Overview of CMOS Annealing Machines

Masanao Yamaoka
Senior Researcher
Research & Development Group, Hitachi, Ltd.

1. Introduction

As the Internet of Things (IoT) becomes commonplace, it requires more computing capacity. In computers with a von Neumann architecture, the scaling of semiconductor devices has so far supported exponential increases in processing capability, thereby ensuring that it has always been possible to deliver the required performance. However, it is now being said that the scaling of semiconductors is coming to an end, and this will make it difficult to improve the performance of conventional von Neumann type computers. Also, considering the future needs of the IoT era, it will be necessary to implement diverse forms of control in a wide variety of systems used by society. This will require the optimal setting of multiple system control parameters, which will entail processing combinatorial optimization problems at high speed.

One method that has been proposed for efficiently solving combinatorial optimization problems involves using an annealing machine based on the Ising model[1][2][3]. Although annealing machines have been implemented in various different ways, we have proposed a CMOS annealing machine that uses semiconductor circuits to simulate an Ising model[4][5][6]. We built a prototype of this CMOS annealing machine and confirmed that it can efficiently process a type of combinatorial optimization problem called the "maximum cut" problem. We also built a prototype second-generation CMOS annealing machine using FPGAs and were able to confirm not only that it can solve more complex combinatorial optimization problems, but also that even larger-scale problems can be solved by connecting multiple second-generation CMOS annealing machines.

2. Combinatorial optimization problems and annealing machines

Combinatorial optimization problems are problems that involve searching for a solution comprising a set of parameters that maximizes (or minimizes) an evaluation function under a given set of conditions. Such problems are characterized in that the number of candidate solutions grows explosively as the number of parameters to be determined increases. In the future, we can expect social systems to become larger in scale and more interconnected, and the number of parameters to be optimized will tend to increase.

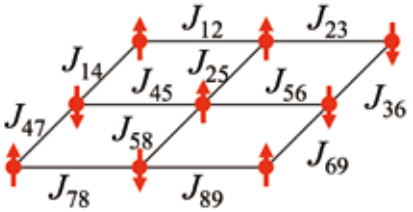
It has been suggested that these combinatorial optimization problems could be solved by using an annealing machine based on the Ising model, which is a statistical mechanics model representing the spin behavior of magnetic bodies. The Ising model is shown in Figure 1. The Ising model consists of "up" and "down" spin states σ_i representing the properties of a magnetic material, an interaction coefficient J_{ij} representing the interaction forces between these two spin states, and an external magnetic field coefficient h_j representing the forces of an externally applied magnetic field. The energy H of the Ising model is expressed by the equation shown in Figure 1. In the Ising model, the spin state is updated to minimize the energy H, eventually yielding the minimal value of H. Combinatorial optimization problems can be solved by using the Ising model as follows. First, the problem is mapped so that the evaluation function of the combinatorial optimization problem corresponds to the energy of the Ising model. Here, the parameters of the optimization problem correspond to the spin values of the Ising model. Next, convergence operations are applied to the Ising model, resulting in a combination of spin states that minimizes its energy. By observing these spin values and mapping their state back to the original optimization problem, it is possible to ascertain the combination of parameters that minimizes the evaluation function, i.e., the solution to the combinatorial optimization problem.

3. CMOS annealing machine

We have proposed a method for simulating an annealing machine in semiconductor CMOS circuits and using it to process combinatorial optimization problems. Since it uses CMOS circuits, it is easy to manufacture, highly scalable, and easy to use.

