Power consumption: roadblock for AI moving forward

Shunpei Yamazaki Ph.D.

President, Semiconductor Energy Laboratory Co., Ltd.

SEMICONDUCTOR ENERGY LABORATORY

aiworld
State of the Practice of Enterprise AI

Introducing SEL

- Company name: Semiconductor Energy Laboratory Co., Ltd. (SEL)
- Location: 398, Hase, Atsugi-shi, Kanagawa, Japan
- Established on July 1, 1980
- Number of employees: 793 (as of April 1, 2017)
- Representative: President Shunpei Yamazaki
- Capital: JPY 4,348,000,000
- Business area: Research and development of transistors, semiconductor devices, and integrated circuits using crystalline oxide semiconductor materials

Flexible OLED display

Outline

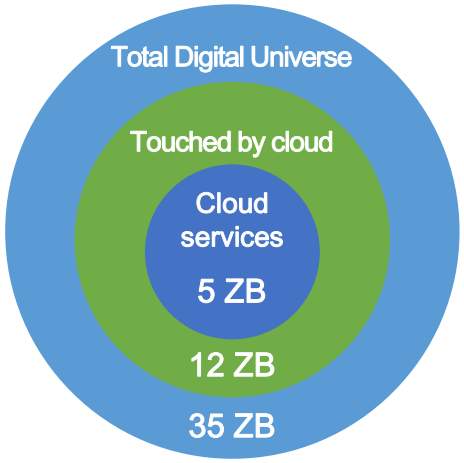
1. What are the roadblocks of AI development?
2. Applying crystalline oxide semiconductors to AI
   - 2-1. DOSRAM (DRAM fabricated with OSFETs)
   - 2-2. Nonvolatile memory, NOSRAM: multi-level memory with no degradation in principle
   - 2-3. Multiply-accumulate operation circuit
   - 2-4. CPU (Noff CPU)
   - 2-5. LSICs with crystalline oxide semiconductor technology
3. Summary

1. What are the roadblocks of AI development?

1) Necessity of large-capacity memory
2) Latency (software solutions underway)
3) Power consumption

Issues surrounding AI

1. Data explosion

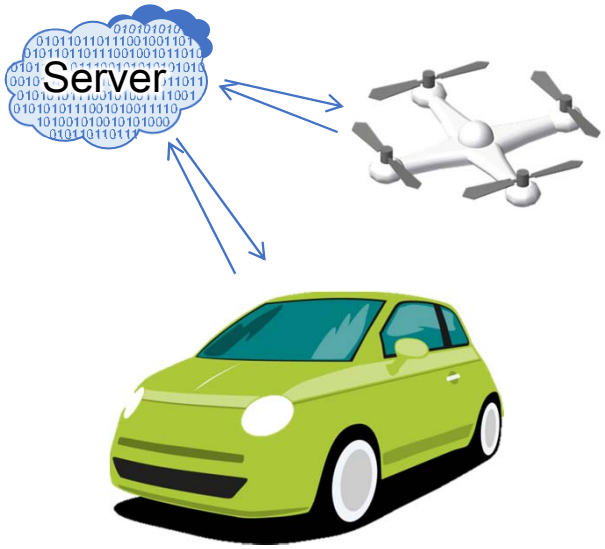


1) IDC “Digital Universe Decade-Are You Ready?” https://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf

* ZB: zettabyte=10^{21}byte

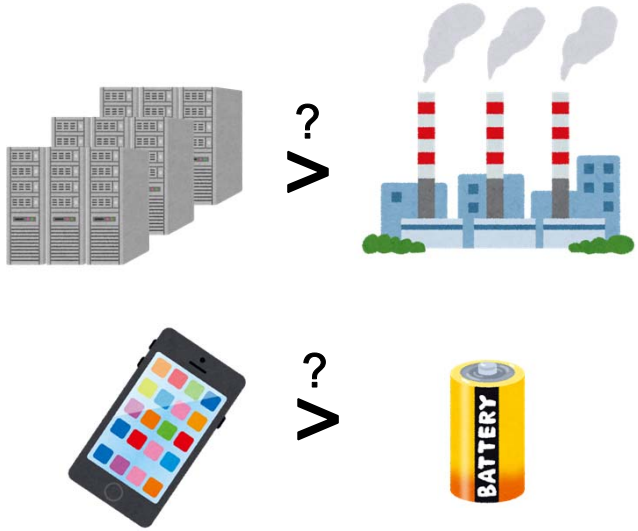
<u>Edge computing necessary</u> as global data use will increase 10x by 2020[1)]

