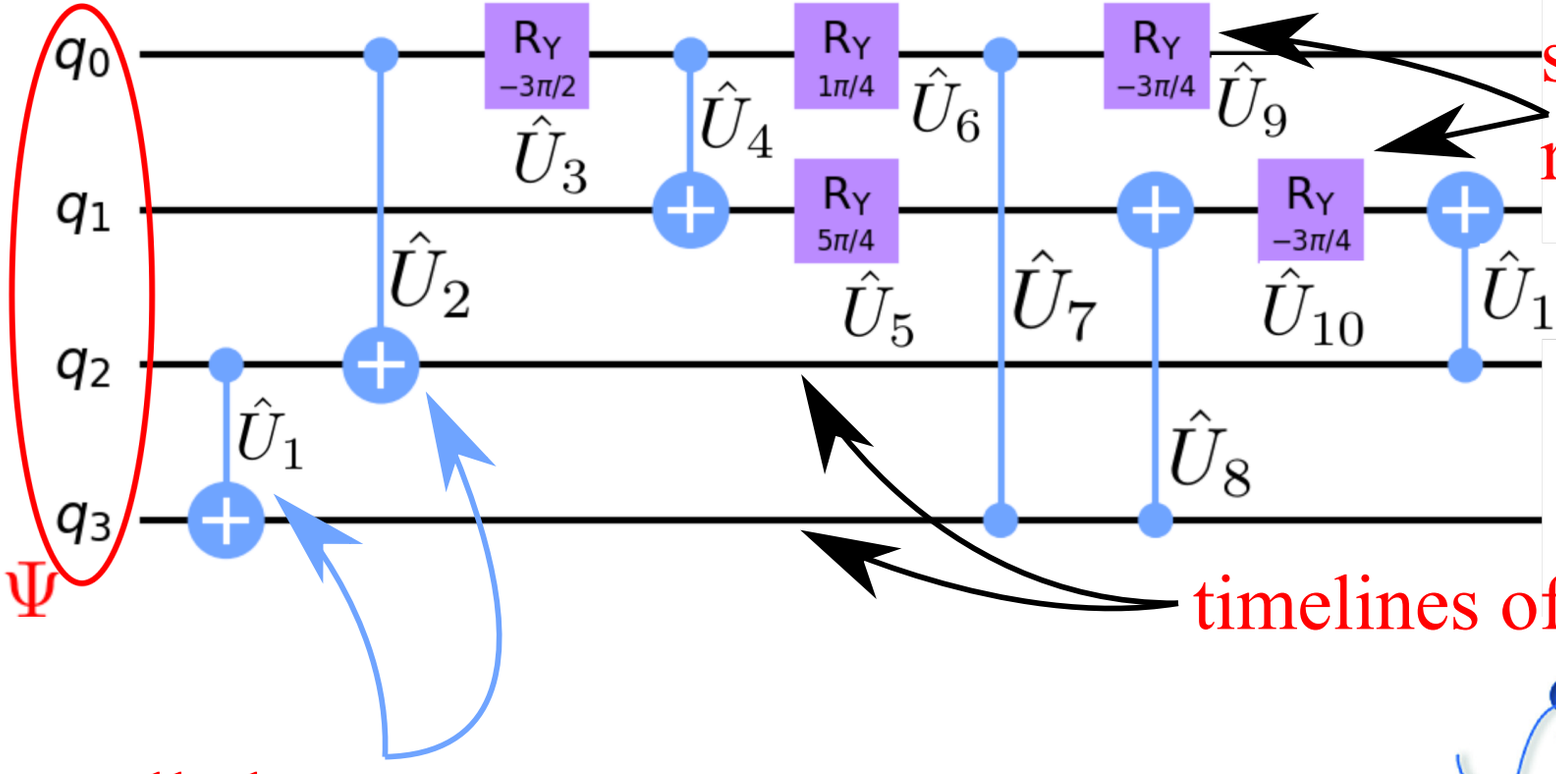
MAXELER Technologies
Maximum Performance Computing

Qubit based architecture

Quantum Information National Laboratory HUNGARY

quantum program (unitary)

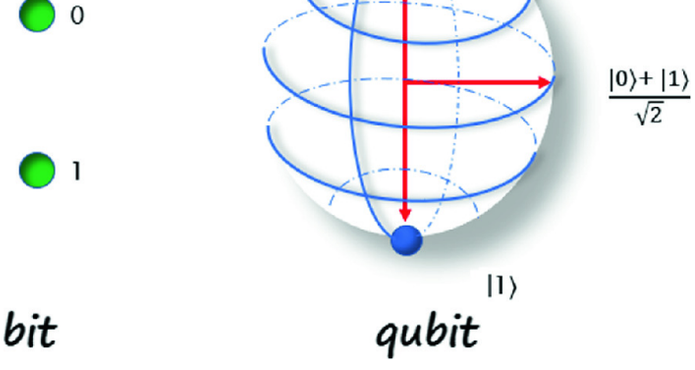
$$\hat{U} = \hat{U}_{11} \cdot \hat{U}_{10} \cdot \hat{U}_9 \cdot \hat{U}_8 \cdot \hat{U}_7 \cdot \hat{U}_6 \cdot \hat{U}_5 \cdot \hat{U}_4 \cdot \hat{U}_3 \cdot \hat{U}_2 \cdot \hat{U}_1$$



