24.3 20k-spin Ising Chip for Combinational Optimization Problem with CMOS Annealing

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

Hitachi, Tokyo, Japan

In the near future, the performance growth of Neumann-architecture computers will slow down due to the end of semiconductor scaling. Presently a new computing paradigm, so-called natural computing, which maps problems to physical models and solves the problem by its own convergence property, is expected. The analog computer using superconductivity from D-Wave [1] is one of those computers. A neuron chip [2] is also one of them. We proposed a CMOS-type Ising computer [3]. The Ising computer maps problems to an Ising model, a model to express the behavior of magnetic spins (the upper left diagram in Fig. 24.3.1), and solves the problems by ground-state search operations. The energy of the system is expressed by the formula in the diagram. Computing flows are expressed in the lower flow chart in Fig. 24.3.1. In the conventional Neumann architecture, the problem is sequentially and repeatedly calculated, and therefore, the number of computing steps drastically increases as the problem size grows. In the Ising computer, in the first step, the problem is mapped to the Ising model. In the next steps, an annealing operation, the ground-state search by interactions between spins, are activated and the state transitions to the ground state where the energy of the system is minimized. The interacting operation between spins is decided by the interaction coefficients, which are set to each connection. Here, the configuration of the interaction coefficients is decided by the problem, and therefore, the interaction coefficients are equivalent to the programming in the conventional computing paradigm. The ground state corresponds to the solution of the original problem, and the solution is acquired by observing the ground state. The interactions for the annealing are performed in parallel, and the necessary steps for the annealing are smaller than that used by a sequential computing, Neumann architecture. As the table in Fig. 24.3.1, our Ising computer uses CMOS circuits to express the Ising model, and acquires the scalability and operation at room temperature.

Figure 24.3.2 shows an approach to realize an Ising computer using CMOS LSI. The topology of the Ising model is 2-layer 2-dimensional lattice as shown in the left diagram. N000 – N122 represents the Ising spins, and the neighbor spins are connected to each other. The energy of whole system goes to a lower level by the interaction of connected spins. This Ising model is mapped to SRAM arrays. The spin has +1 state indicated by a memory cell value of “0”, or, -1 state indicated by a value of “1”. One spin cell, indicated as N000 to N122 in Fig. 24.3.2, includes 13 memory cells for the spin value itself and the values of 12 coefficients for interactions. The memory cells can be accessed by general SRAM access circuits, word lines, bit lines, WL drivers and IO circuits. Each spin also has circuits to model the interaction with adjacent cells as white arrows in the right diagram in Fig. 24.3.2. As shown in Fig. 24.3.2, the 3-dimensional lattice is embedded into the 2-dimensional SRAM array.

Fig. 24.3.3 show the detail of the interaction between spins. To lower the system energy, the interaction between adjacent spins is achieved by the rules indicated at the lower left in Fig. 24.3.3. In the upper left circuit diagram in Fig. 24.3.3 shows a structure of one spin, N is the spin value, NU, NL, NR, ND, and NF are the values from the adjacent spins. Ix0 (x=, U, L, R, D, F) represents the absolute value of coefficient, 1 or 0. Ix1 represents the sign of coefficient, + or -. IS0 and IS1 are the external magnetic coefficient that is the unique value of the spin. Here, the values of EXOR of input values of adjacent spins and coefficients are calculated, and then evaluates which signs are majority. If the number of +1 is greater than that of -1, the next value of the spin becomes +1, and vice versa. The function of EXOR and majority voting are performed by the CMOS circuits indicated at the right in Fig. 24.3.3. We refer to the operation as CMOS annealing. The lines, CT and CB, are precharged at the beginning of the evaluation period, then CLKI is activated and the currents of EXOR flow through the serial NMOS FETs. More activated NMOS lines drive the CT / CB to a lower voltage, and at the next period, a sense amplifier is activated and evaluates the high and low of CT and CB. The result is written to the spin cell, N. To reduce the effect of RDF of the NMOS FETs, a longer gate is used for the serial NMOS FETs. The longer gate also contributes to avoiding rapid current change of the flows, and reduces the effect of noise on the NMOS gates.

