# Energy efficient and Cognitive computing using memristors 

Computer; where is it from and heading to?

## Cheol Seong Hwang

2018. September 12 $^{\text {th }}$ Physics, SNU

## Energy crisis

## ~40,000 exabytes in 2018


by 2040
Total binary operation of

## 10

Digital Information Created Each Year, Globally


## 2,000\%

Expected increase in
global data by 2020
III
Megabytes
Video and photos stored by Facebook, per user
$75 \%$
Percentage of all digital data created by consumers

Energy consumption of U.S data centers in 2013

That's equivalent to the output and pollution of

(Annual output of a plant=500MW) NRDC,Aug. 2014

## Ultimate energy needed for one logic operation

## kT In2 ~ 3x10 ${ }^{-21}$ Joule

## $10 \mu \mathrm{~A} \times 10 \mathrm{~ns} \times 1 \mathrm{~V} \sim 10^{-13}$ Joule

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

> Decrease computation itself

1. Introduction

IBM Journal, July 1961, p. 183

## From mathematics to computer



Kurt Gödel
Incompleteness theorem


Alan Turing
Turing machine


David Hilbert
Formalism

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 ?

## Automatic Computing Machine (ACE)



Dielectric Thin Film Lab, Seoul National University

## 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

| AND |  | OR |  | NOT |  | NOR |  | NAND |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $x^{x}=\square-{ }^{F=x \cdot y}$ |  | $x-\underbrace{F=x+y}$ |  | $x-$ - $-\bar{x}$ |  | $x^{x} \mathrm{D}^{--^{F=\bar{x}+y}}$ |  | $x=\square^{0-m=\overline{x, y}}$ |  |
| A B | C | A B | C | A | B | A B | C | A B | C |
| 00 | 0 | 00 | 0 | 0 | 1 | 00 | 1 | 00 | 1 |
| 01 | 0 | 01 | 1 | 1 | 0 | 01 | 0 | 01 | 1 |
| 10 | 0 | 10 | 1 |  |  | 10 | 0 | 10 | 1 |
| 11 | 1 | 11 | 1 |  |  | 11 | 0 | 11 | 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

More than Moore: Upgrade.
"New technologies for heterogeneous integration of multiple functions"


More Moore: Keep going.

"Conventional dimensional and functional scaling"

Beyond CMOS: Totally new.
"Invention of a new information processing paradigm"


## Future of Lithography for scaling

- Next technology : EUV (Extreme Ultraviolet)


## EUV tool principle


$\rightarrow$ The higher NA (or lower $k_{1}$ ) EUV tools enable resolutions down to 11 nm
$>$ ASML EUV Product Roadmap


- Current Status : 30~40 wf/1hr
- Throughput $\geq 500 \mathrm{wf} /$ day (Target $\geq 1200 \mathrm{wf} / \mathrm{day}$ ) (ArF Immersion 250 wf/1hr)

| Cost | EUR 42 million <br> (\$60 million) | (maybe) $\$ 170$ million |
| :--- | :---: | :---: |

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

## Storage memory: VNAND vs PNAND

## Planar NAND



128 Gb PNAND Chip size : $173.3 \mathrm{~mm}^{2}$

## Vertical NAND

24 layer stacked, 128 Gb VNAND Chip size: $133 \mathrm{~mm}^{2}$


- 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 256 Gb , 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

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


## New computer system organization



## Adv. Electron. Mater., 2015, 1400056

# Prospective of Semiconductor Memory Devices: from Memory System to Materials 

Sheol Song Huang*

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, fabricatimon technology, and materials points of view. Based on this knowledge, the limitations of the present technologies are described, and the possible solutons suggested up to now are reassessed. Finally, a shift in the fundamental computational paradigm from vo 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 44000 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

## New computers?

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


## Von Neumann computer vs. Human brain



Dielectric Thin Film Lab, Seoul National University

## Energy for $2+3$ using Von Neumann computing

Most of energy on conversion


## Neuron and synapse



Dielectric Thin Film Lab, Seoul National University

## Neuromorphic computing architecture



## Cognitive computing



Google AI computer beats human champion of complex Go boardgame


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 Nature 529, 484-489 (28 January 2016 the results of the training. supportea by $<$ billon neurosynaptıc cores contaınıng 330 Improved from billion neurons and 100 trillion synapses running only 1542 times slower than real time.

o stacerul
density crisis

Human brain inspired computing based on Memristor (cell size< $\mathrm{F}^{\mathbf{2}}$ )

## "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 5 ${ }^{\text {th }}, 2016$

## Memristor: ReRAM

## Memristor-The Missing Circuit Element <br> IEEE TRANSACTIONS ON CIRCUIT THEORY, VOL. CT-18, NO. 5, SEPTEMBER 1971 <br> LEON O. CHUA, SENOR MEMBER, IEEE

Although no physical memristor has yet been discovered, quasi-static electroms notion that a memristor
make plausible the notic device could be invented. if not discovered accidentally!
It is perhaps not unreasonable to suppose that such a device might already have been fabricated as a laboratory curiosity but was
improperly identified! ...

