2023 VLSI SYMPOSIUM

Simulated bifurcation machines:
combinatorial optimization accelerators based on a quantum-inspired parallelizable algorithm

Kosuke Tatsumura

Toshiba Corporation

Outline

- Introduction
- Simulated bifurcation (SB)
- Implementation & Performance
- Application
- Conclusion

Combinatorial optimization

Economically important but computationally hard

Practical problems

decision-making, planning, etc

Finance

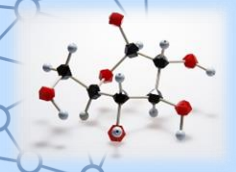
Manufacture

Medicine

Logistics

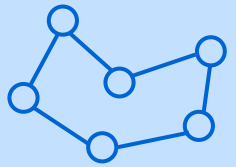
Management

Material

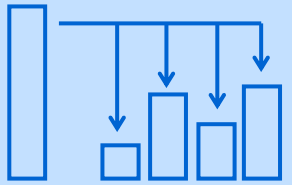


Combinatorial optimization

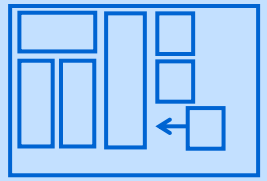
Routing



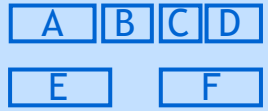
Allocation

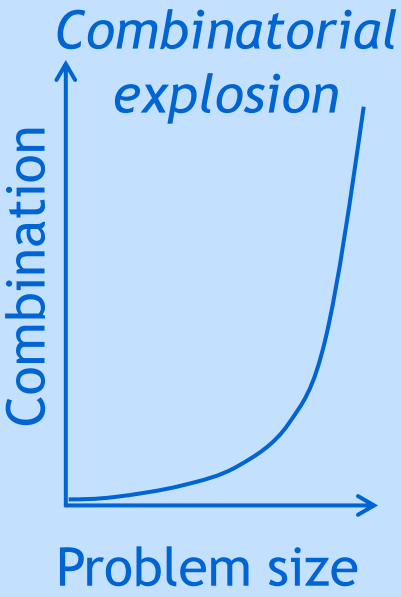


Placement



Scheduling





Nondeterministic polynomial time (NP)-hard

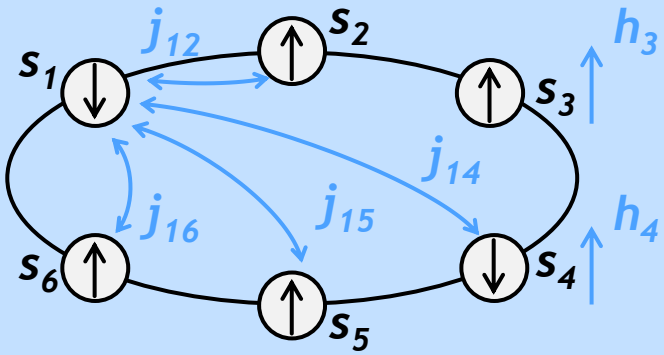
Ising machine

Special-purpose computer for combinatorial optimization

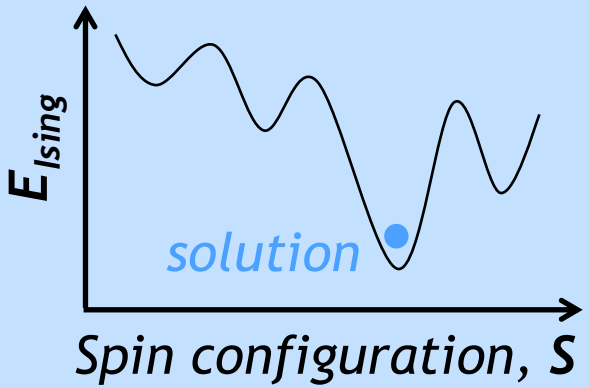
Ising problem

search for the lowest-*E* state of Ising models

$$E = -\sum j_{ij} s_i s_j + \sum h_i s_i$$



spin: binary variable



Combinatorial optimization

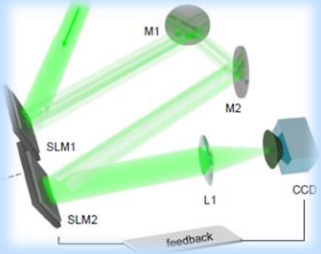
Ising machines

Quantum annealer*1

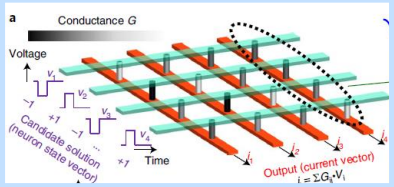
CMOS annealer*2

Digital annealer*3

Optical Ising machines*4,5



Memristor HNN*6



and more …

+ Simulated bifurcation machine (2019)

*1 https://www.dwavesys.com/d-wave-two-system
*2 https://www.hitachi.co.jp/New/cnews/month/2019/02/0219.html
*3 https://www.fujitsu.com/global/about/resources/news/press-releases/2018/0515-01.html
*4 https://www.ntt.co.jp/news2017/1711e/171120a.html
*5 D. Pierangeli, et al., Phys. Rev. Lett. **122**, 213902 (2019).
*6 F. Cai, et al., Nature Electronics **3**, 409 (2020).

Simulated bifurcation machine (SBM)

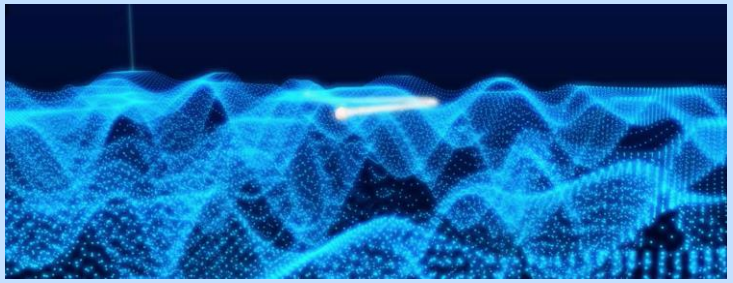
Algorithm

Quantum-inspired

Quantum bifurcation machine
in a quantum principle

Discovery

Simulated bifurcation algorithm
in a new classical principle

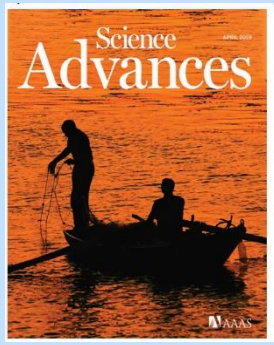


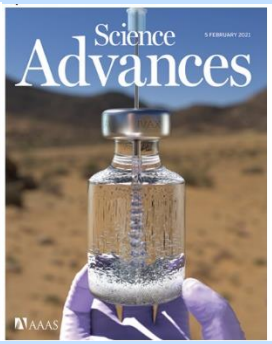
Highly parallelizable

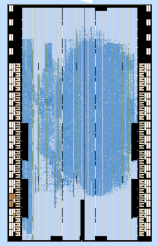
Implementation

High performance

single-chip







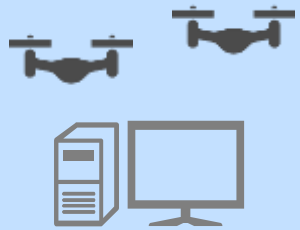
Scalable

multi-chip

Application

Very practical

edge/embedded



cloud

Innovative

ex.real-time systems

Outline

- Introduction
- Simulated bifurcation (SB)
- Implementation & Performance
- Application
- Conclusion

Quantum-inspired algorithm

Quantum Bifurcation (QB) machine

[H. Goto *et al.*, *Sci. Rep.*, (2016)]

Hamiltonian describing adiabatic bifurcation process in a nonlinear oscillator network

$$H_q(t) = \hbar \sum_{i=1}^{N} \left[\frac{K}{2} a_i^{\dagger 2} a_i^2 - \frac{p(t)}{2} \left(a_i^{\dagger 2} + a_i^2 \right) + \Delta_i a_i^{\dagger} a_i \right] - \hbar \xi_0 \sum_{i=1}^{N} \sum_{j=1}^{N} J_{i,j} a_i^{\dagger} a_j$$

Classical Bifurcation (CB) machine

classicization of state variables

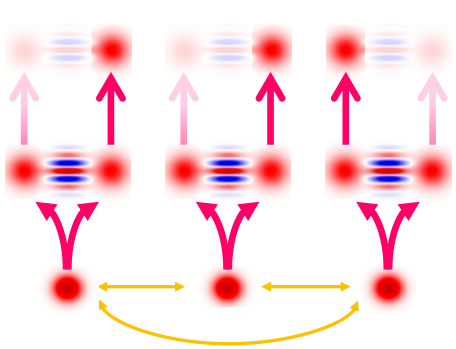
$$H_c(\mathbf{x}, \mathbf{y}, t) = \sum_{i=1}^{N} \left[\frac{K}{4} \left(x_i^2 + y_i^2 \right)^2 - \frac{p(t)}{2} \left(x_i^2 - y_i^2 \right) + \frac{\Delta_i}{2} \left(x_i^2 + y_i^2 \right) \right] - \frac{\xi_0}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} J_{i,j} \left(x_i x_j + y_i y_j \right)$$

algorithmic twist for speed-up

Simulated Bifurcation (SB) algorithm (2019)

[H. Goto *et al.*, *Sci. Adv.*, (2019)]

Combinatorial optimization based on quantum adiabatic theorem



Classicizing QB that works in a quantum principle...?
Why SB works? What principle? There was a discovery

Simulated bifurcation: Why it works

New classical principle: adiabatic and ergodic search

Dynamical change of energy landscape

a single local minimum

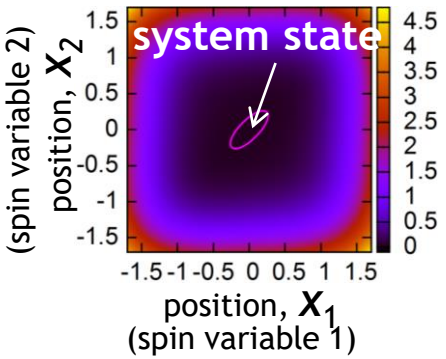
↓

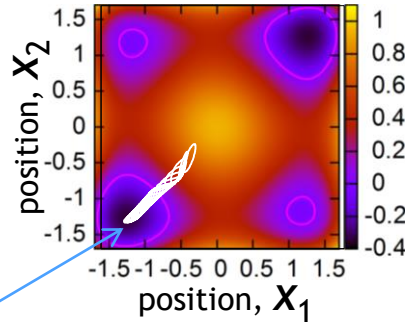
bifurcation
(adiabatic process)