single qubit rotations

timelines of the qubits

controlled not gates

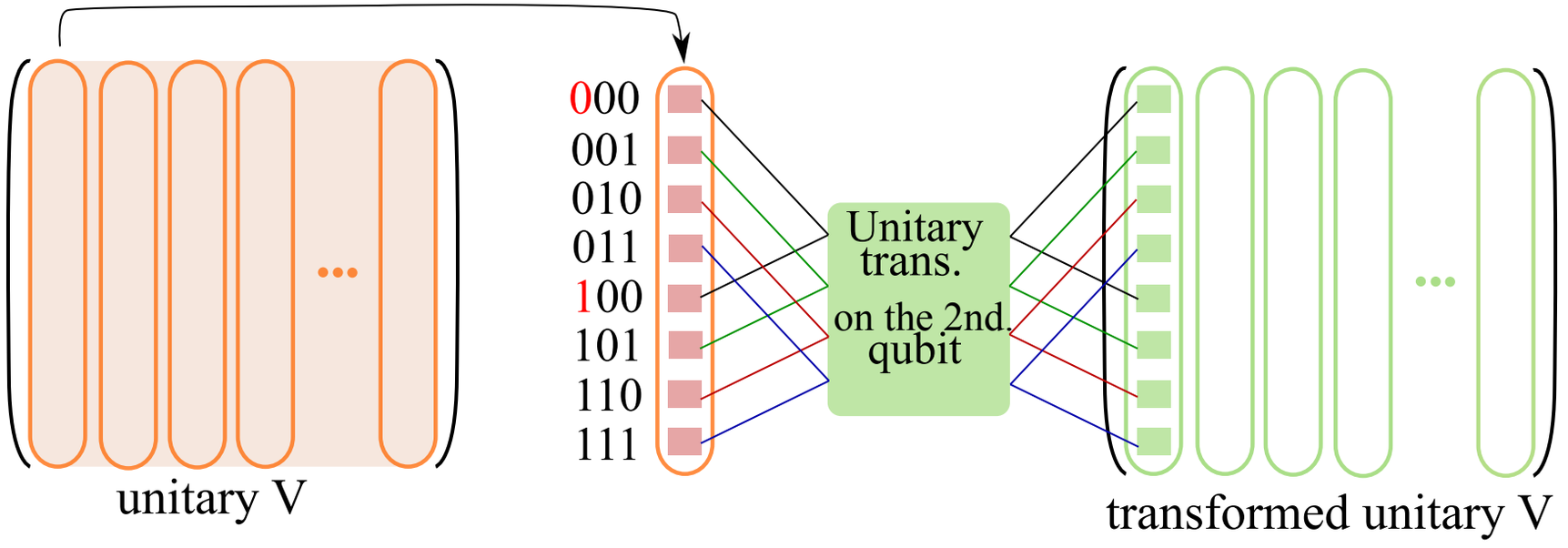


ELTE EÖTVÖS LORÁND UNIVERSITY

WIGNER

DFE flavour of quantum gate operations in quantum gate decomposition

Quantum Information National Laboratory HUNGARY



The elementary gate operations can be represented by sparse unitaries, mixing **element pairs** in the columns of V

Organizing the columns of V into a stream of data → DFE model of gate operations

ELTE EÖTVÖS LORÁND UNIVERSITY

wigner

groq

MAXELER Technologies Maximum Performance Computing

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

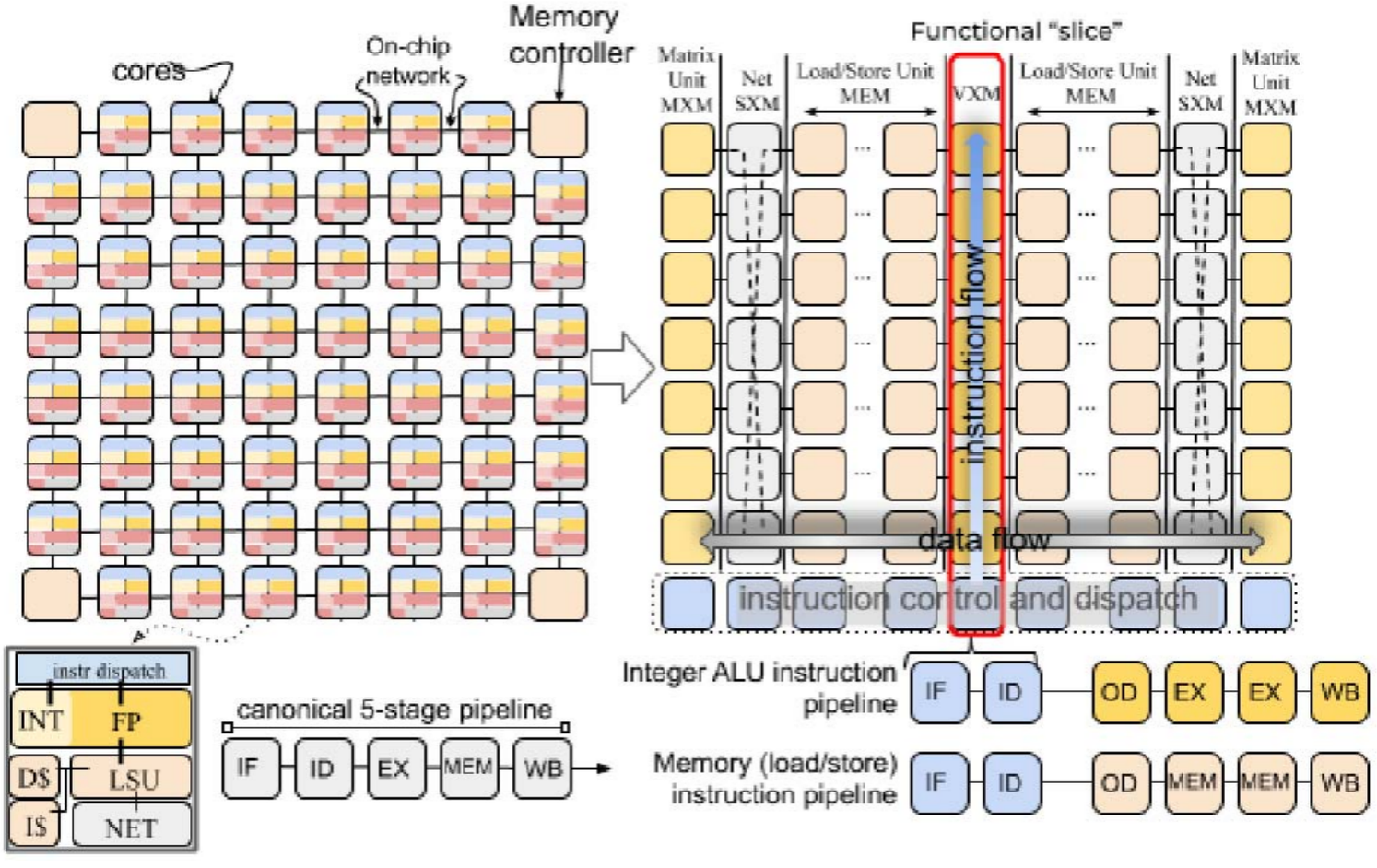
2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)

