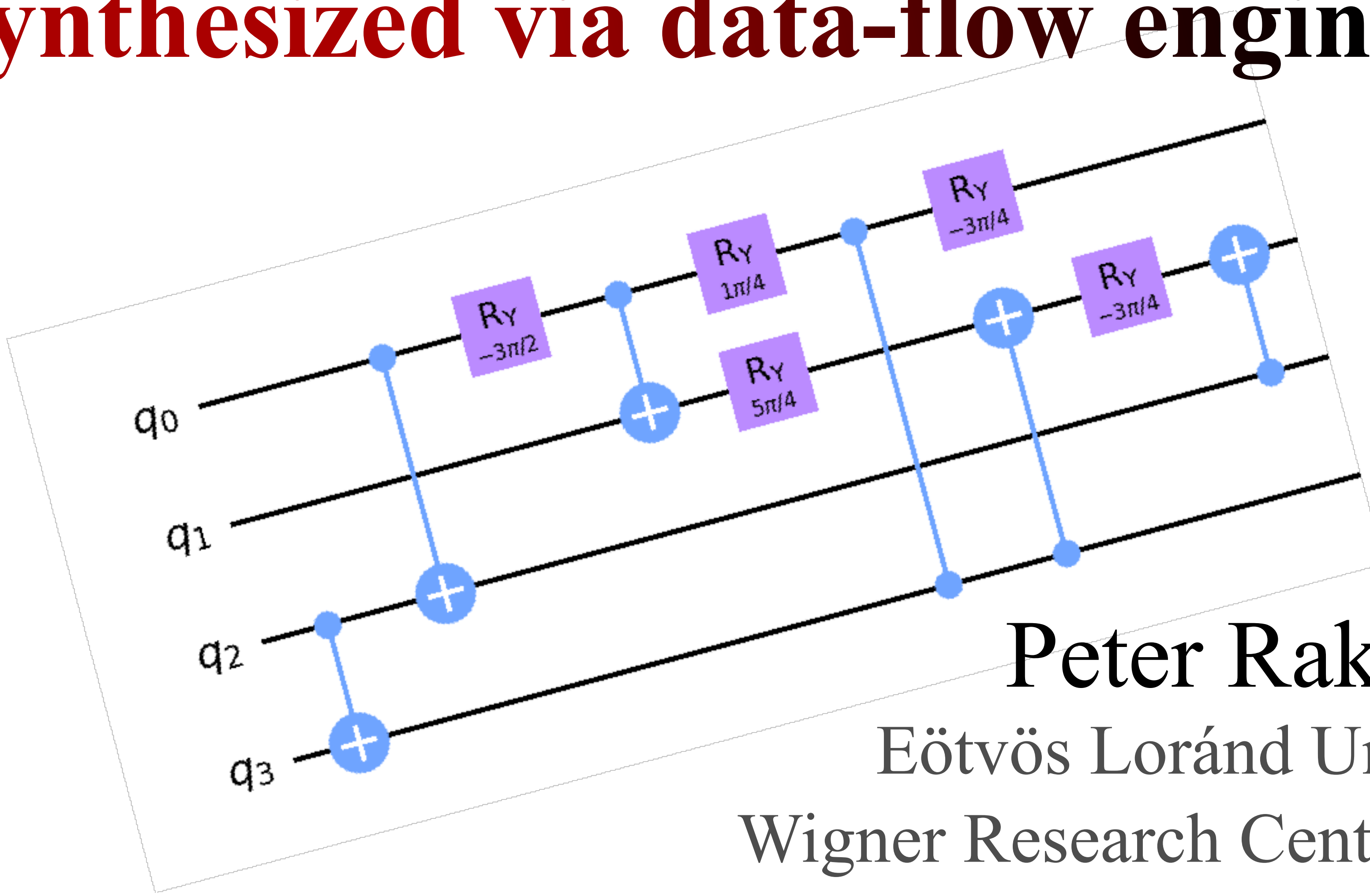
Quantum Information
National Laboratory
HUNGARY

Highly optimized quantum circuits synthesized via data-flow engines



Peter Rakyta
Eötvös Loránd University,
Wigner Research Centre for Physics
Budapest, Hungary

ELTE EÖTVÖS LORÁND UNIVERSITY
wigner
groq
MAXELER Technologies Maximum Performance Computing

Laboratory of Quantum Computer Simulators

Quantum Information National Laboratory HUNGARY

Zoltán Zimborás
- many-body phisics
- quantum computing
- quantum information

Peter Rakyta
- condensed matters
- parallel, hardware oriented programming

Gregory Morse
- software engineer
- machine learning
- parallel, hardware oriented programming

Tamás Kozsik
- functional programming
- programming languages

Zoltán Kolarovszki
- software engeneering
- optical quantum computing
- Python programming

Gábor Vattay
- Complex Systems
- Quantum chaos

Ágoston Kaposi
- algebraic and differential topology
- mathematical network modelling
- C++, python programming

ELTE EÖTVÖS LORÁND UNIVERSITY
Wigner
groq
MAXELER Technologies Maximum Performance Computing

