ASP-DAC 2011
26th January 2011
Yokohama, Japan

Non-Volatile Memory and Normally-Off Computing

T. Kawahara
Central Research Laboratory, Hitachi, Ltd.

Acknowledgements

Original works cited in this presentation were supported in part by;
• The "High-Performance Low-Power Consumption Spin Devices and Storage Systems" program (headed by Professor Hideo Ohno of Tohoku University) under Research and Development for Next-Generation Information Technology of MEXT,
• One of projects on "Fundamental Technologies for the Next Generation Supercomputing" by MEXT, and,
• The Japan Society for the Promotion of Science (JSPS) through its "Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program). "

The author would like to thank to;

Tohoku University: S. Ikeda, T. Meguro, R. Sasaki, M. Yamanouchi, I. Morita, T. Hirata, T. Hanyu, and H. Ohno.

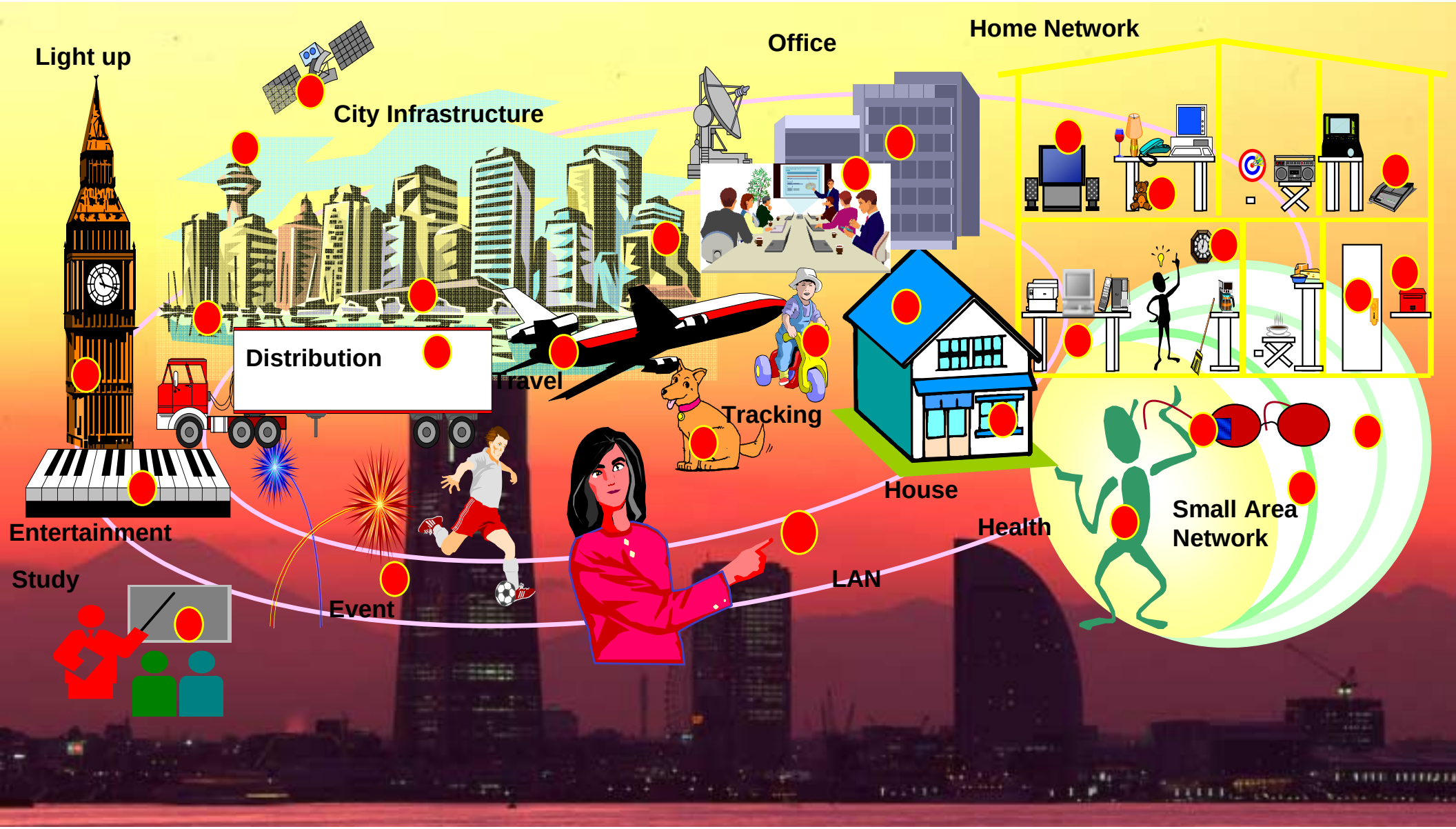
Hitachi: R. Takemura, K. Ono, T. Ishigaki, T. Yamada, A. Kotabe, S. Hanzawa, S. Yamaguchi, K. Miura, H. Yamamoto, J. Hayakawa, N. Matsuzaki, Y. Mouri, K. Ito, H. Takahashi, H. Hasegawa, Y. Goto, N. Osakabe, and H. Matsuoka.

Outline

- Introduction
- Non-volatile memories
- Spin-transfer torque RAM scalability
- Normally-off instant-on computing
- Conclusion

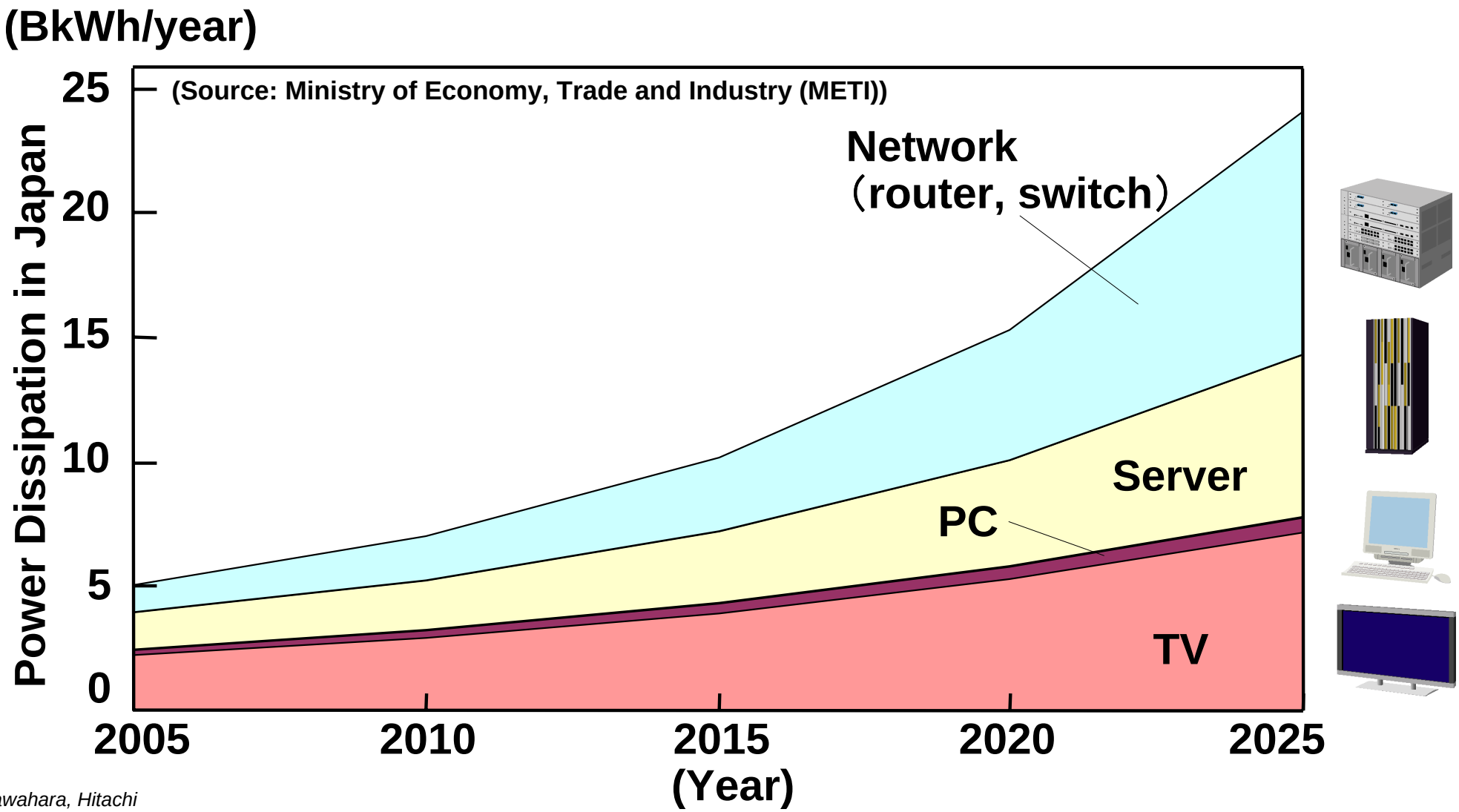
4 Social system needs large power

- **World is power hungry, but...**



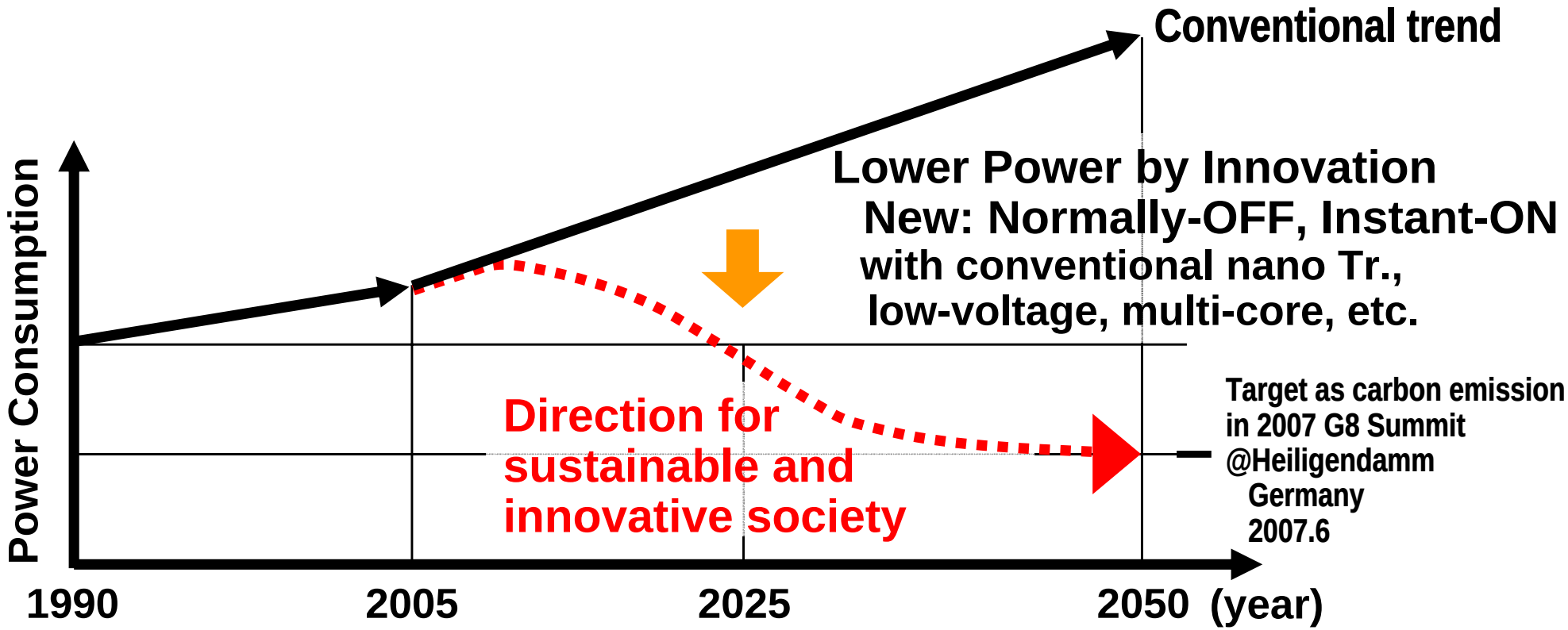
Necessity of low power IT

- **In 2025, five times larger power consumption is expected in Japan than that in 2005.**
- **Will reach nine times in the world.**



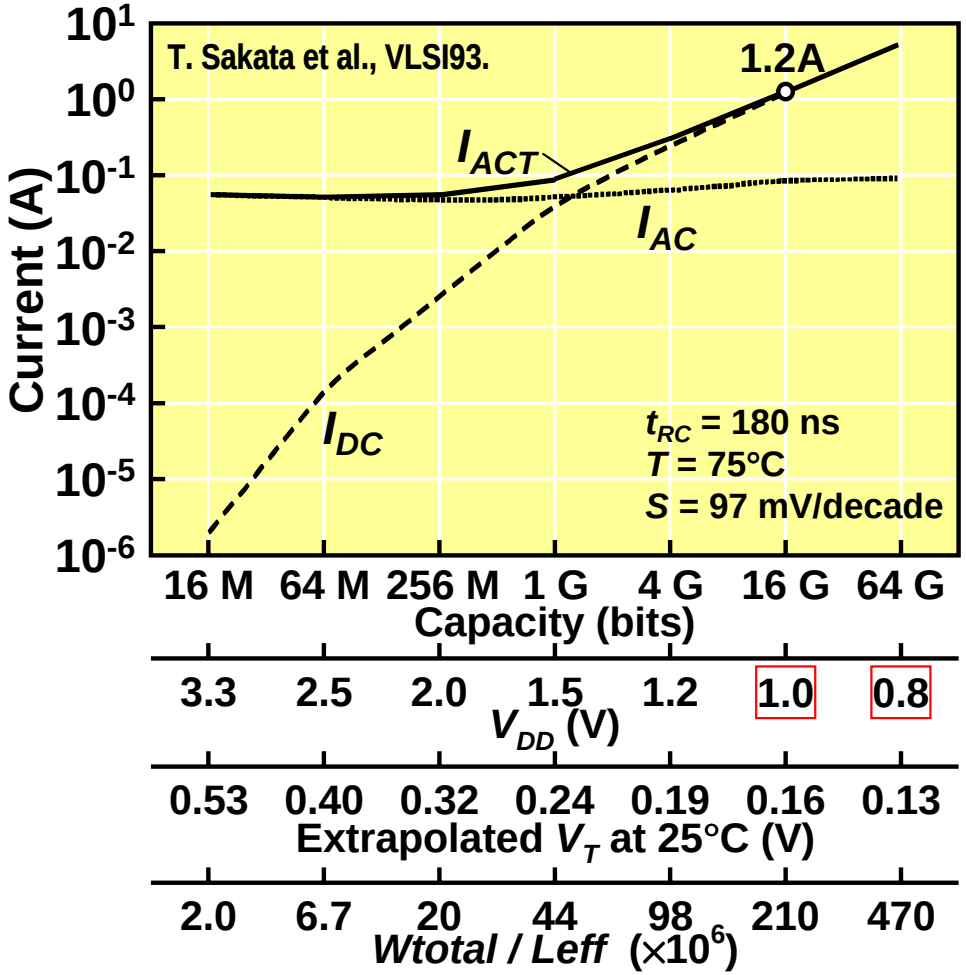
6 Sustainable and innovative society

- **Solutions are not only by LSI technology, but LSI can contribute in many for sustainable world.**

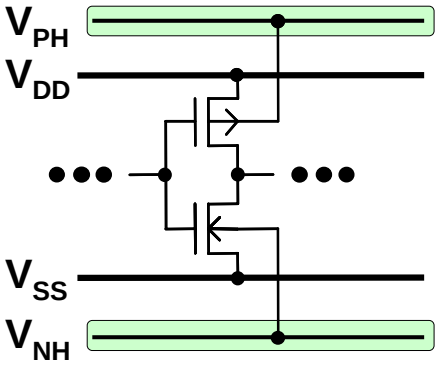


Beyond Low Voltage Operation

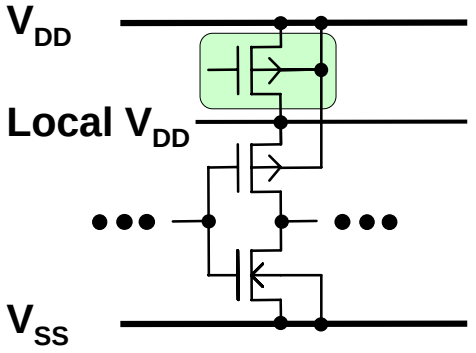
- **Already made extensive progress for low voltage operation with high performance in last twenty years.**
- **Complementarily, a new kind of innovation for further.**



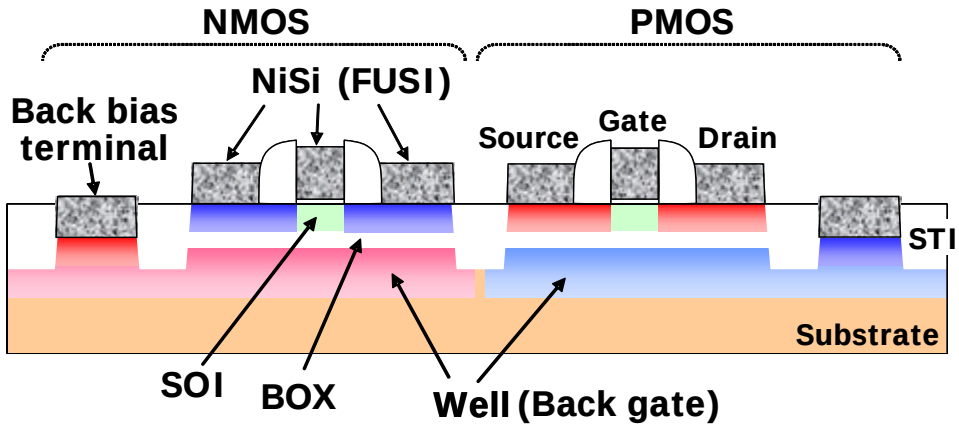
1993: Projected Current Consumption (Hitachi)



1993: Well-driven SA (Mitsubishi)
T. Ooishi, et al., VLSI93.



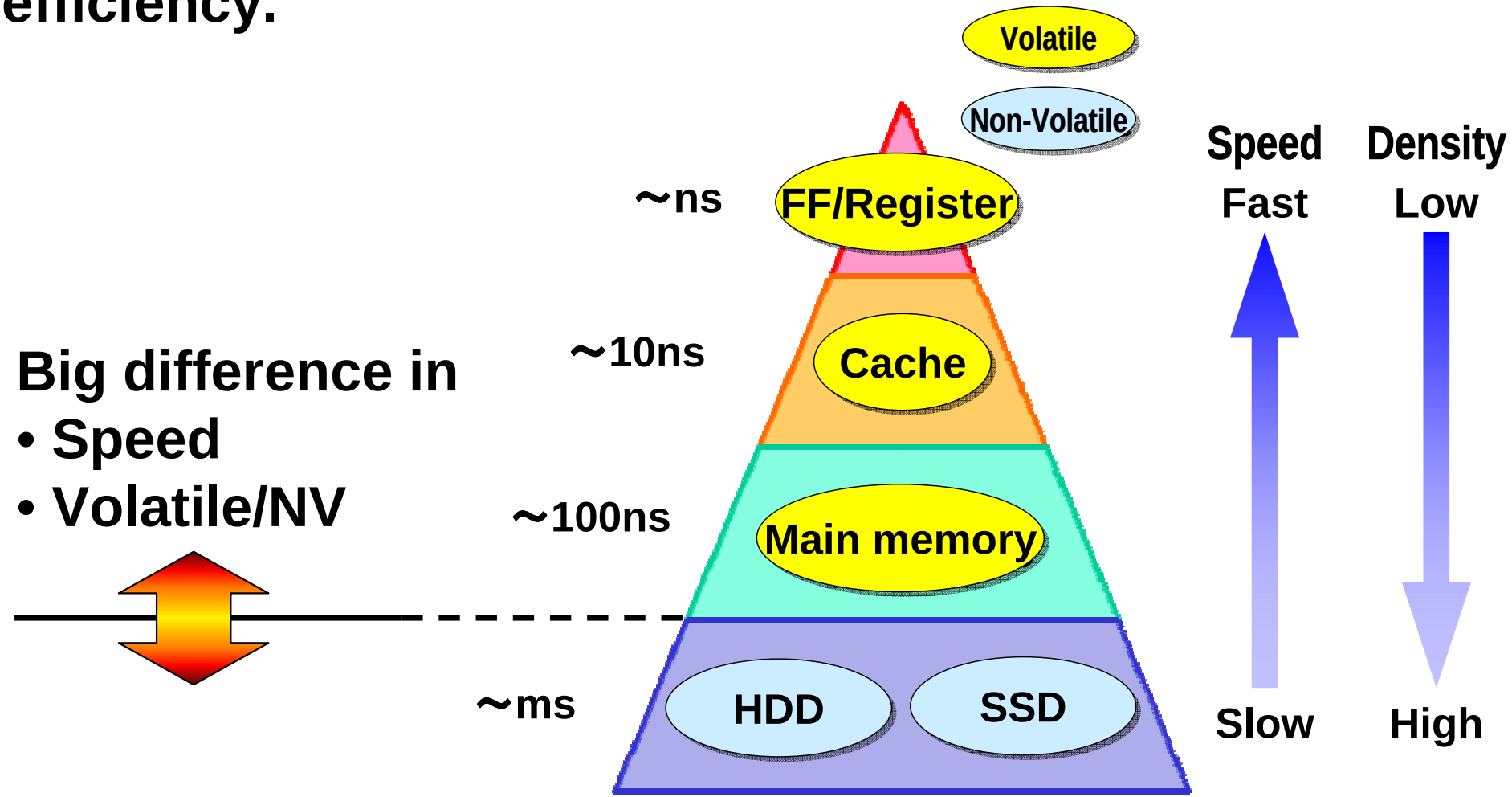
1993: 256Mb DRAM (Hitachi)
G. Kitsukawa, et al., ISSCC93.



2004: Dopeless Channel CMOS (SOTB) (Hitachi)
R. Tsuchiya, et al., IEDM04.

Memory and Computing

- Memory systems are constructed according to a deep hierarchy based on speed and capacity. Volatile memory and nonvolatile memory are often combined
- Leads to slow start-up, long idle times, and low power efficiency.

