International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies (HEART '21), June 21-23, 2021, Online, Germany

Keynote presentation

# Large-scale combinatorial optimization in real-time systems by FPGA-based accelerators for simulated bifurcation

Kosuke TATSUMURA

Chief Research Scientist, Corporate Research and Development Center

Project Manager, New Business Development Office

TOSHIBA

# Contents

01 Introduction:
Combinatorial optimization in real-time systems

02 Simulated bifurcation (SB)

03 Real-time systems that make optimal responses

04 Scaling out Ising machines

# Combinatorial optimization problem*

## Economically valuable but computationally hard

*Find a combination of discrete values, ($s_1$, $s_2$, …),
that minimizes a cost function of the discrete variables, Cost_Func($s_1$, $s_2$, …)

### Enhancing productivity

Finance

Manufacture

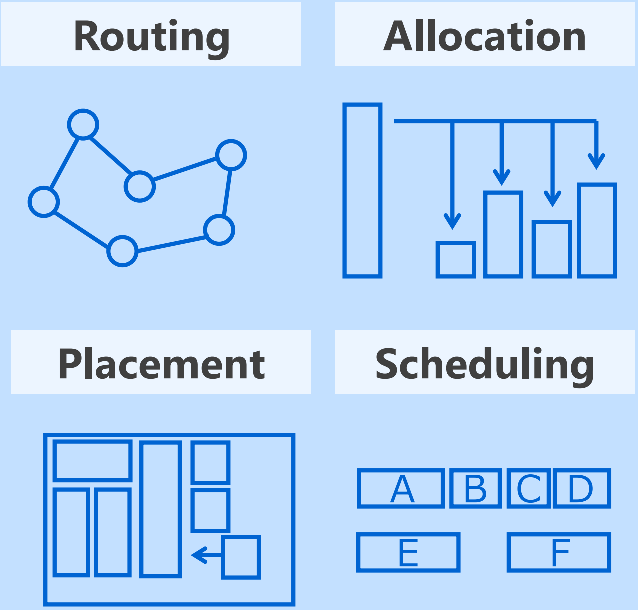
Logistics

Management

Medicine

Material

### Combinatorial optimization

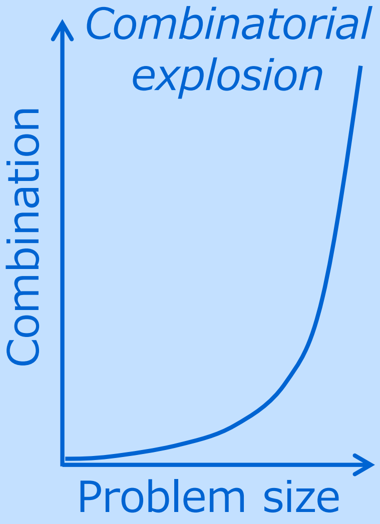


Routing

Allocation

Placement

Scheduling



*Nondeterministic polynomial time (NP)-hard*

Standard approach: Simulated annealing (SA)

# Ising machines

## Special-purpose hardware devices for quickly solving combinatorial optimization

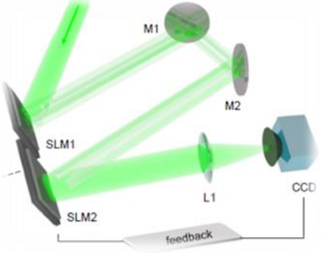
D-Wave Sys.[*1]
2011-
**Quantum Annealer**

HITACHI [*2]
2015-
**CMOS annealing machine**

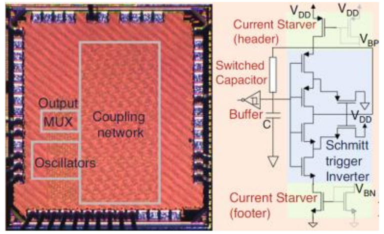
FUJITSU [*3]
2016-
**Digital annealer**

NTT/Stanford/U-Tokyo [*4]
2016-
**Coherent Ising machine (CIM)**

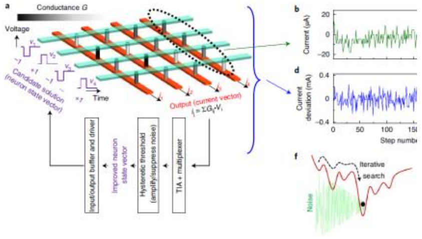
U-Roma[*5]
2019-
**Spatial-photonic Ising machine (SPIM)**



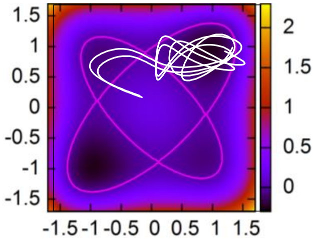
U-Virginia[*6]
2020-
**Coupled oscillators**



HP/U-Michigan[*7]
2020-
**Memristor HNN**


a
Conductance G
Voltage
Time
Candidate solution (neuron state vector)
Output (current vector)
Input/output buffer and driver
Improved neuron state vector
Hysteretic threshold (amplify/suppress noise)
TIA + multiplexer
b
c
d
Current (μA)
Current deviation (mA)
Step number
f
Iterative search
Noise

Toshiba[*8]
2019-
**Simulated bifurcation**



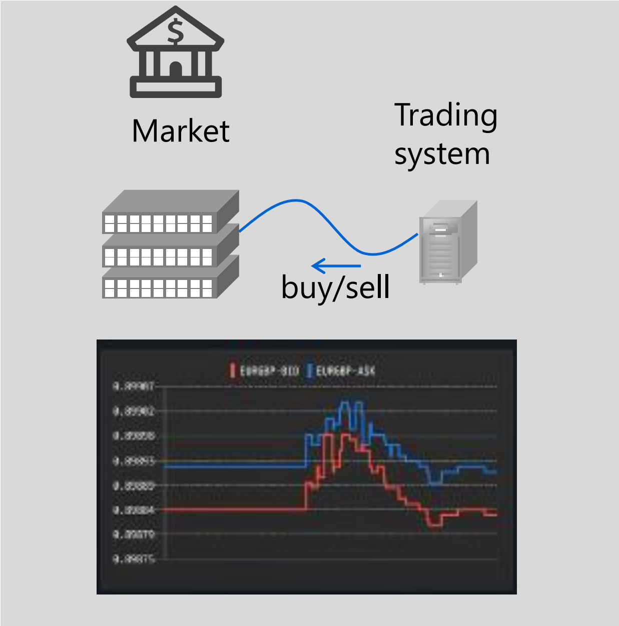
[*1] https://www.dwavesys.com/d-wave-two-system
[*2] https://www.hitachi.co.jp/New/cnews/month/2019/02/0219.html
[*3] https://www.fujitsu.com/global/about/resources/news/press-releases/2018/0515-01.html
[*4] https://www.ntt.co.jp/news2017/1711e/171120a.html
[*5] D. Pierangeli, et al., Phys. Rev. Lett. **122**, 213902 (2019).
[*6] A. Mallick, et al., Nature Communications **11**, 4689 (2020).
[*7] F. Cai, et al., Nature Electronics **3**, 409 (2020).
[*8] https://www.global.toshiba/ww/technology/corporate/rdc/rd/topics/21/2103-03.html

# For real-time systems

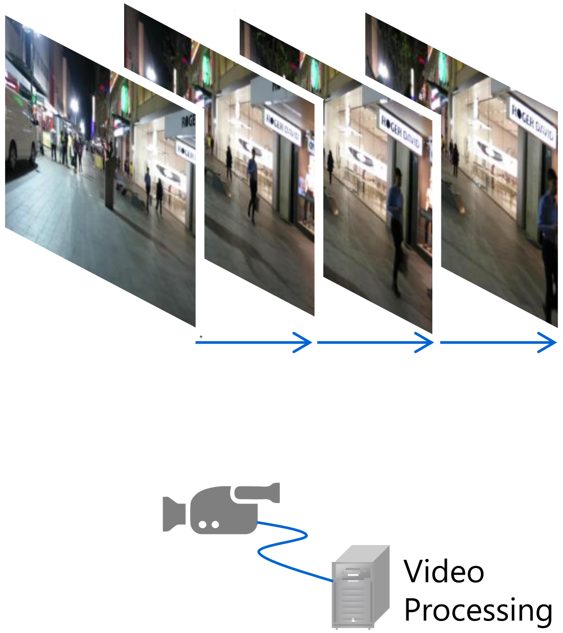
*1: Data source: Integral Development Corporation, [Under the permission from Integral Dev. Corp.]
*2: Data source: https://motchallenge.net/, [Under CC BY-NC-SA 3.0 license for research purpose]

## Ising machines may allow those systems to choose the optimal response from among all the candidates -*Rational* decision-making in real-time systems-

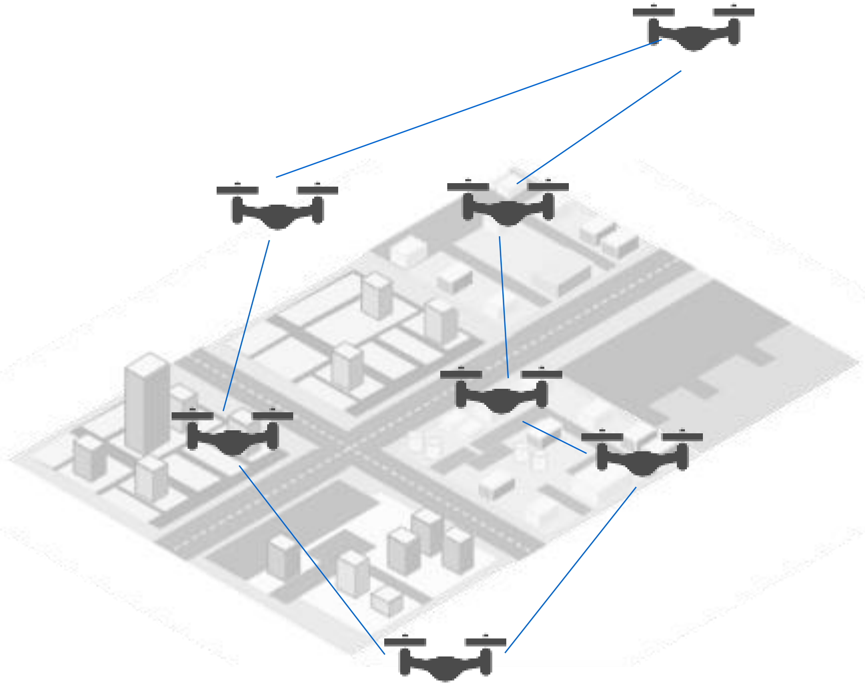
### Financial transaction system*1



### Video processing*2



### Swarm robotics



Real-time systems:

-respond to rapid-changing situations with specified time constrains

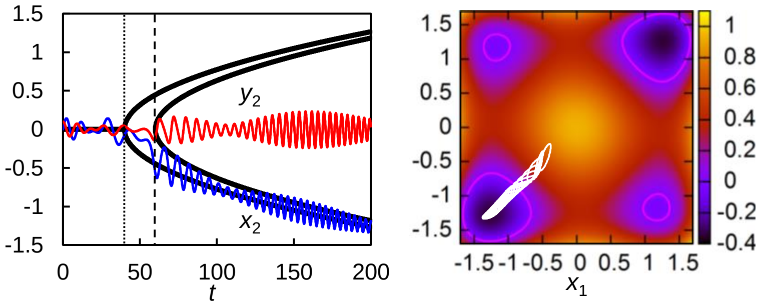
-decision making based on a simple conditional judgement (conventional)

# Simulated bifurcation (SB)

H. Goto, K. Tatsumura, A. R. Dixon, *Sci. Adv.* **5**, eaau0823, ’19
The story of the birth of SB, https://www.toshiba-clip.com/en/detail/p=76

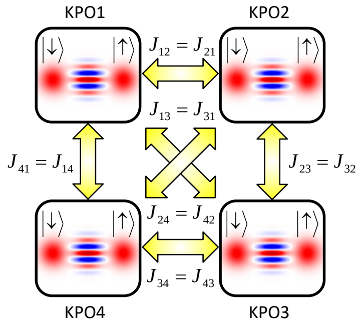
## A quantum-inspired algorithm for combinatorial optimization having *Plentiful Parallelism*

### Simulated Bifurcation (SB)



### Quantum Bifurcation (QB) machine

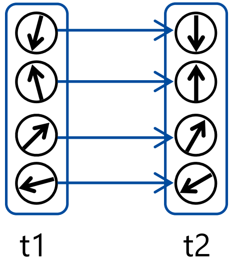
a quantum adiabatic optimization method
[H. Goto, *Sci. Rep* **6**, 21686, ’16]



Derived as the classical counterpart

### *Plentiful parallelism*

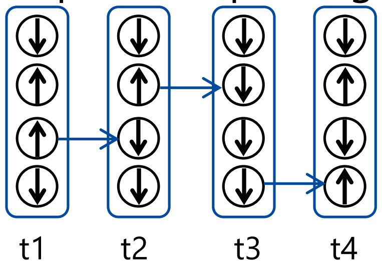
Parallel updating



vs.

### Simulated Annealing (SA)

Sequential updating



***→Substantial speedup by massively parallel processing***

# This talk: Proposal & demonstration

**Real-time systems that make *optimal* responses enabled by FPGA-based SB accelerators**

**Simulated bifurcation (SB) & FPGA-based accelerators for SB**

**Real-time systems that make optimal responses:**
**An ultra-fast financial transaction machine**

**Scale-out architecture of Ising machines with full connectivity using the high parallelism of SB**

# Contents

01 Introduction:
Combinatorial optimization in real-time systems

02 Simulated bifurcation (SB)

03 Real-time systems that make optimal responses

04 Scaling out Ising machines

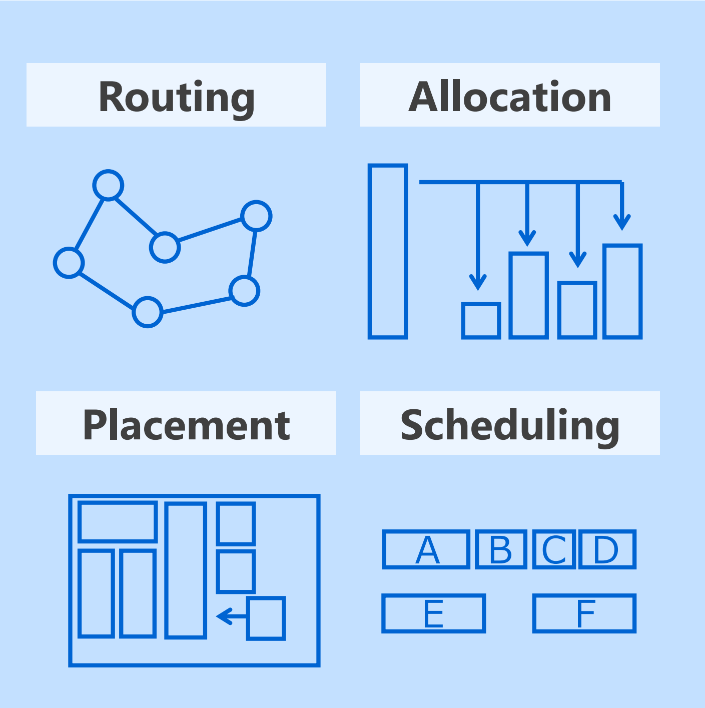
# Ising problem & Ising machine

[Andrew Lucas, *Frontiers in Physics* **2**, 5 (2014)] provides Ising formulations for many NP-complete/hard problems, including all of Karp's 21 NP-complete problems

## 1. Any NP problems can be converted to the Ising problem with P-time
## 2. Ising machine searches for the ground-state of Ising spin model

### Combinatorial optimization

*NP-hard*



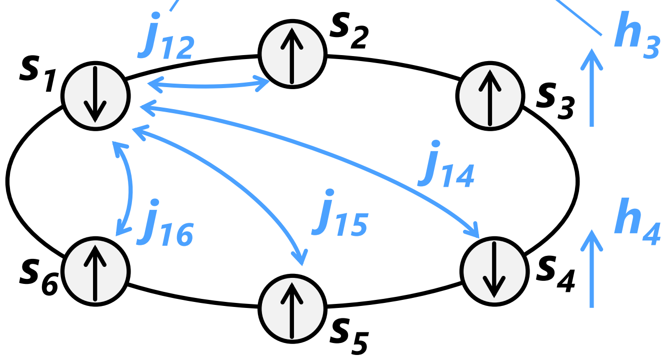
*converted to*

### Ising problem

*NP-hard &*
*NP-complete*

$$J=\begin{bmatrix} 0 & j_{12} & j_{13} \\ j_{21} & 0 & j_{23} \\ j_{31} & j_{32} & 0 \end{bmatrix} \qquad h=\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}$$