Usually the energy profile of the Ising model has peaks and troughs as shown in the upper right graph in Fig. 24.3.4. By the interaction of spins, the energy goes down along the energy profile, and therefore, in many cases, the energy takes on local minimum values (triangles in the graph), and is stuck at these points. It means the solution of the problem, the global minimum, cannot be acquired. To avoid this situation, the energy state is necessary to be randomly changed to other states. In our Ising computer, the spin cell indicated in the upper right diagram in Fig. 24.3.3 has an inversion circuit and occasionally creates an inverse the spin value according to random numbers from an external input. For the other method, we propose to use variations found in SRAM cells. In SRAM circuits, lowering the supply voltage to the memory cells causes error bits. These error bits are used to change the energy states randomly. However, simply reducing the supply voltage of the memory cells also destroys the values of the coefficient. It means the calculation is ultimately false, much like it would be if the program is destroyed in the conventional computing case. Therefore, lowering supply voltage and only activating the memory cell storing the spin data is used. As shown in the lower left circuit diagram in Fig. 24.3.4 shows the situation. The supply voltage to memory cells, VDDM, is lowered from 1.0V to 0.7V, and the WL of the memory cells storing spin value is activated. The SNM of the cell is small and the value is destroyed. On the other hand, the WLs of the memory cells storing interaction coefficients are not activated, and the SNM maintains a high value and the values are not destroyed. The right lower graph of Fig. 24.3.4 shows the measurement results of SRAM array. At 0.7V VDDM, about 30% cells have errors when the WL is activated. On the other hand, no cells have errors when the WL is not activated. These characteristics are used to avoid getting stuck in local minimum states.

We fabricated prototype Ising chips (Fig. 24.3.7) in a 65nm process. The chip size is 4x3=12mm^2 and the spin size is 270μm^2. In the Ising chip, 20 1k-spin sub-arrays with 20k spins are embedded. The logic layout rule is used for the SRAM cells. The left graph in Fig. 24.3.5 shows the energy transition when the CMOS annealing for a combinational optimization problem is operated. As time goes, the energy is lowered. The lower-left pictures in Fig. 24.3.5 show a transition of spin status. White points show spin value “+1”, and black points show spin value “-1”. In this problem, at the global minimum, three characters, “ABC”, is clearly apparent. At the beginning of the operation, the spin values are random and there is no clear picture in the spin status. After 5ms, 500,000 interaction steps, “ABC” is apparent with some noise. The energy at that time is close to the global minimum, but not at the global minimum. After 10ms, 1,000,000 interaction steps, “ABC” is clearly appeared without noise, and this is the global minimum of this problem. The right graph in Fig. 24.3.5 shows the power efficiency when solving a randomly generated problem compared to a conventional method, an approximation algorithm “SG3” run on a CPU. When the problem size is 20k spins, the energy efficiency is 1800 times higher than that of the conventional method. The experimental conditions are described in Fig. 24.3.5. The power efficiency of Ising chip becomes better along with problem size.

The left graph in Fig. 24.3.6 shows the transition of the system energy when using lower-voltage read for destruction of spin values to avoid sticking to a local minimum. From the beginning of the operation to time (a), an interaction operation is performed and the energy is going down. The point (a) is a local minimum and the “ABC” pattern is apparent with noise as shown in the picture (a). Next, a lower-voltage read operation is performed for the spin cells, and the values are randomly destroyed and energy is going up, and the “ABC” pattern is almost diminished as show in the picture (b). Some interaction operations are activated and the energy is going down. At the point (e), a better solution than that at the point (a) and (c) is acquired. The “ABC” pattern of the picture (e) also has less noise than the picture (a). The results show the lower-voltage read contributes to avoiding local minimums.