2. Latency



<u>Latency unacceptable</u> in self-driving cars and drones

3. Power Wall



<u>AI's power consumption impedes</u> edge computing implementation

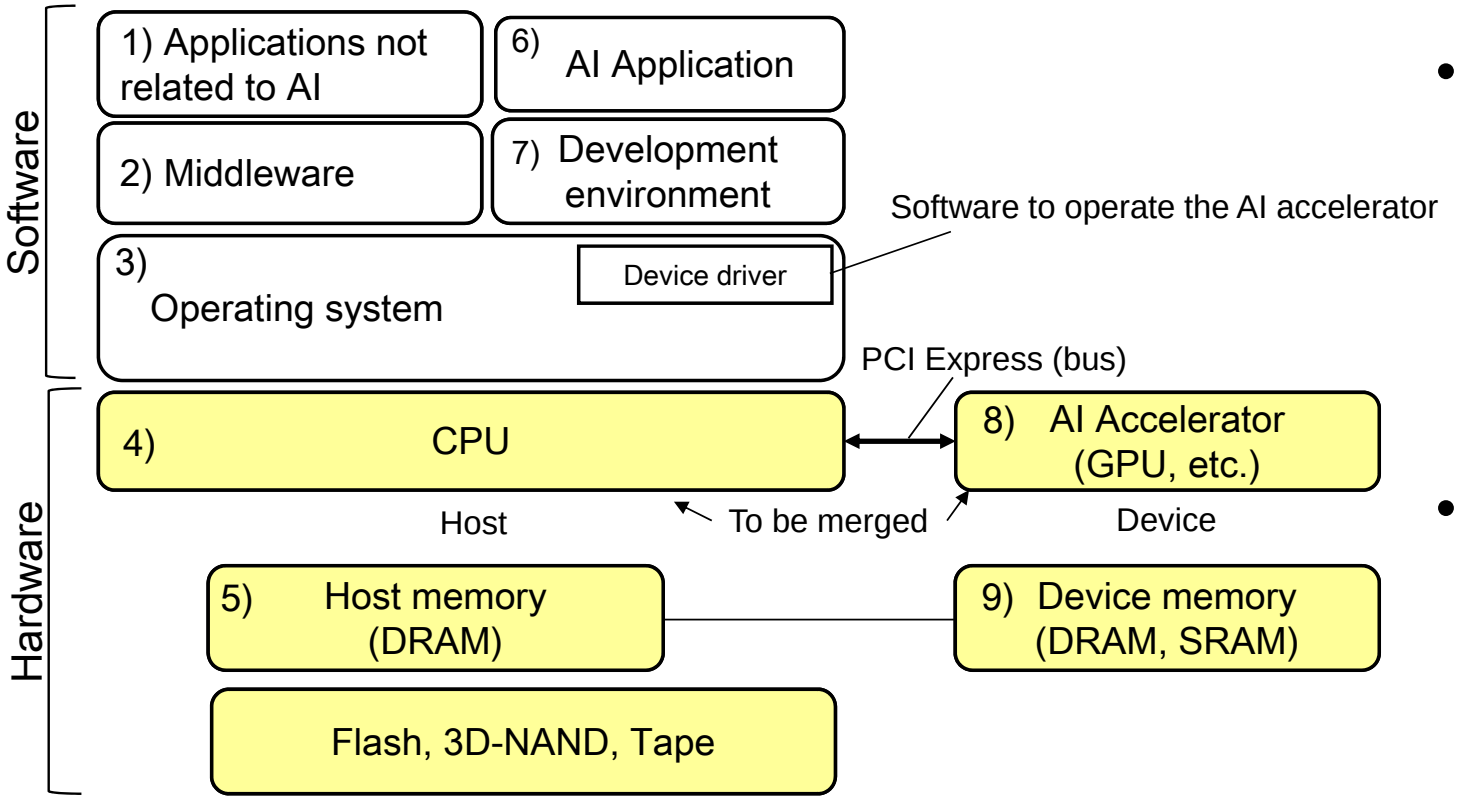
Enormous energy consumption in AI development

- Korea currently is undertaking a big project to speed up their R&D

> *"In order to use actual AI technologies properly in daily lives, experts in related industries say that current amount of electricity consumption needs to be reduced by 1,000 fold."*
>
> (*From ET News dated Aug. 22, 2017* (http://english.etnews.com/20170822200001))

Power consumption of semiconductor devices is the largest factor impeding AI development and extremely low-power AI hardware is needed for future development of AI.

Hardware/software structure of AI



- AI hardware's total power consumption = AI accelerator + CPU + DRAM, etc…
- Other devices around the accelerator need to run efficiently.

How do we achieve the power reduction?

- Size and voltage scaling of Si semiconductor devices could, but they have limits
- New technology is needed to reduce the power consumption of AI solutions.

> *"In order to reduce amount of electricity consumption, next-generation device technology that can replace CMOS (Complimentary Metal-Oxide Semiconductor) ,which is the foundation of current semiconductor device technology, needs to be developed."*
>
> (*From ET News dated Aug. 22, 2017* (http://english.etnews.com/20170822200001))

FETs fabricated with CAAC-IGZO (a crystalline oxide semiconductor material)

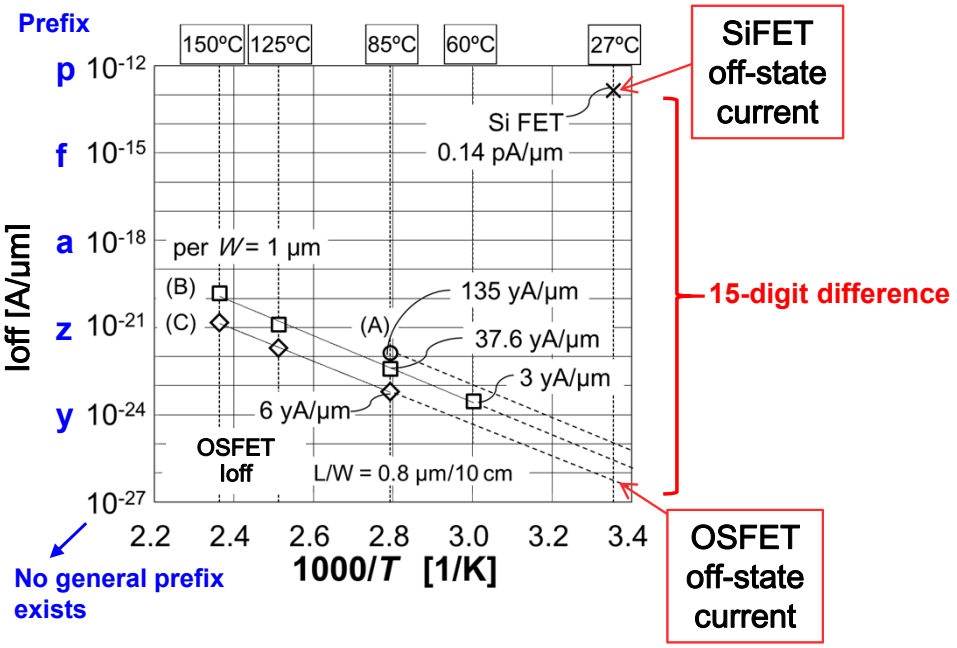
- Off-state current (Ioff) :10^{-24}A/μm (15 digits lower than FETs fabricated with Si,)
- Nonvolatile and high-speed memory devices can be achieved[1)].

