TOSHIBA

IEEE Int’l Workshop on Ising Machines
(Invited) 10:30～11:00, Apr. 17, 2024

Simulated bifurcation machines

Enabling NP-hard optimization-based judgement in real-time systems by quantum-inspired technology

Kosuke Tatsumura

Toshiba Corporation, JAPAN

Outline

01

Introduction

Using quantum-inspired optimization technology in real-time systems

02

Simulated bifurcation machine (SBM): theory, implementation & performance

03

Real-time systems with Simulated bifurcation machines (SBMs): Finance and Automotive vehicle

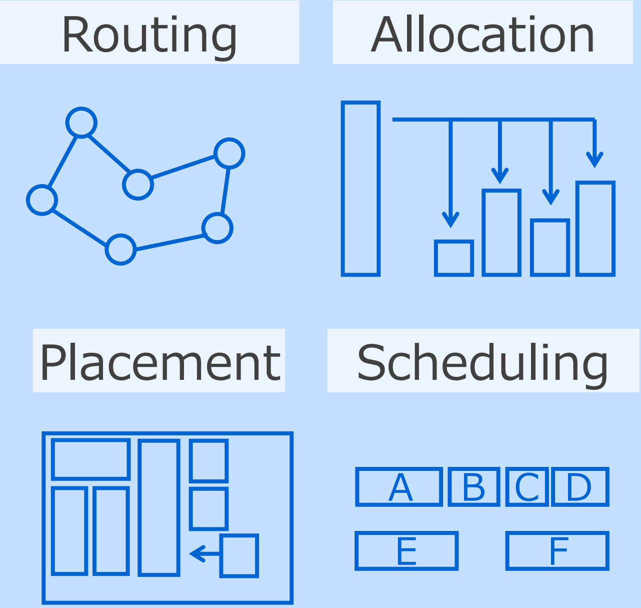
Combinatorial optimization

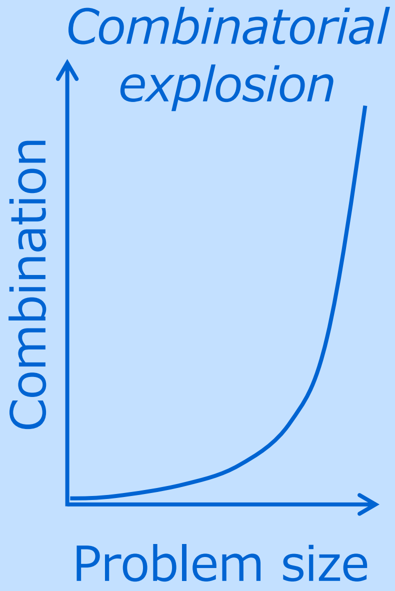
Economically important but computationally hard

Practical problems

decision-making, planning, etc

Combinatorial optimization





Nondeterministic polynomial time (NP)-hard

From the viewpoint of system engineering:

Making *rational* judgment in the situation/at the moment

#1 Andrew Lucas, "Ising formulations of many NP problems", *Frontiers in Physics* **2**, 5 (2014)

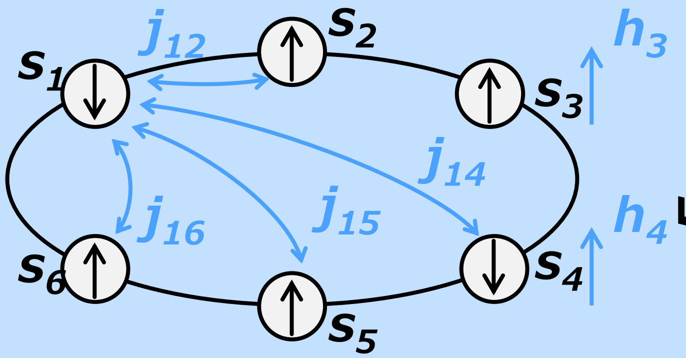
Ising machine

Special-purpose computer for combinatorial optimization

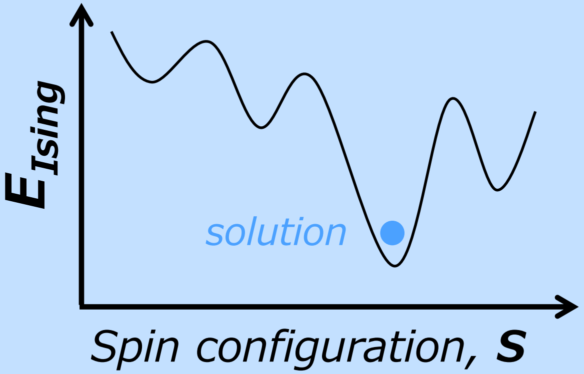
Ising problem

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

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



spin: *binary variable*



Mapping[#1]

Combinatorial optimization
(Quadratic discrete optimization)

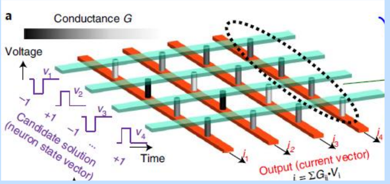
Ising machines

Quantum annealer[*1]
(2011)

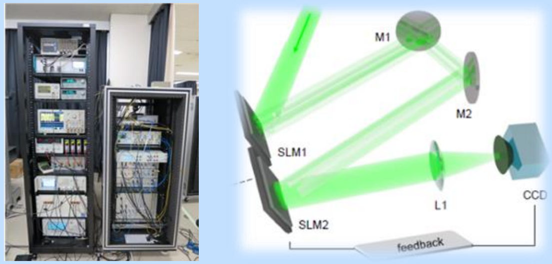
CMOS
annealer[*2]

Digital
annealer[*3]

Memristor HNN[*6]



Optical Ising machines[*4,5]



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).

Quantum-inspired optimization technology

*1 H. Goto *et al.*, *Sci. Adv.* **5**, eaav2372, '19
*2 H. Goto *et al.*, *Sci. Adv.* **7**, eabe7953, '21
*3 K. Tatsumura *et al.*, *Nat. Ele.* **4**, 208–217, '21
*4 T. Kashimata *et al.*, *IEEE Access* **12**, '24

Simulated Bifurcation Machine, SBM

Algorithm

Quantum-inspired

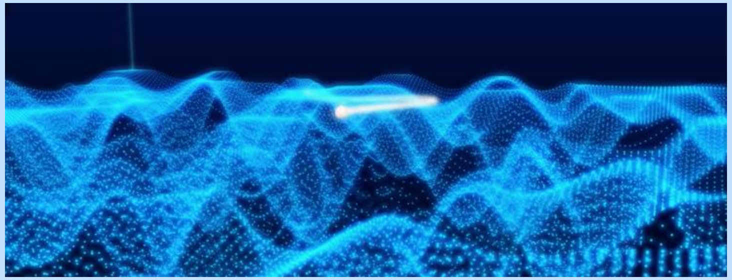
Quantum bifurcation machine

in a quantum principle

Discovery

Simulated bifurcation algorithm

in a new classical principle



Highly parallelizable

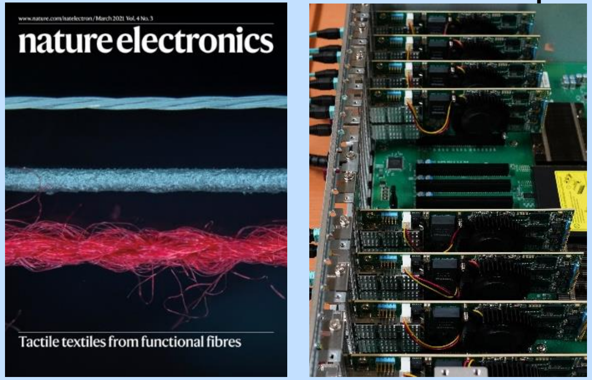
Implementation

High performance*1,2

single-chip

Scalable*3,4

multi-chip

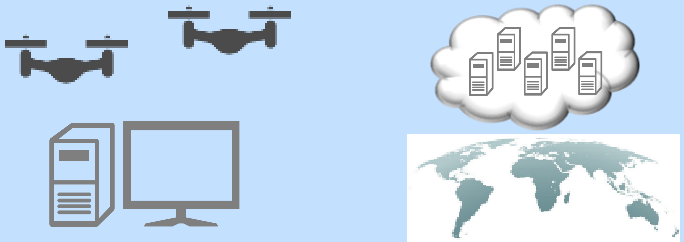


Application

Very practical

edge/embedded

cloud



Innovative

ex. real-time systems

Edge application of SBM
Simulated bifurcation machine

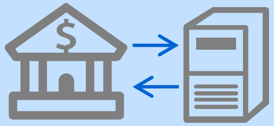
*1 K. Tatsumura *et al.*, "FPGA-Based Simulated Bifurcation Machine," *IEEE Field-Programmable Logic and Apps.* (FPL), 2019
*2 K. Tatsumura, "Large-scale combinatorial optimization in real-time systems by FPGA-based accelerators for simulated bifurcation," *Int'l Symp. on Highly Efficient Accelerators and Reconfigurable Technologies* (HEART), 2021

Enables real-time systems that make more rational judgments*[1,2]
Constrains in the time domain

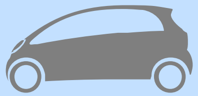
Real-time system

- Must respond within the critically defined time constrain
- Respond time needed for "high-speed" real-time systems:
  Typically less than 1 ms
- (Conventional) simple conditional-judgement

Financial trading

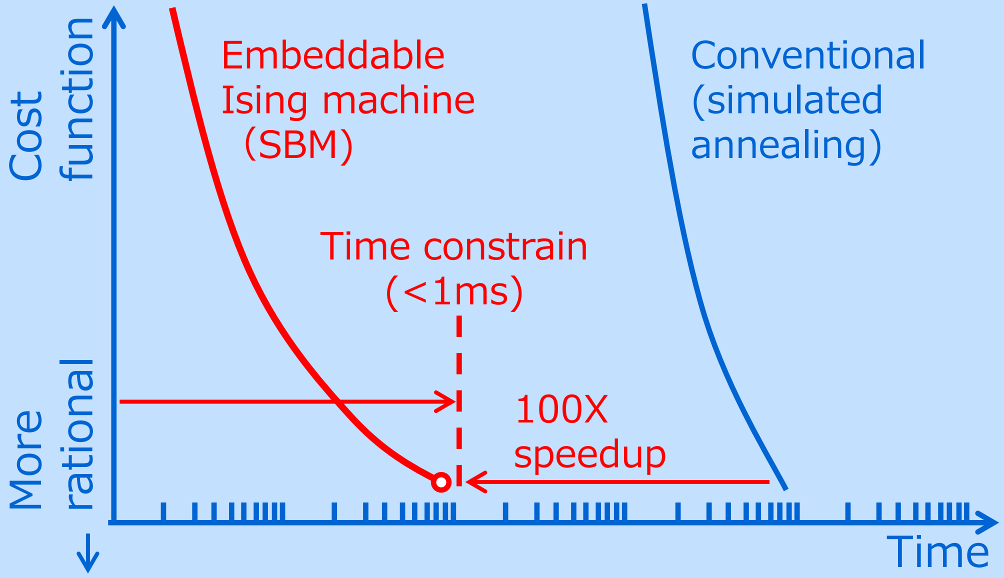


Automotive Vehicle

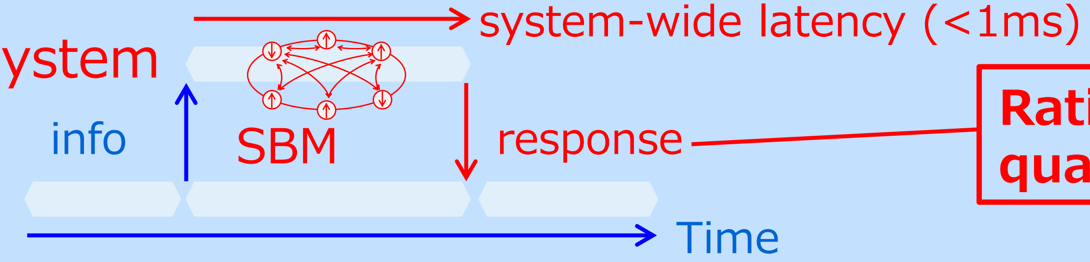


Embedding

Embeddable Ising machine (SBM)



Real-time system with embeddable SBM



Environment
・Financial market
・Surrounding situation

Rational response based on quadratic discrete optimization

Outline

01 Introduction
Using quantum-inspired optimization technology in real-time systems

02 Simulated bifurcation machine (SBM): theory, implementation & performance

03 Real-time systems with Simulated bifurcation machines (SBMs): Finance and Automotive vehicle

SB theory: How it was born

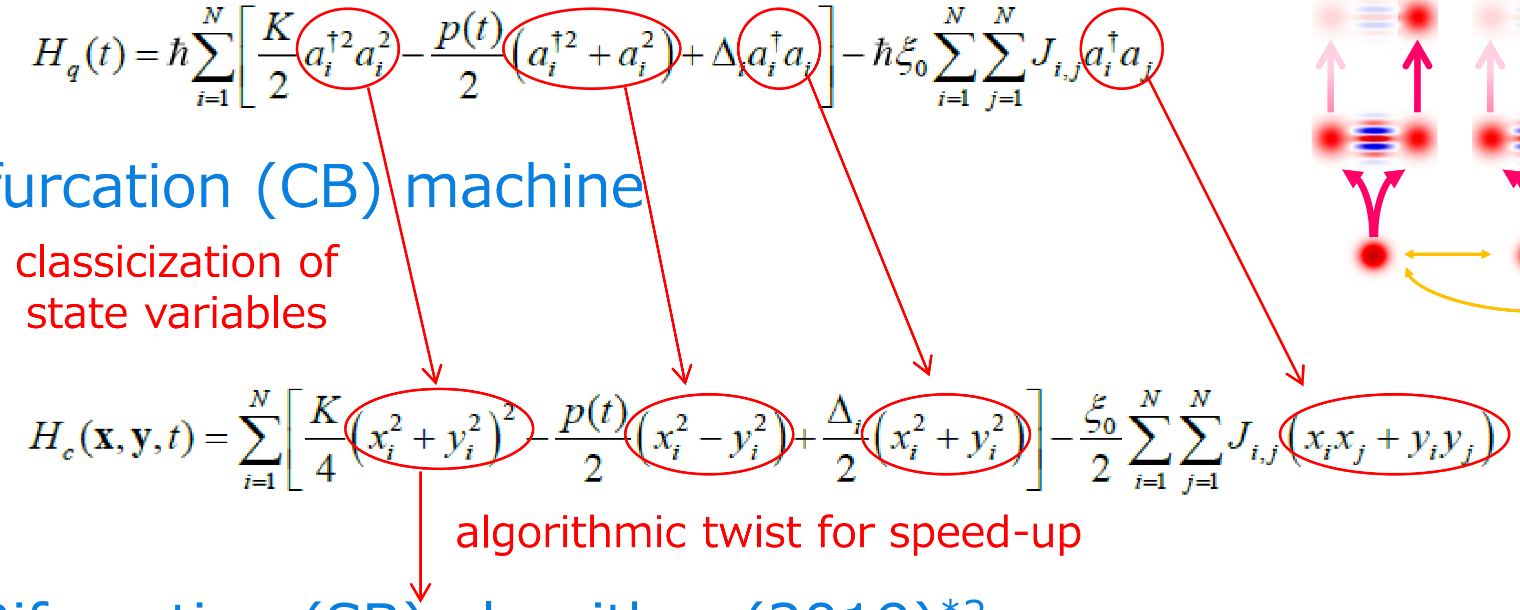
*1 H. Goto *et al.*, *Scientific Reports* **6**, 21686 (2016)