Charge-controlled memristance
$v=\frac{d \varphi}{d t}=\frac{d \hat{\varphi}(q)}{d q} \frac{d q}{d t}=R(q) i$
where
$R(q) \triangleq \frac{d \hat{\varphi}(q)}{d q}$

Flux-controlled memductance

$$
i=\frac{d q}{d t}=\frac{d \hat{q}(\varphi)}{d \varphi} \frac{d \varphi}{d t}=G(\varphi) v
$$

where

$$
G(\varphi) \triangleq \frac{d \hat{q}(\varphi)}{d \varphi}
$$

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



- 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.



## Memristors for new computing paradigms

## Neuristor by HP group

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


NATURE MATERIALS | VOL 12 | FEBRUARY 2013

## $\checkmark$ Synaptic adaptation by S.H. Jo et al.

Spike-timing-dependent plasticity of potentiation and depression in synapse.


Nano Lett. 2010, 10, 1297-1301
$\checkmark$ 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

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

Julien Borghetti ${ }^{1}$, Gregory S. Snider ${ }^{1}$, Philip J. Kuekes ${ }^{1}$, J. Joshua Yang ${ }^{1}$, Duncan R. Stewart ${ }^{1} \dagger$ \& R. Stanley Williams ${ }^{1}$

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

lemristors are non's original or state for the ce of a switch. les both store me nanoscale tes or memory n a nanoscale themselves. 1d memristive 770s, but were 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 | $=((p$ IMP $q)$ IMP $((q$ IMP $p)$ IMP 0) $)$ IMP 0 |

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


## Logic cascading: Full adder



|  |  | $\mathbf{P}$ | $\mathbf{Q}$ | $\mathbf{T}$ |
| :--- | :--- | :---: | :---: | :---: |
|  | Input | $(p ?)$ | $(q$ ? $)$ | $(t ?)$ |
| Step 1 | Write $\mathrm{P}=(p q), \mathrm{Q}=(q p)$ | $(p q)$ | $(q p)$ | $(t ?)$ |
| Step 2 | Execute $p^{\prime} \leftarrow(p$ XOR $q), q^{\prime} \leftarrow(p$ AND $q)$ | $\left(p^{\prime} x\right)$ | $\left(q^{\prime} y\right)$ | $(t ?)$ |
| Step 3 | Write $\mathrm{P}=\left(p^{\prime} t\right), \mathrm{T}=\left(t p^{\prime}\right)$ | $\left(p^{\prime} t\right)$ | $\left(q^{\prime} y\right)$ | $\left(t p^{\prime}\right)$ |
| Step 4 | Execute $p^{\prime \prime} \leftarrow\left(p^{\prime}\right.$ XOR $\left.t\right)=s, t^{\prime} \leftarrow\left(p^{\prime}\right.$ AND $\left.t\right)$ | $\left(s x^{\prime}\right)$ | $\left(q^{\prime} y\right)$ | $\left(t^{\prime} z\right)$ |
| Step 5 | Write $\mathrm{Q}=\left(q^{\prime} t^{\prime}\right)$ | $\left(s x^{\prime}\right)$ | $\left(q^{\prime} t^{\prime}\right)$ | $\left(t^{\prime} z\right)$ |
| Step 6 | Execute $q^{\prime \prime} \leftarrow\left(q^{\prime}\right.$ OR $\left.t^{\prime}\right)=c$ | $\left(s x^{\prime}\right)$ | $\left(c y^{\prime}\right)$ | $\left(t^{\prime} z\right)$ |

## 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


## 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)


Memristors for Energy-Efficient New Computing Paradigms
Doo Seok Jeong, Kyung Min Kim, Sungho Kim, ByungJoon Choi, and Cheol Seong Hwang*

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 vo 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 circu 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) also saves the energy required for the data refresh. Neuromorphic (cognitiv
computing refers to a computing paradigm that mimics the human brain. 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 rec ognition utilizing the recently highlighted deep learning algorithm. However the human brain typically consumes $\approx 10-20$ 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 substan tially decrease the energy consumption for futuristic information technology.
that the total amount of digital data in 2040 (only 25 years from now) will be $=10^{28}$ bytes (1 byte $=8$ bits), which is approximately one million times greater than the current total. ${ }^{[2]}$ A more interesting (and also sobering) expectation is that the total number of binary operations in 2040 will be $=10^{0-1}$, which is an astro 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 $=10^{27}$ 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 $=k T$ (Boltzmann constant ture), which is $=10^{-21}$ Joules at room