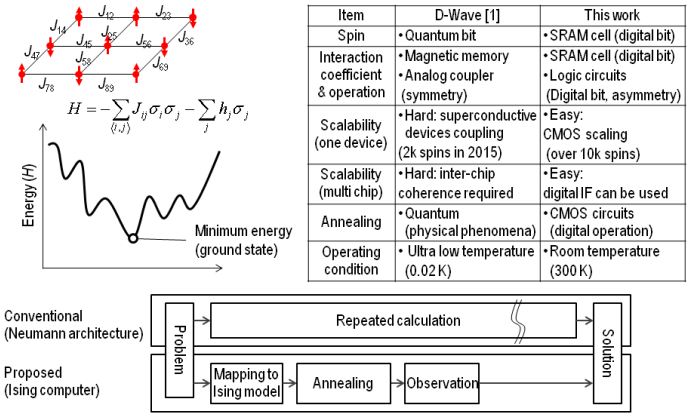
References:

[1] M. W. Johnson, *et al.,* “Quantum annealing with manufactured spins,” *Nature*, Vol. 473, pp. 194–198, May 12, 2011.

[2] R. F. Service, “The brain chip,” *Science*, Vol. 345, no. 6197, pp. 614-616, Aug. 8, 2014.

[3] C. Yoshimura, *et al.,* “Spatial computing architecture using randomness of memory cell stability under voltage control”, *21st European Conference on Circuit Theory and Design*, Sept. 2013.

 978-1-4799-6224-2/15/$31.00



Item	D-Wave [1]	This work
Spin	•Quantum bit	•SRAM cell (digital bit)
Interaction coefficient & operation	•Magnetic memory •Analog coupler (symmetry)	•SRAM cell (digital bit) •Logic circuits (Digital bit, asymmetry)
Scalability (one device)	•Hard: superconductive devices coupling (2k spins in 2015)	•Easy: CMOS scaling (over 10k spins)
Scalability (multi chip)	•Hard: inter-chip coherence required	•Easy: digital IF can be used
Annealing	•Quantum (physical phenomena)	•CMOS circuits (digital operation)
Operating condition	• Ultra low temperature (0.02 K)	•Room temperature (300 K)

Figure 24.3.1: Paradigm shift from the Neumann computer.

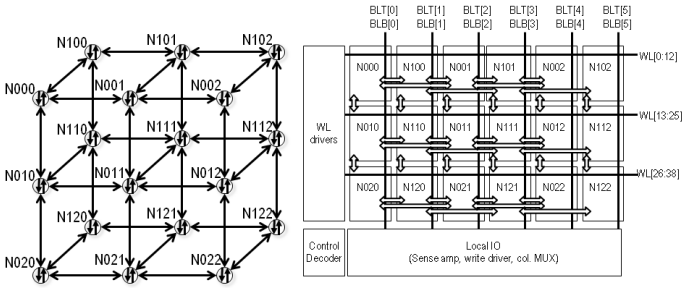


Figure 24.3.2: Topology and implementation to SRAM.

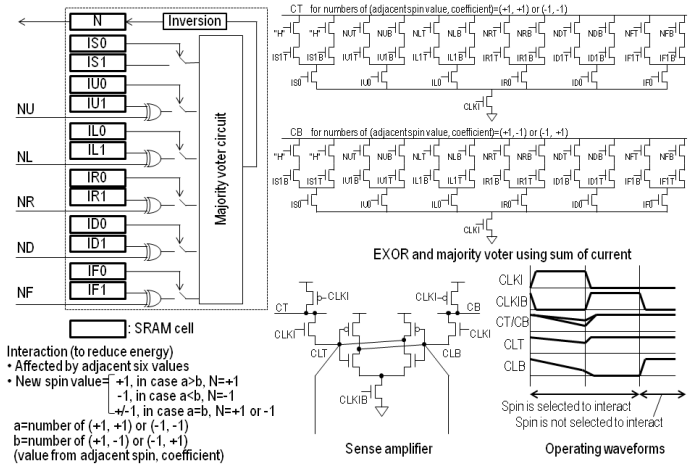


Figure 24.3.3: Interaction calculation.