*2 H. Goto *et al.*, *Science Advances* **5**, eaav2372 (2019)

Quantum-inspired algorithm

Quantum Bifurcation (QB) machine[*1]

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

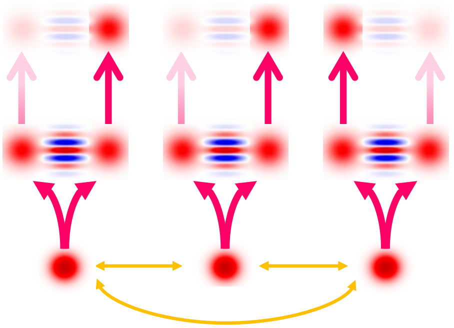
classicization of
state variables

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

algorithmic twist for speed-up

Simulated Bifurcation (SB) algorithm (2019)[*2]

Combinatorial optimization
based on quantum adiabatic theorem



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

SB theory: Why it works

*1 H. Goto *et al.*, *Science Advances* **5**, eaav2372 (2019)
*2 G. Finocchio *et al.*, *Nano Futures* **8**, 012001 (2024)
*3 H. Goto *et al.*, *Science Advances* **7**, eabe7953 (2021)
*4 T. Kanao *et al.*, *Comm. Phys.* **5**, 153 (2022)
*5 T. Kanao *et al.*, *Applied Physics Express* **16**, 014501(2023)

New classical principle: Adiabatic and ergodic search^{*1}

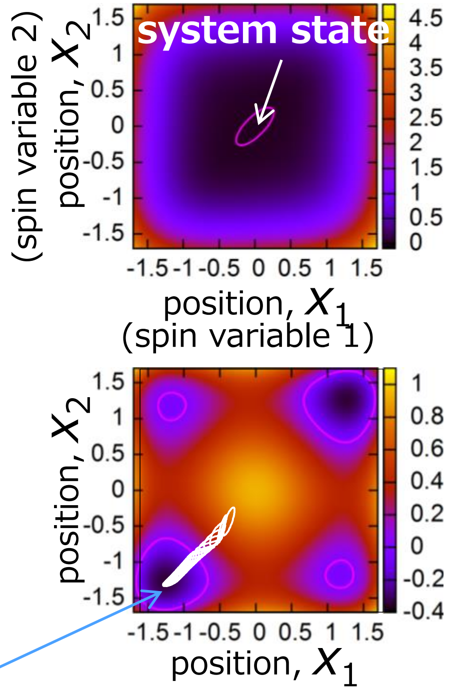
Dynamical change of energy landscape

a single local minimum

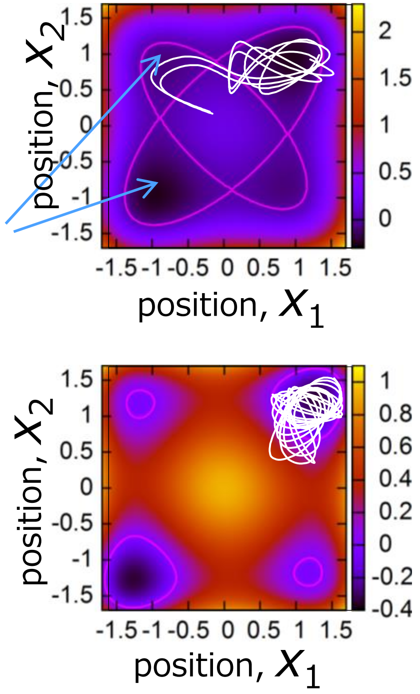
bifurcation
(adiabatic process)

multiple local minima
(target cost function)

Energy landscape (N_{spin}=2)



Multiple minima in the energetically allowable region



best solution
(-1,-1)

adiabatic search
chase one of the minima

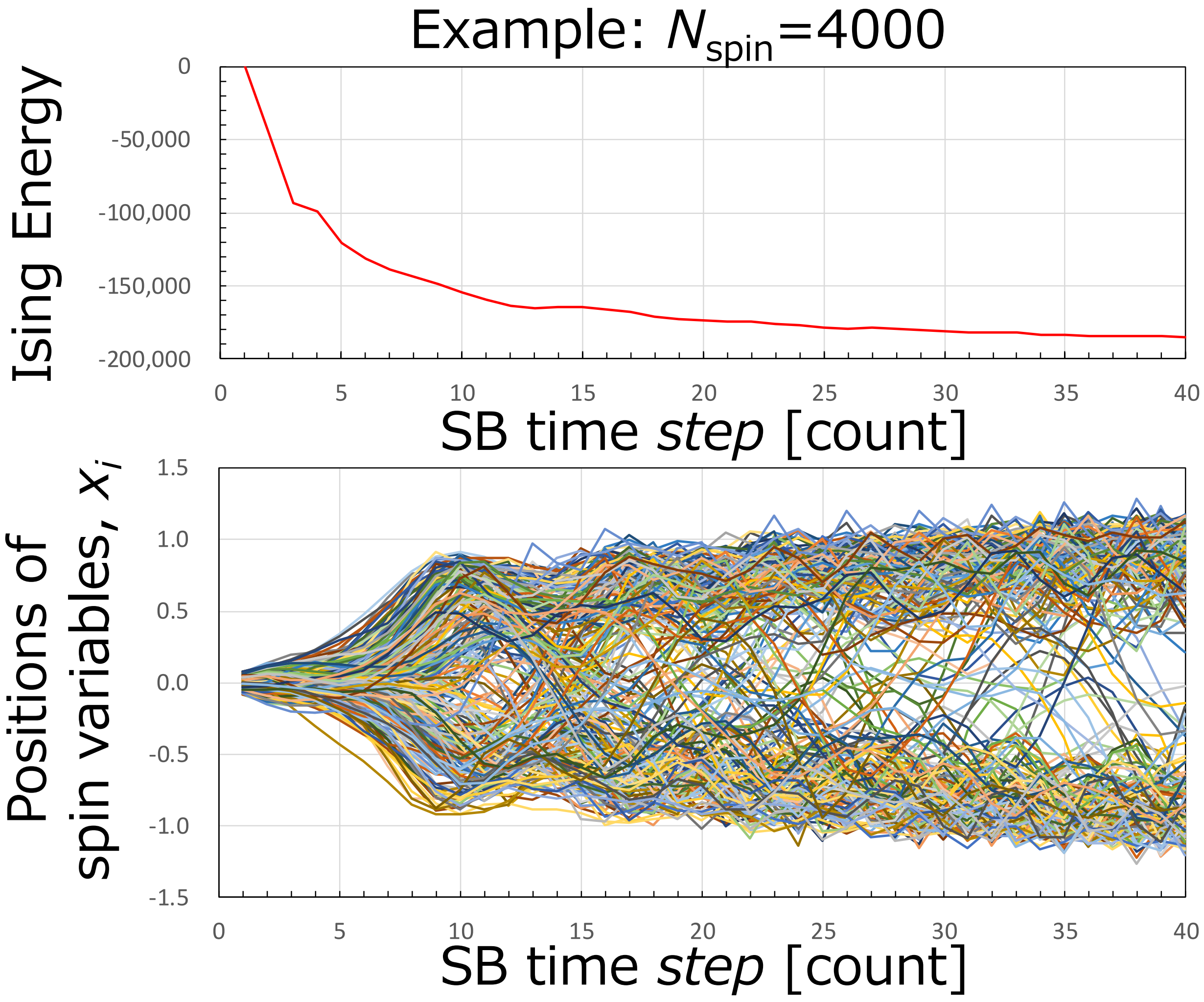
ergodic search
find a better one with a higher probability

Implying *classical* adiabatic theorem (corresponding to quantum adiabatic theorem)^{*2} has been extended to 2$^{\mathrm{nd}}$-gen SB^{*3}, heat-assisted SB^{*4}, and higher-order SB^{*5}

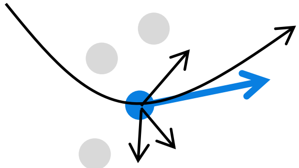
SB algorithm: How it works

“*N*-body”-type algorithmic structure[*1] → Highly parallelizable

Example: N_{spin}=4000



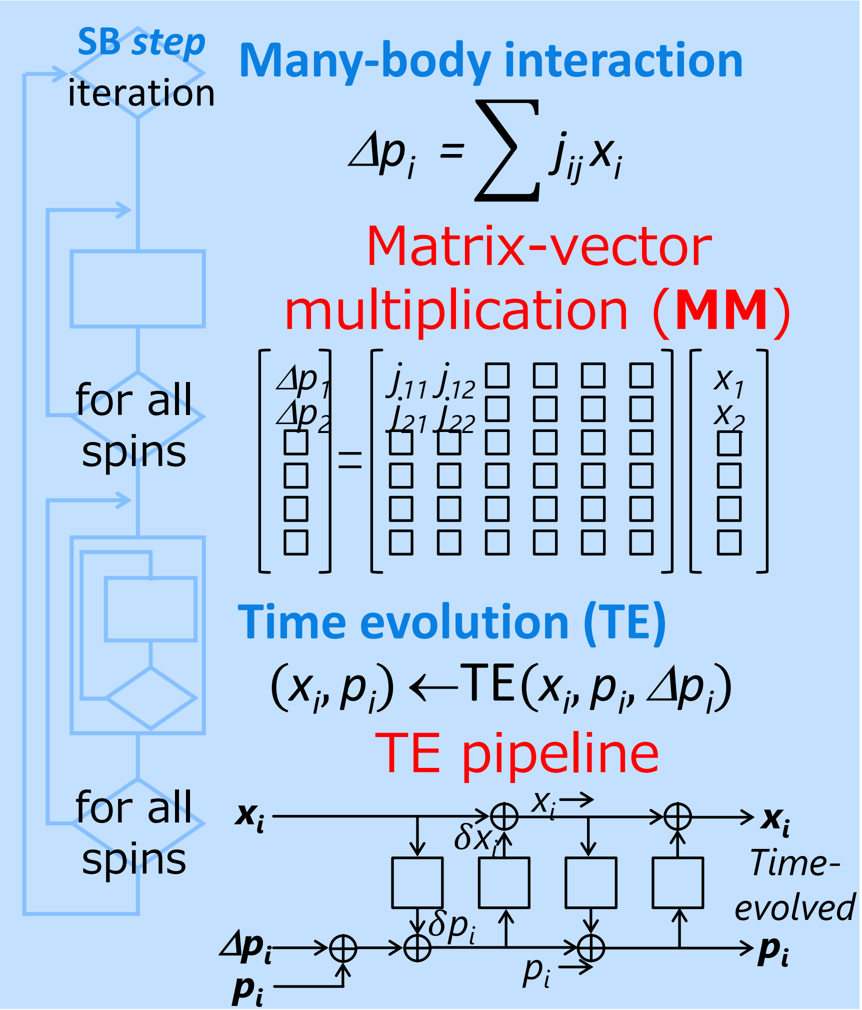
better solution



Joint force

Many-body interaction

depending on all the other spins



Many-body interaction

$$\Delta p_i = \sum j_{ij} x_i$$

Matrix-vector multiplication (**MM**)

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

Time evolution (TE)

$$(x_i, p_i) \leftarrow \mathrm{TE}(x_i, p_i, \Delta p_i)$$

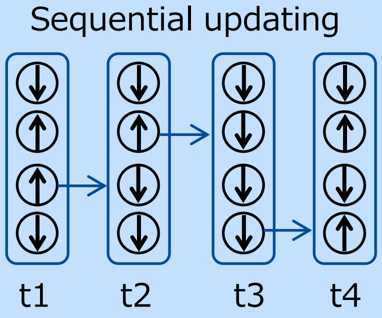
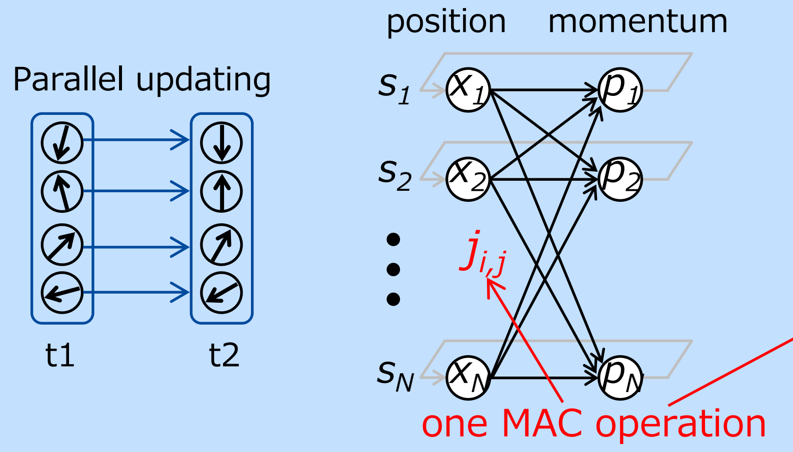
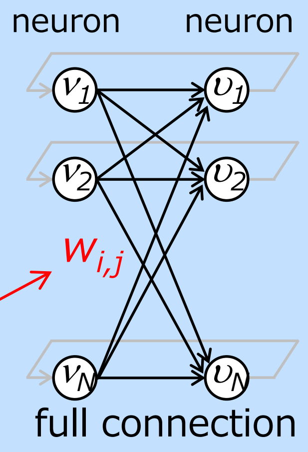
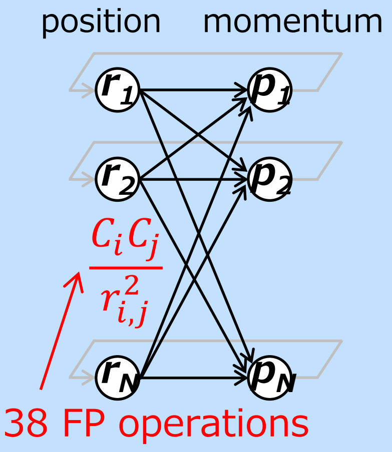
TE pipeline

[*1] Time-evolution simulation of *N*-particle systems with long-range pairwise interaction, like celestial systems with gravitational interaction or molecular systems with Coulomb interaction

*1 FPGA (Field programmable gate array), GPU (Graphic processing unit), NPU (Neural processing unit), ASIC (application specific integrated circuit)

SB algorithm : Characteristics

More parallelizable than SA, Can be accelerated with FPGAs/GPUs/NPUs

	SA simulated annealing	SB simulated bifurcation	R-NN recurrent neural network	*N*-body gravitational (/Coulomb)-force
Structure				
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

