Energy efficient and Cognitive computing using memristors

Computer; where is it from and heading to?



2018. September 12th Physics, SNU









Ultimate energy needed for one logic operation

kT In2 ~ 3x10⁻²¹ Joule

VS.

10μA x 10ns x1V ~ 10⁻¹³Joule

tu.

exact logical history. Two simple, but representative, models of bistable devices are subjected to a more

Decrease computation itself

IBM Journal, July 1961, p.183

1. Introduction



From mathematics to computer



B E H I N D E V E R C O D E I S A N E N L G M A



Alan Turing Turing machine



David Hilbert Formalism





A. M. TURING

[Nov. 12,

ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM

By A. M. TURING.

[Received 28 May, 1936.—Read 12 November, 1936.]

The "computable" numbers may be described briefly as the real numbers whose expressions as a decimal are calculable by finite means. Although the subject of this paper is ostensibly the computable *numbers*.



How does a computer work?







George Boole, Claude Shannon, Johan von Neumann

- George Boole (1815 1864): Boolean algebra
- Whitehead and Russell (1910): AND, OR, NOT and IMP for Boolean algebra
- Alan Turing (1936): Universal computing machine
- Claude Shannon (1937): Switching logic = Boolean algebra



Von Neumann Architecture





Logic processes: Alan Turing, Johan von Neumann, Claude Shannon







Α	В
0	1
1	0

	-
AB	С
0 0	1
0 1	0
1 0	0
1 1	0

A B	С
0 0	1
0 1	1
1 0	1
1 1	0

Complementary MOS technology in late 1960s



Semiconductors in old era: Moore's law

"The number of transistors on a chip will double approximately every two years."



Gordon E. Moore

Co-founder and Chairman Emeritus of Intel

The smaller, the better, the cheaper



Kurzweil's extension of Moore's law

http://en.wikipedia.org/



Then, what to do?

International Technology Roadmap for Semiconductor

Future of Lithography for scaling

Next technology : EUV (Extreme Ultraviolet)

EUV tool principle

Intermediate focus Reticle-stage Technical challenges: Optics fabrication ASML Coating of EUV mirrors EUV system metrology illuminator Collecto **Design Example** Source-Module **Projection optics** 2012 2013 2010 Wafer stage Current Status : 30~40 wf/1hr 0.5 Node \ NA 0.25 0.32 Throughput ≥500wf/day (Target ≥ 1200wf/day) 1.19 32 nm 0.59 constant k (ArF Immersion 250 wf/1hr) 22 nm 0.41 0.52 EUR 42 million 16 nm 0.30 0.37 0.59 Cost (maybe) \$170 million (\$60 million) 11 nm 0.20 0.27 0.41

→ The higher NA (or lower k₁) EUV tools enable resolutions down to 11 nm

Therefore, need alternative memories due to high cost and very low throughput

> ASML EUV Product Roadmap

Storage memory: VNAND vs PNAND

Planar NAND

Vertical NAND

24 layer stacked, 128 Gb VNAND Chip size: 133 mm²

: 173.3 mm²

- Large portion for contact area
- Difficulty in reducing hole size and hole to hole space
- Low performance of poly-Si channel

 \rightarrow Samsung has begun mass production of 256Gb, 48 layer stacked VNAND.

3D XPoint Technology By Intel and Micron

Intel and Micron announced that they developed 128Gb 3D Xpoint memory technology.

Robert B. Crooke, Senior Vice President of intel (left) and D. Mark Durcan, CEO of micron (right)

Cell efficiency can be inferred from picture of a wafer

- ~250ea chips in a 300mm wafer. \rightarrow ~280mm² for one chip.
- Cell size : 2F² = 8 x 10⁻¹⁰mm² (F=20nm, 20nm process used)
- 350Gb will be achieved for 100% cell efficiency (chip area / cell size)
- → 3D Xpoint has ~35% cell efficiency (128Gb / 350Gb)

