HITACHI
Inspire the Next

An In-memory Computing Accelerator, CMOS Annealing Machine, to Solve Combinatorial Optimization Problems

Masanao Yamaoka

Research & Development group, Hitachi, Ltd., Tokyo, Japan

07/11/2019

MPSoC Forum 2019

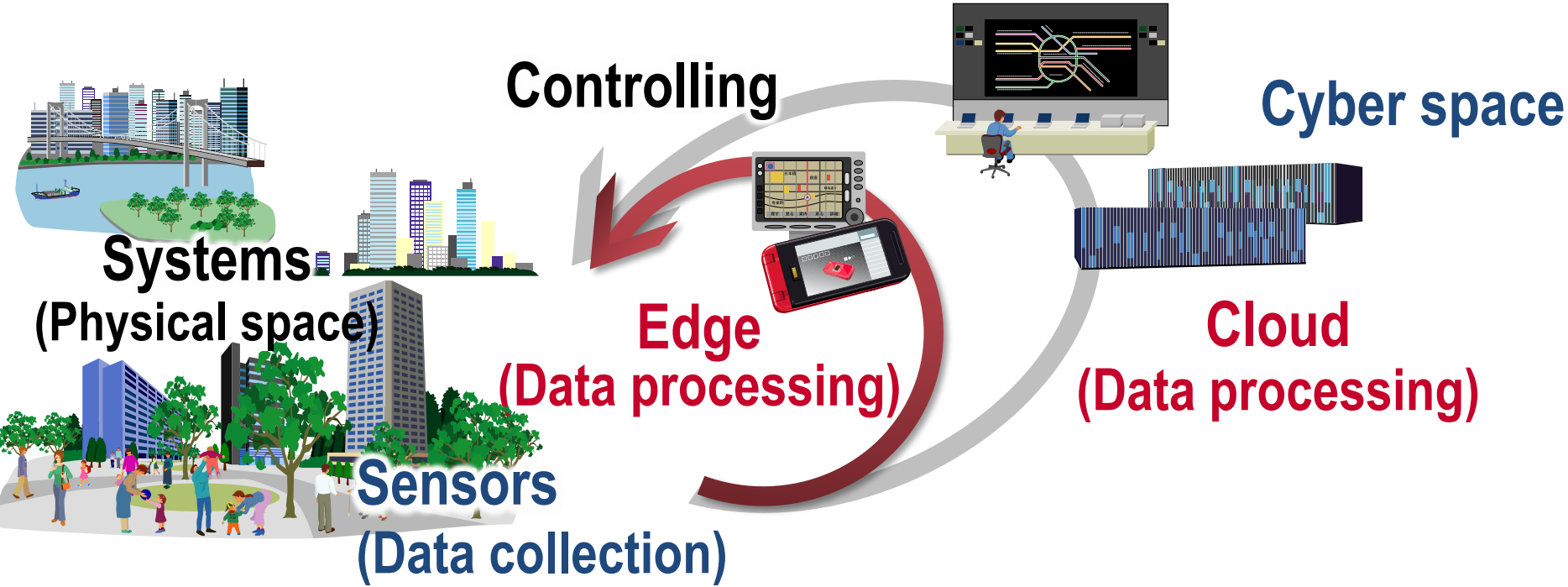
Outline

- Motivations
- Overview of CMOS annealing machine
- Prototypes of CMOS annealing machine
- Related necessary technologies
- Conclusion

System of IoT era

HITACHI
Inspire the Next

- Collecting and analyzing a lot of data
- Results used for controlling systems



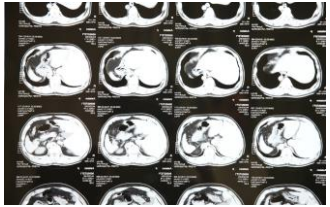
Importance of optimization

HITACHI
Inspire the Next

- **Optimization processing used in various industries**
- **Optimization problem necessary to acquire optimum parameters**

Logistics operation

VLSI design



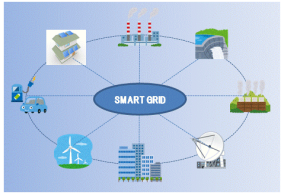
Medical diagnosis with images

Management strategy

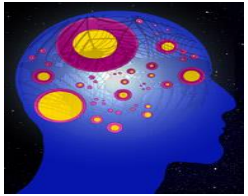
Route selection (logistics/services)

Scheduling at wide-scale disaster

Learning plan customizing



Smart-grid control

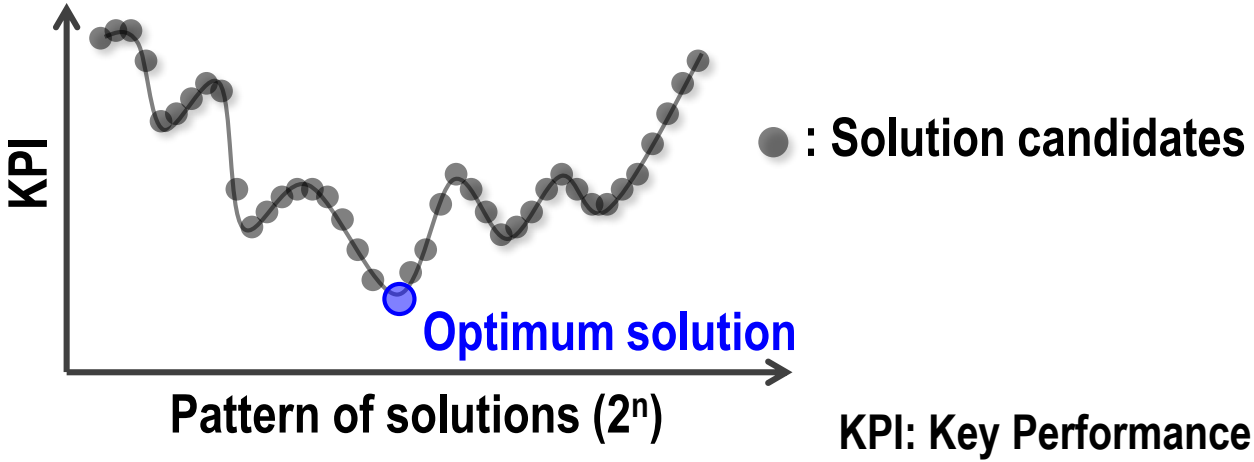


Robotics control

Combinatorial optimization problem

HITACHI
Inspire the Next

- Problem to explore an optimum solution for minimum KPI in given conditions
- Enormous candidates of solution with large number of parameters: 2^n patterns

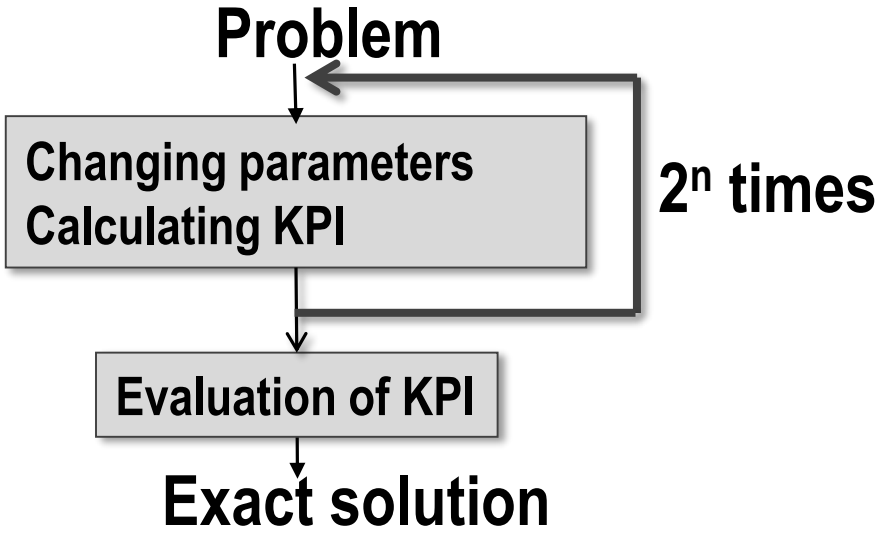


KPI: Key Performance Indicator

Solving combinatorial optimization problem

HITACHI
Inspire the Next

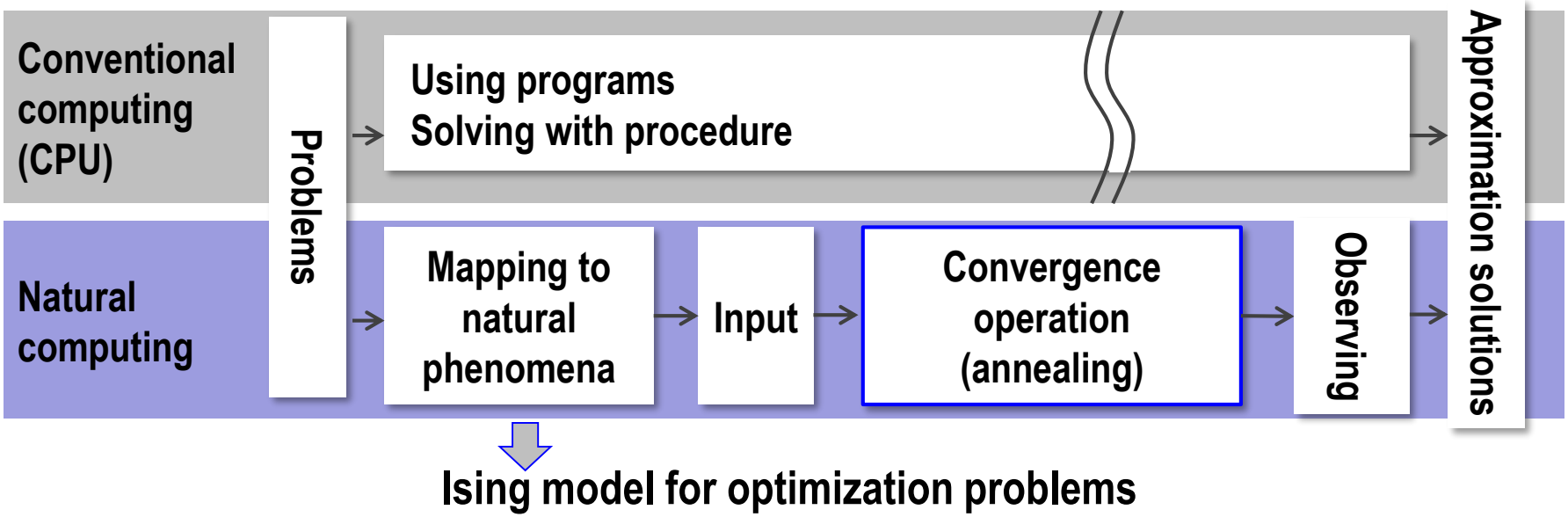
- Conventional computing calculating all KPIs and deciding combination of parameters for the best KPI
- Enormous calculation required when n is large
- Practically approximation algorithms used



New-paradigm, natural, computing

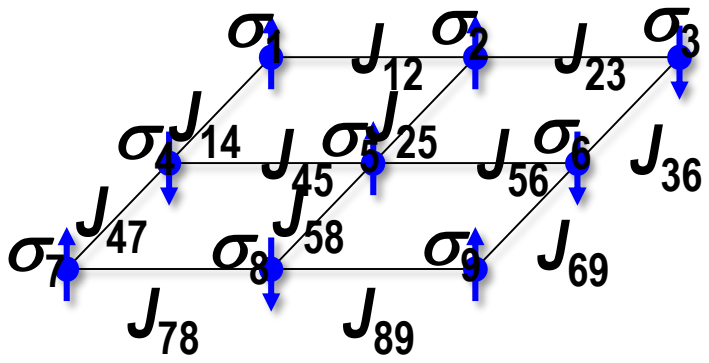
HITACHI
Inspire the Next

- Mapping problems to natural phenomena
- Solution acquired by convergence operation of natural phenomena
- Ising model used for optimization problems



Ising model

- Ising model: expressing behavior of magnetic spins, upper or lower directions
- Spin status updated by interaction between spins to minimize system energy



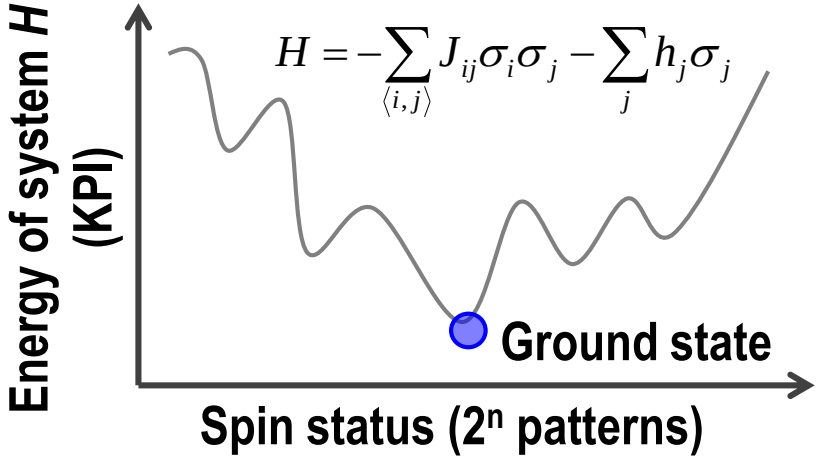
$$H = -\sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j - \sum_j h_j\sigma_j$$