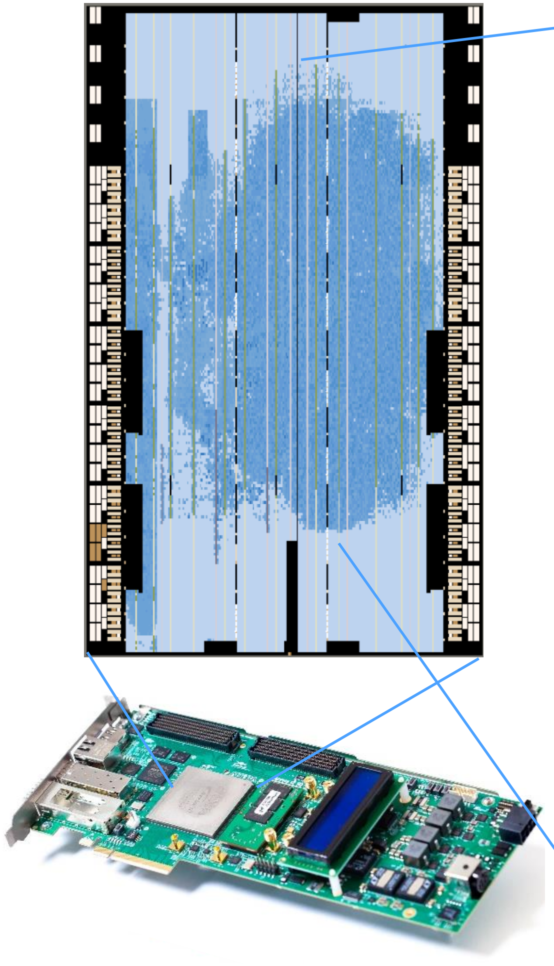
SB can be accelerated with FPGAs/GPUs (not limited to special ASICs)
Many AI chips (NPUs) are beneficial also to SB

*1 K. Tatsumura *et al.*, "FPGA-Based Simulated Bifurcation Machine," *IEEE Field-Programmable Logic and Apps.* (FPL), 2019

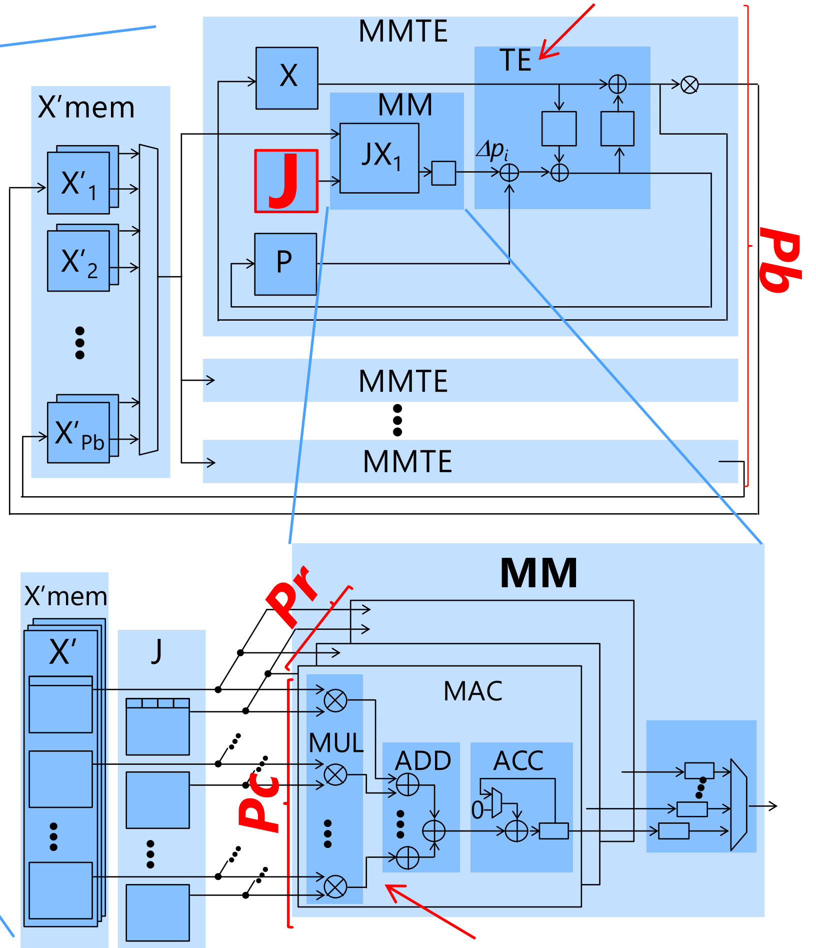
FPGA-based accelerator for SB*1

Large-scale, massively parallel, and high utilization

Arria10 GX1150 FPGA



Temporal parallelization



Spatial parallelization

Problem	complete-graph MAX-CUT
Machine size	4,096 spins full spin-spin connectivity
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)

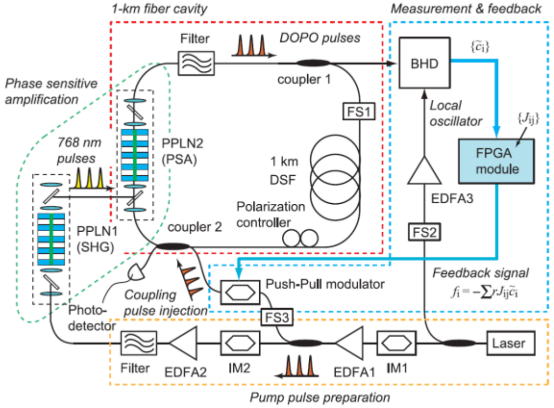
*1 H. Goto *et al.*, *Science Advances* **5**, eaav2372 (2019)

Performance*1 (2019)

Faster and more efficient than the state-of-the-art one

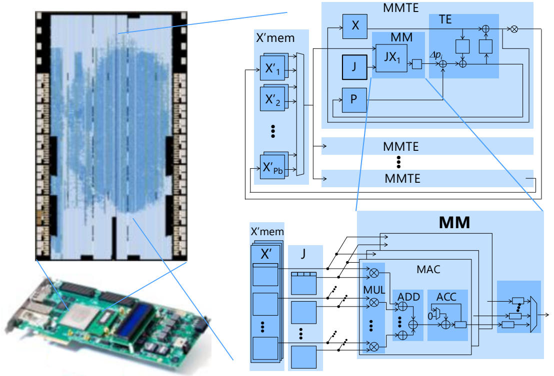
Coherent Ising Machine

800 GMAC/s @ 1000 W

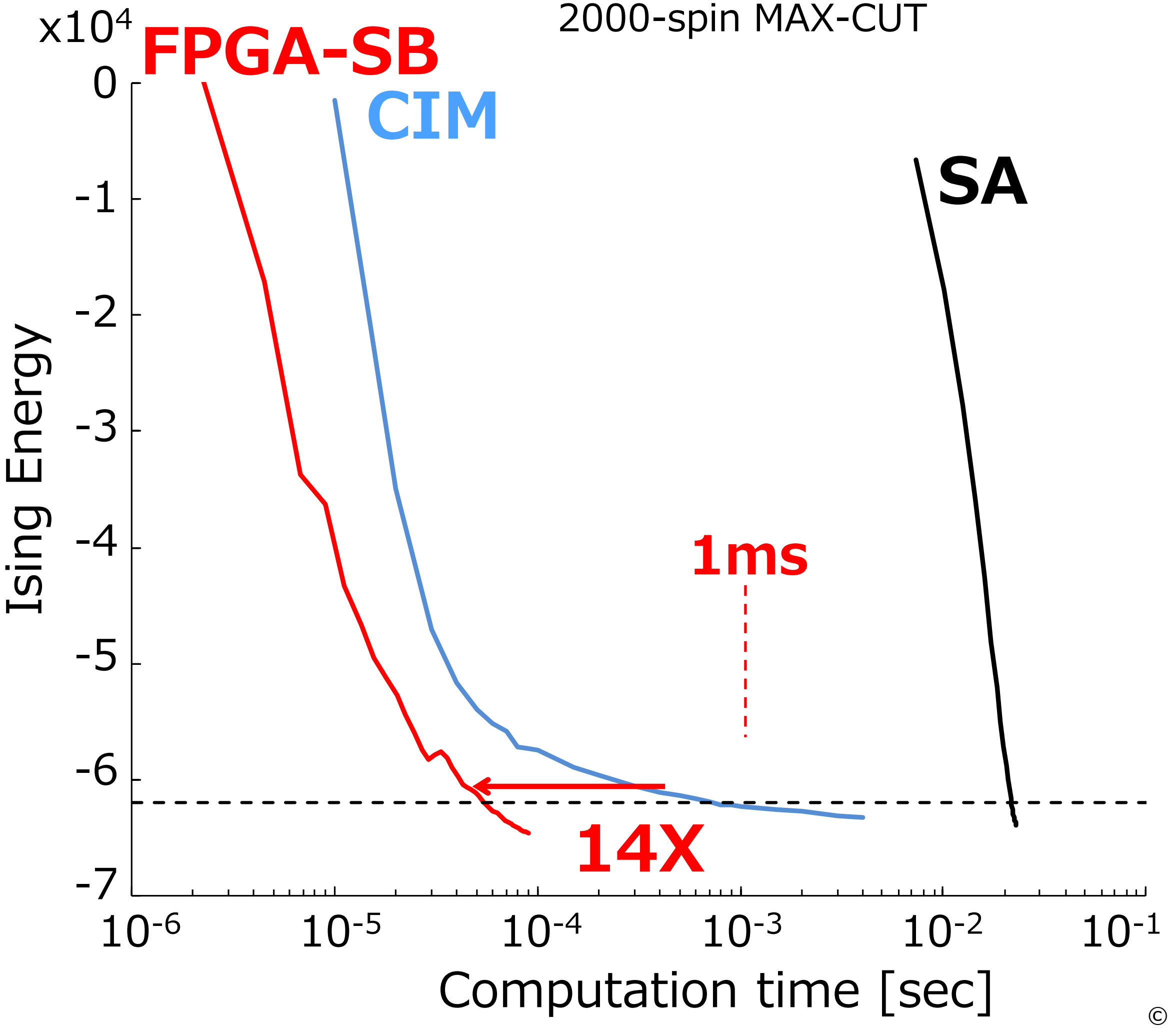


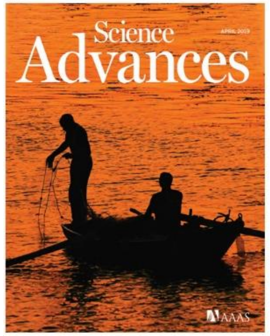
[T. Inagaki, Science 354, 603, '16]

FPGA-SB 1,873 GMAC/s @ 49 W
(288X more energy efficient)



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





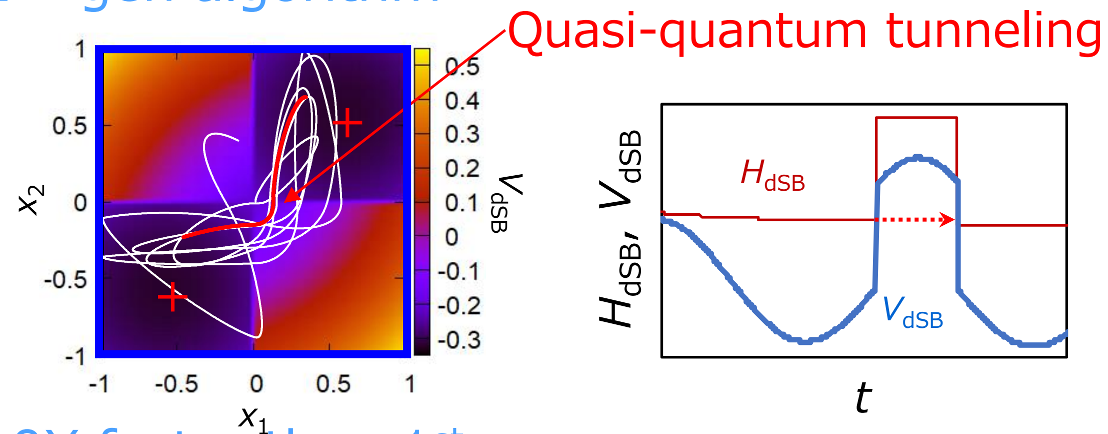
*1 H. Goto *et al.*, *Science Advances* **7**, eabe7953 (2021)

Performance*1 (2021)

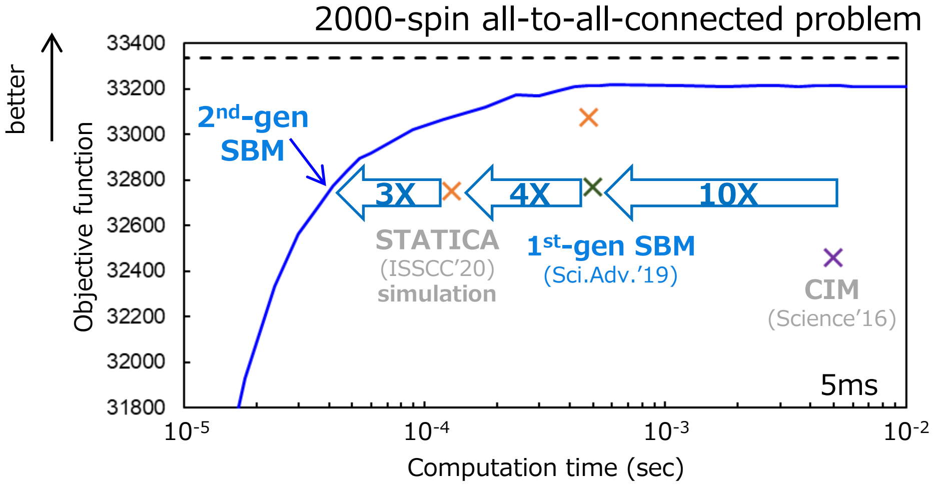
Comprehensive comparison →

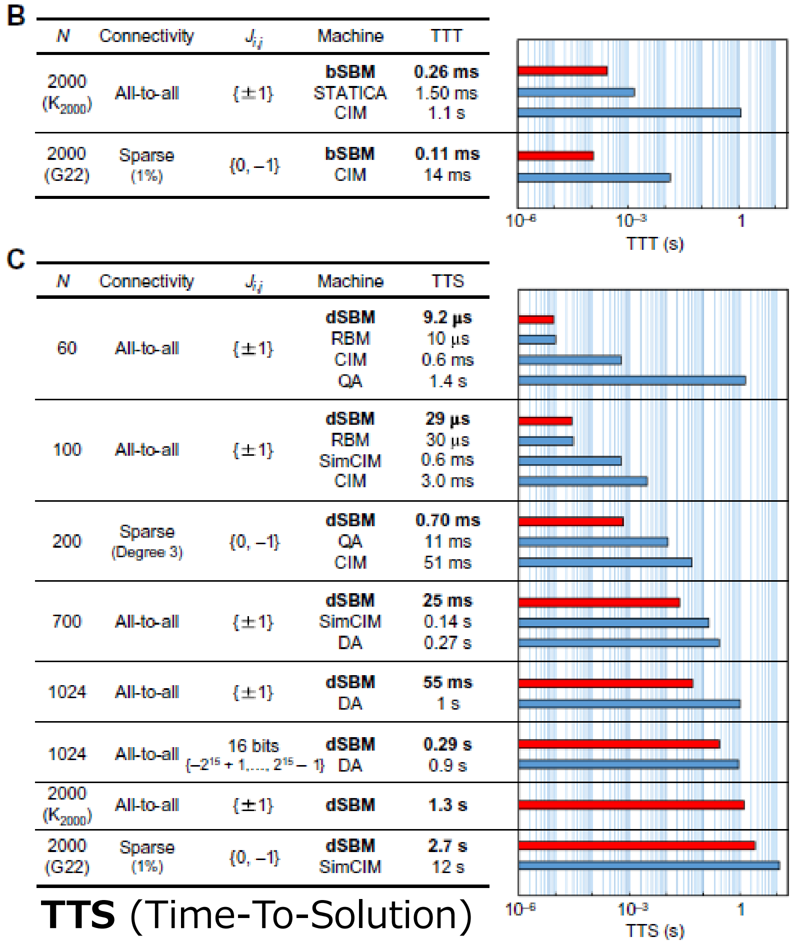
Very competitive with state-of-the-art Ising machines