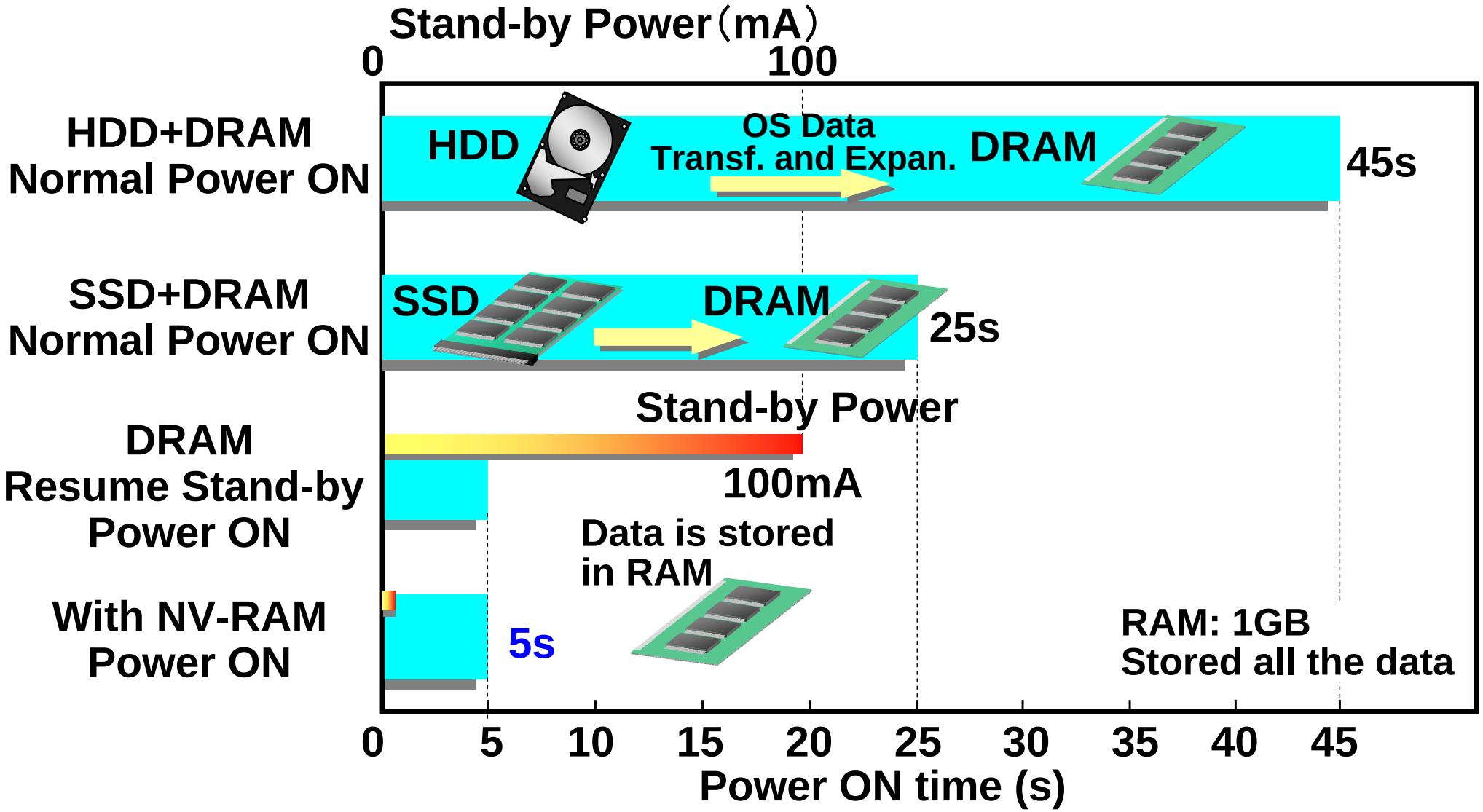


9 Normally OFF, Instant ON Computing

- New innovation in addition to low-voltage.
- Normally OFF, Instant ON
  - Turn off anytime when not in use,
  - but operate instantly with full performance when needed.
- Need data stored in the state of operation before any turned off, with zero power for retaining data.
  - > Nonvolatile RAM (NV-RAM)
- NV-RAM: Infinite number of fast write and read operations with non-volatility.
- Adaptive solutions in wide time domain, and wide area domain. LSI level to System level.

Example (but not ideal yet)

- Power ON time is improved to 1/9 with NV-RAM.
- Nonvolatility achieves low power also.

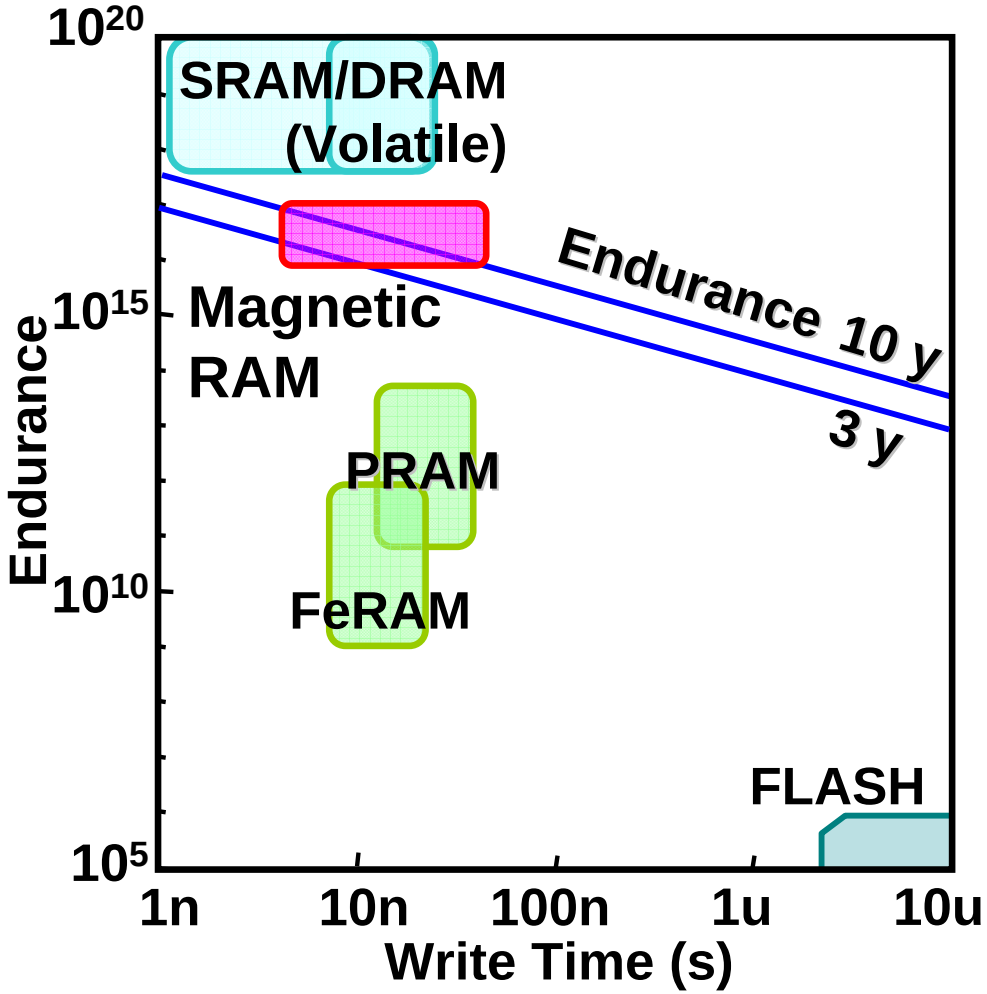


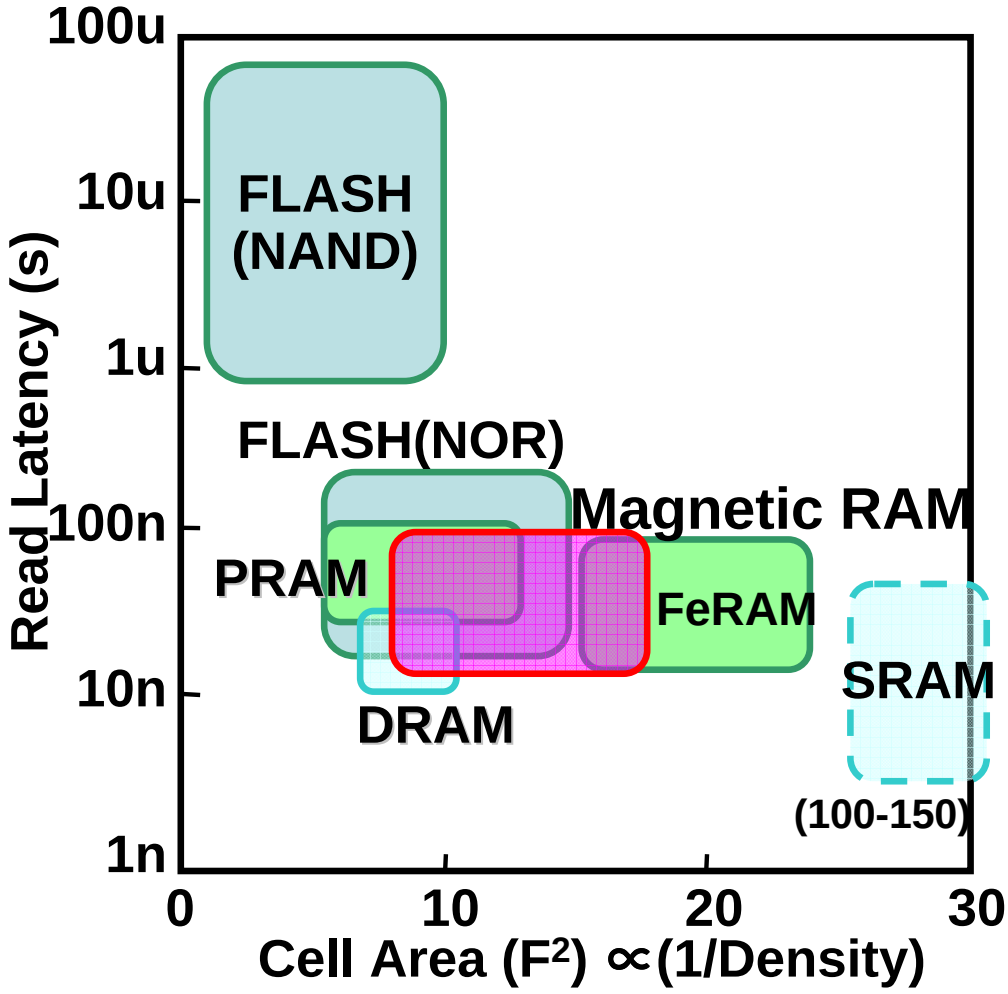
Outline

- Introduction
- **Non-volatile memories**
- Spin-transfer torque RAM scalability
- Normally-off instant-on computing
- Conclusion

RAM Comparison

- NV-RAM: Infinite # of write cycle, non-volatile, fast R/W.
- Magnetic RAMs are good candidates for NV-RAM.



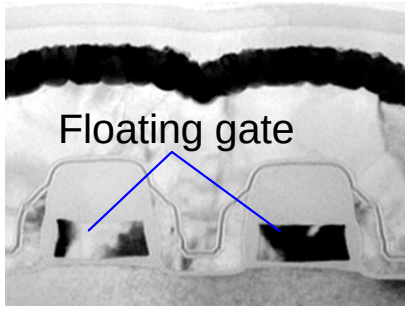


General Memory Performance Map

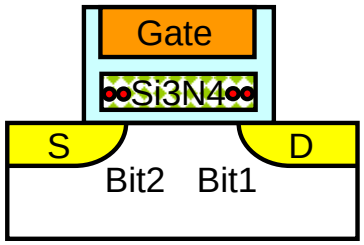
Origin of Non-volatility

- **Various phenomena is applicable: insulator barrier, change of structure (bistable), and alignment of electron condition (bistable).**

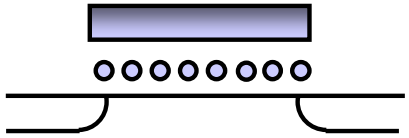
Insulator barrier



Floating gate

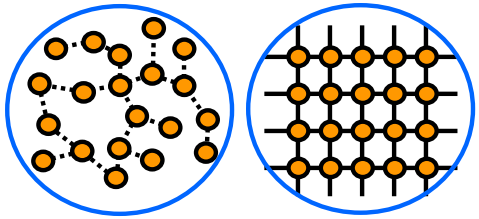


Silicon nitride

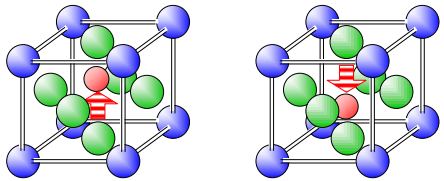


Nano-dots

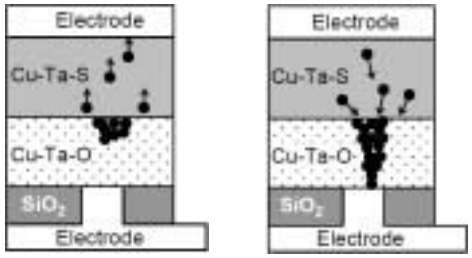
Change of structure



Phase change

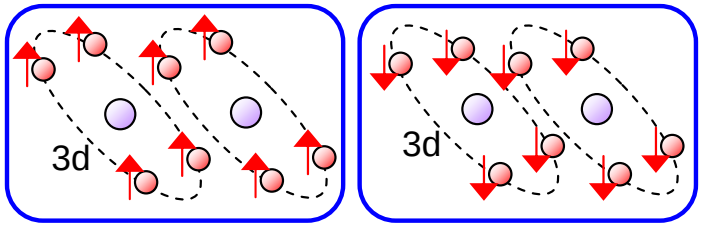


Ferroelectrics



Ionization

Alignment of electron condition



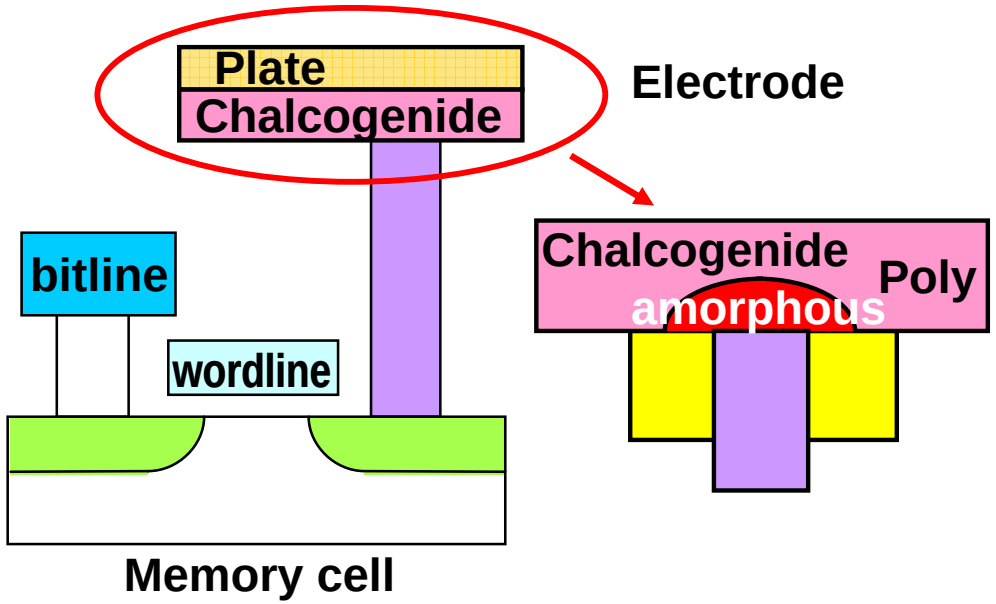
Ferromagnetism

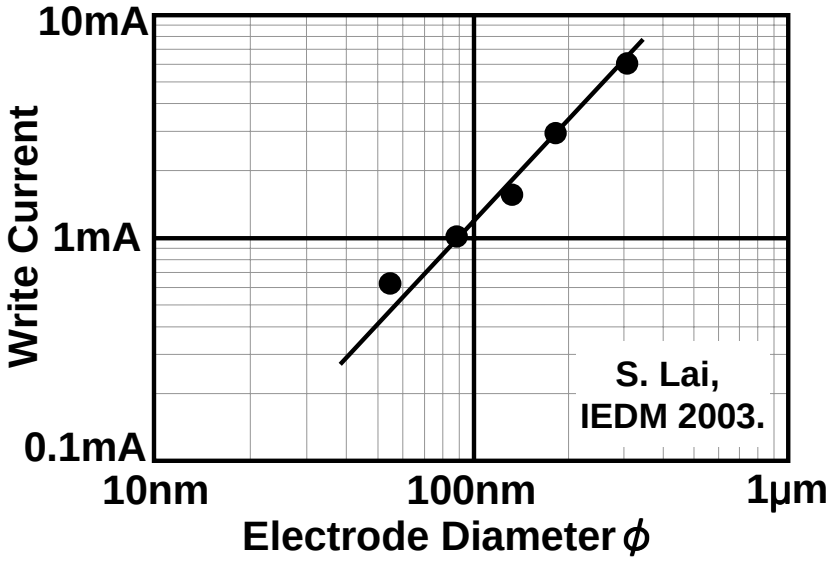
(+ Magneto resistive effect for memory)

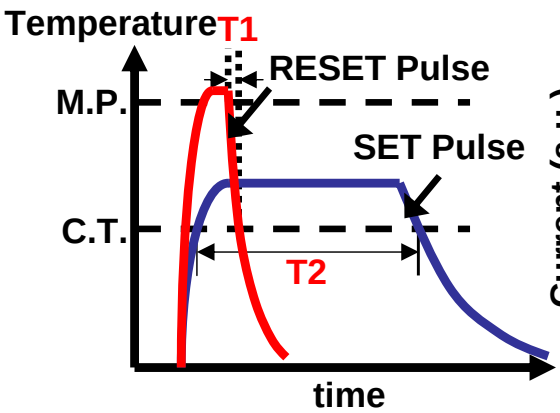
-> Easier to achieve infinite write cycle (endurance).

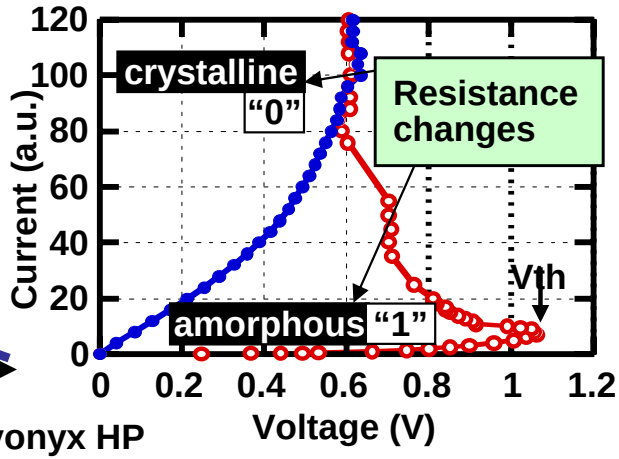
14 Phase Change Memory

- **Resistance change between amorphous and poly of chalcogenide material. Large resistance change.**
- **Good scalability.**

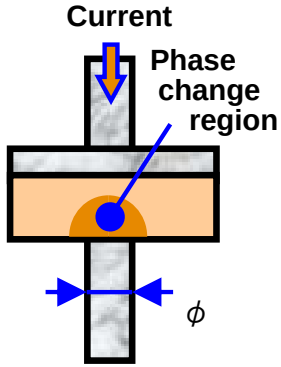


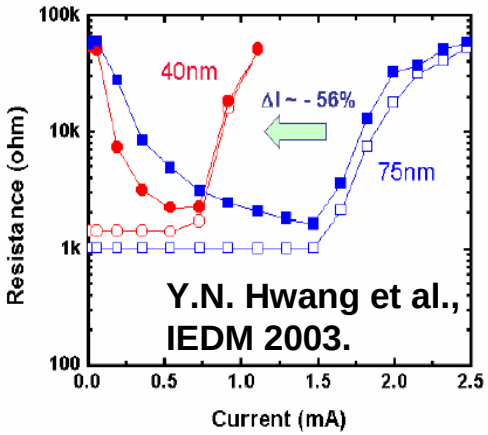






from Ovonyx HP

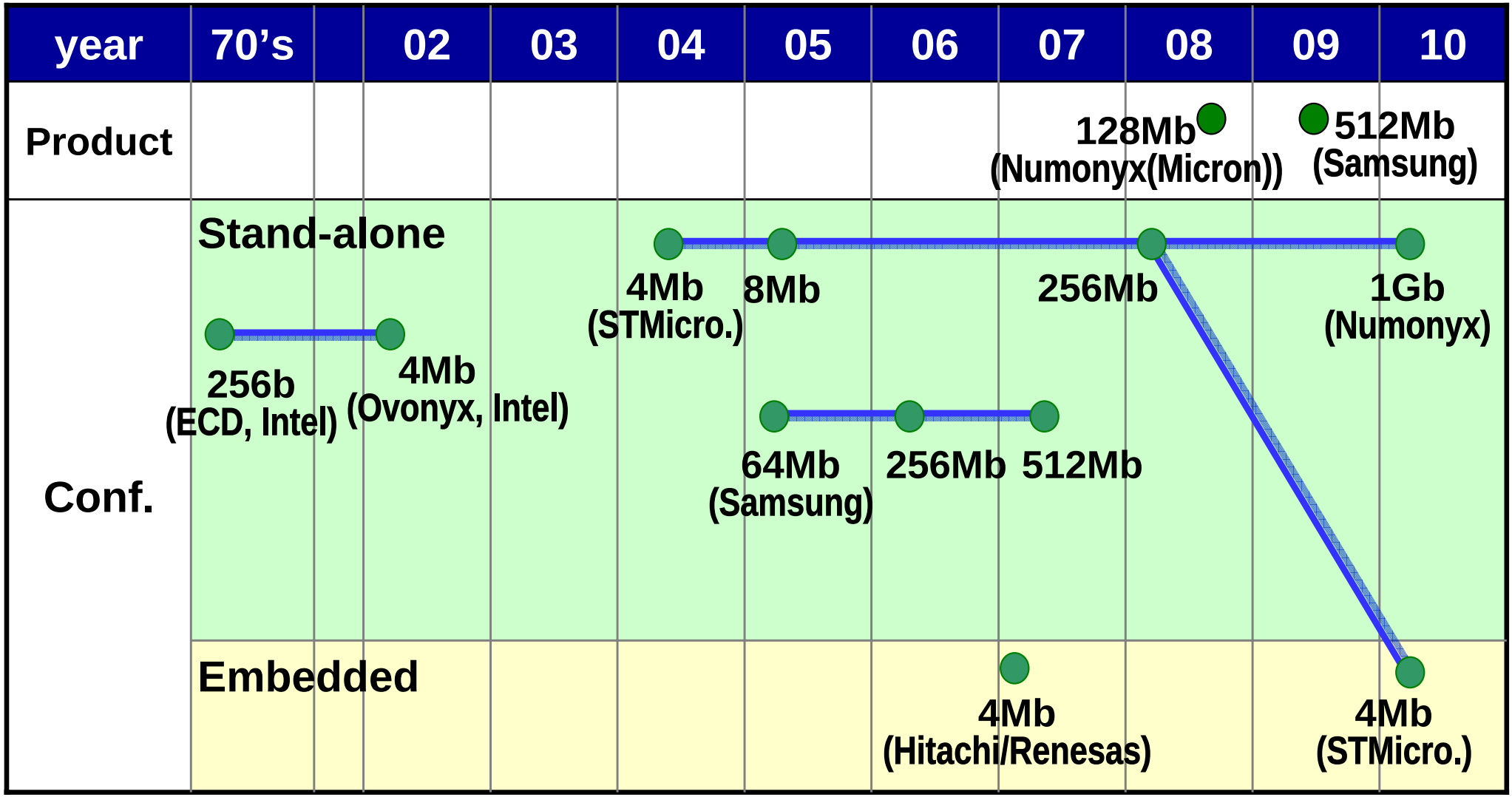




T.Kawahara, Hitachi

15 Development of Phase Change Memory

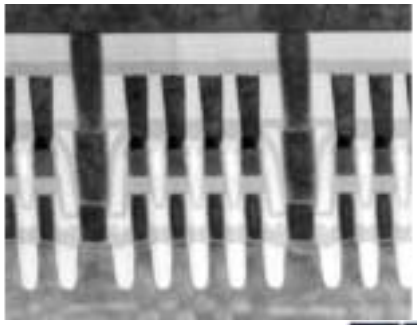
- Principle was announced in the 70's.
- Production as NOR flash replacement in recent years.