- H: Energy of Ising model
- σ_i: Spin status (+1/-1)
- J_{ij}: Interaction coefficient
- h_j: External magnetic coefficient

Computing with Ising model: annealing machine

HITACHI
Inspire the Next

- Shape of landscape of Ising model energy same as KPI plot of combinatorial optimization problem
- By mapping original problems to Ising model, optimum solution acquired as ground state of model



n: number of spins

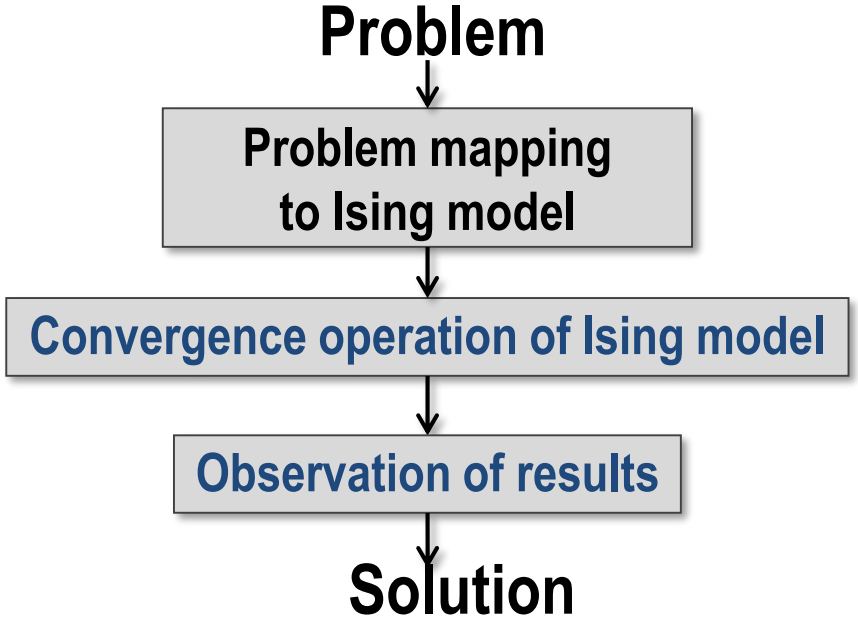
Correspondence of parameters

Ising model	Optimization problems
Energy H	KPI
Spin status σ_i	Control parameters
Interaction coefficient J_{ij}	Input data (sensor data, etc)

Natural computing using Ising model: annealing machine

HITACHI
Inspire the Next

- Mapping optimization problems to Ising model
- Convergence operation of Ising model
- Solution acquired by observing convergence results



Various implementation of annealing machines

HITACHI
Inspire the Next

- Quantum annealing machine, coherent Ising machine, CMOS annealing machine are proposed
- All machines based on Ising model

	Quantum annealing machine	Coherent Ising machine	CMOS annealing machine
Principle	Quantum annealing	Parametron	Classical annealing
Implementation	Superconductor	Laser oscillator	CMOS
Power	Large for cooling	Good	Better
Scalability	2kbit (2017)	2kbit (2016)	100kbit (2018)
Topology	Sparse	All to all	Sparse

Comparison of annealing machines

HITACHI
Inspire the Next

- For edge: Low cost, real time, room temperature, low power
- For cloud: High speed, low power, large scale, expandability

Method	Conventional computer	Quantum annealing machine	CMOS annealing machine
Devices	CPU	Superconductor	Semiconductor (CMOS)
Approach	Digital	Quantum bit	Digital
Calculation time	×	◎	○
Temperature	Room temp.	15mK	Room temp.
Operation power	10-1,000W	15,000W (with cooling)	0.05W
Scalability	Enable	512 ('12) → 2048 ('17)	100kbit (’18)
Expandability	Enable	-	Multiple-chip

Outline

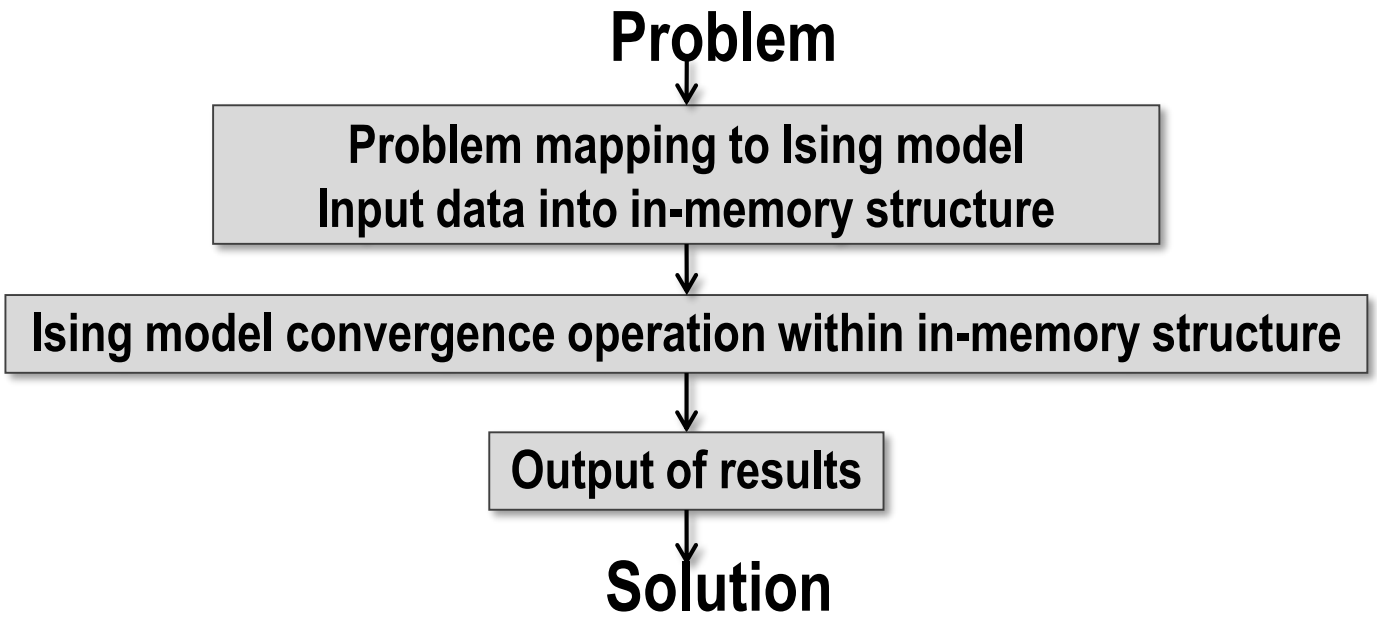
HITACHI
Inspire the Next

- Motivations
- **Overview of CMOS annealing machine**
- Prototypes of CMOS annealing machine
- Related necessary technologies
- Conclusion

In-memory computing for Ising-model computing

HITACHI
Inspire the Next

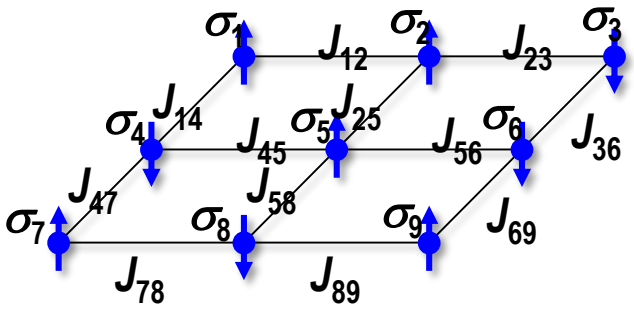
- In-memory structure suitable for mimicking Ising model convergence operation
- No memory access required in convergence operation



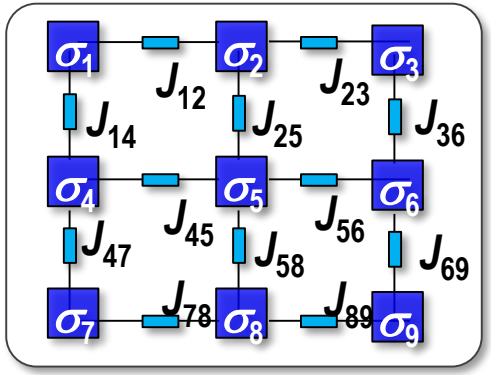
In-memory computing for Ising model

HITACHI
Inspire the Next

- Mimicking physical Ising model with in-memory structure
- Spin status updated by logic circuits implemented in memory



CMOS circuit mimicking Ising model

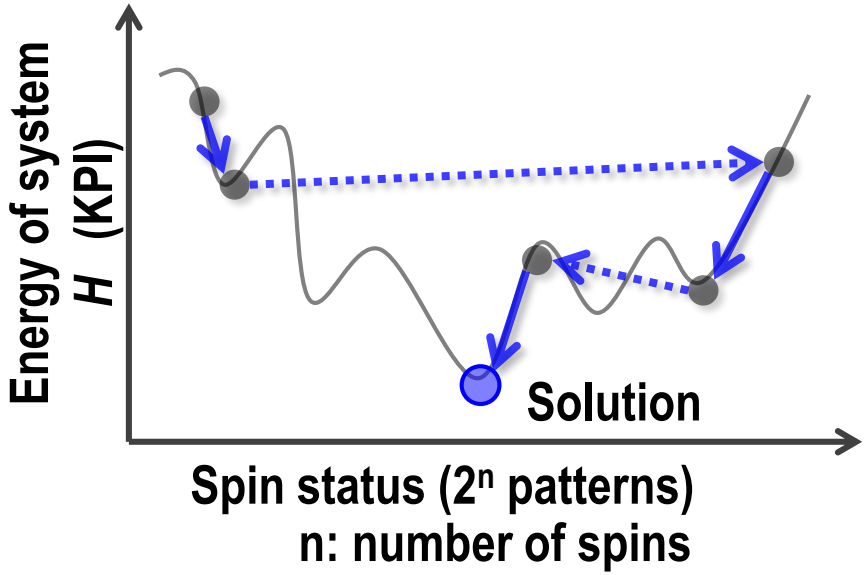


Ising model	CMOS
Spin status σ_i, +1/-1	Memory "1"/"0"
Interaction J_{ij}	Interaction coefficient: memory Interaction operation: logic circuits

CMOS annealing

HITACHI
Inspire the Next

- Only digital operation, spin status stuck at local minimum status
- To avoid local minimum sticking, random status transition used
- Optimum solution not always acquired



- Transition to lower energy (adjacent spin interaction) →
- Avoidance of local minimum (random transition) ⇢

Rule of spin status update

HITACHI
Inspire the Next

- Spin status updated to lower Ising model energy
- Coefficient +: same direction
  Coefficient - : opposite direction
- Majority of adjacent spins effect accepted

$$H = -\sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j - \sum_j h_j\sigma_j$$

Spin status:

$J_{ij} > 0$: same direction

$J_{ij} < 0$: opposite direction

Spin update rules

Next spin status:

in case a>b, σ_5=+1

in case a<b, σ_5 =-1

in case a=b, σ_5 =+1 or -1

a=number of (+1, +1) or (-1, -1)