Quantum Information National Laboratory HUNGARY

Rethink the computational design

two-dimensional mesh of cores

organized into functionally sliced architecture: TSP



ELTE EÖTVÖS LORÁND UNIVERSITY

wigner

groq

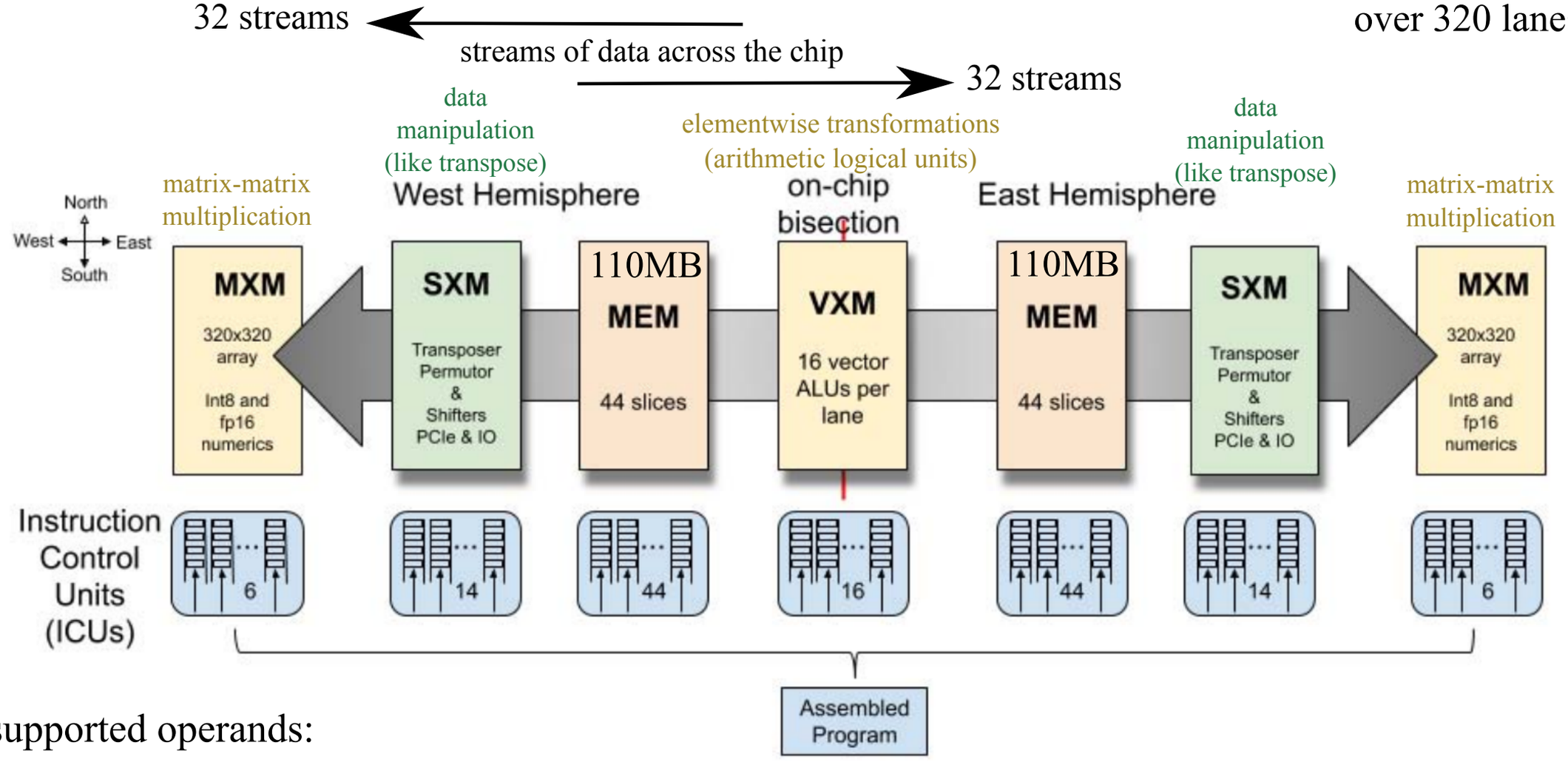
MAXELER Technologies Maximum Performance Computing

