Featured Articles

Advanced Research into AI
Ising Computer

Masanao Yamaoka, Ph.D.
Chihiro Yoshimura
Masato Hayashi
Takuya Okuyama
Hidetaka Aoki
Hiroyuki Mizuno, Ph.D.

OVERVIEW: A major challenge facing AI is the enormous computational load it imposes, of which combinational optimization makes up a large part. Hitachi has devised a computing technology based on a new paradigm that is capable of solving combinatorial optimization problems efficiently using an Ising model, and has built a prototype 20k-spin Ising computer chip using a 65-nm process. An Ising chip represents a combinatorial optimization problem by mapping it onto an Ising model based on the spin of magnetic materials, and solves the problem by taking advantage of the system's natural tendency to converge. This convergence is implemented using a CMOS circuit. In addition to demonstrating its ability to solve combinatorial optimization problems and operate at 100 MHz, the prototype chip has been demonstrated to consume approximately 1,800 times less power to obtain the solution than would be required by a conventional computer with a von Neumann architecture.

INTRODUCTION

A major challenge facing artificial intelligence (AI) is the enormous computational load it imposes. This is because, in contrast to the conventional practice of mechanical execution of an algorithm defined by hand in the form of a program, AI learns automatically from data and uses this as the basis for realtime decision-making. Combinational optimization forms a large part of the heavy processing load associated with both the learning and decision-making steps. When an AI learns from data, for example, it needs to optimize the model parameters in order to minimize error. Similarly, when subsequently using the model for decision-making, the AI needs to optimize the decision parameters in order to maximize a performance function. In both cases, this combinational optimization requires finding the parameters that best satisfy the conditions out of a large number of possibilities, a problem that is difficult to solve efficiently using conventional computing practices.

Accordingly, Hitachi has developed a new concept in computing that can efficiently solve combinatorial optimization problems by using an Ising model, a statistical mechanics model that mimics the behavior of a magnetic material. Tests conducted on a prototype demonstrated that combinational optimization problems could be solved with an efficiency three orders of magnitude or greater compared to conventional computing practices. This article describes this Ising computer.

COMBINATIONAL OPTIMIZATION PROBLEMS

A combinatorial optimization problem involves finding a solution that maximizes (or minimizes) a performance index under given conditions. A characteristic of combinatorial optimization problems is that the number of candidate solutions increases explosively the greater the number of parameters that define the problem. As the number of parameters in AI computation is increasing, the number of candidate solutions to combinational optimization problems is expected to increase dramatically in the future.

The solution of combinatorial optimization problems using existing computing techniques involves calculating the performance index for all parameter combinations and then selecting the combination that results in the minimum performance index [see Fig. 1 (a)]. Because the number of combinations for a problem with n parameters is 2^n, a 1,000 parameter problem requires the performance indices to be calculated for 2^{1000} parameter combinations (roughly

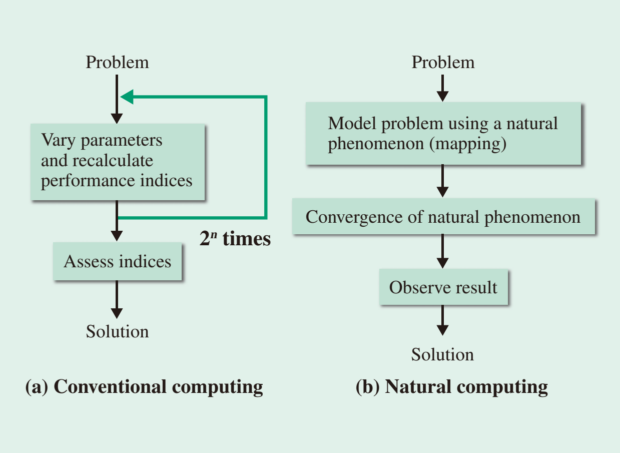
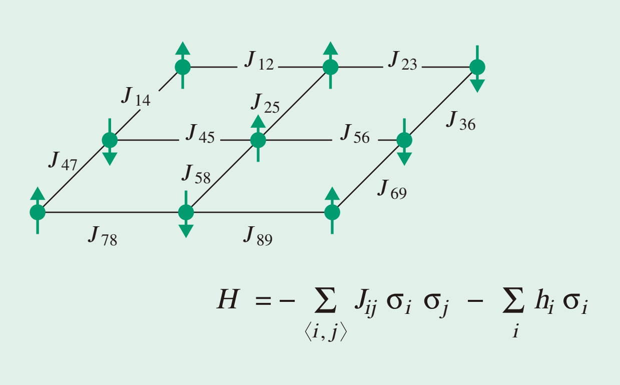