*coupling* *bias*



*spin: binary variable*

*input*

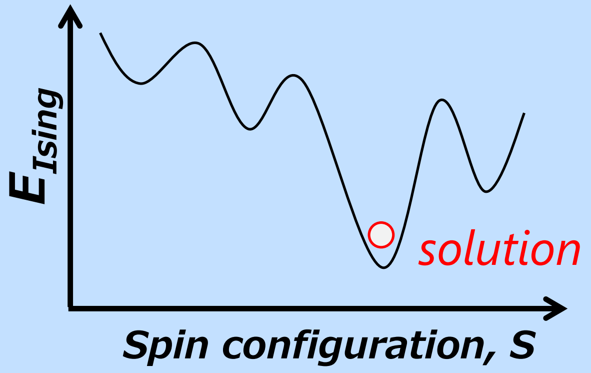
### Ising machine

*Special-purpose*

search for ground-state ***s*** minimizing ***E***

***Ising energy***

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



# The story of the birth of SB

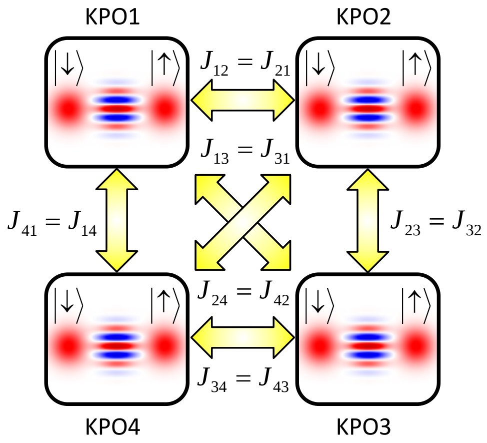
Toshiba Chip 2020/02/12
https://www.toshiba-clip.com/en/detail/p=76

## Simulated bifurcation was “discovered” from a quantum computer

### Quantum Bifurcation (QB) machine [H. Goto, Sci. Rep. 2016]

Based on
the quantum adiabatic theorem

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



### Classical Bifurcation (CB) machine [H. Goto, Sci. Rep. 2016]

Classicization

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

↓ Algorithmic twist
for speed-up

### Simulated Bifurcation (SB) algorithm

[Goto, Tatsumura, Dixon, Science Advances **5**, eaav2372 (2019)]

**There was No guarantee. We found that CB works very well and has an outstanding characteristics, i.e. parallelism.**

Science Advances

"Combinatorial optimization by simulating adiabatic bifurcations in nonlinear Hamiltonian systems"

H. Goto et al.
eaav2372

April 19, 2019

# How it works: Simulated Bifurcation (SB)

## *N*-body system dynamically searches for a good solution

### Movement of the system in *N*-dimensional space

Example: ***N*=2**

a single local minimum

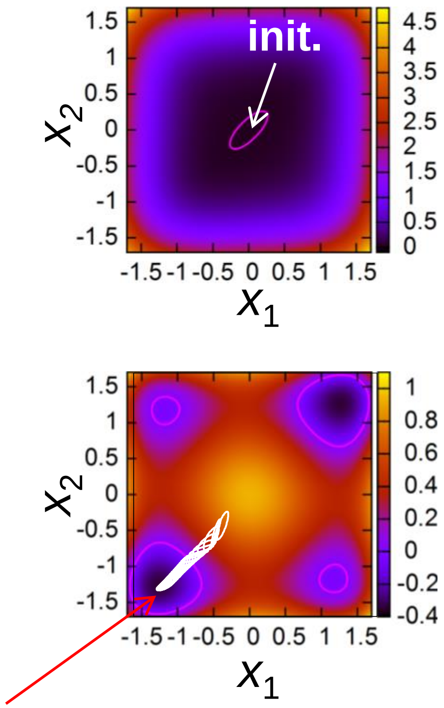
**Bifurcation**

(adiabatic process)

multiple local minima (target cost function)

**best solution**

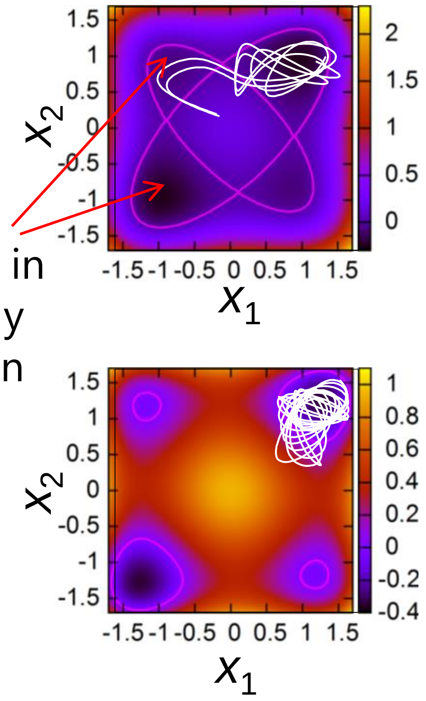
(-1,-1)



**Adiabatic Search**

chase one of the minima

Multiple minima in the energetically allowable region



**Ergodic Search**

find better one with higher probability

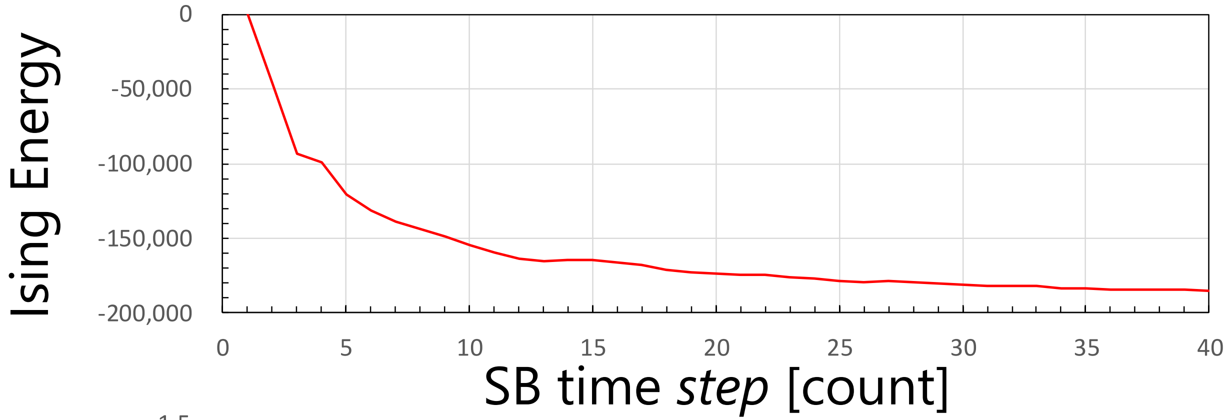
**If *N* is large**,

find a global minimum (or a local minimum close to the grand-state) from among $2^N$ local minima

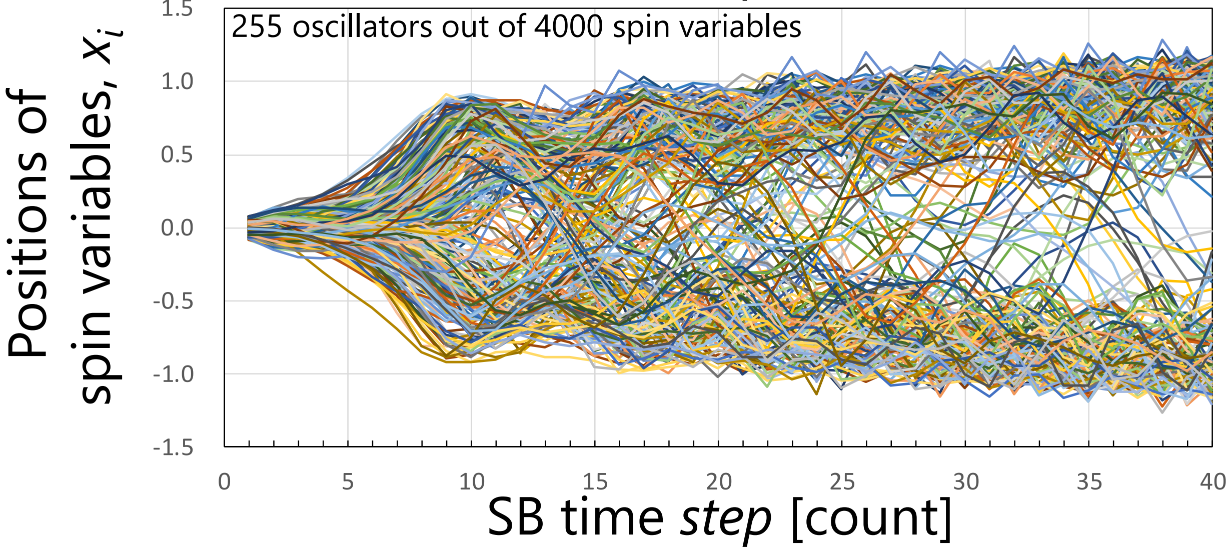
# How it works: Simulated Bifurcation (SB)

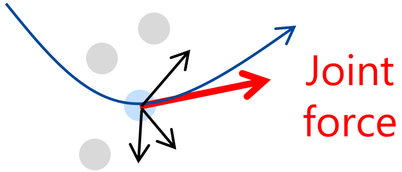
## Time evolution of *N*-body system

### Movements of *N*(=4000) spin-variables as a function of time



**better solution**

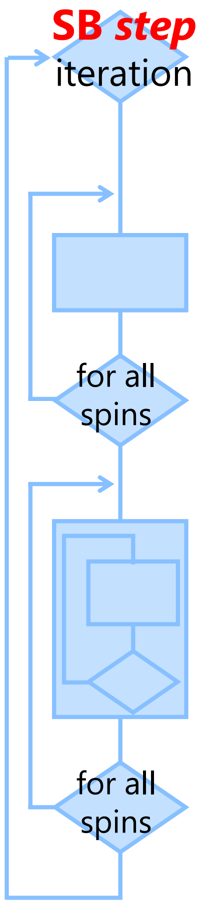




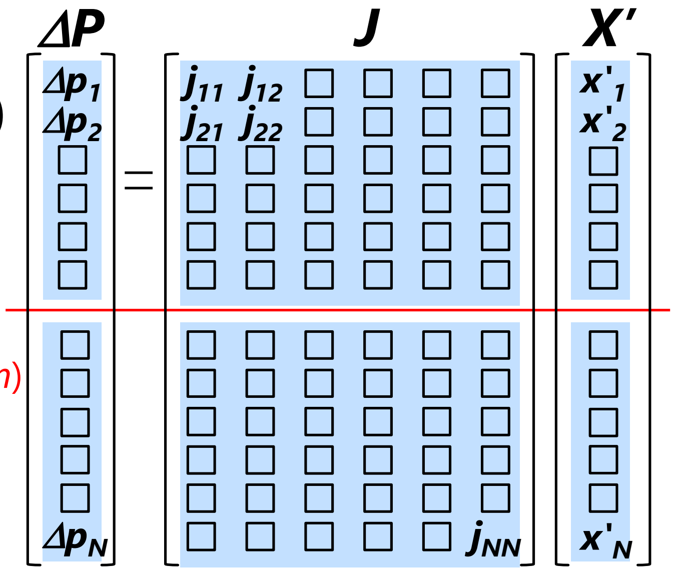
**Many-body interactions** depending on all the other spins

# Algorithm of SB and it's parallelism

## SB *step*: spin state at $t_{n+1}$ ← the previous state at $t_n$



Matrix-Vector Multiplication (**MM**)



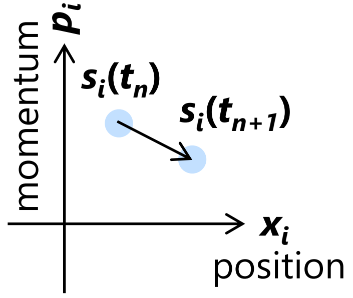
**Many-body interaction**

$$\Delta p_i = \sum j_{ij} x'_i$$

*Horizontal partition (Block parallelism)*

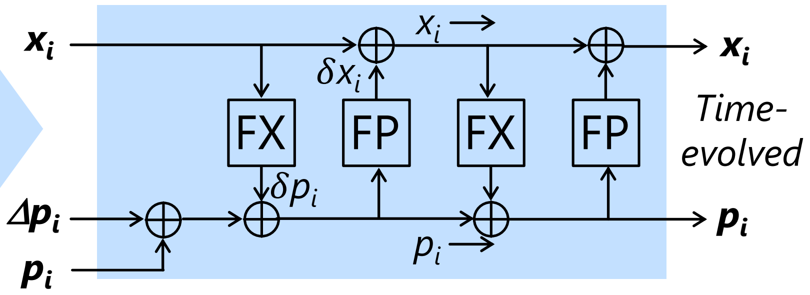
**Time evolution (TE)**

$$(x_i, p_i)_{@tn+1} = \mathrm{TE}(x_i, p_i, \Delta p_i)_{@tn}$$



*Sequential but independent for each spin*

TE pipeline

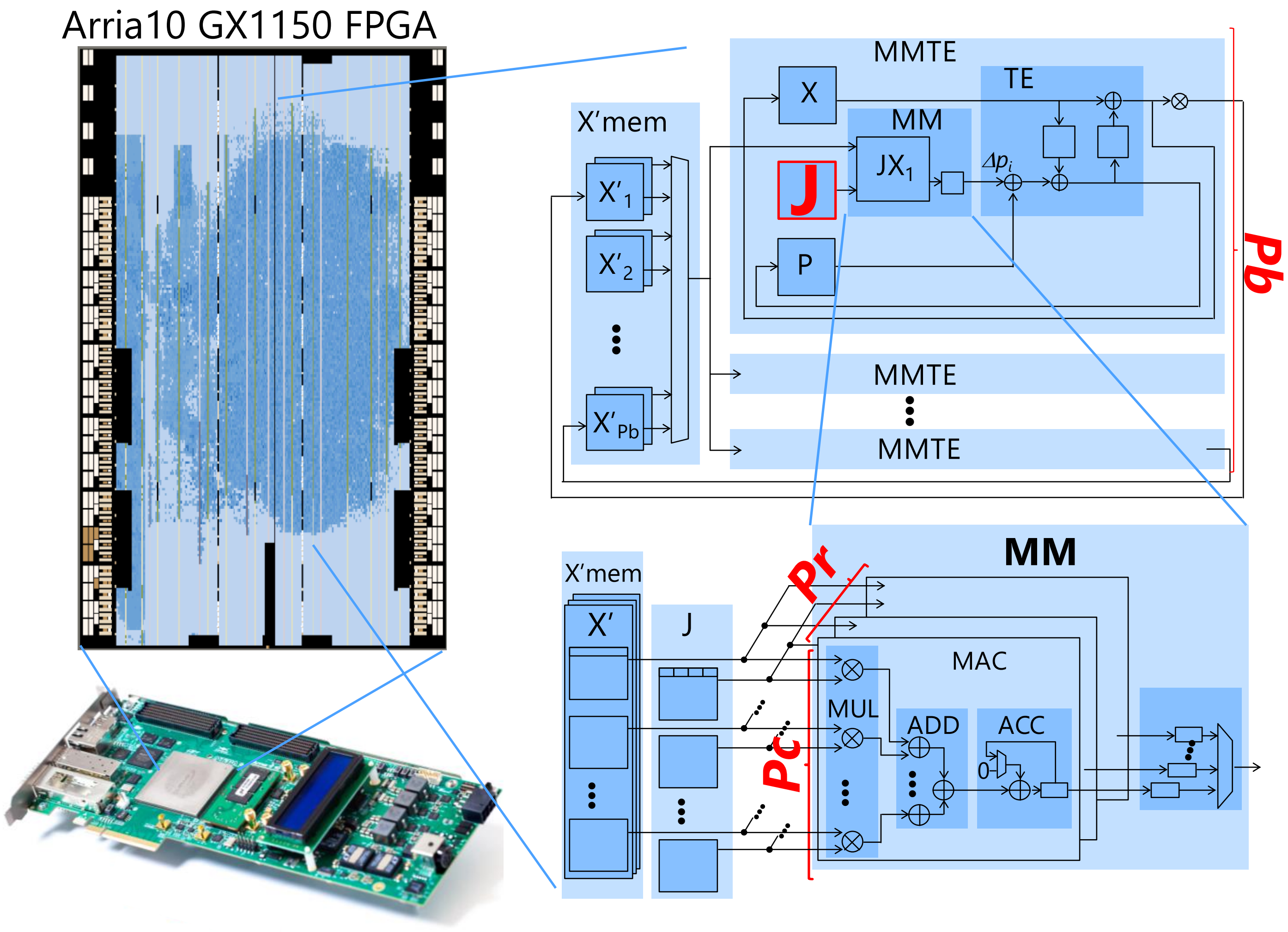


Top-level parallelism: **Simultaneous update of *N* spins is possible**

[K. Tatsumura, A. R. Dixon, H. Goto, *FPL19*]

# FPGA-based accelerator for simulated bifurcation

## Large-scale, massively parallel, and high utilization



| Problem | complete-graph MAX-CUT |
|---|---|
| Machine size | 4,096-size<br>(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** |