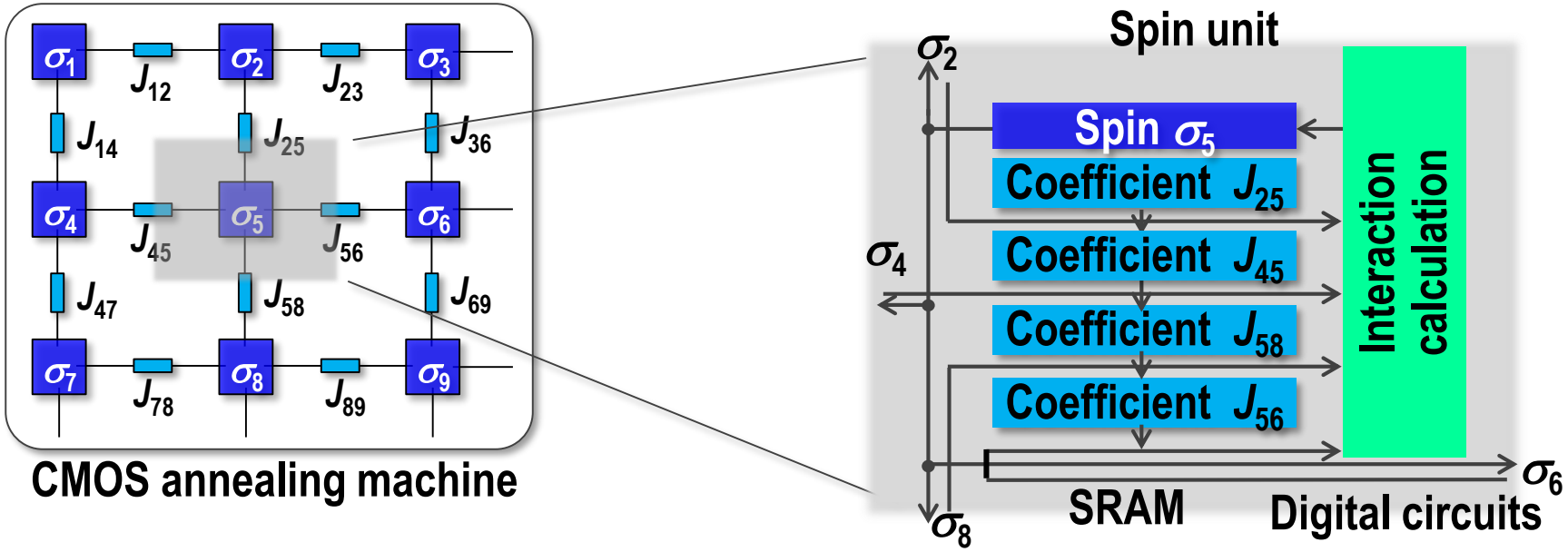
b=number of (+1, -1) or (-1, +1)

(value from adjacent spin, coefficient)

In-memory structure of CMOS annealing machine

HITACHI
Inspire the Next

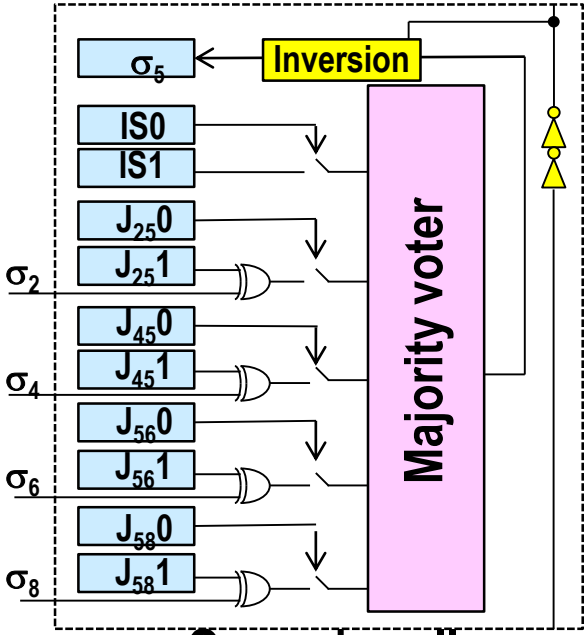
- Spin-update rule achieved by majority voter
- Each spin has update circuit and updated in parallel
- Short calculation time even when number of spins is large



Circuit structure for 1 spin unit

HITACHI
Inspire the Next

- Next spin status digitally calculated
- Majority voter circuits for efficient calculation



One spin cell

$$H = -\sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j - \sum_j h_j\sigma_j$$

Spin update rules

Next spin status:

in case a>b, σ_5=+1
in case a<b, σ_5 =-1
in case a=b, σ_5 =+1 or -1

a=number of (+1, +1) or (-1, -1)
b=number of (+1, -1) or (-1, +1)
(value from adjacent spin, coefficient)

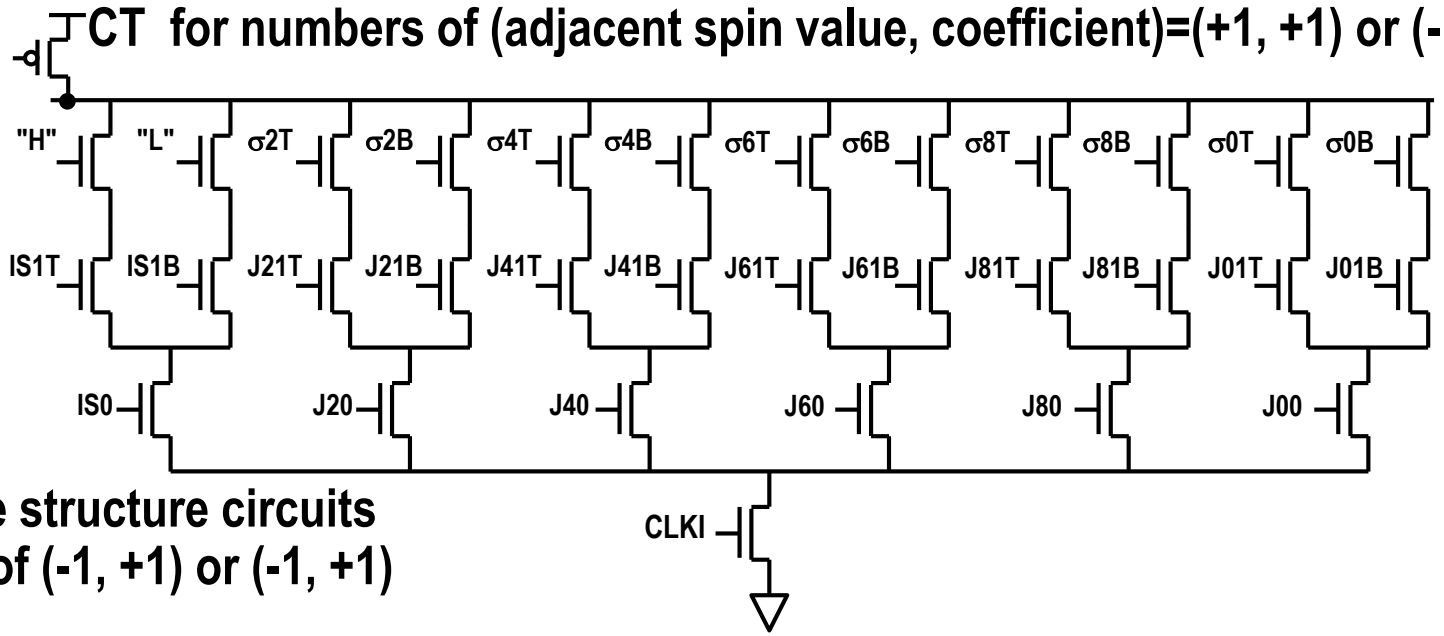
SRAM memory cell

EXOR and majority voter circuit

HITACHI
Inspire the Next

- EXOR and sum of results of EXOR ("0" / "1") calculated by sum of currents
- Voltage of common lines evaluated by sense amplifiers

Precharge



for numbers of (adjacent spin value, coefficient)=(+1, +1) or (-1, -1)

CB with same structure circuits
for numbers of (-1, +1) or (-1, +1)

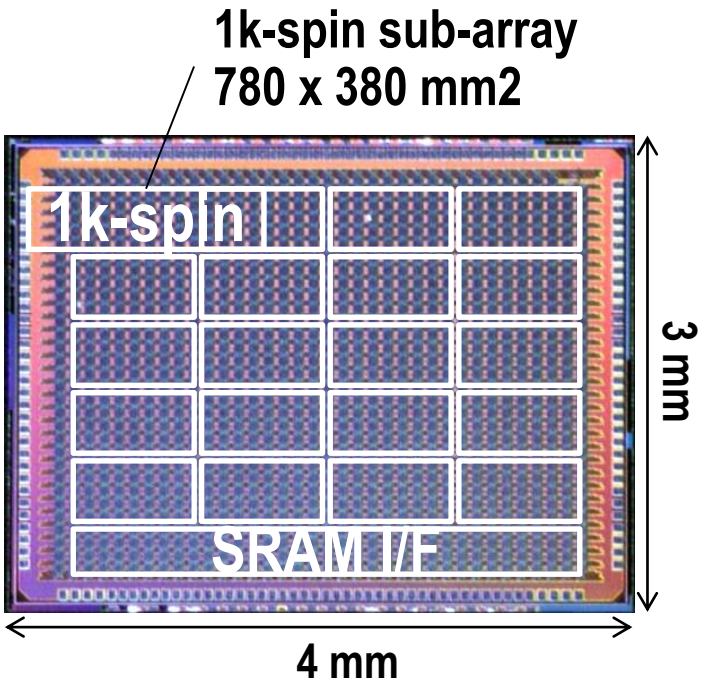
Outline

HITACHI
Inspire the Next

- Motivations
- Overview of CMOS annealing machine
- **Prototypes of CMOS annealing machine**
- Related necessary technologies
- Conclusion

Fabrication results: CMOS annealing chip

HITACHI
Inspire the Next



Items	Values
Number of spins	20k (80 x 128 x 2)
Process	65 nm
Chip area	4x3=12 mm^2
Number of SRAM cells	260k bits Spin value: 20k bits Interaction coefficient: 240k bits
Memory IF	100 MHz
Interaction speed	100 MHz
Operating current of core circuits (1.1 V)	Write: 2.0 mA Read: 6.0 mA Interaction: 44.6 mA

1st generation prototype

HITACHI
Inspire the Next

- Ising chips installed on computing node
- FPGA installed to control Ising chip
- Accessed via LAN cable from PCs/servers

Computing node

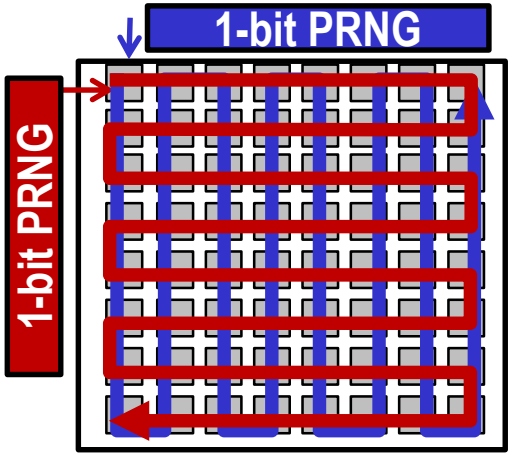
Configuration	2U rack mount
Operation frequency	100 MHz
Number of spins	40k (2chips)
OS	Linux

Random transition: random numbers

HITACHI
Inspire the Next

- Two sequences of 1-bit random numbers input, propagated, and evaluated at spin interaction
- Only two PRNGs provide randomness to whole chip

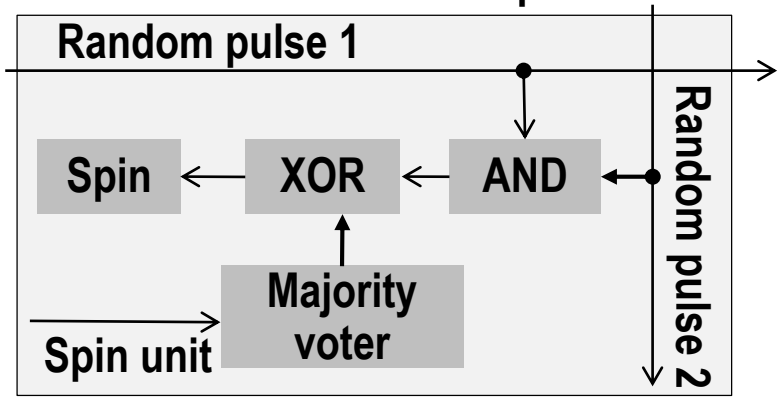
Zigzag paths



PRNG: Pseudo Random Number Generator

Spin inversion using random pulse

➢ Output of the interaction circuit will be inverted if both of the random pulses are '1'