↓

multiple local minima
(target cost function)

Energy landscape (N_{spin}=2)





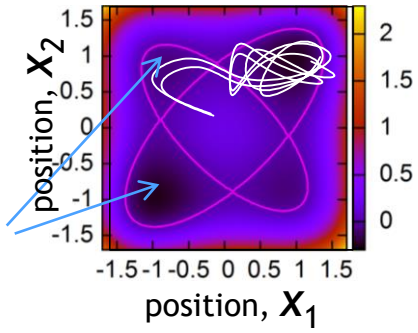
best solution
(-1,-1)

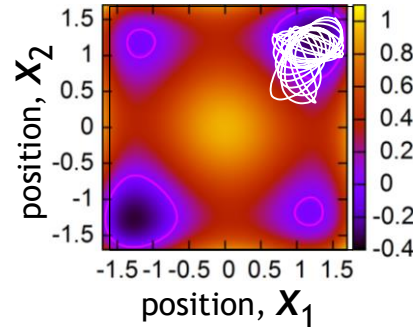
x_1

adiabatic search

chase one of the minima

Multiple minima in the energetically allowable region



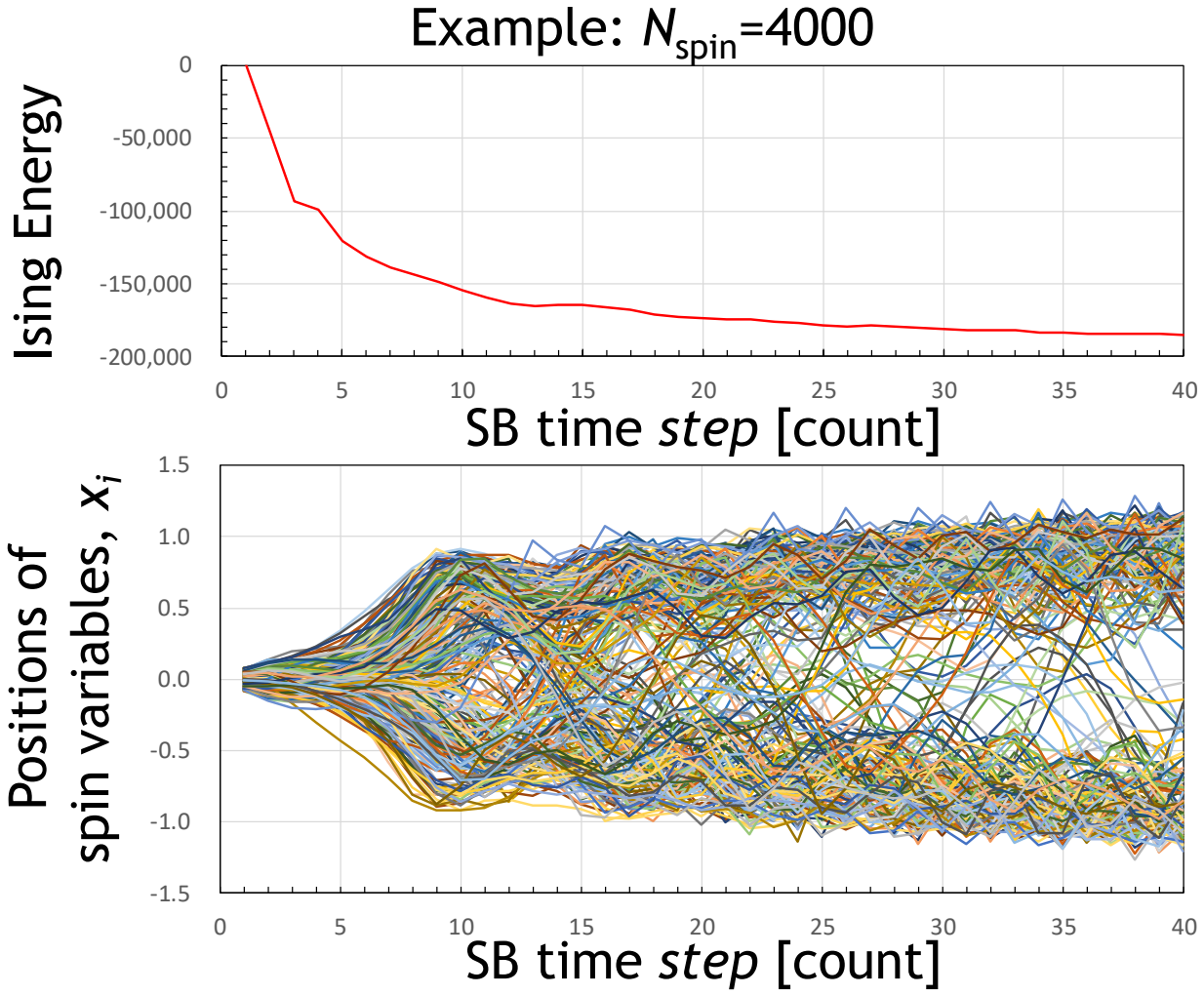


ergodic search

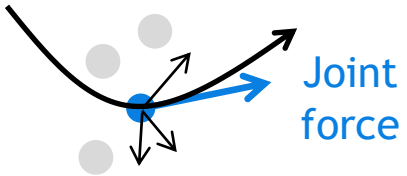
find a better one with a higher probability

Simulated bifurcation: How it works

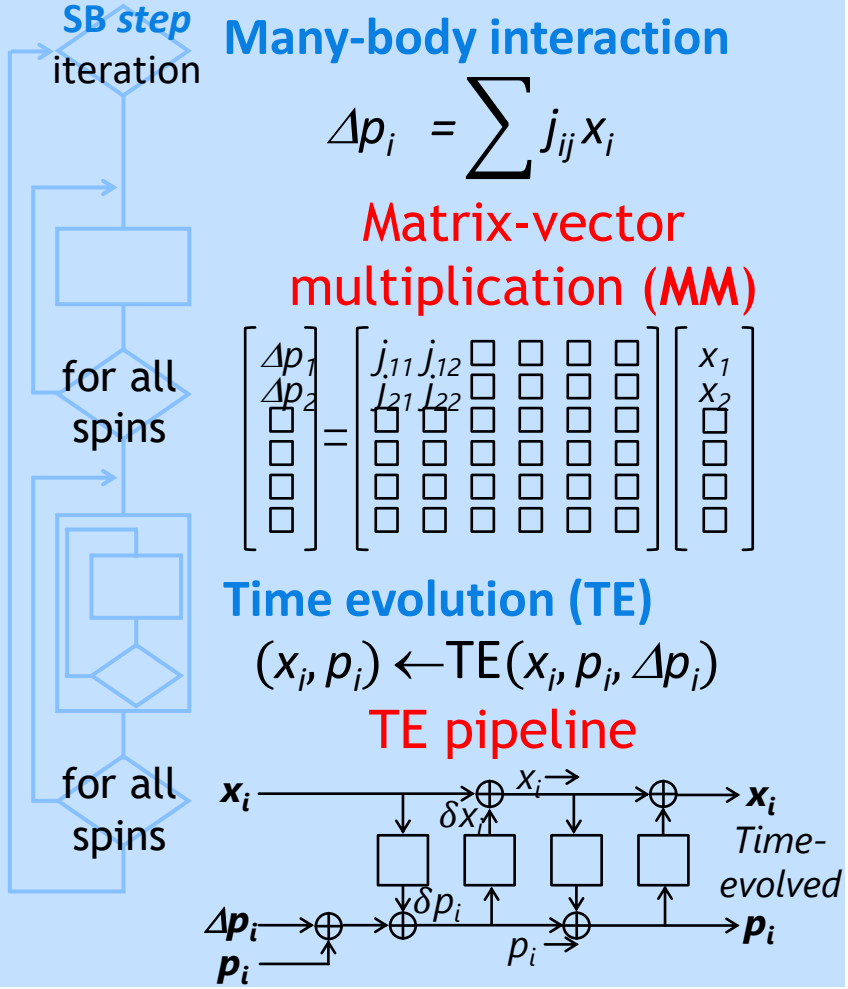
"*N*-body"-type algorithmic structure → highly parallelizable



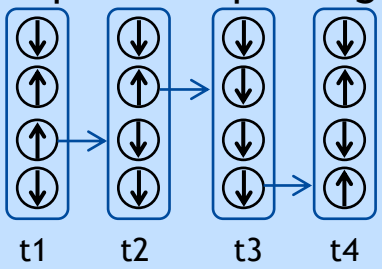
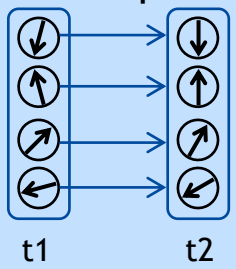
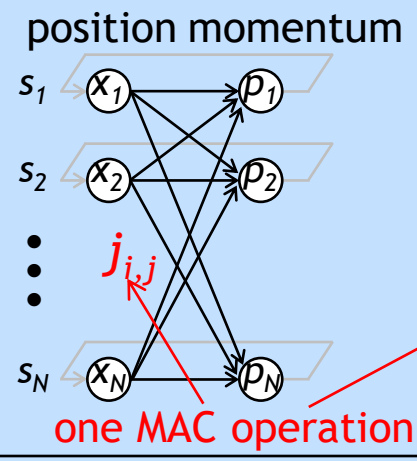
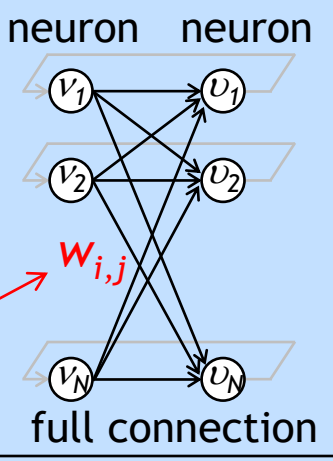
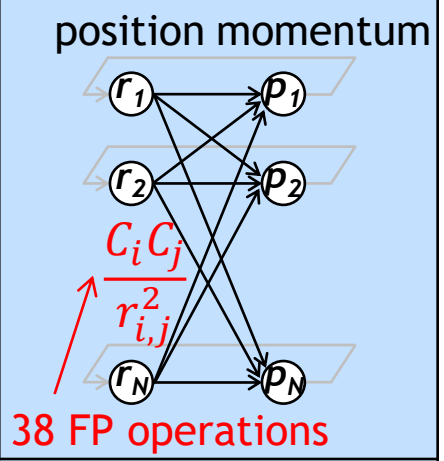
better solution