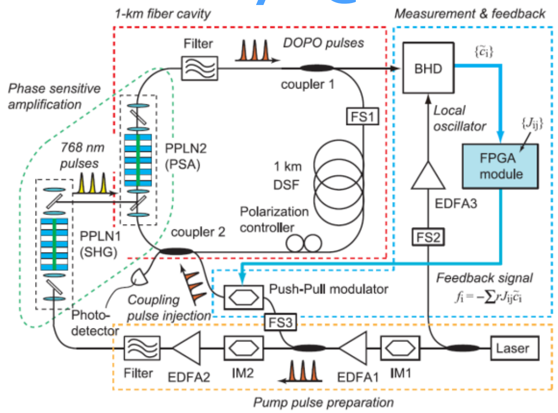
[K. Tatsumura, A. R. Dixon, H. Goto, *FPL19*]

# Evaluation: FPGA-SB vs. CIM

## 14X faster, 288X more energy efficient than CIM

**Coherent Ising Machine**

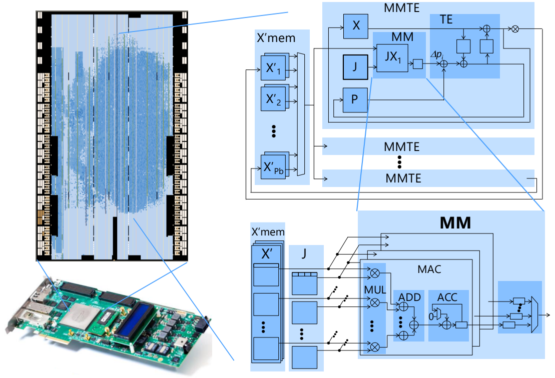
**800 GMAC/s @ 1000 W**



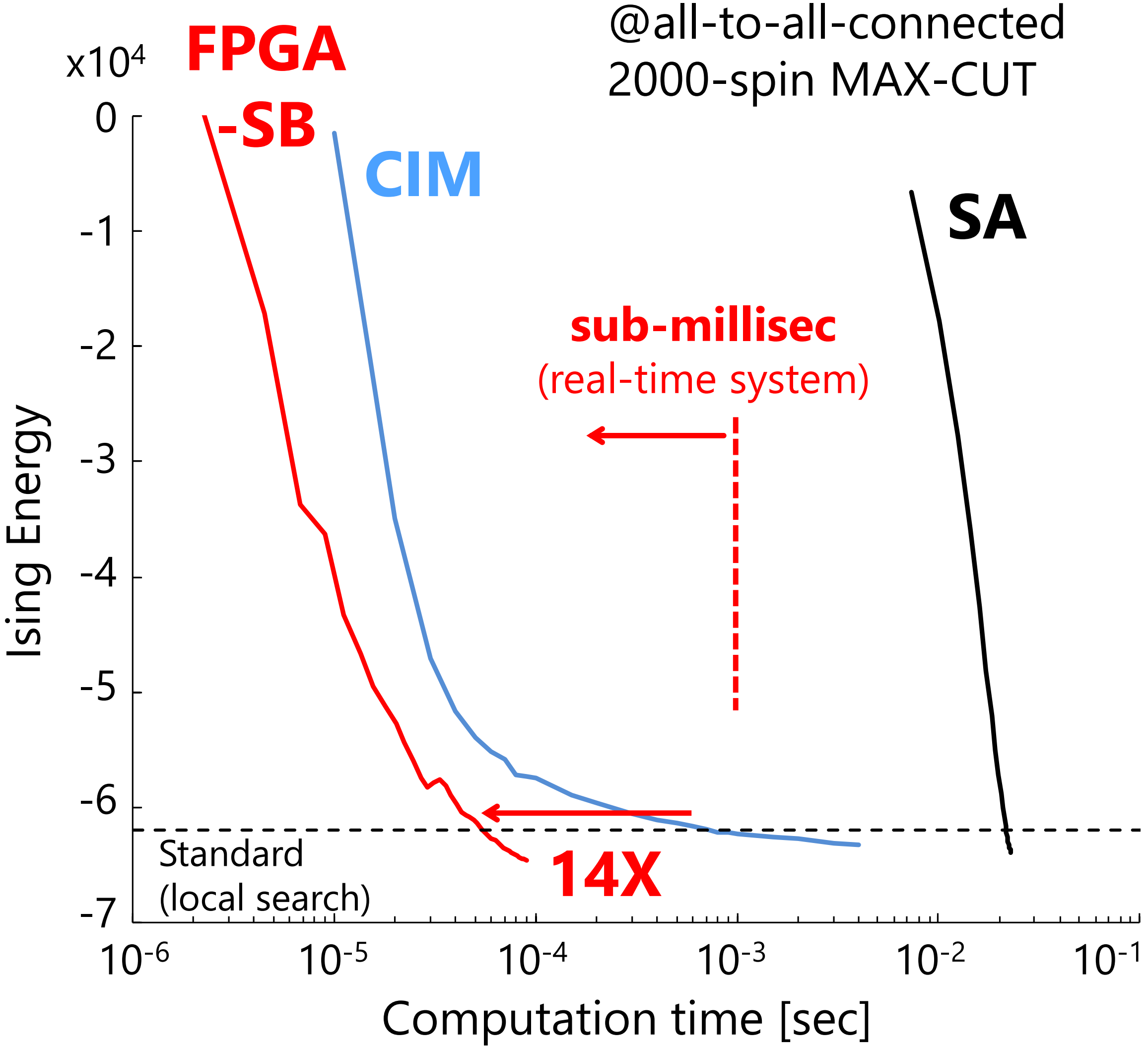
[T. Inagaki, Science 354, 603, '16]

**FPGA-SB**

**1,873 GMAC/s @ 49 W**



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

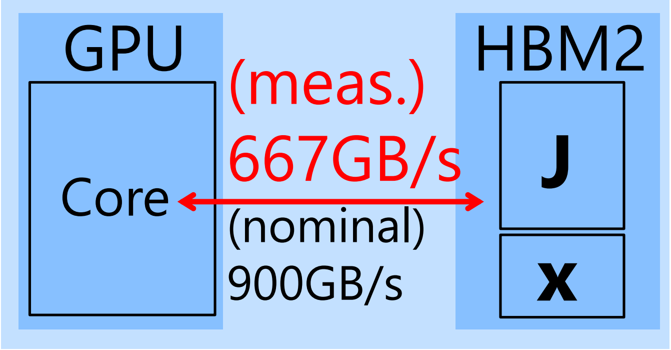


[K. Tatsumura, A. R. Dixon, H. Goto, *FPL19*]

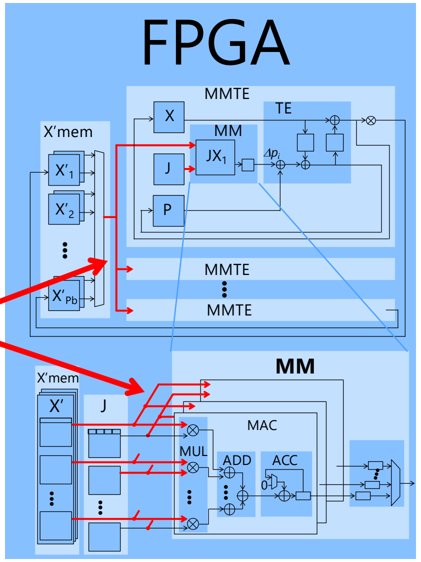
# Evaluation: FPGA-SB vs. GPU-SB

**FPGA is computation-bound, GPU memory-bound**
**-11X faster, 26X more energy efficient than GPU-SB**

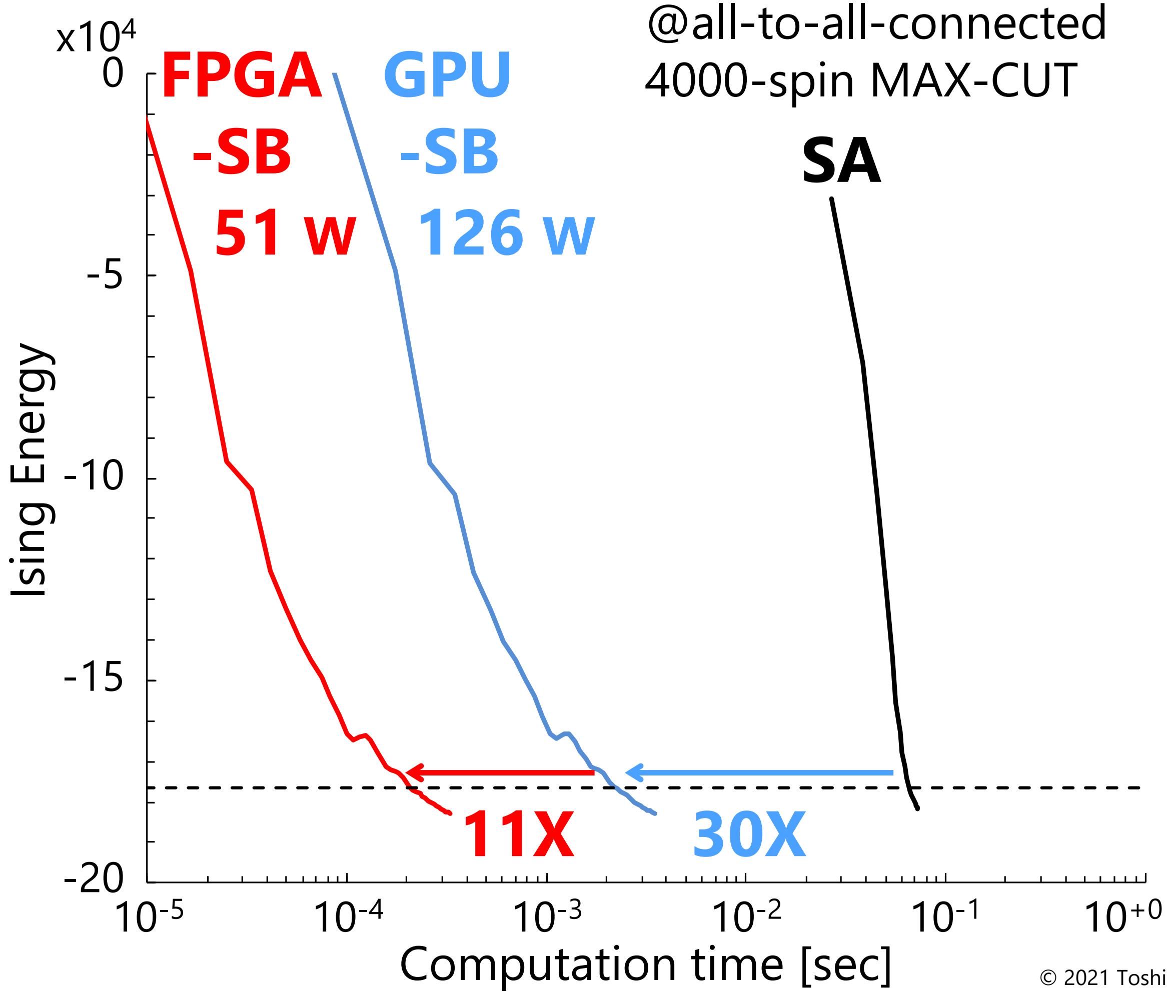
## GPU-SB (Tesla V100)



## FPGA-SB



peak rate
4,678GB/s



@all-to-all-connected
4000-spin MAX-CUT

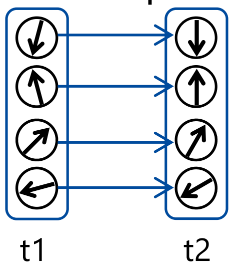
# Why SB is faster than SA

For the detailed discussion, see
[K. Tatsumura, *Nature Electronics* **4**, 208(2021)]

## Maximum computation parallelism: $N^2$ for SB, $N$ for SA
## SB can be further accelerated by more parallel computing

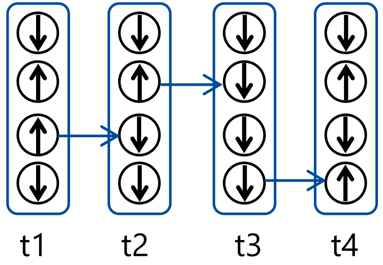
### Simulated bifurcation (SB)

Parallel updating

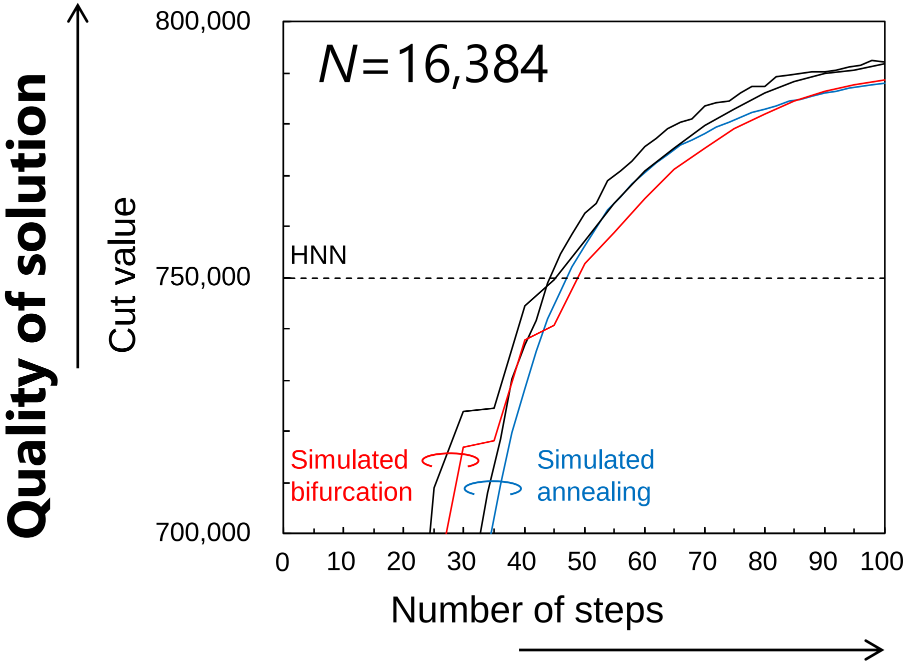


### Simulated Annealing (SA)

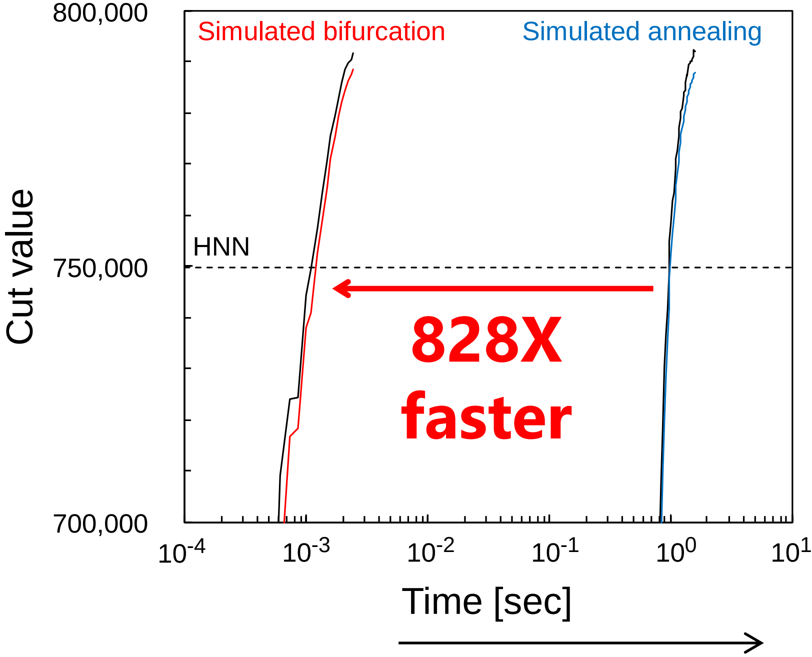
Sequential updating



our best-optimized implementation



**The amount of computation**
(The amount of pair interactions evaluated)



**Computation time**

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

# 2nd-gen simulated bifurcation technology (Feb, ‘21)

## Incorporated quasi-quantum effects, got further faster, larger & higher-quality

Science Advances

AAAS open-access journal

“High-performance combinatorial optimization based on classical mechanics”