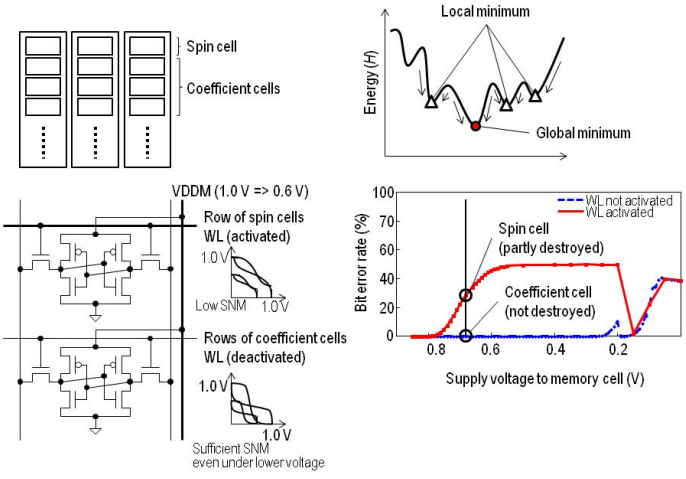


Figure 24.3.4: Positive usage of SRAM-cell variation.

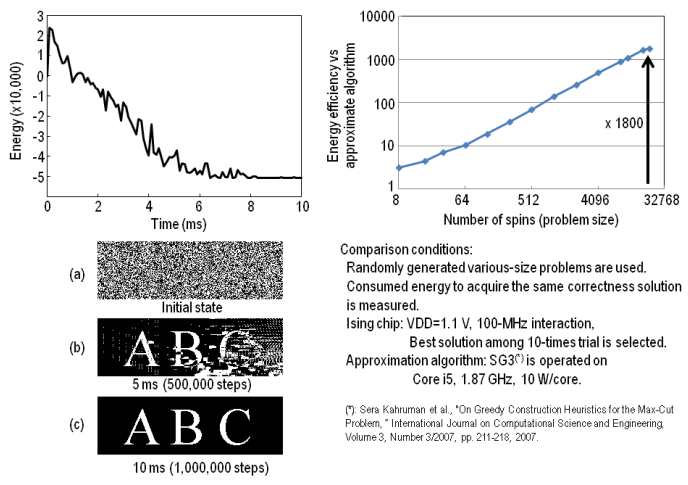


Figure 24.3.5: Measurement; solving optimization problems.

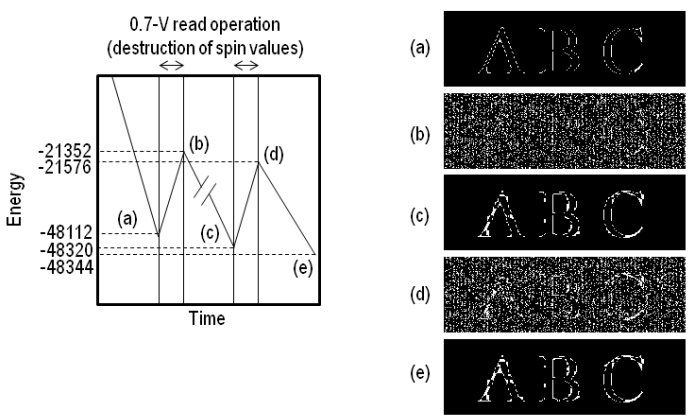
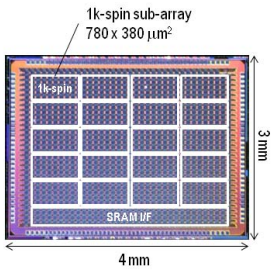


Figure 24.3.6: Measurement; SRAM-cell variation.



Items	Value
Number of spins	20k (80 x 256)
Process	65 nm
Chip area	4x3=12 mm²
Area of spin	11.27 x 23.94 =270 μm²
Number of SRAM cells	260k bits Spin value: 1 bit Interaction factor: 2 bit x 5=10 bits External magnetic coefficient: 2 bits
Memory IF	100 MHz
Interaction speed	100 MHz
Operating current of core circuits (1.1 V) Do not include IO	Write: 2.0 mA Read: 6.0 mA Interaction: 44.6 mA

Figure 24.3.7: Ising chip specifications.

978-1-4799-6224-2/15/$31.00