Many-body interaction depending on all the other spins



Characteristics

	SA simulated annealing	SB simulated bifurcation	R-NN recurrent neural network	*N*-body gravitational (/Coulomb)-force
Structure	Sequential updating  	Parallel updating  	 	 	
Parallelism	O(*N*)	O(N^2)		

More parallelizable

Intensive memory access
J/W matrix (NxN matrix)

Very similar

More PEs per chip
PE: pairwise interaction

SB can be accelerated by FPGAs/GPUs (not limited to special ASICs)
Many AI chips (AI ASSPs) are beneficial also to SB

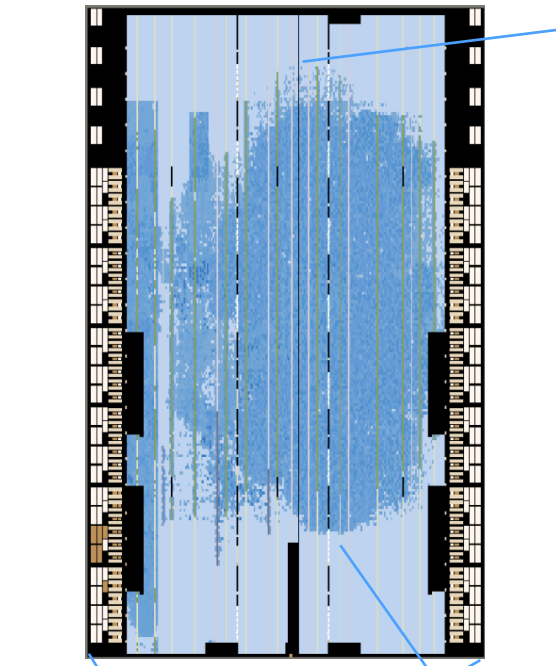
Outline

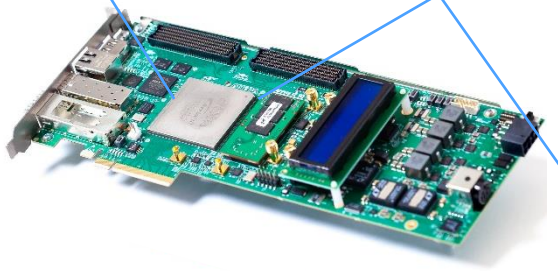
- Introduction
- Simulated bifurcation (SB)
- Implementation & Performance
- Application
- Conclusion

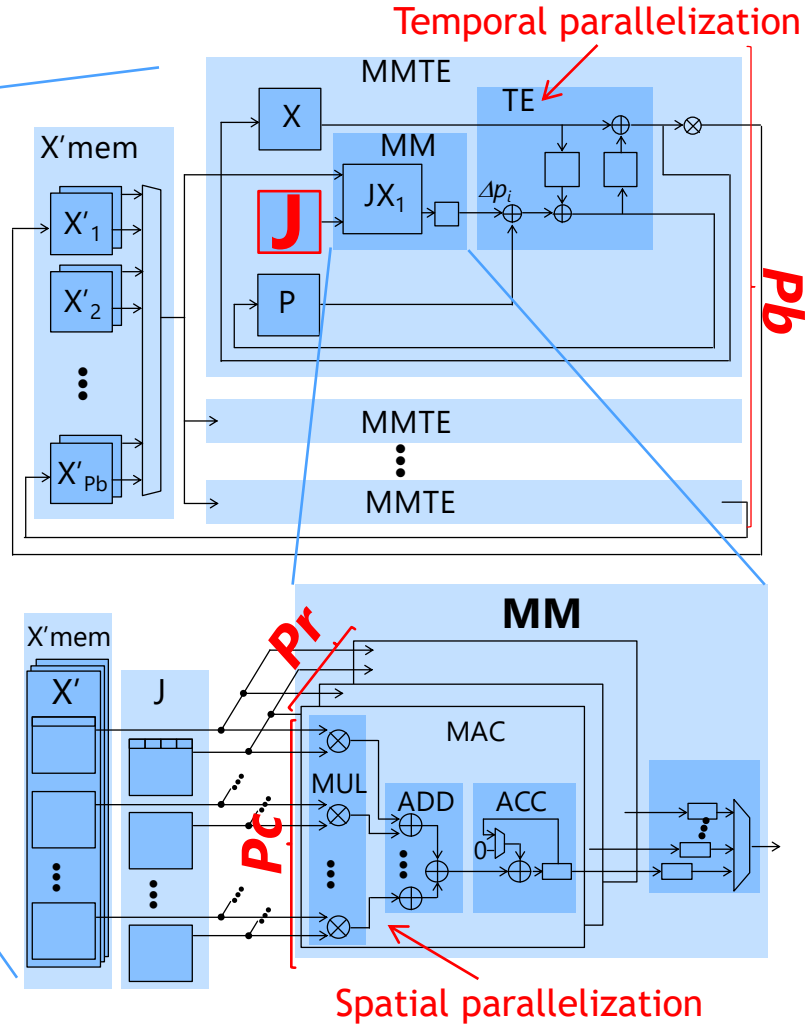
FPGA-based accelerator for simulated bifurcation

Large-scale, massively parallel, and high utilization

Arria10 GX1150 FPGA







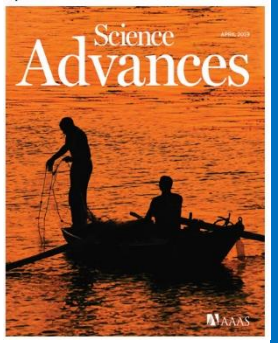
[K. Tatsumura *et al.*, *IEEE FPL*, (2019)]

Problem	complete-graph MAX-CUT
Machine size	4,096 spins (on Arria10 FPGA)
Architecture	
Pr/Pc/Pb	32/32/8
# of MAC PEs	**8,192**
Effective activity	**92%**
Resource	
ALM	40%
BRAM	56%
DSP	7%
System Clock	[MHz]
Fsys	**269**

#PEs > *N* (not achievable for SA)

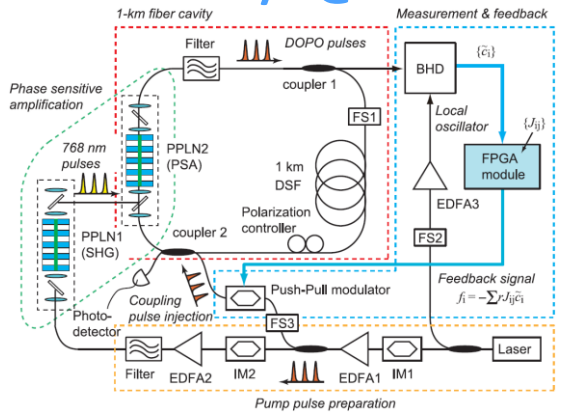
[H. Goto *et al.*, *Sci. Adv.*, (2019)]

Performance (2019)



Coherent Ising Machine

800 GMAC/s @ 1000 W

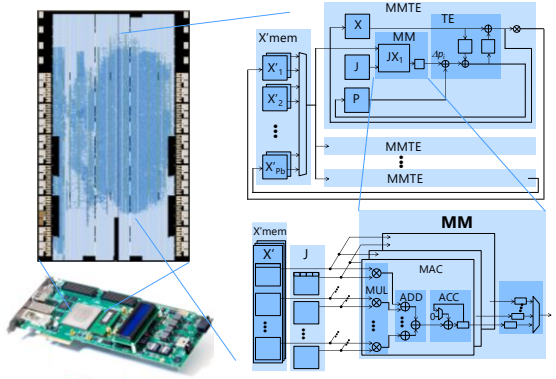


[T. Inagaki, Science 354, 603, '16]