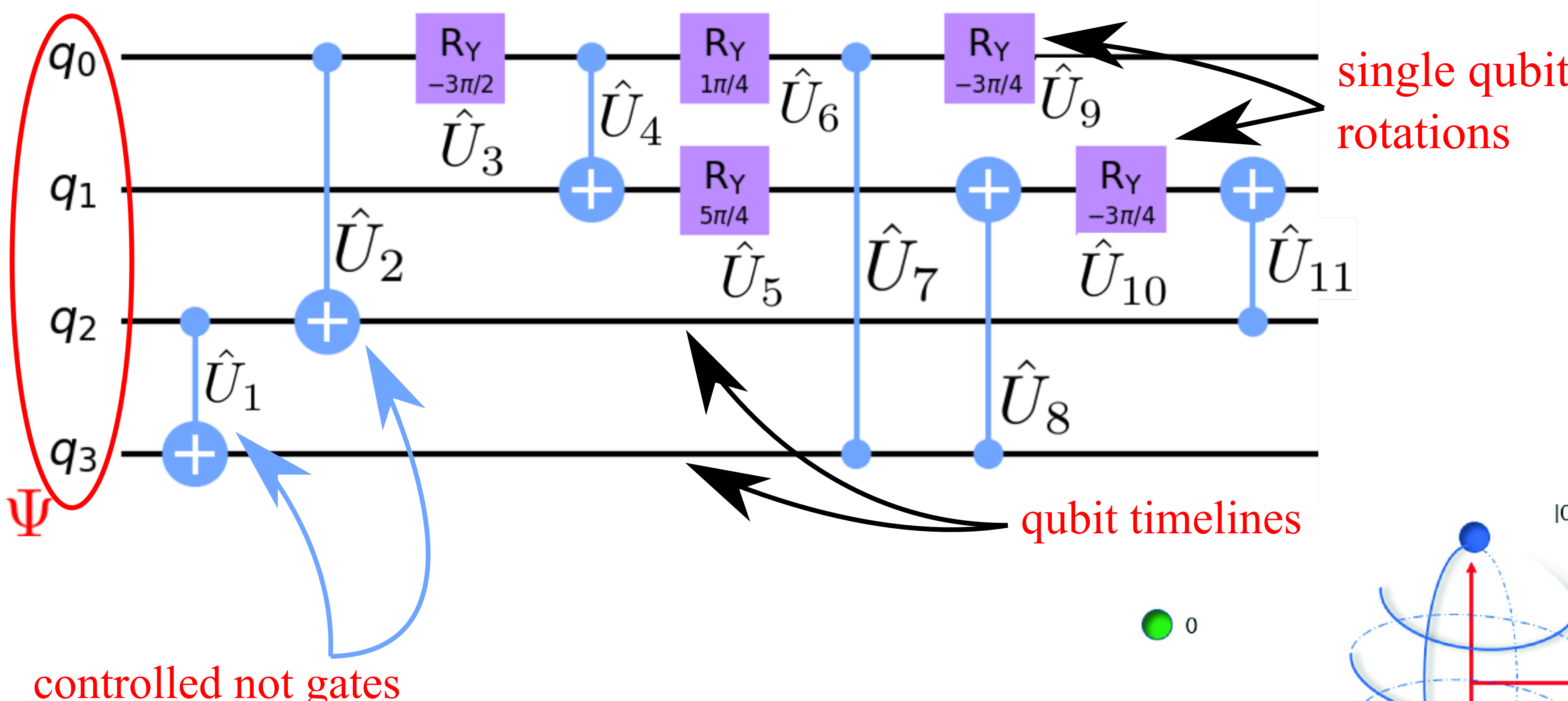
Quantum gate decomposition

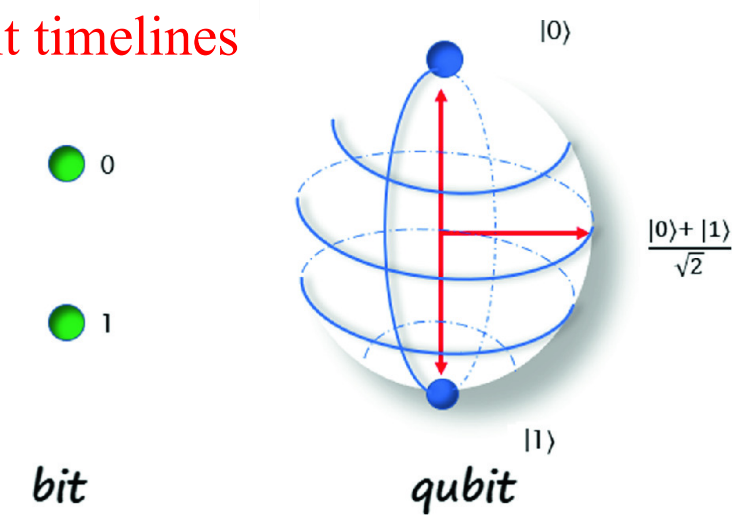
Quantum Information National Laboratory HUNGARY

quantum program (unitary): $UU^{\dagger} = \mathbf{1}$

preserves the norm of the state: $\langle U\Psi|U\Psi\rangle = \langle\Psi|U^{\dagger}U|\Psi\rangle = \langle\Psi|\Psi\rangle$

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





Optimized quantum circuit synthesis

Quantum Information
National Laboratory
HUNGARY

How to find an optimal gate decomposition?

- fewest gate count?
- smallest depth?

Available gate decomposition utilities:

- Quantum Fast Approximate Synthesis Tool (QFAST)
- QSearch + LEAP

(Lawrence Berkeley National Laboratory)

- UniversalQCompiler (incorporated into QISKIT)

(ETH Zürich, University of York, TUM)

T|ket>: A Retargetable Compiler for NISQ Devices

(Cambridge Quantum Computing Ltd., University of Strathclyde

How close is an approximation to the exact one?

exact evolution: U

$$|\psi(U)\rangle := U|\psi\rangle$$

approximate evolution: V

$$|\psi(V)\rangle := V|\psi\rangle$$

The fidelity of the approximation:

$$\overline{F}(U,V) := \int_{\psi} |\langle\psi(V)|\psi(U)\rangle|^2 \, \mathrm{d}\psi$$

average taken over the Haar distribution

The cost function of the optimization:

Hilbert-Schmidt test:

$$C_{HST}(U,V) = 1 - \frac{1}{d^2}\left|\mathrm{Tr}\left(V^\dagger U\right)\right|^2$$

size of the matrices

$$\overline{F}(U,V) = 1 - \frac{d}{d+1} C_{HST}(U,V)$$

for exact decmposition:

$$C_{HST}(U,V) = 0 \quad \overline{F}(U,V) = 1$$

Quantum Information
National Laboratory
HUNGARY

How close is an approximation to the exact one?

Frobenius-norm based fidelity

$$\|A\|_{\mathrm{F}} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}}$$

The cost function of the optimization:

$$f(U,V) = \frac{1}{2} \|V - U\|_F^2 = d - \mathrm{Re}\left[\mathrm{Tr}(U^\dagger V)\right]$$

The Fidelity:

$$\overline{F}_F(U,V) = 1 - \frac{d}{d+1} + \frac{1}{d(d+1)} \left(d - f(U,V)\right)^2$$

$$\overline{F}_F(U,V) \leq \overline{F}(U,V)$$

"Best Approximate Quantum Compiling Problems"

Liam Madden (University of Colorado),
Andrea Simonetto (IBM Research Ireland)
arXiv:2106.05649

Quantum Information
National Laboratory
HUNGARY

Heuristics for Quantum Compiling with a Continuous Gate Set (QSearch)

Quantum Information National Laboratory HUNGARY

arXiv > cs > arXiv:1912.02727

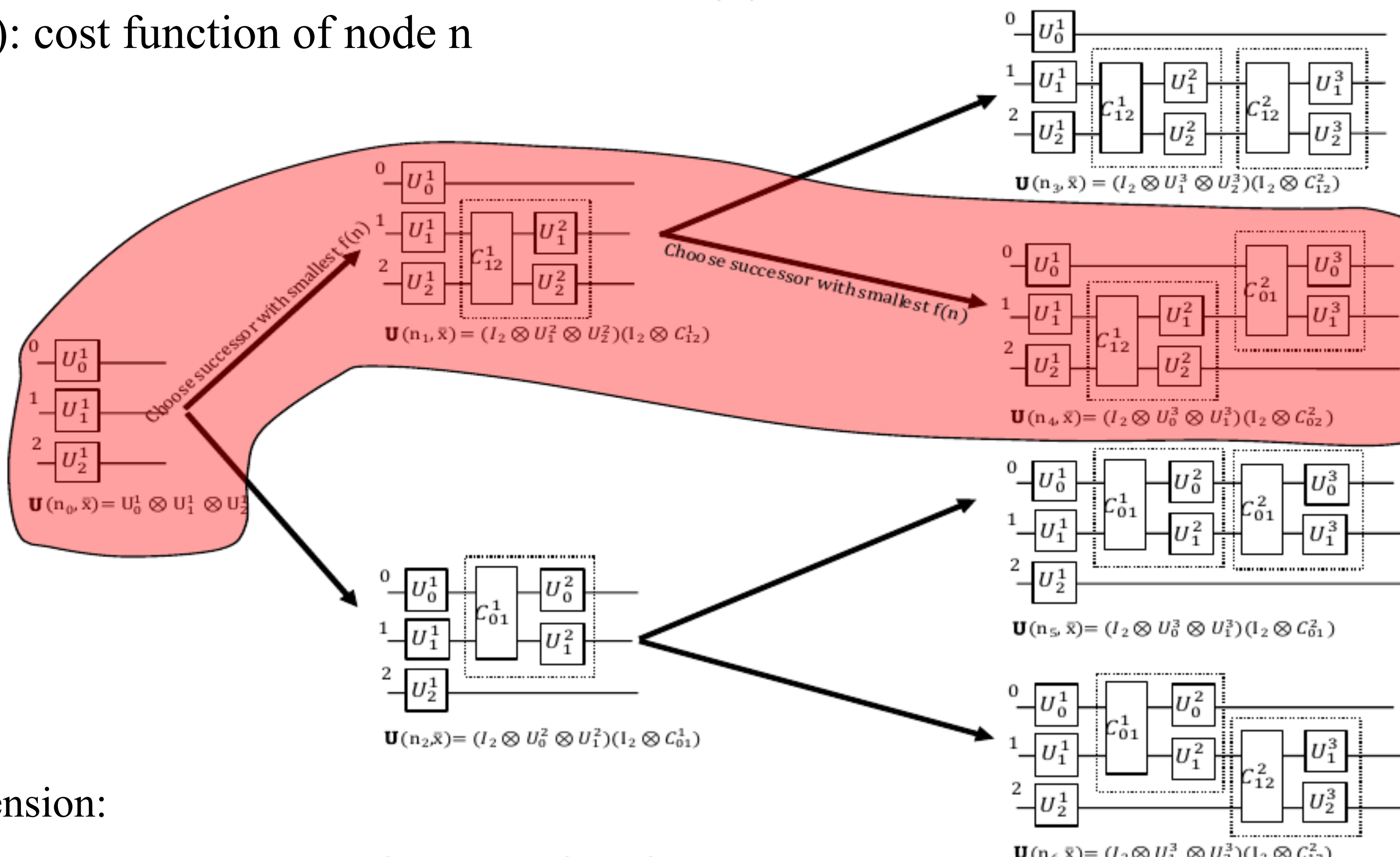
a) Optimization of continuous parameters in U3 gates

b) **Combinatorial optimization problem** of gate structure

partial solutions in graph nodes

f(n): cost function of node n

(A)