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

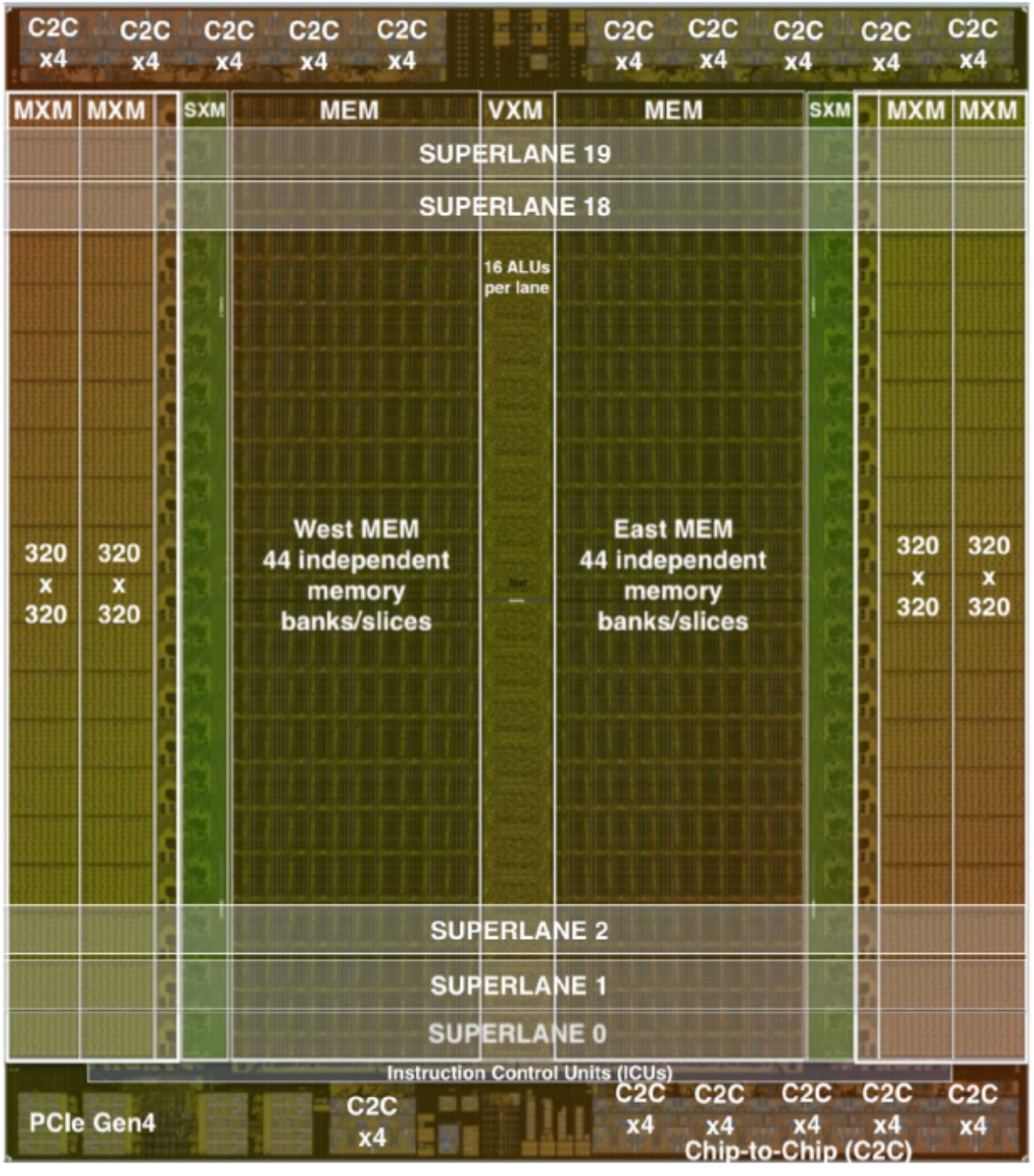
Wigner

groq

MAXELER Technologies Maximum Performance Computing

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

ELTE EÖTVÖS LORÁND UNIVERSITY

Wigner

groq

MAXELER Technologies
Maximum Performance Computing

VXM slices

Quantum Information National Laboratory HUNGARY

supported operations

- add, sub, mul, neg, exp2, log2, tanh, cast
- equal, not_equal, less, greater, bitwise_or
- max, min, left_shift, right_shift, mask,

supported operands

- UINT8, INT8, BOOL8
- UINT16, INT16, BOOL16
- BOOL32, UINT32, INT32
- FLOAT16, FLOAT32

Operations over ALUs can be chined up (pipelined)

- the connectivity between the ALUs is limited
- The ALU chain needs to be planned carefully

ELTE EÖTVÖS LORÁND UNIVERSITY

wigner

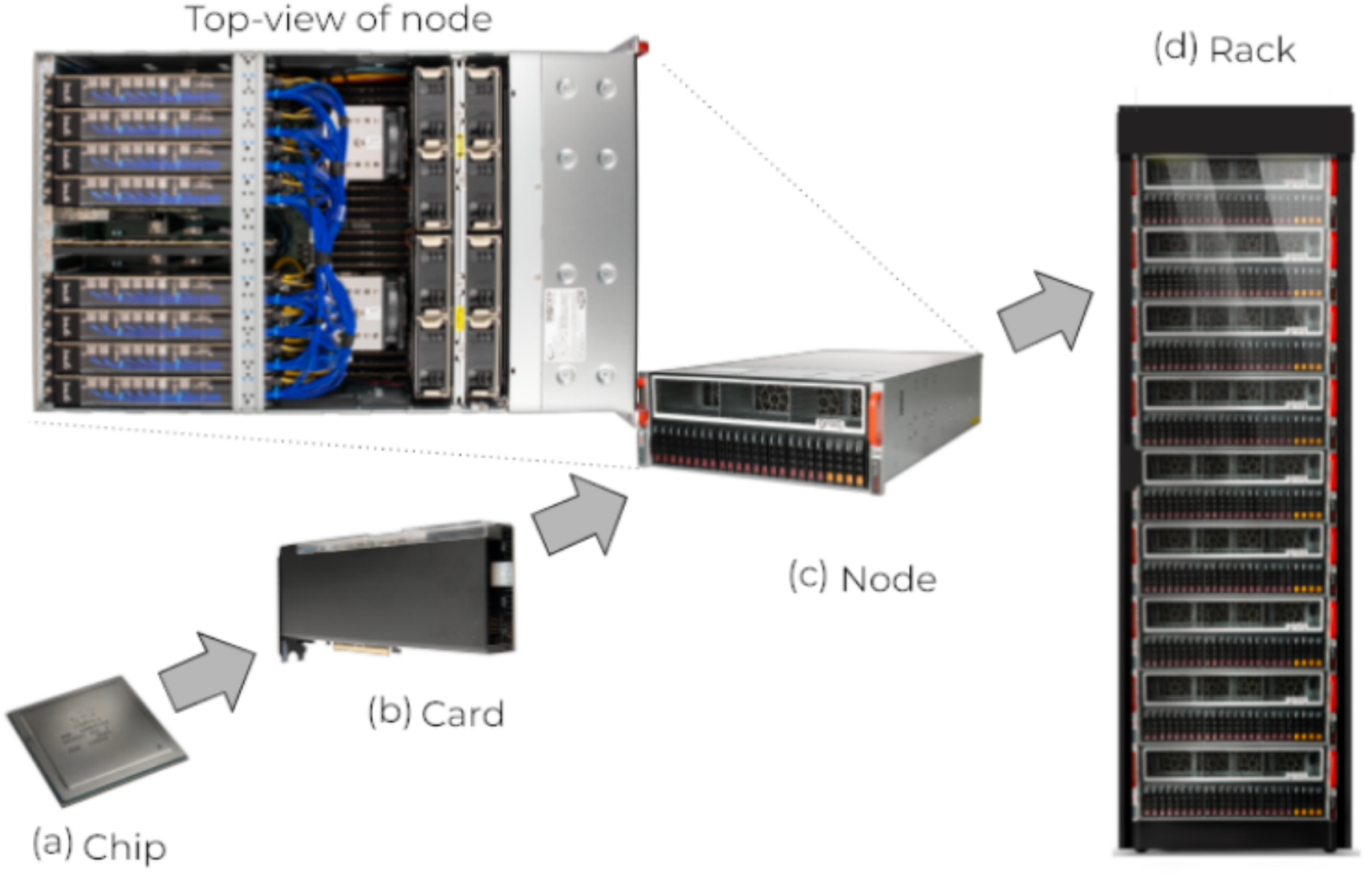
groq

MAXELER Technologies Maximum Performance Computing

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



ELTE EÖTVÖS LORÁND UNIVERSITY

Wigner

groq

MAXELER Technologies Maximum Performance Computing

Groq adapts Meta's chatbot for its own chips in race against Nvidia

The move is significant because Meta's researchers originally developed LLaMA using chips from Nvidia Corp , which has a market share of nearly 90% for AI computing according to some estimates. Showing that a cutting-edge model can be moved to Groq's chips easily could help the startup prove that its products are a viable alternative to Nvidia.