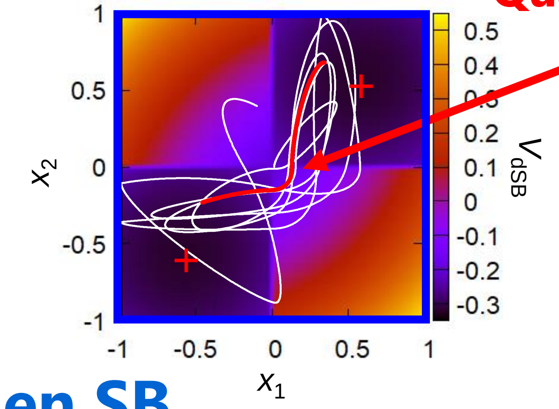
H. Goto et al

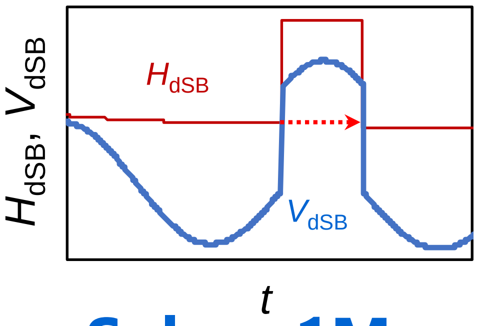
eabe7953

February 03, 2021

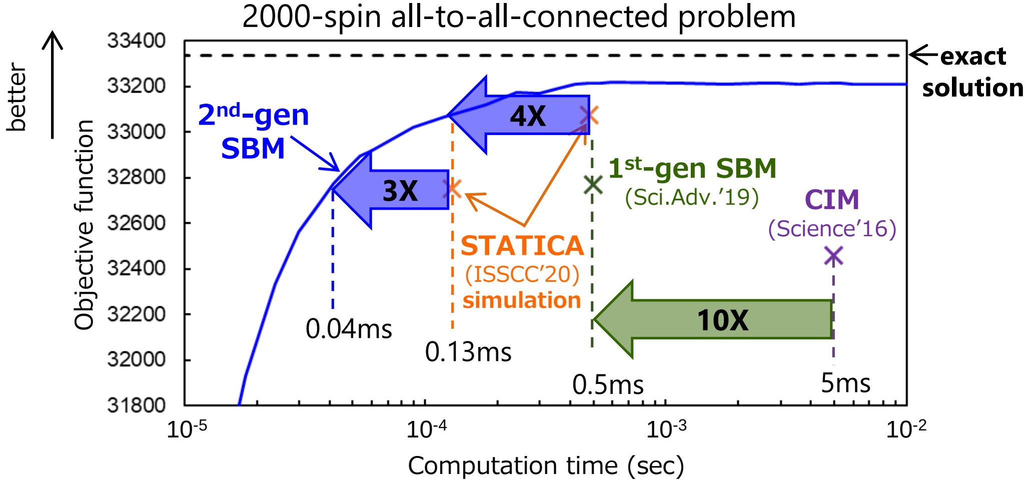
## Improvement of Quality-of-Solution

Quasi-quantum tunneling effects

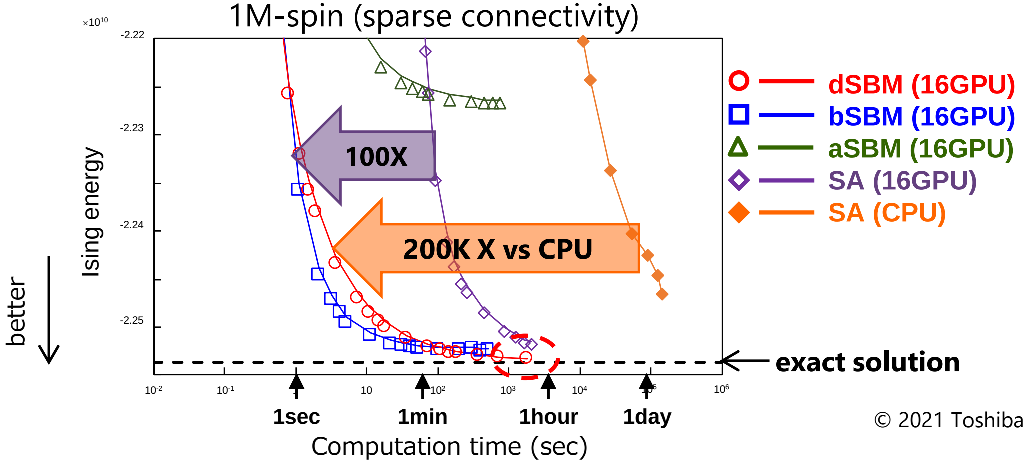




## 10X faster than 1st-gen SB


2000-spin all-to-all-connected problem

## Solves 1M-spin problem in 30 min


1M-spin (sparse connectivity)

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

# 2nd-gen simulated bifurcation technology (Feb, '21)

## Comprehensive comparison with state-of-the-art Ising machines

Science Advances

AAAS | open access journal

"High-performance combinatorial optimization based on classical mechanics"

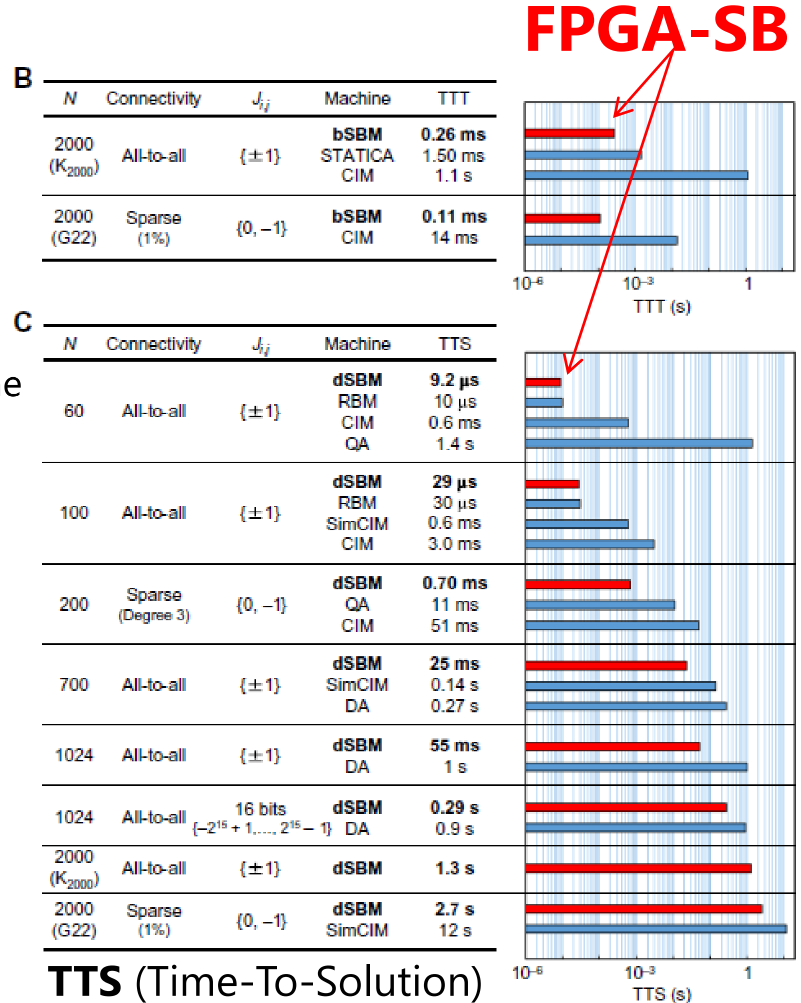
H. Goto et al

eabe7953

February 03, 2021

### Competitors

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



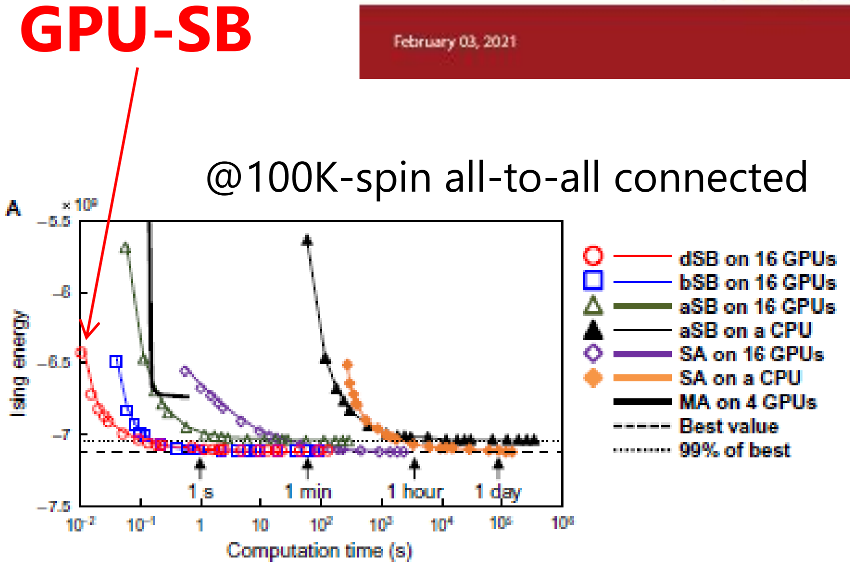
B

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

C

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

**TTS** (Time-To-Solution)



@100K-spin all-to-all connected

**SB technology is competitive**

# Contents

01 Introduction:
Combinatorial optimization in real-time systems

02 Simulated bifurcation (SB)

03 Real-time systems that make optimal responses

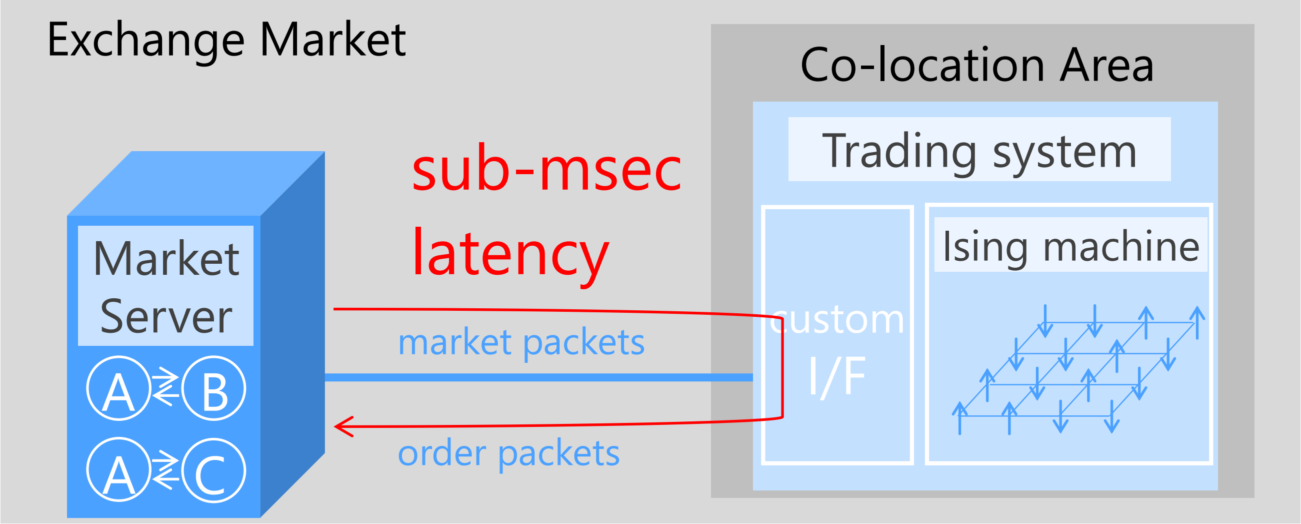
04 Scaling out Ising machines

# FPGA accelerator for SB: For what?

**Real-time systems that make optimal responses**

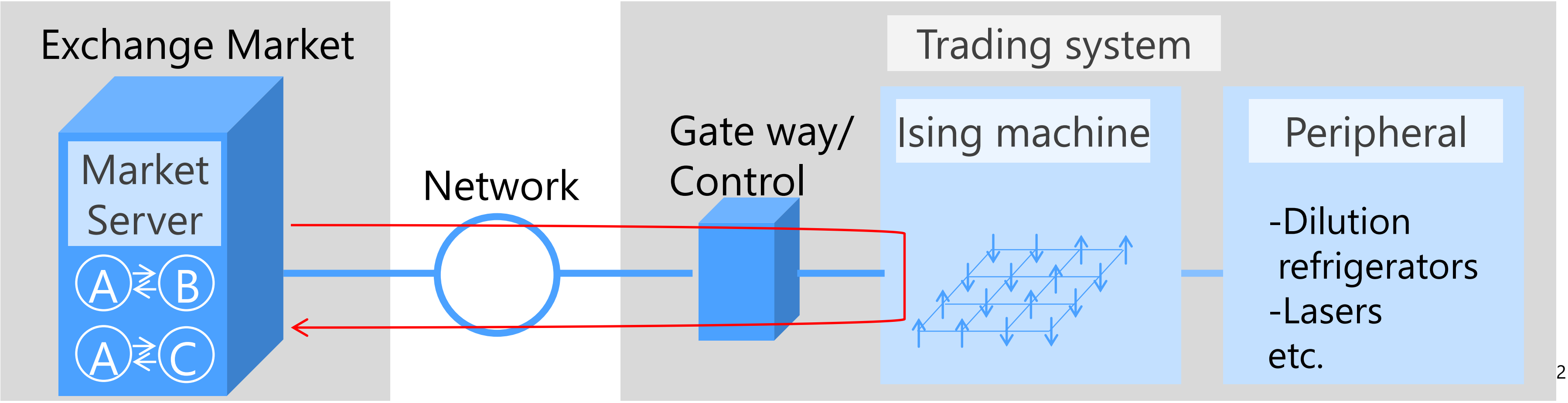
System-wide response time (in sub-ms) is CRITICAL

Ultralow latency



Sub-ms latency has not yet been demonstrated for any Ising machines

Conv.

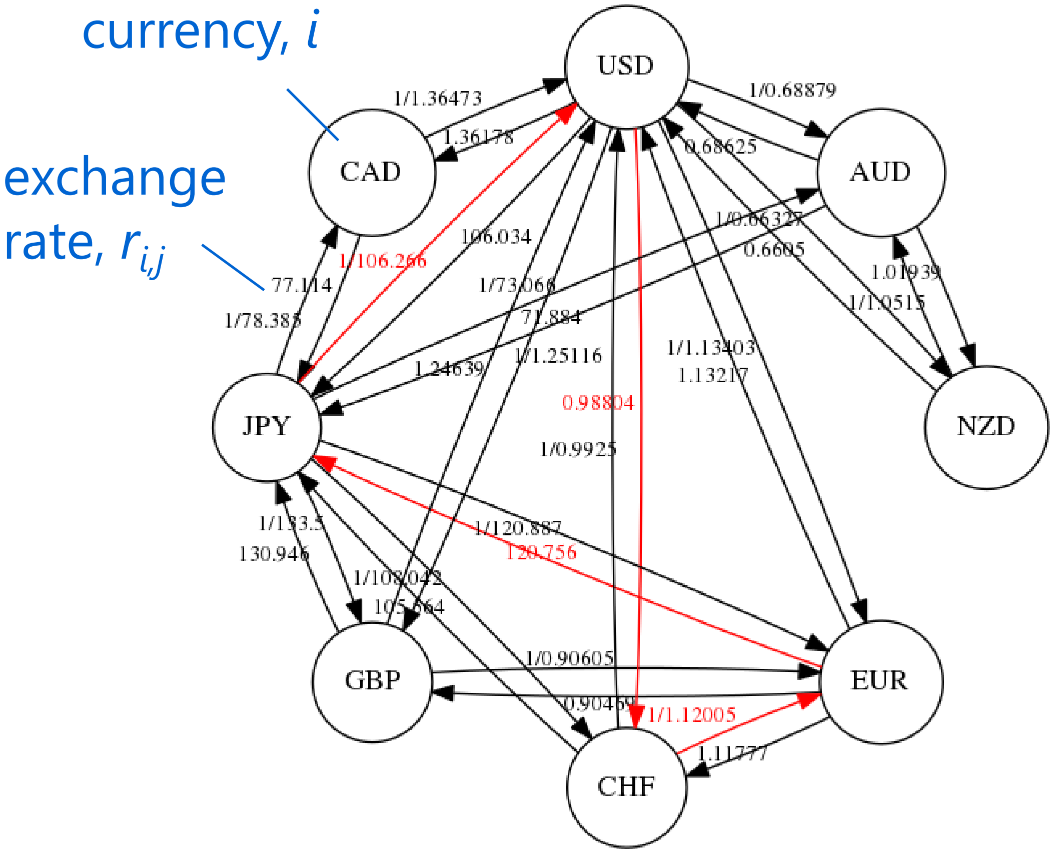


Tatsumura *et al.*, A Currency Arbitrage Machine Based on the Simulated Bifurcation Algorithm for Ultrafast Detection of Optimal Opportunity, IEEE Int'l Symp. on Circuits and Systems (ISCAS), 1-5 (2020). https://doi.org/10.1109/ISCAS45731.2020.9181114

# Detection of cross-currency arbitrage opportunity

## Optimal path search in a directed graph -a combinatorial optimization problem-

### Market Graph



[W. Soon et al., *Int'l J. Math. Edu. Sci. Tech.* **42**, 369-376 (2011)]