New computer system organization

Adv. Electron. Mater., 2015, 1400056

Materials

www.MaterialsViews.com

Prospective of Semiconductor Memory Devices: from Memory System to Materials

Cheol Seong Hwang*

The ever-increasing demand for higher-capacity digital memory shows no sign of declining. The conventional strategy for meeting such demand, i.e. shrinking of the memory cell size, will no longer be useful at some point in the future, owing to economic reasons and performance degradation. Nevertheless, performance of computing systems will keep improving for the next generation information technology. This indicates the necessity to consider a fundamentally disparate approach to enhance memory technology. Here, the current status of computer memory chips is reviewed and the pros and cons of the present technology are discussed from computing system, fabrication technology, and materials points of view. Based on this knowledge, the limitations of the present technologies are described, and the possible solutions suggested up to now are reassessed. Finally, a shift in the fundamental computational paradigm from von Neumann computing to other alternatives the Facebook server with a total storage size of >500 petabytes at the end of 2014, and Google handled >7300 petabytes in a year.^[3] Despite this already huge data size, the volume of data is expected to further increase at even a faster rate. Google is actually a misspelling of Googol, which means 10^{100} . As stated by Bekenstein, approximately 10^{100} , although never proven, corresponds to the total number of particles in our universe.^[4] It is anticipated that the produced and the stored data of 2020 will be 44 000 exabytes ($\approx 10^{18}$ bytes), suggesting that there is still a plenty of room for improvements in data storage.

Computer used to be a "computing machine" but they are now exploited as

Dielectric Thin Film Lab, Seoul National University

REVIEW

New computers?

- Neuromorphic computing
- Stateful logic computing
- DNA memory (not a computer)

Von Neumann computer vs. Human brain

Х

Y

Energy for 2 + 3 using Von Neumann computing

Neuron and synapse

Neuromorphic computing architecture

Cognitive computing

AlphaGo by Google Deepmind

At last – a computer program that can beat a champion Go player MEE 480 ALL SYSTEMS GO

 CONSERVATION
 DESCRIPTIONES
 Description
 <thDescription</th>
 Description
 <thDescription</th>
 <thDescription</th>

Google AI computer beats human champion of complex Go boardgame

Fan Hui, three-time champion of the east Asian board game, lost to DeepMind's program AlphaGo in five straight games

Hui, the researchers used a larger network of computers that spanned about 170 GPU cards and 1,200 standard processors, or CPUs. This larger computer network both trained the system and played the actual game, drawing on the results of the training

Nature 529, 484–489 (28 January 2016) the results of the training.

supported by 2 billion neurosynaptic cores containing 550 billion neurons and 100 trillion synapses running only 1542 times slower than real time.

density crisis

Human brain inspired computing based on Memristor (cell size<F²)

Dielectric Thin Film Lab, Seoul National University

million

"IBM Watson Saves a Patient's Life by Correcting Doctor's Diagnosis"

When the condition of Mrs. Yamashita, a 66-year-old patient previously diagnosed with acute myeloid leukemia, was aggravated even after several months of cancer treatment, doctors turned to IBM's Watson for assistance.

Doctors entered the patient's genetic information into Watson with a database of 25 million clinical and 15 million medical studies. It took only 10 minutes for Watson to diagnose a different, rare form of leukemia and suggest proper changes in the original cancer treatment. With the new treatment provided by the analytical supercomputer, Mrs. Yamashita could leave the hospital last June.

Reported on August 5th, 2016

Memristor: ReRAM

where

where

$$R(q) \stackrel{\Delta}{=} \frac{d\hat{\varphi}(q)}{dq}$$

 $G(\varphi) \stackrel{\Delta}{=} \frac{d\hat{q}(\varphi)}{d\omega}$

nature nanotechnology | VOL 3 | JULY 2008

Memristive switching mechanism for metal/oxide/metal nanodevices

J. JOSHUA YANG, MATTHEW D. PICKETT, XUEMA LI, DOUGLAS A. A. OHLBERG, DUNCAN R. STEWART* AND R. STANLEY WILLIAMS

 Physical mechanism for a memristor based on the dopant migration theory.

Any kind of material system that shows pinched I – V characteristics can be a memristor, meaning that most of the ReRAM system, or even phase change memory material, can be considered as the memristor.

