博士学位論文要約（令和３年３月）

# 確率的演算のための超常磁性磁気トンネル接合の設計と実現

ボーダーズ　ウィリアム
指導教員：深見　俊輔

# Design and Fabrication of Stochastic Magnetic Tunnel Junctions for Probabilistic Computing

William BORDERS
Supervisor: Shunsuke FUKAMI

Combinatorial optimization problems are computationally complex problems that conventional computers of a deterministic nature require increasingly longer solution times for increasing problem sizes, leading to research in non-deterministic methods. This work presents a naturally non-deterministic paradigm, probabilistic computing, developed with stochastic magnetic tunnel junctions (s-MTJs). The s-MTJ is designed to show resistance fluctuations over time and when connected with complementary metal-oxide-semiconductor (CMOS) components, represents a probabilistic bit (p-bit) similar to that of a binary stochastic neuron (BSN) used in stochastic neural network applications. By mapping a adiabatic quantum computing-inspired algorithm to a p-bit network connected through a microcontroller and digital-to-analog controller (DAC), the system can represent the lowest energy state of the cost function as the most probable state visited. As an example of combinatorial optimization, eight p-bits are used to solve integer factorization for integers up to 945.

## 1. Introduction

Computationally complex tasks such as integer factorization, the traveling salesman problem (TSP), and the knapsack problem belong to a class of problems known as combinatorial optimization which not only produces a solution, but is tasked with producing the most optimal one. Due to the vast real-world applications of these problems, a large effort in quantum computing [1] research has begun in an effort to accelerate solutions, because conventional computers using deterministic bits 0 and 1 are ill-matched in nature to asses a problem with an exponentially large number of solutions as the problem size increases. In theory, quantum computing has the capability of simultaneously assessing all possible solutions on a computer consisting of qubits, or the linear superposition of 0 and 1, but suffers from fundamental challenges such as decoherence and a requirement for cryogenic operation temperatures that currently limit realization of large-scale quantum computers.

This work focuses on probabilistic computing [2], a classical system with p-bits fluctuating in time between 0 and 1. The central concept is that by mapping a cost function to a hardware with probabilistic nature, the most probable solution follows the Boltzmann distribution and can be represented by:

$$P(E,T) = \frac{1}{Z}\exp\left[\frac{E}{T}\right]$$

where $Z$ represents the partition function, $E$ the cost function, and $T$ which represents the dimensionless ‘temperature’ of the system. This probabilistic nature makes the network well equipped to tackle combinatorial optimization problems, without the challenges faced by quantum computing. The purpose of this work is to show how s-MTJs can be designed from market-ready spin-transfer torque magnetoresistive random access memory (STT-MRAM) as the source of probabilistic nature in the p-bit, and used as an essential building block for scalable solutions to large-scale combinatorial optimization problems.

## 2. Characterization of s-MTJs

An s-MTJ is comprised of a reference layer (RL) with fixed magnetization direction, a free layer (FL) with reversible magnetization and an insulating layer between the two ferromagnetic layers acting as a tunneling barrier (Fig. 1a). s-MTJs are characterized by the tunneling magnetoresistance (TMR), which produces a high(low) resistance when the magnetic orientations of the FL and RL are antiparallel AP(parallel P), the two stable orientations of the system. In typical spin-transfer torque

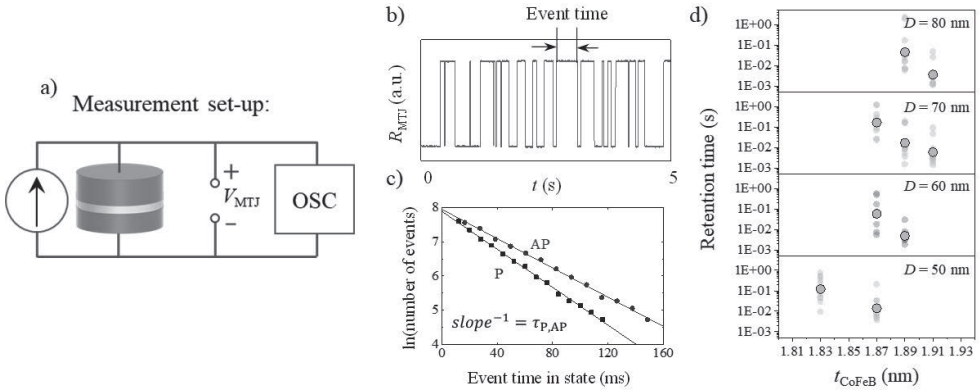


Fig. 1 a) Set-up for measuring the RTN of s-MTJs. b) Example of RTN captured on the oscilloscope. The event time is defined as the time the s-MTJ spends in either the AP or P states before switching. c) example of the histogram of the natural log of counts versus event time. d) summary of the resulting retention times for s-MTJs with various $t_{CoFeB}$ and $D$.