Problem
Vary parameters and recalculate performance indices
2^n times
Assess indices
Solution
(a) Conventional computing
Problem
Model problem using a natural phenomenon (mapping)
Convergence of natural phenomenon
Observe result
Solution
(b) Natural computing

Fig. 1—Procedure for Solving Optimization Problems.
Conventional practice has been to repeatedly calculate all performance indices and assess the values obtained. Natural computing, in contrast, reduces the number of calculation iterations by taking advantage of the tendency for a natural system to converge.



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

Fig. 2—Ising Model.
An Ising model represents the properties of ferromagnetic materials in terms of statistical mechanics. It consists of a lattice of points (spins), each of which can occupy one of two orientation states, and reaches stability when the energy H is at a minimum, taking account of interactions between adjacent points in the lattice.

10^{300}). Calculating performance indices for such a huge number of combinations is impossible in practice.

What is actually done in situations like this is that, rather than calculating the performance indices for all combinations, an approximation algorithm is used to obtain a roughly optimal combination of parameters. Unfortunately, as the number of parameters increases, finding even an approximate solution becomes difficult. Furthermore, semiconductor scaling has enabled the computational methods used in the past to deal with larger problems by improving the performance of the central processing units (CPUs) used for the calculations. However, progress on semiconductor scaling appears to have plateaued in recent years, and in practice there have been no further improvements in CPU clock speeds since the late 2000s. In other words, optimizing the larger and more complex systems of the future will require new computing techniques that do not rely on the practices of the past.

NEW COMPUTING CONCEPT

Conventional computers break problems down into a collection of programs (procedures) and solve the problems by executing these programs sequentially. As noted above, however, the difficulty with solving combinatorial optimization problems is the explosive growth in the number of procedures required for program execution. Accordingly, Hitachi has proposed adopting a different computing concept, namely natural computing.

Fig. 1 (b) shows the calculation procedure using natural computing. Natural computing works by using a natural phenomenon to model the problem to be solved (mapping) and taking advantage of the convergence inherent in this natural phenomenon to converge on the solution to the problem. The problem can then be solved by observing this converged result.

An Ising model, meanwhile, represents the behavior of magnetic spin in a magnetic material in terms of statistical mechanics and has been proposed as a suitable technique for solving combinatorial optimization problems. Fig. 2 shows an Ising model. The properties of a magnetic material are determined by magnetic spins, which can be oriented up or down. An Ising model is expressed in terms of the individual spin states (σ_i), the interaction coefficients (J_{ij}) that represent the strength of the interactions between different pairs of spin states, and the external magnetic coefficients (h_i) that represent the strength of the external magnetic field. The figure also includes the equation for the energy (H) of the Ising model. One property of an Ising model is that the spins shift to the states that minimize this energy, ultimately leaving the model in this minimum state. If a combinatorial optimization problem is mapped onto an Ising model in such a way that its performance index corresponds to the model's energy, the Ising model is allowed to converge so that the spin states adopt the minimum-energy configuration. This is equivalent to obtaining the combination of parameters that minimizes the performance index of the original optimization problem.