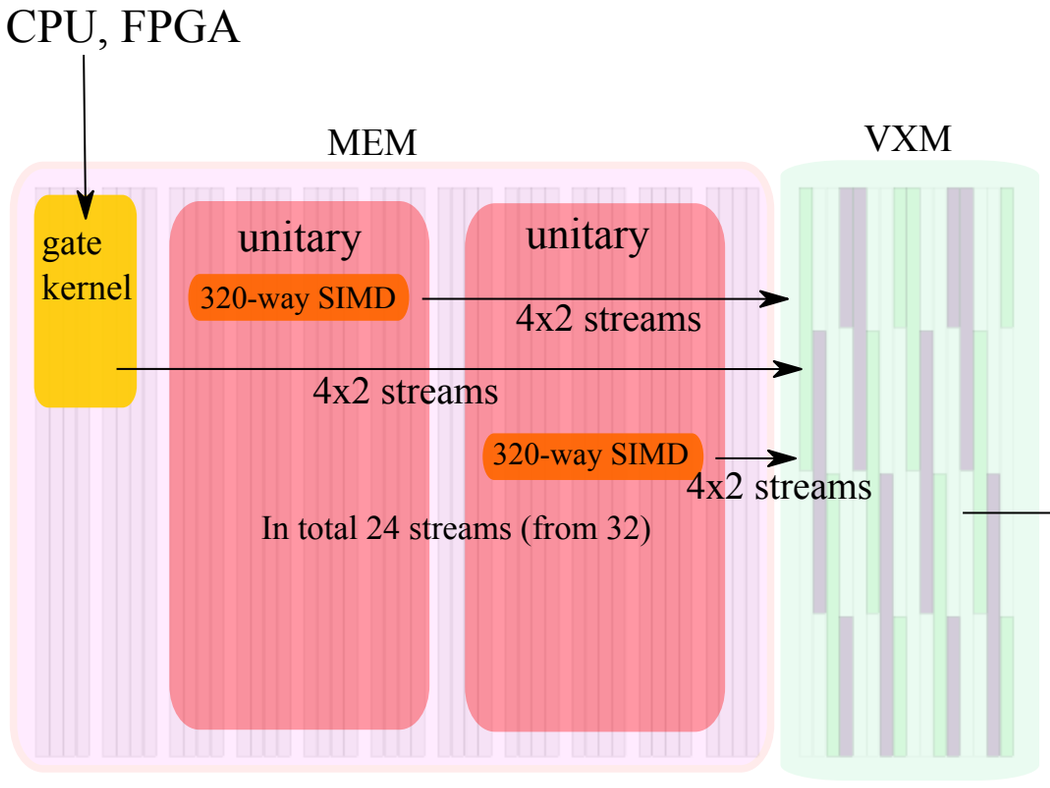
US Army Validation Report Confirms Entanglement AI Cybersecurity Solution on Groq Tech

October 25, 2022

MOUNTAIN VIEW, Calif., Oct. 25, 2022 — The United States Army has released a Validation Report confirming that Entanglement AI's cybersecurity solution on Groq technology – specifically a **GroqNode** – and simultaneously using quantum and classical algorithms for anomaly detection, is running the world's fastest Quadratic Unconstrained Binary Optimization (QUBO) Solver, noted in the report as "dramatically faster and more accurate… – with far fewer false positives – than any known technology."

Resource planing of Groq QC simulator

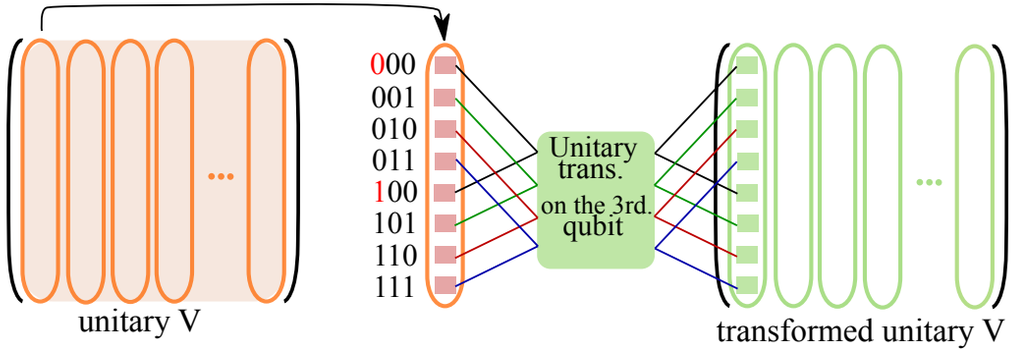
Quantum Information National Laboratory HUNGARY



$$C_\alpha = U_{\alpha,0}C_0 + U_{\alpha,1}C_1$$

- 8x 32-bit float multiplications
- 6x 32-bit float additions
- in toal: 14 ALU units (from 16)

unitary transformation:



multiplication of complex numbers

$$A \times B = (a_0 + \mathrm{i}a_1) \times (b_0 + \mathrm{i}b_1) =$$
$$(a_0b_0 - a_1b_1) + \mathrm{i}(a_0b_1 + a_1b_0)$$