2nd-gen algorithm



10X faster than 1st-gen





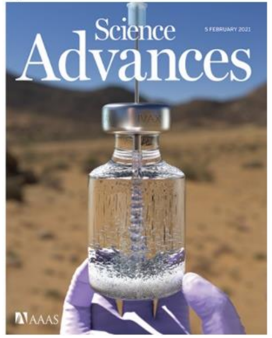
B

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

C

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

Outline

01

Introduction

Using quantum-inspired optimization technology in real-time systems

02

Simulated bifurcation machine (SBM): theory, implementation & performance

03

Real-time systems with Simulated bifurcation machines (SBMs): Finance and Automotive vehicle

Application of SBMs[1]

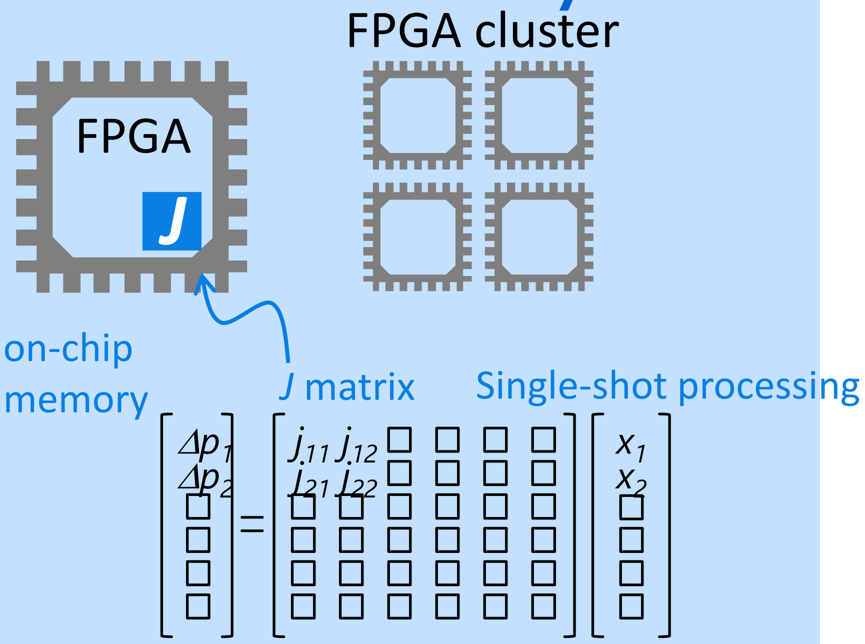
Simulated bifurcation machines

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

for Edge (high-speed real-time) **and Cloud** (large/wide-area)

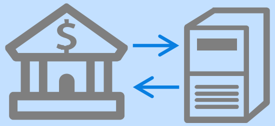
edge/embedded

Lower latency ⟷ Larger problem

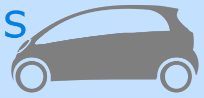


High-speed real-time systems

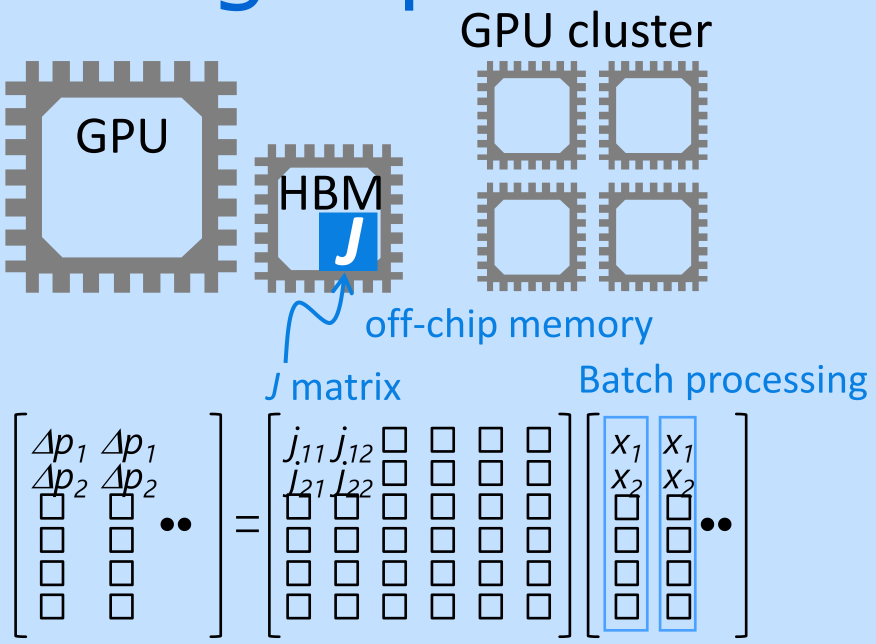
Financial trading



Autonomous control



cloud



Large/wide-area systems

Drug design

Planning

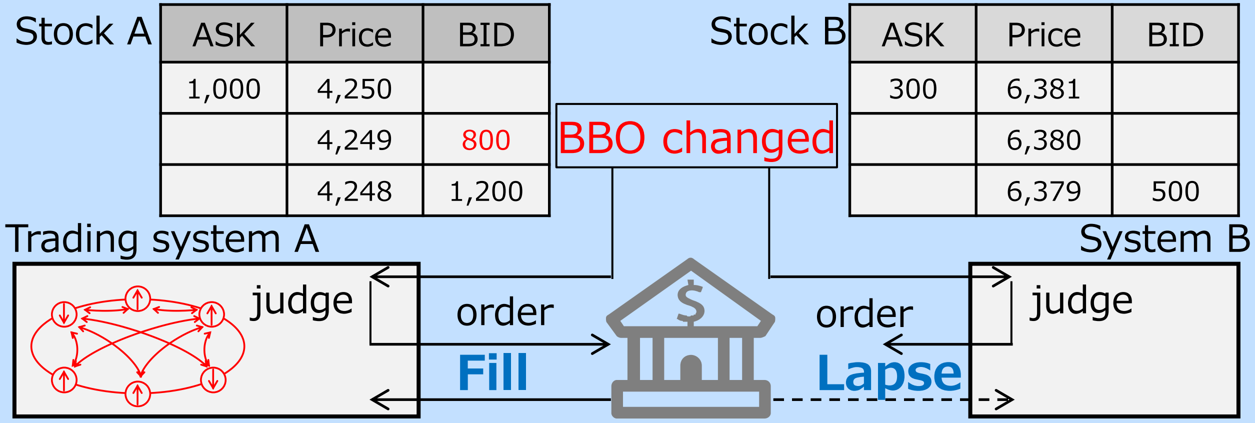
Embeddable SBM
Simulated bifurcation machine

Enabling NP-hard optimization in real-time systems

rational judgment

High-speed real-time systems

Financial trading system



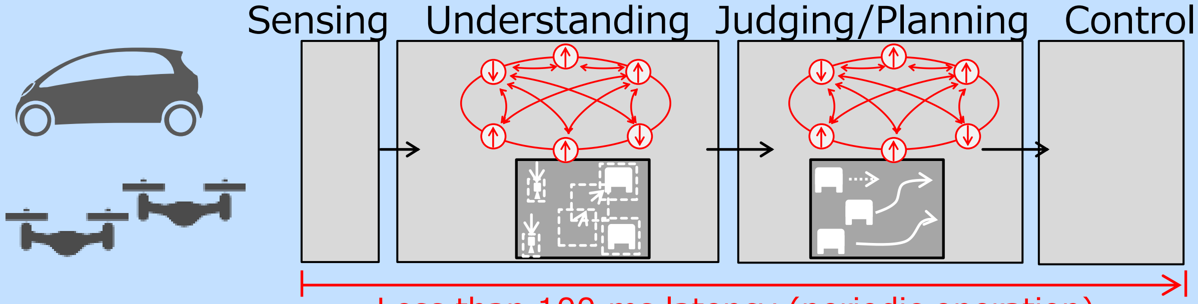
Stock A

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

Autonomous control

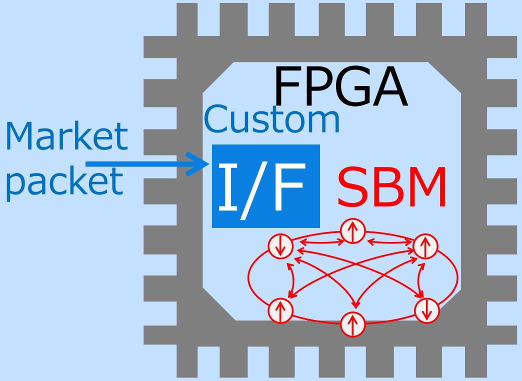


Less than 100 ms latency (periodic operation)

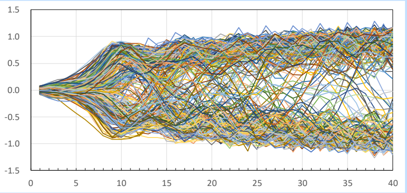
FPGA-based SBMs

Ultralow latency (sub-msec)

Deterministic latency



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



Predetermined #step

---

*1 K. Tatsumura *et al.*, "FPGA-Based Simulated Bifurcation Machine," *IEEE Field-Programmable Logic and Apps. (FPL)*, 2019

*2 K. 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

Applications in Finance

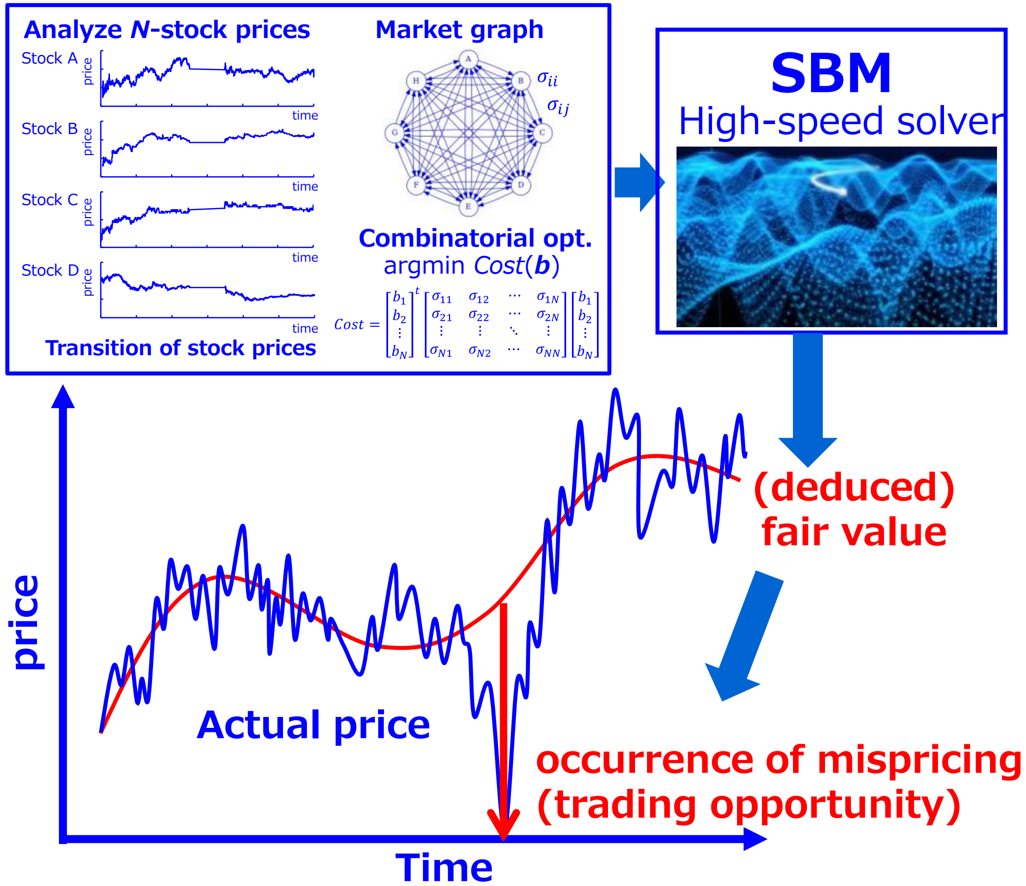
#1 The world's first demonstration of systems that execute unprecedented stock trading strategies based on computationally-hard quadratic discrete optimization by using quantum-inspired computer

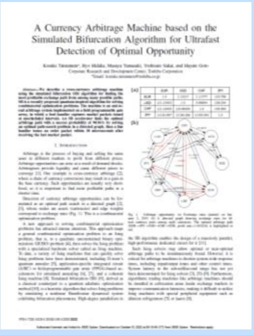
*1 https://doi.org/10.1109/ISCAS45731.2020.9181114
*2 https://doi.org/10.1109/ACCESS.2023.3316727
*3 https://doi.org/10.1109/ACCESS.2023.3326816
*4 https://doi.org/10.1109/ACCESS.2023.3341422

New strategies based on detection of ever-untargeted trading opportunities by SBMs

Even if your strategy is splendid, if there is a competitor executing the same one, you may lose

Traders are essentially ***technology-hungry***, pursuing a new strategy by new technologies



Type	High-speed real-time trading			Asset management
Target	Currency	Stock		
Strategy	Cross-currency arbitrage*1	Extended pair-trade*2	High-speed basket trade*3	Correlation-diversified portfolio*4
Opt.	Optimal path search in market graph	Optimal path search in market graph	Discrete portfolio optimization	Maximum independent set
Paper	 A Currency Arbitrage Machine based on the Simulated Bifurcation Algorithm for Ultrafast Detection of Optimal Opportunity	 RESEARCH ARTICLE Pairs-Trading System Using Quantum-Inspired Combinatorial Optimization Accelerator for Optimal Path Search in Market Graphs	 IEEE Access RESEARCH ARTICLE Real-Time Trading System Based on Selections of Potentially Profitable, Uncorrelated, and Balanced Stocks by NP-Hard Combinatorial Optimization	 IEEE Access RESEARCH ARTICLE Correlation-Diversified Portfolio Construction by Finding Maximum Independent Set in Large-Scale Market Graph
Publish	Oct. 12, 2020	Sep. 18, 2023	Oct. 23, 2023	Dec. 12, 2023

*1 K. Tatsumura *et al.*, "Real-time Trading System based on Selections of Potentially Profitable, Uncorrelated, and Balanced Stocks by NP-hard Combinatorial Optimization," *IEEE Access* **11**, pp. 120023 - 120033 (2023).