Phase Change Memory Chips

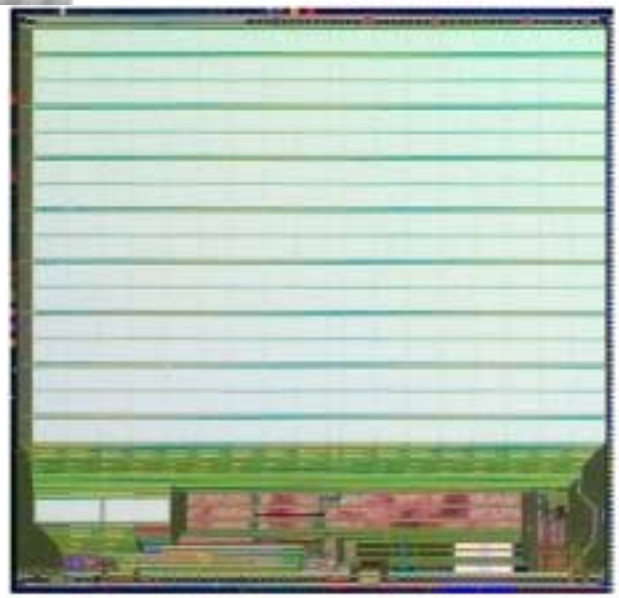
Stand-alone (Numonyx)

- Process: 45 nm
- Power Supply: 1.8 V
- Density: 1Gb
- Chip size: 37.5 mm^2
- Cell size: 0.011 μm^2 (5.5F^2)
- Cell Tr: BJT
- Endurance: ~10^9



Memory Cell

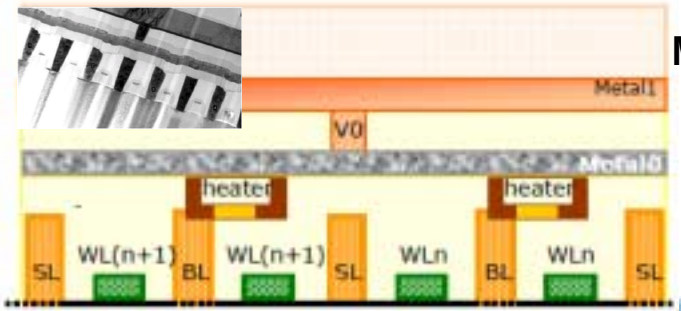
Chip photo



C. Villa, et al., ISSCC 2010
G. Servalli, IEDM 2009

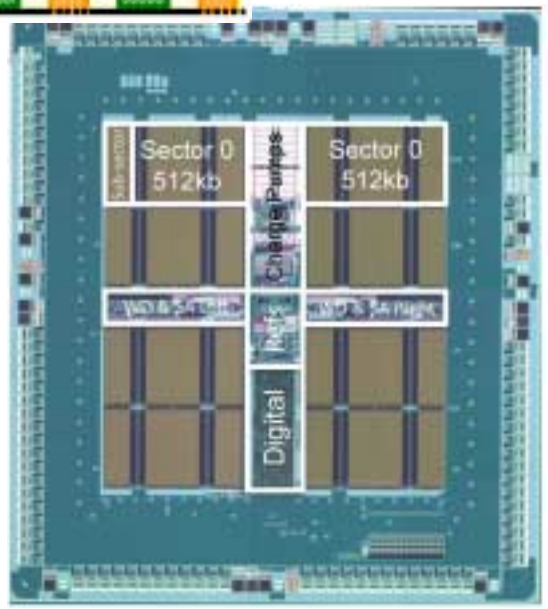
Embedded (STMicroelectronics)

- Process: 90 nm
- Power Supply: 1.2 V
- Density: 4Mb
- Chip size: 3 mm^2
- Cell size: 0.29 μm^2 (36F^2)
- Cell Tr: MOS
- Endurance: ~10^7



Memory Cell

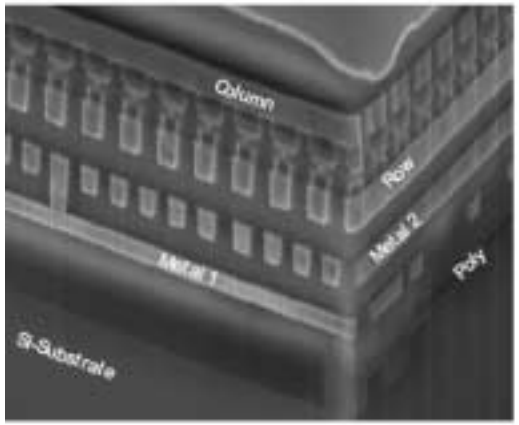
Chip photo



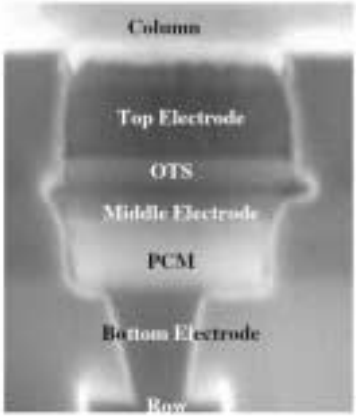
G. De Sandre, et al., ISSCC 2010
R. Annunziata, et al., IEDM 2009

17 Stacked Phase Change Memory

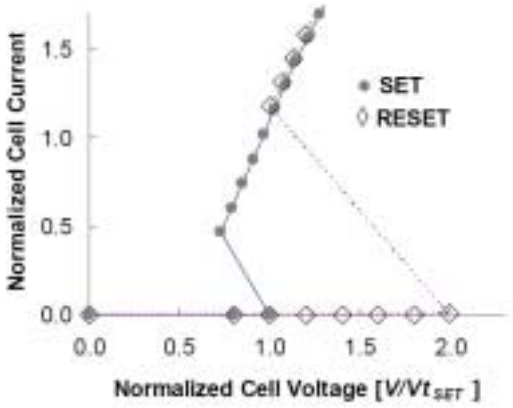
- **Selecting element and the memory device are stacked.**
- **Selecting element by non-substrate Si, formed on wiring.**



Chip



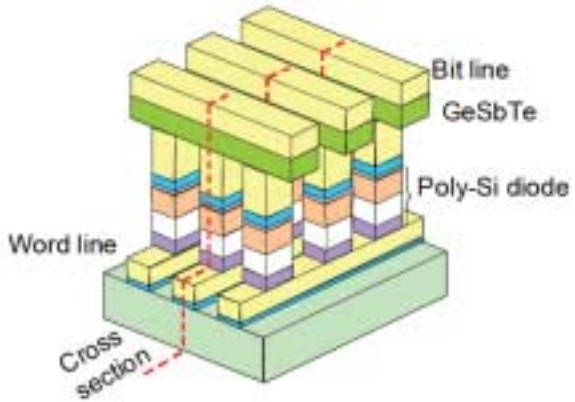
Memory cell



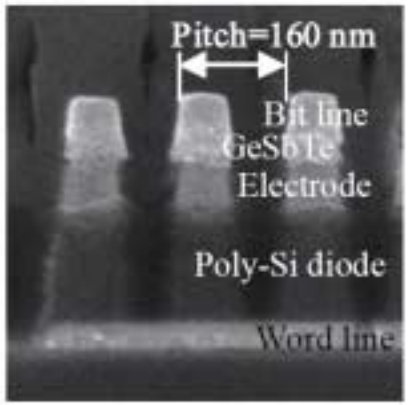
I-V Characteristics

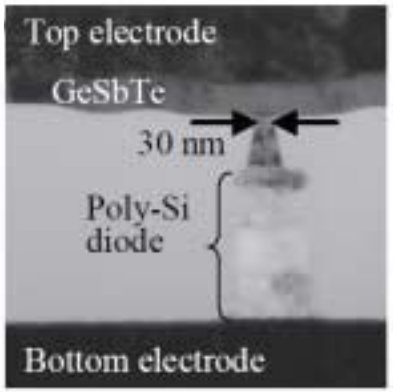
DerChang Kau, et al., IEDM 2009

A switch made of phase-change material itself (Numonyx).



Cross sectional view





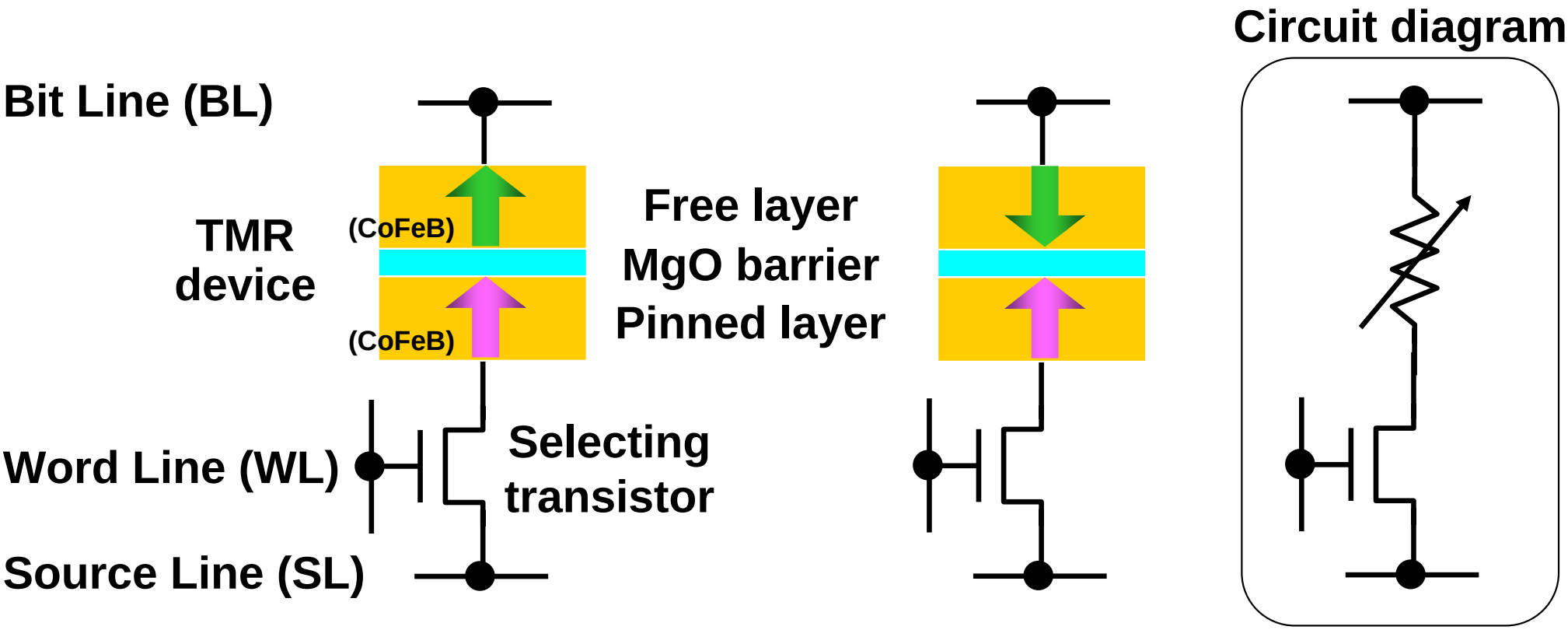
Memory cell

Y. Sasago, et al., VLSI Tech. 2009

Polysilicon diode formed on wiring. Current drivability: 160uA@30nm.

TMR Device and Memory cell

- **Resistance changes between parallel and anti-parallel.**
- **Use this two-state as an information bit.**



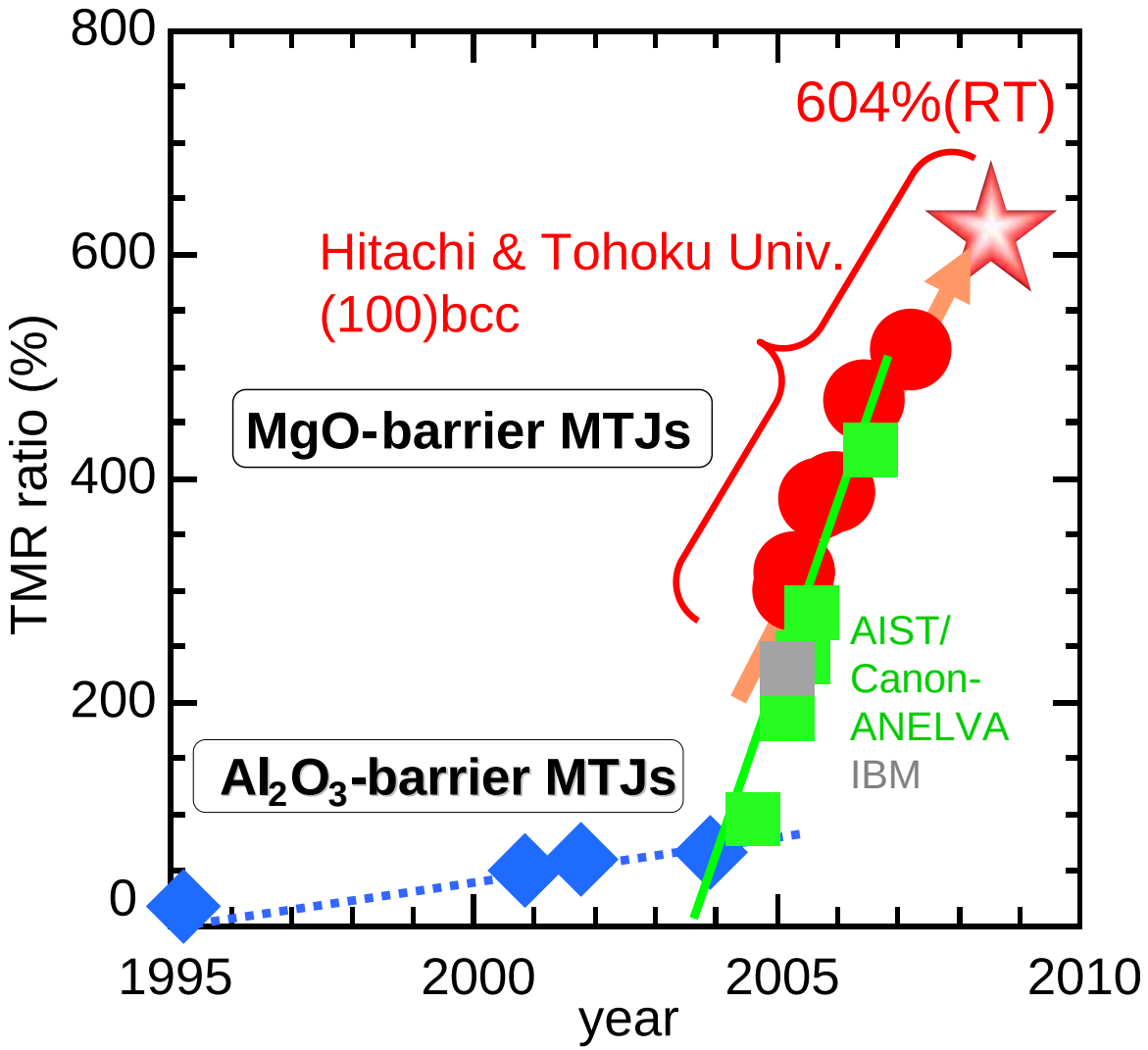
Parallelized State (Low Resistance: R_P)

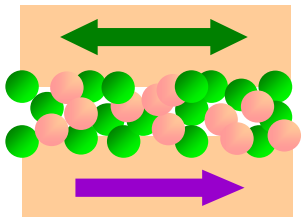
Anti-Parallelized State (High Resistance: R_{AP})

$$\text{TMR ratio} = (R_{AP} - R_P) / R_P$$

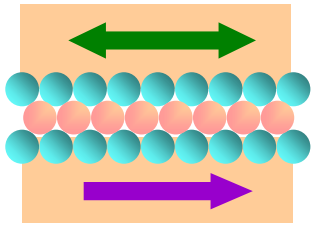
Innovation in TMR device: MgO

- **MgO barrier provides two breakthroughs.**
- **High TMR ratio and small write current.**

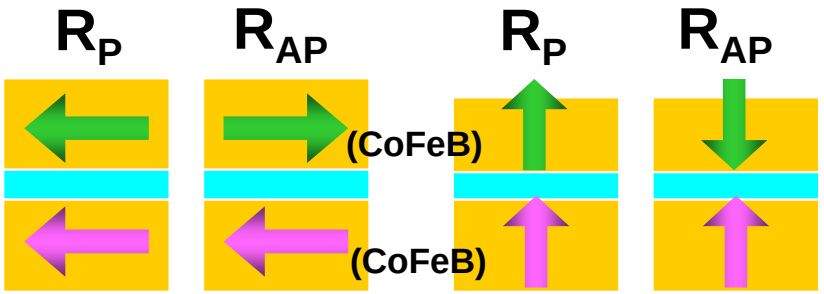




Free layer
AlO barrier (amorphous)
Pinned layer



Free layer
MgO barrier (crystal)
Pinned layer

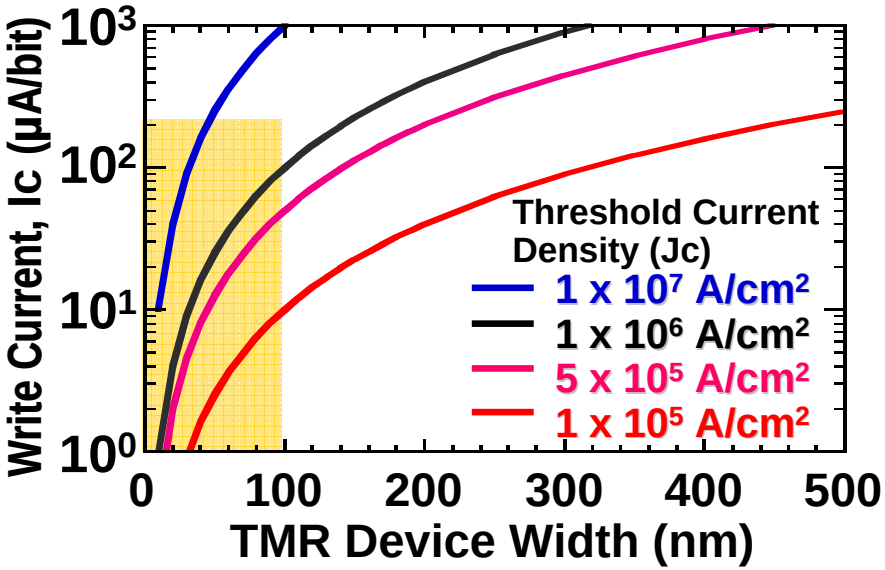


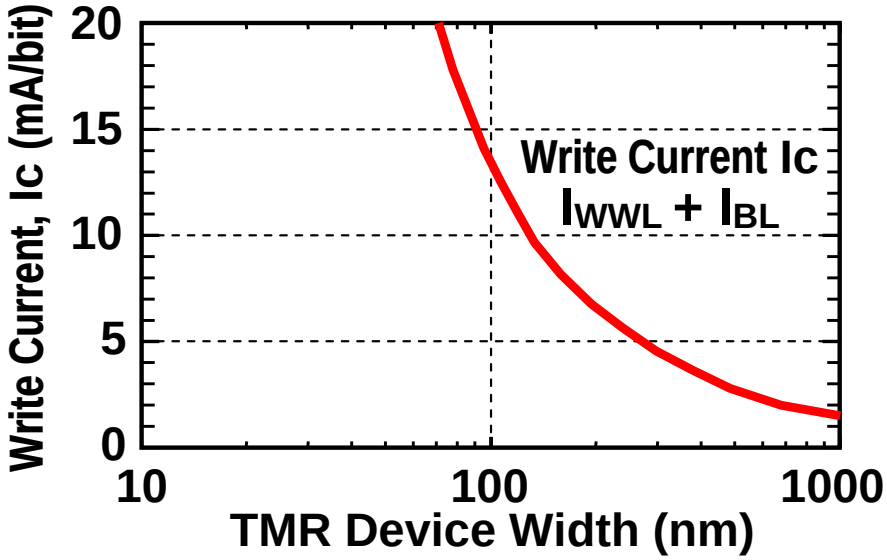
In-plane **Perpendicular**

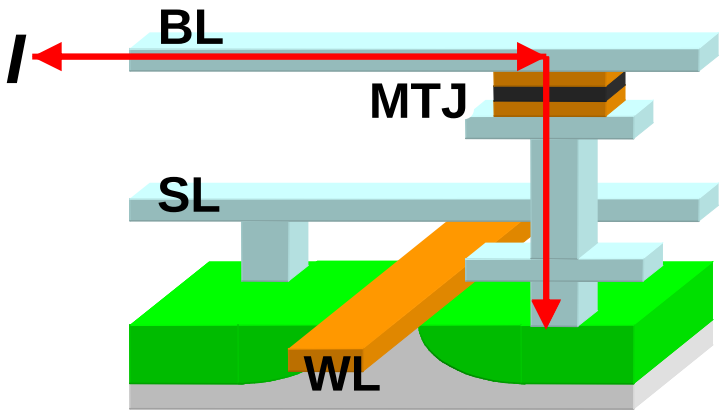
TMR ratio = $(R_{AP} - R_P) / R_P$

SPRAM (SPin-transfer torque RAM)

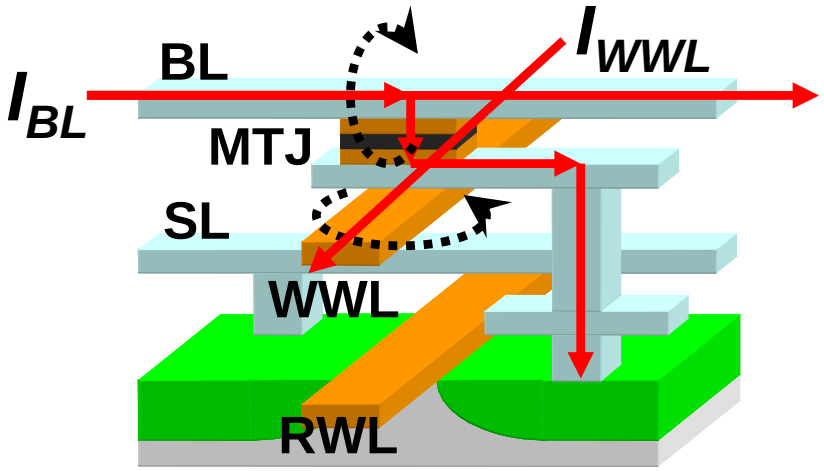
- **MRAM: Matured in product. Inferior in scalability.**
- **SPRAM: Good potential in scalability.**







SPRAM

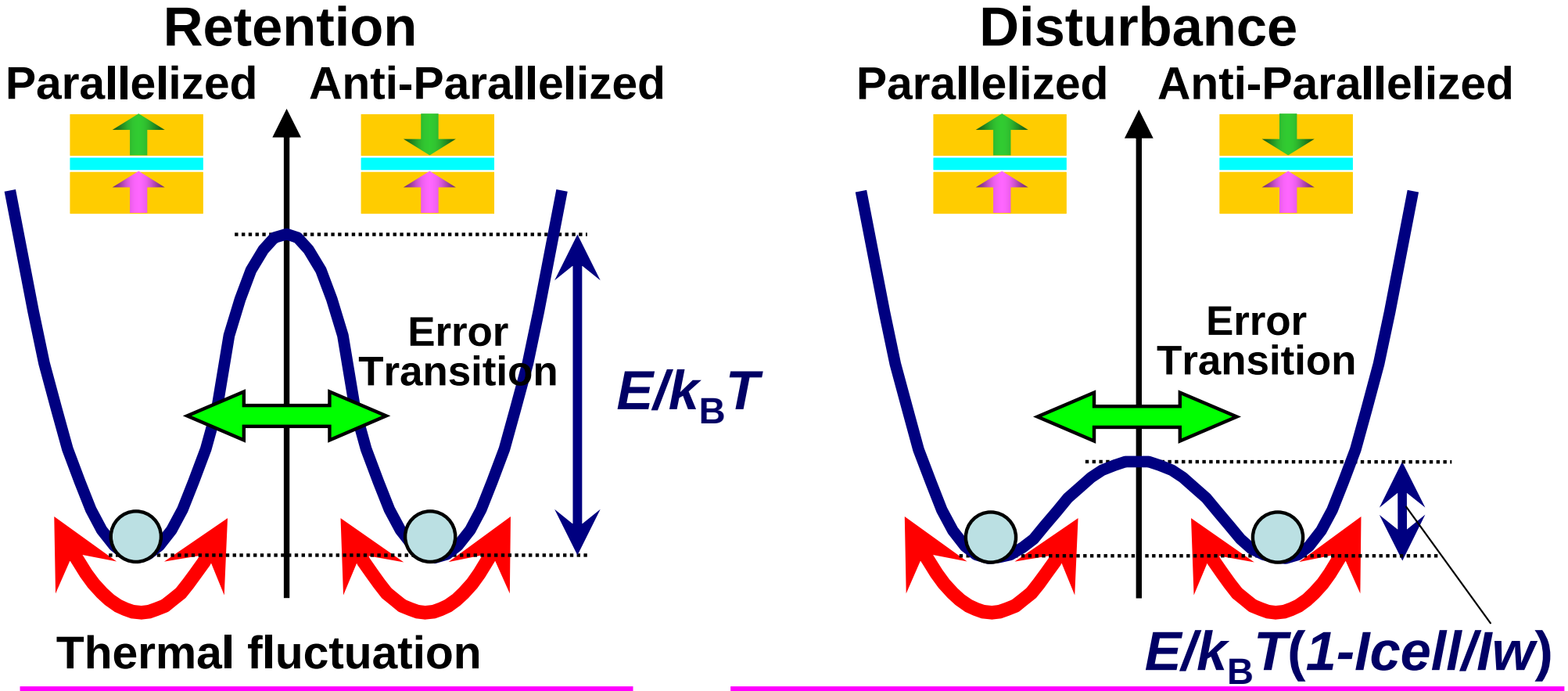


Conventional MRAM

Thermal Stability Factor: E/k_BT

- Thermal field affects retention and disturbance.
- High E/k_BT attainable by perpendicular magnetization.
- Perpendicular TMR with typical CoFeB was reported.