4x multiplications and
2x additions

ELTE EÖTVÖS LORÁND UNIVERSITY

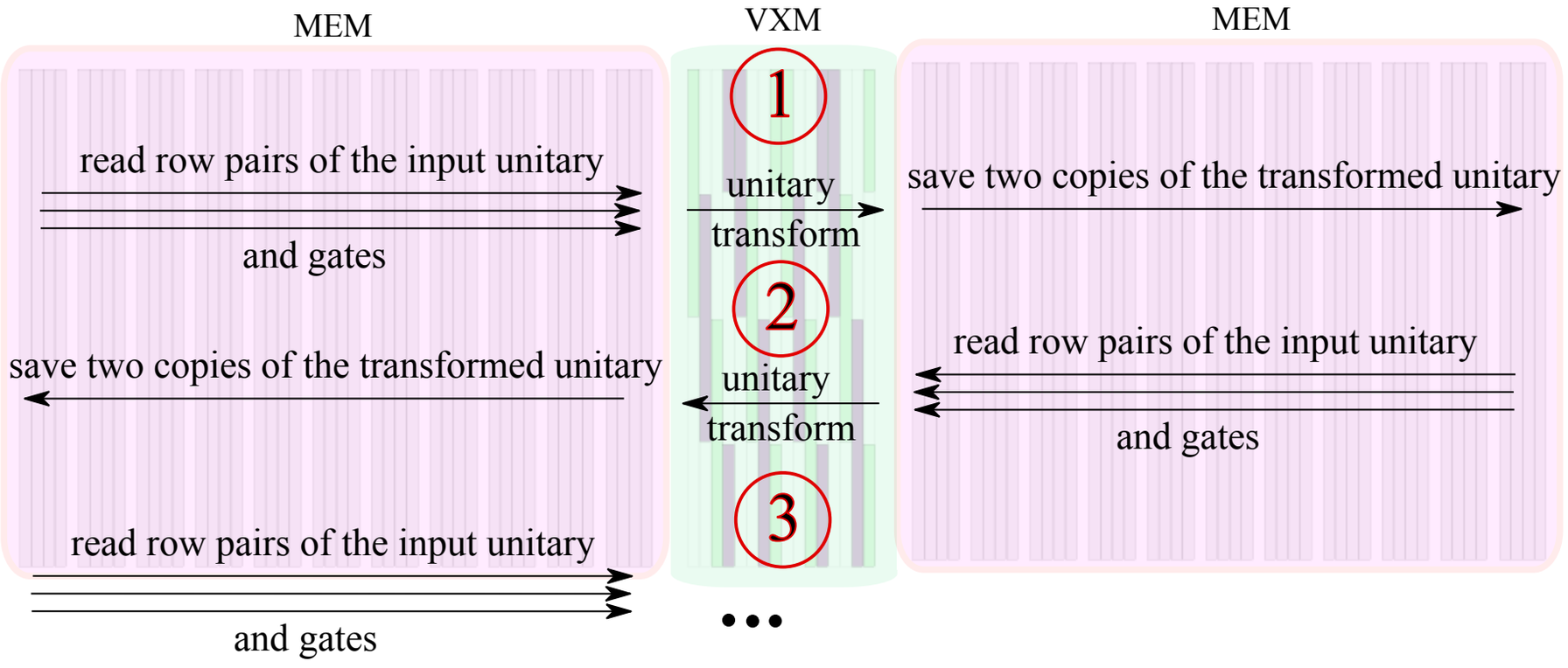
WIGNER

groq

MAXELER Technologies Maximum Performance Computing

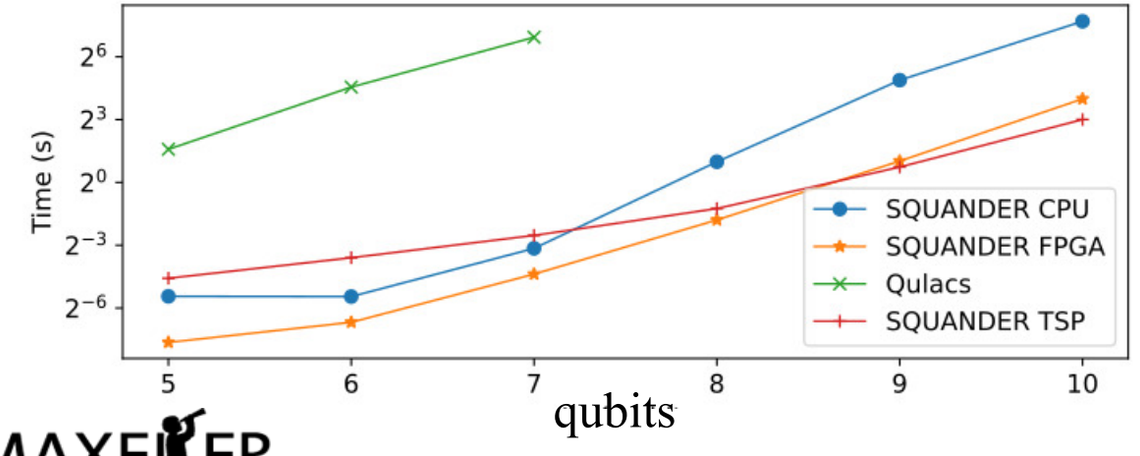
Sequence of quantum gate operations

Quantum Information National Laboratory HUNGARY



- alternating application of the quantum gates on the input unitary
- gates are distributet between the hemispheres during the initial IO

Performance benchmark:



ELTE EÖTVÖS LORÁND UNIVERSITY

Wigner

groq

MAXELER Technologies Maximum Performance Computing

Issue of long compilation

Quantum Information National Laboratory HUNGARY

- gates with different target/control qubits involves different row pairs
  - need different gate program for each target/control qubit?
    - each circuit needs to be compiled individually
      - the compilation takes to much time: ~mins
        - not practical in most of the use casaes
  - to invoke single program takes to much time ~ms
    - need to chain up gate programs to amortize the init IO

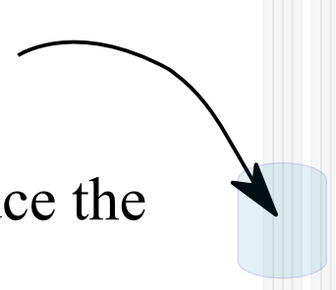
ELTE EÖTVÖS LORÁND UNIVERSITY · WIGNER · groq · MAXELER Technologies Maximum Performance Computing

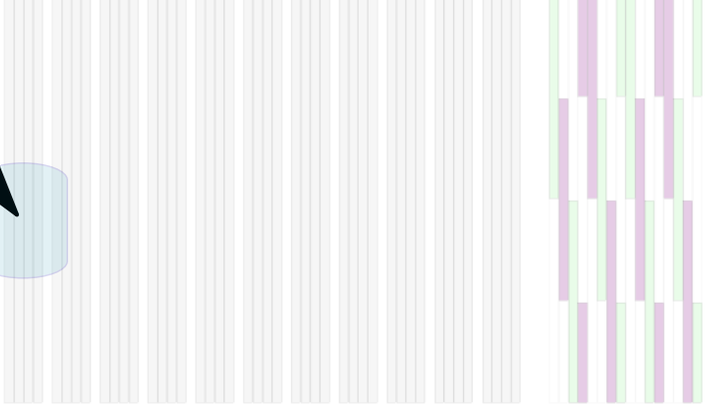
Implementation for universal gate operations