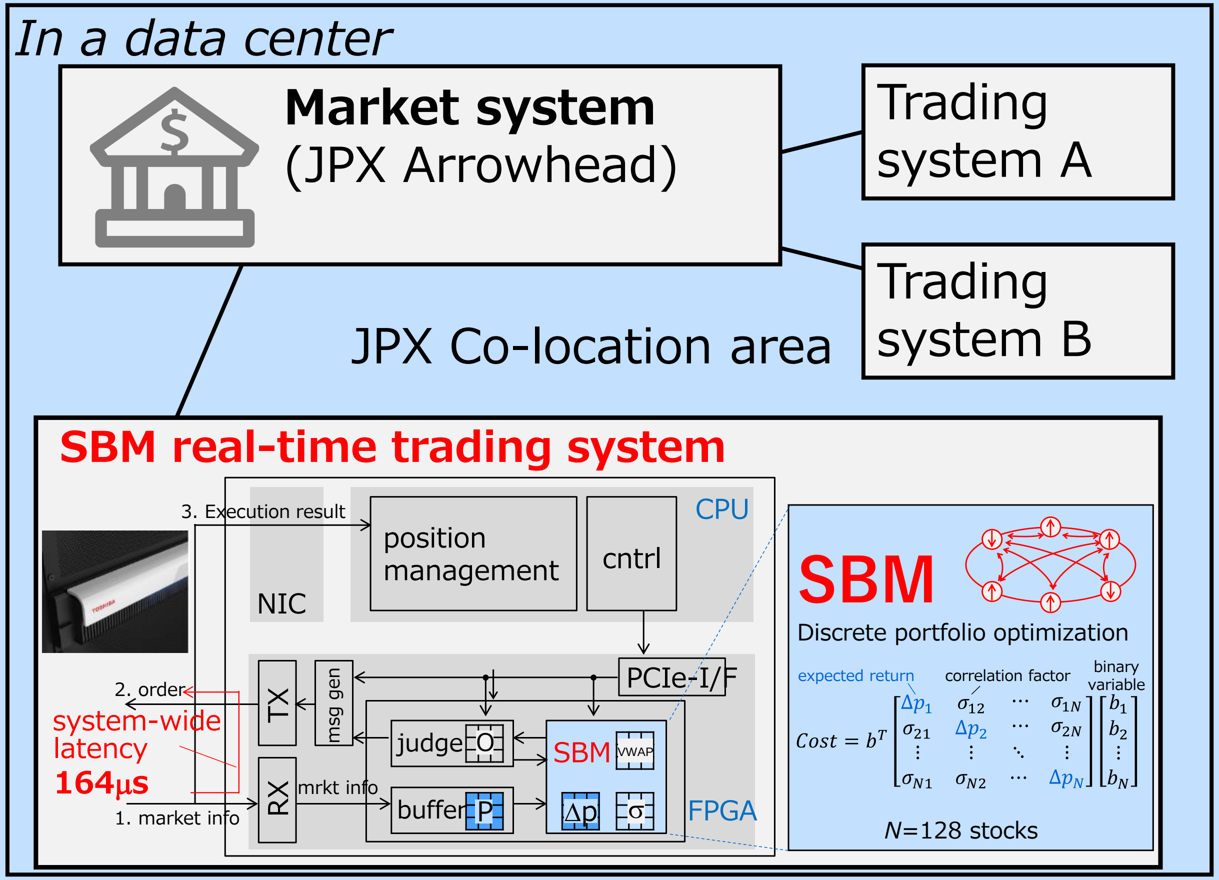
System –High-speed basket trade*1-

Real-time trading system with embedded SBM (164µs latency), installed at the JPX Co-location area of the TSE

JPX Co-location area: special area for high-speed trading

TSE: TOKYO Stock Exchange

System configuration

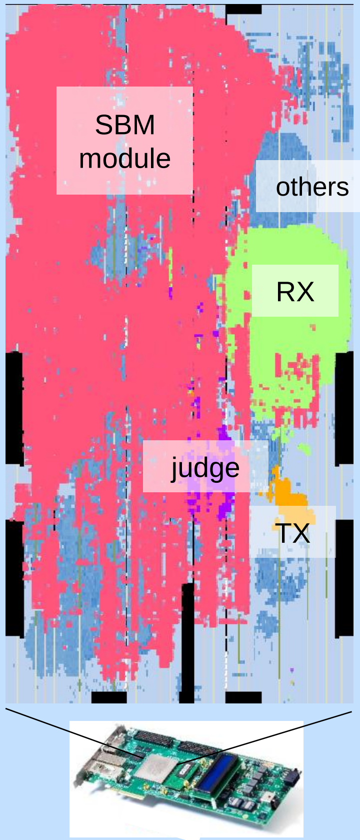


$$Cost = b^T \begin{bmatrix} \Delta p_1 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \Delta p_2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \Delta p_N \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}$$

N=128 stocks

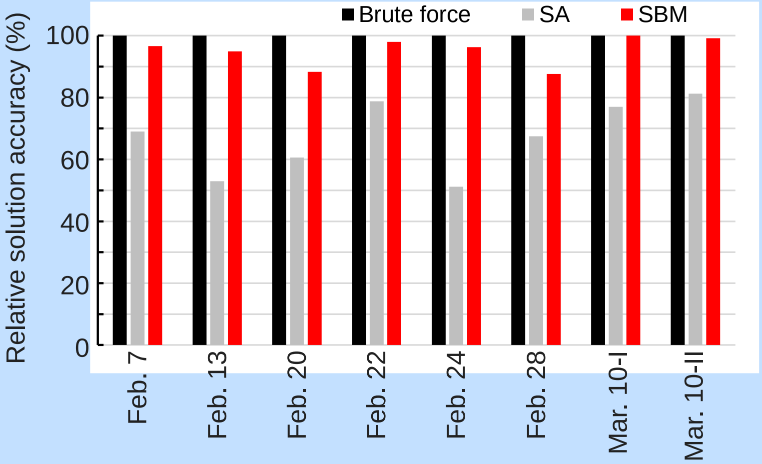
Embedded SBM

FPGA



SB alg. ver	Ballistic SB (bSB)
Machine size	Fully-connected 128 spins
Comp. precision	32-bit FP
Time per SBM run	162 µS

Single-shot solution accuracy

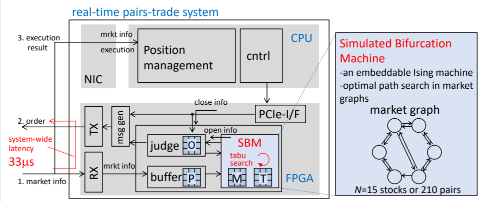
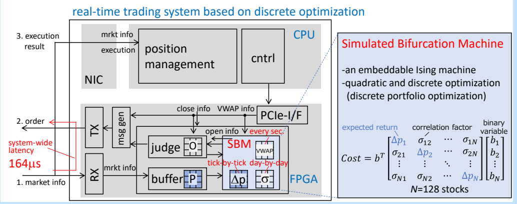
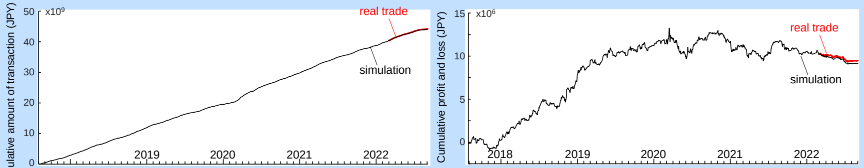
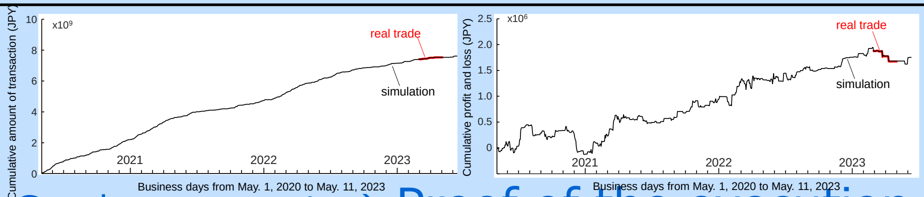


Market situations from Feb. 7, 2023 to Mar. 10, 2023

*1 K. Tatsumura et al., *IEEE Access 11*, Sep. 18, 2023
*2 K. Tatsumura et al., *IEEE Access 11*, Oct. 23, 2023

Demonstration at the Tokyo Stock Exchange

Execution capability of NP-hard optimization-based trading strategy through real-time/real-money transactions

Strategy & System	**Extended pair-trade*1**	**High-speed basket trade*2**
	 Optimal path search in market graph (a tabu search)	 Discrete portfolio optimization
Real-time transactions vs. backsimulation assuming 100% fill rate *Execution at intended prices & volumes*	 Good agreement → Proof of the execution capability in terms of speed	 Good agreement → Proof of the execution capability in terms of speed
System-wide latency	33 µsec for 210-pair universe	164 µsec for 128-stock universe
Sharpe ratio (annualized return/risk)	0.79 (=7.5% / 9.5%)	1.23 (=3.6% / 2.9%)
Cumulative amounts of transaction	3,817,201,458 JPY *(Total)4 billion JPY transactions, 1000-hour no-error operation*	118,956,828 JPY
Real-time trading hours	125 days or 750 hours (No errors)	42 days or 252 hours (No errors)

Strategy –High-speed basket trade[*1]-

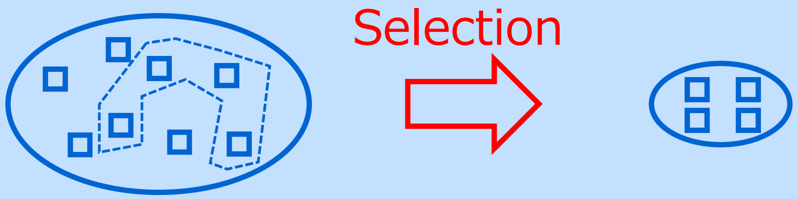
[*1] K. Tatsumura *et al.*, "Real-time Trading System based on Selections of Potentially Profitable, Uncorrelated, and Balanced Stocks by NP-hard Combinatorial Optimization," *IEEE Access* **11**, pp. 120023 - 120033 (2023).

[*2] VWAP: Volume-Weighted Average Price

Select a delta-neutral basket of stocks to maximize return and minimize correlation (risk) for improving Sharpe ratio

Formulation of QUBO

A typical "quadratic discrete optimization problem" (NP-hard)



Tradable stocks ***N*** (***N***=128)

Selection

Selected basket $\boldsymbol{N_s}$

Binary variables

$$\boldsymbol{b} = [b_1 \quad b_2 \quad \cdots \quad b_N] \qquad b_i \in \{\overset{\text{Unselected}}{0}, \overset{\text{Buy/Sell}}{1}\}$$

Total cost function

$$H_{QUBO} = \sum_{i}^{N}\sum_{j}^{N} Q_{i,j} = H_{cost} + H_{penalty}$$

Cost

Expected return

Correlation (risk)

$$H_{cost} = [b_1 \quad b_2 \quad \cdots \quad b_N]\begin{bmatrix} \Delta P_1 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \Delta P_2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \Delta P_N \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}$$

Penalty

$$H_{penalty} = c_2\left(\left(\sum_{i}^{N} b_i\right) - N_s\right)^2 + c_3\left(\sum_{i}^{N} sgn(\Delta p_i) b_i\right)^2$$

Maximize

Sum of expected return

Minimize

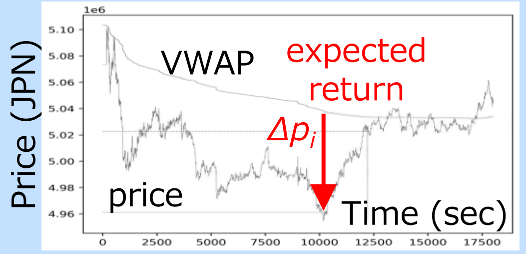
Sum of correlation factors (for diversified-portfolio)

Constrain I

Number of selected stocks, N_s

Constrain II

Delta neutral ($N_{\text{buy}} = N_{\text{sell}}$)



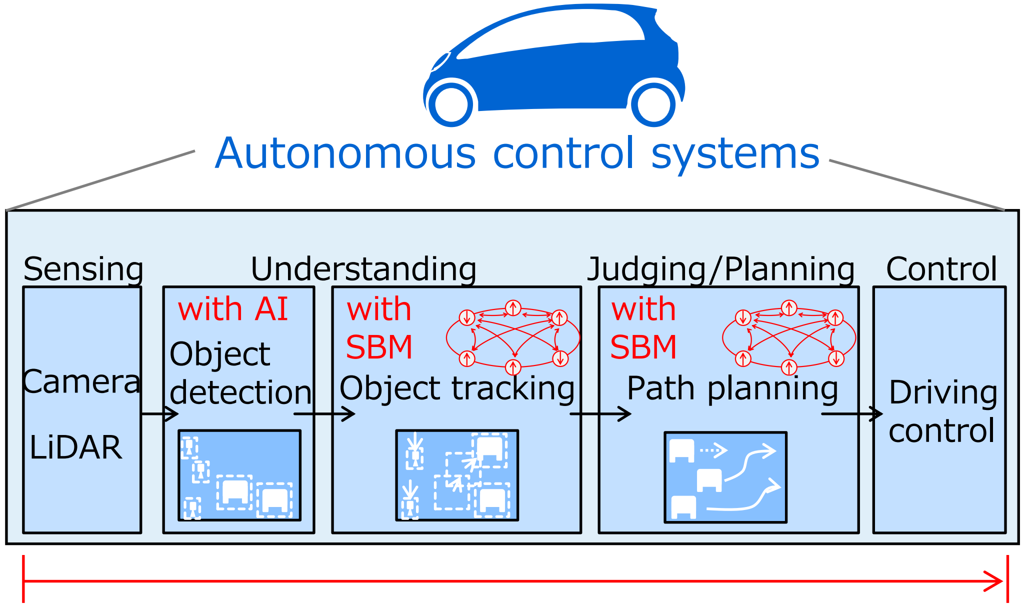
Applications in Automotive vehicle*1

*1 K. Oya et al., "Proposal and prototyping of automotive computing platform with Quantum inspired Processing Unit", Trans. of Society of Automotive Engineers of Japan 54, pp. 1216-1221 (2023). https://doi.org/10.11351/jsaeronbun.54.1216
https://trid.trb.org/View/2270193

Toward autonomous car/driving-support tech that quickly and optimally respond to surrounding situation

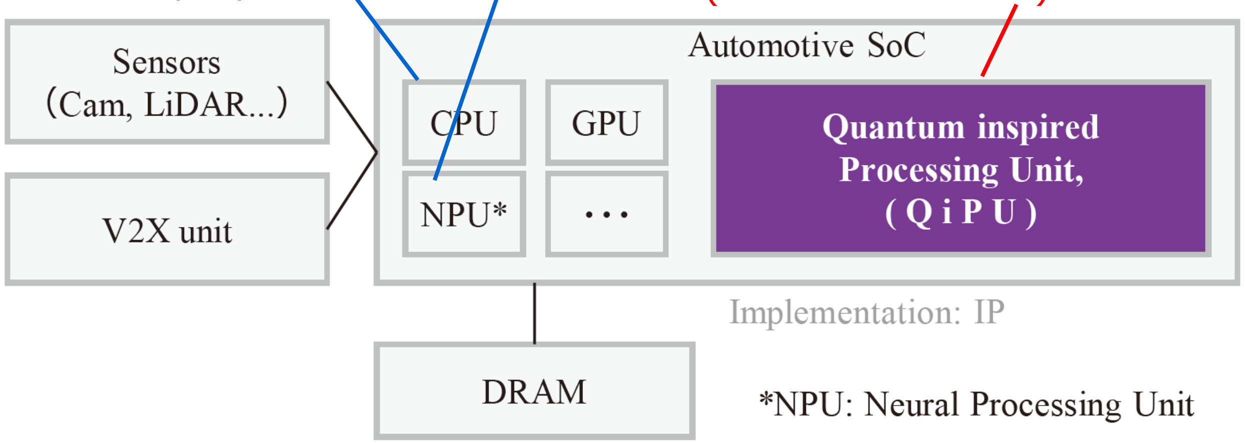
Autonomous control systems: a typical mission-critical, high-speed, real-time system