(S. Ikeda et. al, Nature Materials, Volume 9, pp.721–724 (2010))



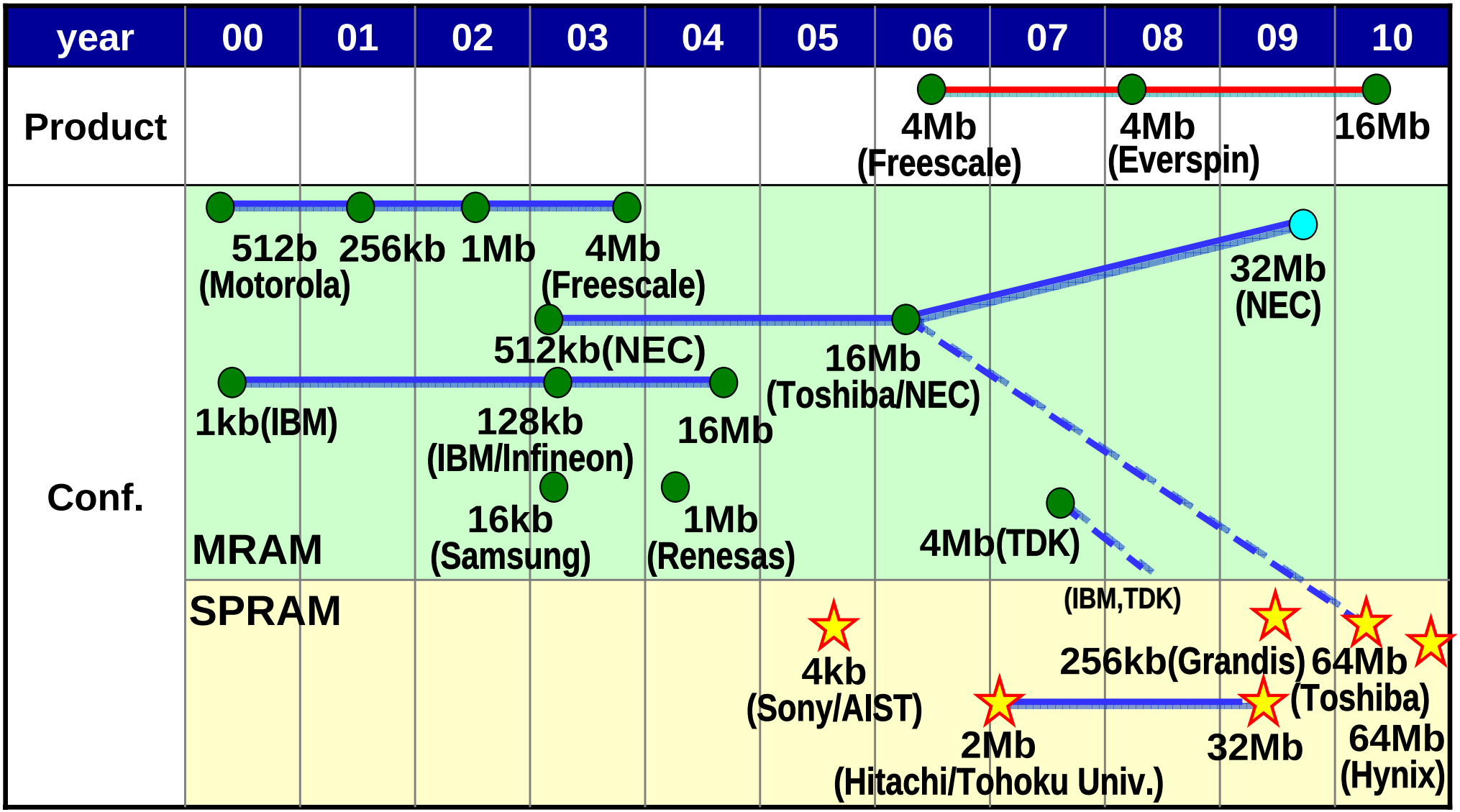
$$P=1-\exp\{-(t/\tau_0)\exp[-E/k_BT]\}$$

$$P=1-\exp\{-(t/\tau_0)\exp[-E/k_BT(1-Icell/Iw)]\}$$

Icell: Read Current, Iw: Write Current

Development of Magnetoresistive Memory

- MRAM: Matured in product.
- R&D moved to SPRAM.

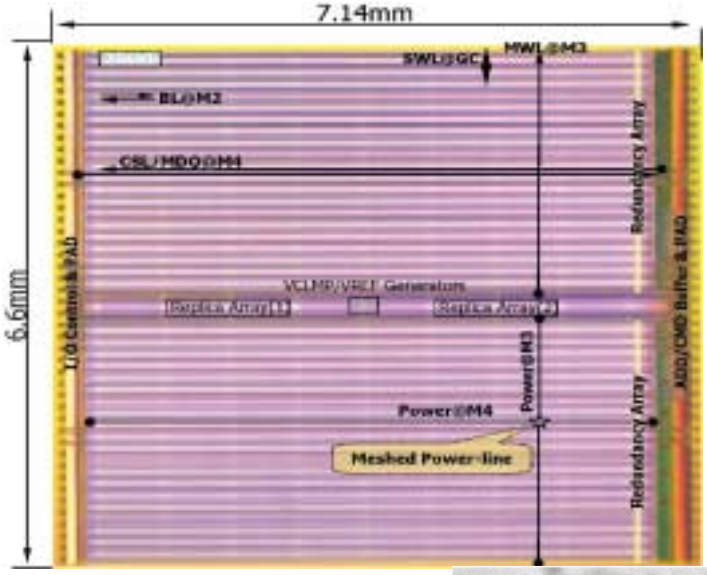


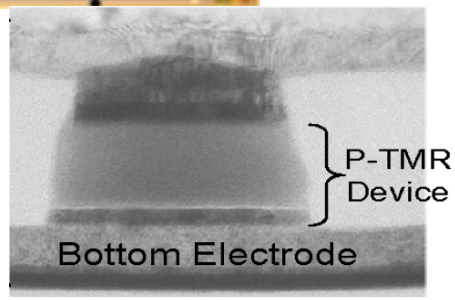
T.Kawahara, Hitachi

23 SPRAM Chips

Perpendicular TMR (Toshiba)

- Process: 65 nm
- Power Supply: 1.2 V
- Density: 64 Mb
- Chip size: 47.12 mm^2
- Cell size: 0.3584 μm^2 (84.8F^2)
- Write Current: 50 uA
- TMR device: Perpendicular

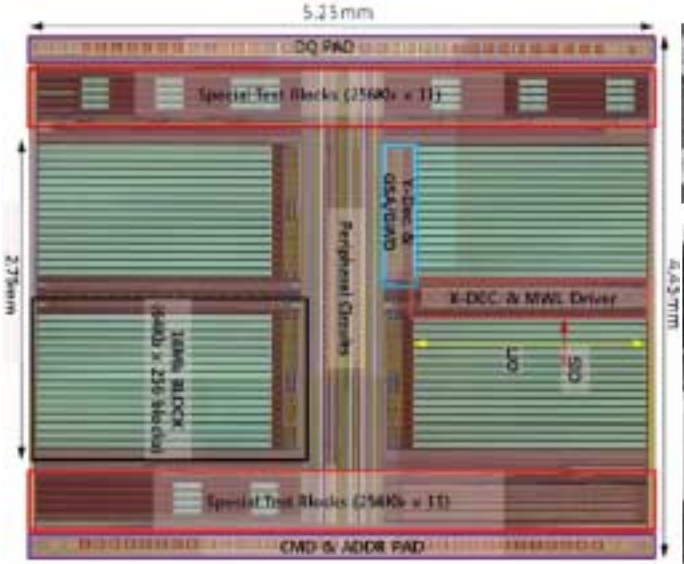


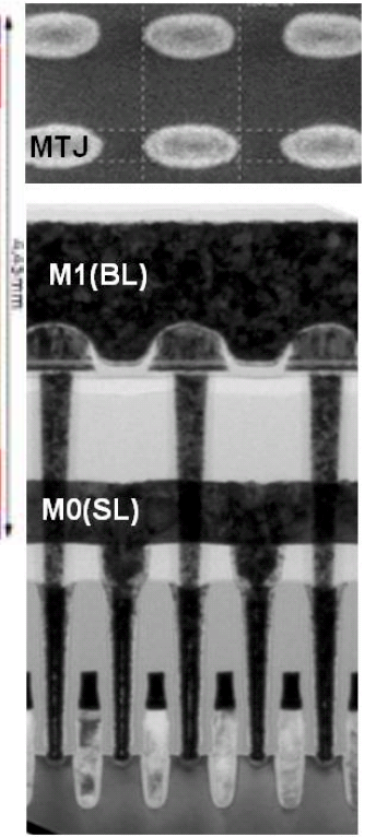


K. Tsuchida, et al.,
ISSCC 2010

Modified DRAM Process (Hynix, Grandis)

- Process: 54 nm
- Power Supply: 1.8 V
- Density: 64 Mb
- Chip size: 23.36 mm^2
- Cell size: 0.041 μm^2 (14F^2)
- Write Current: 140 uA
- TMR device: In-plane



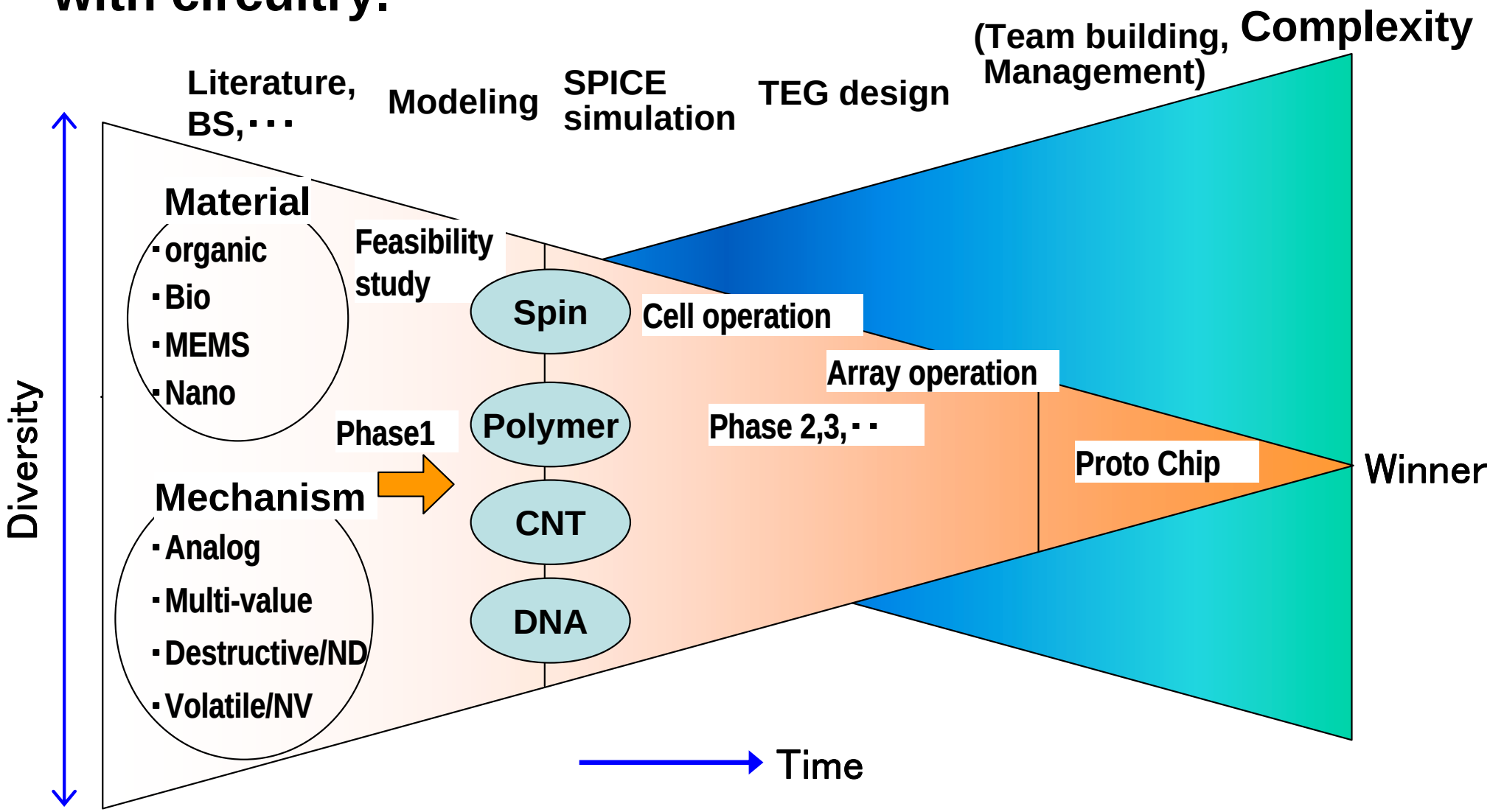


S. Chung, et al.,
IEDM 2010

T.Kawahara, Hitachi

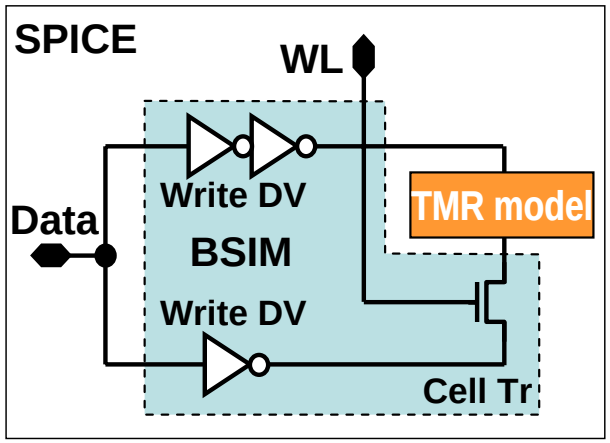
Progress in memory development

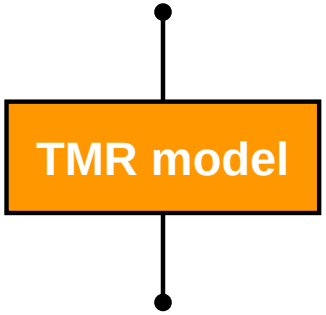
- **Memory tied up closely to deep material and device tech.**
- **Modeling is indispensable to quantitative cooperation with circuitry.**



Ex. Modeling of TMR Memory Cell (1/2)

- **Physics was built into analog and mixed-signal (AMS) simulation, and simulated with CMOS circuits.**
- **Dependence of basic physical characteristic to memory function can be directory simulated.**





$$G(V, T, s)=A(V, T, s)G_P(V)$$

G_P: P-state conductance
A: modulation term
***V*: voltage**
***T*: temperature**
***s*: spin state**

K. Ono, et al., IEDM 2009

- **Non-linear *I-R* behavior**
- ***T*-dependent *I-R* behavior**
  - **P-state: $G_P \propto V^2$ (Simmons)**
  - **AP-state: $\Delta R/R \propto \exp(-|V|/Vh)$**
  - **$\Delta R/R(V=0)$: function of *T*-dependent polarization**
  - **Non-linear parameter *Vh* : *T*-independent**

- **Numerical solution of stochastic-LLG equation**
- **Include thermal fluctuation of macro-spin**

$$\frac{\partial \mathbf{s}}{\partial \tau}=\left[\mathbf{s}\times\left(\mathbf{h}_{\mathrm{eff}}+\mathbf{h}_{\mathrm{fl}}\right)-\alpha\mathbf{s}\times\left(\mathbf{s}\times\left(\mathbf{h}_{\mathrm{eff}}+\mathbf{h}_{\mathrm{fl}}\right)\right)\right]$$

$$\Delta\tau=\frac{\gamma_0 H_{\mathrm{K}}}{1+\alpha^2}\Delta t$$

$$\mathbf{h}_{\mathrm{fl}}=\sqrt{\frac{\alpha}{\left(1+\alpha^2\right)}\frac{2k_{\mathrm{B}}T}{M_{\mathrm{S}}H_{\mathrm{K}}Vo}}\varsigma_{\mathrm{fl}}$$

$$\left\langle\varsigma_{\mathrm{fl}}(t)\right\rangle=0,\ \left\langle\varsigma_{\mathrm{fl}}^{\varphi}(t)\varsigma_{\mathrm{fl}}^{\theta}(t')\right\rangle=2\delta_{\varphi\theta}\delta(t-t')$$

s: normalized spin vector
H_K: easy-plane anisotropy field
γ_0: gyromagnetic ratio
h_{eff}: effective field
h_{fl}: stochastic field

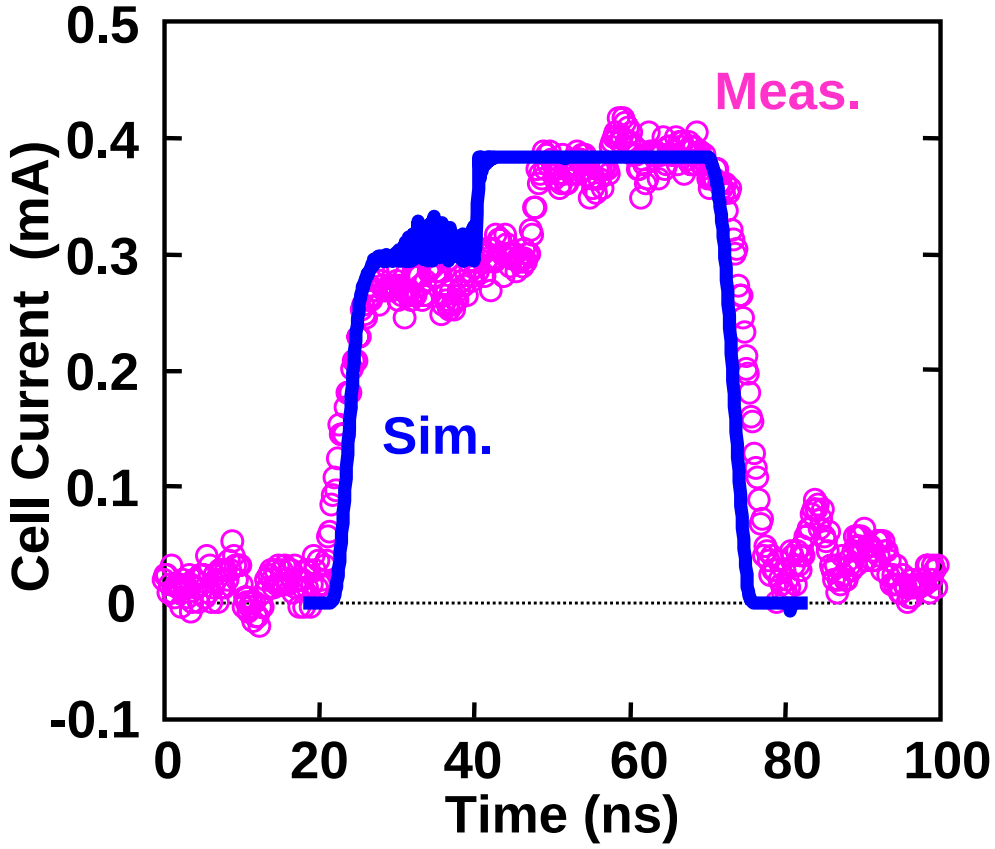
***M*s: saturation magnetization**
α : damping coefficient
***Vo*: free-layer volume**
k_B : Boltzmann's constant
***T*: temperature**

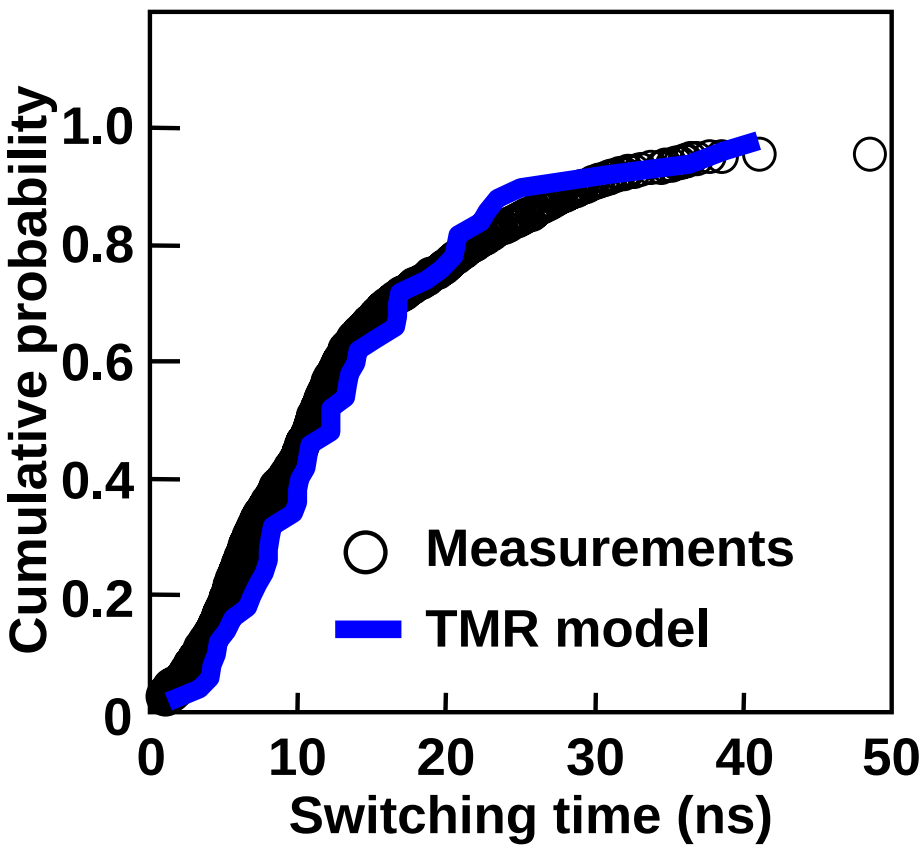
Ex. Modeling of TMR Memory Cell (2/2)

- **Physics was built into analog and mixed-signal (AMS) simulation, and simulated with CMOS circuits.**
- **Dependence of basic physical characteristic to memory function can be directory simulated.**

Transient write current and stochastic switching

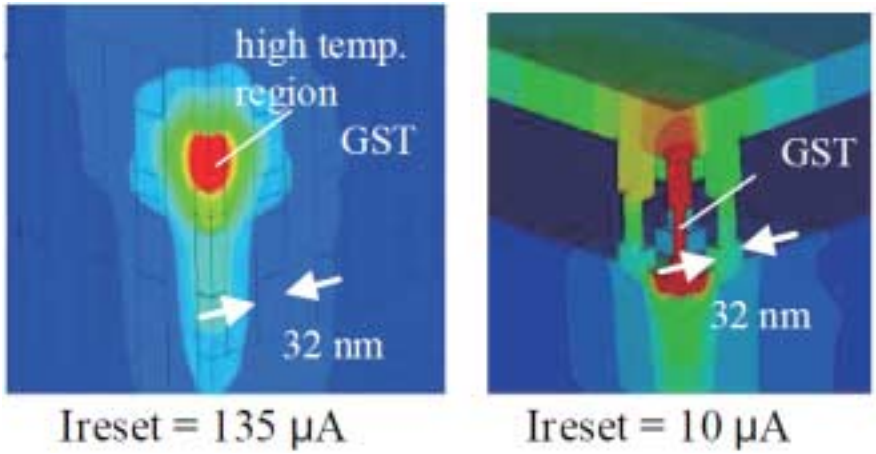
K. Ono, et al., IEDM 2009



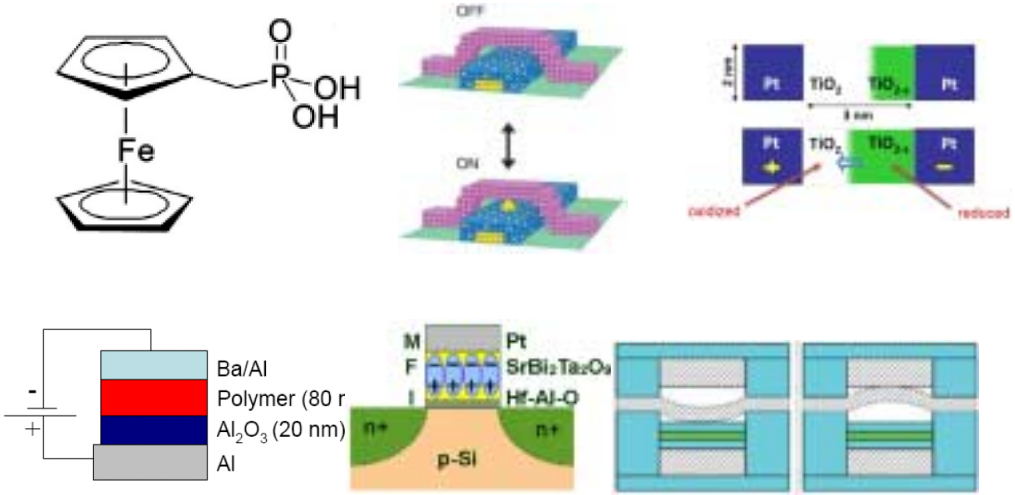


For emerging memory simulation

- **Need path to generate physical model with new materials. Memory tech. should handle physical phenomena.**
- **Estimation of phenomena between adjacent cells and the size effect are important.**



Y. Sasago, et al., VLSI Tech. 2009



Emerging Memories
(from ITRS ERD/ERM WG,2010)

"Visualization": important to intuitive understanding of peculiar phenomenon to the material and grasping the disturbance between adjacent cells in the array.