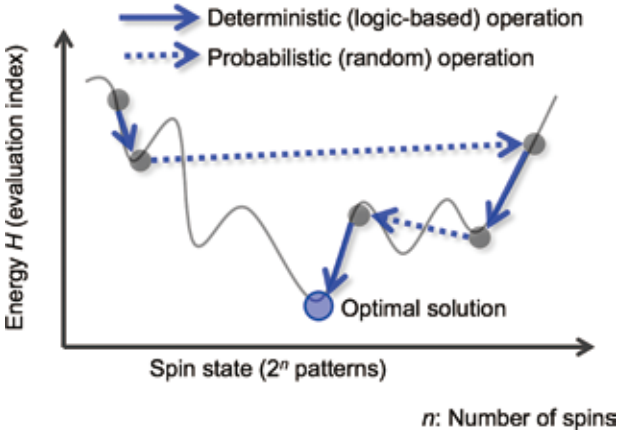
Annealing is an operation that searches for a low-energy state, and is used when searching for the ground state of the Ising model. To perform annealing in CMOS semiconductor devices, it has to be implemented as two operations as shown in Figure 2.

Figure 1: Ising model



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

Figure 2: CMOS annealing



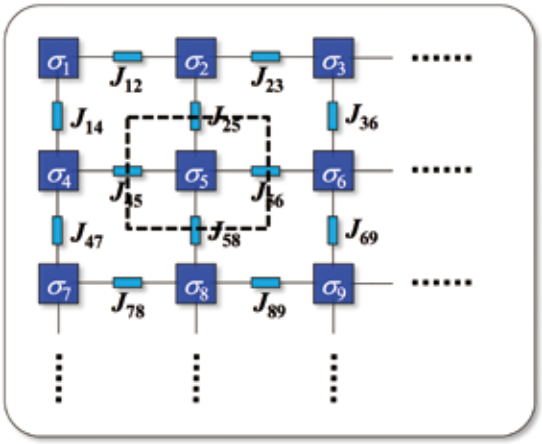
The first operation involves transitioning to a lower-energy state in the energy landscape by a deterministic action as indicated by the solid arrow in Figure 2. With deterministic behavior alone, the algorithm is liable to become trapped in localized energy valleys and will be unable to find other low-energy states. Therefore, in order to escape from these local solutions, the energy state is randomly perturbed by probabilistic operations to search for the lowest possible energy state. The process of searching for a low-energy state by combining these two operations is called CMOS annealing. In CMOS annealing, deterministic operations are performed by modeling interactions between spin states in digital circuits, and the probabilistic operations are performed based on random numbers.

Since CMOS annealing relies on random numbers, it may not always find the optimal solution. However, when this computing technology is used for the optimization of real social systems, it is considered to be acceptable even if it does not always find the absolute optimum value. For example, when searching for traffic routes, a solution can be regarded as acceptable from the viewpoint of system optimization even if it delivers routes that have slightly higher values than the best possible solution. It could therefore be argued that this CMOS annealing technology is geared more towards practicality than strict academic precision.

In the Ising model, the spin states have to be stored as binary values, and thus semiconductor circuits are used to keep them in SRAM. The SRAM also stores the interaction coefficients representing the strength of spin interactions, and the external magnetic field coefficients representing the strength of the external magnetic field. Furthermore, the effects of interactions that update the spin values are simulated by digital circuit operations.

Figure 3 shows the system configuration for implementing CMOS annealing. In this figure, σ_n represents a spin, which is stored as a value of +1 or –1, and J_{ij} represents an interaction between spins. The parts surrounded by the dotted line in this figure represent a single spin interaction processing circuit. This not only stores the actual spin information, but also includes circuits for updating the spin state using the spin effects connected with this spin. CMOS annealing is performed by updating the spin states in these spin circuits. This updating of spin states can be performed simultaneously in parallel for spins that are not connected. For example, in the configuration shown in Figure 3, spins σ_1, σ_3, σ_5, σ_7 and σ_9 are not connected to one another, and can therefore be updated simultaneously. Similarly, spins σ_2, σ_4, σ_6 and σ_8 are also not connected to one another, and can thus be updated simultaneously. In this way, it is possible to update half the total spin states simultaneously in a structure having the (spin connection) topology shown in Figure 3. That is, with this

Figure 3: Configuration of CMOS annealing machine



configuration, any number of spin states can be updated in just two cycles. This means it is possible to suppress increases in the time required for processing even when the scale of the device is increased.