Making understanding/judgment ***more rational within critical time-constrains***



Less than 100 ms latency (periodic operation)

Quantum-inspired co-processor for discrete optimization (embeddable SBM)



*NPU: Neural Processing Unit

Fig.2 Architecture of automotive computing platform with Quantum inspired Processing Unit

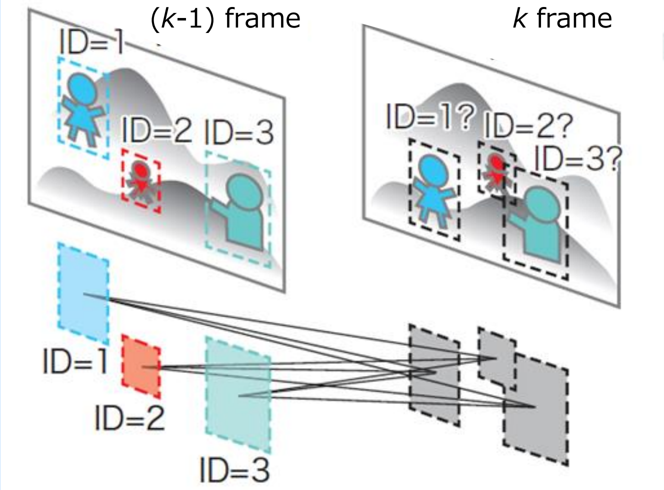
Multiple Object Tracking with **SBM**[*1]

Simulated bifurcation machine

[*1] K. Oya et al., “Proposal and prototyping of automotive computing platform with Quantum inspired Processing Unit”, Trans. of Society of Automotive Engineers of Japan 54, pp. 1216-1221 (2023). https://doi.org/10.11351/jsaeronbun.54.1216

Flexible matching functions by SBMs to realize tracking through multiple long-term occlusion events

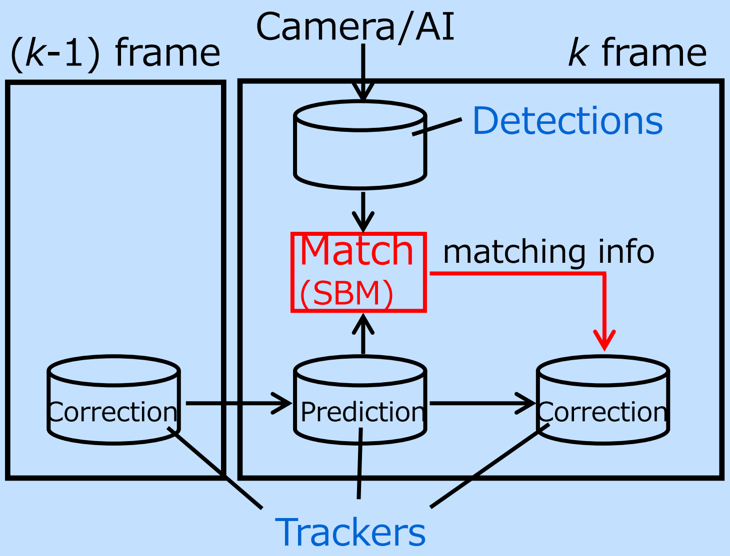
Extended MOT



To realize tracking of objects through multiple long-term occlusion events

QUBO formulation

- Matching between *Detections* and *Trackers*
- Detect the occurrences & locations of occlusion events

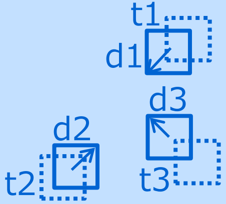


- Execute SBM ***twice*** while changing the penalty weight (***c***) for prohibiting double-match

$$H_{QUBO} = \sum_{i}^{N}\sum_{j}^{N} Q_{i,j} = \overset{\text{Maximize intersections}}{H_{cost}} + \boldsymbol{c}\overset{\text{Prohibiting double-match}}{H_{penalty}}$$

w/o occlusion

(#Detections=#Trackers)

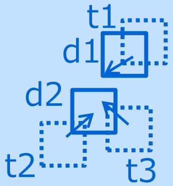


	t1	t2	t3
d1	1		
d2		1	
d3			1

one-to-one matching

with occlusions

(#Detections<#Trackers)



	t1	t2	t3
d1	1		
d2		1	1
d3			

one-to-many matching

Demonstration*1

*1 K. Oya et al., "Proposal and prototyping of automotive computing platform with Quantum inspired Processing Unit", Trans. of Society of Automotive Engineers of Japan 54, pp. 1216-1221 (2023).
https://doi.org/10.11351/jsaeronbun.54.1216

System throughput of 20 FPS and SBM-unique functions with vehicle-mountable FPGA boards

Vehicle-mountable boards

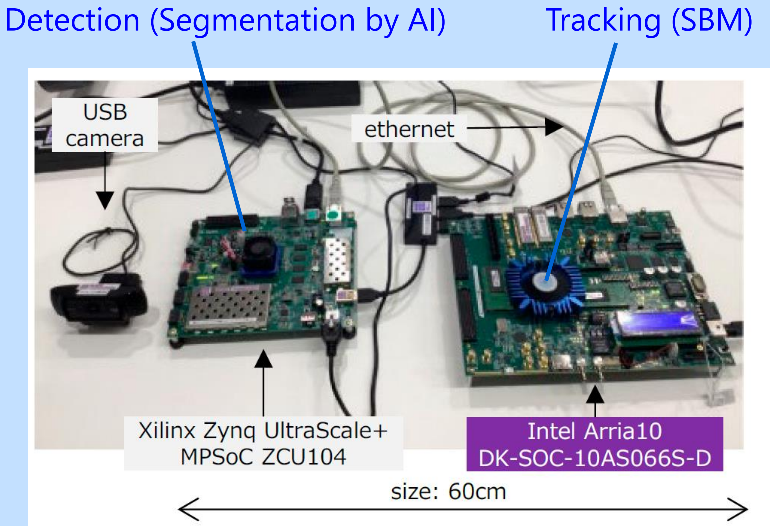


Fig.6 Appearance of prototype

SB alg. ver	Ballistic SB (bSB)
Machine size	Fully-connected 512 spins
# of MAC-PE	2,048
Time per SBM run	284 μS

Throughput (>10FPS)

Table 3. Tracking performance of SORT with matching methods of Hungarian and SBM on MOT benchmark sequences (19)

MOT algorithm	Matching	MOTA↑	HOTA↑
Original SORT	Hungarian	48.77	44.80
Original SORT	SBM	48.76	44.83

Rectangles indicates Detection result, and Vectors in Rectangles indicates Tracking result.
*This image is processed for privacy protection.

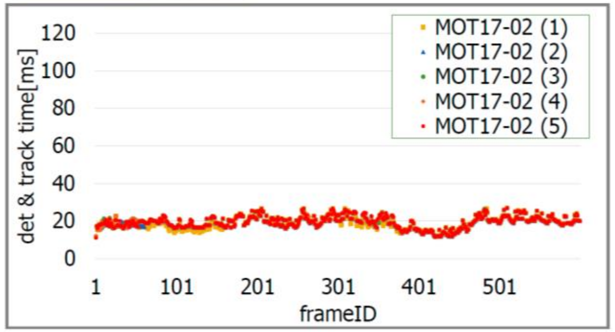


Fig.7 Evaluation result of processing time of prototype

System throughput of approx. 20 FPS

SBM-unique function

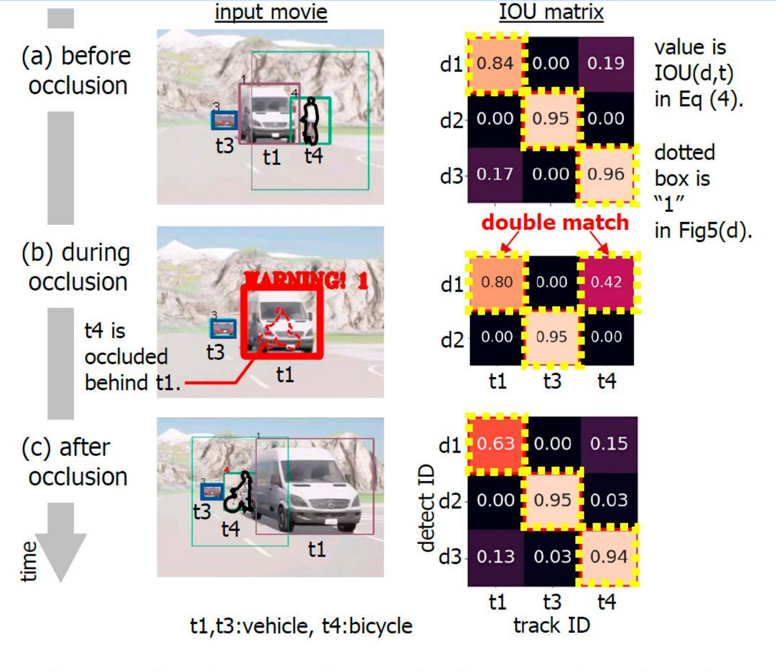


Fig.8 Function demonstration result of tracking through occlusion

Flexible matching functions for tracking through multiple long-term occlusion events

Summary

Simulated bifurcation machines

Simulated bifurcation (SB):

・quantum-inspired, highly-parallelizable algorithm for combinatorial optimization

・Can be accelerated with FPGAs/GPUs/NPUs → Very practical

High-speed real-time systems that make more rational judgments

・enabled by embeddable SBMs, for *innovative applications*

Demonstration

Financial trading

Execution capability of NP-hard optimization-based trading strategy through real-time/real-money transactions

Automotive vehicle

System throughput of 20 FPS (frames per second) and SBM-unique functions with vehicle-mountable FPGA boards

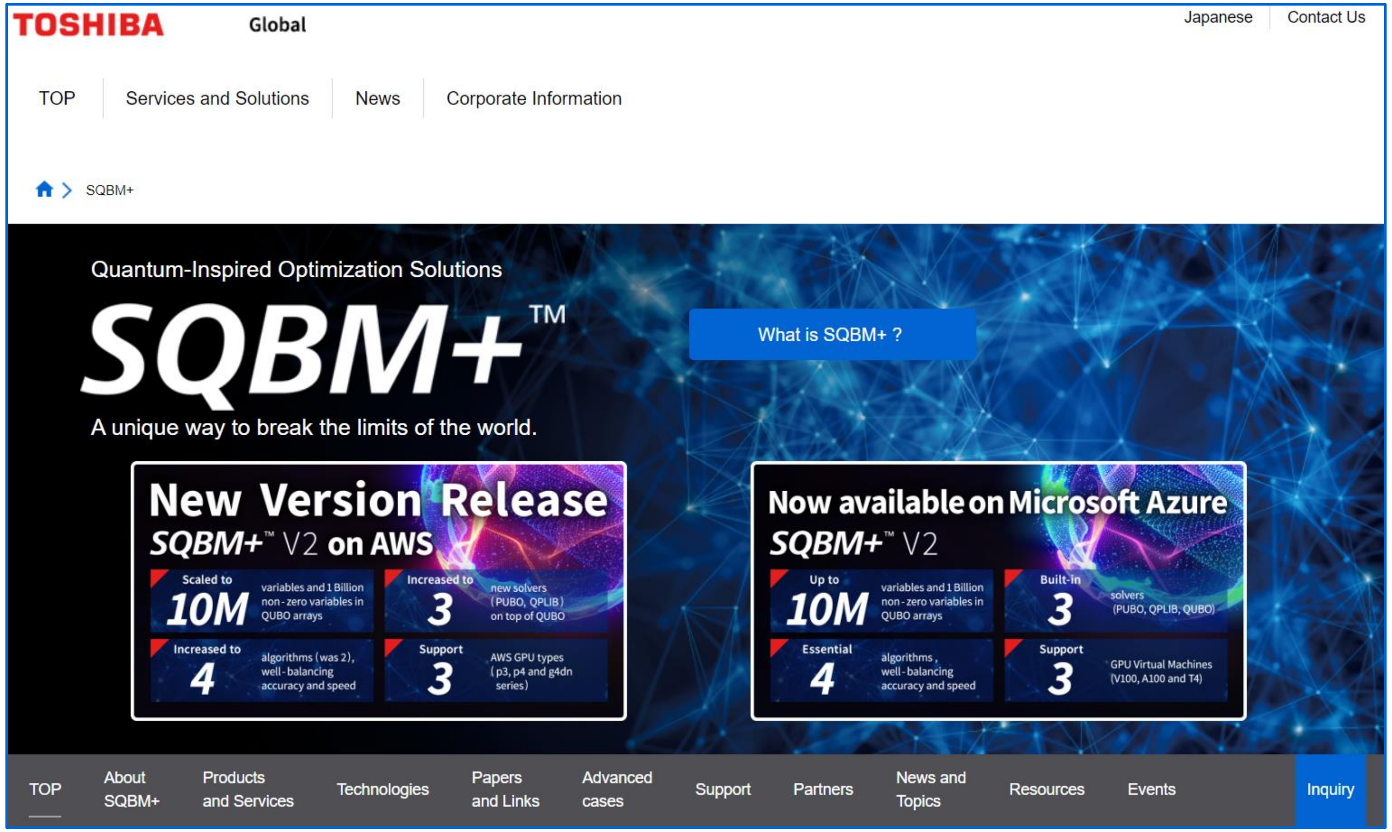
For further information

*1 Toshiba’s website “SQBM+^{TM}” https://www.global.toshiba/ww/products-solutions/ai-iot/sbm.html
SQBM+^{TM} : (Commercial service) Quantum-inspired Optimization Solution based on SBM technology

Search by “SQBM+”

List of Papers, Presentation slides, Manuals

The Slides of *this* presentation are available



Quantum-Inspired Optimization Solutions

SQBM+™

A unique way to break the limits of the world.