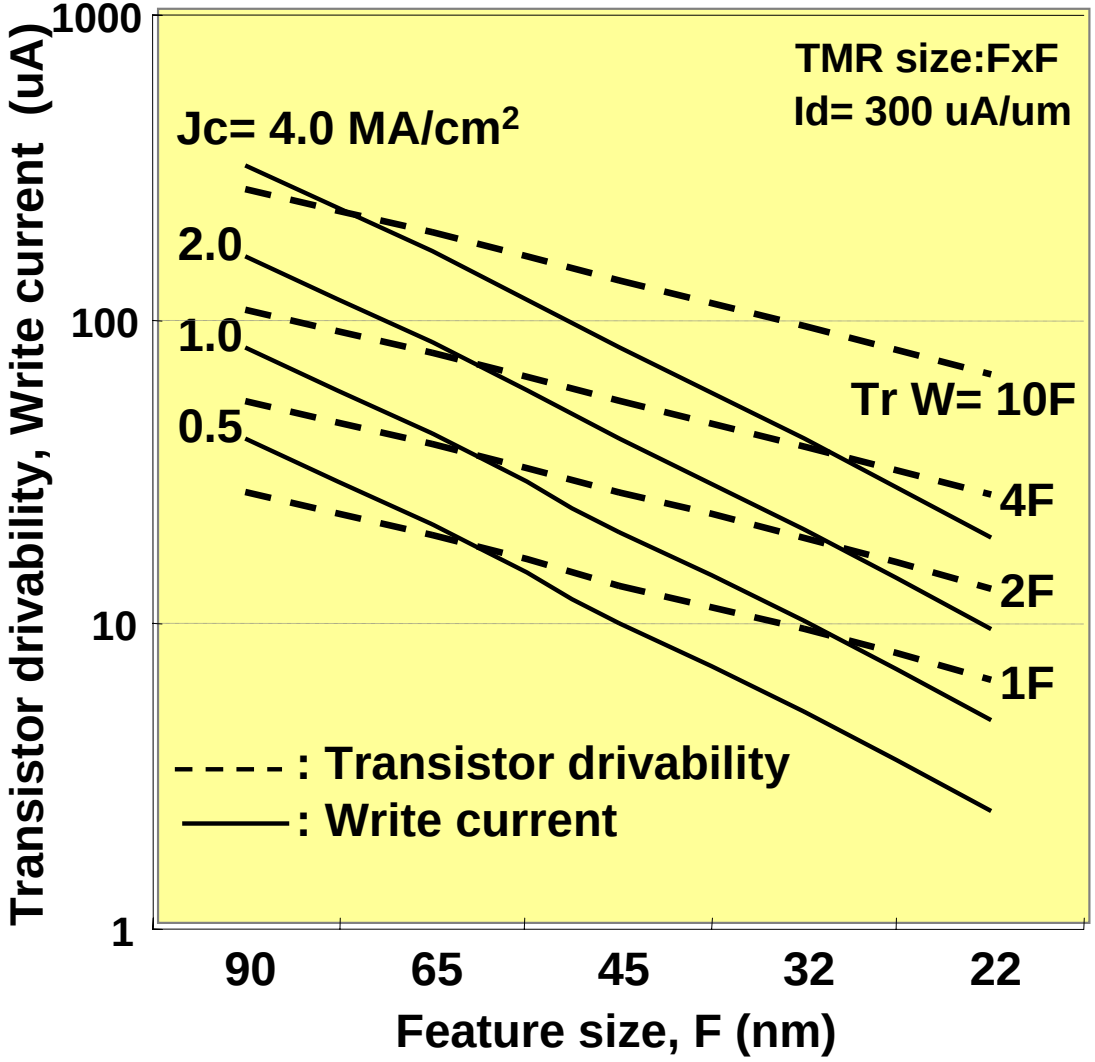
"Size dependent phenomena": TCAD necessary to handle materials more than current Si, otherwise size dependence, that is important, cannot be estimated.

Outline

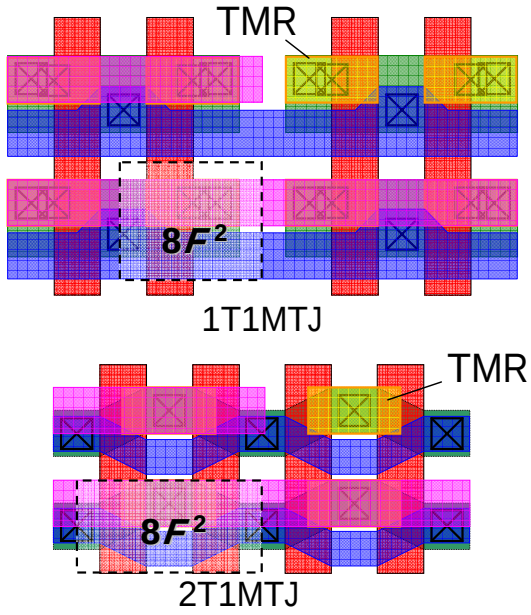
- Introduction
- Non-volatile memories
- **Spin-transfer torque RAM scalability**
- Normally-off instant-on computing
- Conclusion

Memory cell size scalability

- Transistor drivability ∝F, but TMR write current ∝F².
- Cell becomes easy to shrink with advancing feature size.



TMR size	F²	F²	F²	F²
Gate width	10F	4F	2F	F
Cell size	40F²	16F²-	8-6F²	4F²

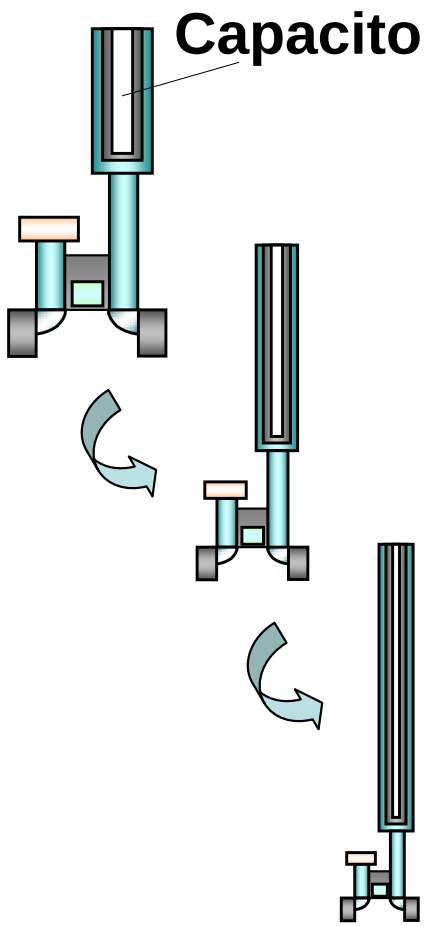


T.Kawahara, Hitachi

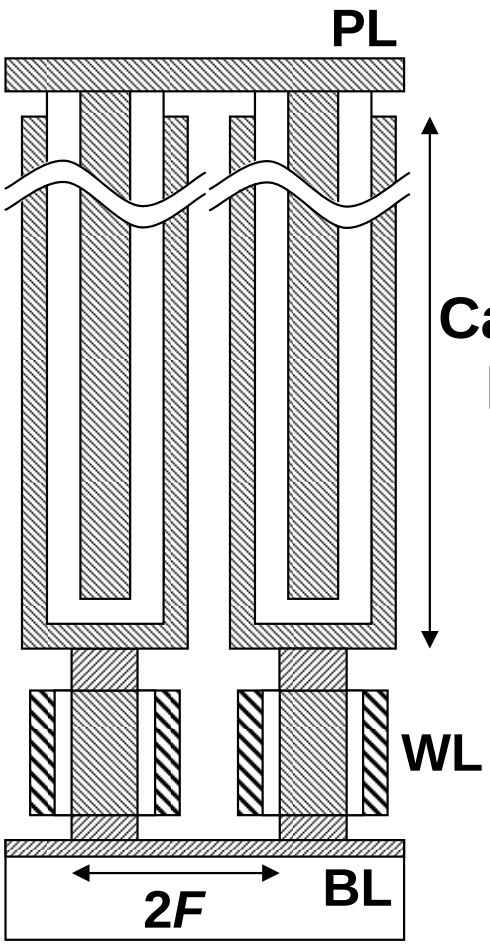
Scalable $4F^2$ Memory-cell configurations

- **TMR memory cell has better scalability than DRAM cell due to low aspect ratio.**

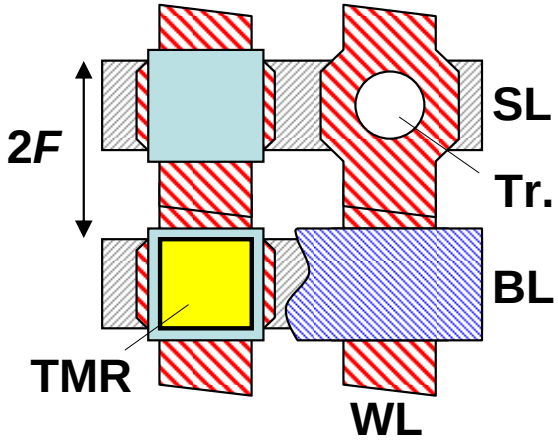
R. Takemura, et al., IMW 2010



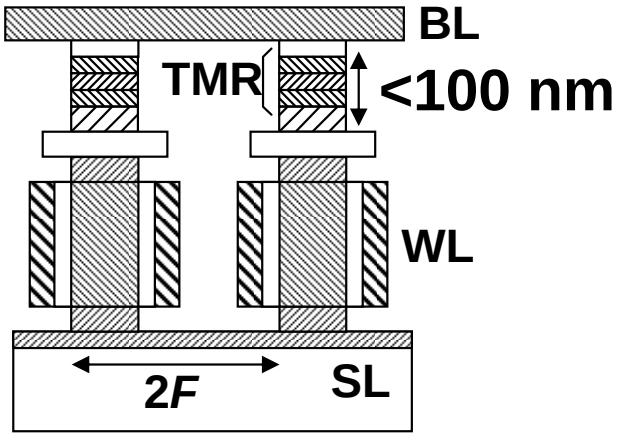
Scaling in DRAM cell's capacitor



$4F^2$ DRAM cell



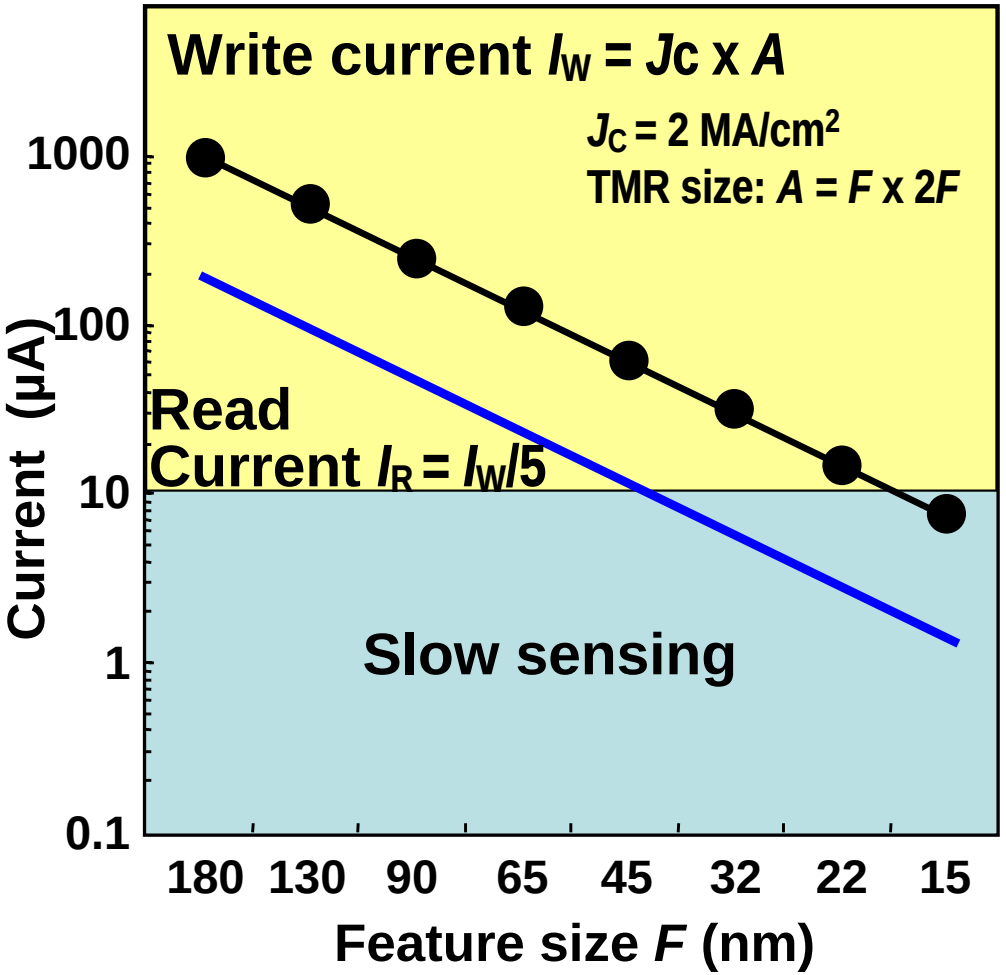
Cell layout



$4F^2$ SPRAM cell

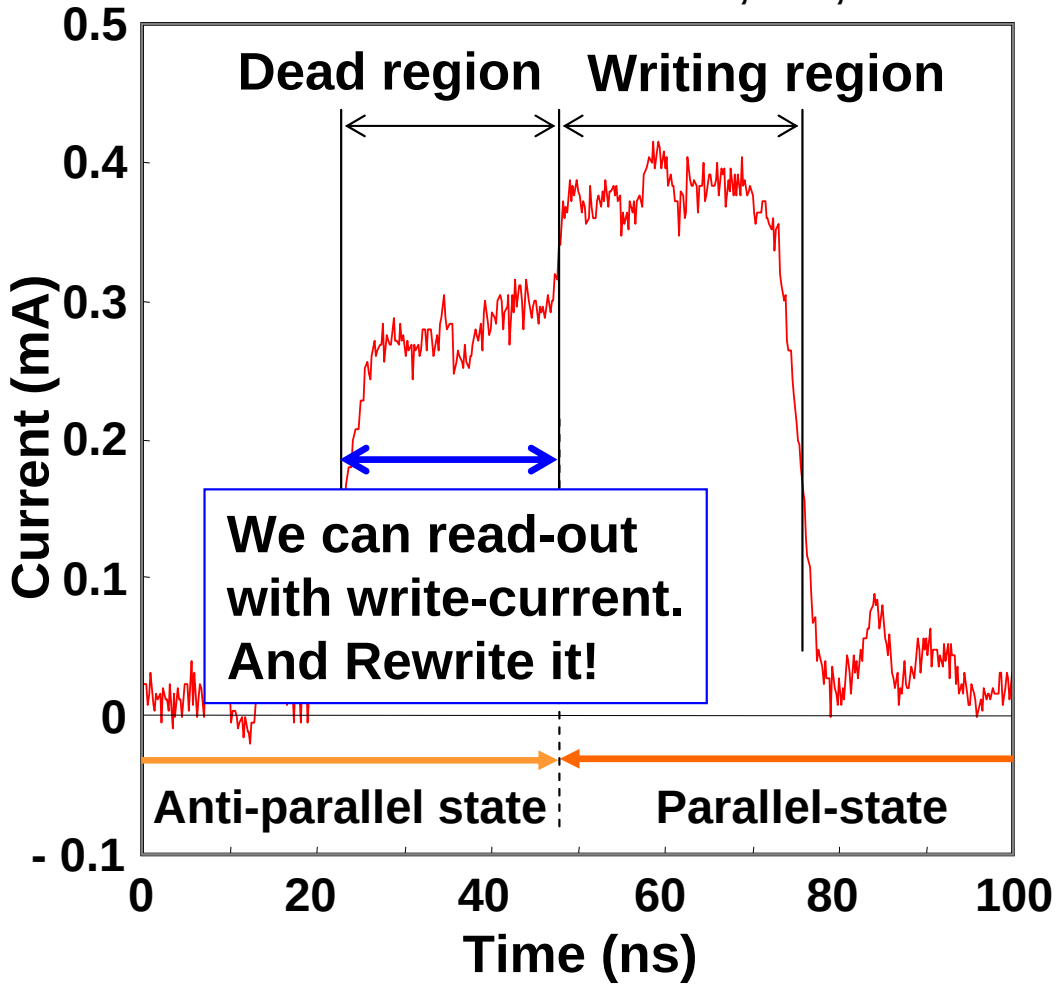
Scalable write current vs read current

- Good scalability causes small read current
- Disruptive reading by using dead region in write.



Scaling of write and read current

R. Takemura, et al., IMW 2010

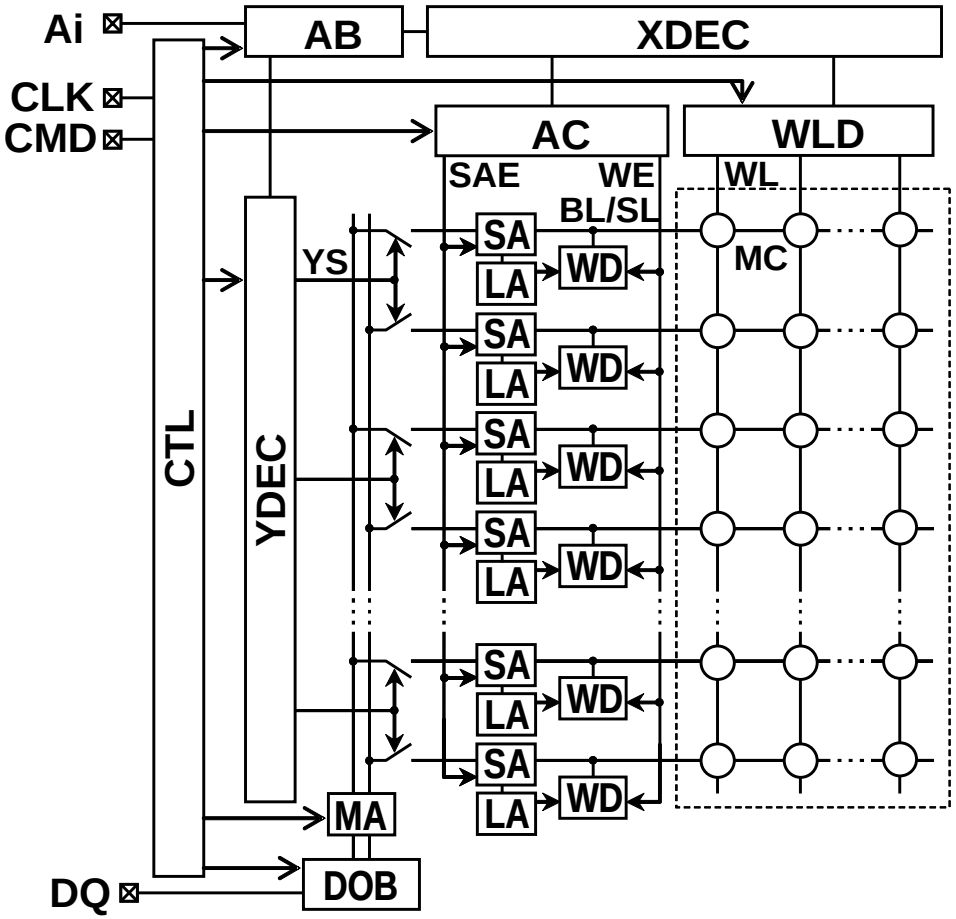


Dead region time in write operation

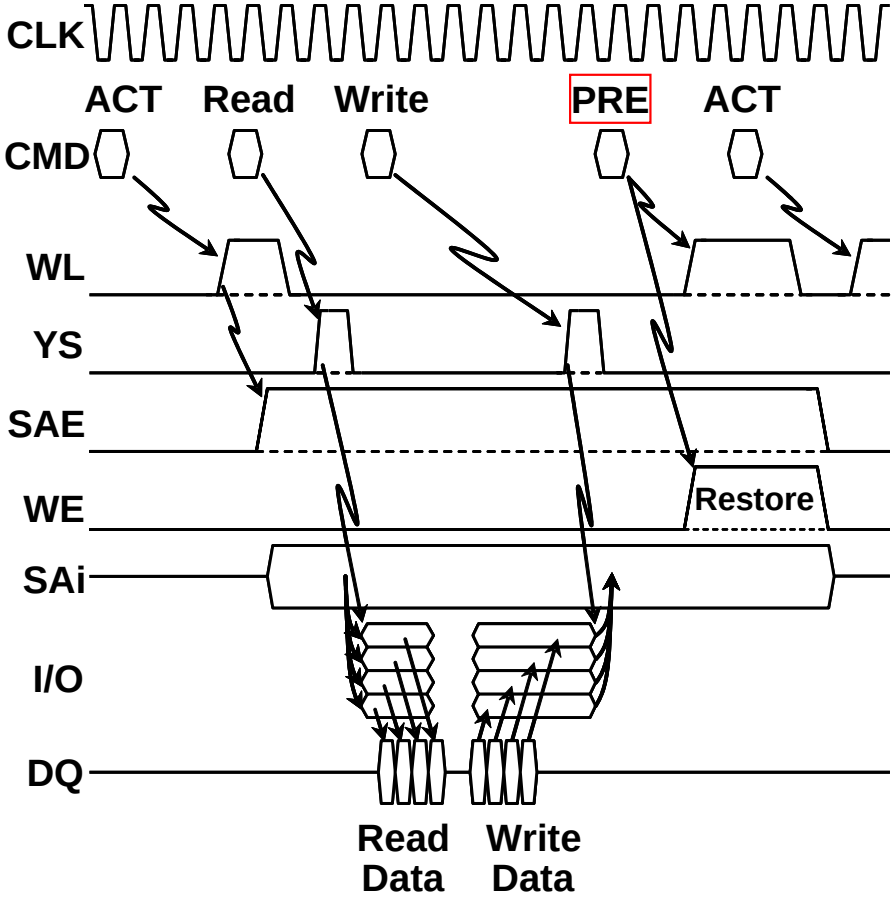
DDRx-SDRAM-compatible operation

- **Restore operation can be performed with PRE.**
- **NV-RAM or long-retention DRAM for super LP.**