## 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
temperature This estimate means that the aforementioned energy consumption in 2040 could be decreased by a factor of $=10^{8}$, 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


Korea Institute of Science and Technology
5 Hwarang-ro 14 -gil, Seongbuk-gu, Seoul 02792 , Republic of Korea Dr. K. M. Kim
Hewlett Packa
Hewlett Packard Laboratories
Hewlett Packard Enterprise
Palo Alto, California 94304, USA
Prof. S. Kim
Department of
Sejong University DOI: 10.1002/aelm. 201600090

## Future (von Neumann) computer ? -overcoming quantum limitation-tunneling

- Processing: electron - power, speed
- Communication: photon - speed, power
- Memory: jo 1 - 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?

or Nick Goldman, the sce
DNA started out as a oke
It was Wednesday 16 February 2011, and It was Wednesday 16 February 2011, and
Goldman was at a hotel in Hamburg, Germany, Goldman was at a hotel in Hamburg, Germany,
talkting with some of his fllow bolonformatictists
about how they could a fford to store the reams of genome sequences and other data the world was getting so frustrated by the expense and limittations of conventional computing technology that they started kiddding about sc1-fa laternaThen the laughter stopped "It was a lightbulb moment," says
Goldman, a group leader at the European Boolniormatics Instltute (EBD)
 sticcon memory chy. It would take hours to encode data by symthestz-
ing DNA strings with a specific pattern of hase, and stll more hourt recover that tiformation using a a sequencing machine. But with DNA,
whole humman genome fits into a cell that is Ifvistle to the naked eye Fi sheer denstry of informatton storage, DNA could be orders of magnttude -We sat down it the ber th nather and
and started crribbiling ideas "What would you have to do to to make that work" The researchers' bggest worry was that DNA synthesss and
sequencting made mistakes as often as in th very 100 nucleotides. Thls
they could And a workable error-ccrrection scheme. Could they encode
bitts into base parts in a way that would allow them to detetet and undo the mistakes? " Wuthin the course of an evening." says Goldman, "we
knew that you could" knew that you could
He and hs $\operatorname{sic}$
He and hst EBI colleague Ewan Birney took the Idea back to ther labs
and two years later announced that they had succesfully used DNA and two years later announced that they had successfully ysed DNA
to encode five fles, including Shakespeares sonnets and a snlppet of
Martin Uther Kings Thave adraan' ppeect Mo encode five files, Incluading Shakespeares. sonnets and a snlppet of
Martmn Luther King 5 Thave a drean's ppeech'. By then, blologstst George Church and his team at Harvard Untversty in Cambridge, Massachesetts had unveled an independdent demonstration of DNA encoading
Butat 739 kllobytes $(\mathrm{kB})$, the EBI ille comprised thelargest DNA archi/ But at 739 kilobytes (kB), the EBS flles comprised thelargest DNA ardh
ver produced - untll July 2016 , when researchers from Micrusoft and the Unviveratty of Washing ton damened a leeap to 2000 megsibtes (MB). The latest experiment signals that interest In using DNA as a storage
medum is surging far beyond genommos the whole world is acing adata)
crunch Counting everything from astronomical images and journal crunch. Counting every thing from astronomical Images and Journa 44 trillion gigabytes (GB) by 2020 a tenfold increase over 2013. By 2000 , If every thing were stored for instant access in, say, the flash memory the expeted supply of microchlp-grade stilcon?
That Is one reason why permanent archlves of rarely accessed data currently rely on old-fashloned magnettc tapes. This medium pads tor read Yet even that approach Is becoming unsustannable, syys David Advanced Research Putrojects nectrosclientist at the US Intell pence It ts posstble to Imagine a data centre holding an erabyte (one bilic gigay bullion over 10 years to bulld and malntafn, as well as hundred of megawatts of power. "Molecular data storage has the potential to
reduce all of those requirements by up to three orders of magnitude" syys Markowitr. If nifformathon could be packaged as densely as it is 1
 could be met by about a allogram of DNA (see 'Storage limits).
Achteving that potenttal wont be easy. Before DNA can become
viable compettor to conventional torage technologtes,
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.




## Consciousness?

## shadows of the mind <br> 信 ROGER PENROSE



Dielectric Thin Film Lab, Seoul National University

## RMCEIR Plinilst



> A SEARCH FIR THE MISSING SCIIACE OF COUSCIOUSNESS
$\mathscr{A}$. All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations.
$\mathscr{B}$. 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.
$\mathscr{C}$. Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally.
$\mathscr{D}$. Awareness cannot be explained by physical, computational, or any other scientific terms.

## "John Conway" Game of Life

(1)


Cellular Automata