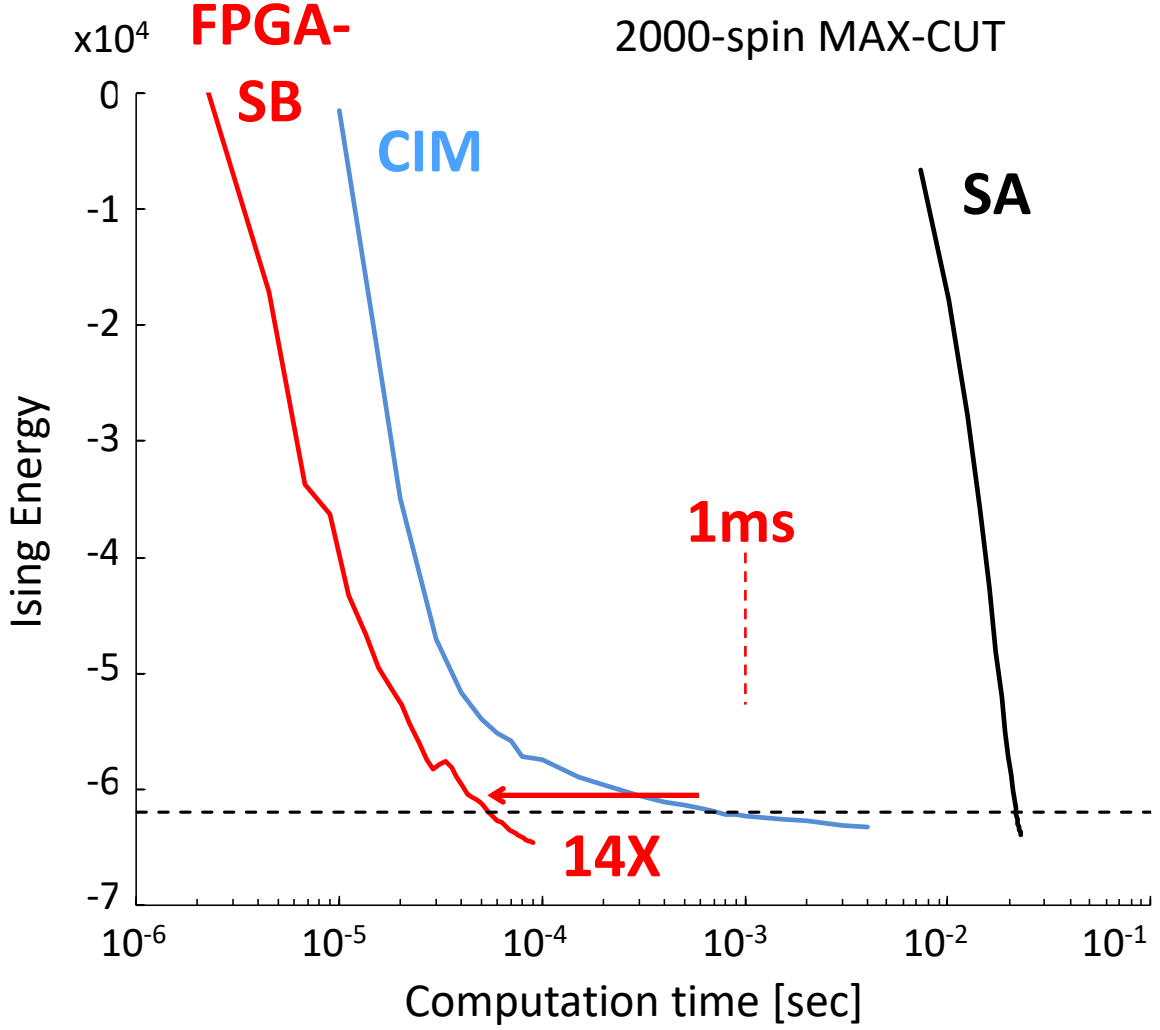
FPGA-SB

1,873 GMAC/s @ 49 W

(288X more energy efficient)



@all-to-all-connected
2000-spin MAX-CUT

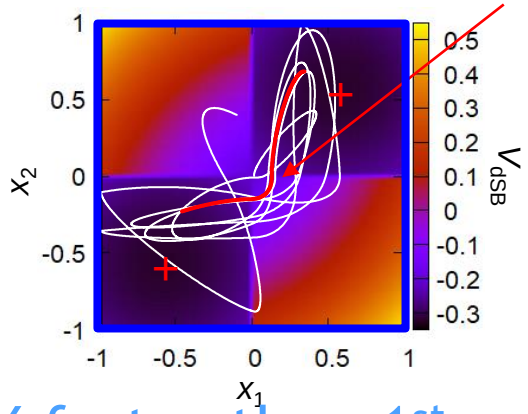


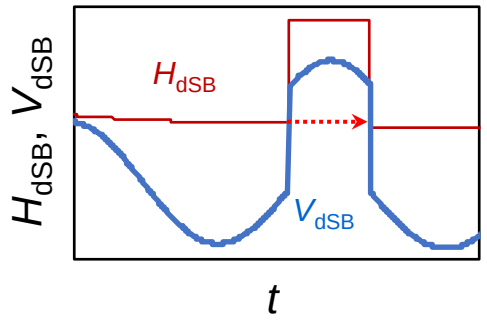
[H. Goto *et al.*, *Sci. Adv.*, (2021)]

Performance (2021)

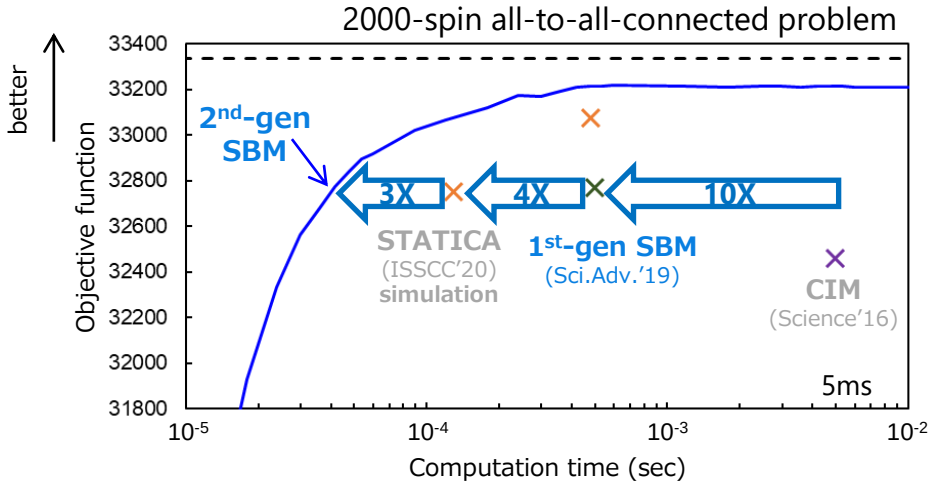
SB is very competitive with state-of-the-art Ising machines

2nd-gen algorithm Quasi-quantum tunneling



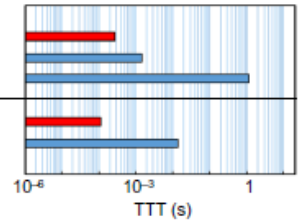


10X faster than 1st-gen



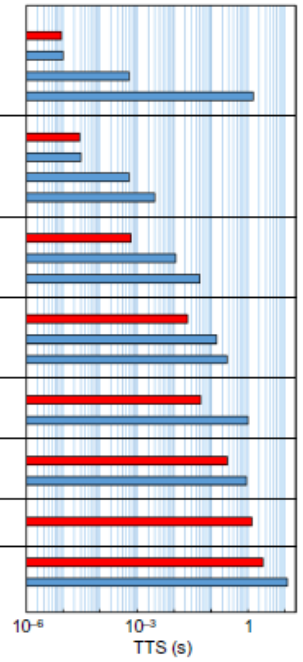
B

N	Connectivity	J_{ij}	Machine	TTT
2000 (K_{2000})	All-to-all	{±1}	**bSBM**	**0.26 ms**
			STATICA	1.50 ms
			CIM	1.1 s
2000 (G22)	Sparse (1%)	{0, −1}	**bSBM**	**0.11 ms**
			CIM	14 ms



C

N	Connectivity	J_{ij}	Machine	TTS
60	All-to-all	{±1}	**dSBM**	**9.2 µs**
			RBM	10 µs
			CIM	0.6 ms
			QA	1.4 s
100	All-to-all	{±1}	**dSBM**	**29 µs**
			RBM	30 µs
			SimCIM	0.6 ms
			CIM	3.0 ms
200	Sparse (Degree 3)	{0, −1}	**dSBM**	**0.70 ms**
			QA	11 ms
			CIM	51 ms
700	All-to-all	{±1}	**dSBM**	**25 ms**
			SimCIM	0.14 s
			DA	0.27 s
1024	All-to-all	{±1}	**dSBM**	**55 ms**
			DA	1 s
1024	All-to-all	16 bits $\{-2^{15}+1, \ldots, 2^{15}-1\}$	**dSBM**	**0.29 s**
			DA	0.9 s
2000 (K_{2000})	All-to-all	{±1}	**dSBM**	**1.3 s**
2000 (G22)	Sparse (1%)	{0, −1}	**dSBM**	**2.7 s**
			SimCIM	12 s



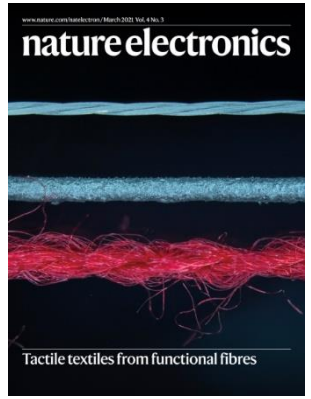
TTS (Time-To-Solution)

Competitors

SB: Simulated bifurcation
QA: Quantum annealer
CIM: Coherent Ising machine
DA: Digital annealer
SimCIM: Simulated CIM
RBM: Restricted Boltzmann machine
MA: Momentum annealing

[K. Tatsumura *et al.*, *Nat. Ele.*, (2021)]

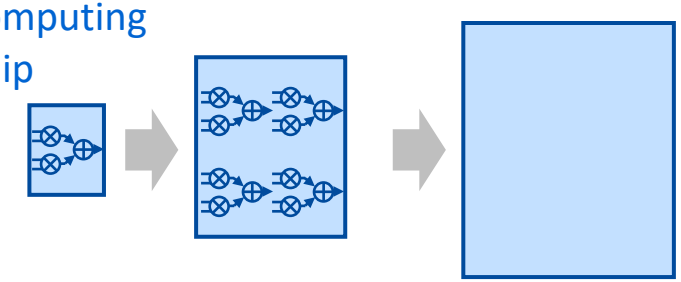
Scalability (2021)