R. Takemura, et al., IMW 2010



Chip configuration

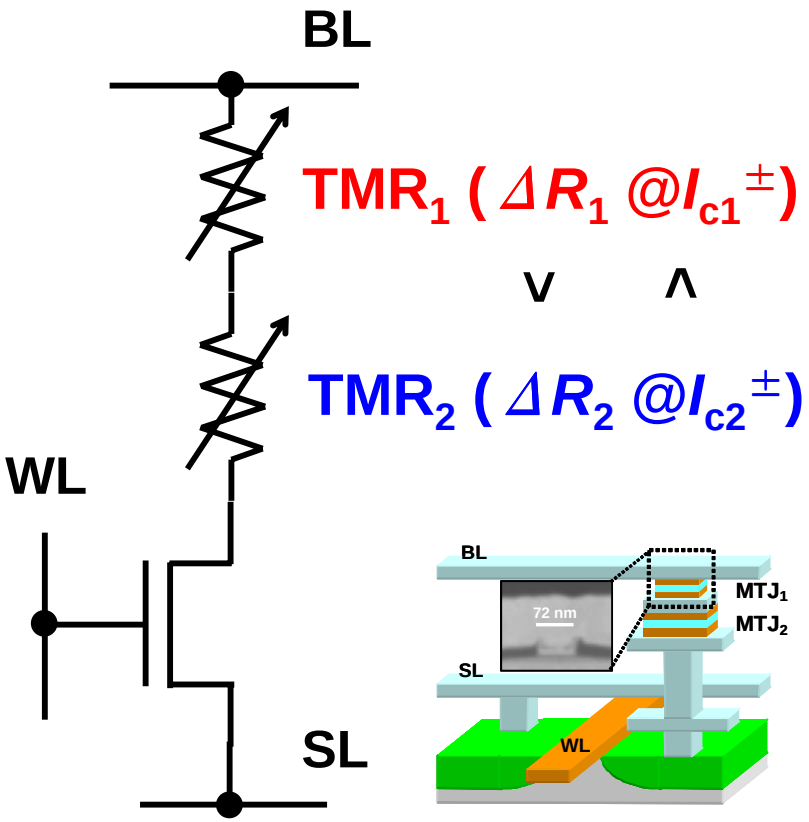


Timing diagram

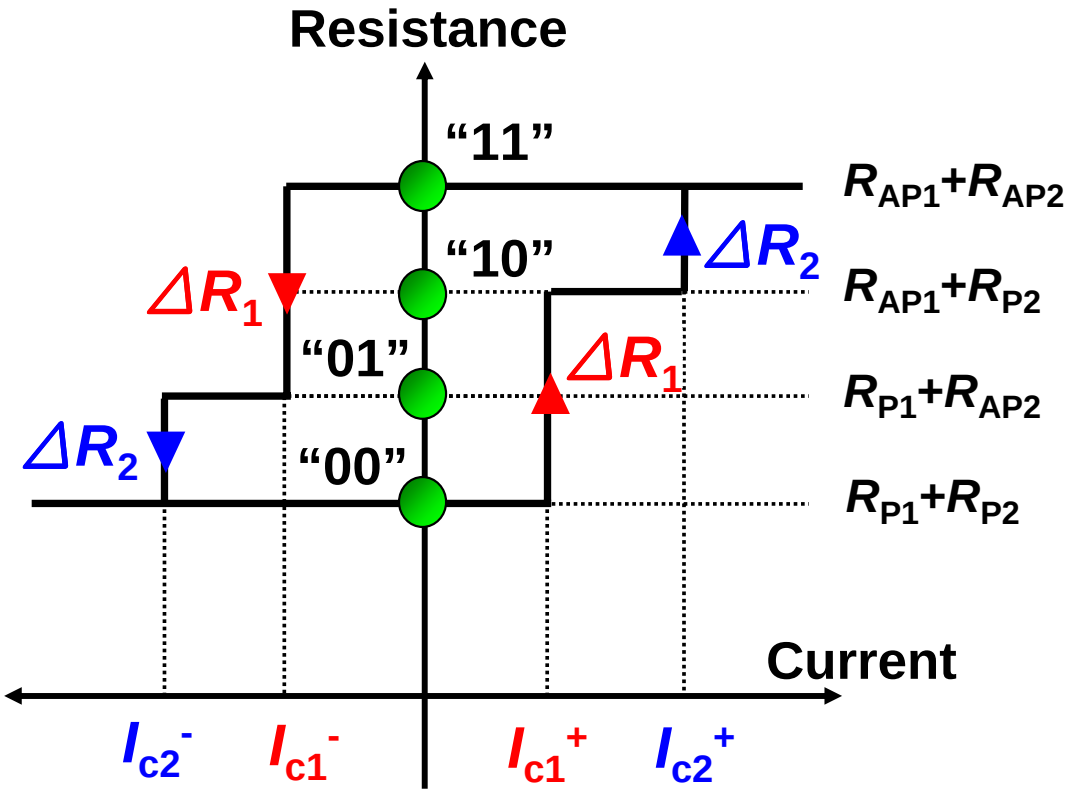
Concept of MLC-SPRAM

- **Series 2-TMRs with different *Ic* and *ΔR* provide digital levels of resistance, possible <3F²/bit.**
- **Proven basic operation by measurement.**

T. Ishigaki, et al., VLSI Tech. 2010



Memory cell configuration

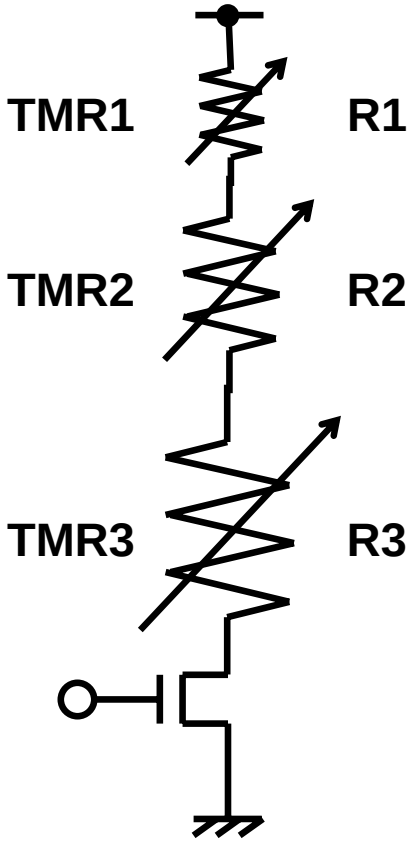


Characteristic

Toward $2F^2$/bit and More…

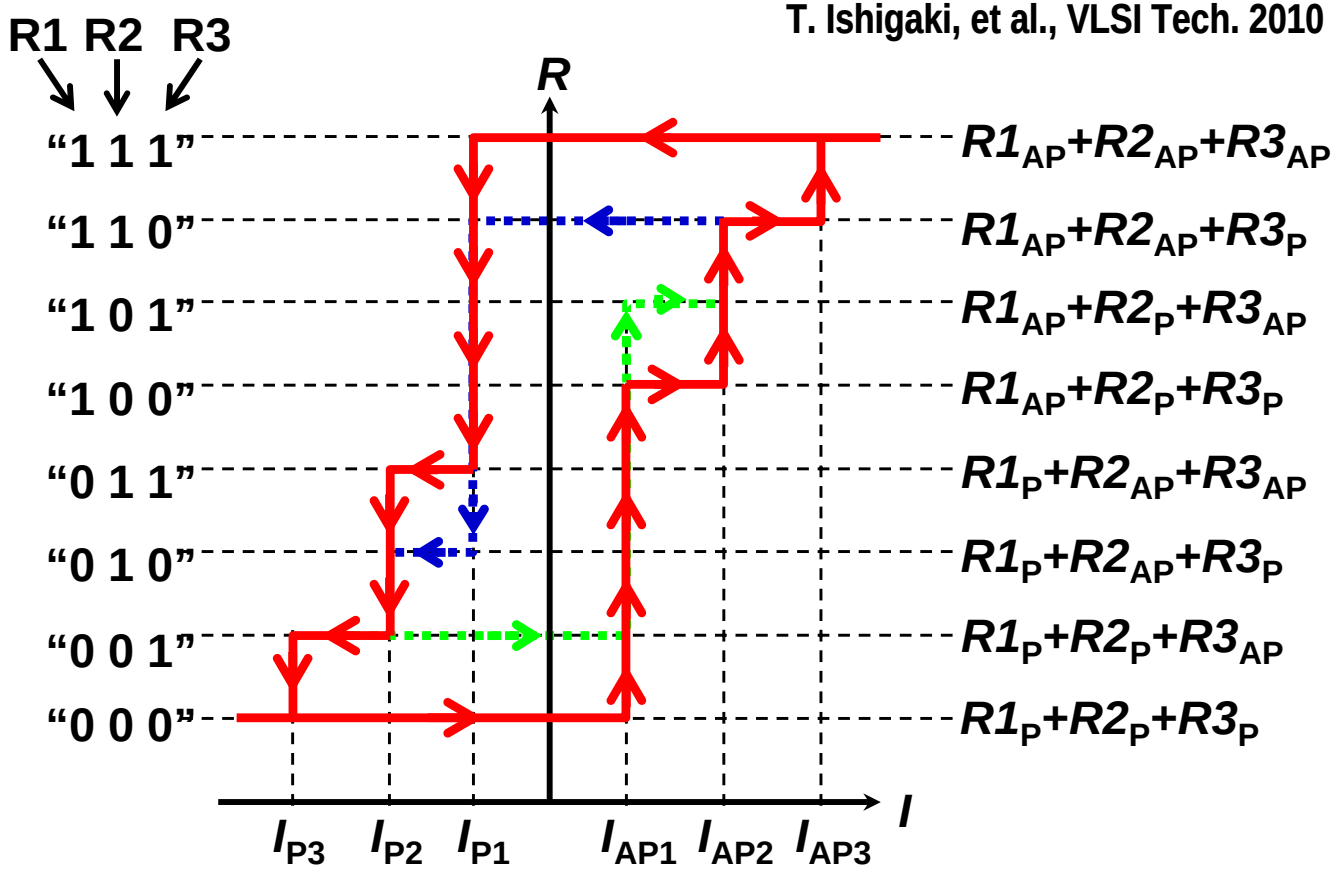
- 2^n value level can achieve n-bits per cell.
- Needs high E/k_BT and sufficiently small RA.

(RA: resistance area product of TMR)



Memory cell circuit

T. Ishigaki, et al., VLSI Tech. 2010



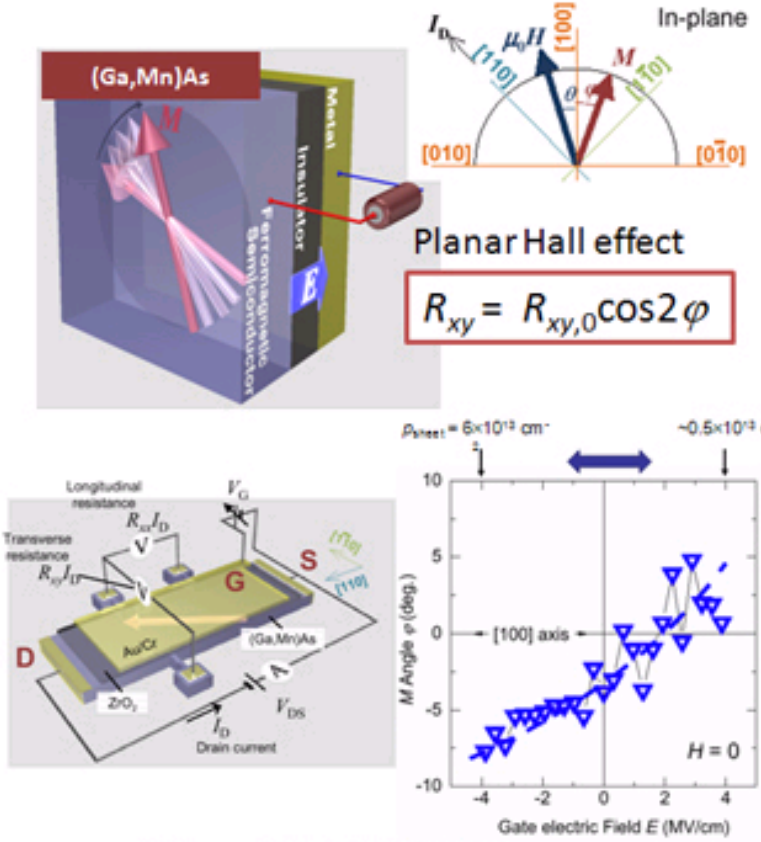
3-bit/cell state transition scheme

Future tech.: Voltage-driven magnetic RAM

- **Basic research is progressing for voltage-driven magnetic RAM.**

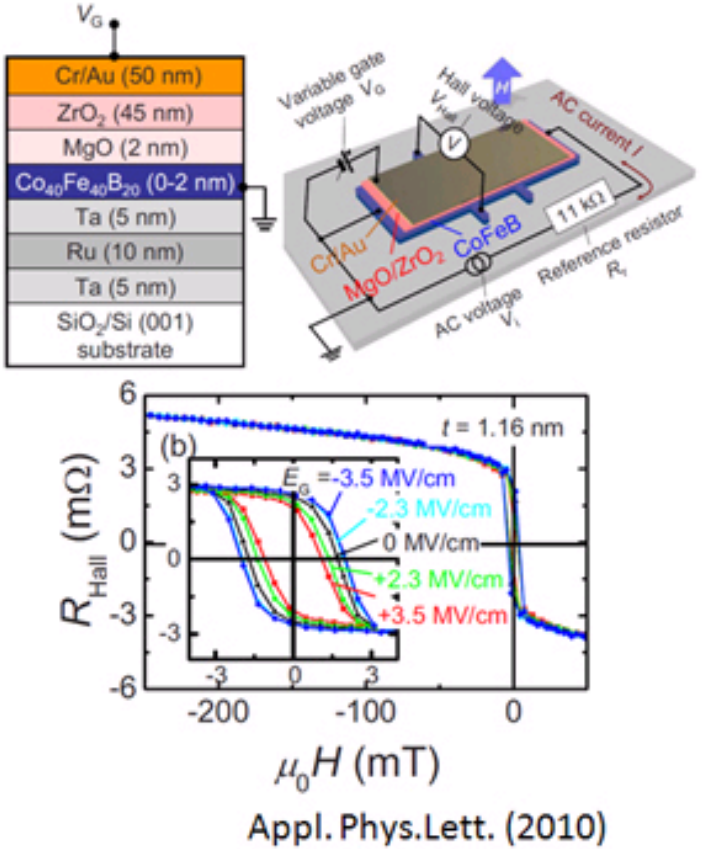
Current-driven device should face power consumption concern, voltage drop in line, and ability for current supply switch. -> Solved by Voltage-driven.

Ferromagnetic semiconductor



Nature 455, 515 (2008).

Sputtered MgO/CoFeB/Ta stack structure



Appl. Phys.Lett. (2010)

※This work was supported by Research and Development for Next-Generation Information Technology of MEXT.

Ref.: http://www.csis.tohoku.ac.jp/ (Center for Spintronics Integrated Systems)

Outline

- Introduction
- Non-volatile memories
- Spin-transfer torque RAM scalability
- **Normally-off instant-on computing**
- Conclusion

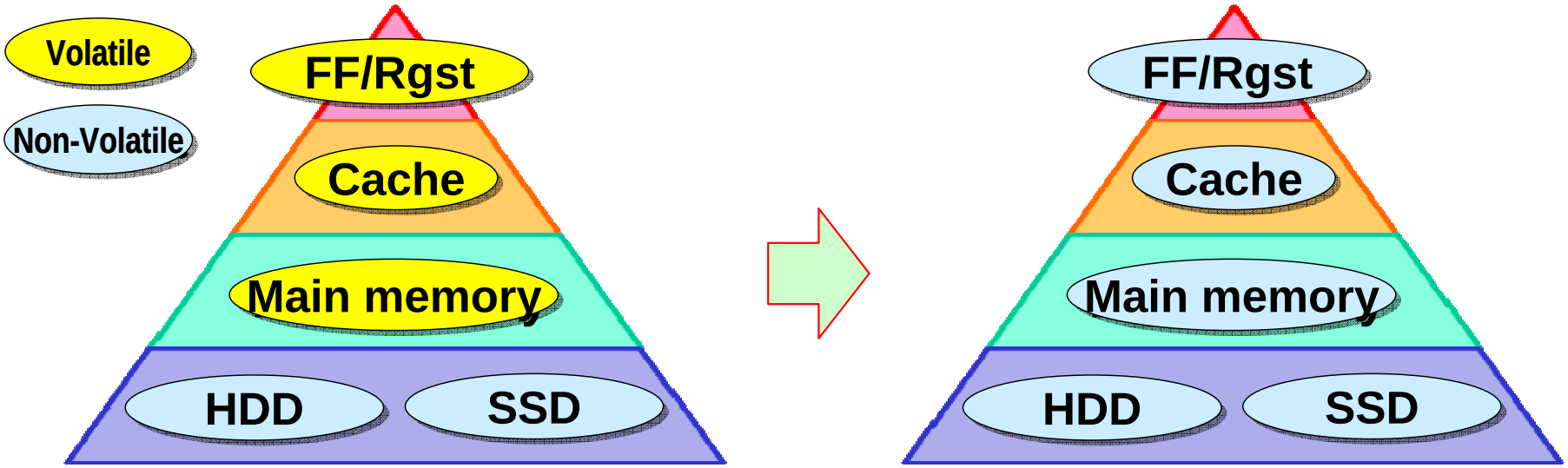
Non-Volatile Architecture

Merits for non-volatile architecture (non-volatile computing)

- **Disappear the difference between suspend and hibernation.**
  - **Easy to control On/Off of server due to the load.**
- **Assured write back operation.**
  - **Fast and quick restarting.**

Issues

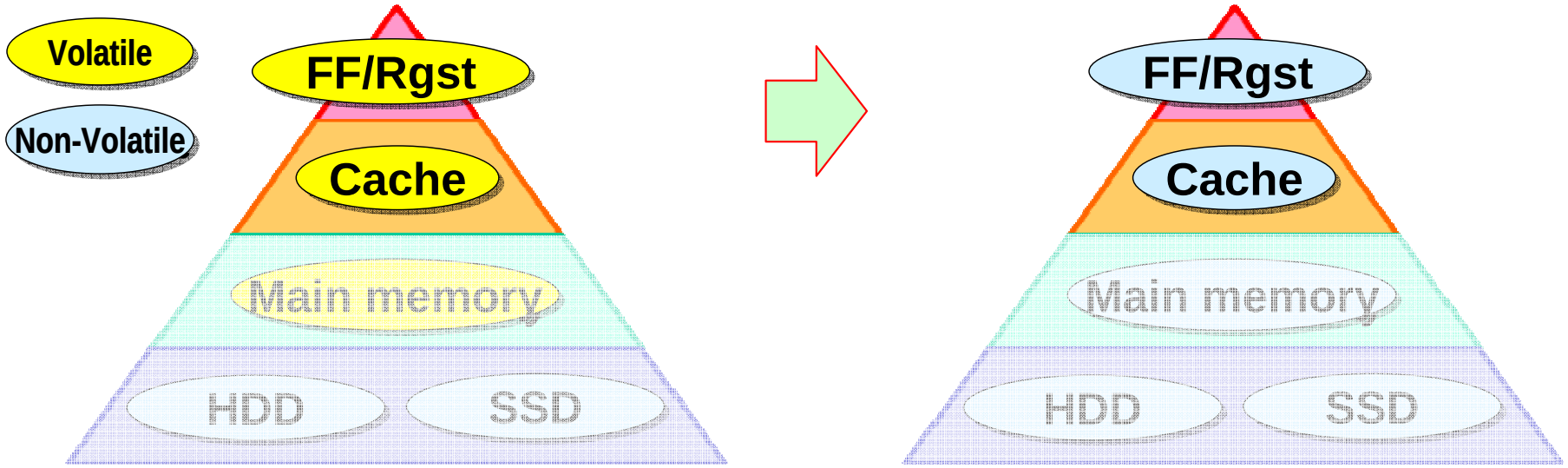
- **New OS for the best use of non-volatile memory.**
  - **When restarting, activate the contents of memory.**
  - **For security issue, handle the erase of contents.**
  - **Same as uninterruptible computation.**
- **Cooperation with processor vender and OS vender.**
  - **Ultimately, compiler schedules On/Off.**



Conventional Architecture **Non-Volatile Architecture**

In CPU and Basic Circuitry Layer

- **Achieving a normally-off state as much as possible by detecting any standby operation, when the power needed is preferably lower than required for transition.**
- **Eliminating the power required for communication between the memory and logic circuits.**
- **High energy efficiency in operation and no DC power consumed in CPU's long standby state, as occurs often in most business applications.**

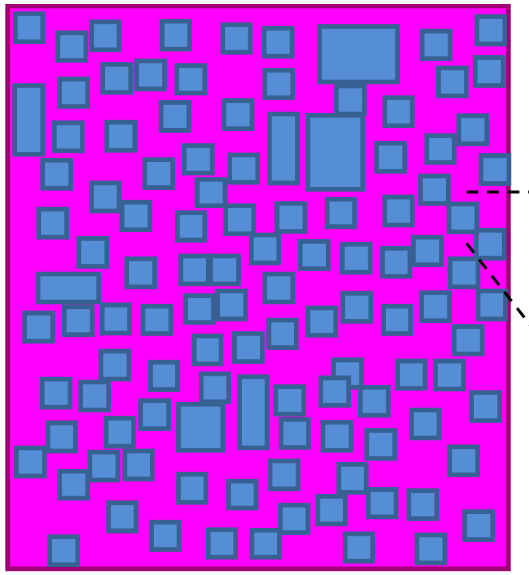


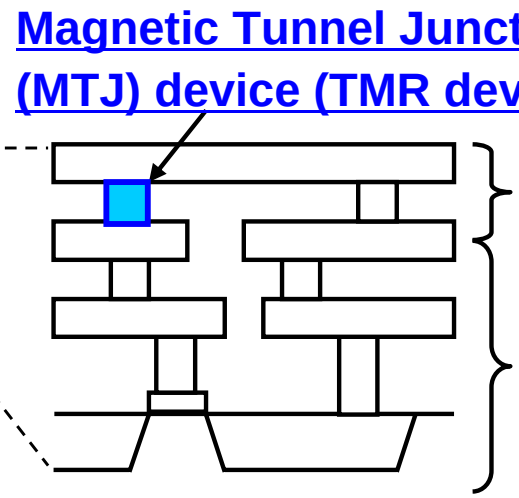
Conventional Architecture **Non-Volatile Architecture**

Nonvolatile Logic-in-Memory Architecture

- **Logic-in-Memory Architecture (proposed in 1969):**

Storage elements are distributed over a logic-circuit plane.