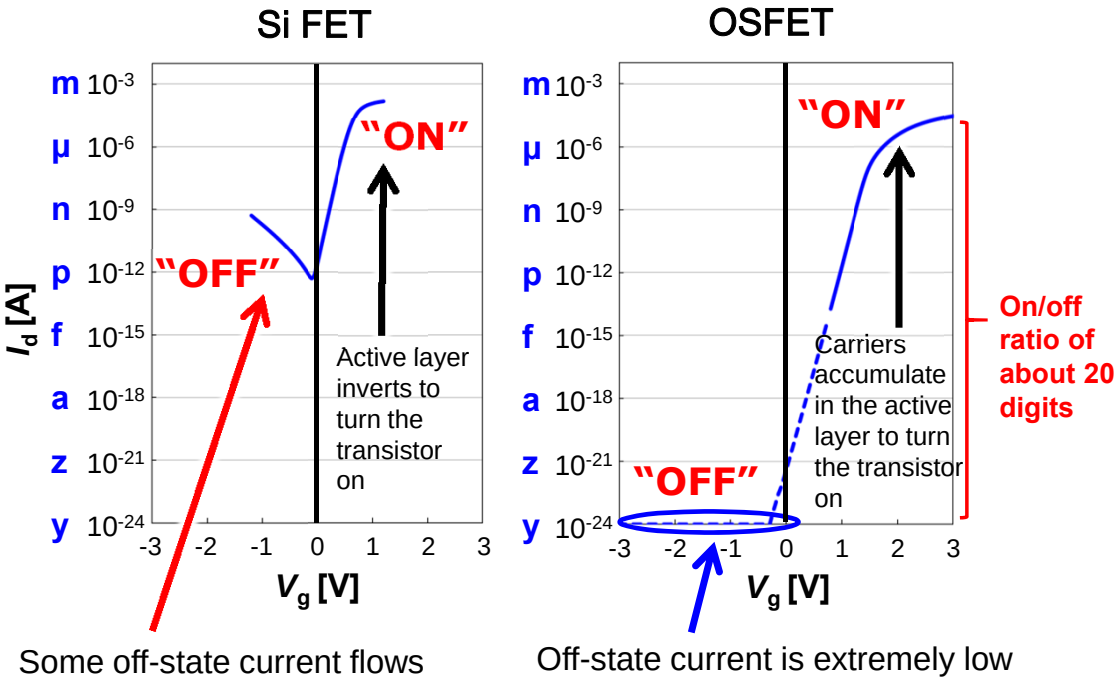
CAAC-IGZO is the most fitting technology that can challenge Si-CMOS

1) "Physics and Technology of Crystalline Oxide Semiconductor CAAC-IGZO: Application to LSI" (2017) John Wiley.

Extremely low off-state current (Ioff)

- CAAC-IGZO FET has an extremely low Ioff and can drastically reduce the energy consumption of devices.

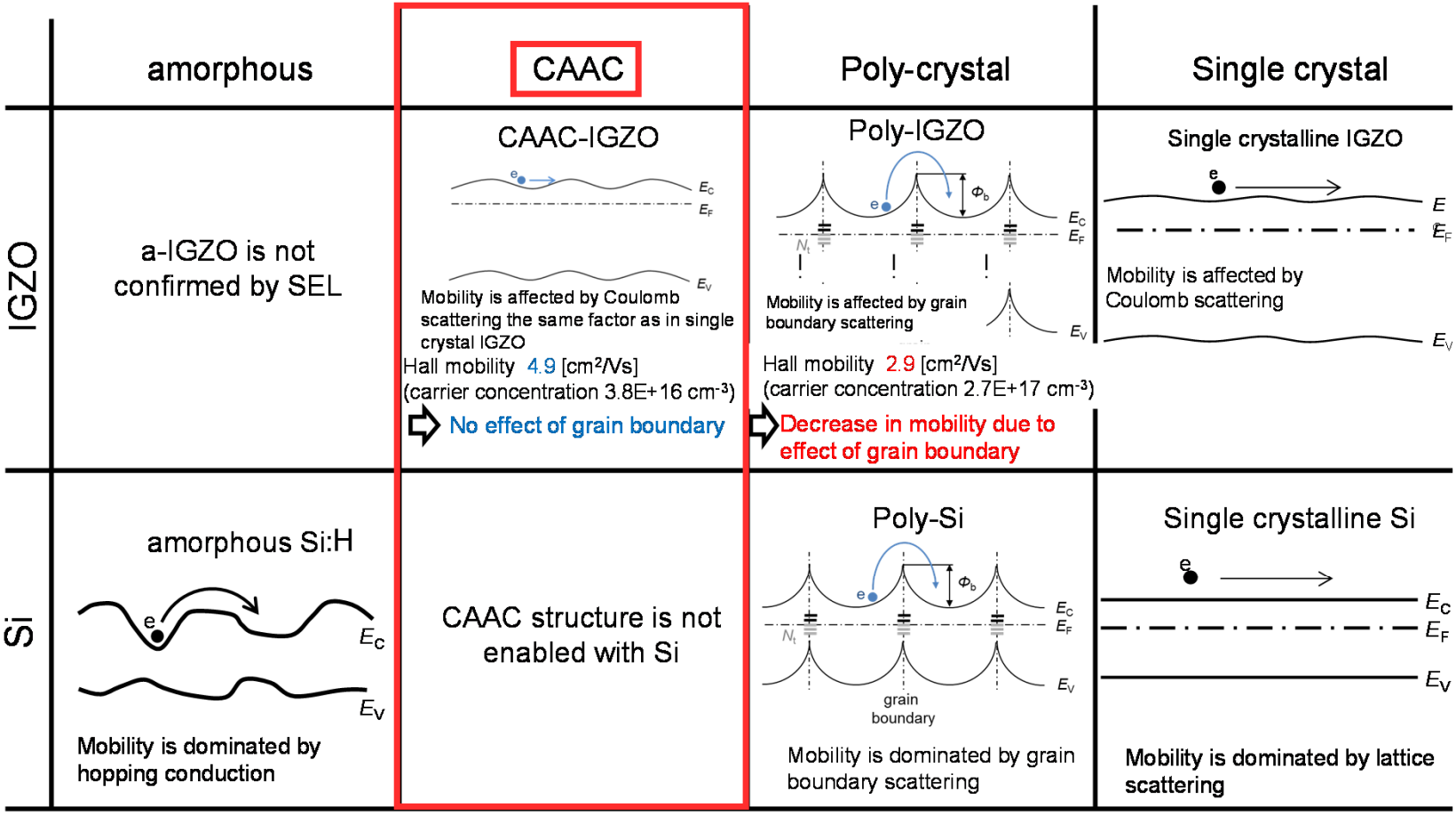




- OSFETs: FETs that use crystalline oxide semiconductors (e.g. CAAC-IGZO)

(A) K. Kato, *et al.*, Jpn. J. Appl. Phys. 51, (2012) 021201.
(B) Reported at FPD International '13
(C) H. Tamura, *et al.*, *IEEE micro,* **34**,6, (2014) 42.