The probabilistic operation of the CMOS annealing operation in Figure 2 introduces random number sequences into the spin circuits. By evaluating these random number sequences, the spin values are stochastically flipped. This causes random transitions to unrelated states as shown by the dotted lines in Figure 2[7][8]. By performing CMOS annealing in combination with the interaction of spin states and probabilistic state transition actions, it is possible to find as many low-energy states of the Ising model as possible.

4. Prototype CMOS annealing machine

To demonstrate the operation of the proposed CMOS annealing machine, we used 65 nm CMOS process technology to fabricate an Ising chip to reproduce the CMOS annealing operations. A photograph of this chip is shown in Figure 4(a). It contains 20,000 spin simulator circuits in a chip measuring 3 mm × 4 mm. Each spin simulator circuit measures 11.27 × 23.94 = 270 μm^2. The interface circuit for reading and writing the spin states and interaction coefficients from outside operates at 100 MHz, which is the same as the rate of the interaction operations that update the spin values.

This Ising chip incorporates a three-dimensional Ising model consisting of two interconnected layers of two-dimensional lattice Ising models. The three-dimensional Ising model is embedded in a two-dimensional memory structure. Semiconductor chips use two-dimensional structures to achieve a high integration density, and our Ising chip achieves a high integration density in the same way, allowing it to simulate a large number of spin states. Figure 4(b) shows a prototype Ising computing node with two Ising chips. This prototype can solve optimization problems supplied to it from a PC or a server via a LAN.

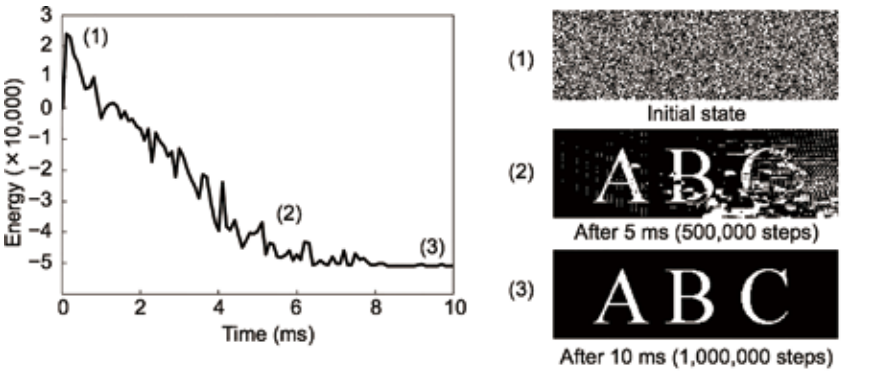
Figure 5 shows the results of using the Ising chip to solve a maximum cut problem (a kind of NP-complete combinatorial optimization problem). Figure 5(a) shows how the energy of the Ising model changes when solving this problem. During CMOS annealing, it can be seen that the energy decreases with time, and finally reaches the minimum energy after 10 ms. The changes of spin state that occur when solving this problem are illustrated by the black and white images in Figure 5(b). Here, white and black points represent "up" and "down" spins, respectively. The problem being solved in this example was chosen so that the letters "ABC" would appear clearly when the spin states corresponding to the optimum solution had been found. As the changes of spin state in this picture clearly show, the spin states start off in a random initial state with an irregular arrangement of white and black points. After 5 ms, the energy of the Ising model has decreased, and the characters ABC have started to emerge from

Figure 4: First generation prototype

(a) Prototype chip

(b) Prototype Ising computing node

Figure 5: First generation prototype measurement results



(a) Variation of Ising model energy when solving maximum cut problems

(b) Variation of Ising model spin states when solving maximum cut problems

the noise. However, the inclusion of noise shows that this state is only a local solution. When the CMOS annealing operations are continued, the energy falls further still. After 10 ms, the ABC characters can clearly be seen without any noise. This is the minimum energy state, showing that the optimum solution to the maximum cut problem has been found. Although it was possible to obtain the optimum solution in this example, it is not always possible to obtain the optimum solution due the probabilistic nature of CMOS annealing as described above. However, we have confirmed that this operation results in a reduction of energy and is capable of finding as good a solution as possible to combinatorial optimization problems.