Scaling out Ising machines with full spin-to-spin connectivity

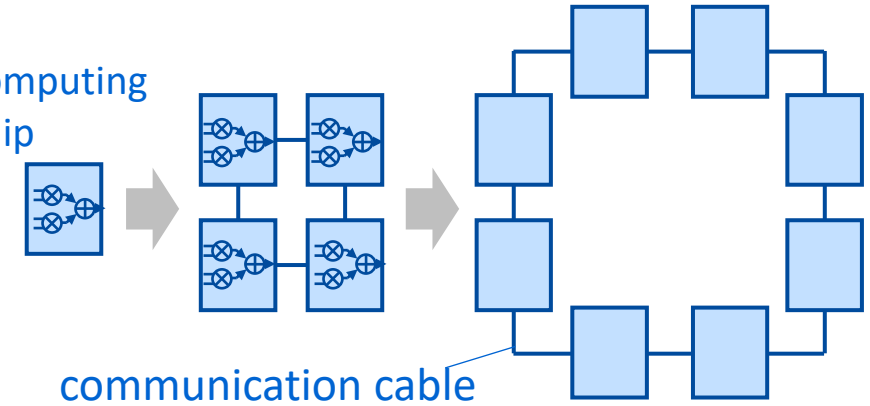
Scale-up

computing chip

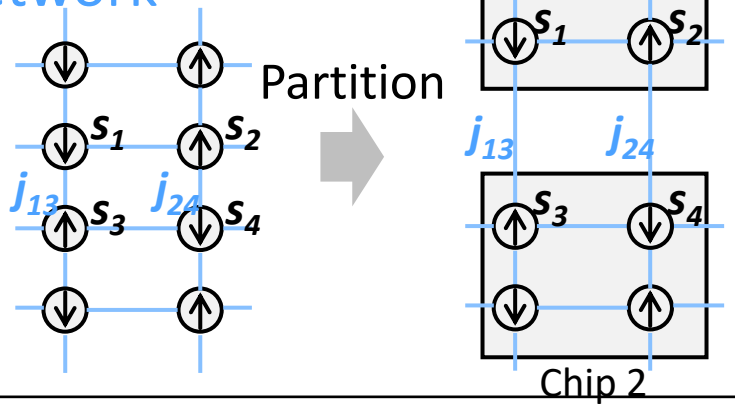


Our aim Scale-out

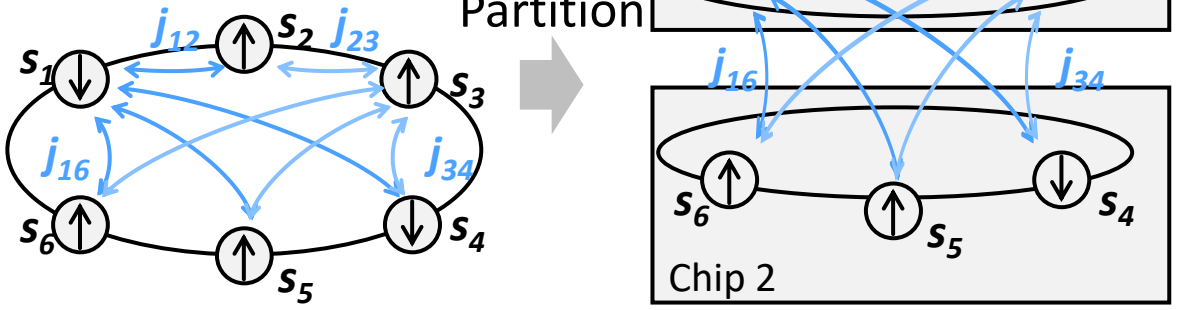
computing chip



Locally-connected spin network



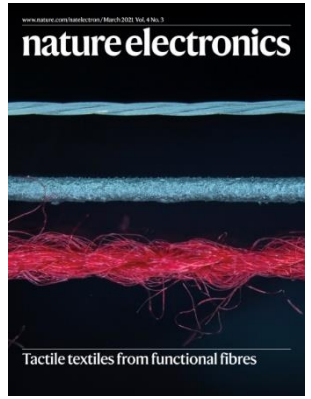
Our aim Fully-connected spin network



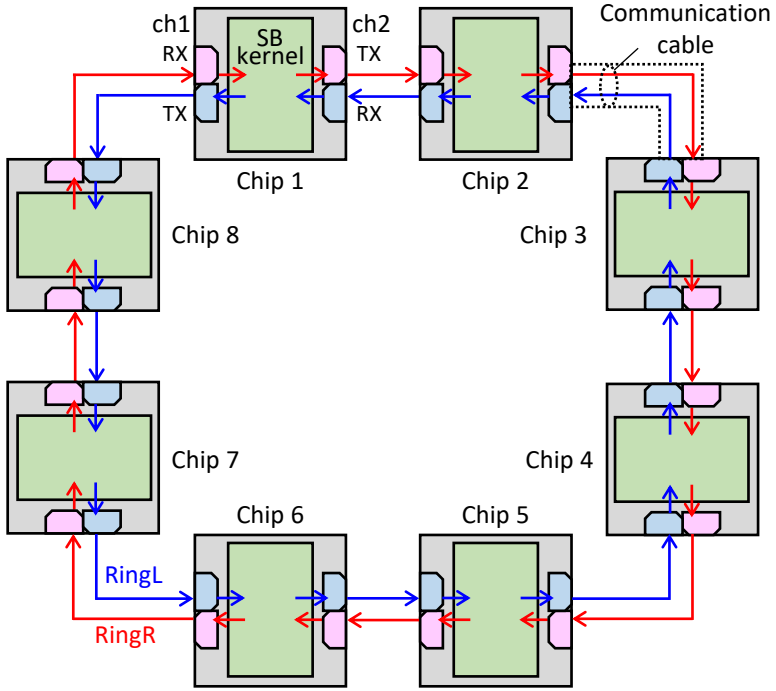
[K. Tatsumura *et al.*, *Nat. Ele.*, (2021)]

Scalability (2021)

Multi-chip architecture based on partitioned SB

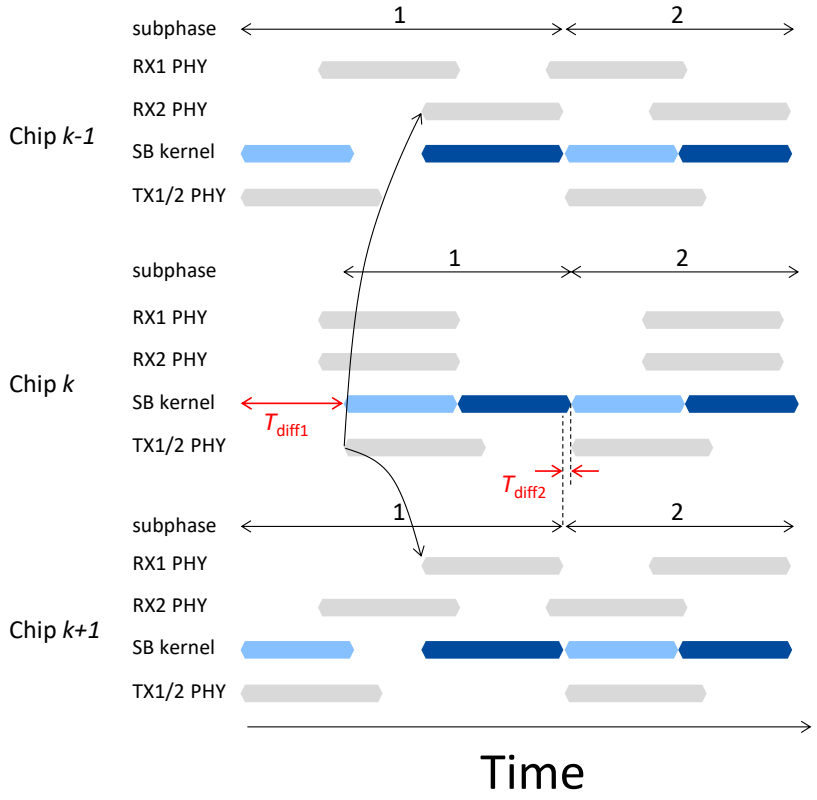


Bidirectional ring-network cluster
without any centralized features

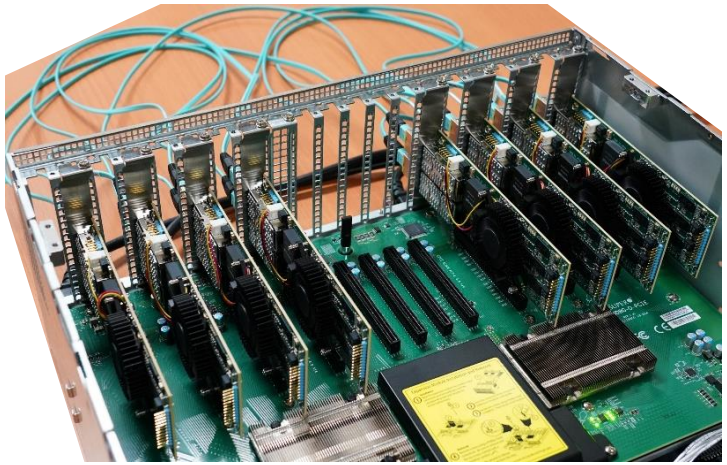


All chips are
autonomous, homogeneous and symmetric

Autonomous synchronization mechanism
(No clock-sharing, No central-HUB)



P_{chip}=8



[K. Tatsumura *et al.*, *Nat. Ele.*, (2021)]

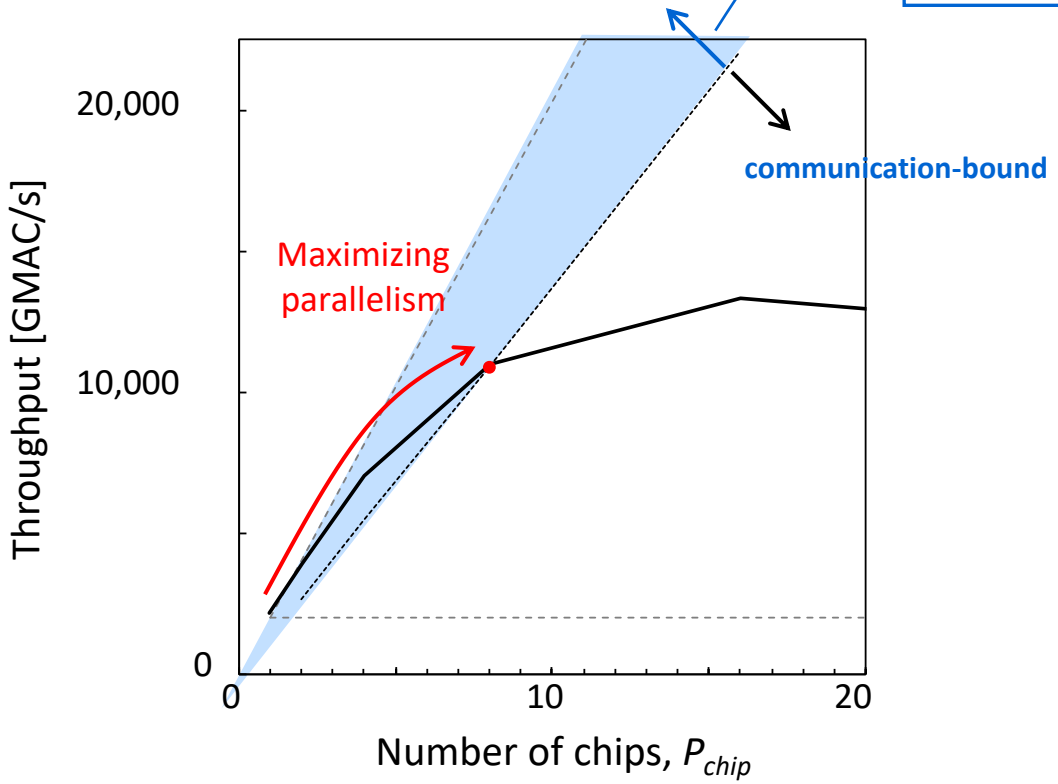
Scalability (2021)