Leap extension:

ACM Transactions on Quantum Computing • Accepted on June 2022 • https://doi.org/10.1145/3548693

Quantum Fast Approximate Synthesis Tool (QFAST)

Quantum Information National Laboratory HUNGARY

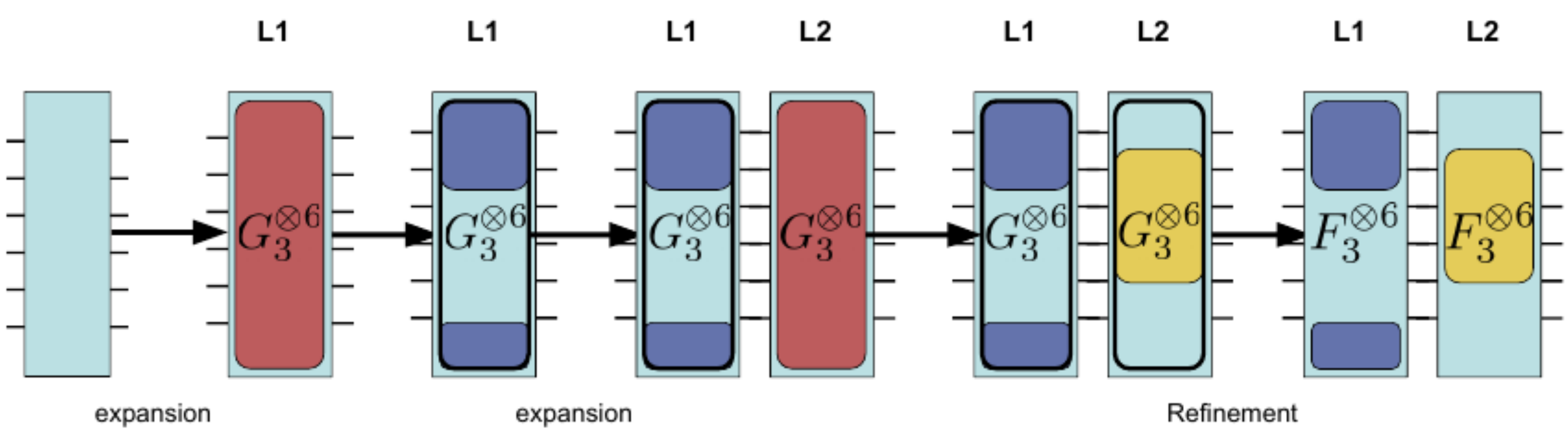
arXiv > quant-ph > arXiv:2003.04462

n-qubit unitary $\longrightarrow \frac{n}{2}$ -qubit unitaries $\longrightarrow \frac{n}{4}$ -qubit unitaries

$\longrightarrow$ 1 and 2-qubit unitaries

$$U(2^n) = \{e^{\mathrm{i}(\boldsymbol{\alpha}\cdot\boldsymbol{\sigma}^{\otimes n})} \mid \boldsymbol{\alpha} \in \mathbb{R}^{4^n}\}$$

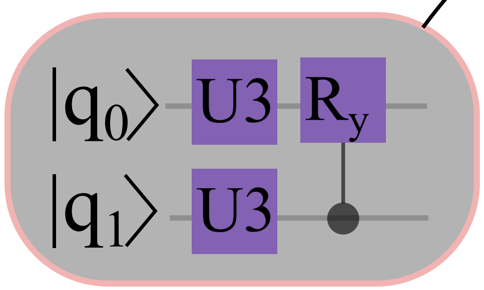
$$\boldsymbol{\sigma}^{\otimes n} = \{\sigma_j \otimes \sigma_k \mid \sigma_j \in \boldsymbol{\sigma}, \sigma_k \in \boldsymbol{\sigma}^{\otimes n-1}\} \qquad \boldsymbol{\sigma} = \{\sigma_i, \sigma_x, \sigma_y, \sigma_z\}$$



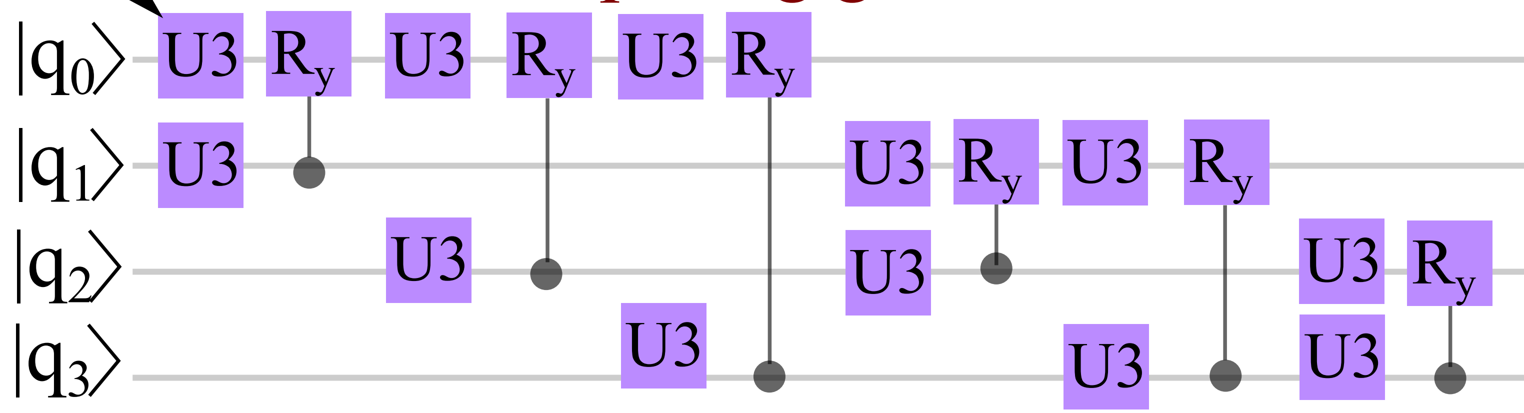
Adaptive quantum gate decomposition (SQUANDER)