**Arbitrage Problem**

find a closed path that maximizes the profit

Cost function

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

Constraint

Must be a closed path

**Requirement**

find more profitable paths in a shorter time

Tatsumura *et al.*, A Currency Arbitrage Machine Based on the Simulated Bifurcation Algorithm for Ultrafast Detection of Optimal Opportunity, IEEE Int'l Symp. on Circuits and Systems (ISCAS), 1-5 (2020). https://doi.org/10.1109/ISCAS45731.2020.9181114

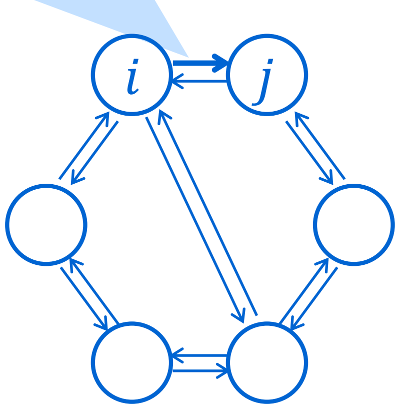
# **Problem formulation: QUBO** (Quadratic Unconstrained Binary Optimization)

## QUBO formulation of the arbitrage problem

For each edge,

Decision variable (binary) $b_{i,j}$

Exchange rate $r_{i,j}$



**Cost function** (high-order polynomial)

**Cost function** (linear)

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

**Penalty function** (quadratic)

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

outflow < 1

inflow < 1

outflow=inflow

forbids traversing the same edge twice

**Total cost function** (quadratic)

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

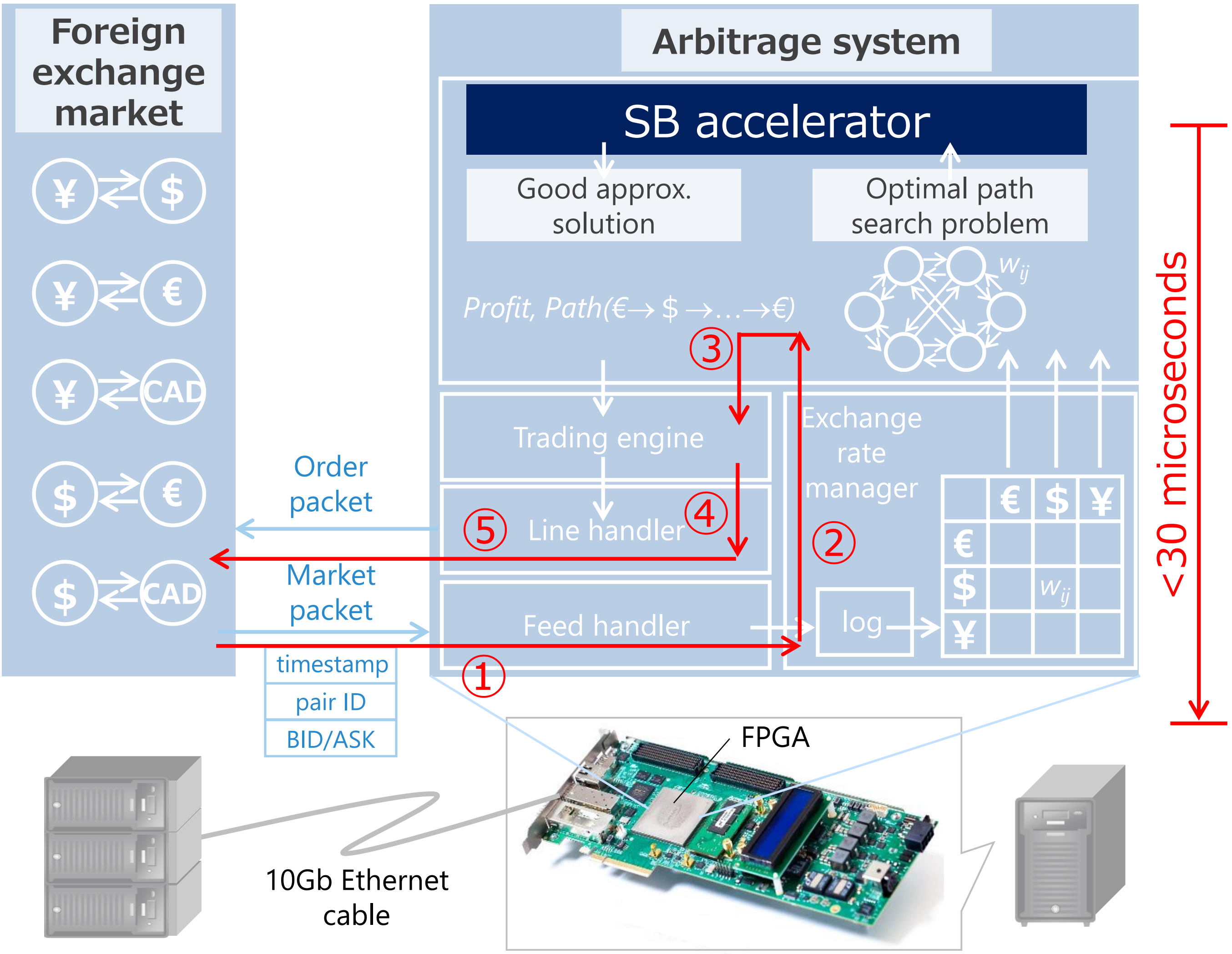
**Arbitrage problem as a QUBO:**

Optimize the bit configuration $\{b_{i,j}\}$ to minimize the quadratic cost function $C_{tot}$

Tatsumura *et al.*, A Currency Arbitrage Machine Based on the Simulated Bifurcation Algorithm for Ultrafast Detection of Optimal Opportunity, IEEE Int'l Symp. on Circuits and Systems (ISCAS), 1-5 (2020). https://doi.org/10.1109/ISCAS45731.2020.9181114

# System configuration

## An end-to-end FPGA-based arbitrage system



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

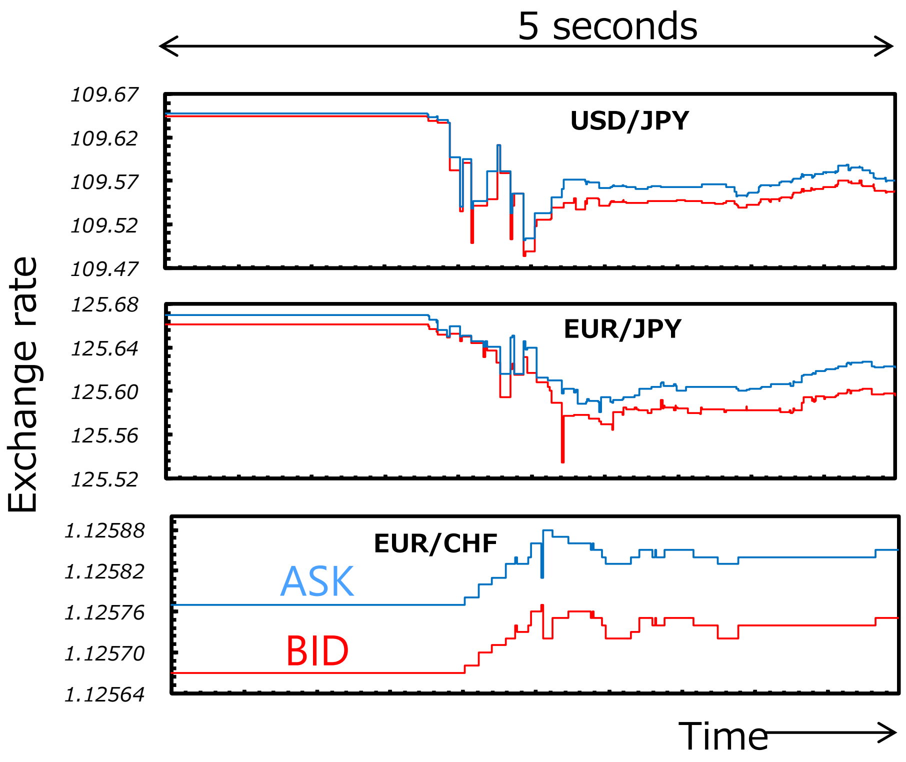
⑤Custom I/F (line handler)
issues the order packets

Tatsumura *et al.*, A Currency Arbitrage Machine Based on the Simulated Bifurcation Algorithm for Ultrafast Detection of Optimal Opportunity, IEEE Int'l Symp. on Circuits and Systems (ISCAS), 1-5 (2020). https://doi.org/10.1109/ISCAS45731.2020.9181114

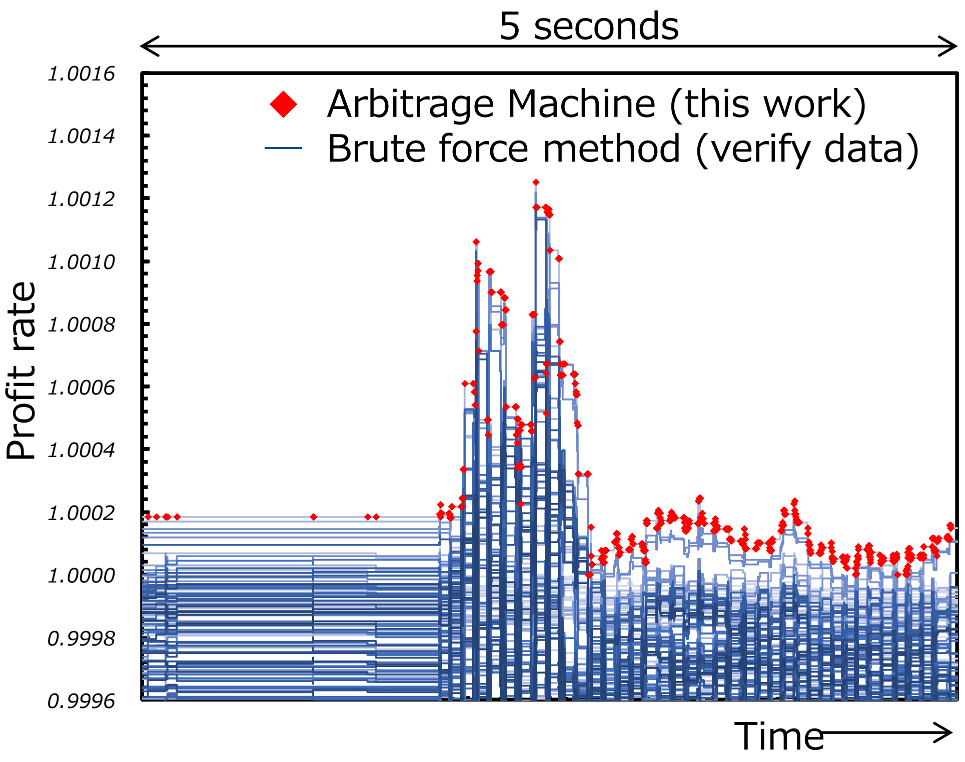
# Demonstration: How it works

## The system's responses to real market situation on January 2nd, 2019

Exchange rates



Profit rates for arbitrage paths



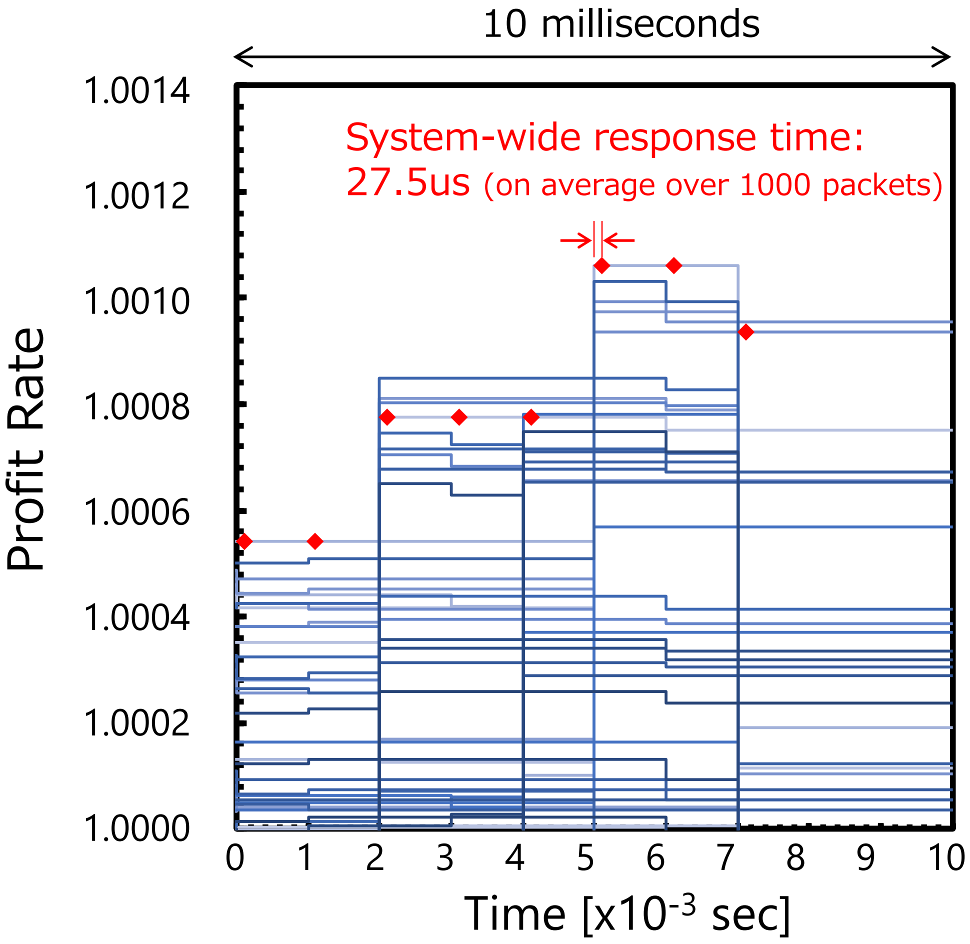
Red markers on the top blue line
→Finding *optimal* path *in real-time*

Tatsumura *et al.*, A Currency Arbitrage Machine Based on the Simulated Bifurcation Algorithm for Ultrafast Detection of Optimal Opportunity, IEEE Int'l Symp. on Circuits and Systems (ISCAS), 1-5 (2020). https://doi.org/10.1109/ISCAS45731.2020.9181114

# Performance: Response time & Accuracy

## <30us System Latency & 91% Top-1 Probability

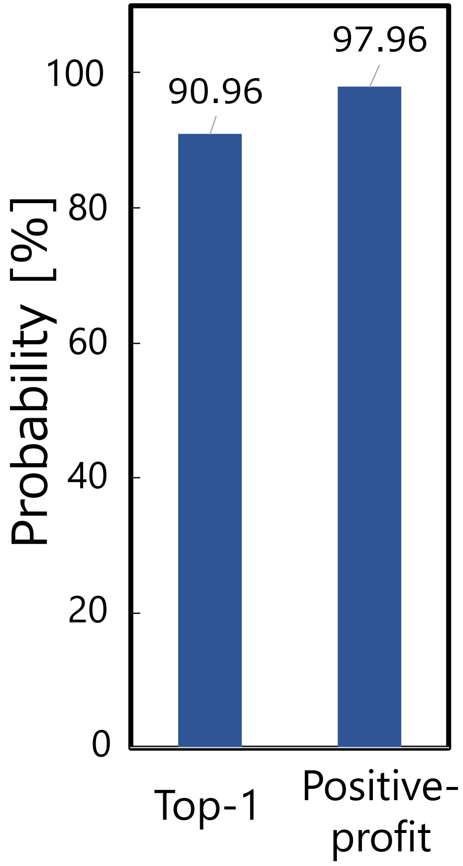
### Response time



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

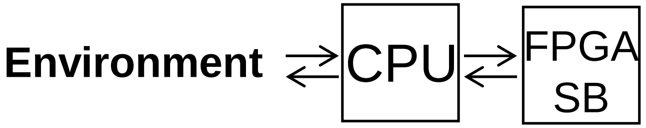
### Solution accuracy

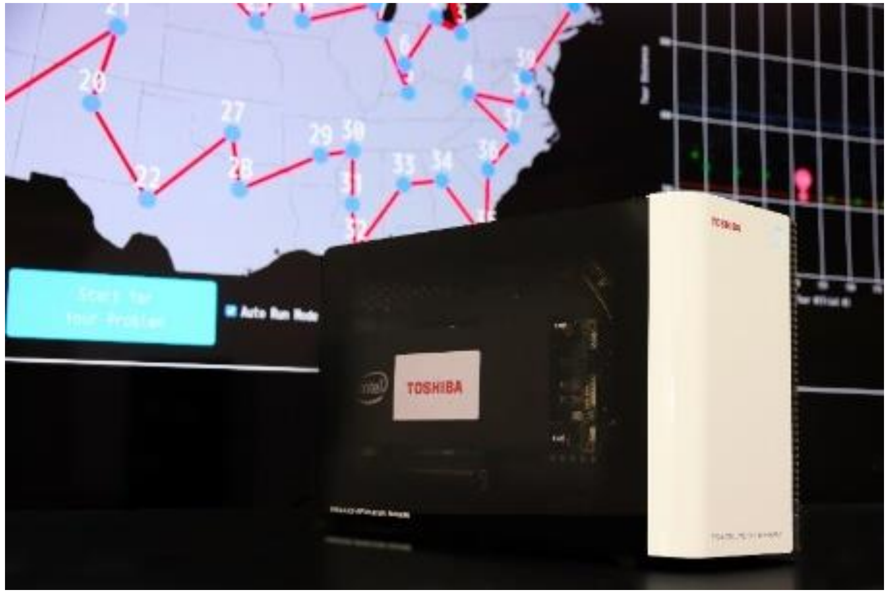
In the one-month data, 34,471,865 distinct events,
21.3% (7,355,698) were profitable (at least one path with a profit rate >1.0)