Strong scaling

Increase P_{chip} at a fixed problem size (N)

Computation-bound

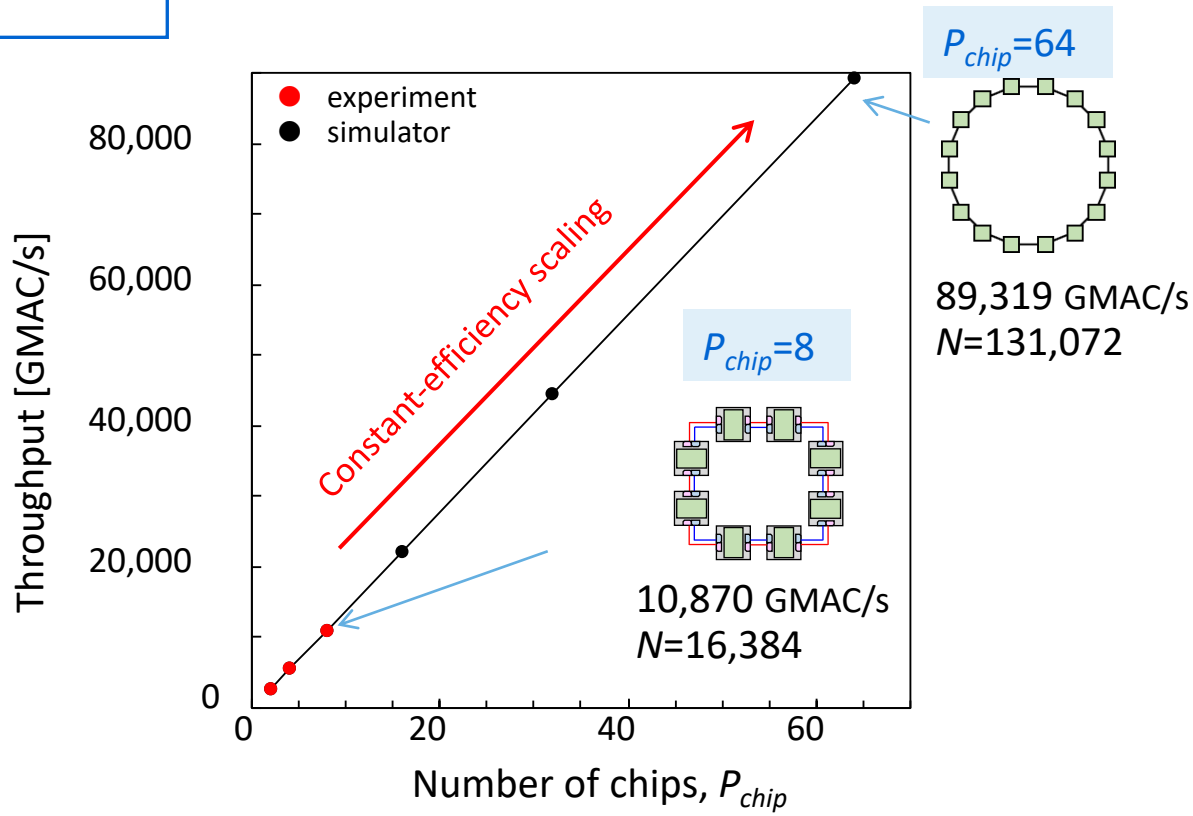
$$\frac{T_{computation}}{T_{communication}} > 1$$



Throughput enhancement to the vicinity of an ideal upper limit determined by the communication tech.

Weak scaling

Increase P_{chip} and N in the same proportion



Constant-efficiency scaling at the maximized computation parallelism (at the strong scaling limit)

Outline

- Introduction
- Simulated bifurcation (SB)
- Implementation & Performance
- Application
- Conclusion

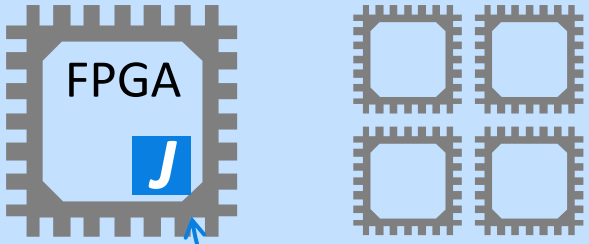
Application of SBMs

Toshiba's website "SQBM+™"
https://www.global.toshiba/ww/products-solutions/ai-iot/sbm.html

edge/embedded

Lower latency

FPGA cluster



on-chip memory

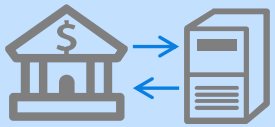
J matrix

$$\begin{bmatrix} \Delta p_1 \\ \Delta p_2 \\ \square \\ \square \\ \square \end{bmatrix} = \begin{bmatrix} j_{11}\, j_{12} & \square & \square & \square & \square \\ j_{21}\, j_{22} & \square & \square & \square & \square \\ \square & \square & \square & \square & \square \\ \square & \square & \square & \square & \square \\ \square & \square & \square & \square & \square \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \square \\ \square \\ \square \end{bmatrix}$$

Single-shot processing

High-speed real-time systems

Financial trading

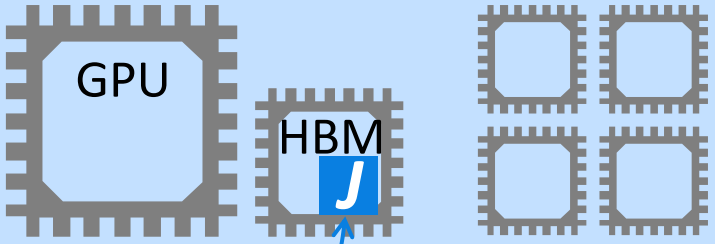


Autonomous control

cloud

Larger problem

GPU cluster



off-chip memory

J matrix

Independent trials

$$\begin{bmatrix} \Delta p_1 & \Delta p_1 \\ \Delta p_2 & \Delta p_2 \\ \square & \square \\ \square & \square \\ \square & \square \end{bmatrix}\!\!\bullet\bullet = \begin{bmatrix} j_{11}\, j_{12} & \square & \square & \square & \square \\ j_{21}\, j_{22} & \square & \square & \square & \square \\ \square & \square & \square & \square & \square \\ \square & \square & \square & \square & \square \\ \square & \square & \square & \square & \square \end{bmatrix} \begin{bmatrix} x_1 & x_1 \\ x_2 & x_2 \\ \square & \square \\ \square & \square \\ \square & \square \end{bmatrix}\!\!\bullet\bullet$$