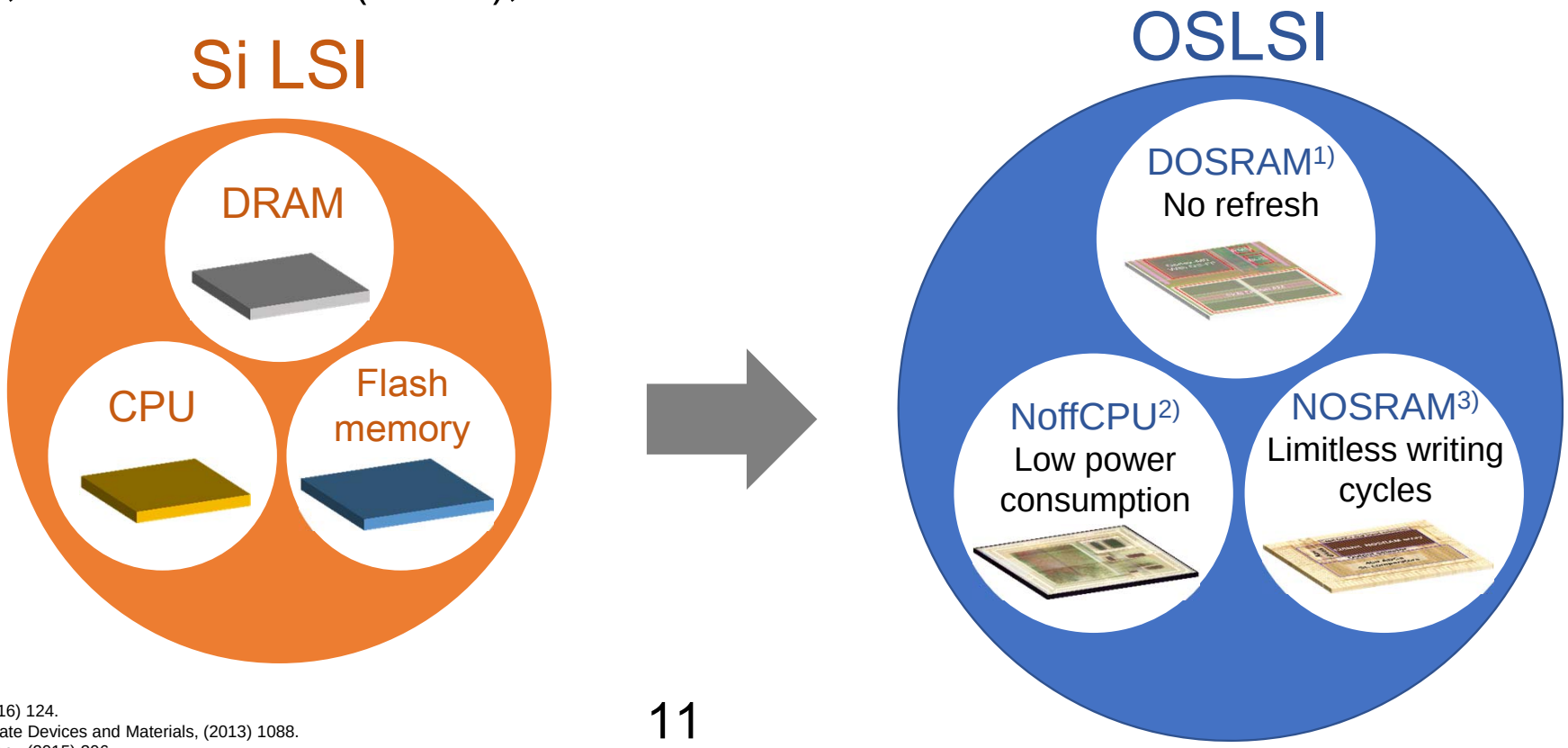
IGZO's crystal structure and an image of its electrical conduction



CAAC is a new crystal structure

OSLSI is applicable to a variety of hardware

- LSICs using CAAC-IGZO (OSLSI) are applicable to FPGAs, CPUs (NoffCPUs), DRAMs, AI accelerators (GPUs), etc.

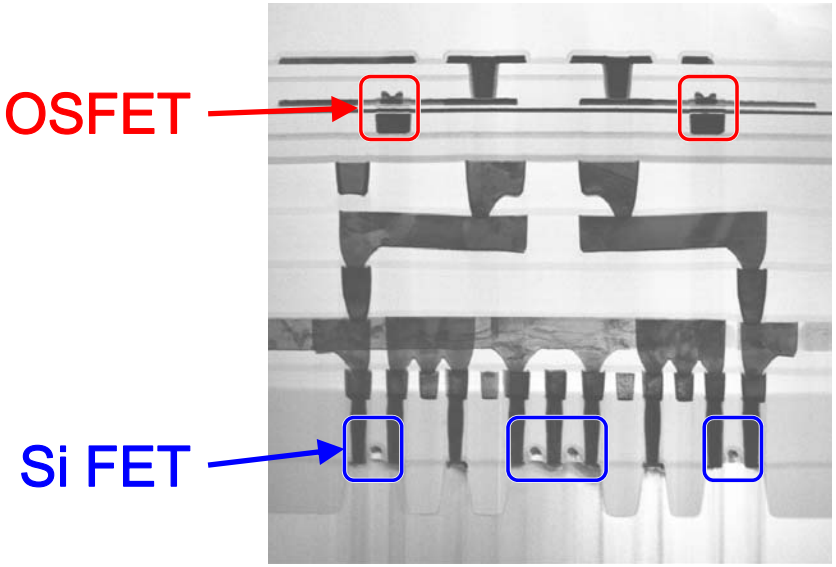


1) T. Onuki *et al*., *Symp*. *VLSI Circuits*, (2016) 124.
2) Niclas Sjokvist *et al.*, Ext. Abstr. Solid-State Devices and Materials, (2013) 1088.
3) T. Matsuzaki *et al*., ISSCC Dig. Tech. Pap., (2015) 306 .