Quantum Information National Laboratory HUNGARY

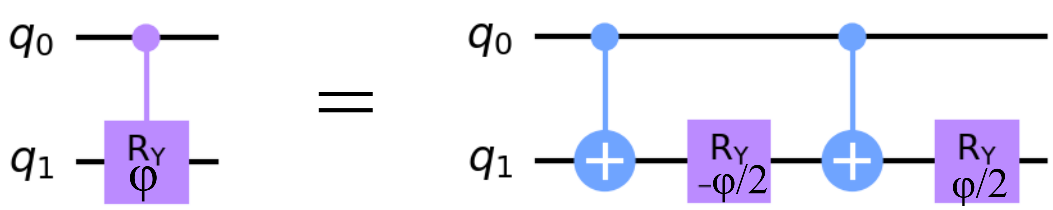
Bulding block



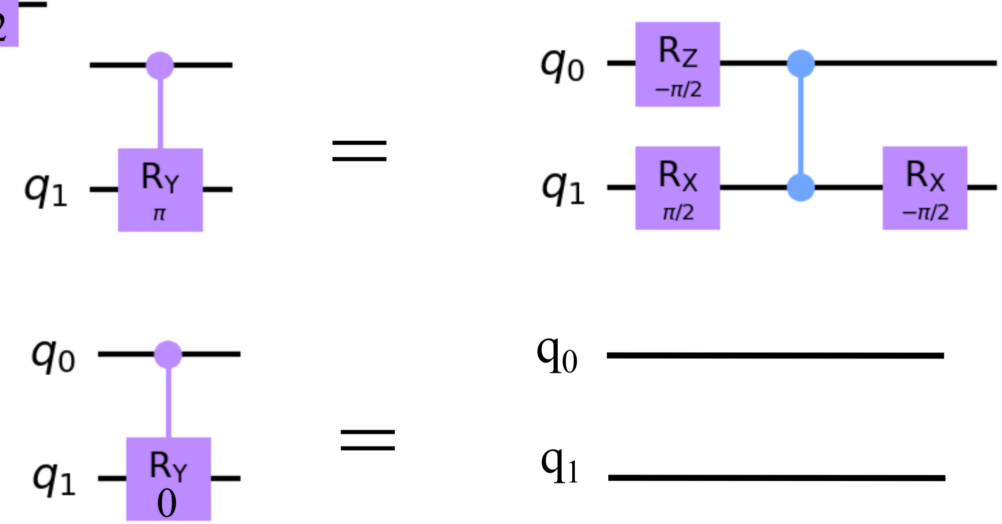
Decomposing gate structure



Expansion of controlled R_y rotations



Special case



GitHub

arXiv > quant-ph > arXiv:2203.04426

SQUANDER vs QFAST vs QSearch

Quantum Information National Laboratory HUNGARY

Journals & Magazines > IEEE Transactions on Computer... > Volume: 38 Issue: 7

An Efficient Methodology for Mapping Quantum Circuits to the IBM QX Architectures

Publisher: IEEE Cite This PDF

Alwin Zulehner ; Alexandru Paler ; Robert Wille All Authors

- In the benchmark we tested the decomposition of 3, 4 and 5-qubit unitaries from online database containing series of circuits published as part of the Qiskit Developer Challenge, a public competition to design a better routing algorithm.

- quantum circuits of well known algorithms:
  - Grover search,
  - Quantum Fourier Transformation (QFT)
  - Quantum Approximate Optimization Algorithm (QAOA),
  - Quantum variational eigensolver (VQE)

GitHub https://github.com/iic-jku/ibm_qx_mapping

Gate synthesis benchmark

Quantum Information National Laboratory HUNGARY

File name	Initial	QISKIT	SQUANDER		QFAST		QSEARCH	
	$CNOT$	$CNOT$	$CNOT$	$\overline{T}\,[s]$	$CNOT$	$\overline{T}\,[s]$	$CNOT$	$\overline{T}\,[s]$
4gt5_77	58	**338**	**23**	1293	**26**	332	-	-
4gt13_91	49	**187**	**23**	1296	**25**	732	**48**	2324
ham3_102	11	**15**	**6**	4.9	**7**	3.2	**8**	2.6
4gt5_76	46	**529**	**24**	1711	**29**	476	-	-
alu-v0_26	38	**204**	**23**	7900	**42**	912	**29**	9284
miller_11	23	**18**	**8**	7	**9**	5.4	**10**	4.5
rd32_v1_68	16	**66**	**9**	23.9	**13**	21.6	**13**	615
4mod5-v0_20	10	**526**	**9**	3650	**17**	166	**16**	14508
alu-v0_27	17	**212**	**17**	3452	**30**	674	**34**	3801
mod5mils_65	16	**73**	**12**	11162	**20**	405	-	-
ex-1_166	9	**20**	**9**	4.4	**8**	4.7	**8**	5.9
decod24-v1_41	38	**130**	**20**	2414	**36**	413	**24**	349
alu-v3_34	24	**237**	**25**	6090	**37**	1814	**27**	7834
3_17_13	17	**23**	**7**	6.5	**9**	4.2	**9**	4.3
4gt11_84	9	**163**	**9**	642	**20**	318	-	-
decod24-v0_38	23	**48**	**14**	62	**23**	58	**15**	285
4mod5-v0_19	16	**75**	**13**	701	**21**	375	-	-
4mod5-v1_22	11	**168**	**9**	962	**13**	52	**17**	82

gate fidelity: $\overline{F}_F = 1 - \varepsilon \qquad \varepsilon \approx 10^{-9}$

Gate synthesis benchmark

Quantum Information National Laboratory HUNGARY

File name	Initial	QISKIT	SQUANDER		QFAST		QSEARCH	
	$CNOT$	$CNOT$	$CNOT$	$\overline{T}\,[s]$	$CNOT$	$\overline{T}\,[s]$	$CNOT$	$\overline{T}\,[s]$
alu-v1_29	17	**240**	**19**	3820	**33**	801	-	-
alu-v1_28	18	**331**	**19**	2488	**36**	607	-	-
4mod5-v1_23	32	**74**	**13**	946	**40**	702	-	-
4mod5-v0_18	31	**671**	**15**	1134	**31**	266	-	-
rd32_270	36	**522**	**14**	893	**27**	627	-	-
rd32-v0_66	16	**66**	**10**	29	**16**	25	**13**	443
alu-v3_35	18	**249**	**20**	3655	**31**	1050	-	-
4gt13-v1_93	30	**218**	**23**	2408	**38**	466	**33**	21315
4mod5-v1_24	16	**241**	**14**	5081	**33**	210	**52**	3968
mod5d1_63	13	**76**	**13**	867	**29**	304	-	-
alu-v4_36	51	**193**	**40**	11090	**49**	2343	-	-
4gt11_82	18	**419**	**15**	883	**22**	698	**19**	1003
4gt5_75	38	**259**	**25**	7002	**37**	429	**49**	33246
alu-v2_33	17	**358**	**17**	2339	**31**	665	**23**	6520
4gt11_83	14	**151**	**13**	1994	**15**	98	**19**	1107
decod24-v2_43	22	**46**	**9**	93	**19**	44	**17**	1390
4gt13_92	30	**161**	**24**	1767	**46**	1830	-	-
alu-v4_37	18	**276**	**18**	3509	**37**	837	**32**	2142
mod5d2_64	25	**129**	**14**	846	**26**	104	**16**	256

gate fidelity: $\overline{F}_F = 1 - \varepsilon \qquad \varepsilon \approx 10^{-9}$

Complexity analysis of the calculations

Quantum Information National Laboratory HUNGARY

The cost function of the optimization:

$$f(U,V) = \frac{1}{2}\|V-U\|_F^2 = d - \mathrm{Re}\left[\mathrm{Tr}(VU^\dagger)\right]$$

Number of gates

Number of qubits

The computational cost to evaluate $VU^\dagger$ is M x 4^n

trial circuit to synthesize U

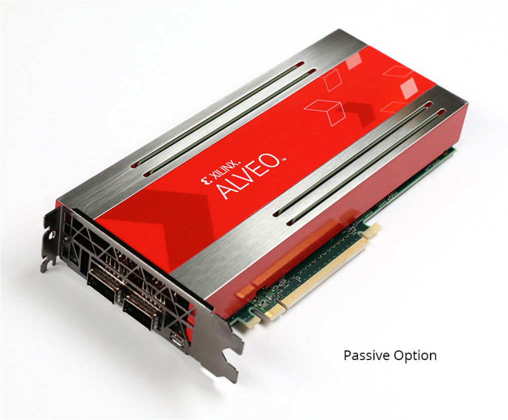
input quantum program 2^nx2^n

Gradient components also need to be calculated

ELTE EÖTVÖS LORÁND UNIVERSITY

MAXELER Technologies Maximum Performance Computing

Hardware accelerator



FPGA implementation of a quantum computer simulator

Quantum Information National Laboratory HUNGARY

computational concurrency
(on-chip multipliers used for multiplications)

number of supported qubits
(converted into on-chip memory usage)

reasonable trade-off

computational accuracy
(fixed point number representation, bitwidth)

- support for **arbitrary quantum circuit** composed of single qubit rotations and conditional two-qubit gates

- don't recompile the FPGA implementation when the gate structure is changed

- High level development framework of MAXELER Technologies Maximum Performance Computing