# On-premises ver. of simulated bifurcation machine$^{TM}$ (Mar., 2021)

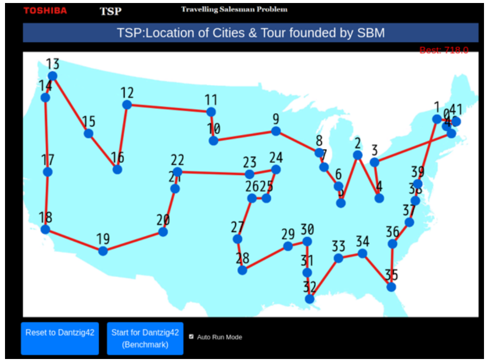
## A look-aside FPGA accelerator for SB



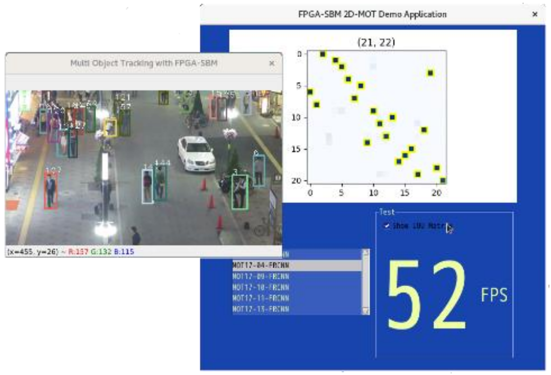


## C/Python APIs for software engineers

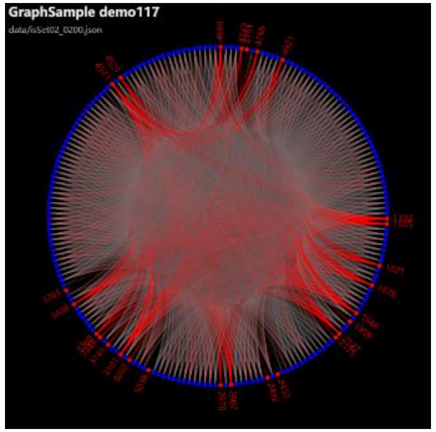
## Reference designs of real-time applications


TOSHIBA
TSP
Travelling Salesman Problem
TSP:Location of Cities & Tour founded by SBM
Reset to Dantzig42
Start for Dantzig42 (Benchmark)
Auto Run Mode

a user-interactive interface for solving the travelling salesman problem


FPGA-SBM 2D-MOT Demo Application
Multi Object Tracking with FPGA-SBM
(21, 22)
52
FPS

multi-object tracking by solving the maximum matching problem



stream data processing of market graph for finding the diversified portfolio through solving the maximum-independent-set problem

# Contents

01 Introduction:
Combinatorial optimization in real-time systems

02 Simulated bifurcation (SB)

03 Real-time systems that make optimal responses

04 Scaling out Ising machines

*1 K. Tatsumura et al., *FPL19*
*2 K. Tatsumura et al., *ISCAS20*

# Enlarging machine size & enhancing processing speeds

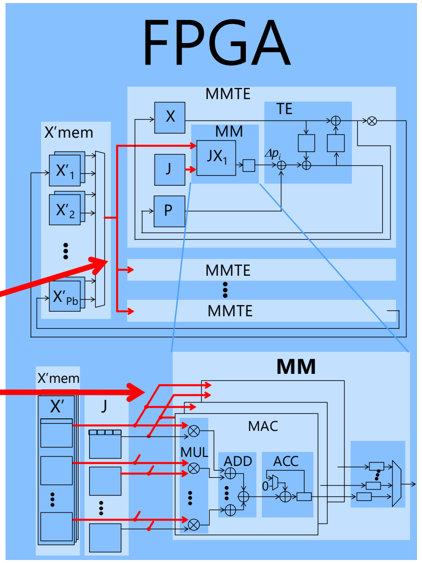
## -Enlarging machine size while keeping computational efficiency-

### Single-FPGA accelerator for SB[*1]

# of PEs: 8,192
Effective activity factor: 92%

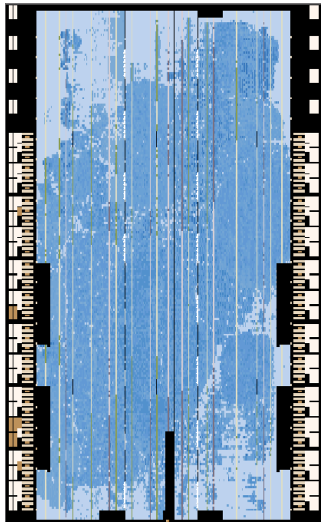
Sufficient data supply to PEs
Peak rate: 4,678GB/s

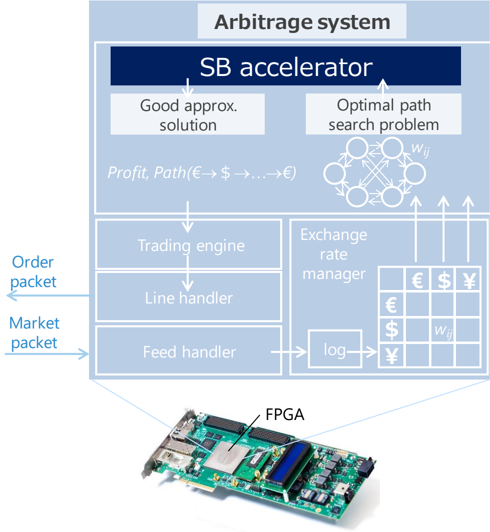
On-chip memory/wiring resources



The machine size is limited by on-chip memory

### Single-FPGA arbitrage machine[*2]





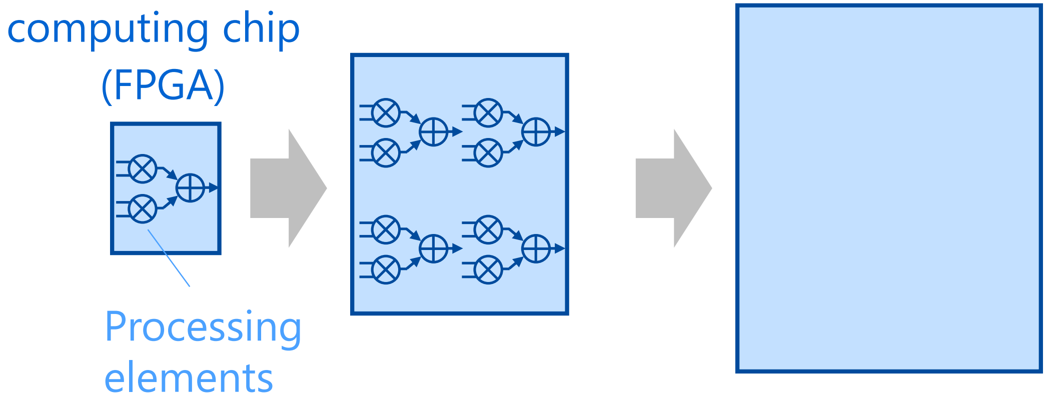
Maximum market graph:
Fully-connected 16-asset problems
(16 nodes, 256 directed edges)

What can we do if we want to take more assets into account?

# Two approaches: scale-up and scale-out

## Scale-up

making a chip larger (denser)

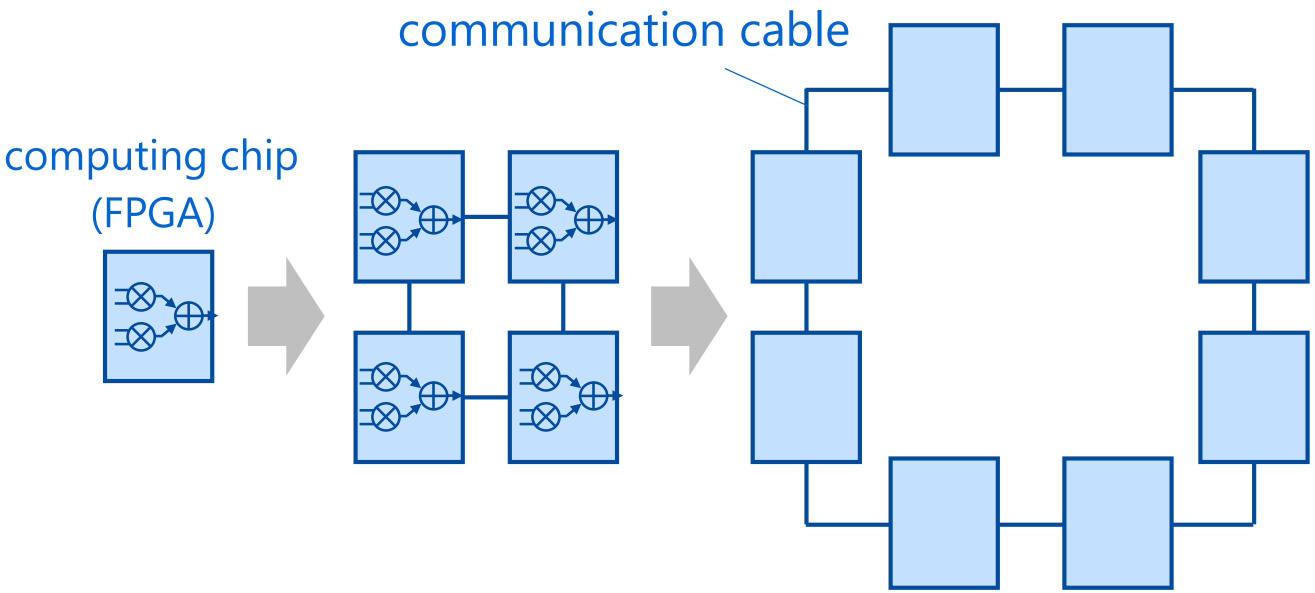


Pros. Reactively easy to enhance the performance
Cons. Need to develop a manufacturing technology

This work

## Scale-out

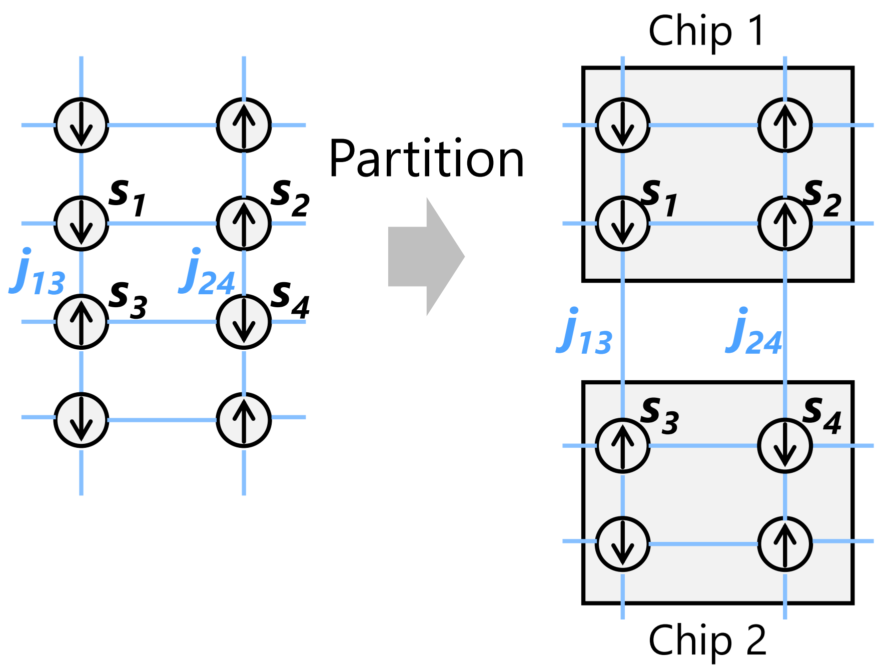
increasing the number of networked chips



Pros. Enables continued enlargement of the computing scale
Cons. Need to develop a cluster architecture to avoid performance saturation due to communication overhead

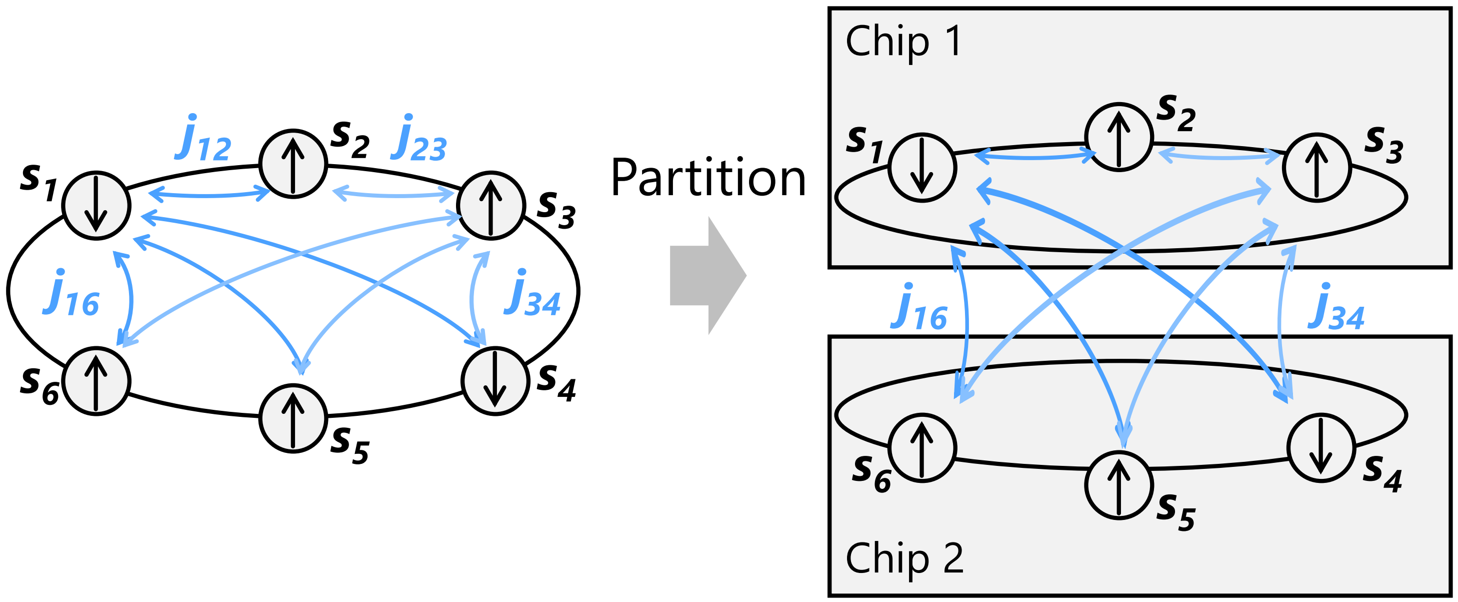
# Partitioning spin networks with local-/full-connectivity

## Locally-connected network



This work

## Fully-connected network



More practical value, but more difficult to partition

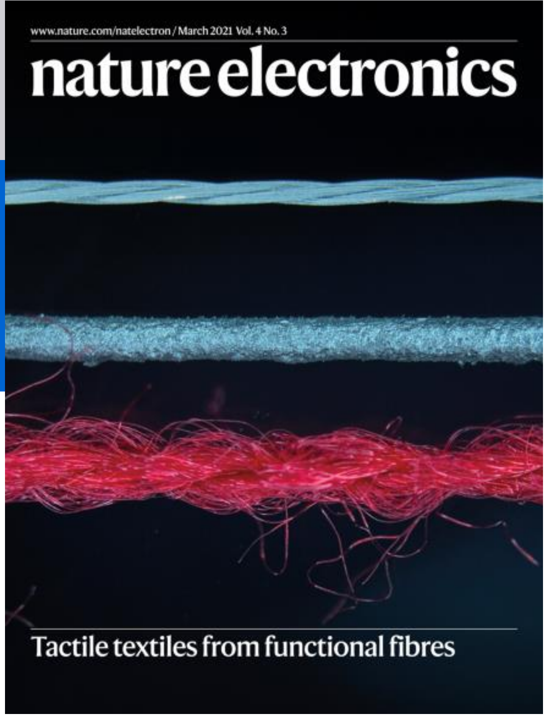
**Issues:** Spin-spin couplings over the subsystems must be incorporated
Partitioned subsystems also have to evolve in a single time domain

→Communication and synchronization can easily degrade the speed performance.