- Non-volatility
- Unlimited endurance
- Fast writability
- Scalability
- CMOS compatibility
- 3-D stack capability

- Storage is nonvolatile:
  (Leakage current is cut off)
- MTJ devices are put
  on the **CMOS** layer
- Storage/logic are merged:
  (global-wire count is reduced)

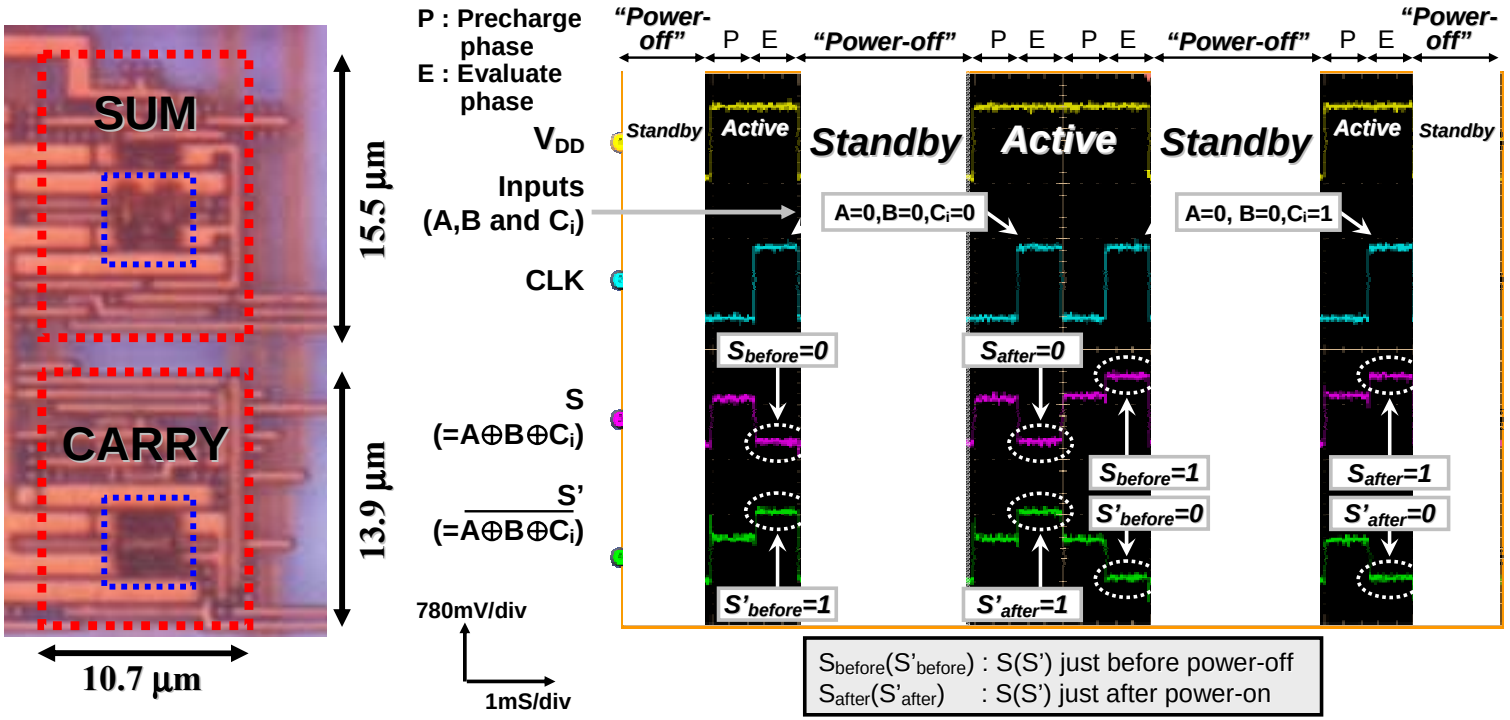
→ Static power is cut off.
→ Chip area is reduced.
→ Wire delay is reduced.
→ Dynamic power is reduced.

Ref.: http://www.csis.tohoku.ac.jp/ (Center for Spintronics Integrated Systems)

Design of a Nonvolatile Full Adder

- **Demonstrated quick on/off operation.**
- **Dynamic power reduce to 1/4 with keeping performance.**

S. Matsunaga et. al., Applied Physics Express 1 (2008) 091301.



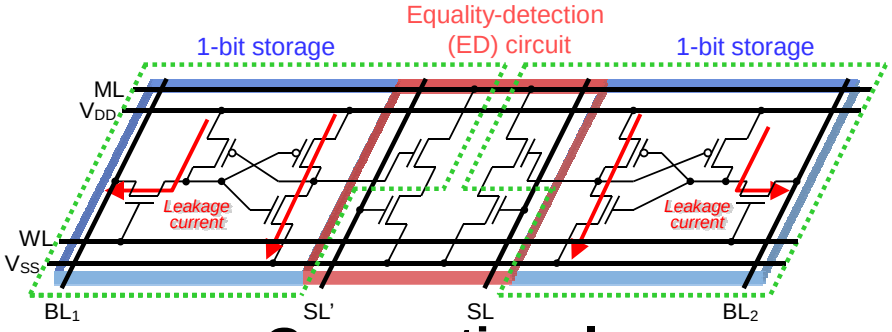
	CMOS	Proposed
Delay	224 ps	219 ps
Dynamic power (@500MHz)	71.1 mW	16.3 mW
Write time	2 ns/bit	10 ns/bit
Write energy	4 pJ/bit	20.9 pJ/bit
Static power	0.9 nW	0.0 nW
Area (Device counts)	333 mm² (42 MOSs)	315 mm² (34 MOSs + 4 MTJs)

Ref.: http://www.csis.tohoku.ac.jp/ (Center for Spintronics Integrated Systems)

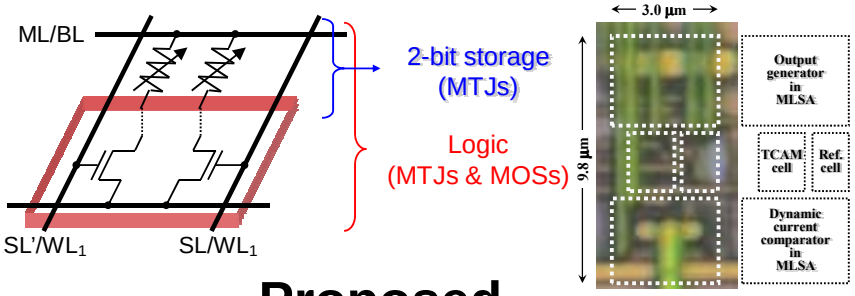
High-Density TCAM Cell Design

- **Halves active power-delay product with 1/100 stand-by.**
- **2bit TMR storage merged into CMOS logic.**

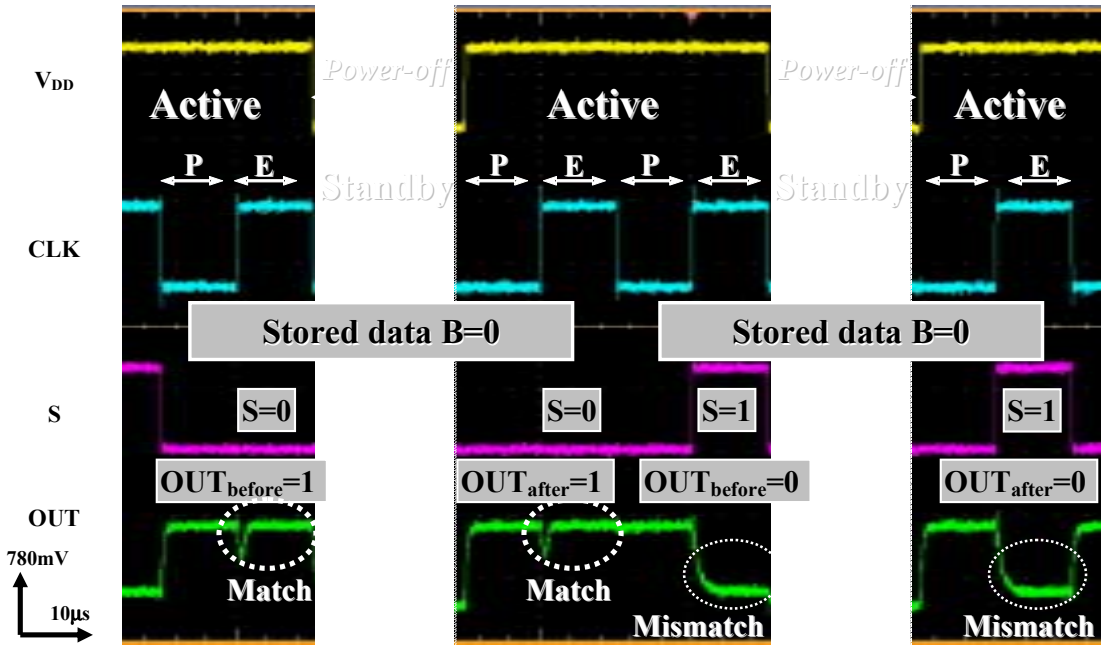
S. Matsunaga, et al., APEX, 2, 2, 023004, Feb.2009.



Conventional



Proposed



		CMOS-only	Proposed
Active Power	Cell array	109.6 μW	107.3 μW
	SA	30.8 μW	9.6 μW
	ACC	3.7 μW	3.7 μW
	CLK	32.7 μW	62.0 μW
Standby Power	Cell array	340.9 μW	1.8 μW
	SA	1.2 μW	2.3 μW
Delay		1.39 ns	0.60 ns

103%
1.2%
43%

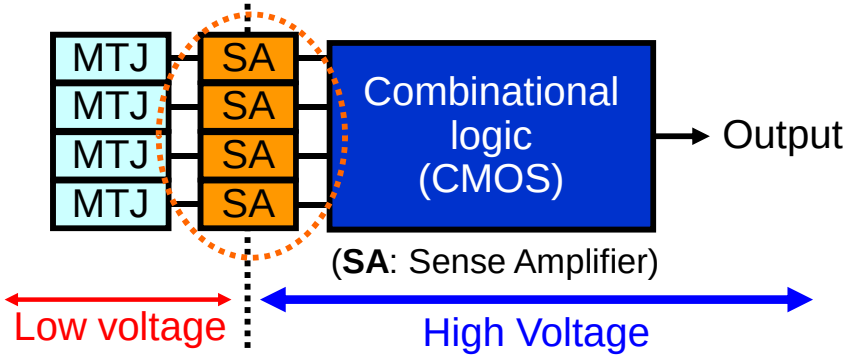
HSPICE simulation under a 90nm CMOS/MTJ technology @125MHz,
R_P : 2kΩ, TMR ratio : 100%

Ref.: http://www.csis.tohoku.ac.jp/ (Center for Spintronics Integrated Systems)

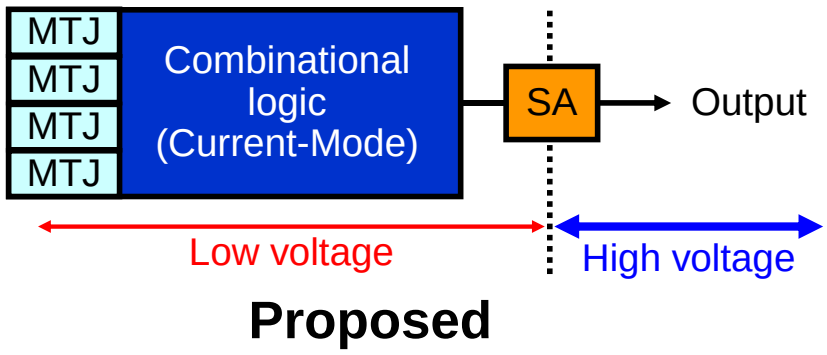
Design of a Compact Nonvolatile FPGA

- **TMR device/logic direct combination by differential current-mode tech. simplifies the circuit, reduces power.**

D. Suzuki, et al., Symp. on VLSI Circuits, 80/81, June 2009.



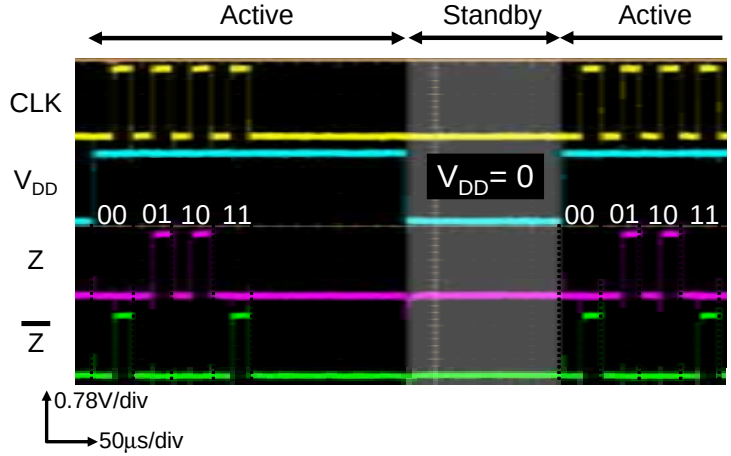
Conventional



Proposed

		Nonvolatile SRAM	Proposed
Device Counts		102 MOSs + 8 MTJs	29 MOSs + 4 MTJs
Area *1)		702 μm²	287μm²
Active	Delay *2)	140 ps	185 ps
	Power*2)	26.7 mW	17.5 mW
Standby	Power	0 mW	0 mW

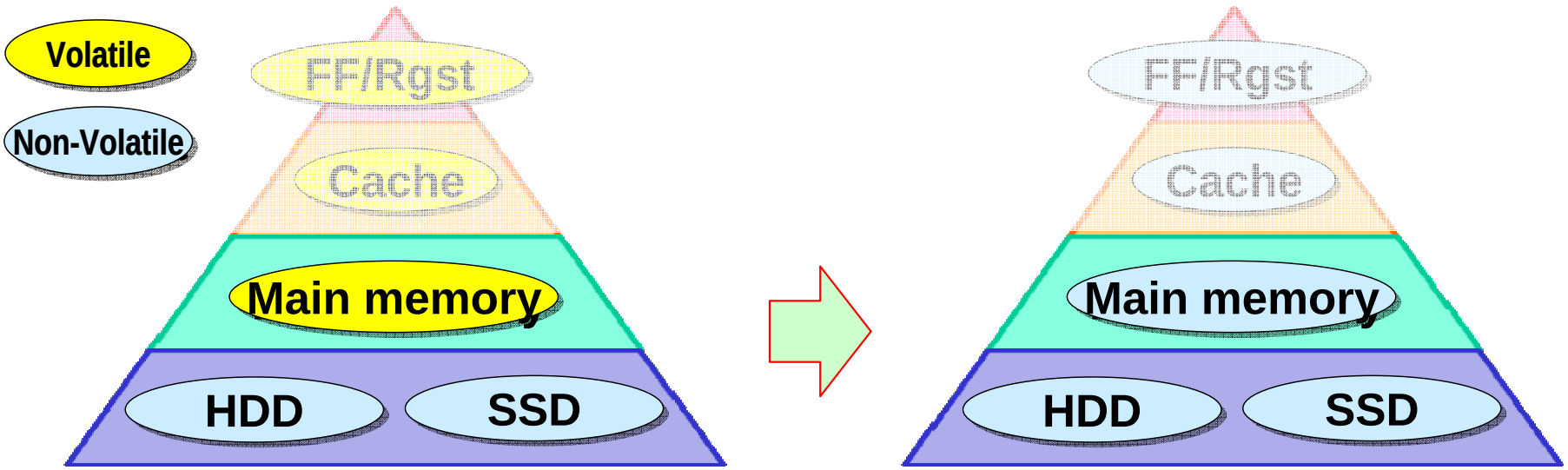
*1) Estimation based on a 0.14mm process
*2) HSPICE simulation based on a 0.14mm MOS/MTJ-hybrid process



Ref.: http://www.csis.tohoku.ac.jp/ (Center for Spintronics Integrated Systems)

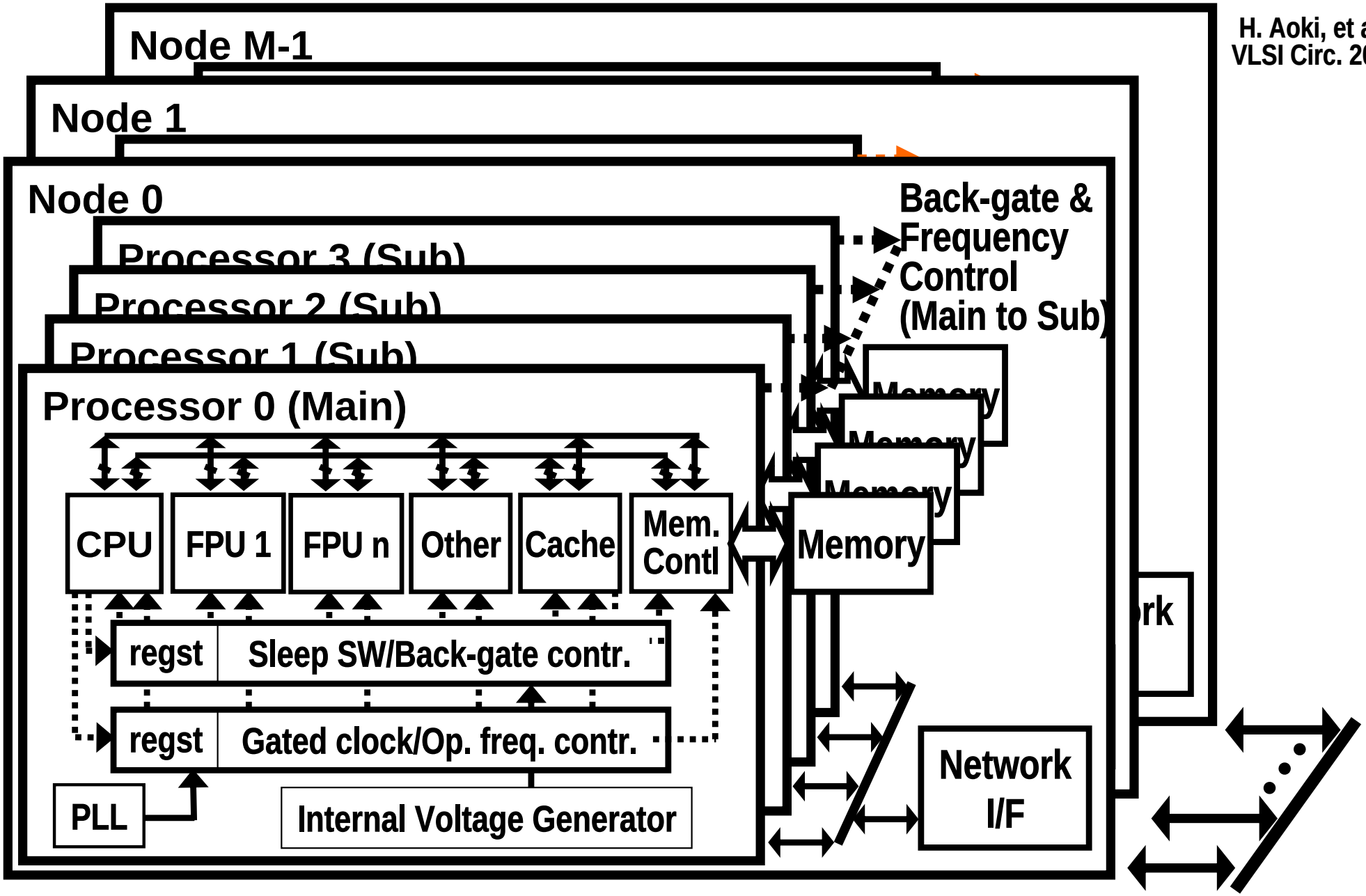
In Main Memory and System Layer

- **For instant-on, keeping the status in main memory in standby without power dissipation.**
  - **Simplified control circuit because the refresh operation is minimized.**
  - **Battery backup in cache logging area is eliminated.**
- **Hot standby configuration (speed) with deep standby mode (power) is attainable.**
  - **Achieving system reliability for mission-critical systems (high SLA) with low-power.**



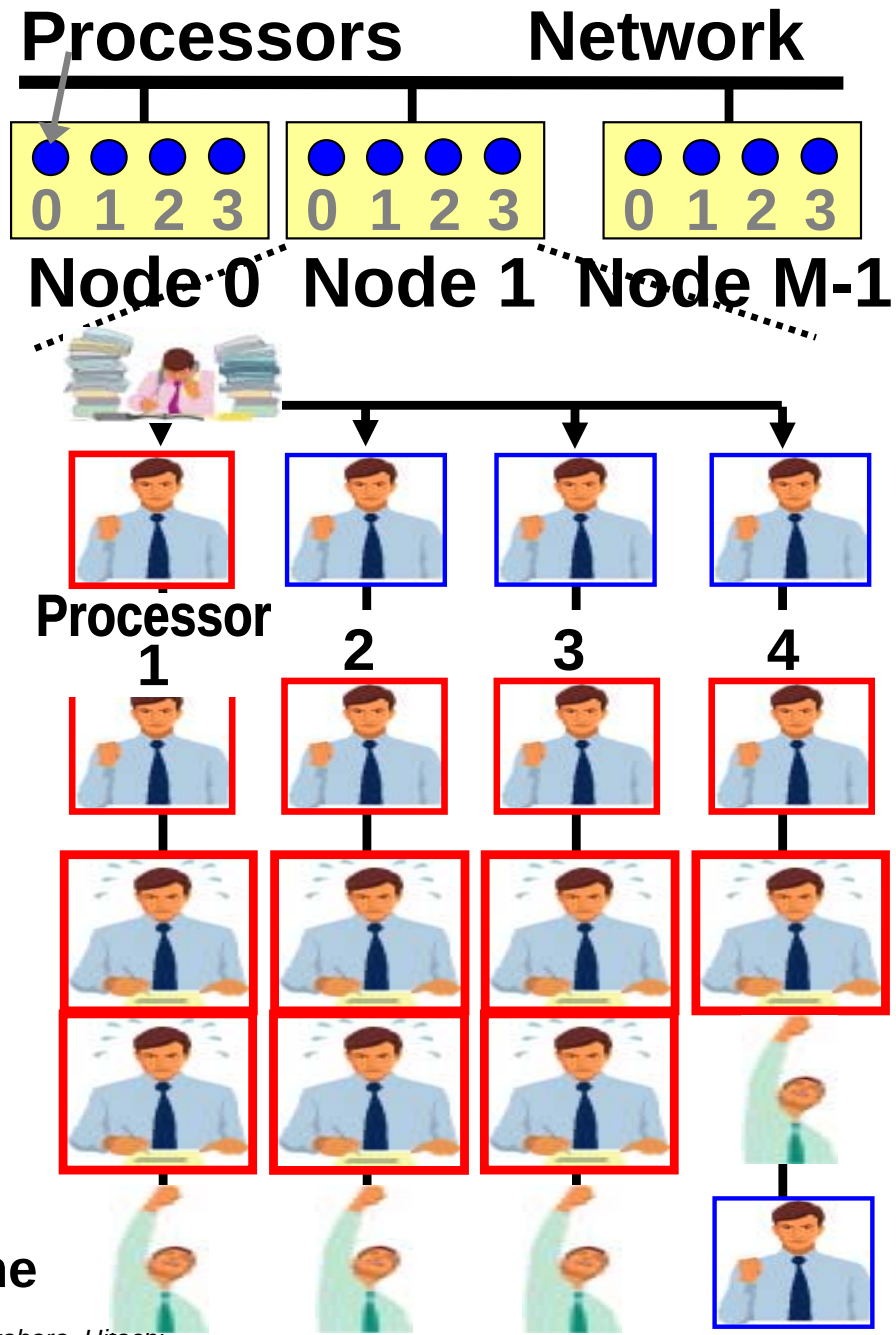
Conventional Architecture **Non-Volatile Architecture**

Ex. In Parallel Processing System



H. Aoki, et al., VLSI Circ. 2008

Power-down during Parallel Processing?