ELTE EÖTVÖS LORÁND UNIVERSITY wigner groq

Data-flow implementation of a quantum computer simulator

Quantum Information National Laboratory HUNGARY

Organize data into streams flowing through the chip

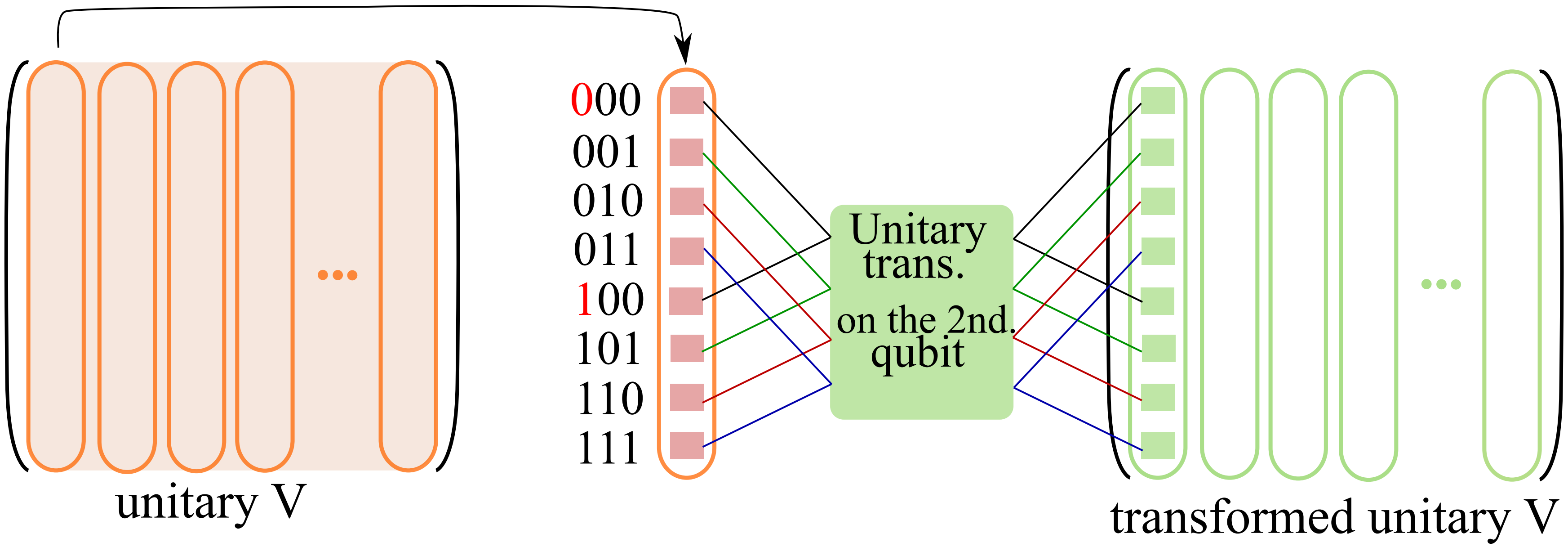
Computations: operations on the elements of a data stream

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

ELTE EÖTVÖS LORÁND UNIVERSITY wigner groq MAXELER Technologies Maximum Performance Computing

DFE flavour of quantum gate operations

Quantum Information
National Laboratory
HUNGARY

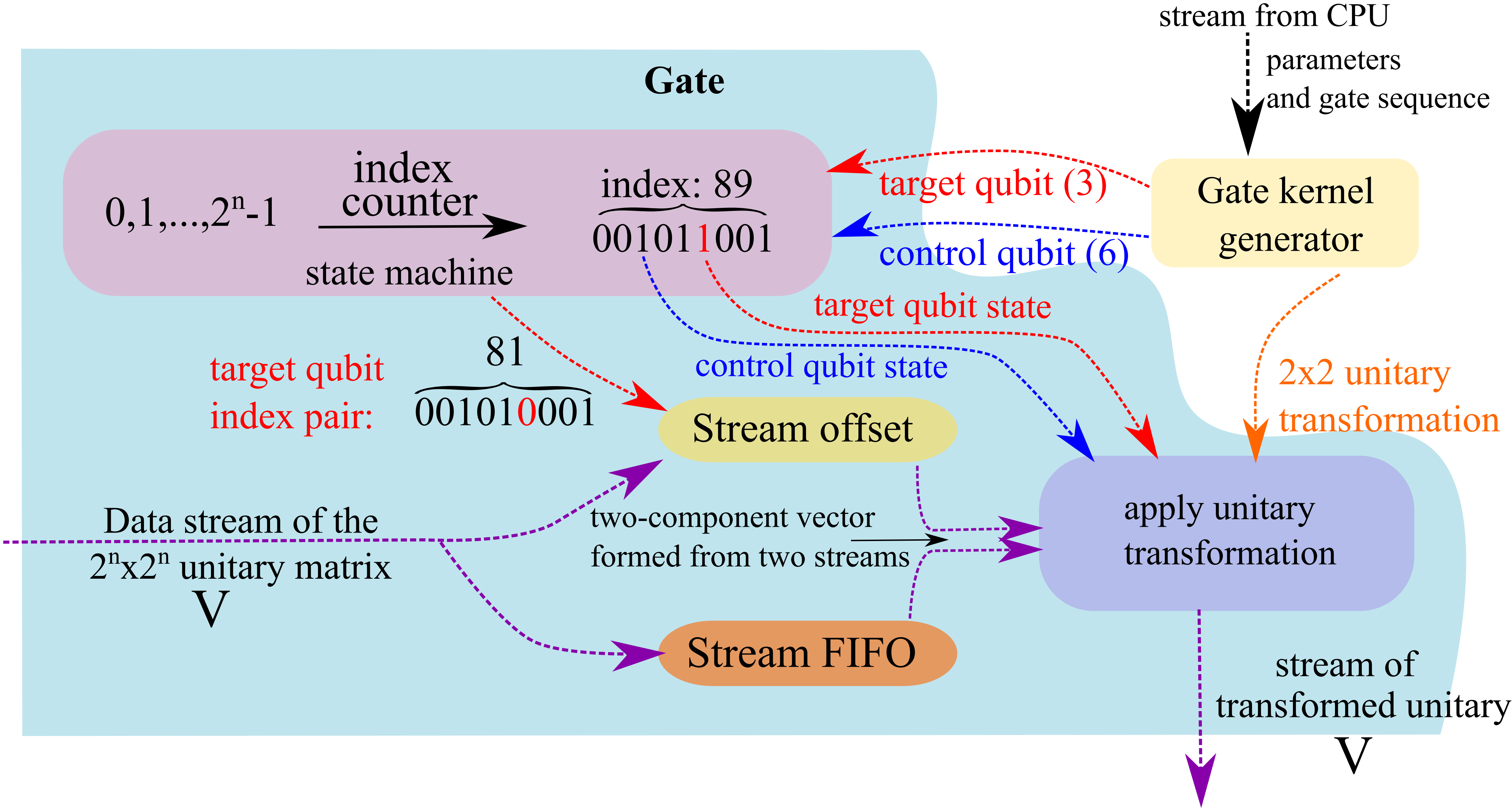


The elementary gate operations can be represented by **sparse unitaries**, mixing **element pairs** in the columns of V

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

DFE implementatio of quantum gate operations



Complexity of a gate operation

Quantum Information National Laboratory HUNGARY

Amplitude transformation

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

2 complex multiplication and 1 complex addition

digital signal processing (DSP) units for multiplicatios have input ports:

18 bit x 27 bit $\longrightarrow$ 32bit multiplications needs to be tiled

Karatsuba multiplication of 32-bit integers (W=16 bits)

$$A \times B = \left(a_1 2^W + a_0\right)\left(b_1 2^W + b_0\right) = \overset{16\times 16}{a_1 b_1} 2^{2W} + \underbrace{\left(a_1 b_0 + a_0 b_1\right)}_{\left(a_0+a_1\right)\left(b_0+b_1\right) - a_0 b_0 - a_1 b_1} 2^W + \overset{16\times 16}{a_0 b_0}$$