Near chaotic behaviors of Neurons and Memristors

Klaus Mainzer • Leon Chua

LOCAL ACTIVITY PRINCIPLE

L. Chua "Biological neurons are poised at the edge of chaos"

- The resting states of neurons are very near chaotic behavior, so even a minute perturbation can make the neurons fire with apparently chaotic behavior.
- It is similar to the *drastic change in resistive* status of a memristor driven by minor change in input voltage.

Typical Negative differential resistance (NDR) in memristors

Memristors for new computing paradigms

✓ <u>Neuristor by HP group</u>

Memristive Hodgkin huxley model describing the action potentials within the squid giant axon, or a neuron.

Synaptic adaptation by S.H. Jo et al.
 Spike-timing-dependent plasticity of potentiation and depression in synapse.

 Stateful Logic (first by HP group) combining the logic and memory chips, fundamentally eliminates the energy cost accompanied with the data input/output step.

Stateful logic

Vol 464 8 April 2010 doi:10.1038/nature08940

nature

LETTERS

'Memristive' switches enable 'stateful' logic operations via material implication

Julien Borghetti¹, Gregory S. Snider¹, Philip J. Kuekes¹, J. Joshua Yang¹, Duncan R. Stewart¹[†] & R. Stanley Williams¹

The author Semiconduc for advancii 🗸 lenged the c state variabl new architec those availa resistive me conductor (strated^{7–12}. A been identif ristive devic

Stateful Logic (first by HP group) combining the logic and memory chips, fundamentally eliminates the energy cost accompanied with the data input/output step.

when the link

The 16 binary Boolean operations using IMP

Table S1. Computational Universality of IMP (Material Implication) & FALSE Operations: the 16 distinct binary Boolean operations on two logic values.

Operation	Truth Table			Equivalent Operation	
p	1	1	0	0	= p
q	1	0	1	0	= q
TRUE	1	1	1	1	= p IMP p
p OR q	1	1	1	0	= (p IMP 0) IMP q
q IMP p	1	1	0	1	= q IMP p
p	1	1	0	0	= (p IMP 0) IMP 0
p IMP q	1	0	1	1	= p IMP q
q	1	0	1	0	= (q IMP 0) IMP 0
p EQUAL q	1	0	0	1	= (($p \text{ IMP } q$) IMP (($q \text{ IMP } p$) IMP 0)) IMP 0

 Stateful Logic (first by HP group) combining the logic and memory chips, fundamentally eliminates the energy cost accompanied with the data input/output step.

FALSE

Logic cascading: Full adder

a	Step1 Step2 Step3 b			
	AND q' OR C	P C		Mu MB
		Р	Q	т
	Input	(p?)	(q?)	(t?)
Step 1	Write $P = (pq)$, $Q = (qp)$	(pq)	(<i>qp</i>)	(t?)
Step 2	Execute $p' \leftarrow (p \text{ XOR } q), q' \leftarrow (p \text{ AND } q)$	(p'x)	(q'y)	(t?)
Step 3	Write $P = (p't)$, $T = (tp')$	(p't)	(q'y)	(tp')
Step 4	Execute $p'' \leftarrow (p' \text{ XOR } t) = s, t' \leftarrow (p' \text{ AND } t)$	(sx')	(q'y)	(ťź)
Step 5	Write $Q = (q't')$	(sx')	(q't')	(ťź)
Step 6	Execute $q'' \leftarrow (q' \text{ OR } t') = c$	(sx')	(cy')	(ť2)

Stateful logic vs. CMOS logic

New computing paradigm

Logic vs. memory evolution

- Mammalian brain and computer show a similar evolution trend between the data processing elements and memory. (The increasing rate of memory density is faster than that of CPU.)
- Memory density should be further increased in order to mimic the human brain.

Adv. Electron. Mater. 2016, 1600090

www.MaterialsViews.com

Memristors for Energy-Efficient New Computing Paradigms

Doo Seok Jeong, Kyung Min Kim, Sungho Kim, Byung Joon Choi, and Cheol Seong Hwang*

NEUROMORPHIC COMPUTING

In article number 1600090, D. S. Jeong et al. review memristors and their potential application in new, energy saving forms of computation, including stateful logic and neuromorphic computing. Memristors in a cross-bar array format (background image) create a two-terminal voltage- or charge-driven non-volatile memory and logic component, serving as the critical circuit element for mimicking the human brain. Their discussions on stateful logic are based on material implication logic (center image).