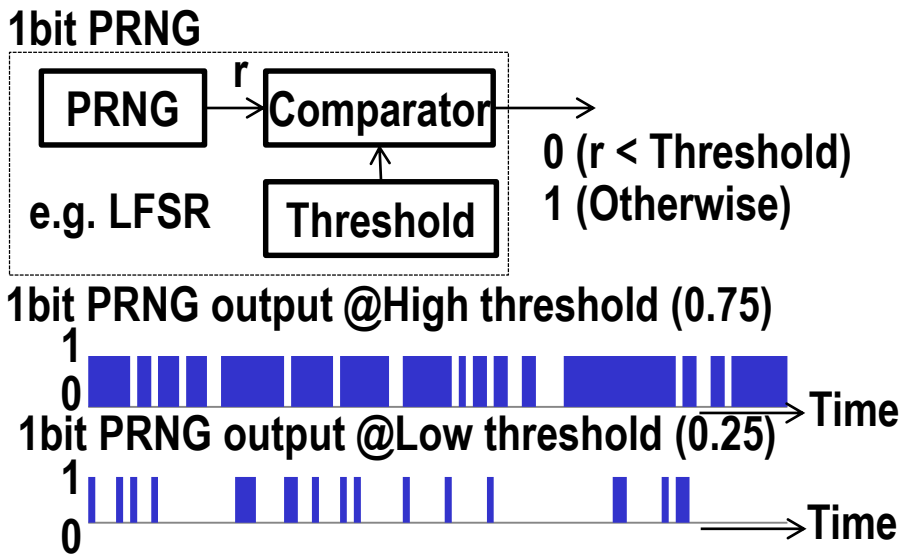


Random pulse control

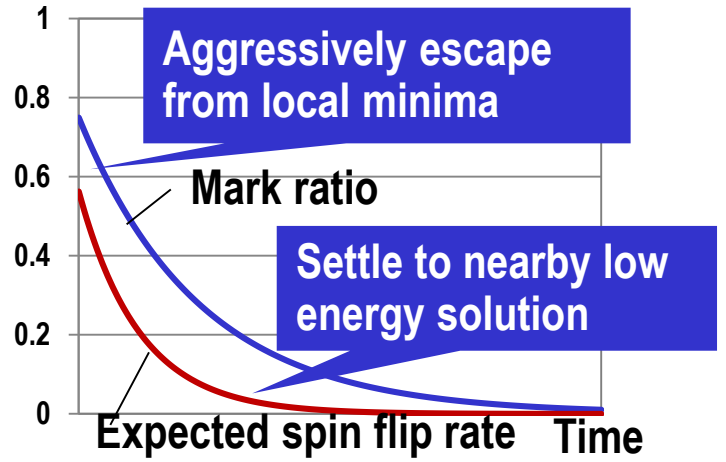
HITACHI
Inspire the Next

- Spin flip rate gradually lowered for annealing
- Spin flip rate controlled by mark ratio of 1bit PRNGs

Controlling mark ratio of 1bit PRNG



Mark ratio during ground-state search

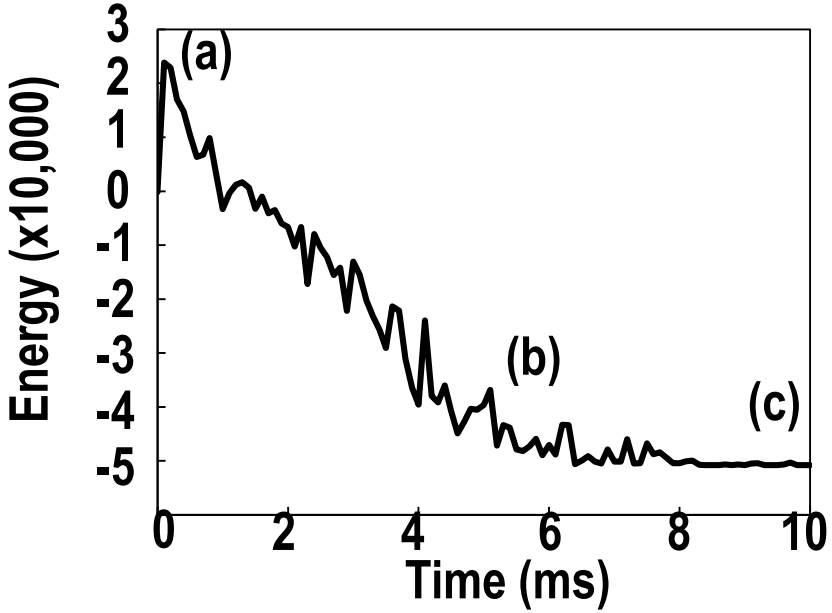


➢ Decreasing mark ratio has similar effect to cooling schedule in Simulated Annealing

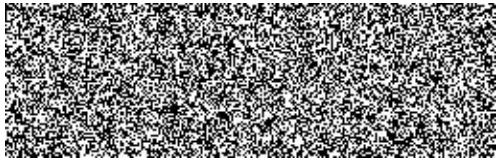
Measurement results of CMOS annealing

HITACHI
Inspire the Next

- MAX-cut problem, NP-complete problem, with 20k-spin solved
- Coefficient values set "ABC" appeared at optimum
- Optimum solution not always acquired



(a)



Initial state

(b)

5 ms (500,000 steps)

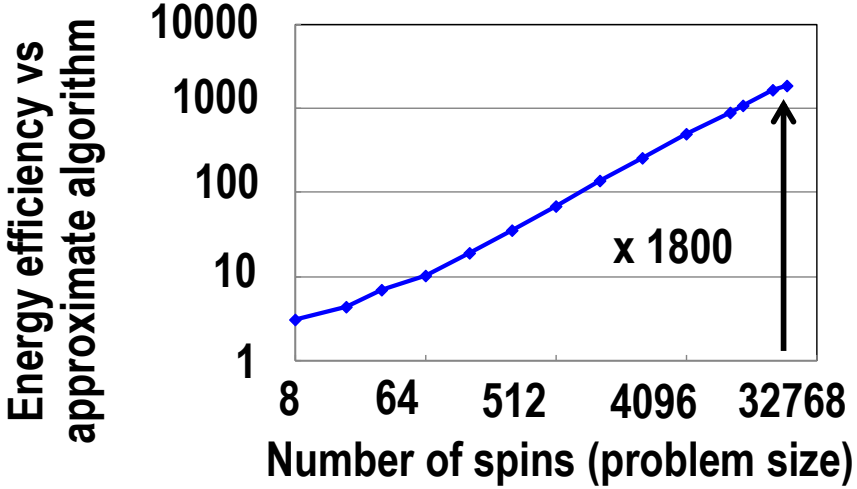
(c)

10 ms (1,000,000 steps)

Operating energy

HITACHI
Inspire the Next

- 1,800 times higher energy efficiency for 20k spin problem than approximation algorithm on CPU



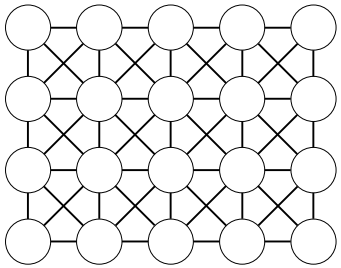
Conditions: Randomly generated Maximum-cut problems, energy for same accuracy solution Ising chip: VDD=1.1 V, 100-MHz interaction, best solution among 10-times trial is selected. Approximation algorithm: SG3(*) is operated on Core i5, 1.87 GHz, 10 W/core.

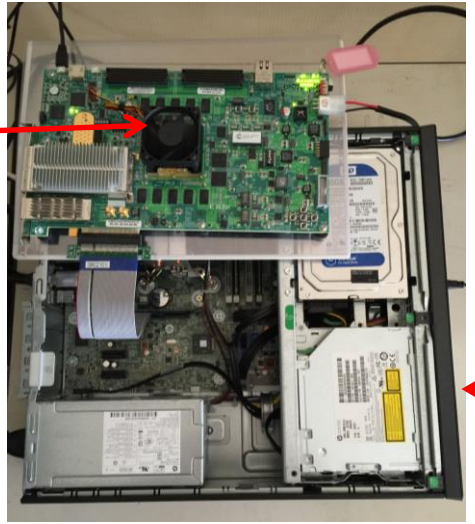
2nd generation prototype

HITACHI
Inspire the Next

- For software development, FPGA prototype used
- Various structures for trials (various topology, various bit number of coefficient)

Reconfigurable FPGA used to trial various structures (King's graph)





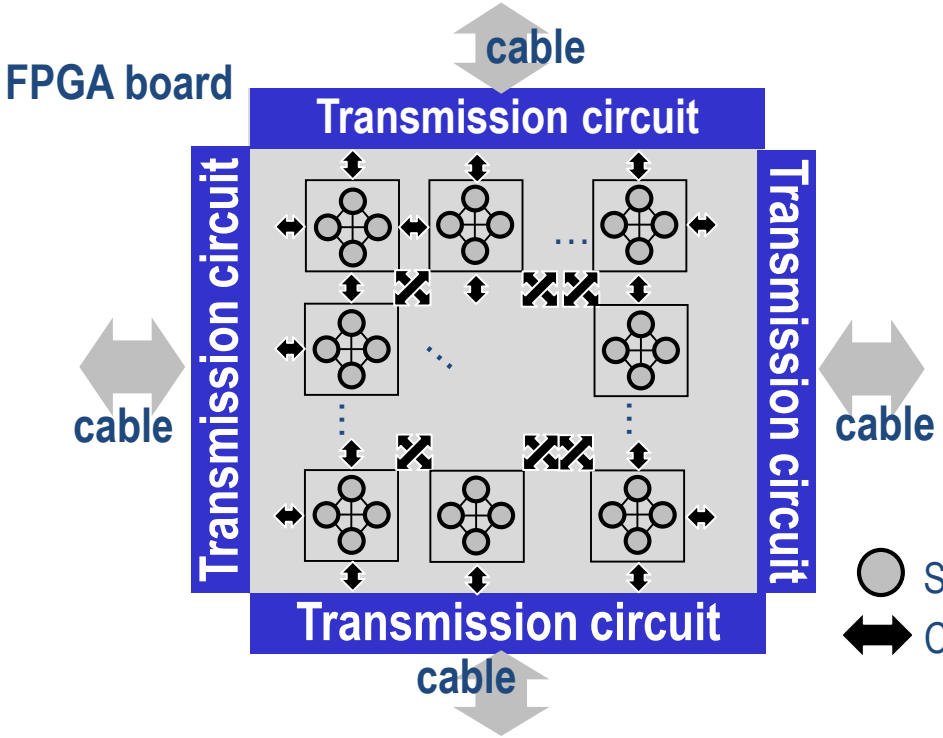
Control PC
- FPGA control
- Graph embedding algorithm

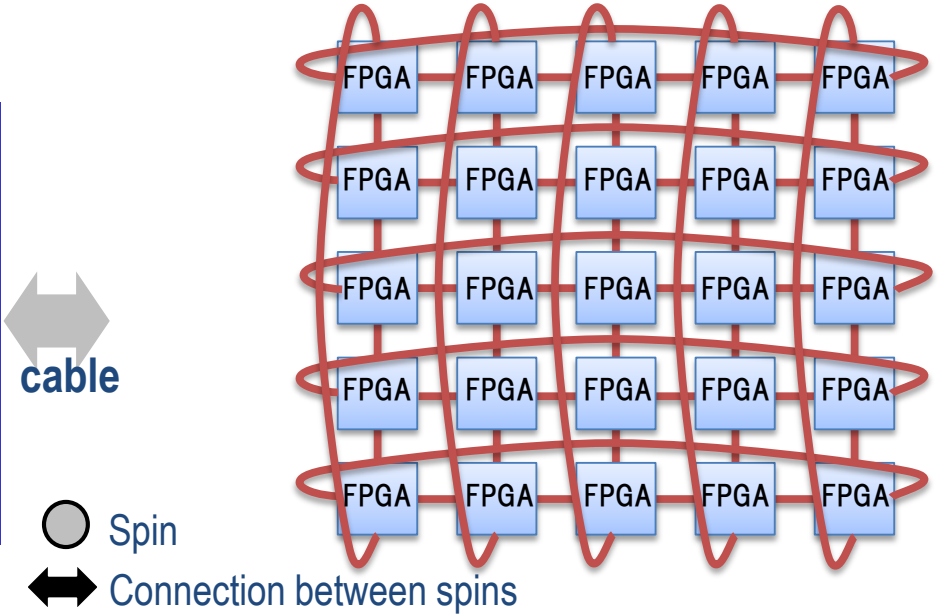
FPGA: Field Programmable Gate Array

Scalable inter-FPGA connectivity

HITACHI
Inspire the Next

- Operate as a large-scale machine using chip-to-chip connection
- Local transmission for scalable connection of many FPGA boards



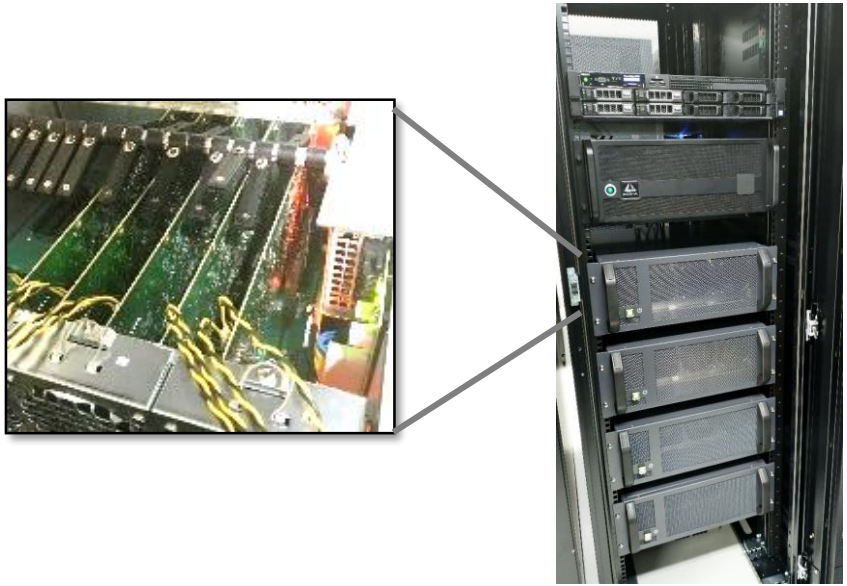


2nd generation 100kbit prototype

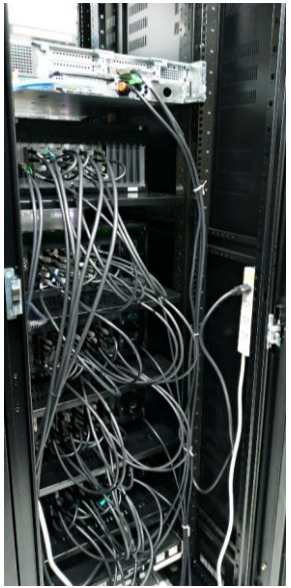
HITACHI
Inspire the Next

- Connect 5x5 FPGAs and demonstrate 100k bit operation with the largest number of parameters