Batch processing

Addressing complex/global issues

Drug design

Planning

Enabling NP-hard optimization in real-time systems

Must respond within critically defined time constraints

→Enabling *rational* judgment based on combinatorial optimization

High-speed real-time systems

Financial trading system

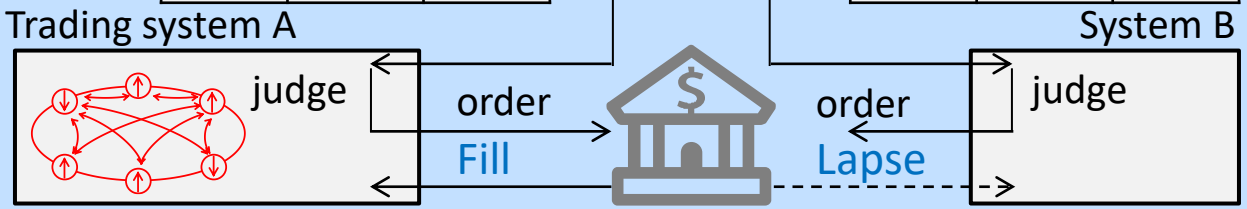
Stock A

ASK	Price	BID
1,000	4,250	
	4,249	800
	4,248	1,200

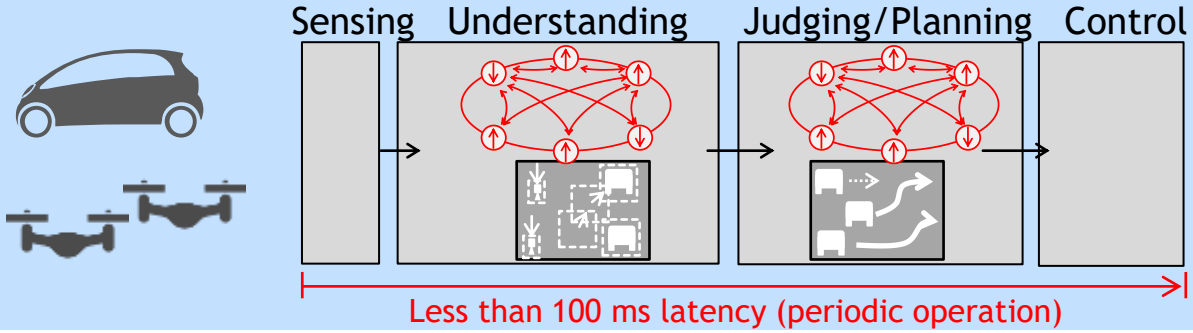
Stock B

ASK	Price	BID
300	6,381	
	6,380	
	6,379	500

BBO changed

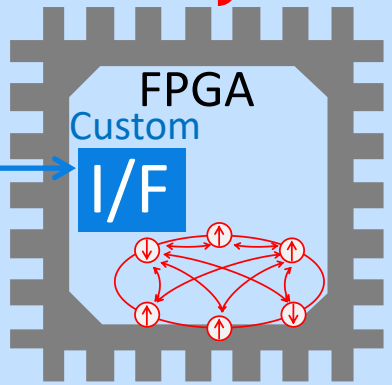


Autonomous control

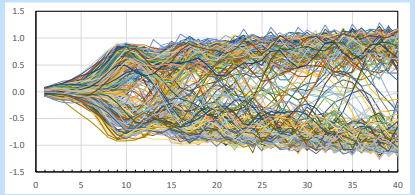


FPGA-based SBMs

Ultralow latency (sub-msec)
Deterministic latency



1. Embeddable
2. Custom I/F
3. Custom circuit
   (No software interrupt)



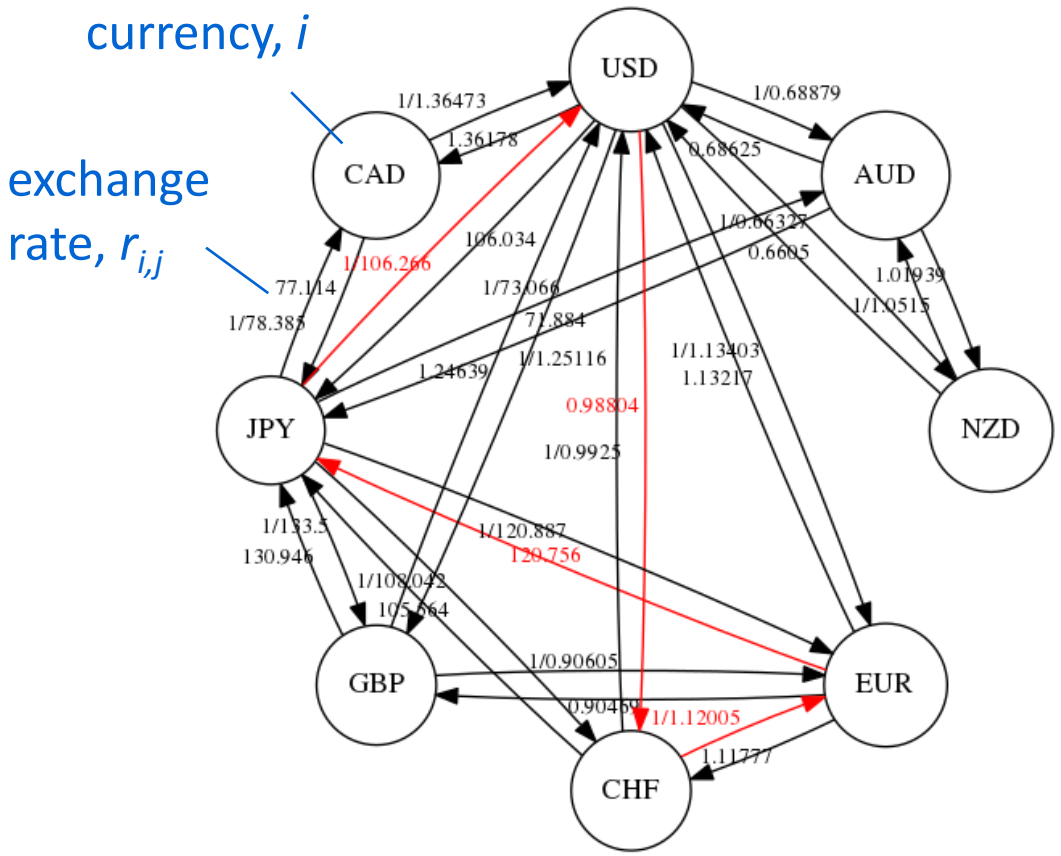
Predetermined #step

[K. Tatsumura *et al.*, *IEEE ISCAS.*, (2020)]

Trading system for cross-currency arbitrage

Optimal path search in a directed graph (a typical combinatorial problem)

Market Graph



Arbitrage Problem

find a closed path
that maximizes the profit

Cost function

$$Profit = \prod_{i,j \in path} r_{i,j}$$

Constraint

Must be
a closed path

Ising (QUBO) formulation

$$C_{tot} = m_c C + m_p P$$

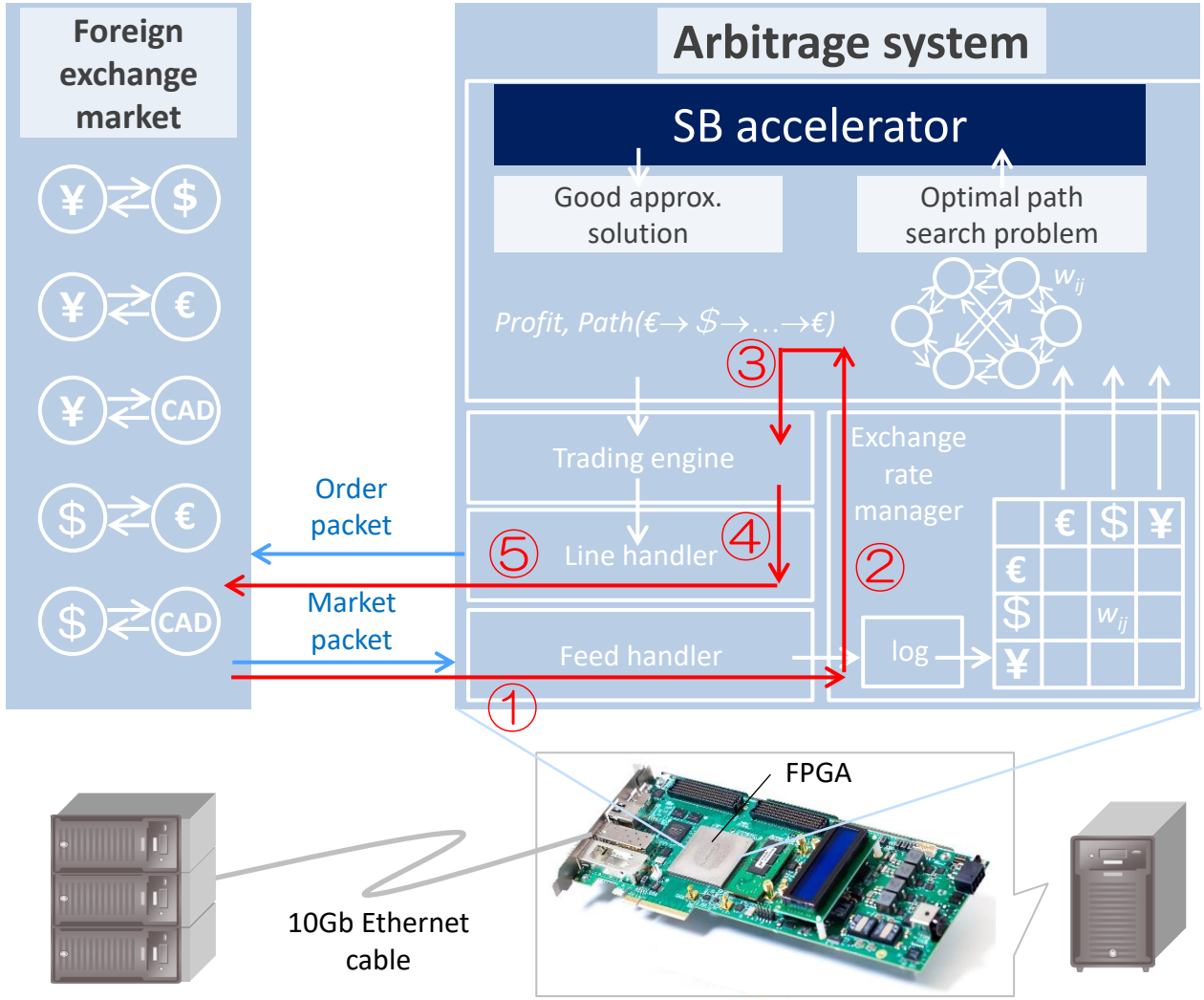
$$C' = \prod r_{i,j}{}^{b_{i,j}} \quad \xrightarrow{w_{i,j} = -\log r_{i,j}} \quad C = \sum w_{i,j}\, b_{i,j}$$

$$P = \sum_i \sum_{j \neq j'} b_{i,j}\, b_{i,j'} + \sum_j \sum_{i \neq i'} b_{i,j}\, b_{i',j} + \sum_i \left(\sum_j b_{i,j} - \sum_j b_{j,i} \right)^2 + \sum_{i,j} b_{i,j}\, b_{j,i}$$

[K. Tatsumura *et al.*, *IEEE ISCAS.*, (2020)]

Trading system for cross-currency arbitrage

An end-to-end FPGA-based arbitrage system



<30 microseconds

①Custom I/F (feed handler)
captures market feeds
at unscheduled intervals

②Exchange rate manager
updates an NxN *wij* matrix,
outputs all weights in a single clock

③SB accelerator
searches for an optimal path from all possible paths

④Trading engine
prepares order packets

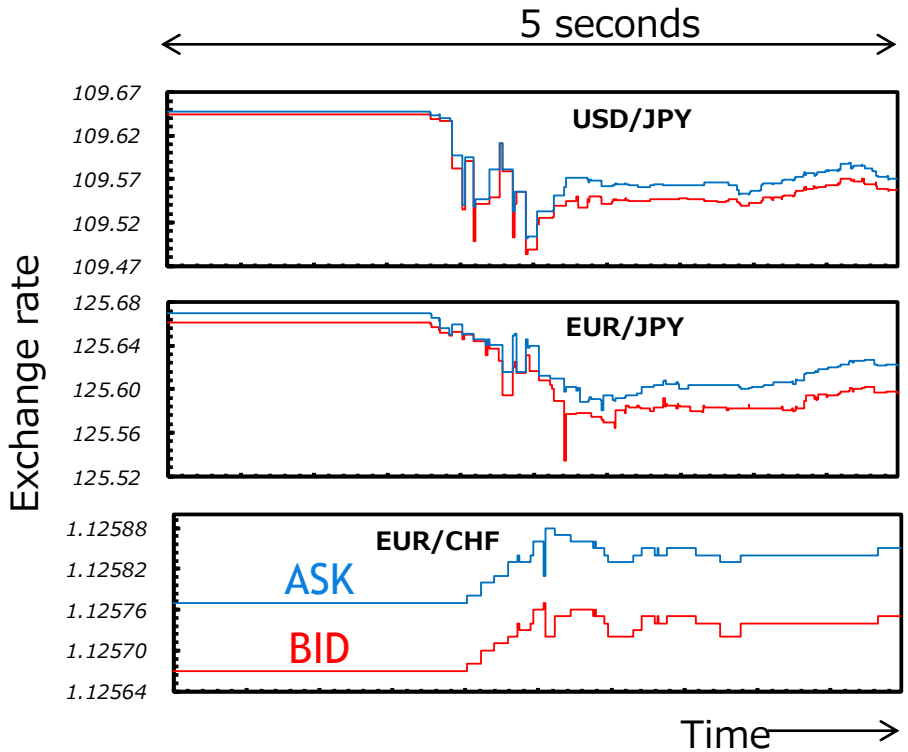
⑤Custom I/F (line handler)
issues the order packets