17×17

3 multiplications instead of 4
and 5 additions

Use Karatsuba strategy for complex multiplications as well

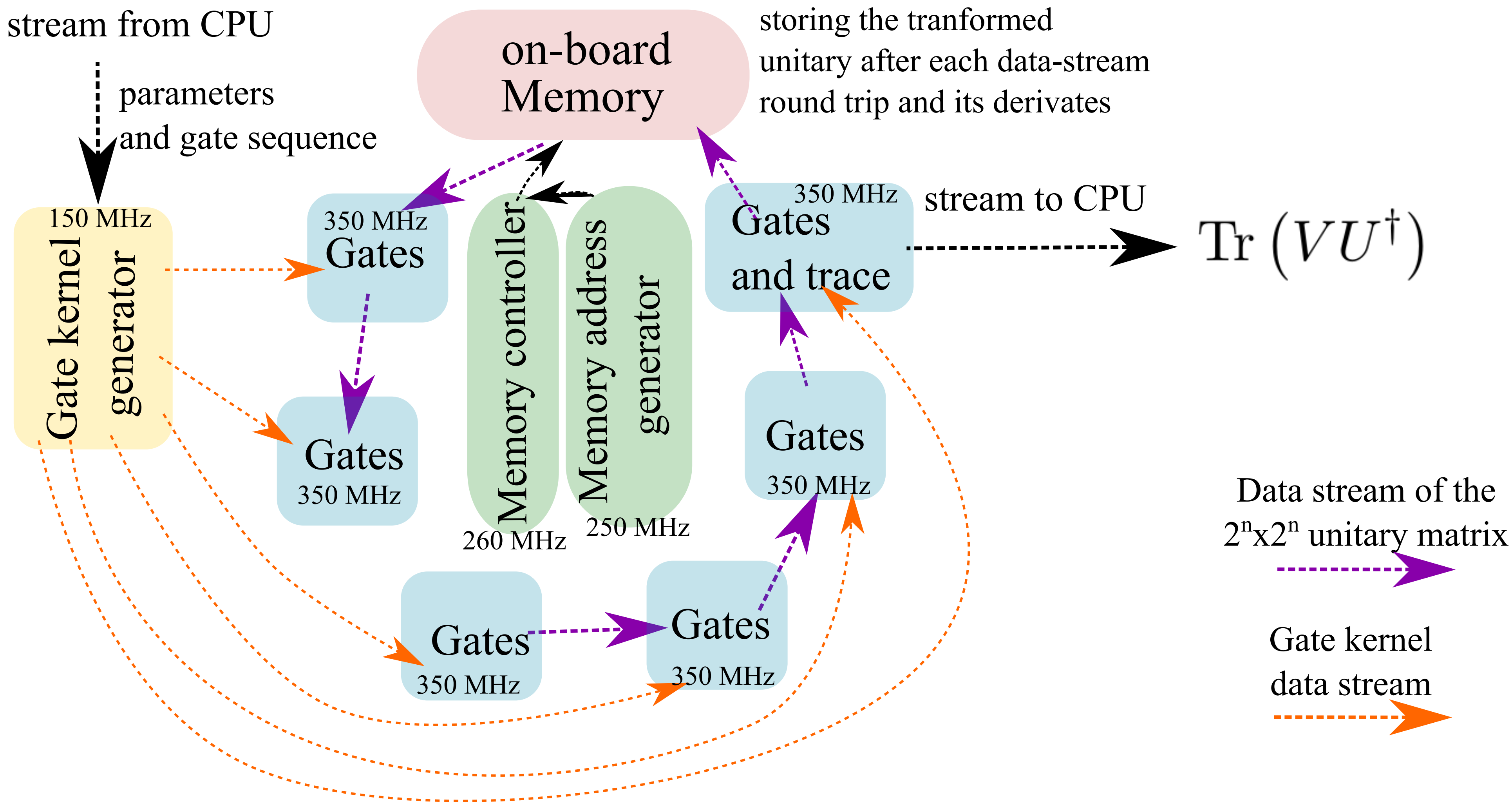
In total: 2 x 3 x 3 = 18 multiplications $\longrightarrow$ 18 DSP units are needed

+ look-up-tables (LUTs)

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

DFE quantum computer simulator

Quantum Information National Laboratory HUNGARY



- Chain up successive gate operations to increase computational concurrency
- Buffer the transformed unitary into the on-board memory (64 GB)

ELTE EÖTVÖS LORÁND UNIVERSITY Wigner groq MAXELER Technologies Maximum Performance Computing

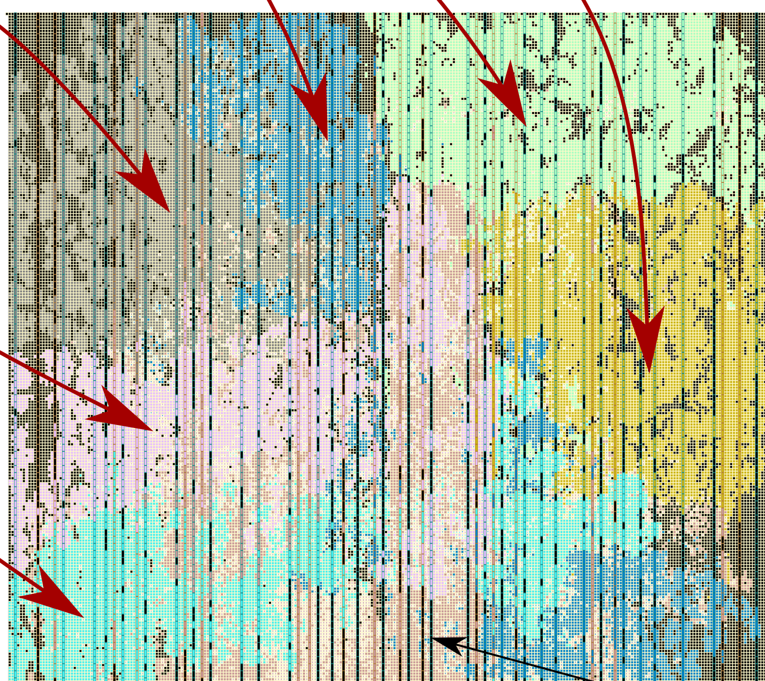
DFE quantum computer simulator

Quantum Information National Laboratory HUNGARY

6 x 18 = 108 gates on each Super Logic Region (SLR)

6 asynchronously operating gate blocks

Each block contains 18 synchronously operating gate operations



SLR

Memory controller

Xilinx Alveo U250 FPGA chips contain 4 SLRs

In total 432 parallel gate operations on a single FPGA chip

ELTE EÖTVÖS LORÁND UNIVERSITY wigner groq MAXELER Technologies Maximum Performance Computing

DFE QC simulator performance

Quantum Information National Laboratory HUNGARY

cost function

$$f = 2^n - \mathrm{Re}\left[\mathrm{Tr}(VU^\dagger)\right]$$

and

gradients

$$\frac{\partial f}{\partial x_i} = \mathrm{Re}\left[\mathrm{Tr}\left(\frac{\partial V}{\partial x_i}U^\dagger\right)\right]$$

can be evaluated with DFE

number of free parameters

gate count

initialization overhead

~ms

$$T_{\mathrm{DFE}} = 4^n \frac{N_p + 1}{4 \cdot f_{\mathrm{DFE}}} \cdot \mathrm{ceil}\left(\frac{N_G}{N_{G,chain}}\right) + t_0$$

frequency of gates (350MHz)

gates in the chain

arithmetic operations per second: $18 \cdot 4 \cdot N_{G,chain} \cdot f_{\mathrm{DFE}}$

(excluding all integer logic from the count)

equvalent to 2.72 TOPS

ELTE EÖTVÖS LORÁND UNIVERSITY

wigner

groq

MAXELER Technologies Maximum Performance Computing

DFE vs CPU performance

Quantum Information National Laboratory HUNGARY

DFE vs 32-Core AMD EPYC 7542 Processor (64 threads)