Hybrid structure of OSLSI chip

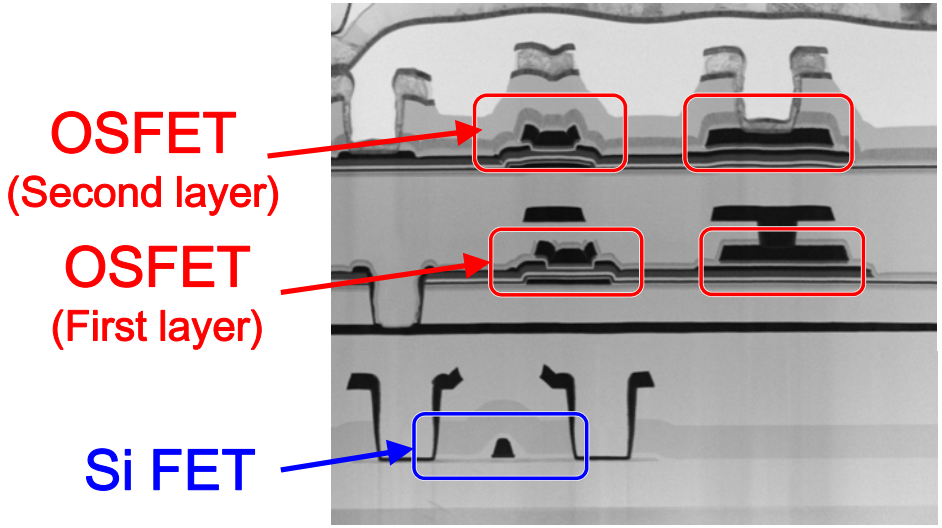
*voids are filled in the photo.

- Single OSFET layer structure



Cross section of SEL 60nm OS / UMC* 65nm Si

- Multi-layer OSFET



Stack OSFETs
→higher area efficiency and memory capacity

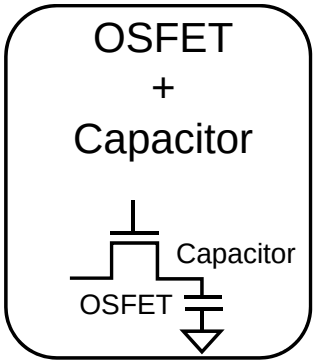
UMC*:United Microelectronics Corporation

2. Applying crystalline oxide semiconductors to AI

CAAC-IGZO (c-axis aligned crystalline IGZO)

2-1. DOSRAM
(DRAM fabricated with OSFETs)

DOSRAM: Dynamic Oxide Semiconductor Random Access Memory

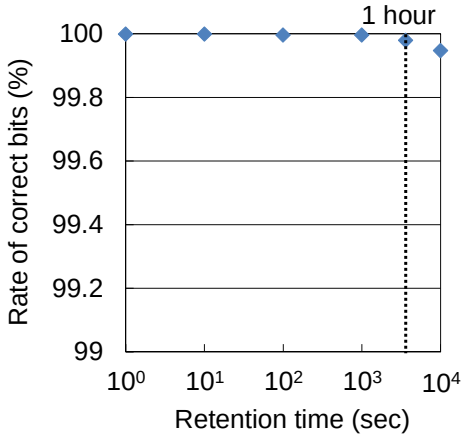


OSFET: - Extremely low Ioff (15 digits lower than Si)
- Adequate Ion/frequency characteristics

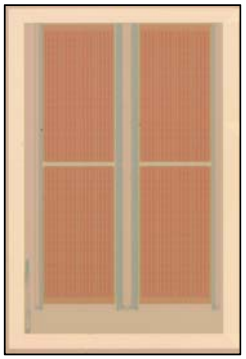
>10 DRAM chips in 1 AI chip

DRAM's refresh period: 60 - 200 ms
DOSRAM's refresh period: 1 h - 1 yr

Non-volatile DRAM



Retention characteristics of DOSRAM[1)]

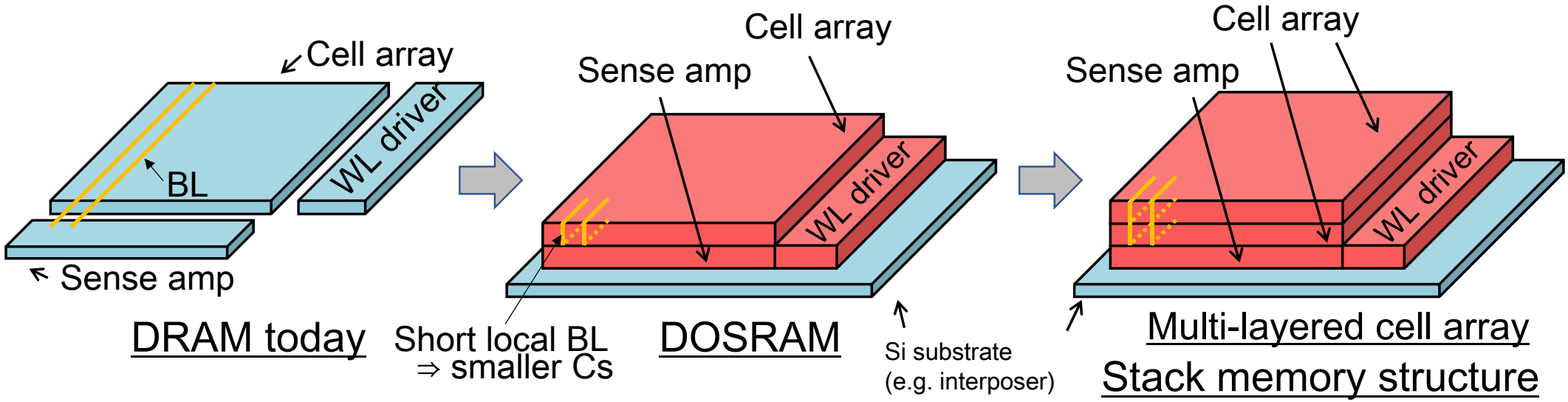


DOSRAM chip micrograph[1)]

1) T. Onuki *et al.*, Symp. VLSI Circuits, (2016) 124.