H. Aoki, et al.,
VLSI Circ. 2008

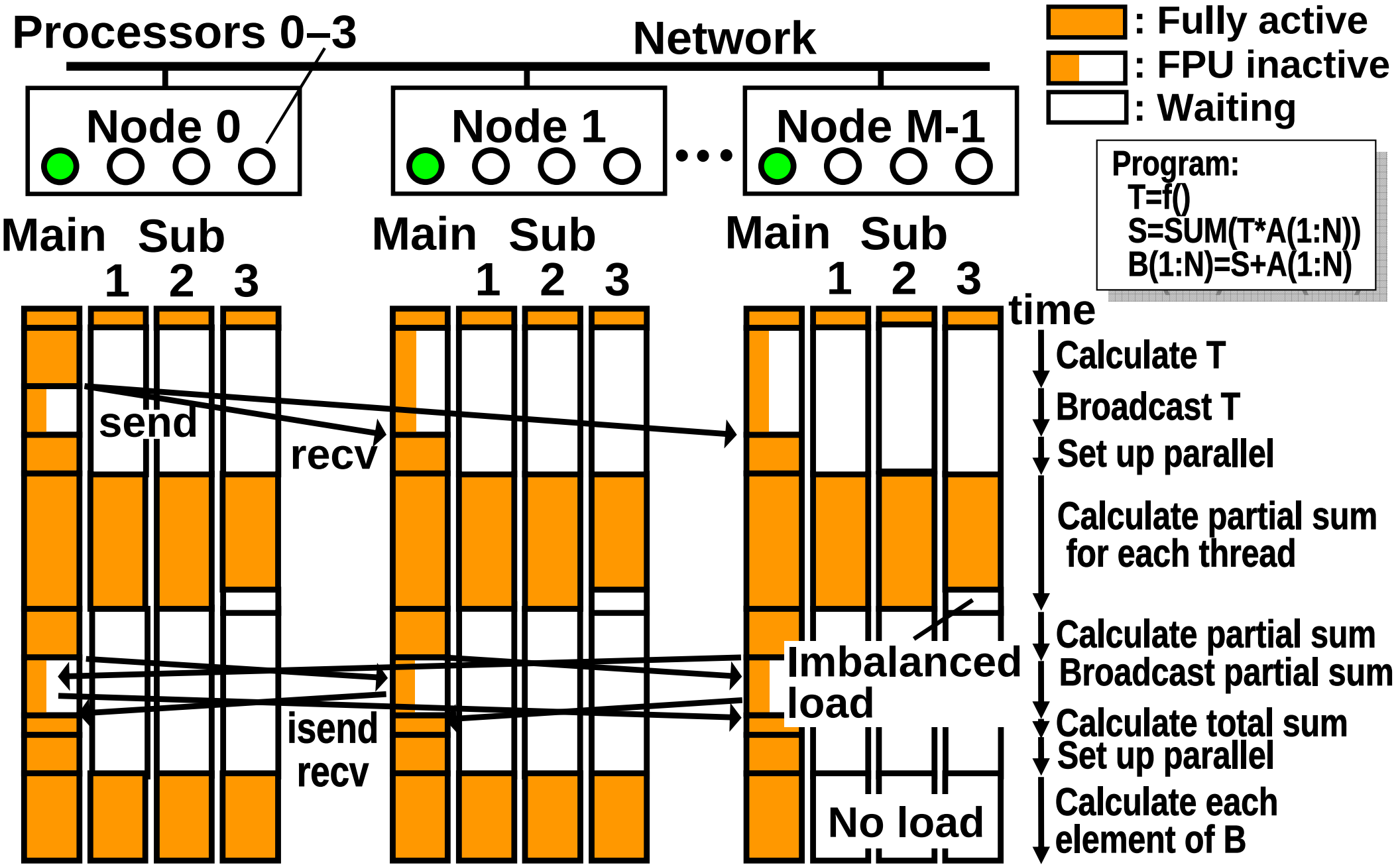
Preparation for processing
Communication with other Proc.
- **Needs μs order**
- **Not scalable to processor speed**
- **Larger parallel, longer time**

-> use for fine-grain power control

In parallel processing
- **Hard to allocate entire processor**
- **Hard to assign equal load either**
- **Hard to assume homogeneity**

-> use for fine-grain power control

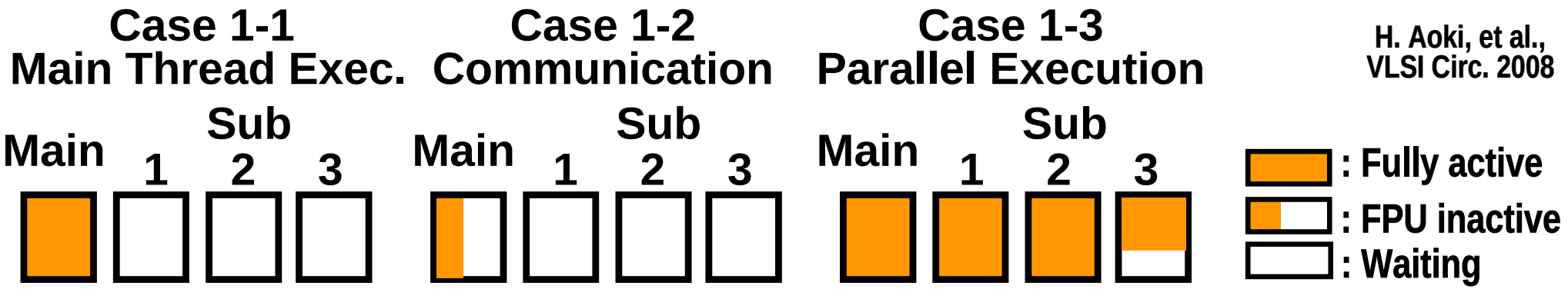
Power-down Controllability



H. Aoki, et al., VLSI Circ. 2008

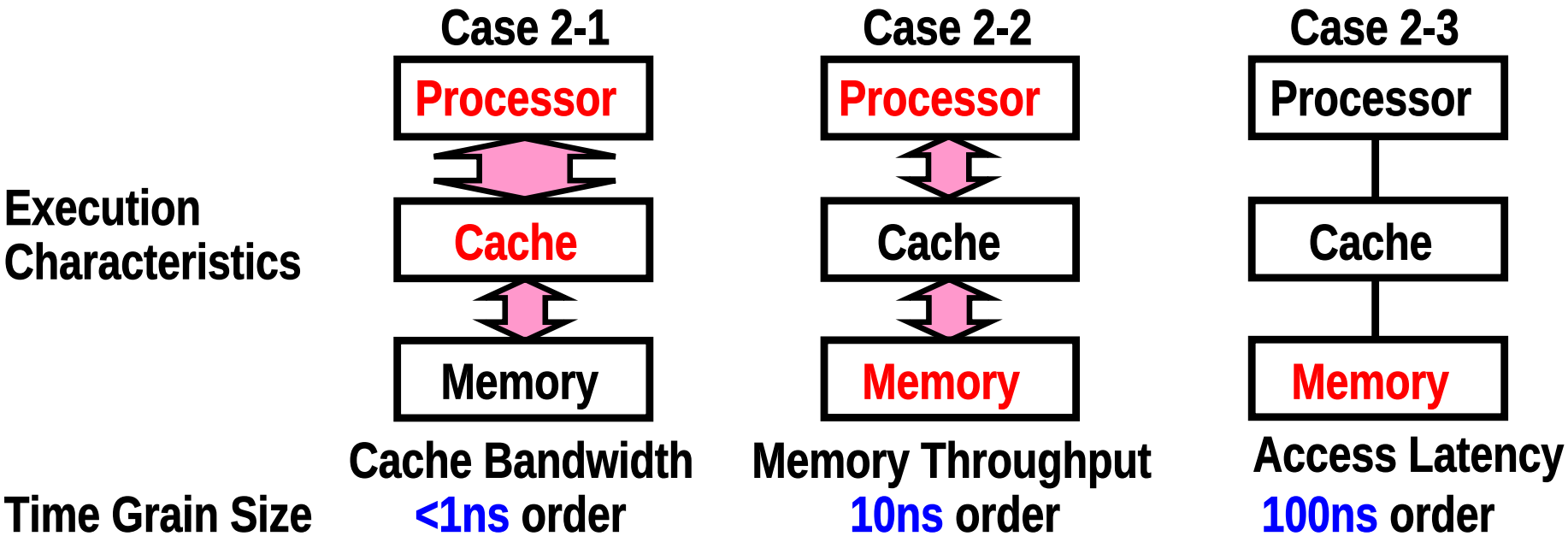
Power Control in Parallel Processing

- High-performance processor capacity is underutilized.
- Can reduce power. Need automatic implementation tools.



Case1-1+1-2, Time Grain Size : μs order

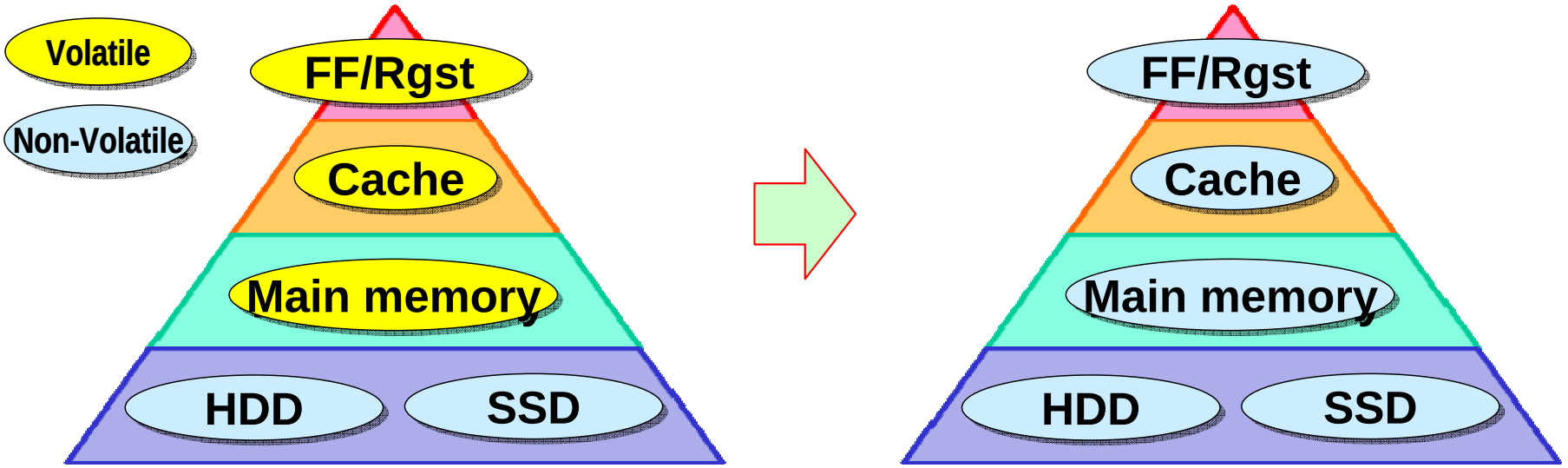
In case 1-3: Execution Cases with Memory Hierarchy



Non-Volatile Architecture

- **Operation of sub-processor has some patterns.**
  - **Need tool that optimizes these to reduce power in the actual flow of system.**
- **Advancement going on for logic level normally-off and instant-on operation.**
  - **Need tool that automatically optimized for logic synthesis and layout included power supply line**
- **Need merged with event-driven, asynchronous computing.**

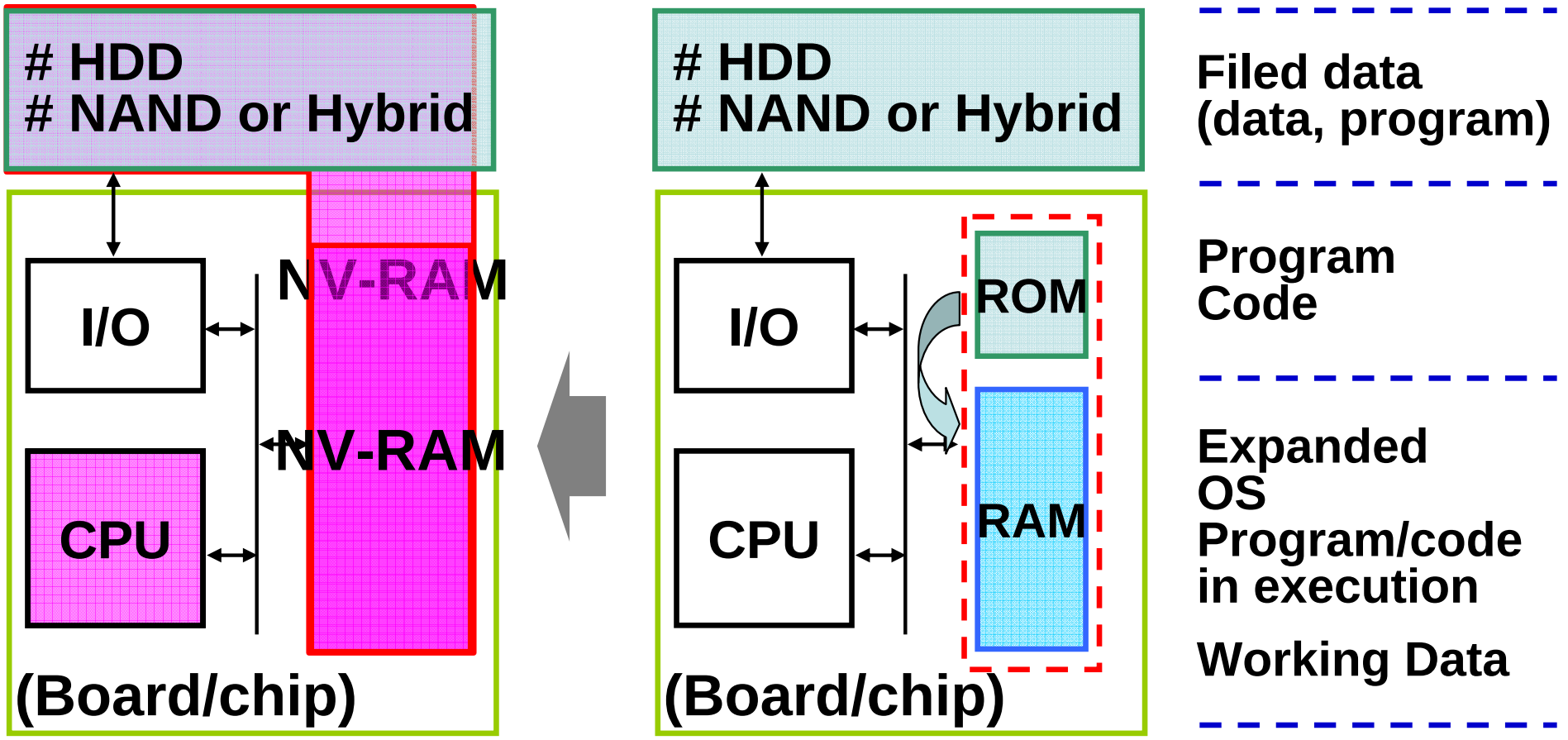
Normally-off and instant-on system



Conventional Architecture **Non-Volatile Architecture**

Normally-off, Instant-on Computing

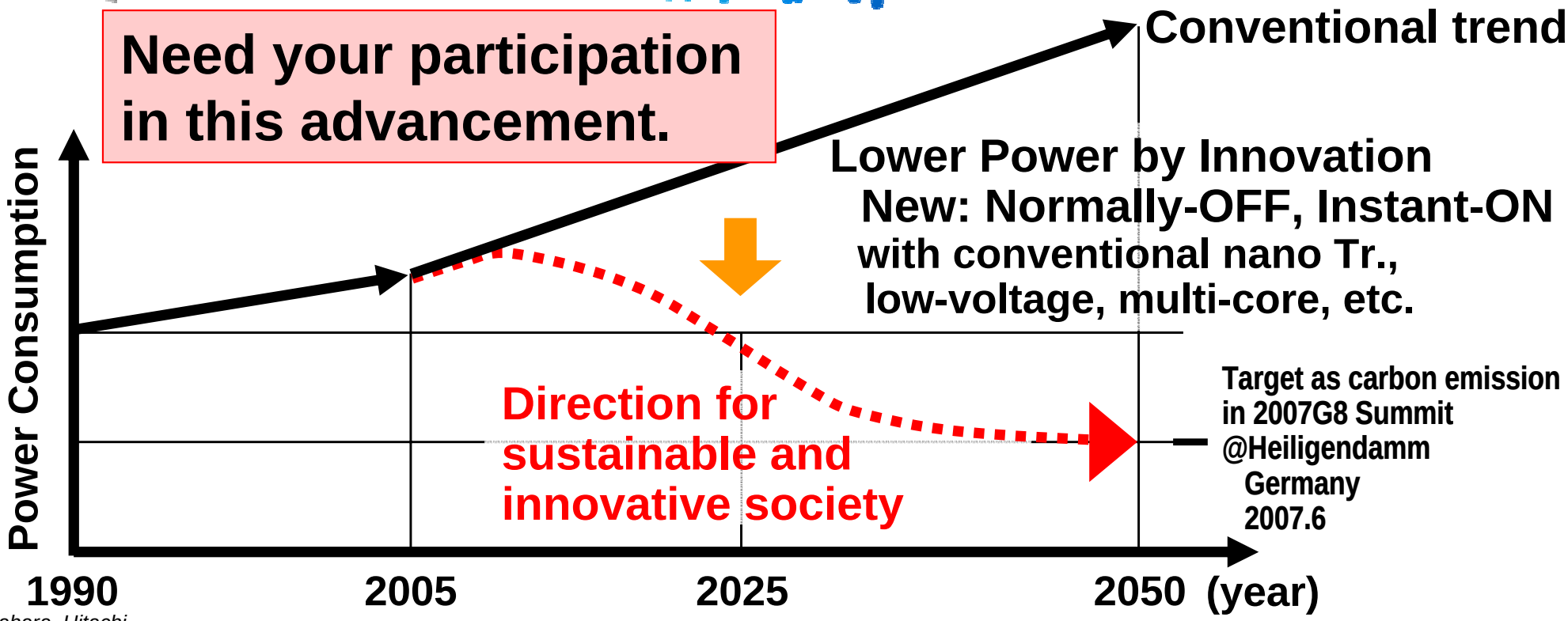
- **IT equipment to be turned on instantly and be resumed the state prior to the interruption.**



Sustainable and innovative society

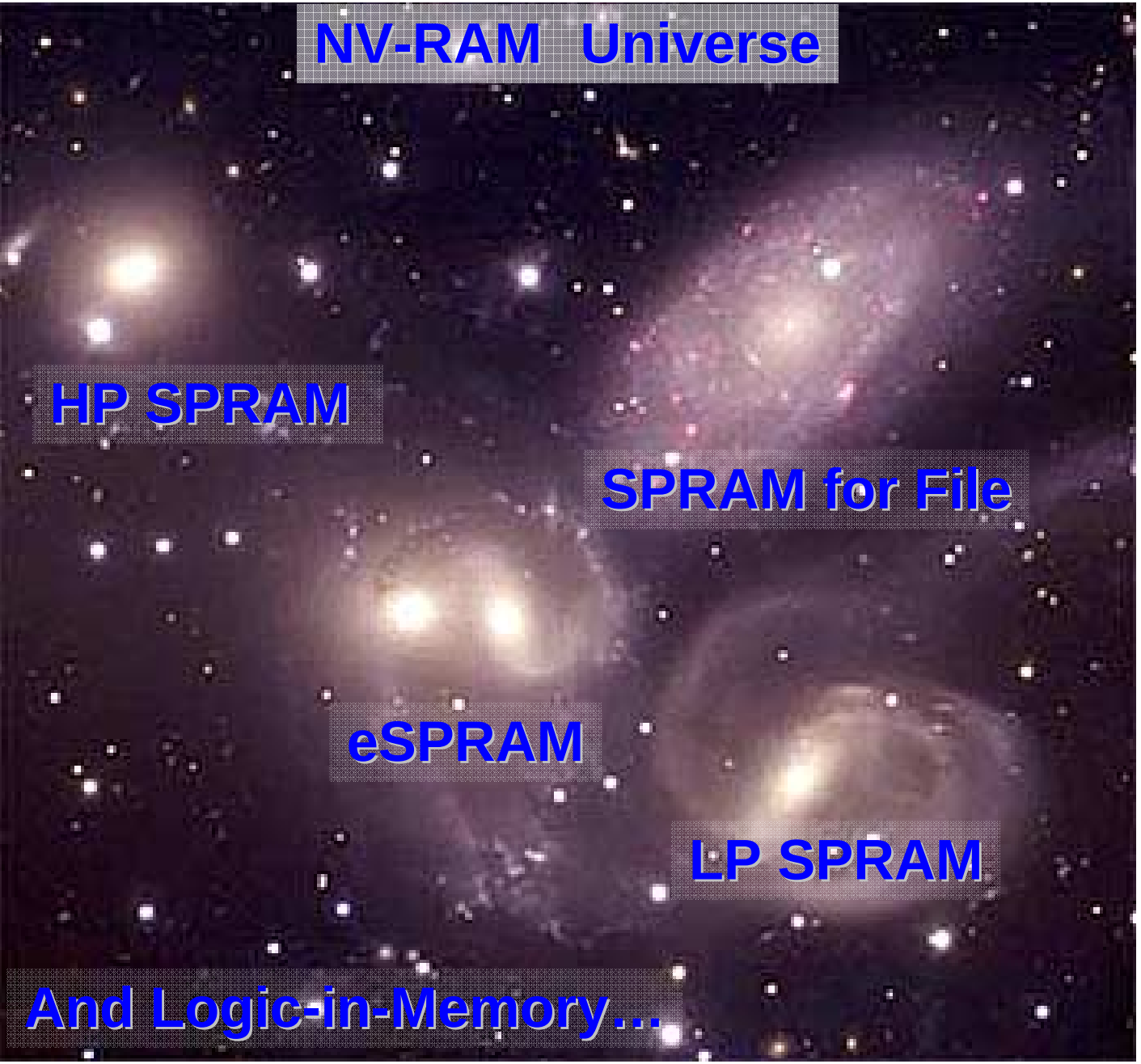
- **Low power innovation is indispensable for sustainability.**
- **Normally-off, Instant-on computing: Harmonizing the power control tech. and the information tech. for our social innovation.**

Need your participation in this advancement.

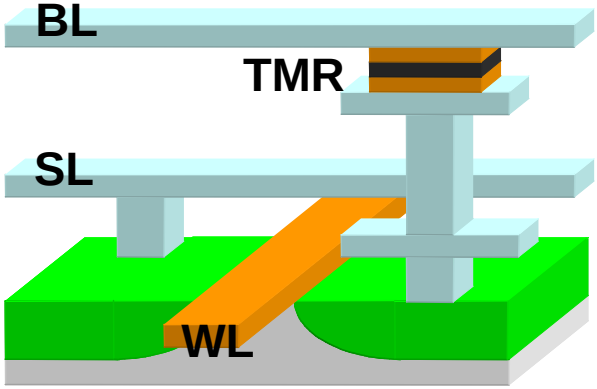


T.Kawahara, Hitachi

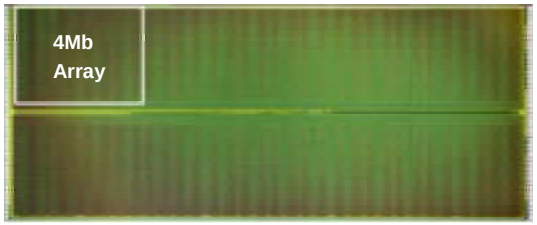
Non-Volatile RAM: a Key Component



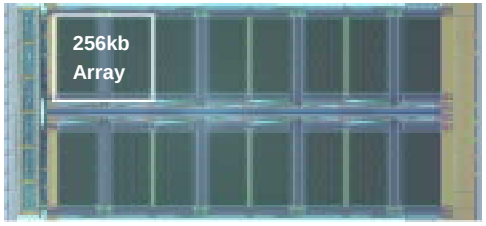
NV-RAM: infinite number of fast write and read operations as well as non-volatility.



Spin-transfer torque RAM (SPRAM)



32Mb Chip (VLSI 2009, Hitachi/Tohoku U)



2Mb Chip (ISSCC 2007, Hitachi/Tohoku U)

Outline

- Introduction
- Non-volatile memories
- Spin-transfer torque RAM scalability
- Normally-off instant-on computing
- **Conclusion**

Conclusion

- **Normally-off, Instant-on Computing: a way to allow computing equipment to normally be turned off when not in use but be able to turn on instantly with full performance when needed.**
- **Non-volatile memory: any internal status of computation is memorized at any time before the power is turned off without consuming power. NV-RAM: Infinite number of fast write and read operations as well as non-volatility.**
- **Tools for physical analysis and implementation with circuits are important for the deign of emerging memory.**
- **Tools that clarify internal computation status and manage fine grain power control combined with NVRAM are necessary.**

Harmonizing power control tech. and information tech. for the social innovation.