WILEY-VCH

In this Review, memristors are examined from the frameworks of both von Neumann and neuromorphic computing architectures. For the former, a new logic computational process based on the material implication is discussed. It consists of several memristors which play roles of combined logic processor and memory, called stateful logic circuit. In this circuit configuration, the logic process flows primarily along a time dimension, whereas in current von Neumann computers it occurs along a spatial dimension. In the stateful logic computation scheme, the energy required for the data transfer between the logic and memory chips can be saved. The non-volatile memory in this circuit also saves the energy required for the data refresh. Neuromorphic (cognitive) computing refers to a computing paradigm that mimics the human brain. Currently, the neuromorphic or cognitive computing mainly relies on the software emulation of several brain functionalities, such as image and voice recognition utilizing the recently highlighted deep learning algorithm. However, the human brain typically consumes ≈10–20 Watts for selected "human-like" tasks, which can be currently mimicked by a supercomputer with power consumption of several tens of kilo- to megawatts. Therefore, hardware implementation of such brain functionality must be eventually sought for power-efficient computation. Several fundamental ideas for utilizing the memristors and their recent progresses in these regards are reviewed. Finally, material and processing issues are dealt with, which is followed by the conclusion and outlook of the field. These technical improvements will substantially decrease the energy consumption for futuristic information technology.

1. Introduction

1.1. The Energy Crisis and Information Technology

The amount of digital data worldwide exceeded that of analog data in 1998 due to the explosive growth of personal computers, smartphones and enterprise systems.^[1] It is expected

Dr. D. S. Jeong Center for Electronic Materials Korea Institute of Science and Technology 5 Hwarang-ro 14-gil, Seongbuk-gu, Seoul 02792, Republic of Korea Dr. K. M. Kim Hewlett Packard Laboratories Hewlett Packard Enterprise Palo Alto, California 94304. USA Prof. S. Kim Department of Electrical Engineering Sejong University Neungdong-ro 209. Gwangiin-gu, Seoul 143–747. Republic of Korea

DOI: 10.1002/aelm.201600090

Adv Electron Mater 2016 1600090

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Prof. B. J. Choi

Prof. C. S. Hwang

Department of Materials

Science and Engineering

Science and Technology

College of Engineering Seoul National University

E-mail: cheolsh@snu.ac.kr

Seoul National University of

Seoul 01811, Republic of Korea

Seoul 151-744, Republic of Korea

Department of Materials Science and Engineering

Inter-university Semiconductor Research Center

wilevonlinelibrary.com

(1 of 27) 1600090

Dielectric Thin Film Lab, Seoul National University

esting (and also sobering) expectation is that the total number of binary operations in 2040 will be $\approx 10^{40}$, which is an astronomically large number.^[2] The average energy consumption of binary digital operations, including logic, memory, and input/output operations between logic and memory chips, is currently =0.1 picojoules (=10⁻¹³ Joules). Therefore, if this rate of energy consumption per binary operation is maintained, the total energy expenditure in 2040 for computer operations will reach ≈10²⁷ Joules, which is far higher than the total energy that humans will be able to produce at that time. In 1961, Landauer published a paper on the theoretical aspects of computation based on digital logic and demonstrated that effective computation must be based on an irreversible process (otherwise, the input and output cannot be distinguished), the unit process of which will require minimum energy on the order of $\approx kT$ (Boltzmann constant \times temperature), which is $\approx 10^{-21}$ Joules at room

that the total amount of digital data in

2040 (only 25 years from now) will be

 $\approx 10^{28}$ bytes (1 byte = 8 bits), which is approximately one million times greater

than the current total.^[2] A more inter-

temperature.^[3] This estimate means that the aforementioned energy consumption in 2040 could be decreased by a factor of $\approx 10^8$, which appears to be quite promising. However, what Landauer showed is that this energy is a sort of fundamental limit (energy for one thermodynamic degree of freedom) without consideration of a detailed method of how binary states can be represented by a physical entity and what type

- Processing: electron power, speed
- Communication: photon speed, power
- Memory: ion density, stability

- From processing-intensive (memory scarce) to Memory-intensive (memory abundance) environment!
- Unlimited access to abundant memory!