CMOS ISING COMPUTING

While computing methods that use superconductors to replicate an Ising model have been proposed in the past, Hitachi has proposed using a complementary metal oxide semiconductor (CMOS) circuit for this purpose. The benefits of using a CMOS circuit are simpler manufacturing, greater scalability, and ease of use.

The updating of actual spin values is performed in accordance with the following rule:

New spin value = +1 (if $a > b$)
−1 (if $a < b$)
+/−1 (if $a = b$)

Here, a is the number of cases in which (adjacent spin value, interaction coefficient) is (+1, +1) or (−1, −1) and b is the number of cases in which it is (+1, −1) or (−1, +1). These interactions cause the energy of the Ising model to fall, following the energy contours (landscape) like that shown in Fig. 3. However, because the energy profile includes peaks and valleys (as shown in the figure), this interaction process operating on its own has the potential to leave the model trapped in a local minimum in a region that is not the overall minimum for the system.

To escape such local minima, the spin states are randomly perturbed. This causes the system to randomly switch to an unrelated state, as indicated by the dotted line in Fig. 3. Collectively, these two processes are called CMOS annealing. By using them, it is possible to identify the state with the lowest energy that can be found.

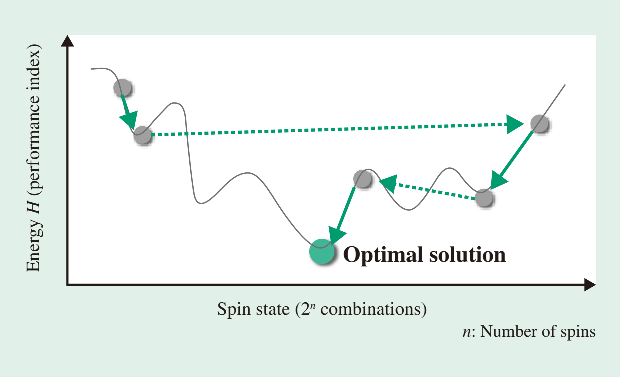

Energy H (performance index)
Optimal solution
Spin state (2^n combinations)
n: Number of spins

Fig. 3—Ising Model Energy Landscape and CMOS Annealing.
In Ising computing, although the energy falls in accordance with the energy contours (landscape) due to the interactions between spins (solid arrows), there is a potential for it to get trapped at a local minimum. This can be prevented by inputting random numbers to deliberately invert spin values (dotted arrows). Called CMOS annealing, this operation obtains a solution with low energy.

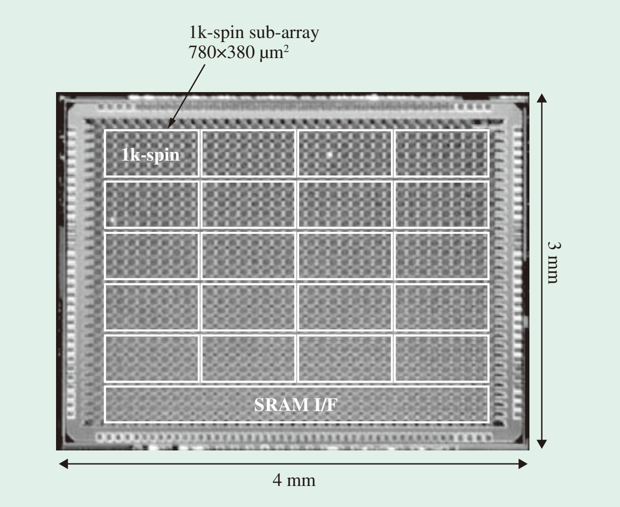
In practice, this use of random numbers means that the solution obtained is not necessarily the optimal one. However, when the computing technique is used for parameter optimization, it is likely that it will not matter if the results obtained are not always optimal. In situations where this computing technique might be deployed, it is possible to anticipate applications where providing a theoretical guarantee that it will produce solutions with 99% or better accuracy, 90% or more of the time, for example, will mean that these solutions can be relied on to not cause any problems for the system.