[K. Tatsumura *et al.*, *IEEE ISCAS.*, (2020)]

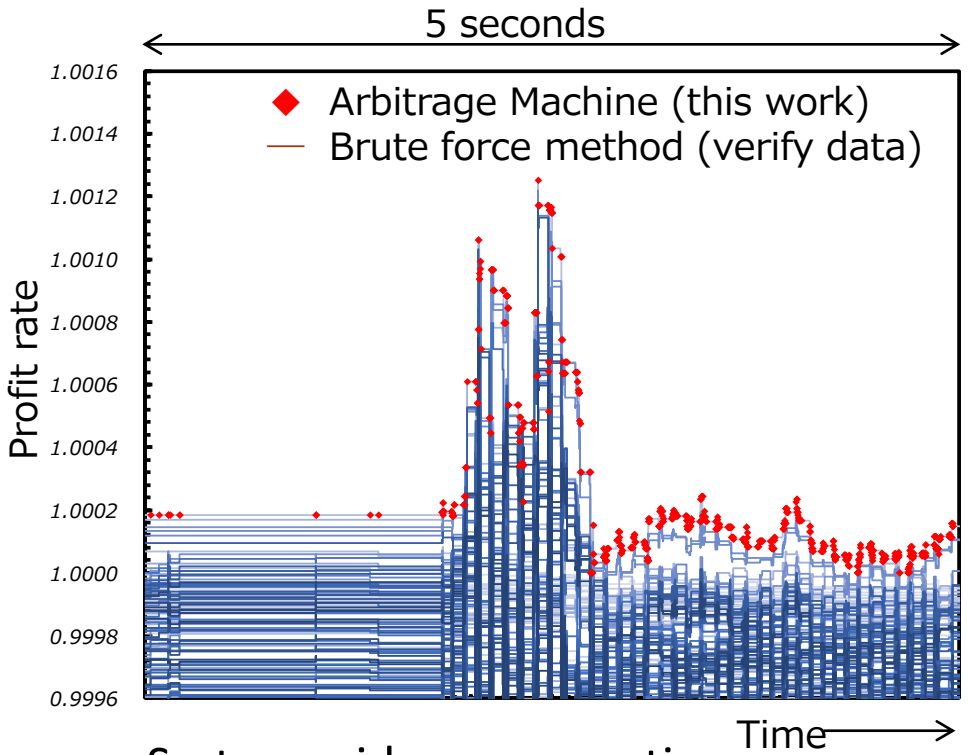
Trading system for cross-currency arbitrage

<30 μs system-wide latency & 91% Top-1 probability

Exchange rates on Jan. 2nd, 2019

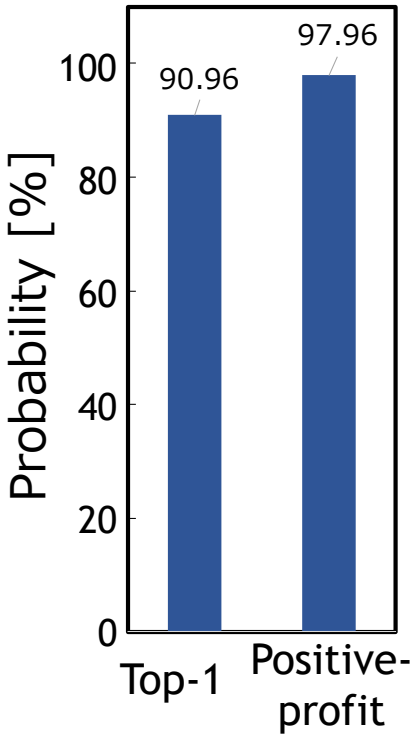


Profit rates for arbitrage paths



System-wide response time:
27.5us (on average over 1000 packets)

Solution accuracy



Conclusion

Simulated bifurcation (SB):

quantum-inspired, highly-parallelizable algorithm for combinatorial optimization

Simulated bifurcation machines (SBM, HW implementation):

efficiently implemented with FPGAs/GPUs, very practical (no refrigerator, no laser)
high performance, very competitive with state-of-the-art Ising machines
embeddable, customizable (FPGA), scalable (FPGA cluster, GPU cluster)
prefer memory-rich architectures, affinity to Al chips

Innovative applications:

Edge(FPGA):
real-time systems that make a rational judgment based on combinatorial optimization
Cloud(GPU):
enabling large/complex combinatorial optimization that was previously impossible

References

Toshiba's website "SQBM+™"
https://www.global.toshiba/ww/products-solutions/ai-iot/sbm.html

[1] Hayato Goto, Kosuke Tatsumura, Alexander R. Dixon, "Combinatorial optimization by simulating adiabatic bifurcations in nonlinear Hamiltonian systems," Science Advances **5**, eaav2372, 2019. https://doi.org/10.1126/sciadv.aav2372

[2] Hayato Goto, Kotaro Endo, Masaru Suzuki, Yoshisato Sakai, Taro Kanao, Yohei Hamakawa, Ryo Hidaka, Masaya Yamasaki, Kosuke Tatsumura, "High-performance combinatorial optimization based on classical mechanics," Science Advances **7**, eabe7953, 2021. https://doi.org//10.1126/sciadv.abe7953

[3] Kosuke Tatsumura, Masaya Yamasaki, Hayato Goto, "Scaling out Ising machines using a multi-chip architecture for simulated bifurcation," Nature Electronics **4**, pp. 208-217, 2021. https://doi.org/10.1038/s41928-021-00546-4

[4] Hayato Goto, "Bifurcation-based adiabatic quantum computation with a nonlinear oscillator network," Scientific Reports **6**, 21686, 2016. https://doi.org/10.1038/srep21686

[5] Hayato Goto, "Quantum Computation Based on Quantum Adiabatic Bifurcations of Kerr-Nonlinear Parametric Oscillators," Journal of the Physical Society of Japan **88**, 061015, 2019. https://doi.org/10.7566/JPSJ.88.061015

[6] Hayato Goto, Taro Kanao, "Chaos in coupled Kerr-nonlinear parametric oscillators," Physical Review Research **3**, 043196, 2021. https://doi.org/10.1103/physrevresearch.3.043196

[7] Taro Kanao, Hayato Goto, "Simulated bifurcation assisted by thermal fluctuation," Communications Physics **5**, 153, 2022. https://www.nature.com/articles/s42005-022-00929-9

[8] Taro Kanao, Hayato Goto, "Simulated bifurcation for higher-order cost functions," Applied Physics Express 16, 014501, 2023. https://doi.org/10.35848/1882-0786/acaba9

[9] Kosuke Tatsumura, Alexander R. Dixon, Hayato Goto, "FPGA-Based Simulated Bifurcation Machine," Proc. of IEEE International Conference on Field Programmable Logic and Applications (FPL), pp. 59-66, 2019. https://doi.org/10.1109/FPL.2019.00019

[10] Kosuke Tatsumura, "Large-scale combinatorial optimization in real-time systems by FPGA-based accelerators for simulated bifurcation," Int'l Symp. on Highly Efficient Accelerators and Reconfigurable Technologies (HEART), 2021. https://doi.org/10.1145/3468044.3468045

[11] Kosuke Tatsumura, Ryo Hidaka, Masaya Yamasaki, Yoshisato Sakai, Hayato Goto, "A Currency Arbitrage Machine based on the Simulated Bifurcation Algorithm for Ultrafast Detection of Optimal Opportunity," Proc. of IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5, 2020. https://doi.org/10.1109/ISCAS45731.2020.9181114

[12] Nasa Matsumoto, Yohei Hamakawa, Kosuke Tatsumura, Kazue Kudo, "Distance-based clustering using QUBO formulations," Scientific Reports **12**, 2669, 2022. https://doi.org/10.1038/s41598-022-06559-z

[13] Kyle Steinhauer, Takahisa Fukadai, Sho Yoshida, "Solving the Optimal Trading Trajectory Problem Using Simulated Bifurcation," arXiv preprint arXiv:2009.08412, 2020. https://doi.org/10.48550/arXiv.2009.08412

[14] Tingting Zhang, Qichao Tao, Jie Han, "Solving Traveling Salesman Problems Using Ising Models with Simulated Bifurcation," Proc. of International SoC Design Conference (ISOCC), pp.288-289, 2021. https://doi.org/10.1109/ISOCC53507.2021.9613918

[15] W. Zhang, Y.-L. Zheng, "Simulated Bifurcation Algorithm for MIMO Detection," arXiv:2210.14660, 2022. https://arxiv.org/abs/2210.14660

[16] G. Finocchio, K. Tatsumura, Hayato Goto et al. "Roadmap for Unconventional Computing with Nanotechnology," arXiv: 2301.06727, 2023. https://doi.org/10.48550/arXiv.2301.06727

[17] Naeimeh Mohseni, Peter L. McMahon, Tim Byrnes, "Ising machines as hardware solvers of combinatorial optimization problems," Nature Reviews Physics, 2022. https://doi.org/10.1038/s42254-022-00440-8

[18] Hiroki Oshiyama, Masayuki Ohzeki, "Benchmark of quantum-inspired heuristic solvers for quadratic unconstrained binary optimization," Scientific Reports **12**, 2146, 2022. https://doi.org/10.1038/s41598-022-06070-5