Quantum Information National Laboratory HUNGARY

- to resolve the long-compilation issue we designed a general gate implementation

- chain up multiple **general** gate operations to amortize the IO overhead
- How to design a general gate implementation?
  - memory gather/scatter via memory maps
    - upload precalculated memory maps to determine which data need to be gathered from the memory to produce the correct row-pair streams





ELTE EÖTVÖS LORÁND UNIVERSITY

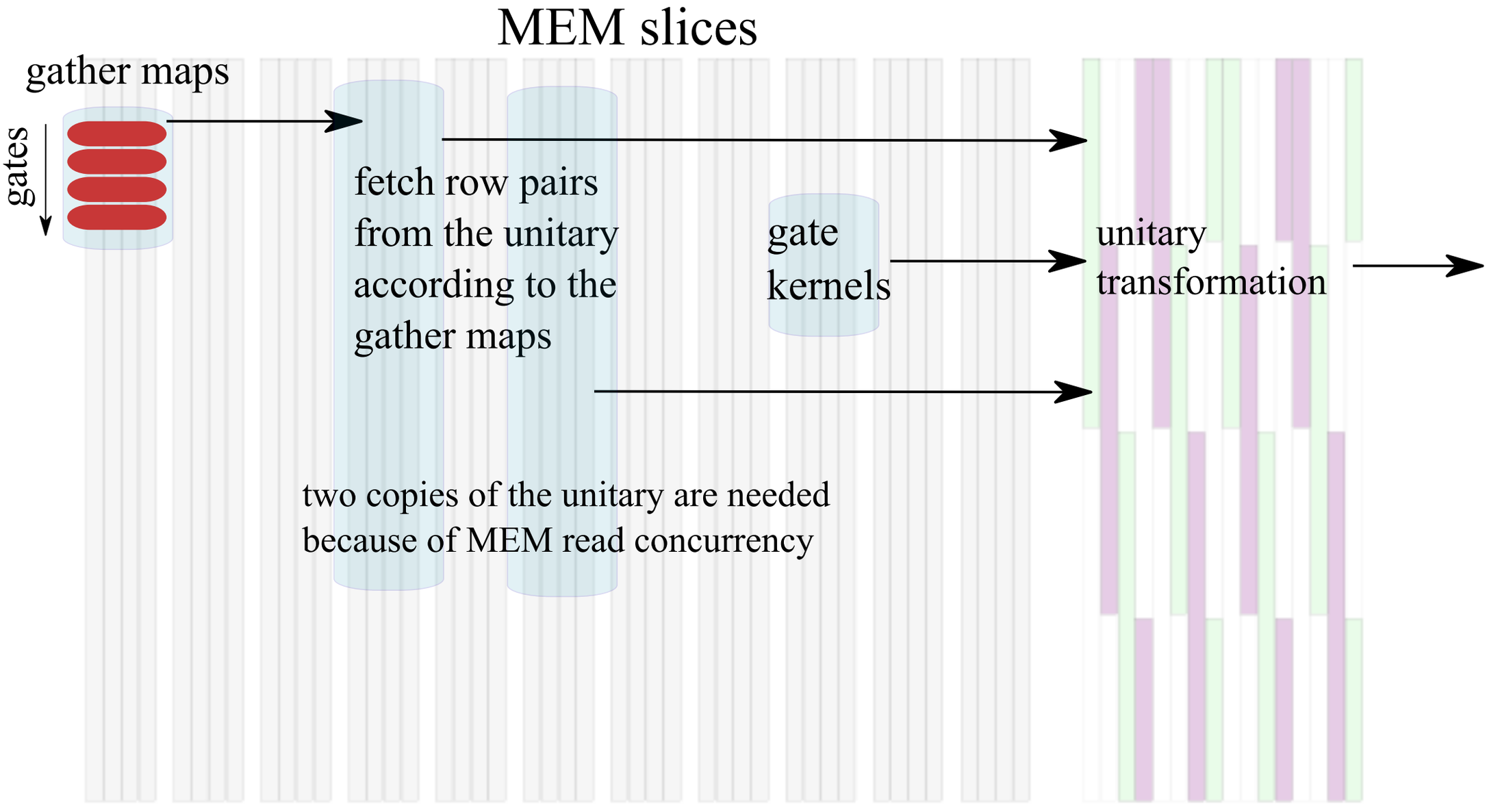
wigner

groq

MAXELER Technologies Maximum Performance Computing

Implementation for universal gate operations

Quantum Information National Laboratory HUNGARY



Indirect memory addressing through maps is a powerful tool on the Groq chip

ELTE EÖTVÖS LORÁND UNIVERSITY

WIGNER

groq

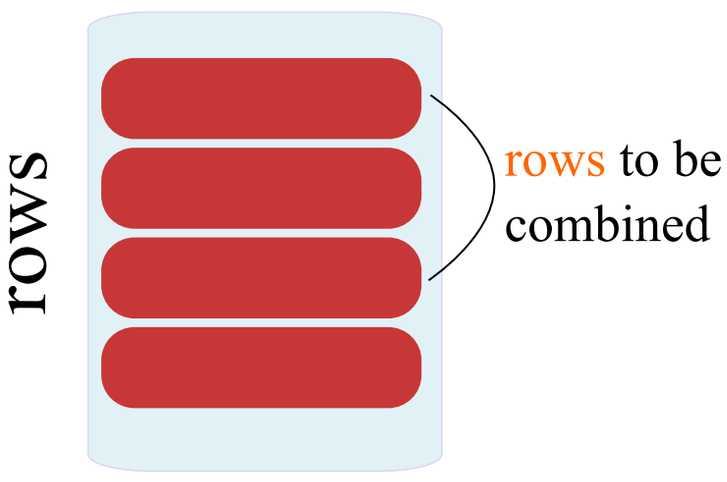
MAXELER Technologies Maximum Performance Computing