OSFET stack (multi-layer) structure of DOSRAM

- OSFET layers (and thus, cell arrays) can be fabricated over other FET layers
  →larger capacity than DRAM
  →smaller Cs (= lower refresh power)



2-2. Nonvolatile memory, NOSRAM[1)]:

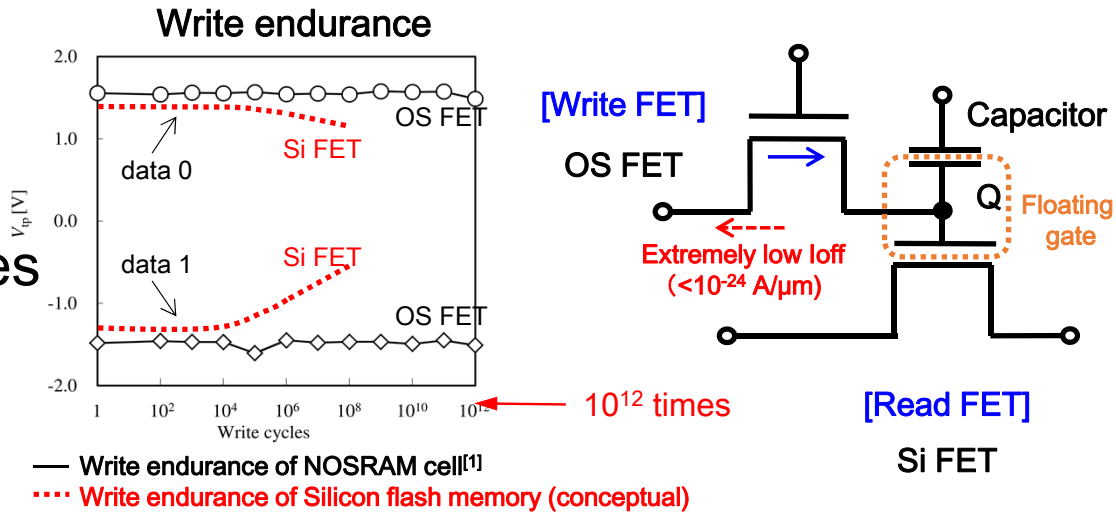
Multi-level memory with no degradation in principle

NOSRAM: Nonvolatile Oxide Semiconductor Random Access Memory
2 transistors, 1 capacitor

- **Si flash memory**:
  Endures only 10^4 to 10^5 writes

NOSRAM:
Endures > 1,000,000,000,000 (10^{12}) writes

- Si read transistor enables fast writes
- OSFETs can be stacked over SiFETs



NOSRAM develops into a MAC circuit applicable to AI (shown next slide)

1) H. Inoue *et al.,* IEEE J. Solid-State Circuits, vol. **47**, no. 9, (2012) 2258.

Application example of NOSRAM to AI operations

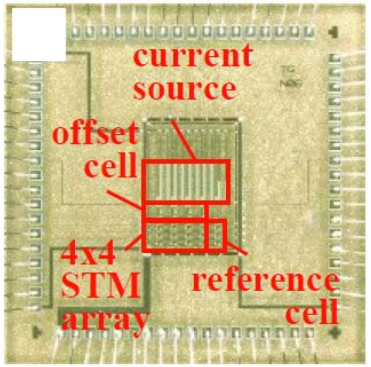
NOSRAM cell

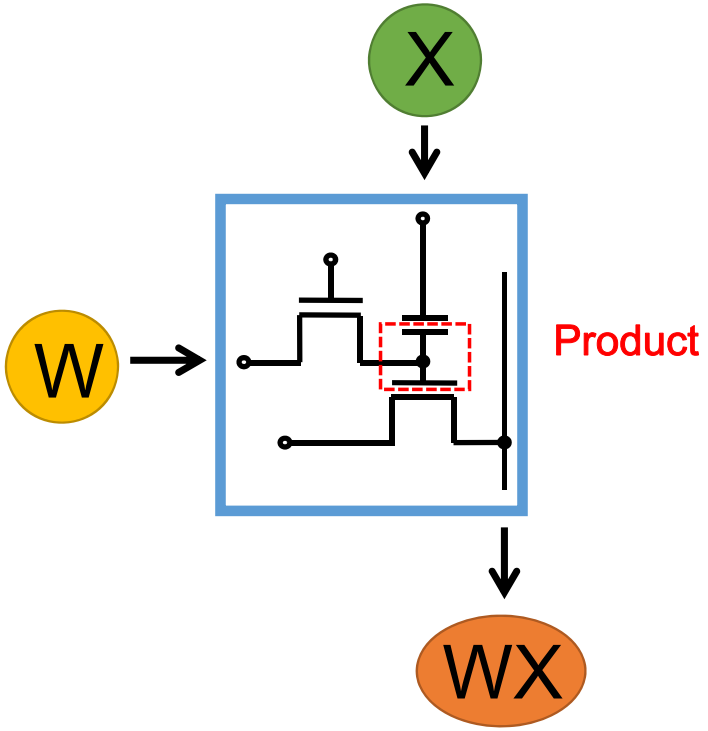
= Applicable to AI operation[1)2)]

NOSRAM cell calculates product of X and W

↓

Product sum obtained by simply connecting the WX outputs





(Details on the next slide)

1) Demonstrated at "1st AI EXPO", June 28 to 30, 2017, Tokyo Big Sight, Japan
2) T.Aoki, *et al.*, Ext. Abstr. Solid State Devices and Materials, (2017) 191

2-3. Multiply-accumulate operation circuit

– I_0 and I_1: current from memory cells

→ **Product** ($W_0 \times X_0$, $W_1 \times X_1$)