DNA memory for digital data archive?

COULD THE MOLECULE KNOWN FOR STORING **GENETIC INFORMATION ALSO** STORE THE WORLD'S DATA?

BY ANDY EXTANCE

or Nick Goldman, the idea of encoding data in DNA started out as a toke

It was Wednesday 16 February 2011, and Goldman was at a hotel in Hamburg, Germany, talking with some of his fellow bioinformaticists about how they could afford to store the reams of throwing at them. He remembers the scientists

getting so frustrated by the expense and limitations of conventional computing technology that they started kidding about sci-fi alternatives. "We thought, 'What's to stop us using DNA to store information?"

Then the laughter stopped. "It was a lightbulb moment," says Goldman, a group leader at the European Bioinformatics Institute (EBI) in Hinxton, UK. True, DNA storage would be pathetically slow compared with the microsecond timescales for reading or writing bits in a silicon memory chip. It would take hours to encode data by synthesizing DNA strings with a specific pattern of bases, and still more hours to recover that information using a sequencing machine. But with DNA, a whole human genome fits into a cell that is invisible to the naked eye. For sheer density of information storage, DNA could be orders of magnitude beyond silicon - perfect for long-term archiving.

"We sat down in the bar with napkins and biros," says Goldman, and started scribbling ideas: "What would you have to do to make that work?" 'The researchers' biggest worry was that DNA synthesis and sequencing made mistakes as often as 1 in every 100 nucleotides. This would render large-scale data storage hopelessly unreliable - unless they could find a workable error-correction scheme. Could they encode bits into base pairs in a way that would allow them to detect and undo the mistakes? "Within the course of an evening," says Goldman, "we knew that you could"

He and his EBI colleague Ewan Birney took the idea back to their labs, and two years later announced that they had successfully used DNA to encode five files, including Shakespeare's sonnets and a snippet of Martin Luther King's 'I have a dream' speech1. By then, biologist George Church and his team at Harvard University in Cambridge, Massachusetts, had unveiled an independent demonstration of DNA encoding But at 739 ktlobytes (kB), the EBI files comprised the largest DNA archive ever produced - until July 2016, when researchers from Microsoft and the University of Washington claimed a leap to 200 megabytes (MB).

The latest experiment signals that interest in using DNA as a storage medium is surging far beyond genomics: the whole world is facing a data crunch. Counting everything from astronomical images and journal articles to You'Tube videos, the global digital archive will hit an estimated genome sequences and other data the world was 44 trillion grgabytes (GB) by 2020, a tenfold increase over 2013. By 2040, if everything were stored for instant access in, say, the flash memory chips used in memory sticks, the archive would consume 10-100 times the expected supply of microchip-grade silicon3.

That is one reason why permanent archives of rarely accessed data currently rely on old-fashioned magnetic tapes. This medium packs in information much more densely than silicon can, but is much slower to read. Yet even that approach is becoming unsustainable, says David Markowitz, a computational neuroscientist at the US Intelligence Advanced Research Protects Activity (IARPA) in Washington DC. It is possible to imagine a data centre holding an exabyte (one billion gigabytes) on tape drives, he says. But such a centre would require US\$1 billion over 10 years to build and maintain, as well as hundreds of megawatts of power. "Molecular data storage has the potential to reduce all of those requirements by up to three orders of magnitude," says Markowitz. If information could be packaged as densely as it is in the genes of the bacterium Escherichia coli, the world's storage needs could be met by about a kilogram of DNA (see 'Storage limits').