Complete system(front)



Complete system(back)



100kbit CMOS annealing machine

HITACHI
Inspire the Next

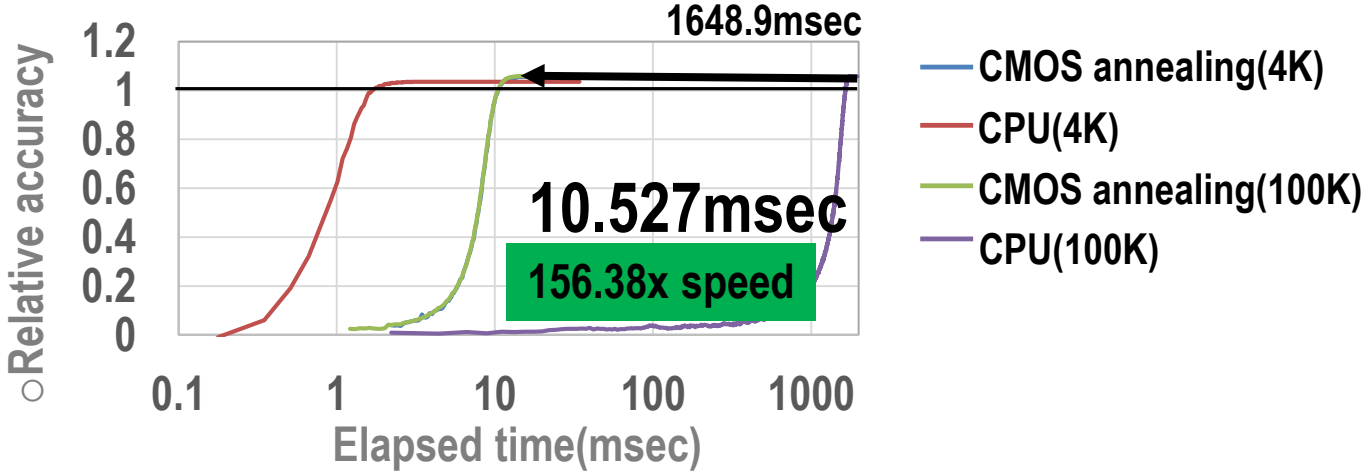
Item	Value
Number of parameters	102,400
Connection of Ising model	Partially coupled
Number of FPGA boards	25 (4,096 per 1FPGA)
FPGA operating frequency	82.5MHz
Parameter resolution	5bit (±15)
FPGA board	Xilinx® UltraScale®
Interconnect of FPGA	Xilinx® Aurora®

Performance comparison

HITACHI
Inspire the Next

- 156x speed improvement compared to conventional computer
- Larger number of parameters, greater speed improvement

Speed comparison with conventional machines



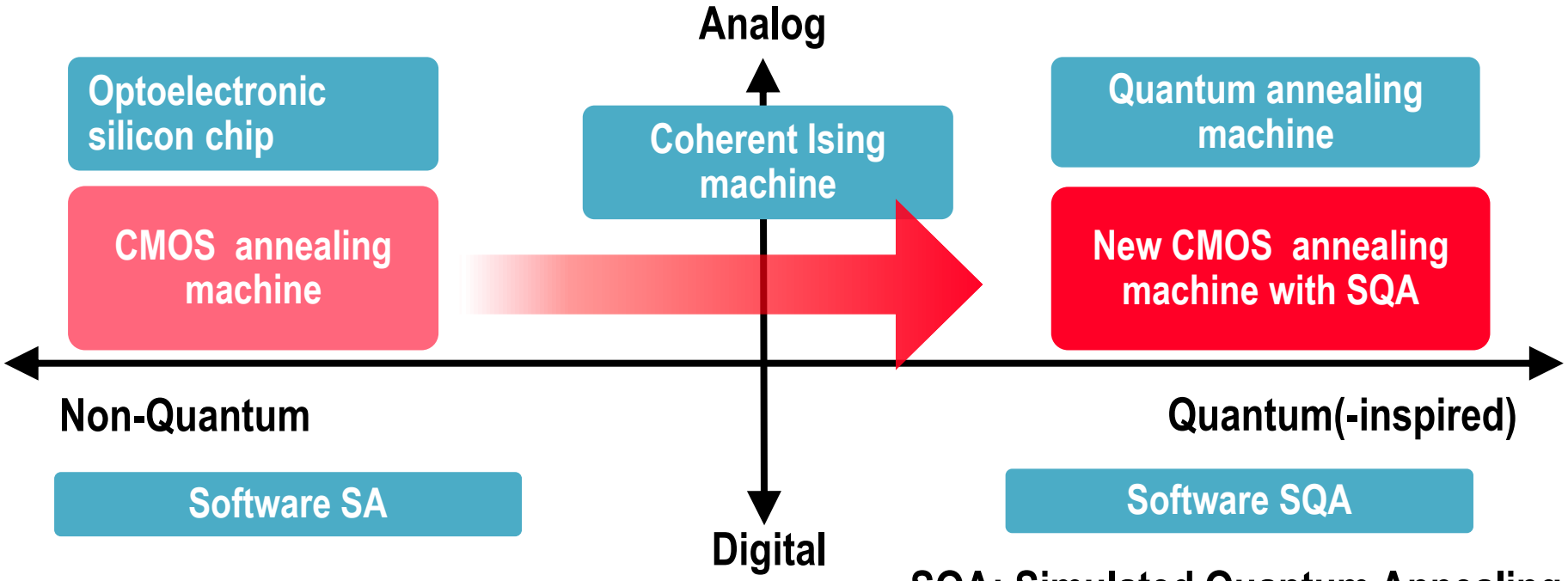
<u>Evaluation condition</u>

Execute ground state search of randomly generated Ising model with SA on CMOS annealing machine and conventional computer (CPU), and compare time to reach reference accuracy obtained by existing algorithm

Annealing technique for higher precision

HITACHI
Inspire the Next

- Map of annealing machine and algorithms

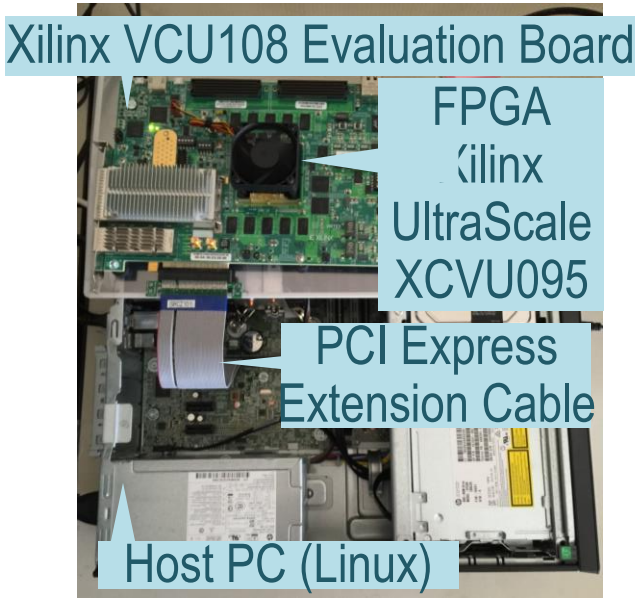


SQA: Simulated Quantum Annealing

3rd generation prototype

HITACHI
Inspire the Next

- Include SQA algorithm to improve performance
- Improve solution accuracy by incorporating pseudo quantum effects



Item	Value
Algorithm	Simulated Quantum Annealing (SQA)
Implementation	FPGA
Topology	King graph
# of spins	2500 (50×50)
# of replicas	32
Coefficient	8 bits (0, ±1, …, ±127)
Interaction	50 MHz

Simulated annealing

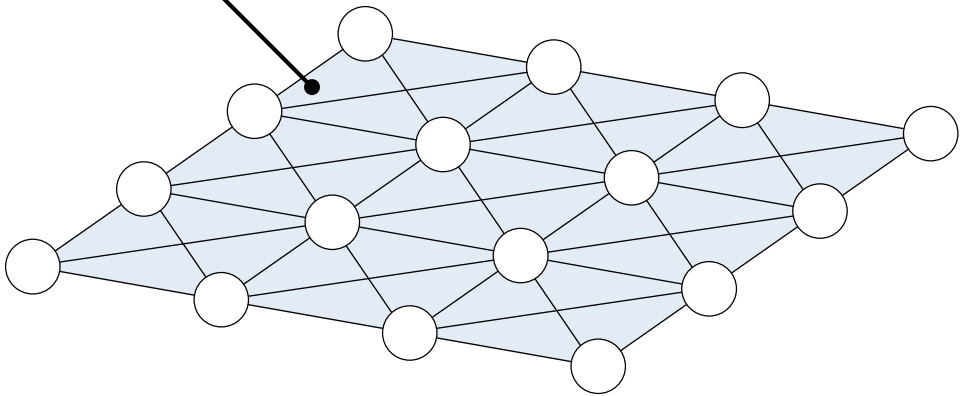
HITACHI
Inspire the Next

- **SA consists of discrete-time Markov processes that converge to Boltzmann distribution**

Hamiltonian for *Simulated Annealing*

$$H_{\mathrm{SA}} = -\sum_{(i,j)} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i$$

Spin update by
Metropolis algorithm



Simulated Quantum Annealing

HITACHI
Inspire the Next

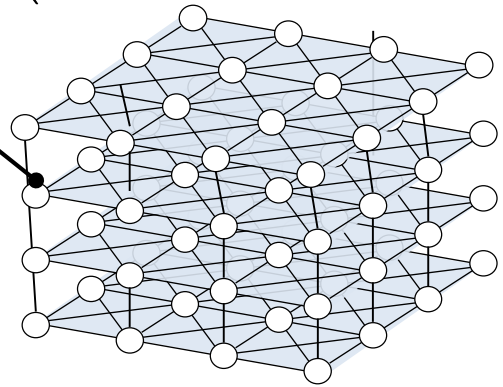
- Reproduce the quantum superposition effect by the SQA method
- Digital circuit implements SQA method

Hamiltonian for *Simulated Quantum Annealing*

monotonically increasing

$$H_{\mathrm{SQA}} = -\sum_{k=1}^{M}\left(\sum_{(i,j)}\frac{J_{ij}}{M}\sigma_{i,k}\sigma_{j,k} + \sum_{i}\frac{h_i}{M}\sigma_{i,k} + \boxed{J^{+}}\sum_{i}\sigma_{i,k}\sigma_{i,k+1}\right)$$