magnetoresistive random access memory (STT-MRAM), the FL and RL are comprised of CoFeB with magnetization easy axis (stable orientation) perpendicular to the plane of the device [3]. For typical non-volatile memory applications, the energy barrier $E$ that separates the P and AP states is increased to levels $E > 40\ k_BT$ so that information stored in memory is retained for years following the Neel-Arrhenius law:

$$\tau = \tau_0 \exp\left(E/_{k_BT}\right)$$

where $\tau_0$ is the attempt time (≈1ns) [4], $k_B$ is the Boltzmann constant, and $T$ is the temperature. In contrast, this work develops s-MTJs with retention times on the order of milliseconds.

To do so, the retention time is measured as a function of s-MTJ FL thickness ($t_{CoFeB}$) and diameter ($D$) (Fig. 1b). First, s-MTJs are fabricated, from the substrate side with the following stack structure: Ta(5)/Pt(5)/[Co(0.3)/Pt(0.4)]$_7$/Co(0.3/Ru(0.45)/[Co(0.3)/Pt(0.4)]$_2$/Co(0.3)/Ta(0.3)/CoFeB(1)/MgO(1.1)/CoFeB($t_{CoFeB}$)/Ta(5)/Ru(5)/Ta(50) where the numbers in parentheses represent nominal thicknesses in nanometers. The films are deposited on thermally oxidized silicon substrates by d.c. and r.f. magnetron sputtering at room temperature and then processed into circular s-MTJs by electron beam lithography and argon ion milling with nominal $D$ varied from 40 to 80 nm. The s-MTJs are then annealed at 300℃ in a vacuum for an hour under a 1.2 T perpendicular magnetic field. Second, the s-MTJ's random telegraph noise (RTN) (Fig. 1b) at a charge current $I_{5050}$ that induces equal time spent in the AP and P states is read across an oscilloscope. To determine $\tau$, a histogram is plotted as the natural log of counts versus the unique event times measured from RTN where the inverse of the slope corresponds to $\tau_{P(AP)}$ (Fig. 1c). Histograms are produced from approximately 1,000 to 10,000 switching events and at a sampling rate ≥10 times faster than the fastest event time to ensure precision of data. Fig. 1d shows the resulting retention times for s-MTJs with varying FL thicknesses and diameters, where an increase(decrease) in $t_{CoFeB}(D)$ leads to a speed up in the time scale of retention time. To match with the speed of the microcontroller and DAC used in this work, s-MTJs with $1 \leq \tau \leq 100$ ms are used, corresponding to a $t_{CoFeB}$ = 1.9 nm and varied $D$. s-MTJs used in the following experiments are then cut from 3-inch wafers and wire bonded onto IC chips.

### 3. P-bit and P-circuit Construction

To form a probabilistic bit, s-MTJs are connected with standard n-type metal-oxide-semiconductor (NMOS) transistors (2N7000), a source resistance, and operational amplifier (AD8692). The output voltage of each p-bit $V_{OUT}$ as a function of the gate voltage at the NMOS $V_{IN}$ can be expressed as:

$$V_{OUT} = \vartheta\left[\sigma\left(\frac{V_{IN-}v_0}{V_{t0}}\right) - r\right]$$

where $\vartheta$ represents the unit-step function and is performed by the operational amplifier, σ represents the sigmoid function performed by the NMOS, $v_0$ represents the center of the $V_{OUT}$ response (1.95 V in this work), $V_{t0}$ represents the transistor scaling