PROTOTYPE COMPUTER

A prototype Ising chip was manufactured using a 65-nm CMOS process to test the proposed Ising computing technique. An Ising node was then built with this Ising chip and its ability to solve optimization problems was demonstrated. This section describes the prototype and the results of its use to solve optimization problems.

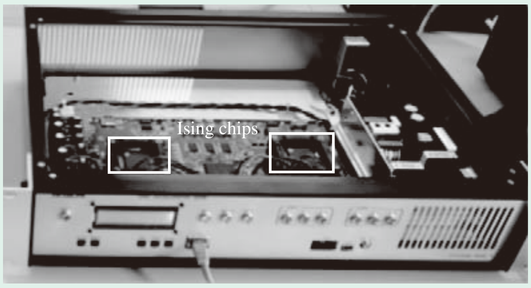
CMOS Ising Chip

The prototype Ising chip was fabricated using a 65-nm semiconductor CMOS process. Fig. 4 shows a photograph of the chip. The 3-mm × 4-mm chip can hold 20,000 spin circuits, each occupying an area of 11.27 µm × 23.94 µm ≈ 270 μm^2. The interface circuit used for reading and writing the spin states and interaction coefficients operates at 100 MHz, as does the interaction process for updating spin values.



SRAM: static random access memory I/F: interface

Fig. 4—Ising Chip Photograph.
The chip has 20,000 spin circuits in a 3-mm × 4-mm = 12-mm^2 area.



PC: personal computer LAN: local area network

Fig. 5—Ising Node.
The photograph shows an Ising node with two Ising chips. The Ising node is connected to a server or PC via a LAN cable and can be used to solve combinatorial optimization problems.

The Ising chip implements a three-dimensional Ising model on a two-dimensional memory lattice. Semiconductor chips achieve a high level of integration by using a two-dimensional layout, and the prototype Ising chip also takes advantage of this to achieve a high level of integration, meaning that it can implement a large number of spin circuits.

Ising Computer

Fig. 5 shows a prototype Ising node fitted with two Ising chips.

The Ising node can be accessed from a personal computer (PC) or server via a local area network (LAN) to input optimization problems and obtain the solutions.

Fig. 6 shows a comparison of the energies required to solve a randomly generated maximum cut problem using the Ising node and using conventional computing. The horizontal axis represents the number of spins in the Ising model. The conventional computing technique used for comparison consisted of executing the SG3 approximation algorithm (which has been optimized for solving maximum cut problems) on a general-purpose CPU. The same problems were solved using both techniques and a comparison was made of the energies consumed in obtaining a solution to an equivalent level of accuracy in each case. Because the SG3 approximation algorithm used for the comparison had been optimized for maximum cut problems that use Ising models, there was no significant difference between the times taken by the two techniques for a problem with 20,000 spins. The amount of energy consumed in solving a 20,000-spin problem, however, was approximately 1,800 times less using the new technique.

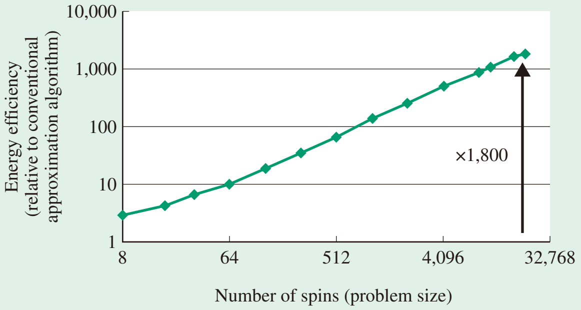


Fig. 6—Energy Efficiency of Solving Randomly Generated Maximum Cut Problem.
The graph shows the relative energy efficiency of the calculation compared to an approximation algorithm executing on a general-purpose CPU. The energy efficiency improves as the size (number of spins) of the problem increases, with the new technique being approximately 1,800 times more efficient for a 20,000-spin problem.