Strength of coupling between replicas J^{+}

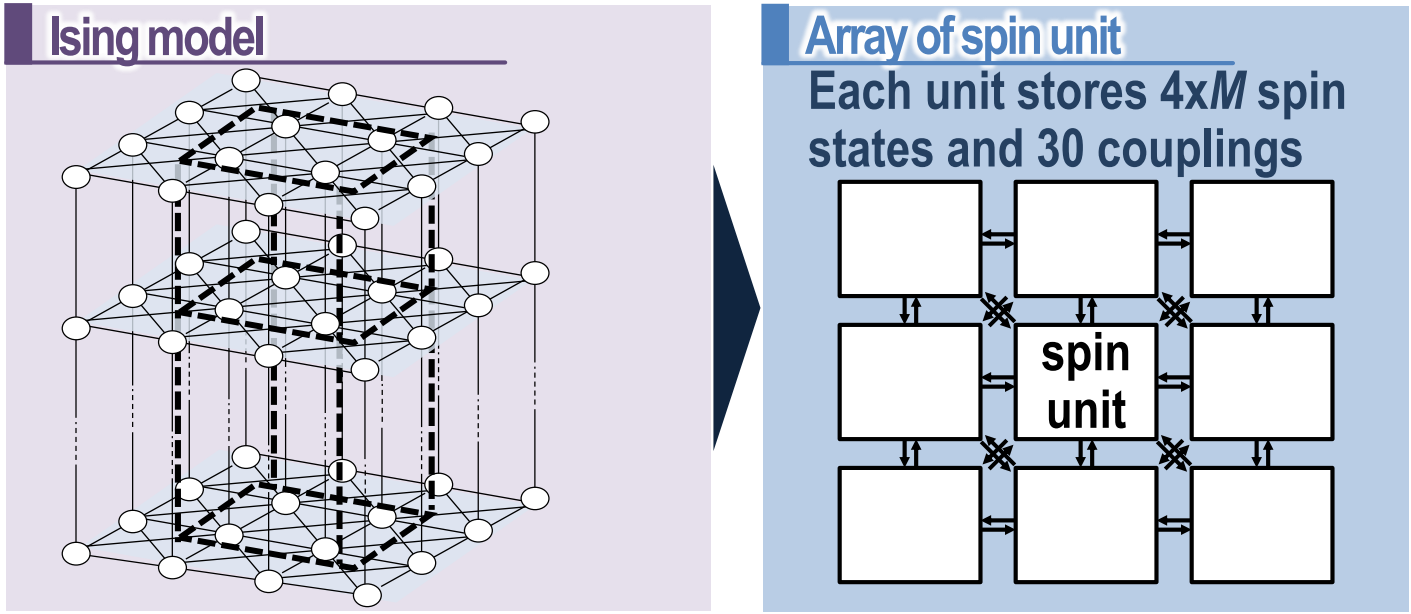


1st replica
2nd replica
⋮
*M*th replica

CMOS annealing machine based on SQA

HITACHI
Inspire the Next

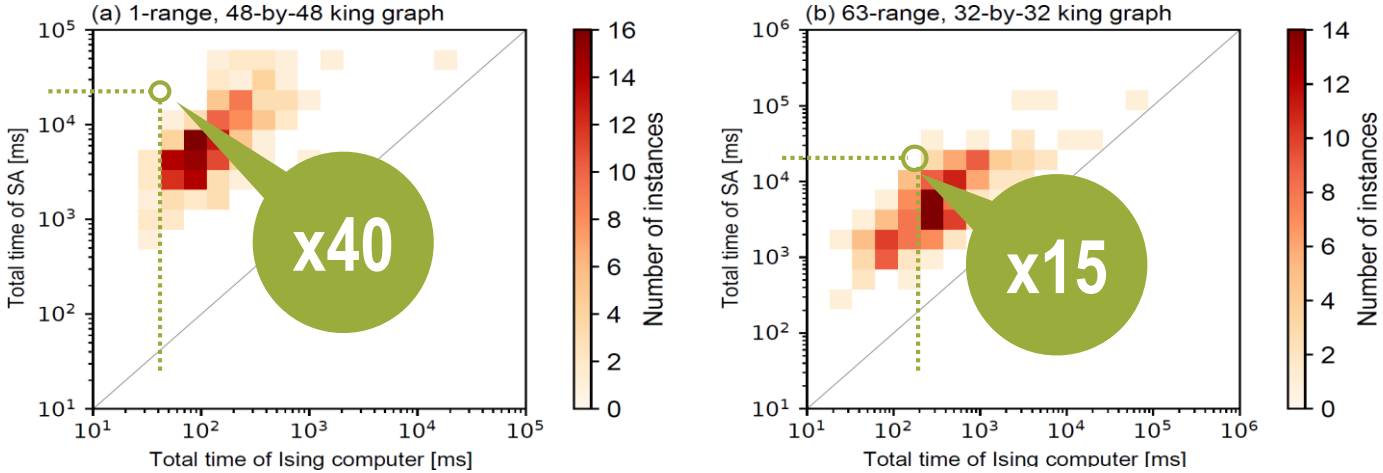
- Each replica has the same combination of couplings and biases
- Memory for interaction coefficients shared



Speed comparison for optimization problem calculation

HITACHI
Inspire the Next

- **Executes optimization processing (SA) 40 times faster than conventional methods**



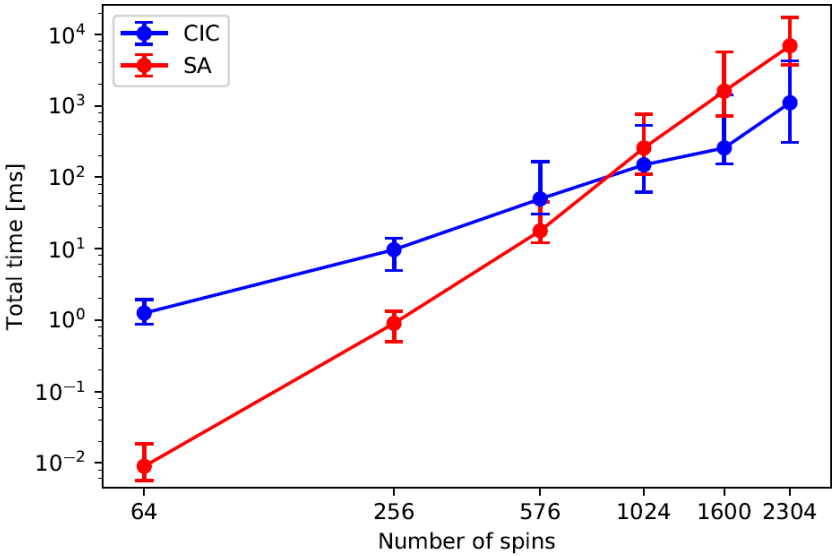
Note:

- Total annealing time means a time to obtain 99.9% solution with a probability of 99%
- 200 different random Ising models on a king graph are used
- We run the optimized SA program on a state-of-the-art CPU (Intel Core i7-6700K, 8 threads)

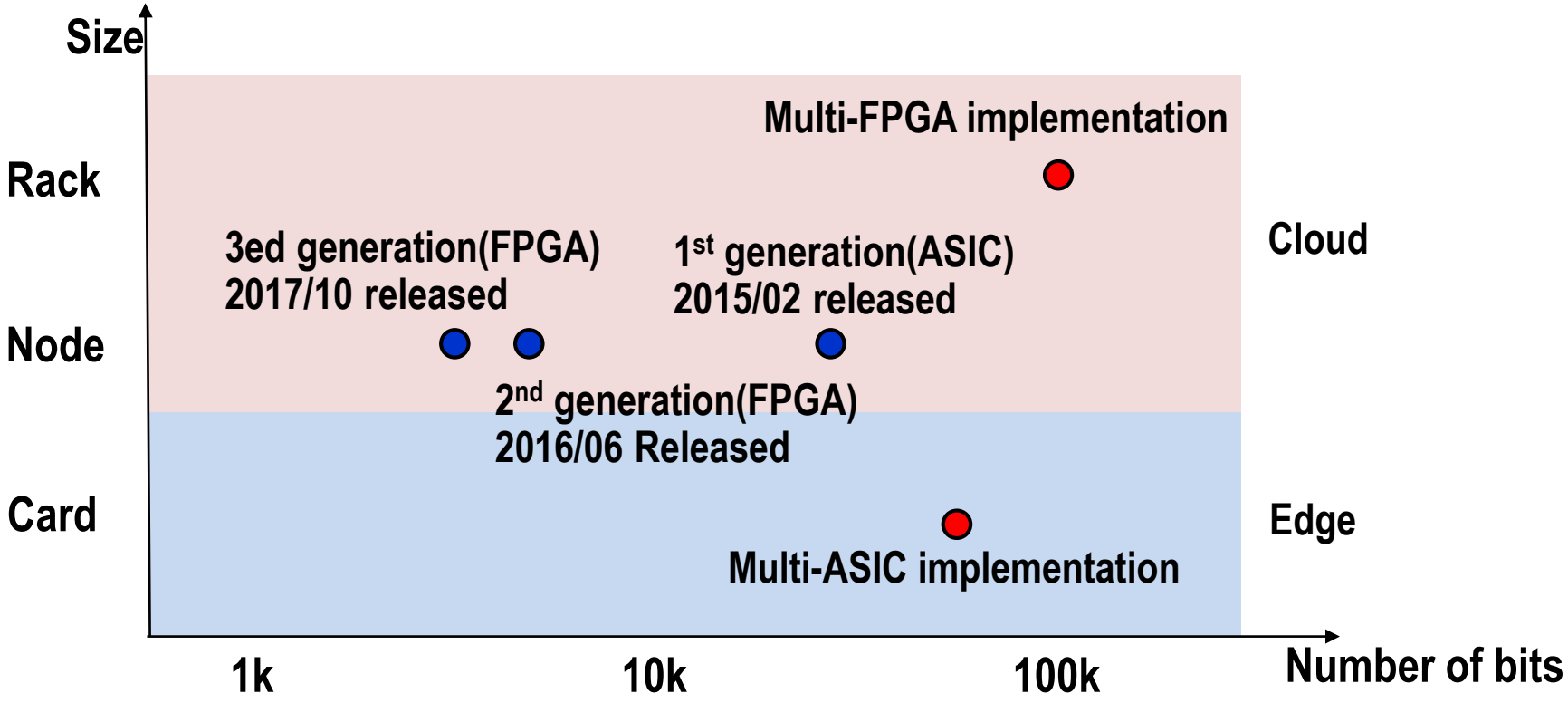
Performance of SQA based CMOS annealing machine

HITACHI
Inspire the Next

- Supremacy over software SA for large size problems
- Computational amount reduced:
  Contributing to higher speed, lower power consumption



For larger-scale implementation

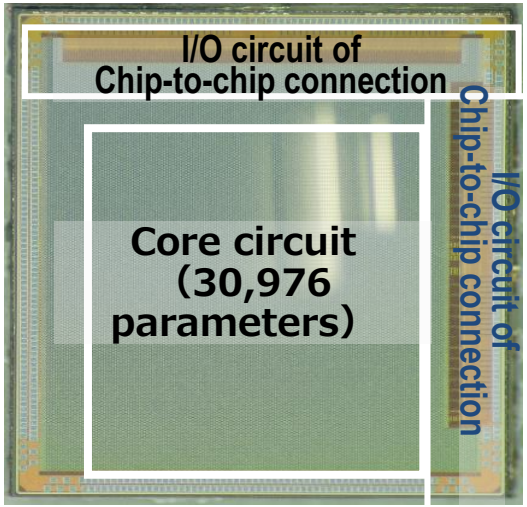


Card-size CMOS annealing machine for edge devices

HITACHI
Inspire the Next

- Prototype of 30-k spin annealing chip in 40 nm CMOS process
- Card sized CMOS annealing machine equipped with 2 chips
（Realized optimization calculation of about 60,000 parameters）