As a second-generation prototype, we also fabricated a CMOS annealing machine from FPGAs. A photograph of this prototype is shown in Figure 6(a). Since this prototype uses FPGAs, it allows various different Ising model topologies and interaction coefficients to be tried out. To take advantage of this flexibility, we developed an embedding algorithm to embed the Ising model in the hardware topology[9]. With this algorithm, it is possible to run real-world combinatorial optimization problems on a CMOS annealing machine.

Another major advantage of CMOS annealing machines is that they can easily be scaled up by connecting multiple chips together. This is because the calculations are performed by digital circuits, so by exchanging digital signals between the chips, it is possible to perform the same operations as would be performed in a single chip. Since the spin states are sparsely connected in the CMOS annealing machine, another advantage is that the information about spins in one chip can be easily sent to another connected chip. To confirm this scalability, we constructed a large-scale 100 kbit machine by connecting together two second-generation prototypes. With this architecture, we were able to solve huge problems 25 times larger than could be solved by the second-generation hardware alone. This means it is possible to handle social problems that will become more prominent in the future (Figure 6(b)).

5. Conclusion

We have built a CMOS annealing machine based on CMOS semiconductor circuits. Our first-generation prototype can simulate about 20,000 spin states. In the future, it will be possible to reproduce even larger Ising models by making use of finer semiconductor processing. Furthermore, we have shown that spin interactions can be calculated using digital values. This means that further increases of scale can easily be achieved by interconnecting multiple chips. Using a second-generation prototype, we confirmed that this multiple chip configuration works properly. From the viewpoint of ease of use and scalability, it can be said that this semiconductor-based approach is a significant engineering achievement. We have confirmed that our prototype CMOS annealing machine is capable of solving actual maximum cut combinatorial optimization problems. It is known that maximum cut problems can be transformed mathematically into other combinatorial optimization problems, so we believe that this architecture can be used in the optimization of real-world systems.

References
[1] W. Johnson et al., "Quantum annealing with manufactured spins," Nature 473, pp. 194–198, 12nd May 2011.
[2] T. Inagaki et al., "A coherent Ising machine for 2000-node optimization problems," Science 20, Oct 2016, DOI: 10.1126/science.aah4242.
[3] P.L. McMahon et al., "A fully-programmable 100-spin coherent Ising machine with all-to-all connections," Science 20, Oct 2016, DOI 10.1126/science.aah5178.
[4] C. Yoshimura et al., "Spatial computing architecture using randomness of memory cell stability under voltage control", 21st European Conference on Circuit Theory and Design, September 2013.
[5] 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.
[6] M. Yamaoka et al., "A 20k-Spin Ising Chip to Solve Combinatorial Optimization Problems With CMOS Annealing," IEEE J. Solid-State Circuits, vol. 51, no. 1, pp. 303–309, Jan. 2016.
[7] M. Hayashi et al., "An Accelerator Chip for Ground-State Searches of the Ising Model with Asynchronous Random Pulse Distribution," 2015 Third International Symposium on Computing and Networking (CANDAR), pp. 542–546, Feb. 2015.
[8] M. Hayashi et al., "Accelerator Chip for Ground-state Searches of Ising Model with Asynchronous Random Pulse Distribution," International Journal of Networking and Computing, vol. 6, no. 2, pp. 195–211, July 2016.
[9] T. Okuyama et al., "Computing architecture to perform approximated simulated annealing for Ising models," International Conference on Rebooting Computing, Oct. 2016.

Figure 6: Second generation FPGA prototype

(a) 4 kbit machine (b) A 100 kbit machine made by connecting 25 FPGAs