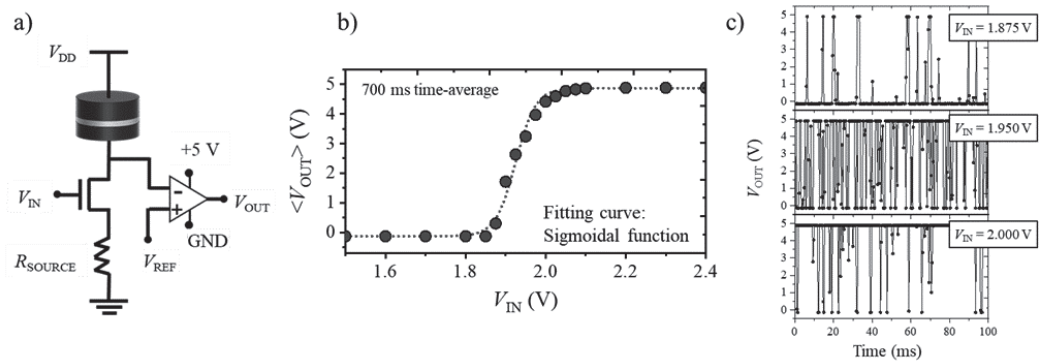


Fig. 2 a) Circuit diagram representing the p-bit. b) time-averaged response of a p-bit's $V_{OUT}$ as a function of $V_{IN}$ taken for 700 ms at each point. c) Time-domain response of a p-bit's output for low, mid, and high input voltages. At $V_{IN}$ = 1.95 V, the charge current flowing through the s-MTJ is approximately equal to $I_{5050}$.

characteristics, and $r$ represents a random number uniformly distributed between 0 and 1 performed by the s-MTJ. Fig. 2a shows the circuit diagram for a single p-bit. $V_{DD}$ is the supply voltage, $R_{SOURCE}$ is used to center s-MTJs with different $I_{5050}$ to a similar range, and $V_{REF}$ is set to $V_{DD} - I_{5050}(R_{AP}+R_P/2)$. Fig. 2b shows the time-averaged characteristics of $V_{OUT}$ as $V_{IN}$ is increased to 2.5 V. Coupled with the time-domain response of $V_{OUT}$ shown in Fig. 2c, the p-bit circuit is capable of representing a hardware implementation of a BSN. P-bits can then be connected together through a microcontroller and DAC implementing the synaptic weight logic in neural networks. Fig. 3 shows a block diagram of the implemented synaptic weight logic. The microcontroller reads the $V_{OUT}$ of each p-bit as digital outputs, updates their inputs based on the cost function, and sends the inputs to the DAC which converts the values to analog voltage input signals.

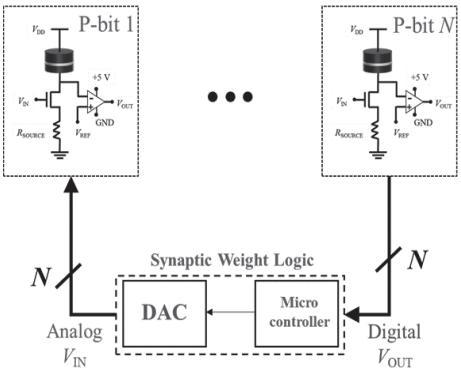


Fig. 3 Block diagram of the p-bit network showing how p-bits are connected through synaptic weight logic implemented on a microcontroller and DAC.

## 4. Integer Factorization Experiment

To perform integer factorization, a cost function inspired by adiabatic quantum computing [5] represents the energy of the system as $E = (XY\text{-}F)^2$. In this representation, the interactions between p-bits are designed so that the overall lowest energy state of the system corresponds to a configuration when the integers $X$ and $Y$ are equal to the integer $F$. By writing the cost function in binary form,

$$E(x_P, \ldots, x_1, y_Q, \ldots, y_1) = \left[\left(\sum_{p=0}^{P} 2^p x_p\right)\left(\sum_{q=0}^{Q} 2^q y_q\right) - F\right]^2$$

with $x_0$ and $y_0$ = 1, and $P$ and $Q$ corresponding to the number of bits used to represent $X$ and $Y$, respectively, one p-bit can be used to represent one bit in either $X$ or $Y$. Similar to Boltzmann machines, the input to each p-bit $I_i$ is determined by the relation $I_{xi,yi} = -\partial E/\partial x_i y_i$, so that the distribution of the cost function follows that of the Boltzmann distribution. The input $I_i$ can be translated to input voltages by the relation $V_{IN,i} = I_0 V_{t0,i} I_i + v_{0,i}$, where $I_0$ corresponds to a dimensionless 'inverse pseudotemperature', which controls the strength of correlation between bits.