New Version Release
SQBM+™ V2 on AWS

Scaled to 10M variables and 1 Billion non-zero variables in QUBO arrays
Increased to 3 new solvers (PUBO, QPLIB) on top of QUBO
Increased to 4 algorithms (was 2), well-balancing accuracy and speed
Support 3 AWS GPU types (p3, p4 and g4dn series)

Now available on Microsoft Azure
SQBM+™ V2

Up to 10M variables and 1 Billion non-zero variables in QUBO arrays
Built-in 3 solvers (PUBO, QPLIB, QUBO)
Essential 4 algorithms, well-balancing accuracy and speed
Support 3 GPU Virtual Machines (V100, A100 and T4)

Appendix

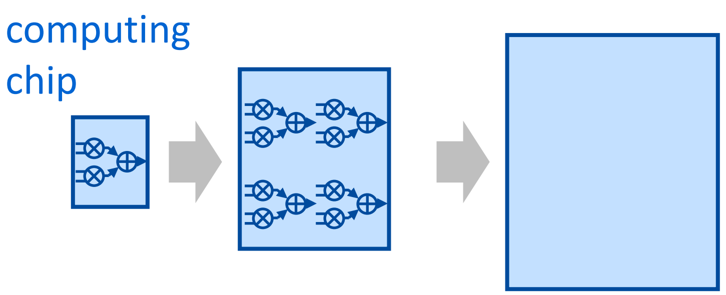
Appendix

*1 K. Tatsumura *et al.*, *Nat. Ele.* **4**, 208–217, '21
*2 (for 2nd gen) T. Kashimata *et al.*, *IEEE Access* **12**, '24

Scalability*1 (2021)

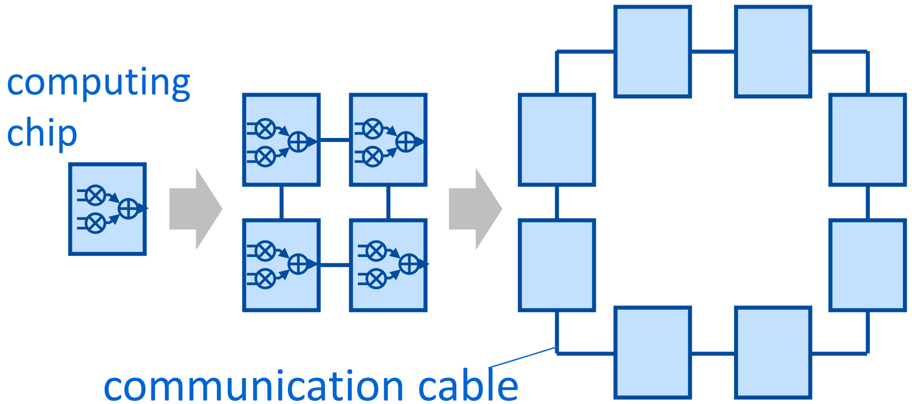
Scaling out Ising machines with full spin-to-spin connectivity

Scale-up

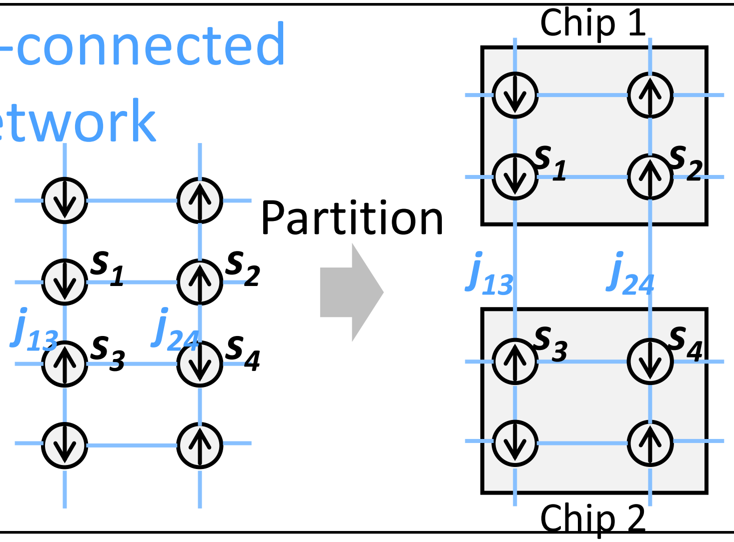


Our aim

Scale-out

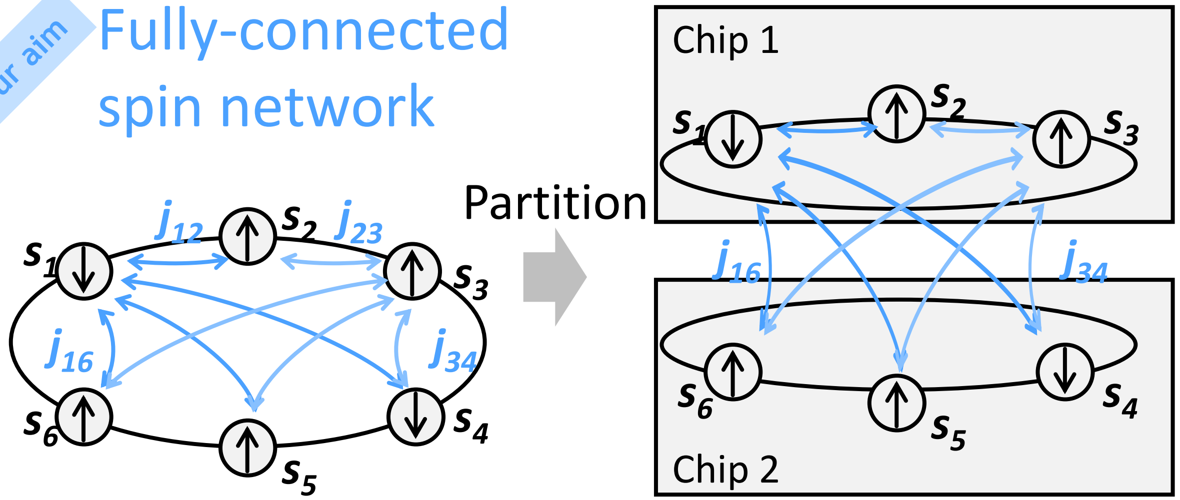


Locally-connected spin network



Our aim

Fully-connected spin network

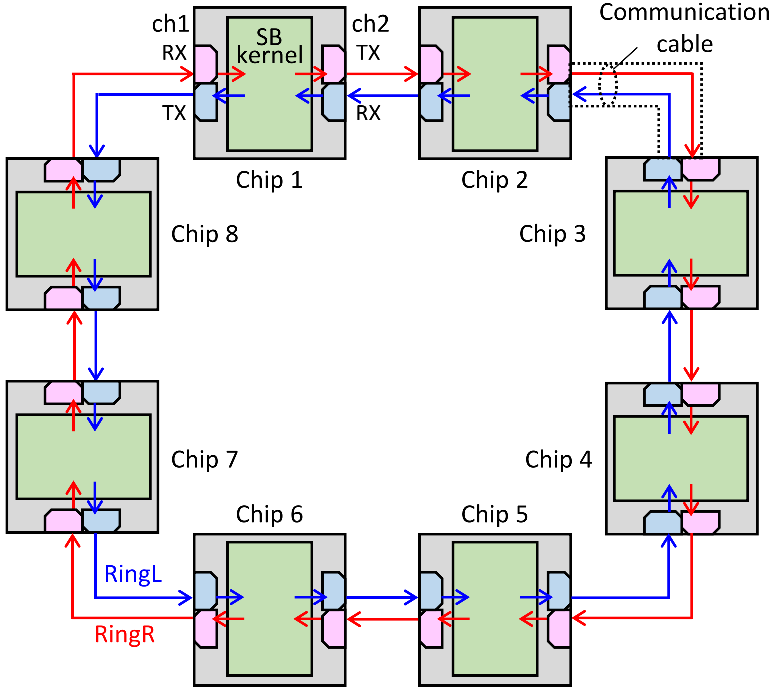


Scalability*1 (2021)

*1 K. Tatsumura *et al.*, *Nat. Ele.* **4**, 208–217, ’21
*2 (for 2nd gen) T. Kashimata *et al.*, *IEEE Access* **12**, ’24

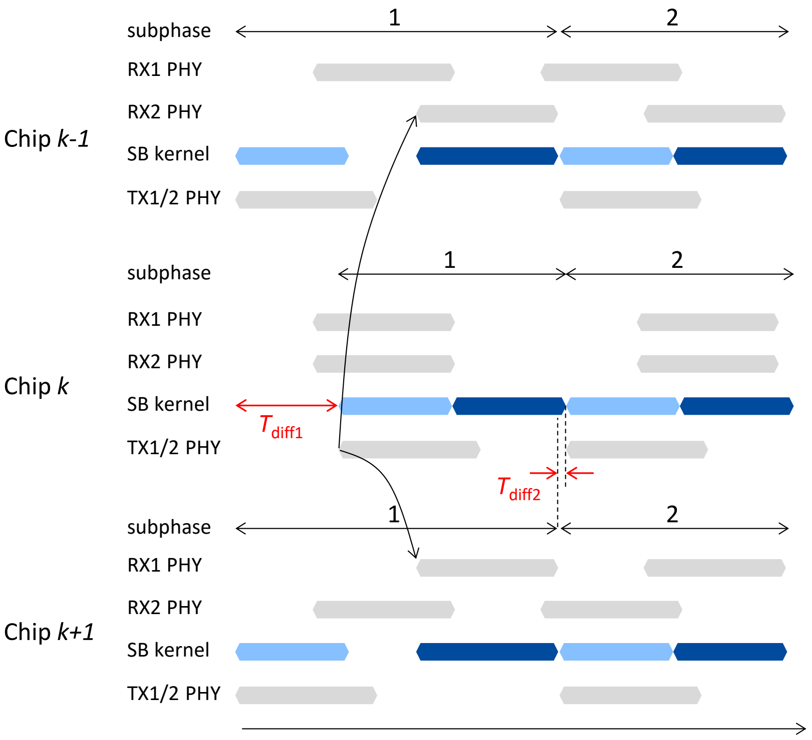
Multi-chip architecture based on partitioned SB

Bidirectional ring-network cluster without any centralized features



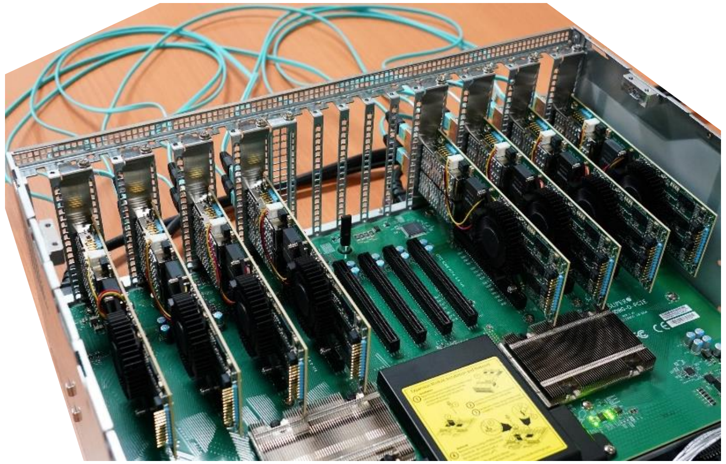
All chips are autonomous, homogeneous and symmetric

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



Time

P_{chip}=8



*1 K. Tatsumura *et al.*, *Nat. Ele.* **4**, 208–217, '21
*2 (for 2nd gen) T. Kashimata *et al.*, *IEEE Access* **12**, '24

Scalability^{*1} (2021)

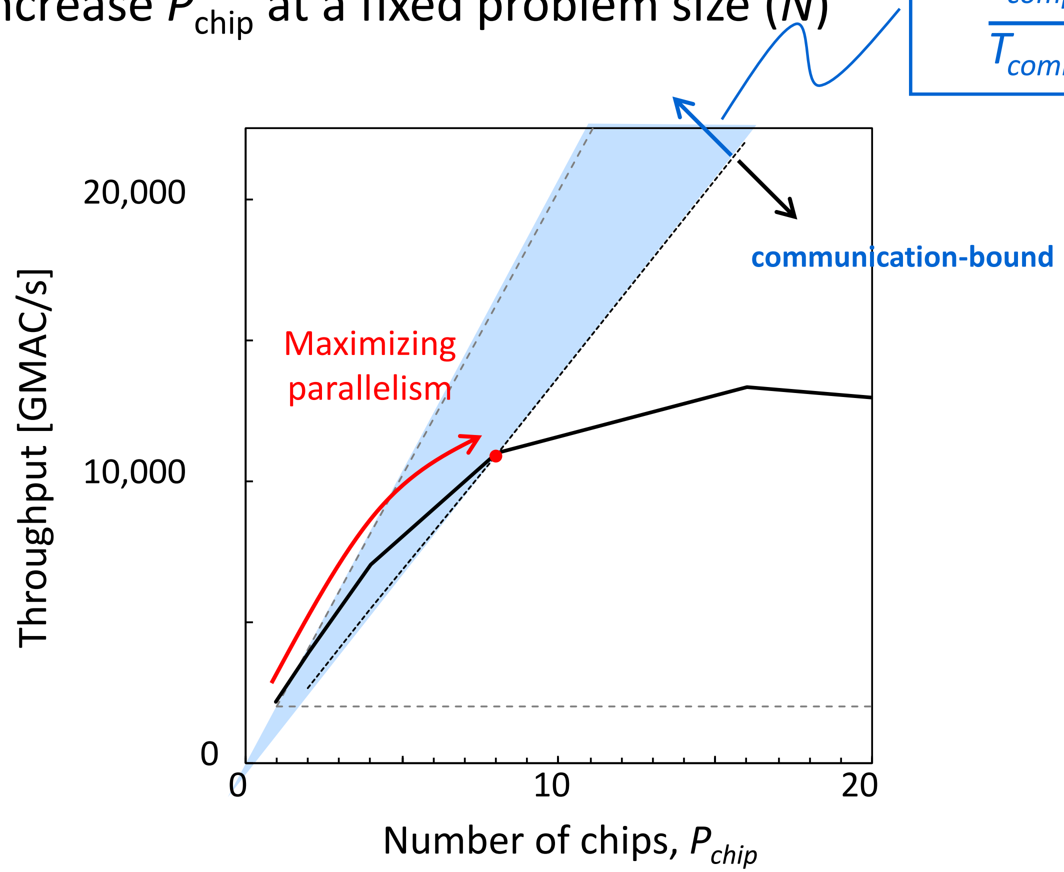
Good strong-scaling & weal-scaling charactersitics