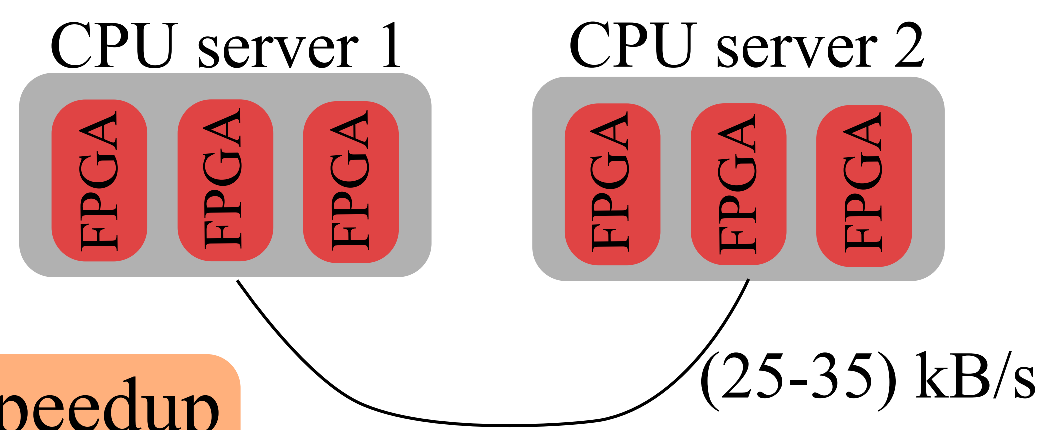
(6-13)x speedup with DFE

6 qubits

9 qubits

Scaling up calculations:

- split gradients over FPGAs
- 3 FPGAs per CPU server
- MPI between CPU servers



up to 78x speedup

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

Gate synthesis benchmark

Quantum Information
National Laboratory
HUNGARY

Circuit name	n	IBM QX (*39*)		QISKIT (*40*)		SQUANDER (*41*)			comp. rate
		CX	D	CX	D	CX	D	f	
4gt12-v0_87	6	112	131	625	1146	47	73	0.0028	93.6%
4gt12-v0_88	6	86	108	853	1647	44	71	0.0072	95.7%
4mod5-bdd_287	7	31	41	1037	1825	26	41	0.012	97.8%
alu-bdd_288	7	38	48	224	408	30	35	0.0038	91.4%
C17_204	7	205	253	2992	5915	104	133	0.0042	97.8%
ex2_227	7	275	355	2852	5554	133	161	0.0128	97.1%
majority_239	7	267	344	4024	7950	143	175	0.0127	97.8%
rd53_131	7	200	261	6538	12320	93	119	0.0129	99.0%
rd53_135	8	134	159	26126	50436	120	147	0.0195	99.7%
rd53_138	8	60	56	18567	35172	87	117	0.061	99.7%
cm82a_208	8	283	337	11246	22284	86	67	0.0129	99.7%
con1_216	9	415	508	55822	109798	205	229	0.118	99.8%

$$\overline{F}_F = 1 - \varepsilon \qquad \varepsilon \approx 10^{-4}$$

ELTE EÖTVÖS LORÁND UNIVERSITY

Wigner

groq

MAXELER Technologies Maximum Performance Computing

Further improvements to the compiler

Quantum Information National Laboratory HUNGARY

1

Construct the initial gate structure

VS

interations to compress the initial gate structure

for 8-9 qubits

- takes a 2-7 days

- takes a 1-30 minutes (each iteration)

- In both cases we need to solve quite similar optimization problem.
- **Why is the big difference in the execution time?**

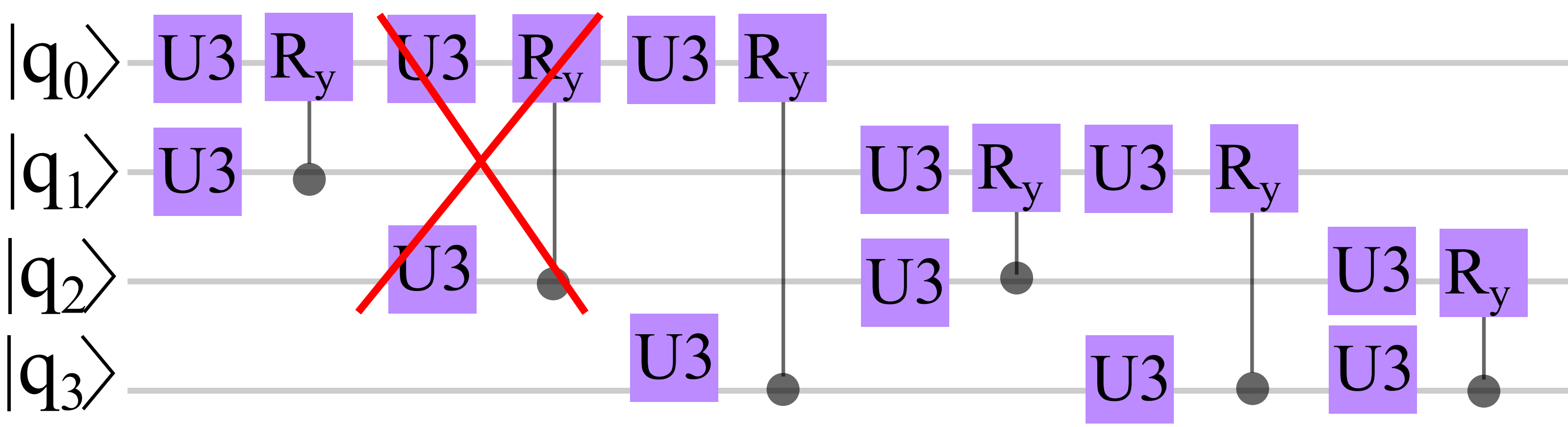
2 Scale up the compiler to circuits with many qubits

- Reorganize circuit to exhibit large 6-10 qubit subblocks that can be optimized

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

Iterative Gate Compression

- Try to remove a random 2-qubit block



- Try to solve the optimization problem for the reduced circuit
- The reduced initial parameter set is derived from the previous solution

$$p_1,p_2,p_3,\cancel{p_4,p_5,p_6,p_7},p_8,p_9,p_{10},p_{11},p_{12},p_{13},p_{14},p_{15},p_{16},p_{17},p_{18},p_{19},p_{20},\ldots$$

- Solve the optimization problem in 1-30 minutes
- A well chosen initial parameter set (i.e. correlations between the qubits) is crucial to speed up the quantum compilation

Quantum Information National Laboratory HUNGARY

ELTE EÖTVÖS LORÁND UNIVERSITY

Wigner

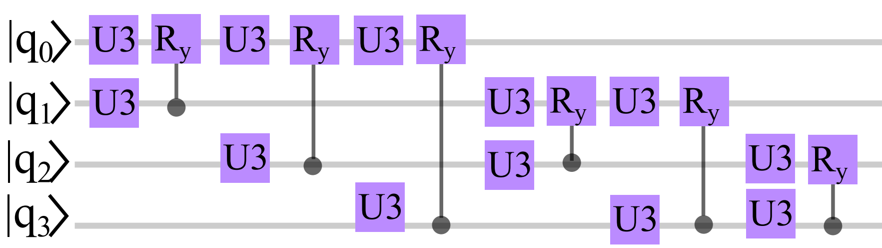
groq

MAXELER Technologies Maximum Performance Computing

How to chose the initial parameter set?