Before performing the integer factorization experiment, an uncorrelated state is established to ensure that the system is not biased towards any particular state. First, the time-averaged $V_{OUT}$ is measured for each p-bit and any shifts of the sigmoid response from the center of 1.95 V is compensated by applying additional d.c. biases to $V_{IN}$. Fig. 4a shows the experimentally measured uncorrelated state histograms after calibrating the p-bits for problem sizes of 4, 6, and 8 p-bits. The uncorrelated state is effectively obtained by setting the correlation parameter $I_0$ = 0. The

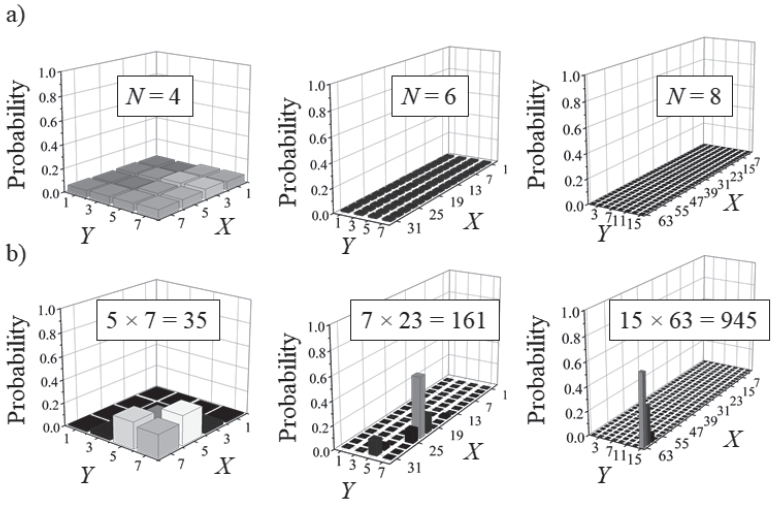


Fig. 4 a) Uncorrelated state for a 4-, 6-, and 8-p-bit system showing approximately equal probability for every state of $X$ and $Y$. b) Correlated state (non-zero $I_0$) showing successful factorization for $F$ = 35, 161, and

histograms are produced by measuring the $V_{OUT}$ of each p-bit and converting the values to decimal form with the relation $X = 2^P x_P + \ldots + 4x_2 + 2x_1 + 1$ and $Y = 2^Q y_Q + \ldots 4y_2 + 2y_1 + 1$ where the $V_{OUT}$ for each p-bit is thresholded to {0,1}.

Once the uncorrelated state is established, integer factorization is performed by applying a non-zero correlation parameter $I_0$ to the system. In the correlated state, 4 p-bits ($P = 2$, $Q = 2$) successfully factor $F = 35$, by showing prominent peaks equivalent to the correct factors of 35, 7×5 and 5×7. Similarly, 6 p-bits ($P = 4$, $Q = 2$) show the correct factors for $F = 161$, and 8 p-bits ($P = 5$, $Q = 3$) show the integer $F = 945$ can be factored into $X = 63$ $Y = 15$ (Fig. 4b) [6].

## 5. Conclusion

In conclusion, this work has presented how s-MTJs can be used as the building block for a hardware implementation of p-bits in solving computationally complex problems such as combinatorial optimization. One of the key advantages of the p-bit implementation is a capability for asynchronous operation, where purely software implementations of Boltzmann machine-like systems require sequential updates for each neuron. For this system, it was found that as long as the synaptic weight logic can be updated faster than the fastest p-bit, the network will perform as expected. To perform the same function as the s-MTJ on purely CMOS-based abstractions, a linear-shift feedback register (LFSR) is often employed. Compared to the s-MTJ, the LFSR requires 10× more energy per random bit and 300× more area.

Compared with quantum computers, p-bits are capable of operating at room temperature, capable of scaling to a large number of bits using STT-MRAM currently manufacturable at the 1 Gb level, and capable of implementing many-body interactions electrically. While p-bits are incapable of implementing quantum interactions, quantum annealing is also capable of being approximated with replicas of p-bit networks for a certain subclass of quantum systems.

These results show the potential for a massively parallel s-MTJ-based p-bit network set to propel large scale probabilistic hardware into a near-future reality.

## References

1) R. P. Feynman, Int. J. Theor. Phys. **21,** 467-488 (1982).
2) K. Y. Camsari, R. Faria, B. M. Sutton, and S. Datta, Phys. Rev. X **7,** 031014 (2017).
3) S. Ikeda, K. Miura, H. Yamamoto, K. Mizunuma, H. D. Gan, M. Endo, S. Kanai, J. Hayakawa, F. Matsukura, and H. Ohno, Nat. Mat. **9**, 721-724 (2010).
4) W. F. Brown, Phys. Rev. **130**, 1677-1686 (1963).
5) T. Albash, D. Lidar, Rev. Mod. Phys. **90**, 015002 (2018).
6) W. A. Borders, A. Z. Pervaiz, S. Fukami, K. Y. Camsari, H. Ohno, and S. Datta, Nature **573,** 390-393 (2019).