CONCLUSIONS

Table 1 shows a comparison with previous Ising computers. Use of a CMOS semiconductor circuit means the computer can operate at room temperature. This means low power consumption for cooling is achieved. While the prototype computer has approximately 20,000 spin circuits, it will be possible to replicate even larger Ising models by using higher levels of semiconductor process scaling.

Furthermore, because the current system uses digital values to calculate spin interactions, it is easy to link a number of chips together and expand the size by using multiple chips.

Although it is anticipated that this use of digital circuits will result in lower solution accuracy

TABLE 1. Comparison with Existing Ising Computers
The new technique is significant in engineering terms because of its suitability for real-world applications, being superior to an existing Ising computer that uses superconductors in terms of things like ease-of-use and scalability.

	New technique	Existing technique
Approach	Ising computing	
	Semiconductor (CMOS)	Superconductor
Operating temperature	Room temperature	20 mK
Power consumption	0.05 W	15,000 W (including cooling)
Scalability (number of spins)	20,000 (65 nm) Can be scaled up by using higher level of scaling	512
Computation time	Milliseconds	Milliseconds (fast in principle)

than can be achieved by previous systems based on superconductors, it is adequate for use in the optimization of actual social systems because it is able to solve problems in practice. Moreover, the approach described here of using a semiconductor is significant in engineering terms for reasons that include ease-of-use and scalability.

This article has described how the prototype Ising computer successfully solved a maximum cut problem, which is a form of combinatorial optimization problem. As it is known that this problem can be translated mathematically into other combinatorial optimization problems, this indicates that the technique has potential for use in actual system optimization. Furthermore, energy measurements demonstrated that the technique can reduce consumption by three or more orders of magnitude compared to conventional computing techniques.

In the future, Hitachi sees the Ising computer as a highly efficient technology for solving combinational optimization problems in AI applications, which are expected to impose increasing processing loads in the future.

REFERENCES

(1) M. W. Johnson et al., "Quantum Annealing with Manufactured Spins," Nature **473**, pp. 194–198 (May 2011).

(2) R. F. Service, "The Brain Chip," Science **345**, Issue 6197 (Aug. 2014).

(3) C. Yoshimura et al., "Spatial Computing Architecture Using Randomness of Memory Cell Stability under Voltage Control," 2013 European Conference on Circuit Theory and Design (Sep. 2013).

(4) M. Yamaoka et al., "20k-spin Ising Chip for Combinational Optimization Problem with CMOS Annealing," ISSCC 2015 digest of technical papers, pp. 432–433 (Feb. 2015).

(5) S. Kahruman et al., "On Greedy Construction Heuristics for the MAX-CUT Problem," International Journal of Computational Science and Engineering **3**, No. 3, pp. 211–218 (2007).

ABOUT THE AUTHORS

Masanao Yamaoka, Ph.D.

Center for Exploratory Research, Research & Development Group, Hitachi, Ltd. He is currently engaged in research into computers based on new concepts. Dr. Yamaoka is a member of the IEEE.

Chihiro Yoshimura

Center for Exploratory Research, Research & Development Group, Hitachi, Ltd. He is currently engaged in research into computers based on new concepts.

Masato Hayashi

Center for Exploratory Research, Research & Development Group, Hitachi, Ltd. He is currently engaged in research into computers based on new concepts.

Takuya Okuyama

Center for Exploratory Research, Research & Development Group, Hitachi, Ltd. He is currently engaged in research into computers based on new concepts.

Hidetaka Aoki

Hitachi Asia (Malaysia) Sdn. Bhd. He is currently engaged in the research and development of green computing. Mr. Aoki is a member of the Information Processing Society of Japan.

Hiroyuki Mizuno, Ph.D.

Center for Technology Innovation–Information and Telecommunications, Research & Development Group, Hitachi, Ltd. He is currently engaged in research into information and telecommunications technology. Dr. Mizuno is a member of the IEEE.