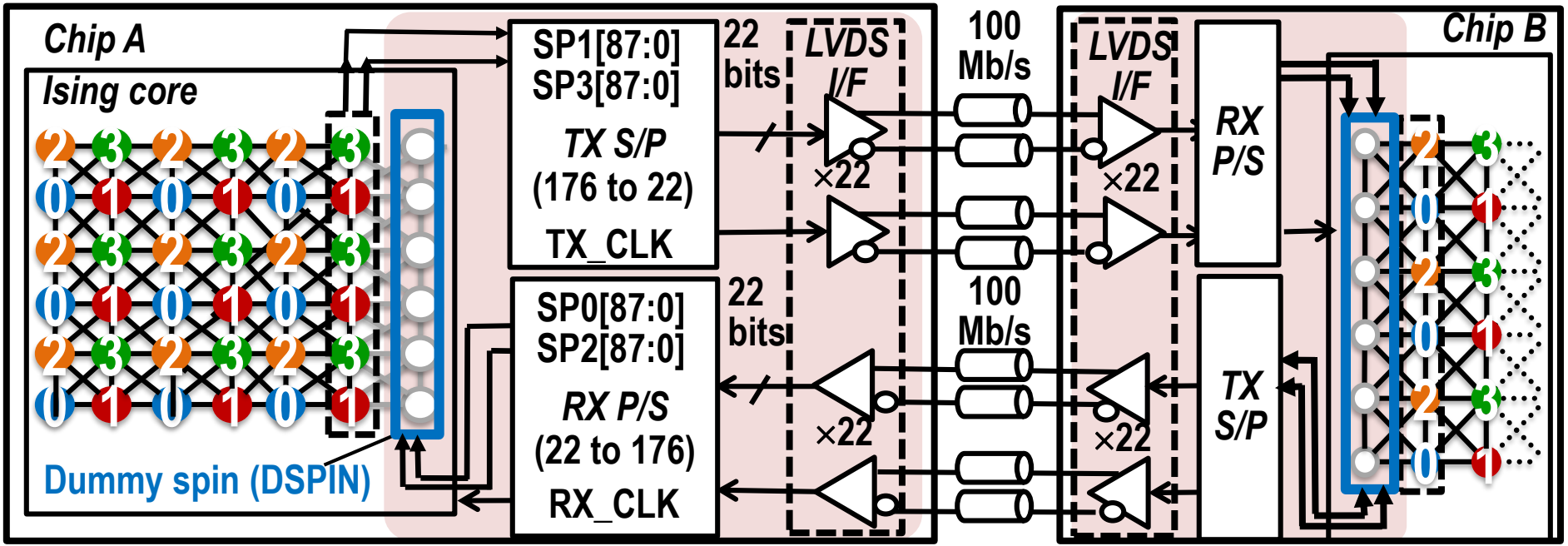
CMOS annealing
（2×30,976 parameters）



Proposed inter-chip Interface

HITACHI
Inspire the Next

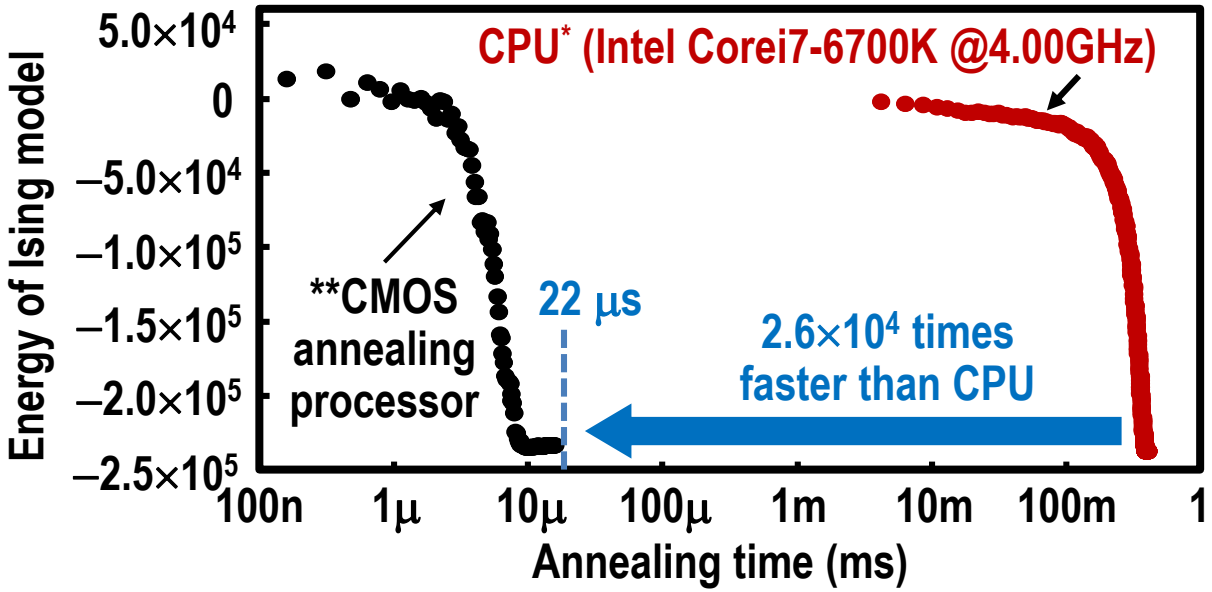
- 100-Mb/s LVDS I/F:2×88-bits data are split into 8×22-bits chunks
- Update values of dummy spin based on spin update rule



Annealing speed compared to CPU

HITACHI
Inspire the Next

- Fast annealing time: 22 μs (=22 clock cycles x 100 annealing steps)
- 2 x 30k spin system Max-Cut problem with randomly allocated coefficients

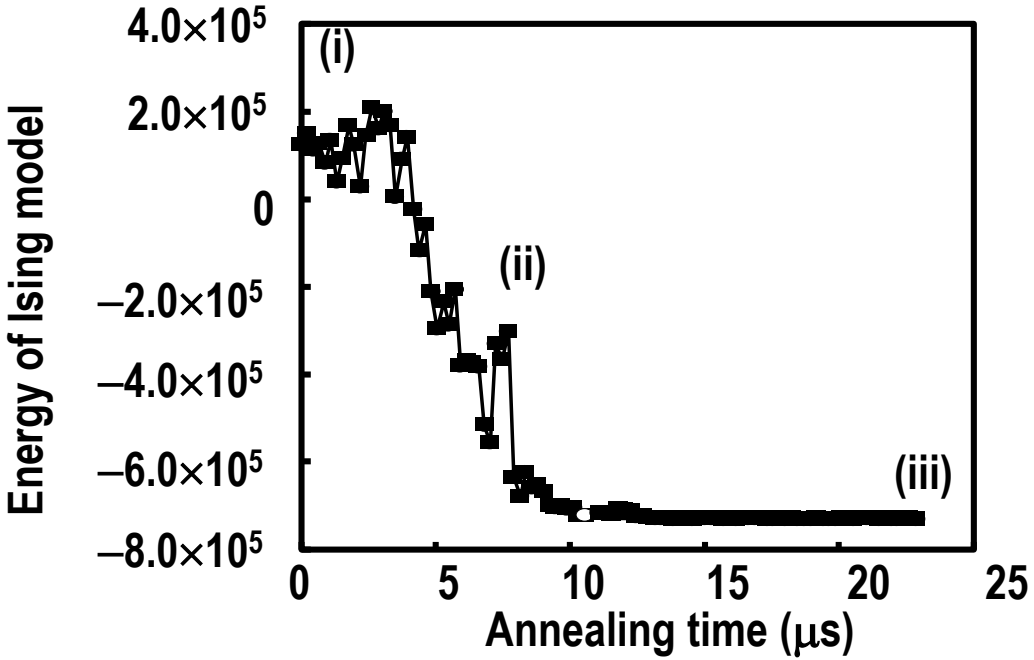


** Without I/O time (23.8 ms) through USB3.0 I/F

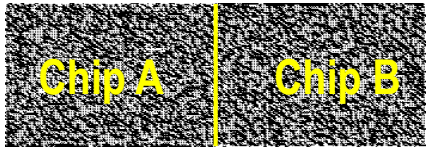
Two-chip operation for Max-Cut problem

HITACHI
Inspire the Next

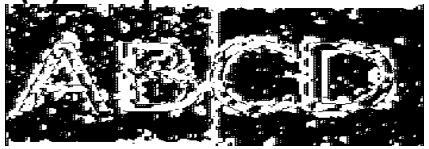
- 2 chip operation confirmed with max cut problem
- Border line between chips disappears in the final low-temperature state



(i) 0 μs (initial state)



(ii) 7.7 μs



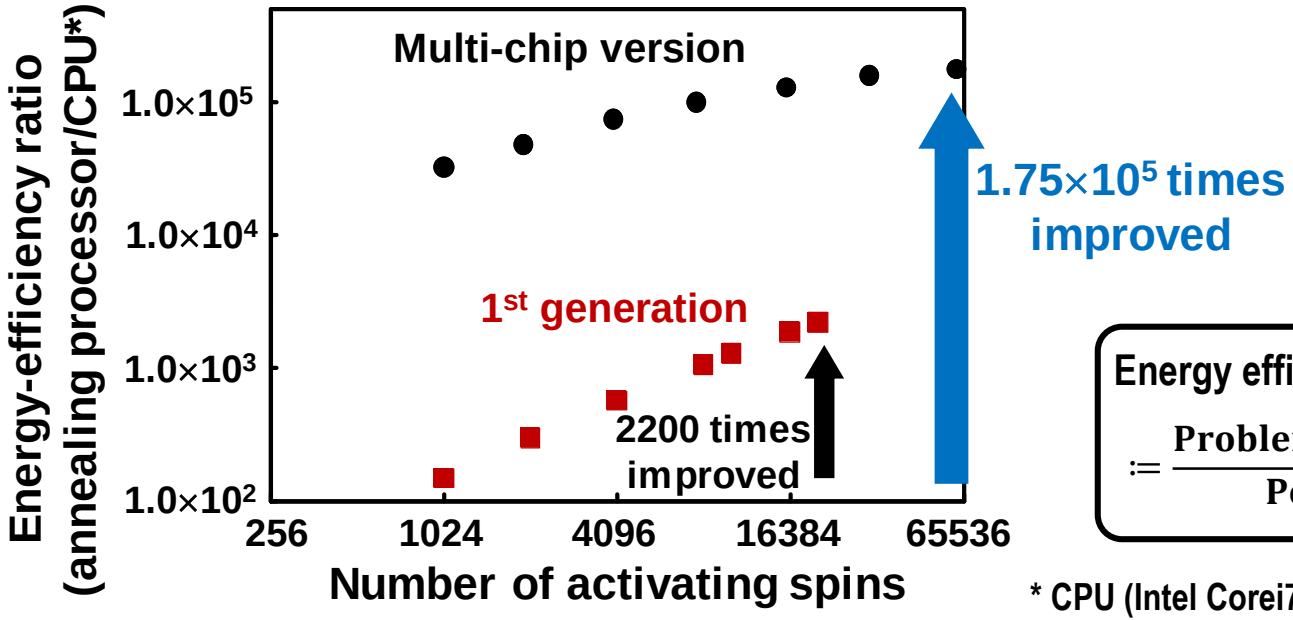
(iii) 21.8 μs (final state)

One pixel is one spin

Energy efficiency compared to CPU

HITACHI
Inspire the Next

- Energy efficiency improves with increasing number of spins
- 2x30k spin system: 1.75×10^5 higher than CPU



Energy efficiency

$$:= \frac{\text{Problem size/Calculation time}}{\text{Power consumption}}$$

* CPU (Intel Corei7-6700K @4.00GHz): SG3 algorithm

Comparison of annealing machines

HITACHI
Inspire the Next

- Multi-chip operation for larger scale confirmed
- Faster operation and higher energy efficiency achieved by new chip

	D-Wave 2000Q	1st gen.	Multi-chip version
Method	Quantum annealing	Simulated annealing	
Accuracy	Better	Not so good	Good
Implementation	Superconductor	65-nm CMOS	40-nm CMOS
Number of chips	1	1	2 (multichip in principle)
Number of spins	2k	20k	2×30k
Annealing time*	-	10 ms***	22 μs
Energy efficiency*, **	-	2200 times***	1.75×10^5 times

* Max-Cut problem is applied. ** These values are evaluated by comparing against running SA on CPU.
*** These values are evaluated under the condition similar to this work.

Benefits vs a traditional computation

HITACHI
Inspire the Next

- In-memory implementation for Ising model computing:
  Lower power operation with local-area process
  Higher operation speed with parallel operation

- Quantum-inspired algorithm:
  Higher speed and lower power consumption with smaller computational amount

- Multi-chip implementation by in-memory structure:
  Large-scale integration for larger problems

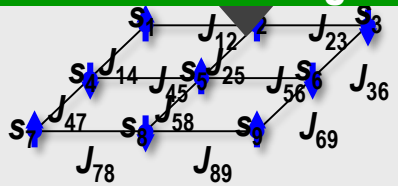
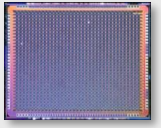
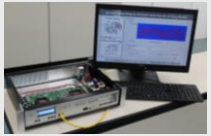
Outline

HITACHI
Inspire the Next

- Motivations
- Overview of CMOS annealing machine
- Prototypes of CMOS annealing machine
- **Related necessary technologies**
- Conclusion

How to solve "real" problems

HITACHI
Inspire the Next

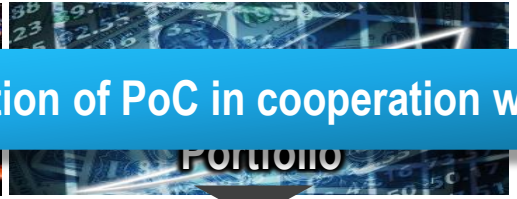
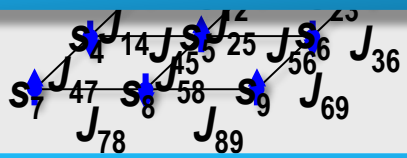
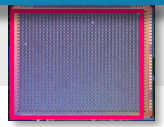
Systems	Supply chain	Finance	Traffic
application			
Software Problem Mapping Graph embedding	Max-flow problem	Knapsack problem	Shortast path problem
	Ising model on computer		J_{12}=x, J_{23}=y, ...
Hardware		Data IO (J_{ij}, h_i) Ising Computer	 

Cost function of optimization problems

Formulation of Ising model

How to solve "real" problems

HITACHI
Inspire the Next

Systems	Supply chain	Finance	Traffic
application			
Software Problem Mapping Graph embedding	Max-flow problem	Knapsack problem	Shortast path problem
	Ising model on computer		J_{12}=x, J_{23}=y, ...
Hardware	Ising Computer		

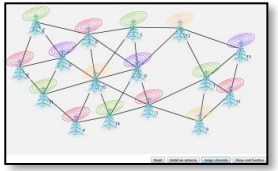
Implementation of PoC in cooperation with users

Promote open development in collaboration with universities etc.