Achieving that potential won't be easy. Before DNA can become a viable competitor to conventional storage technologies, researchers

22 | NATURE | VOL 537 | 1 SEPTEMBER 2016 | CORRECTED 2 SEPTEMBER 2016 © 2016 Macmilian Publishers Limited, part of Springer Nature. All rights reserved commentary

Nucleic acid memory

Victor Zhirnov, Reza M. Zadegan, Gurtej S. Sandhu, George M. Church and William L. Hughes

Nucleic acid memory has a retention time far exceeding electronic memory. As an alternative storage media, DNA surpasses the information density and energy of operation offered by flash memory.

nformation and communication technologies generate vast amounts of data that will far eclipse today's data flows (Fig. 1). Memory materials must therefore be suitable for high-volume manufacturing. At the same time, they must have elevated information stability and limit the energy consumption and trailing environmental impacts that such flows will demand. Analysts estimate that global memory demand — at 3 × 1014 bits — will exceed projected silicon supply in 2040 (Fig. 1b and Supplementary Information sections 1 and 2). To meet such requirements, flashmemory manufacturers would need ~10° kg of silicon wafers even though the total projected wafer supply is ~107-108 kg (Supplementary Figs 1 and 2). Such forecasts motivate an exploration of unconventional materials with cost-competitive performance attributes. With information retention times that range from thousands to millions of years, volumetric density 103 times greater than flash memory and energy of operation 10^s times less, we believe that DNA used as a memory-storage material in nucleic acid memory (NAM) products promises a viable and compelling alternative to electronic memory.

In this Commentary, we discuss the information retention, density and energetics of NAM - specifically related to DNA - for non-biological and non-volatile memory applications, ranging from letters to libraries. The potential of NAM has often been dismîssed, as nucleic acids are believed by some to be fragile and therefore unreliable. This is not the case. For example, the room-temperature half-life of ancient DNA exceeds 100 years^{1,2}. Indeed, the complete genomes of an ~50,000-year-old Neanderthal3 recovered from Siberia and an ~700.000-year-old horse⁴ recovered from the Arctic permafrost (approximate average temperature -4 °C) have been sequenced. Still, the long-term stability of DNA and its decay kinetics are poorly understood at a per-bit (that is, base) level. As an energybarrier model shows (Methods), DNA has a retention time far exceeding electronic memory, and it can store information reliably over time. Through first-principle calculations, DNA has been validated as a model material for future NAM products (Supplementary Information section 8). Therefore, we call for increased cooperation between the biotechnology and semiconductor sectors to pair previously

unfathomable technological advances such as those from the Human Genome Project - with the scaling expertise of the semiconductor industry.

Nucleic acid memory as a material As a material, nucleic acids are negatively charged polyelectrolytes with four monomers (the nucleotides A, T or U C and G). Monomers are covalently bonded to form polymer chains. Once polymerized, an individual chain can hydrogen-bond with itself or with other chains that satisfy base complementarity. These attributes endow nucleic acids with the power of molecular self-assembly, which is made possible by the thermal fluctuations between complementary hydrogen bonds during Watson-Crick hybridization. During DNA hybridization, adenine (A) forms a basepair with thymine (T), and guanine (G) pairs with cytosine (C). In RNA, thymine is substituted by uracil (U). By encoding sequence complementarity, molecular selfassembly can be exploited to pull nucleic acids like a rope^c, weave them like a fabric⁴⁷, decorate them like a scaffold 8,0 and recycle 10 them like a thermoplastic. Beyond their recyclability, nucleic acids and potential

Figure 1 | Change of storage needs over time. a, Timeline of stored analogue, digital and total data (ref. 48) where the percentage values refer to the fraction of stored digital data. b, Projected global memory demand. Actual (filled circles: i, ref. 49; ii, ref. 50; iii, ref. 51) and projected (open circles; iv, ref. 51; v, ref. 52) data points fall between the conservative estimate and the upper bound. See also Supplementary Information section 1.

NATURE MATERIALS | VOL 15 | APRIL 2016 | www.nature.com/haburematerials

© 2016 Macmillan Publishers Limited. All rights reserved

Dielectric Thin Film Lab, Seoul National University

366

Consciousness?

shadows of the mind

GER PENROSE

A SEARCH FOR THE MISSING SCIENCE OF CONSCIOUSNESS

- A. All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations.
- 3. Awareness is a feature of the brain's physical action; and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke awareness.
- C. Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally.
- \mathcal{D} . Awareness cannot be explained by physical, computational, or any other scientific terms.

"John Conway" Game of Life

Cellular Automata

http://www.youtube.com/watch?v=CgOcEZinQ2I Dielectric Thin Film Lab, Seoul National University