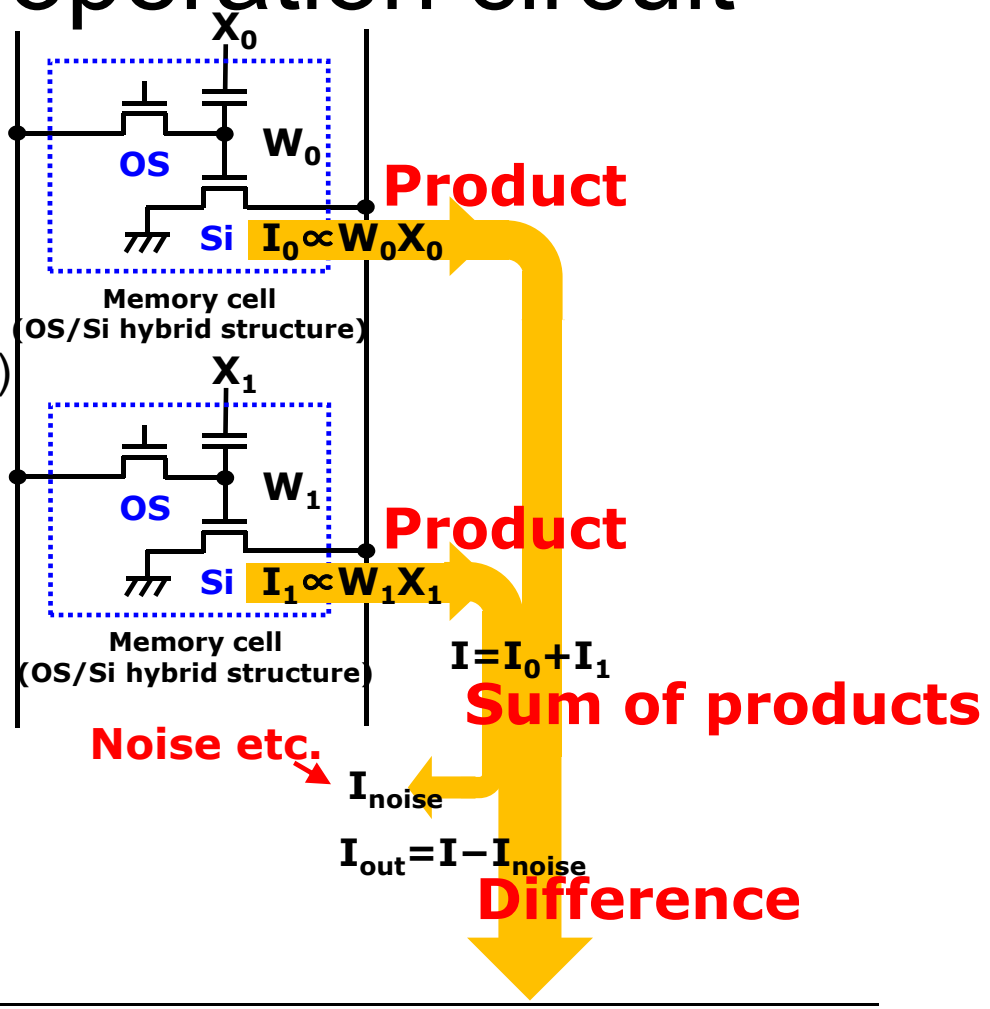
– Current: current I (sum of I_0 and I_1)

→ **Sum of products** ($I = I_0 + I_1$)

– Subtract noise etc. from current I

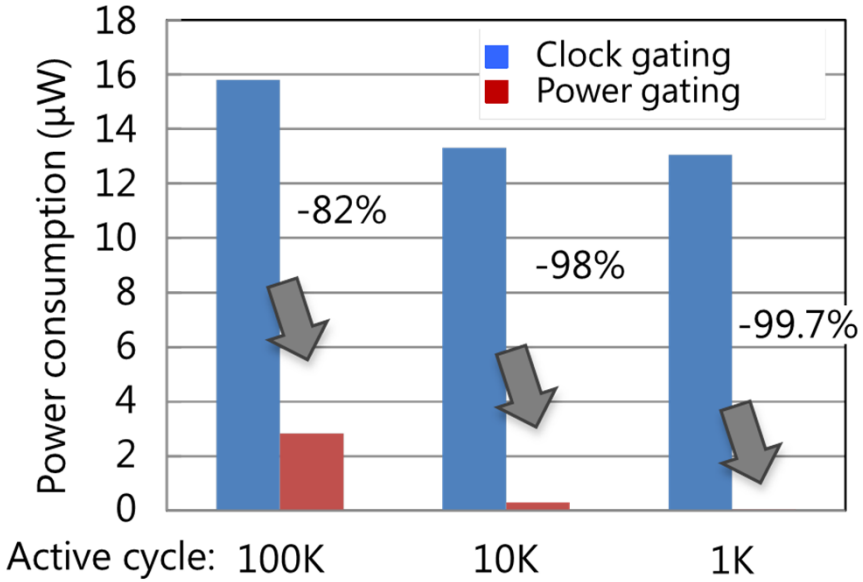
Current I_{out} ($I_{out} = I - I_{noise}$)

→ **Difference**



- Operation accuracy can be improved with difference operation and multiply-accumulate operation.

2-4. Example of CPU with OSLSI (NoffCPU)

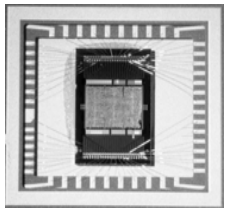


※ 60-nm normally-off CPU + 8KB DOSRAM
※ Clock gating : Sleep mode by stopping a clock
※ Power gating : Sleep mode by shut-off

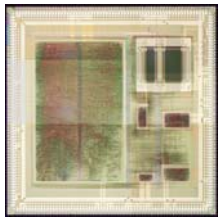
1000x CPU standby power reduction

S.H. Wu *et al.*, *Symp. VLSI Technology*, (2016) 58.