State-vector QC simulation on the Groq chip

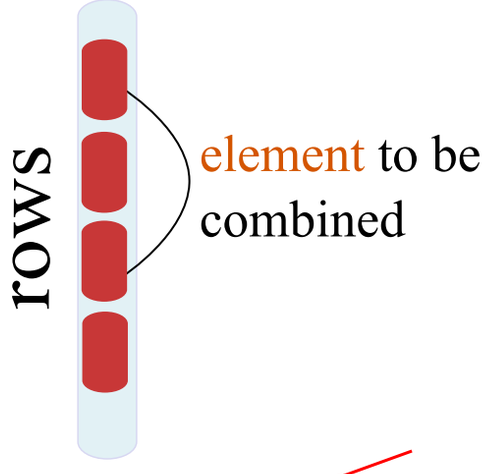
Quantum Information National Laboratory HUNGARY

- unitary simulator VS state-vector sumulator

$$2^n \times 2^n \qquad 2^n \times 1$$



320-way SIMD over the elements in the columns



~~320-way SIMD over the elements in the columns~~

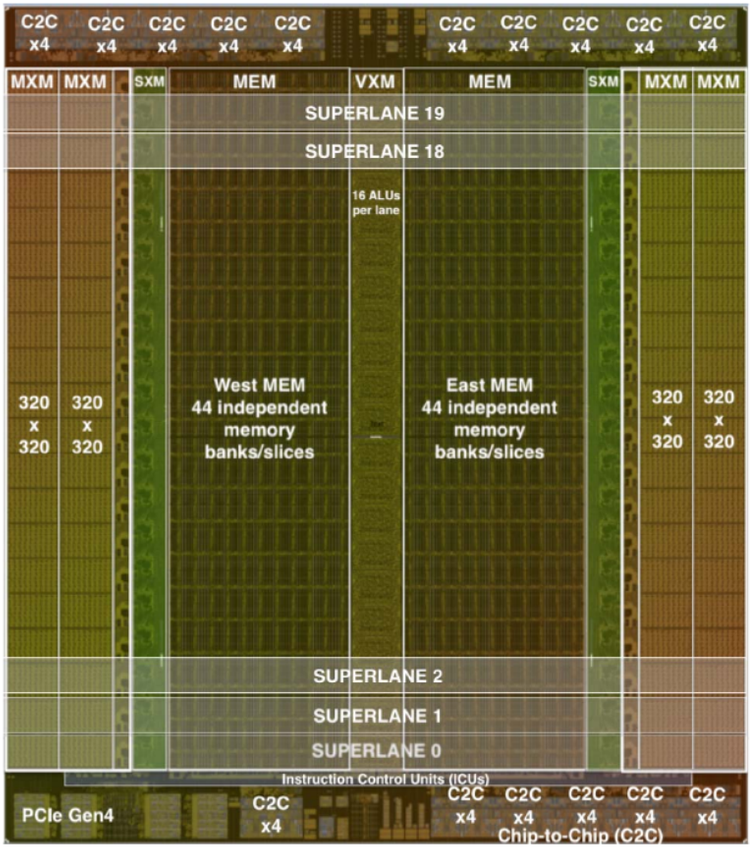
The elements of the state vector need to be permuted within a vector

ELTE EÖTVÖS LORÁND UNIVERSITY · Wigner · groq · MAXELER Technologies Maximum Performance Computing

State-vector QC simulation on the Groq chip

Quantum Information National Laboratory HUNGARY

permutation/shifter slices
to modify the elements within a 320-vector



permuted
state-vector

gate
kernels

unitary
transformation

state-vector

ELTE EÖTVÖS LORÁND UNIVERSITY

wigner

groq

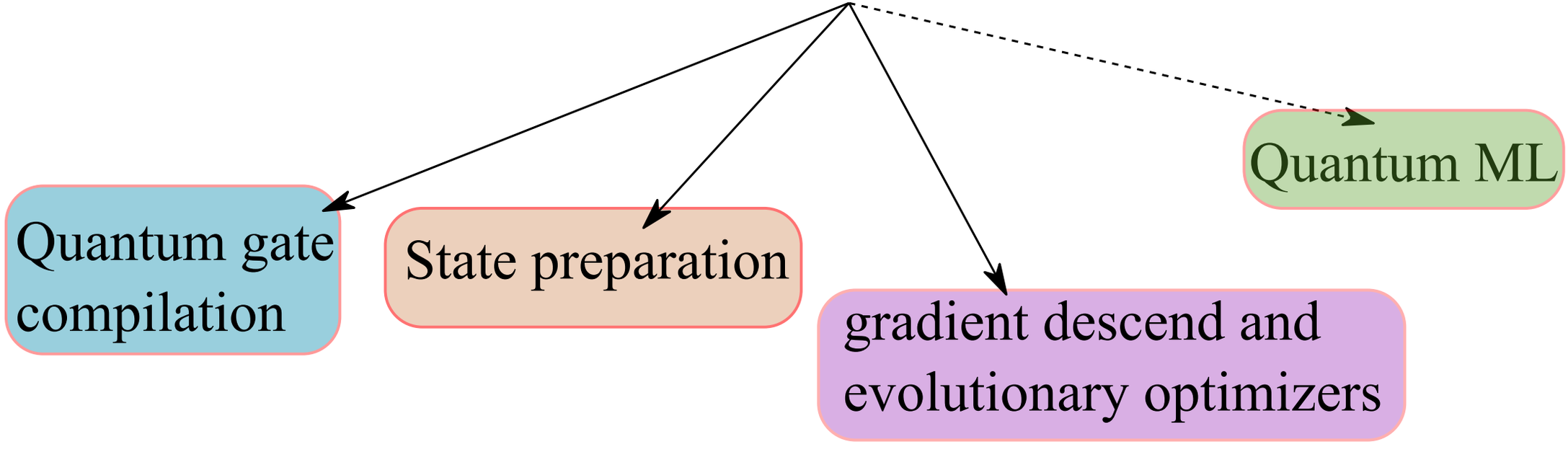
MAXELER Technologies
Maximum Performance Computing

Conclusions and outlooks

Quantum Information National Laboratory HUNGARY

- The Groq unitary simulator integrated into SQUANDER:

Sequantial Quantum Gate Decomposer



GitHub https://github.com/rakytap/sequential-quantum-gate-decomposer

ELTE EÖTVÖS LORÁND UNIVERSITY

WIGNER

groq

MAXELER Technologies Maximum Performance Computing

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. 2020-2.1.1-ED-2021-00179, by the ÚNKP-22-5 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, peter.rakyta@ttk.elte.hu