Tatsumura *et al.*, Nature Electronics **4**, 208 (2021)

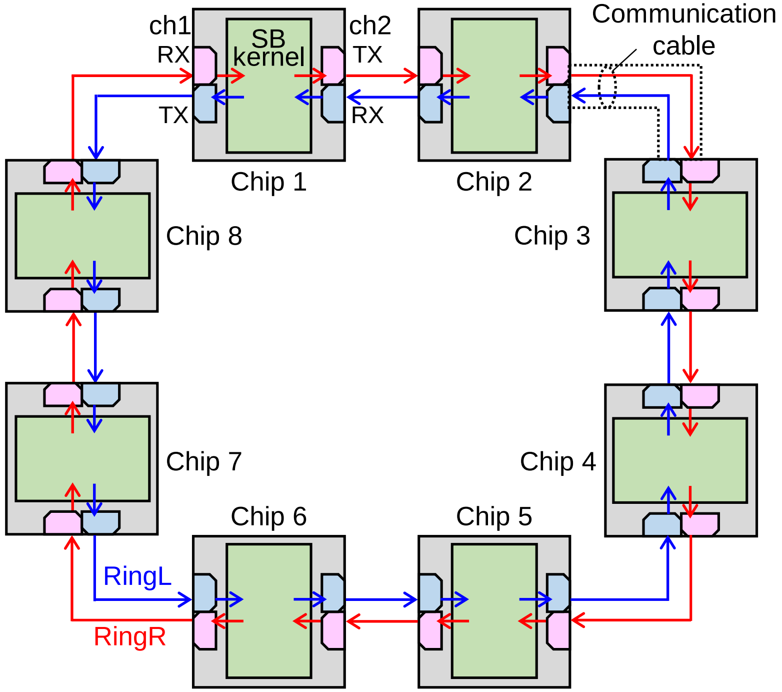
# Scaling out Ising machine with full connectivity

## A multi-chip architecture for SB that enables continued scaling of both machine size and computational throughput



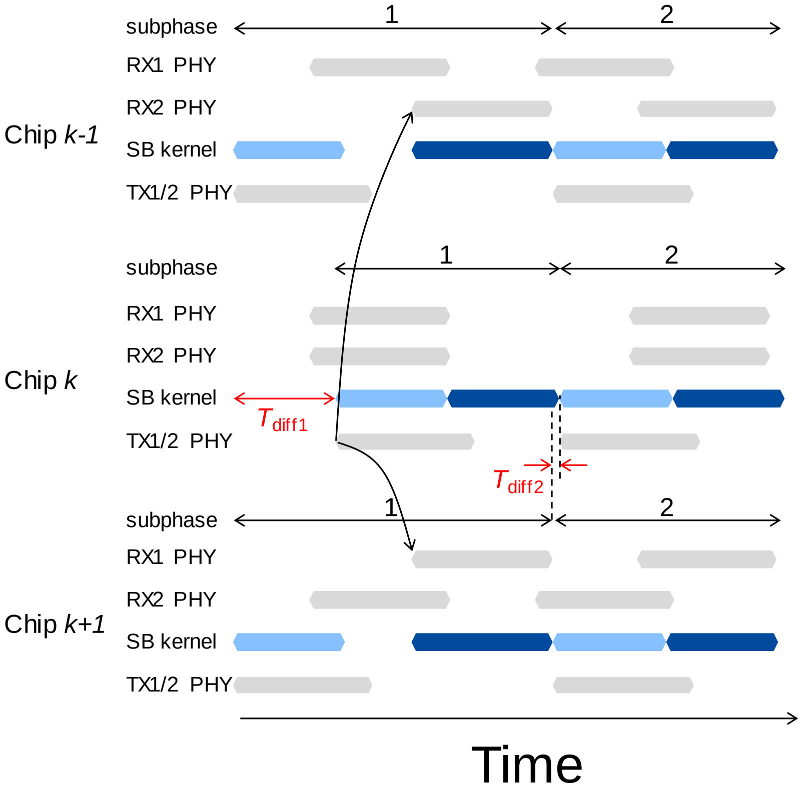
www.nature.com/natelectron / March 2021 Vol. 4 No. 3
nature electronics
Tactile textiles from functional fibres

### Bidirectional ring-network cluster without any centralized features

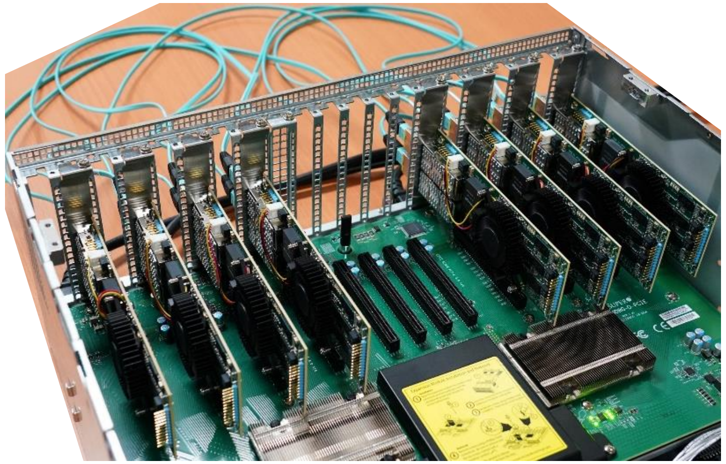


All chips are
Autonomous, homogeneous and symmetric

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



$P_{chip}=8$



Tatsumura *et al.*, Nature Electronics **4**, 208 (2021)

# Demonstration: Scaling out characteristics

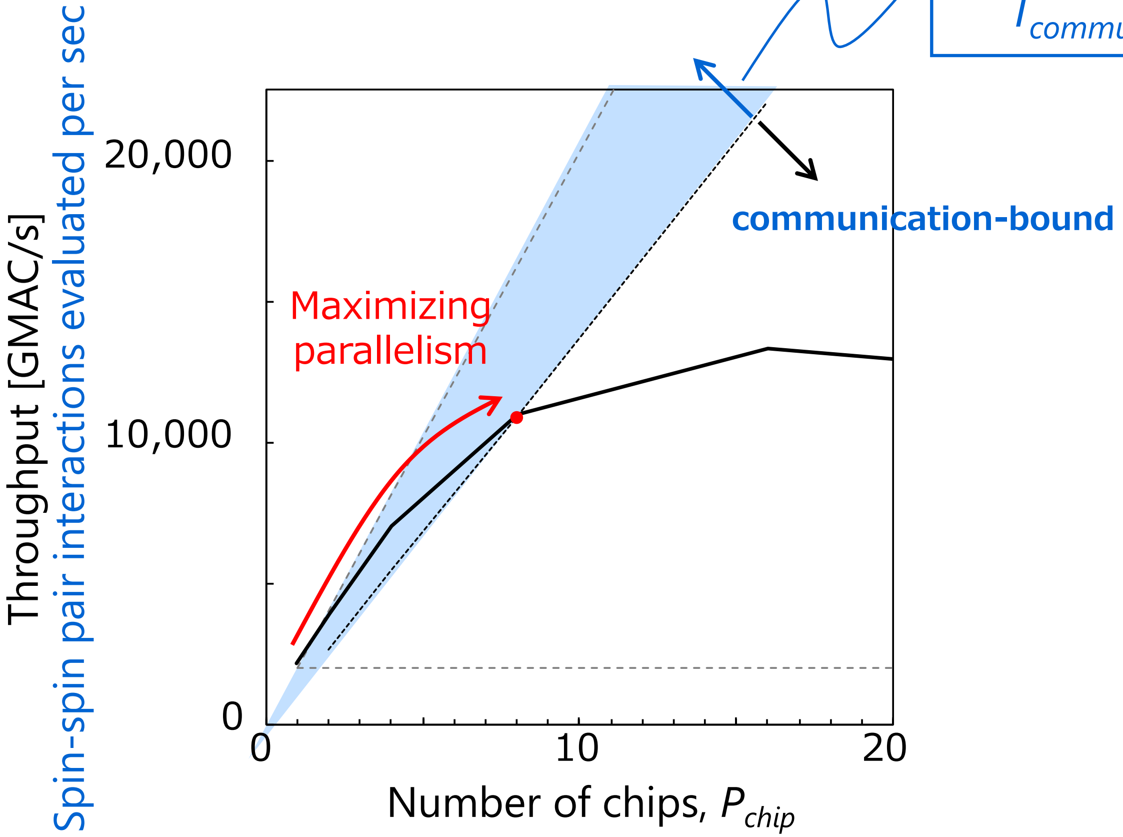
## Constant-efficiency scaling at the maximized computation parallelism

### Strong scaling

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

**Computation-bound**

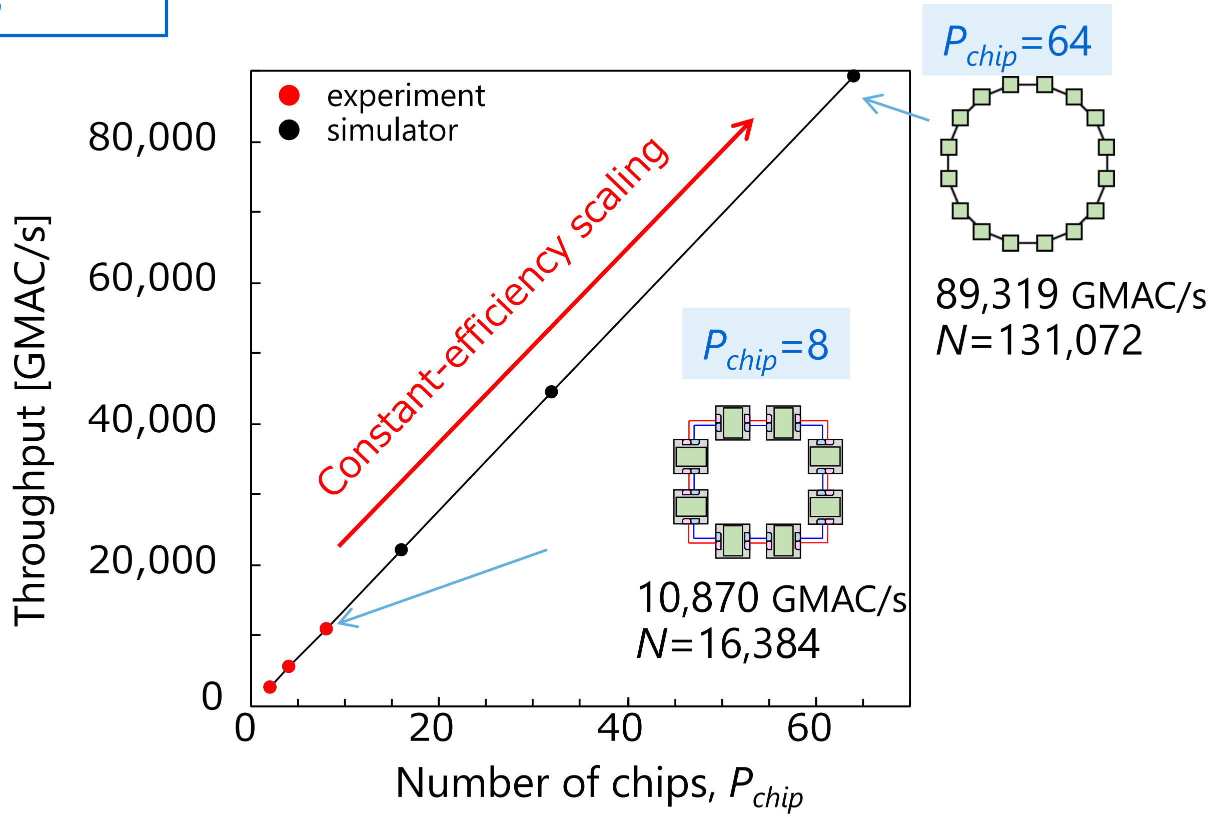
$$\frac{T_{computation}}{T_{communication}} > 1$$



Demonstration: the throughput enhancement to the vicinity of an ideal upper limit determined by $N$ and communication technology

### Weak scaling

Increase $P_{\text{chip}}$ and $N$ in the same proportion



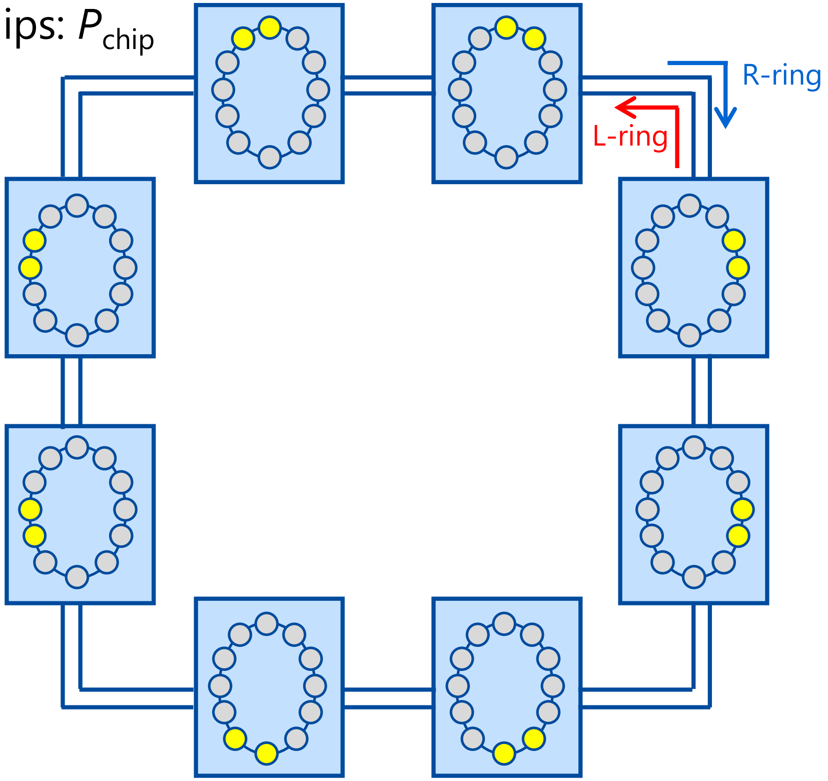
Demonstration: the constant-efficiency scaling at the maximized computation parallelism (at the strong scaling limit)

Tatsumura *et al.*, Nature Electronics **4**, 208 (2021)

# A multi-chip architecture for simulated bifurcation

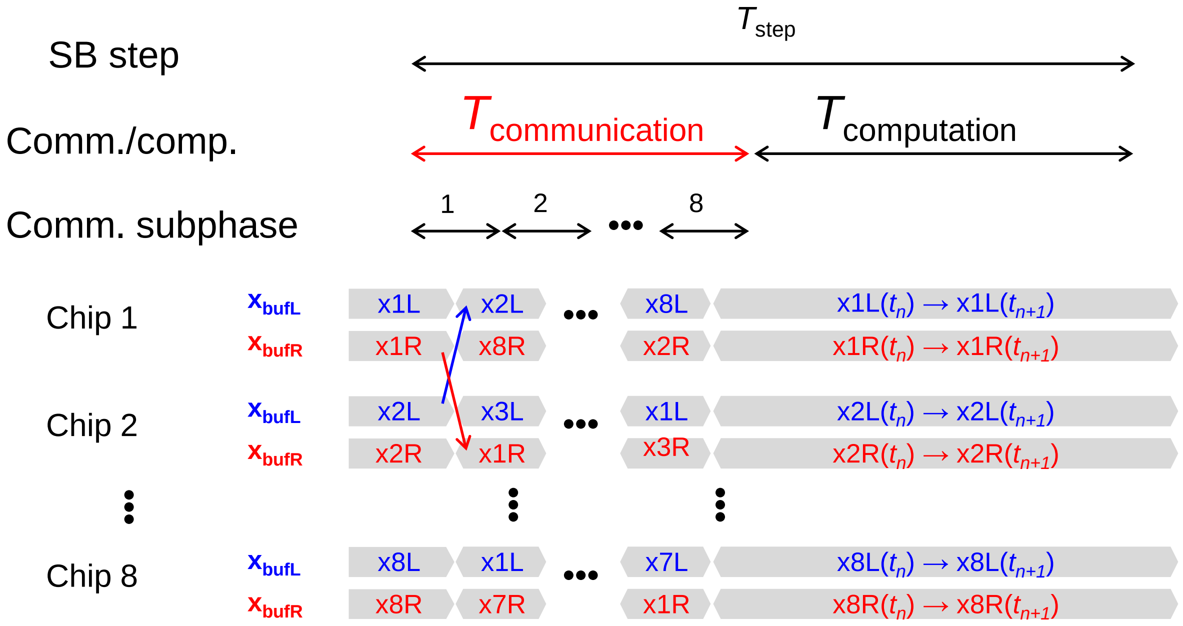
## Partitioned SB algorithm that can be executed simultaneously on multiple chips