Forming a community and creating a current flow for new computers

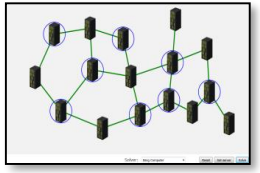
Application of CMOS annealing machine

HITACHI
Inspire the Next



1. Frequency allocation for wireless radio

2. Communication order allocation



3. Server security

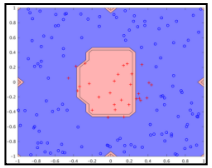
4. Image inpainting

5. Exploring explosion material

6. Image noise reduction

7. Facility allocation

8. Community core detection

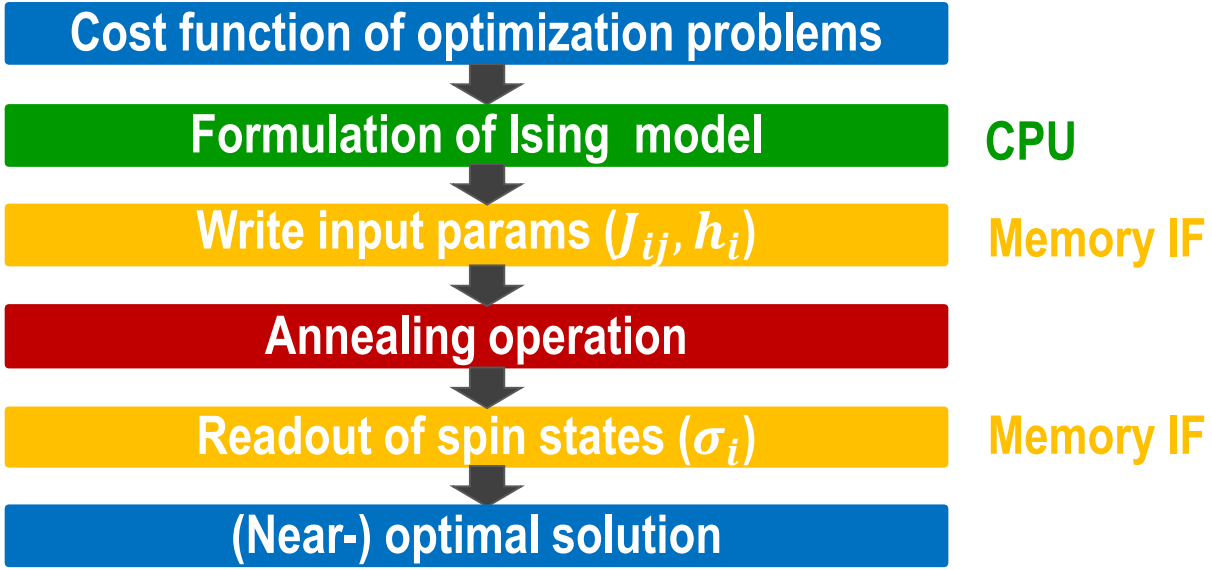


9. Machin learning (Boosting)

Operation procedure

HITACHI
Inspire the Next

- Pre-processing required for application-specific computing using Ising model
- Generated data easily input/output using memory IF



Number Partitioning Problem (NPP)

HITACHI
Inspire the Next

- Finding a division of a set into 2 subsets such that the sums of each subset are as close as possible

S = { 3, 1, 4, 15, 92, 65, 35, 89, 79, 32,
38, 46, 26, 43, 38, 32, 79, 50, 28, 84}

440

1, 65, 35, 38, 46, 26, 38, 79, 28, 84

439

3, 4, 15, 92, 89, 79, 32, 43, 32, 50

Mapping of NPP to Ising model

HITACHI
Inspire the Next

- **Formulation of NPP mapped to Ising model formulation**

Original formulation

$$\min_{\sigma_1,\cdots,\sigma_n \in \pm 1} \sum_i w_i \sigma_i$$

w_i : value of i-th element of the set S
σ_i : label of i-th element

Ising formulation of NPP

$$H(\sigma) = \sum_{i<j} w_i w_j \sigma_i \sigma_j$$

Mapping of constrained NPP to Ising model

HITACHI
Inspire the Next

- **Constrained optimization problem also mapped to Ising model formulation**

Original formulation

$$\min_{\sigma_1,\cdots,\sigma_n \in \pm 1} \sum_i w_i \sigma_i$$

w_i : value of i-th element of the set S
σ_i : label of i-th element

$$\text{subject to} \sum_i \sigma_i = 0$$

Same number of items for each group

Ising formulation of constrained NPP

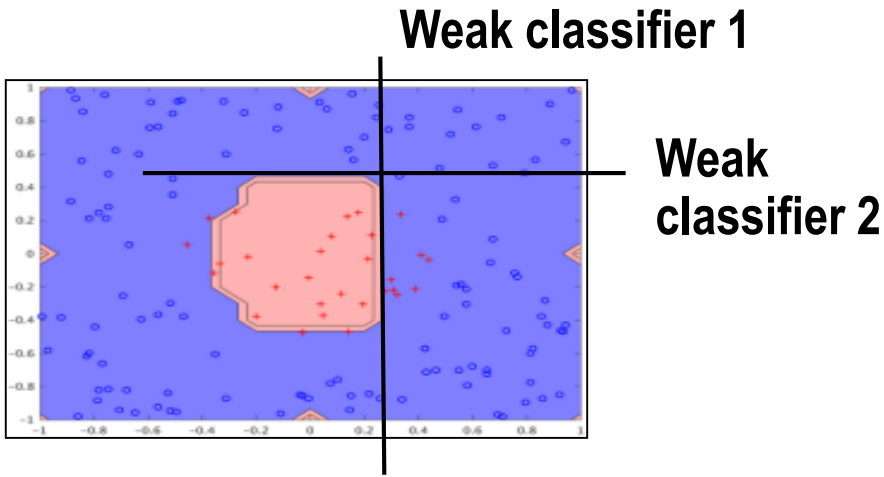
$$H(\sigma) = \sum_{i<j} w_i w_j \sigma_i \sigma_j + \lambda \left(\sum_i \sigma_i \right)^2$$

Machine learning using Boltzmann machine

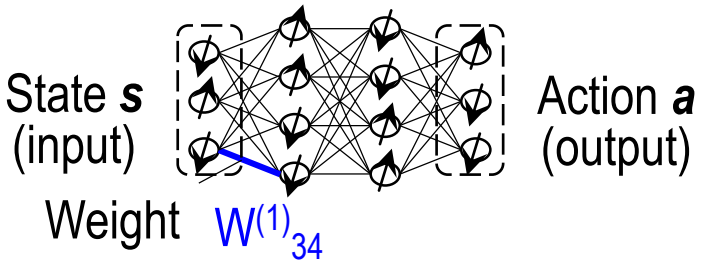
HITACHI
Inspire the Next

- **Boosting technique and reinforcement learning proposed**

Boosting:
finding optimum set of weak classifiers



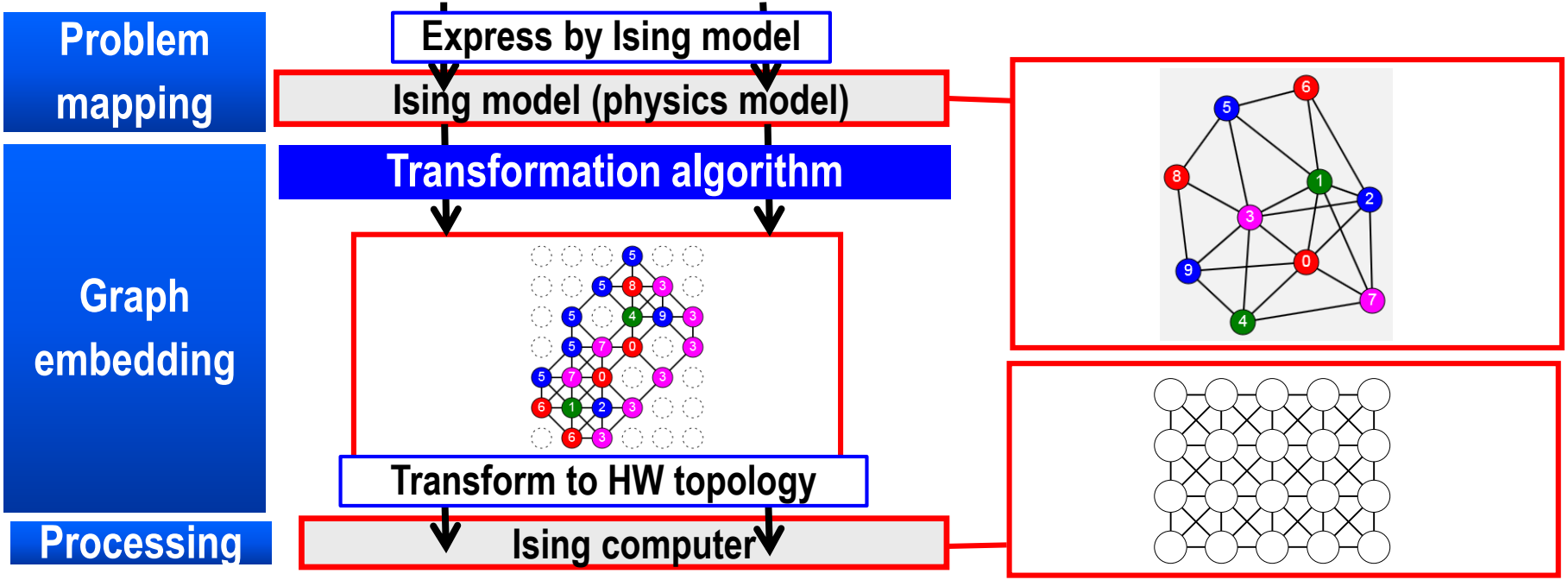
Reinforcement learning:
using Ising model as Boltzmann machine
instead of neural network



Graph embedding required for in-memory structure

HITACHI
Inspire the Next

- Many original problems have complex graph topology
- Graph embedding to use in-memory structure efficiently

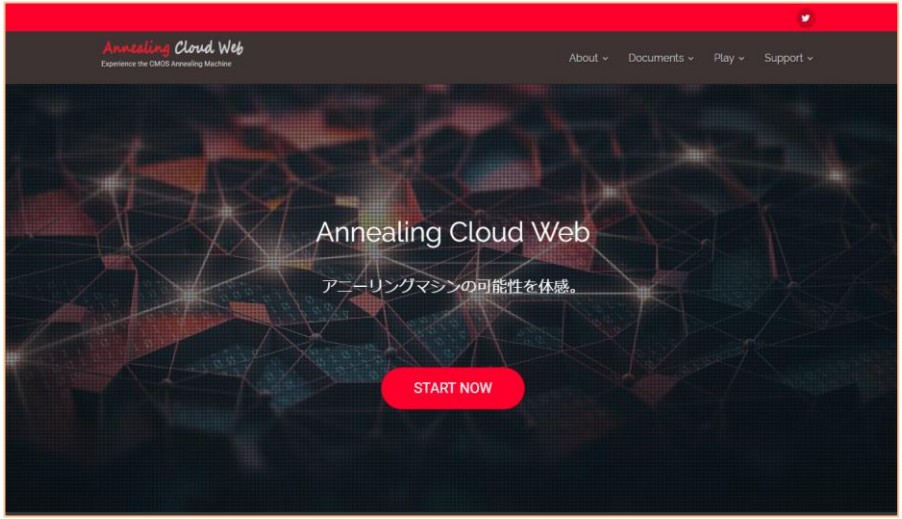


Cloud environment for beginners

HITACHI
Inspire the Next

- Access to CMOS annealing machine via internet
- Tutorials and demos to understand annealing machine

https://annealing-cloud.com/



This web page is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Program contest for annealing machine

HITACHI
Inspire the Next

- Program contest for graph embedding in 2017, and for lowering graph order in 2018
- From the junior high school students to the 50s, answer codes submitted from some countries

Result

1st（entry:973 submitted:296）

順位	ユーザ名	Problem 1	得点 / 時間
1	kurenai3110	761063 (61) 20070:40	761063 (61) 20070:40
2	siman	758012 (108) 20033:09	758012 (108) 20033:09

2nd（entry:446 submitted:126）

順位	ユーザ名	Problem 2	得点 / 時間
1	yosss	20090835 (14) 19916:18	20090835 (14) 19916:18
2	yowa	19375872 (16) 19857:05	19375872 (16) 19857:05
3	Aquarius	19341187 (40) 20088:02	19341187 (40) 20088:02

Postscript of participation

Award ceremony (information processing national convention)

Conclusion

HITACHI
Inspire the Next

- New paradigm, natural computing is necessary for system optimization with large amount of data.
- CMOS annealing machine for combinatorial optimization problem is proposed.
- 1st to 3rd, and multi-chip machines are developed to solve larger problems.
- Software techniques and related hardware techniques are necessary for practical application.

Acknowledgement

HITACHI
Inspire the Next

- Part of this work is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).