Strong scaling

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

Computation-bound

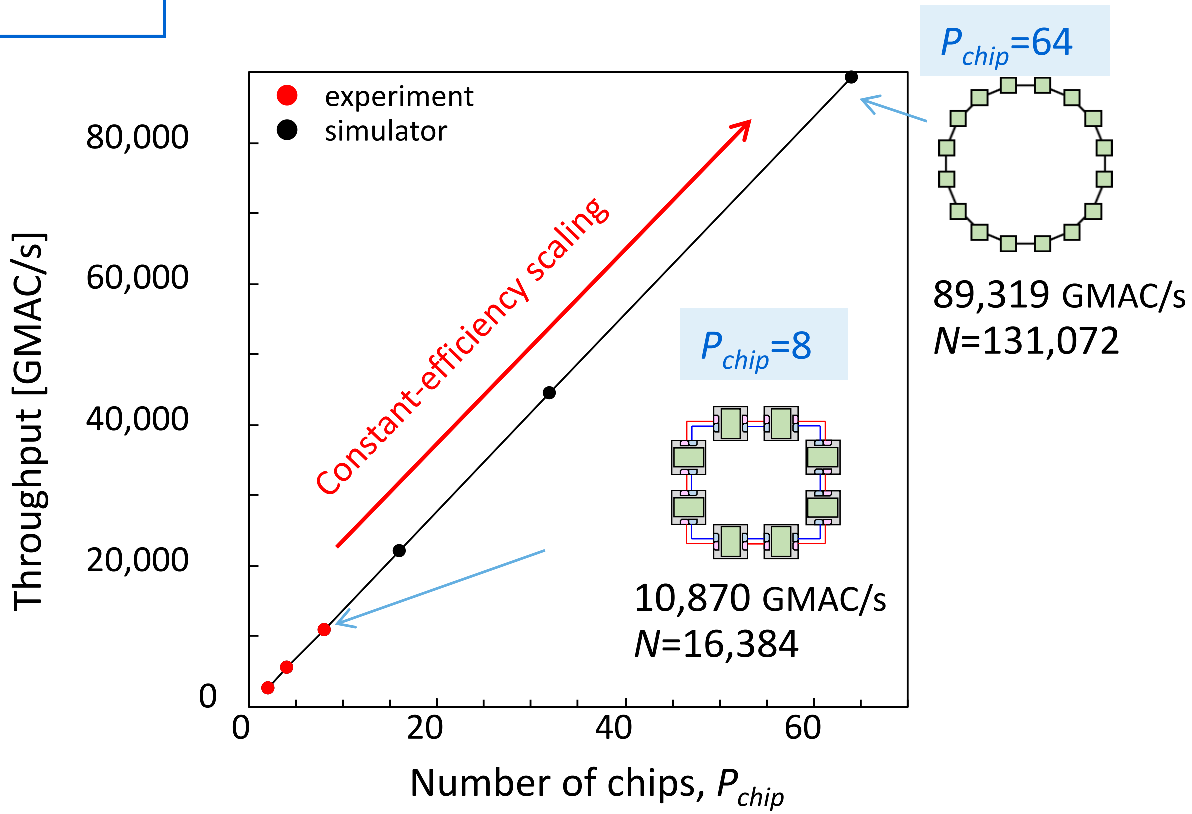
$$\frac{T_{computation}}{T_{communication}} > 1$$



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)

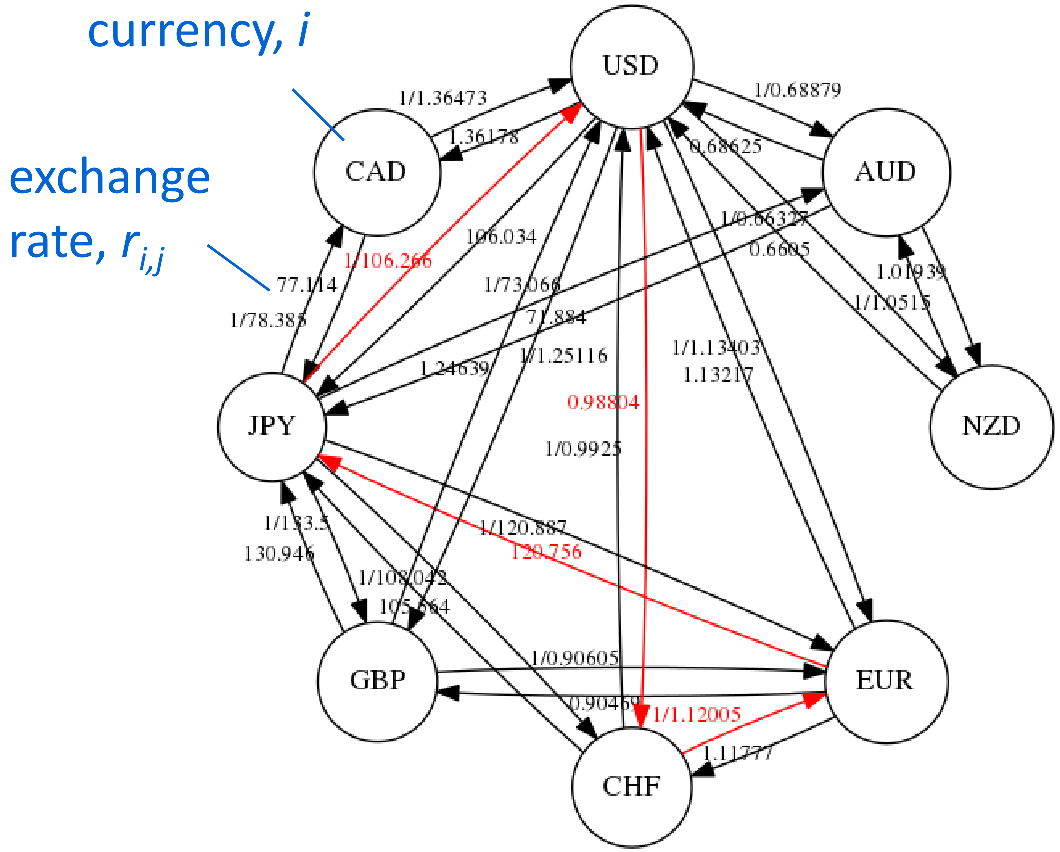
*1 K. Tatsumura et al., IEEE ISCAS., (2020) https://doi.org/10.1109/ISCAS45731.2020.9181114

High-speed real-time trading

Trading system for cross-currency arbitrage*1

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	Constraint
$Profit = \prod_{i,j \in path} r_{i,j}$	Must be a closed path

Ising (QUBO) formulation

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

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

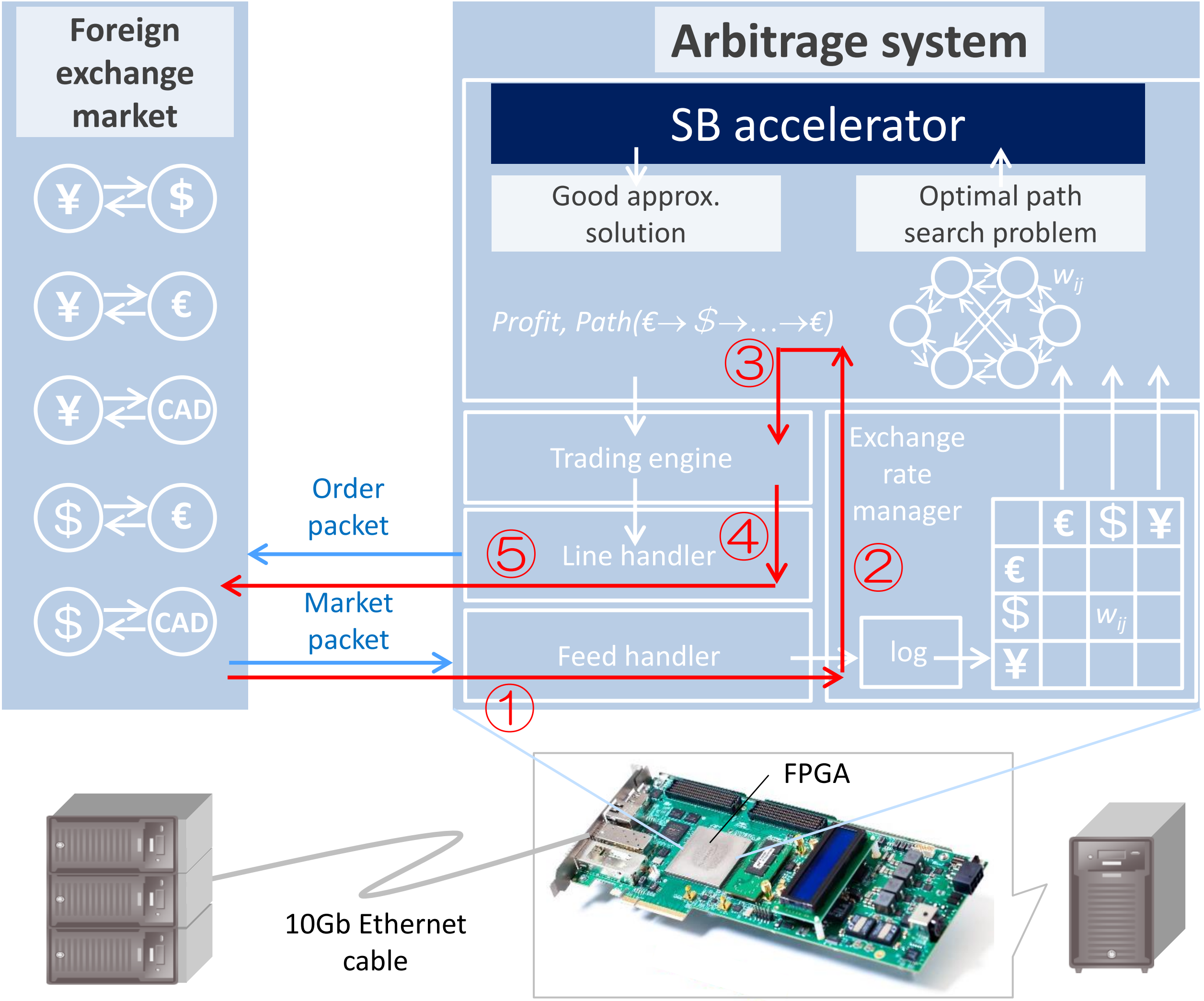
$$P = \sum_i \sum_{j \neq j'} b_{i,j}\, b_{i,j'} + \sum_j \sum_{i \neq i'} b_{i,j}\, b_{i',j} + \sum_i \left(\sum_j b_{i,j} - \sum_j b_{j,i} \right)^2 + \sum_{i,j} b_{i,j}\, b_{j,i}$$

*1 K. Tatsumura et al., IEEE ISCAS., (2020) https://doi.org/10.1109/ISCAS45731.2020.9181114

High-speed real-time trading

Trading system for cross-currency arbitrage*1

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)

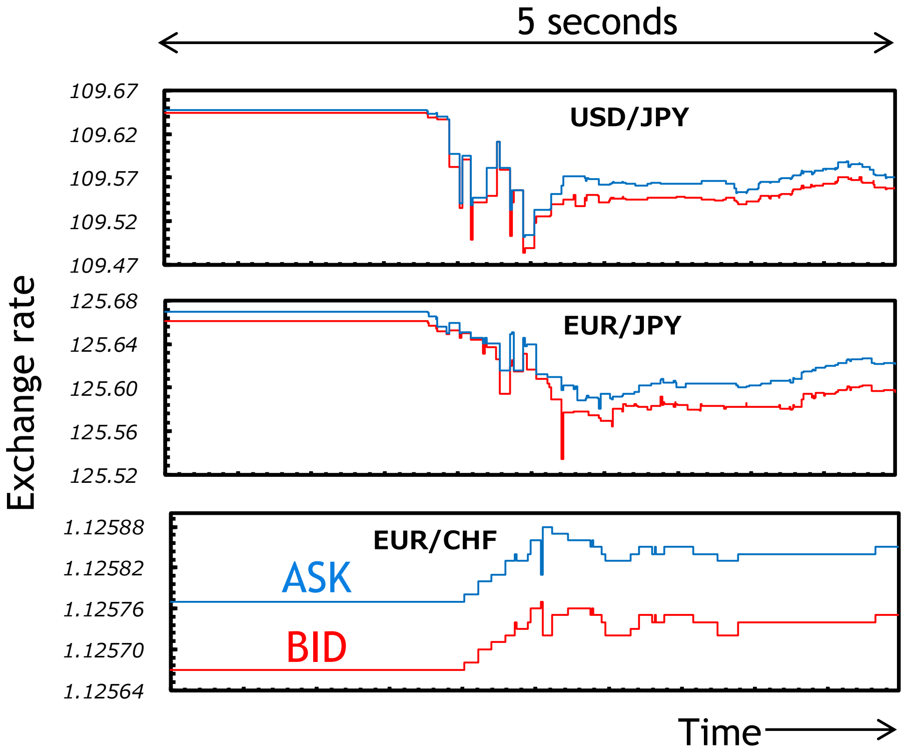
issues the order packets

High-speed real-time trading

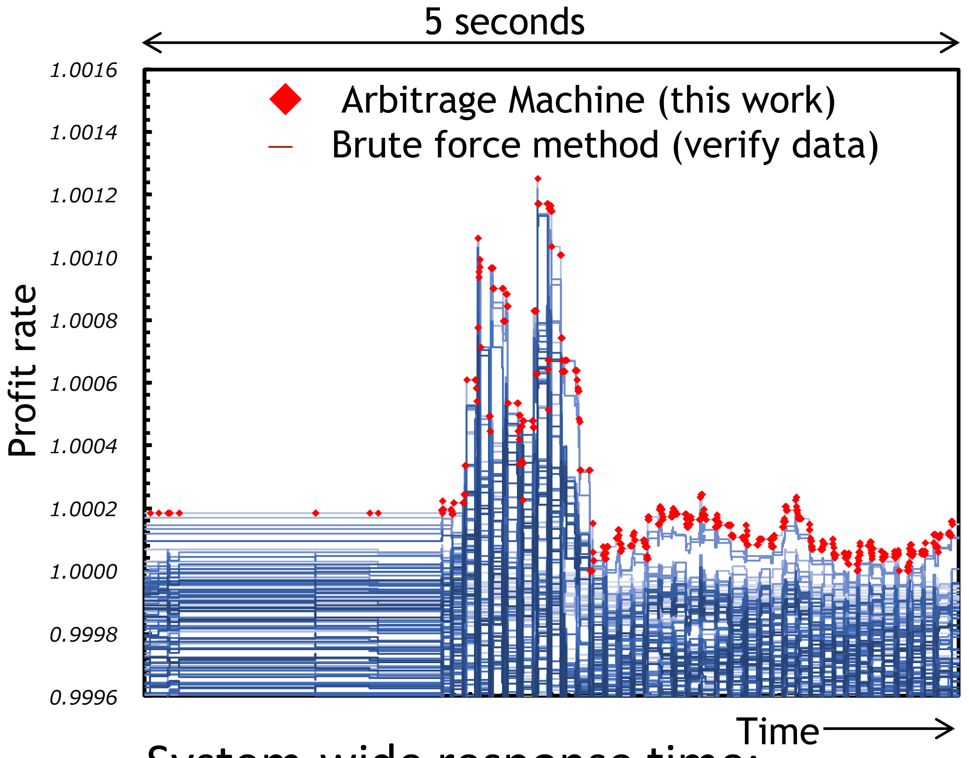
Trading system for cross-currency arbitrage*1

<$30\ \mu$s system-wide latency & 91% Top-1 probability

Exchange rates on Jan. 2nd, 2019

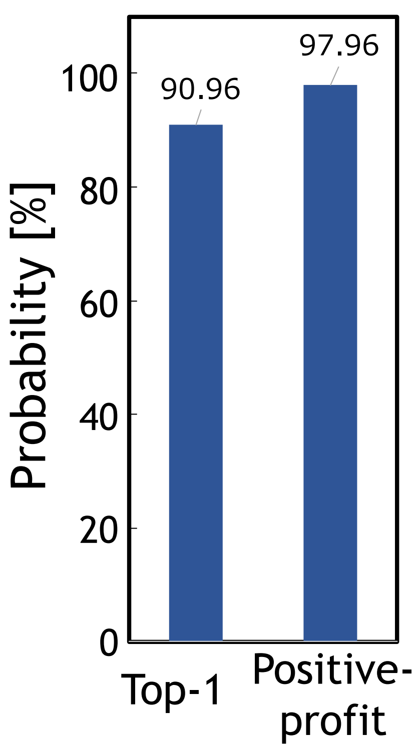


Profit rates for arbitrage paths



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

Solution accuracy



---

*1 K. Tatsumura et al., IEEE ISCAS., (2020) https://doi.org/10.1109/ISCAS45731.2020.9181114