Quantum Information National Laboratory HUNGARY

- Use **machine learning** to predict the initial correlation between the qubits



- Supervised machine learning with labeled data:

Random initial parameters ⟷ Unitary

- Predict initial parameter set from the trained model

Unitary ⟶ Initial parameters

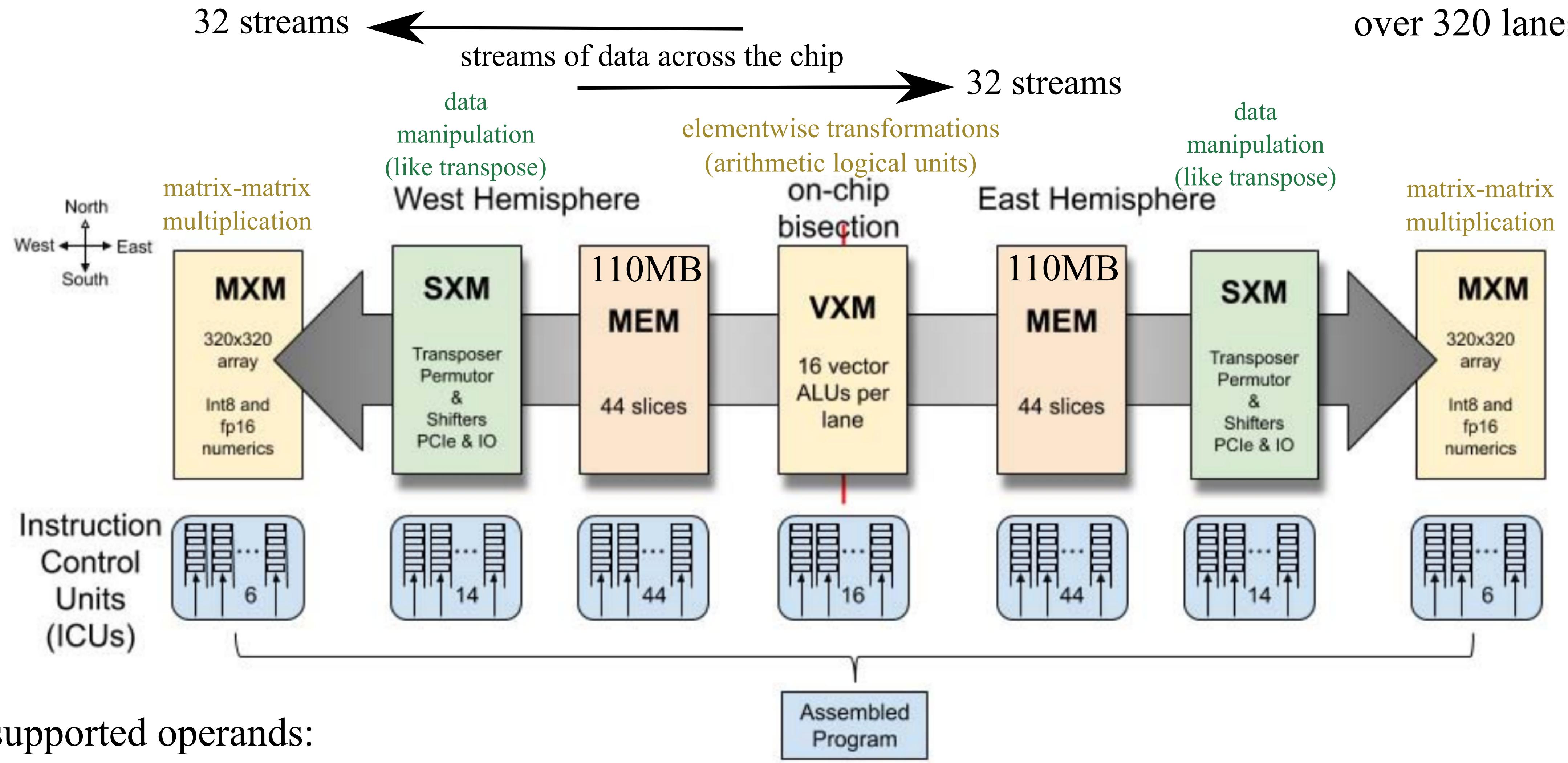
ELTE EÖTVÖS LORÁND UNIVERSITY · Wigner · groq · MAXELER Technologies Maximum Performance Computing

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



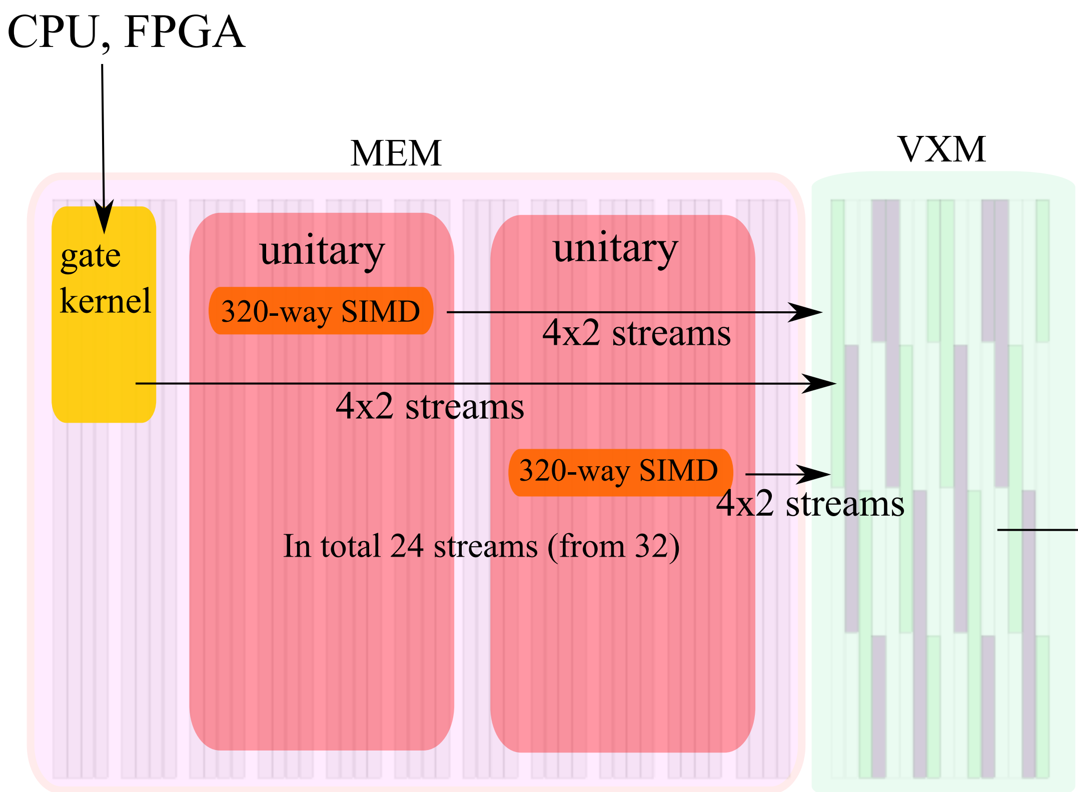
supported operands:

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

Concept of Groq QC simulator



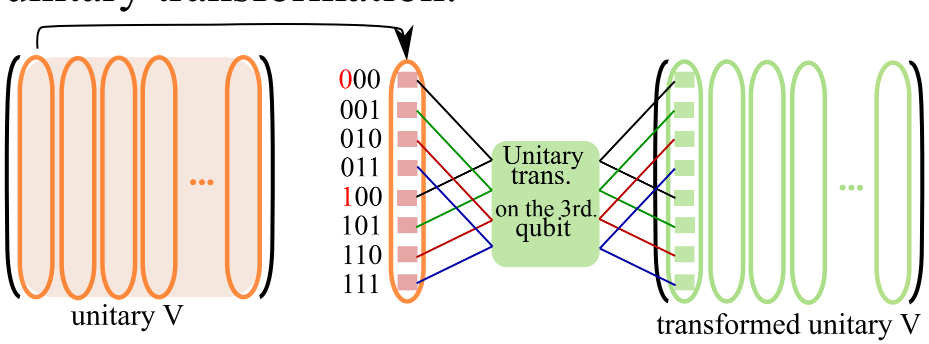
4x2 streams

Save the transformed unitary on the other hemisphere

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

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

unitary transformation:



- In principle 50% more perfromance by Groq compared to DFE
- The preformance can be keep up to 11 qubits (on DFE the limit is 9 qubits)

Quantum Information National Laboratory HUNGARY

ELTE EÖTVÖS LORÁND UNIVERSITY

wigner

groq

MAXELER Technologies Maximum Performance Computing

Conclusions and outlook

Quantum Information
National Laboratory
HUNGARY

We have designed a DFE based QC simulator to speed up the gate synthesis process up to 9 qubit circuits.

Aiming to reduce the execution time by:

- Predict intial parameter set with machine learning
  (achieve competitive execution time with deterministic tools)

- Scale up the decomposition for circuits with more qubits
  (transform the circuit with gate identities, optimize 6-9 qubit blocks)

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