Number of spins: $N$
Number of chips: $P_{\mathrm{chip}}$



Each chip (spin subsystem) is responsible for $N/P_{\mathrm{chip}}$ spins

Each chip needs all the spin information to update the $N/P_{\mathrm{chip}}$ spins (all-to-all connectivity)



| | | 1 | 2 | | 8 | |
|---|---|---|---|---|---|---|
| Chip 1 | $x_{\mathrm{bufL}}$ | x1L | x2L | ••• | x8L | $x1L(t_n) \to x1L(t_{n+1})$ |
| | $x_{\mathrm{bufR}}$ | x1R | x8R | | x2R | $x1R(t_n) \to x1R(t_{n+1})$ |
| Chip 2 | $x_{\mathrm{bufL}}$ | x2L | x3L | ••• | x1L | $x2L(t_n) \to x2L(t_{n+1})$ |
| | $x_{\mathrm{bufR}}$ | x2R | x1R | | x3R | $x2R(t_n) \to x2R(t_{n+1})$ |
| ⋮ | | | ⋮ | | ⋮ | |
| Chip 8 | $x_{\mathrm{bufL}}$ | x8L | x1L | ••• | x7L | $x8L(t_n) \to x8L(t_{n+1})$ |
| | $x_{\mathrm{bufR}}$ | x8R | x7R | | x1R | $x8R(t_n) \to x8R(t_{n+1})$ |

Communication phase
-Share all the spin information
-Repeating exchange processes of spins btw neighboring chips

Computation phase
-compute the time-evolved state in a chip-parallel fashion

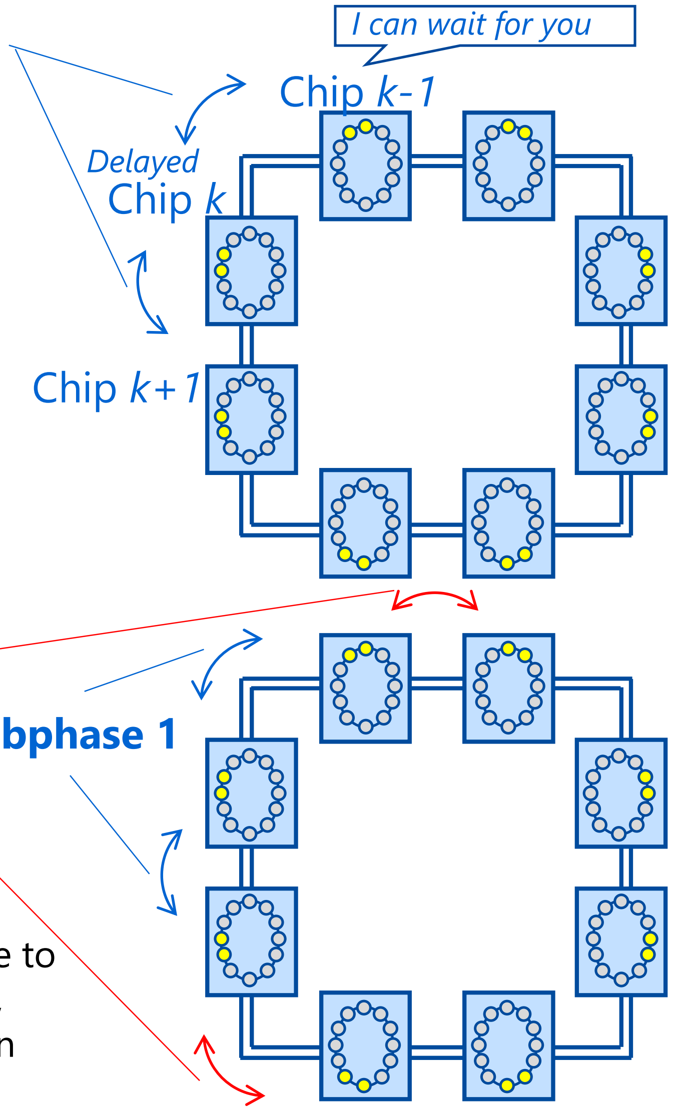
Tatsumura *et al.*, Nature Electronics **4**, 208 (2021)

# A multi-chip architecture for simulated bifurcation

## Autonomous synchronization mechanism

### Local synchronization

When Chip *k* is delayed,
Chip *k±1* wait for Chip *k*
until Chip *k* gets ready



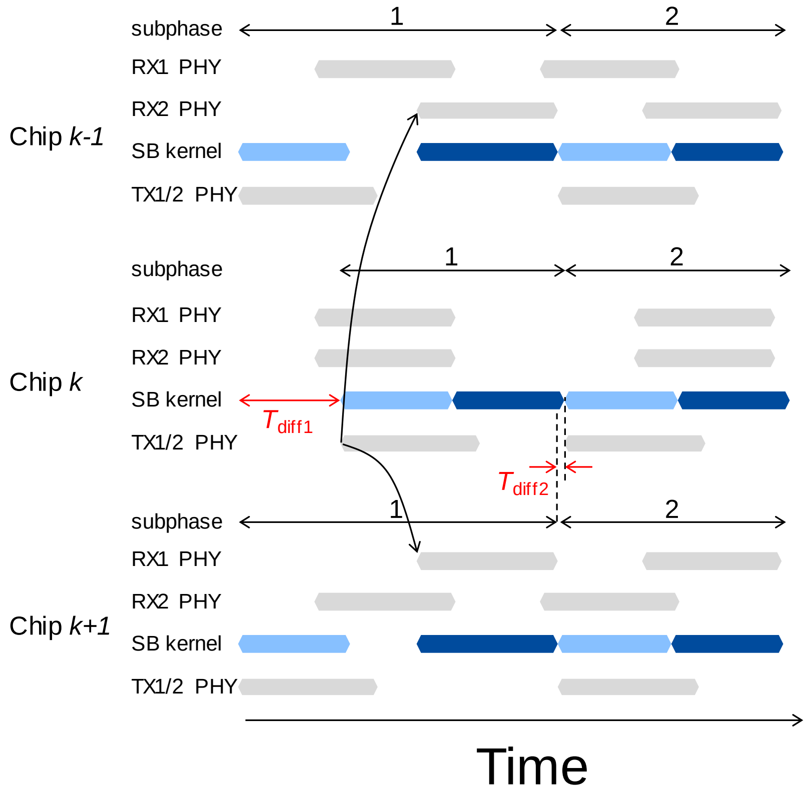
### Global synchronization

subphase 2

subphase 1

Local synchronization propagate to adjacent nodes every subphase, achieving global synchronization

### Autonomous synchronization mechanism



Global synchronization without a centralized control node (a chokepoint)

→ Good scalability of the processing speed

# Conclusions

## FPGA-based accelerators for simulated bifurcation that enables large-scale combinatorial optimization in real-time systems

### Simulated bifurcation (SB):

a quantum-inspired algorithm having plentiful parallelism

### FPGA-based accelerators for SB:

massively-parallel, fully-customized circuit architecture
very practical (no refrigerator, no laser, but in FPGA)
can be scaled out with an autonomously-synchronizable multi-chip architecture

### Real-time systems that make optimal responses

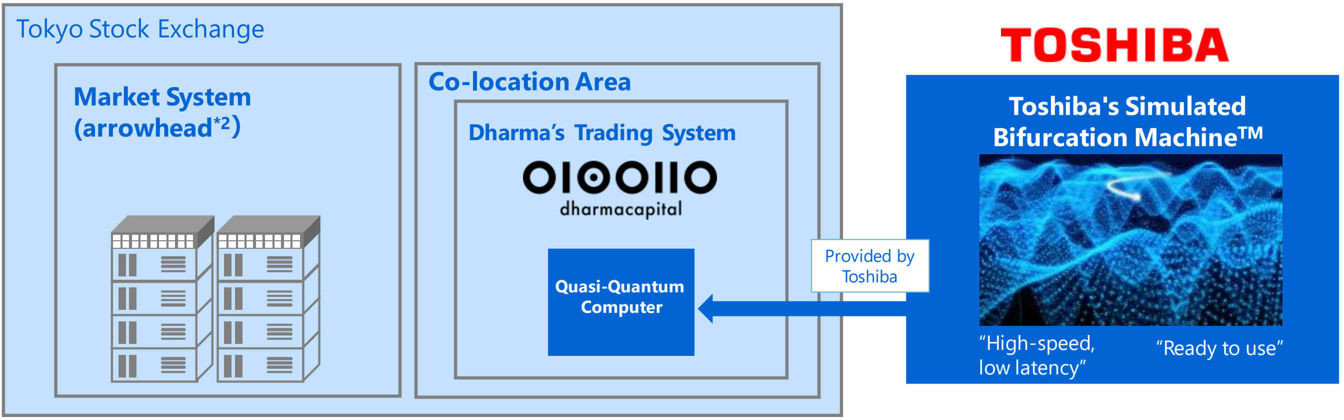
an example: an end-to-end cross-currency arbitrage system
with <30us system latency & 91% top-1 probability

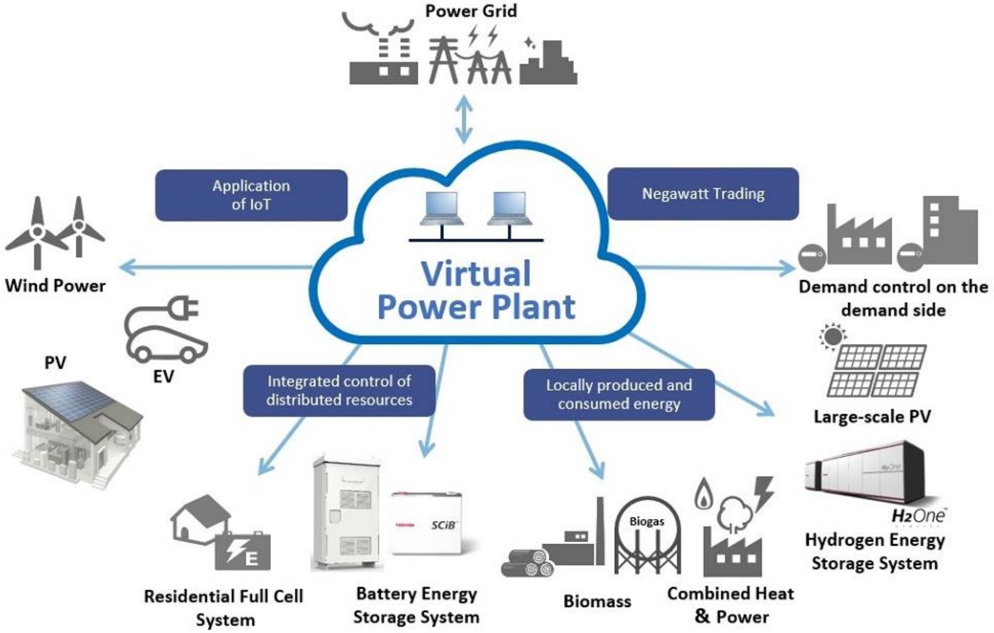
# Future outlook

## Toward creating various innovative real-time systems

**[On-going]** Testing high-frequency trading systems using SB accelerators in the Tokyo stock exchange

**[Future]** High-speed dynamic pricing systems applicable to virtual power plant





Power Grid
Application of IoT
Negawatt Trading
Virtual Power Plant
Wind Power
Demand control on the demand side
PV
EV
Integrated control of distributed resources
Locally produced and consumed energy
Large-scale PV
Hydrogen Energy Storage System
H2One
Residential Full Cell System
Battery Energy Storage System
SCiB
Biomass
Biogas
Combined Heat & Power

https://www.global.toshiba/ww/technology/corporate/rdc/rd/topics/21/2105-01.html

https://www.toshiba-clip.com/en/detail/p=228

# References: Simulated Bifurcation Machine™

**【Official web site】**

Simulated Bifurcation Machine™
https://www.toshiba-sol.co.jp/en/pro/sbm/index.htm

**【Journal papers/Peer-reviewed conference papers/Toshiba’s Press Release】**

[1] The story of the birth of simulated bifurcation machine™
https://www.toshiba-clip.com/en/detail/p=76

[2] 1st Announcement of simulated bifurcation machine™
H. Goto *et al.*, “Combinatorial optimization by simulating adiabatic bifurcations in nonlinear Hamiltonian systems”, *Science Advances* **5**, eaav2372 (2019).
https://doi.org/10.1126/sciadv.aav2372
Toshiba's Breakthrough Algorithm Realizes World's Fastest, Largest-scale Combinatorial Optimization https://www.toshiba.co.jp/rdc/rd/detail_e/e1904_01.html

[3] 2-gen simulated bifurcation machine™
H. Goto *et al.*, “High-performance combinatorial optimization based on classical mechanics”, *Science Advances* **7**, eabe7953 (2021).
https://doi.org//10.1126/sciadv.abe7953
Toshiba’s New Algorithms Quickly Deliver Highly Accurate Solutions to Complex Problems
https://www.toshiba.co.jp/rdc/rd/detail_e/e2102_01.html

[4] Custom accelerator
K. Tatsumura *et al.*, “FPGA-Based Simulated Bifurcation Machine”, *IEEE Int'l Conf. on Field-Programmable Logic and Applications* (FPL), 59-66 (2019).
https://doi.org/10.1109/FPL.2019.00019
Toshiba Develops a Dedicated Massively Parallel Processing Circuit for Simulated Bifurcation Algorithms https://www.toshiba.co.jp/rdc/rd/detail_e/e1909_03.html

[5] PoC for ultra-fast financial transaction machine
K. Tatsumura *et al.*, “A Currency Arbitrage Machine Based on the Simulated Bifurcation Algorithm for Ultrafast Detection of Optimal Opportunity”, *IEEE Int’l Symp. on Circuits and Systems* (ISCAS), 1-5 (2020). https://doi.org/10.1109/ISCAS45731.2020.9181114
M. Yamasaki *et al.*, “Live Demonstration: Capturing Short-Lived Currency Arbitrage Opportunities with a Simulated Bifurcation Algorithm-Based Trading System*”, IEEE Int’l Symp. on Circuits and Systems* (ISCAS), 1-1 (2020).
https://doi.org/10.1109/ISCAS45731.2020.9180679
Toshiba Develops Proof-of-concept Device for Ultra-high-speed Financial Transaction Machine with Simulated Bifurcation Algorithm
https://www.toshiba.co.jp/rdc/rd/detail_e/e1910_02.html

[6] Scaling out Ising machines
K. Tatsumura *et al.*, “Scaling-out Ising machines using a multi-chip architecture for simulated bifurcation”, *Nature Electronics* **4**, 208–217 (2021).
https://doi.org/10.1038/s41928-021-00546-4
Also see BEHIND THE PAPER: https://go.nature.com/2MuGe21
Cutting-edge Scale-Out Technology from Toshiba will Take Fintech and Logistics to New Level
https://www.toshiba.co.jp/rdc/rd/detail_e/e2103_01.html

[7] On-premises version of Simulated Bifurcation Machine™
Toshiba Offers On-premises Simulated Bifurcation Machine™ for Market Trials in Japan
https://www.global.toshiba/ww/technology/corporate/rdc/rd/topics/21/2103-03.html

[8] Testing SB-based financial transaction machines in the Tokyo stock exchange
Toshiba and Dharma Capital’s Joint Experiment in Financial Markets to Verify the Effectiveness of a Quasi-Quantum Computer When Applied to High Frequency Trading
https://www.global.toshiba/ww/technology/corporate/rdc/rd/topics/21/2105-01.html

[9] Real-time systems that make optimal responses
K. Tatsumura *et al.*, “Large-scale combinatorial optimization in real-time systems by FPGA-based accelerators for simulated bifurcation”, *ACM Int’l Symp. on Highly Efficient Accelerators and Reconfigurable Technologies* (HEART), 1-6 (2021).
https://doi.org/10.1145/3468044.3468045

**【News Media 】**

[1] IEEE Spectrum (Dec, ’19)
Toshiba’s Optimization Algorithm Sets Speed Record for Solving Combinatorial Problems
https://spectrum.ieee.org/tech-talk/computing/software/toshiba--optimization-algorithm-speed-record-combinatorial-problems#.XkHhXCOP1GQ.email

[2]Analytics India Magazine (Dec, ’19)
Top 5 Algorithm Breakthroughs In 2019
https://analyticsindiamag.com/top-5-algorithm-breakthroughs-in-2019/

[3] COMMUNICATIONs of the ACM (May, ’21)
Quantum Simulator Beats Quantum Hardware
https://cacm.acm.org/news/252584-quantum-simulator-beats-quantum-hardware/fulltext

[4] Risk.net (Jun, ‘21)
Quantum kit offers HFTs ‘100-fold’ speed boost
https://www.risk.net/derivatives/7840361/quantum-kit-offers-hfts-100-fold-speed-boost