2-5. LSICs with crystalline oxide semiconductor technology (OSLSI)

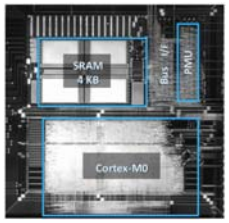


(1) 8bit Noff* CPU
SSDM 2012

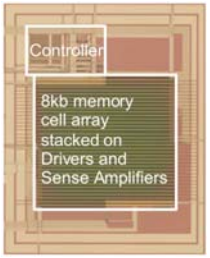


(2) 32bit Noff CPU
SSDM 2013

(3) FPGA
ISSCC 2014

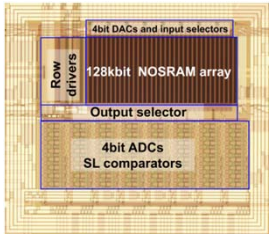


(4) Cortex-M0 with SRAM
Cool Chips 2014

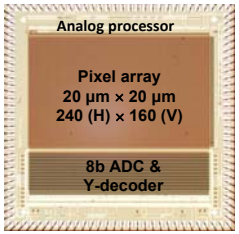


(5) DOSRAM
IMW 2012

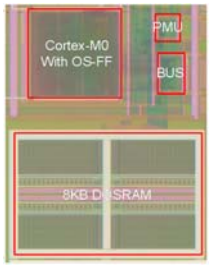
(6) 1Mbit NOSRAM
JSSC 2012



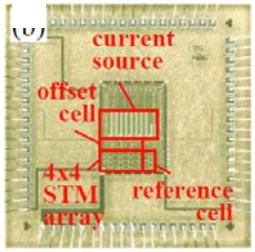
(7) 4bit/cell NOSRAM
ISSCC 2015



(8) Image Sensor
ISSCC 2015



(9) DOSRAM
VLSI 2016

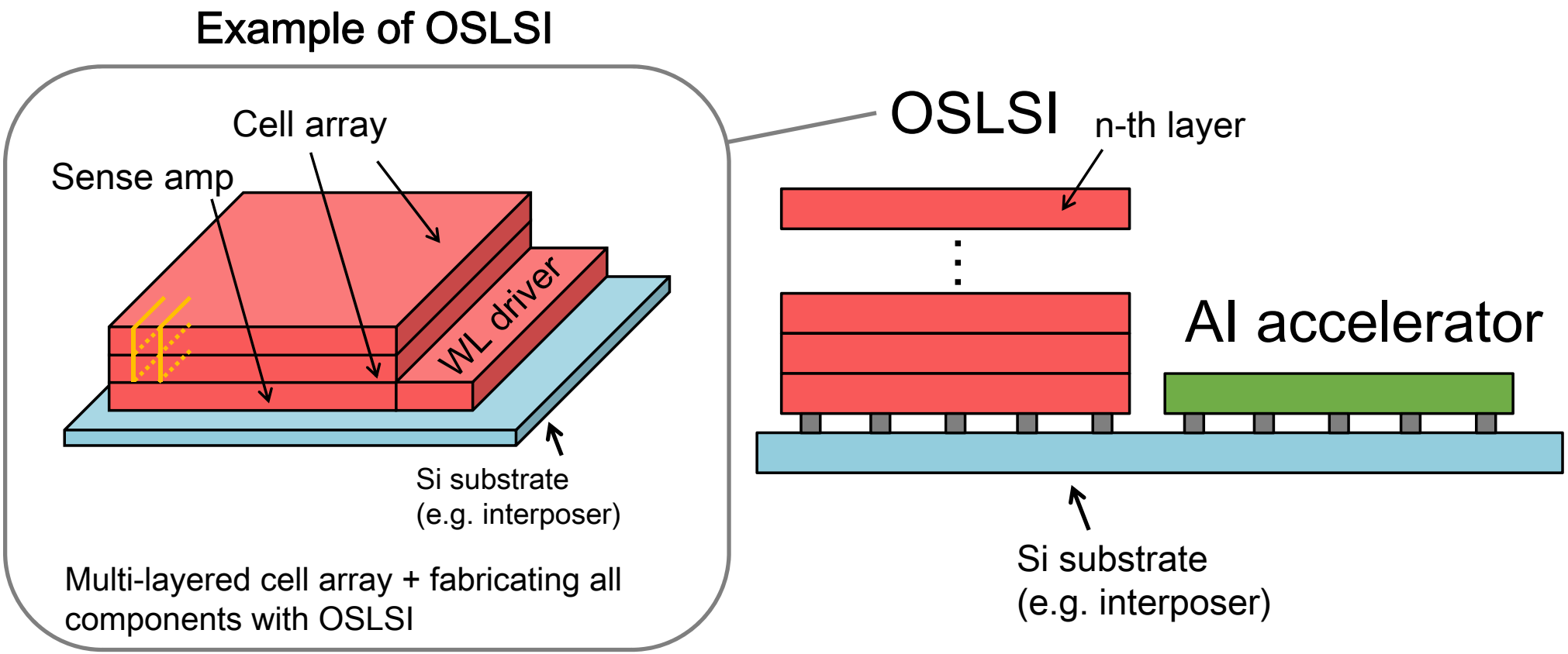


(10) STM
SSDM 2017

*Noff: normally off

(1) T. Ohmaru *et al.*, Ext. Abstr. Solid-State Devices and Materials, (2012) 1144.
(2) Niclas Sjokvist *et al.*, Ext. Abstr. Solid-State Devices and Materials, (2013) 1088.
(3) T. Aoki, *et al.*, ISSCC Dig. Tech. Pap., (2014) 502.
(4) H. Tamura *et al.*, COOL Chips XVII, (2014).
(5) T. Atsumi *et al.*, 4th IEEE Int. Memory Workshop, (2012) 99 .
(6) H. Inoue *et al.*, IEEE J. Solid-State Circuits 47, (2012) 2258 .
(7) T. Matsuzaki *et al.*, ISSCC Dig. Tech. Pap., (2015) 306 .
(8) T. Ohmaru *et al.*, ISSCC Dig. Tech. Pap., (2015) 118.
(9) T. Onuki *et al.*, *Symp. VLSI Circuits*, (2016) 124.
(10) T. Aoki, *et al.*, Ext. Abstr. Solid State Devices and Materials, (2017) 191.

All-OSLSI Stack LSI structure



3. Summary

(1) All AI hardware run on Si circuits now

(2) Crystalline oxide semiconductor materials (e.g., CAAC-IGZO) enable low-power semiconductor devices
→next-generation semiconductor material

(3) Crystalline oxide semiconductor material is the only material with such a low off-state current

(4) OSLSI challenges SiLSI technology

Thank you!