

Normally-Off Computing Project : Challenges and Opportunities



Hiroshi Nakamura (Speaker)

Takashi Nakada

Shinobu Miwa

Graduate School of Information Science and Technology
The University of Tokyo



Introduction of Normally Off Computing Project



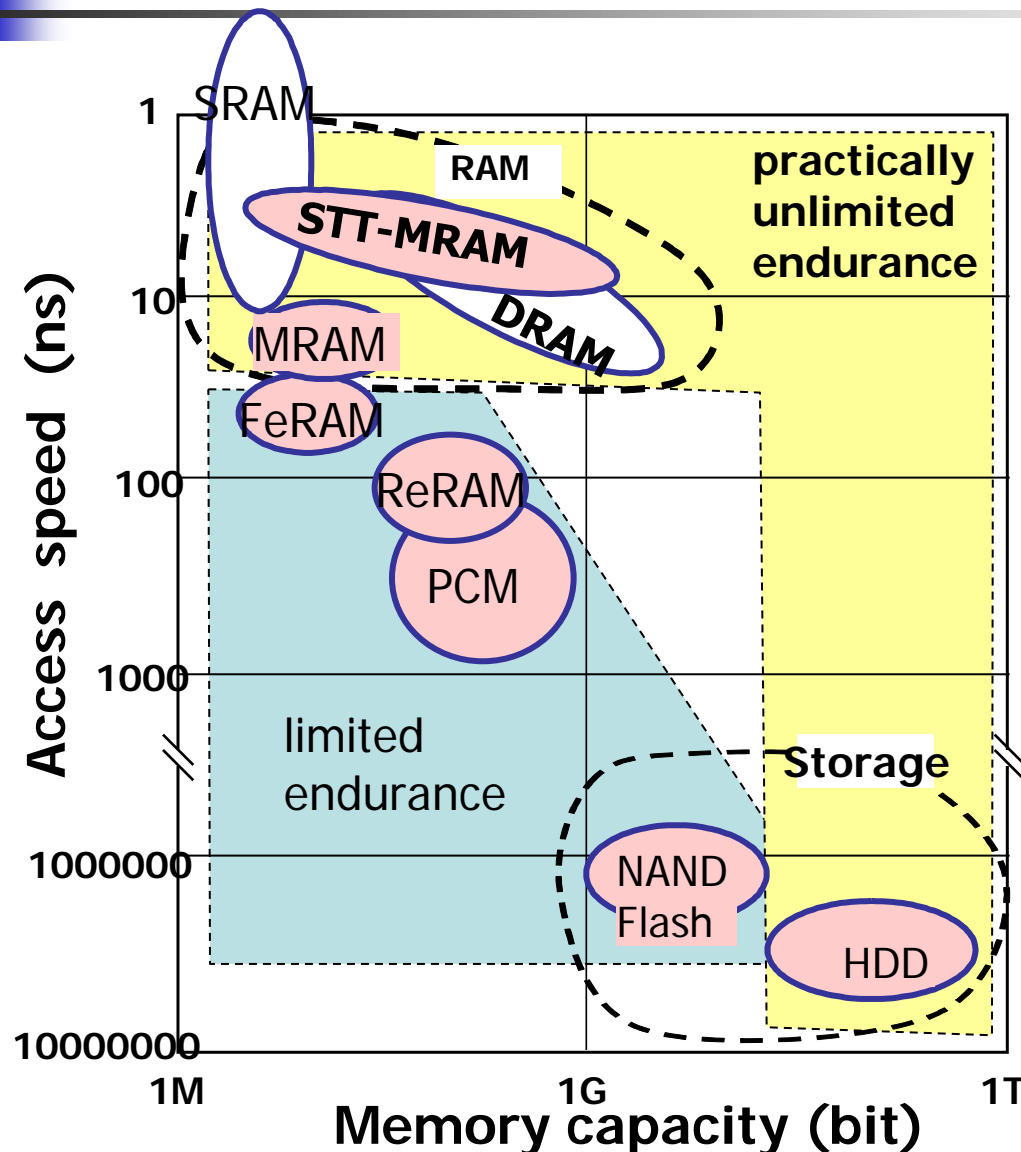
- Project supported by NEDO/METI
 - Participating Corporations: Renesas, Toshiba, Rohm
 - Period : Sep. 2011 – Feb. 2016
 - NEDO : New Energy and Industrial Technology Development Organization
 - METI : Ministry of Economy, Trade and Industry
 - Budget : Half Support by Government
(Approx.) \$7M USD / year by NEDO + \$7M USD / year by Industry
 - Project Leader : Hiroshi Nakamura (U. Tokyo)
 - Project URL: <http://noff-pj.jp/en/>



What is 'Normally-Off Computing'

- Normally-Off: **aggressively powers off** components of computer systems when they need not to operate, even under computation.
- Computing which realizes the 'Normally-Off'
- Key Technology
 - Non-Volatile Memory (MRAM, FeRAM, etc.)
 - Power Management
- Strategy:
 - not a simple combination of these technologies
 - Computing which exploits synergy of the technologies

Status of Volatile / Non-volatile RAM



Modified from the reference:
K. Ando, S. Fujita et al,
“Roles of Non-Volatile
Devices in Future Computer
System: Normally-off
Computer”, Energy-Aware
Systems and Networking for
Sustainable Initiatives, edited
by N. Kaabouch and W.-C.
Hu, published by IGI Global,
June, 2012

Courtesy of Dr. Shinobu
Fujita, Toshiba corp.

Goal of Normally-Off Computing

So Far:

Combinational logic

Volatile RAM

Always ON

Power-off area

Power off as much as possible

Temporally and spatially fine-grained power gating

Coarse-grained power gating

Combinational logic

Volatile RAM

long time for data save

Non-volatile Storage

Characteristics of NV-RAM

- Zero Stand-by Power 😊
- Slow speed ☹️
- Higher write Power ☹️

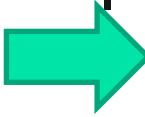
Combinational logic (Power Gating)

Non-volatile RAM

Ideal
Normally-Off



Outline of this presentation

- 
- Challenges of Normally-Off computing
 - Related Work: Low-power techniques
 - Fine-grained power gating processor
 - Overview and Current Status of
'Normally-Off computing Project'



Problem : Granularity

- Problem: Temporal Granularity

- Finer Granularity is preferable for Power Reduction

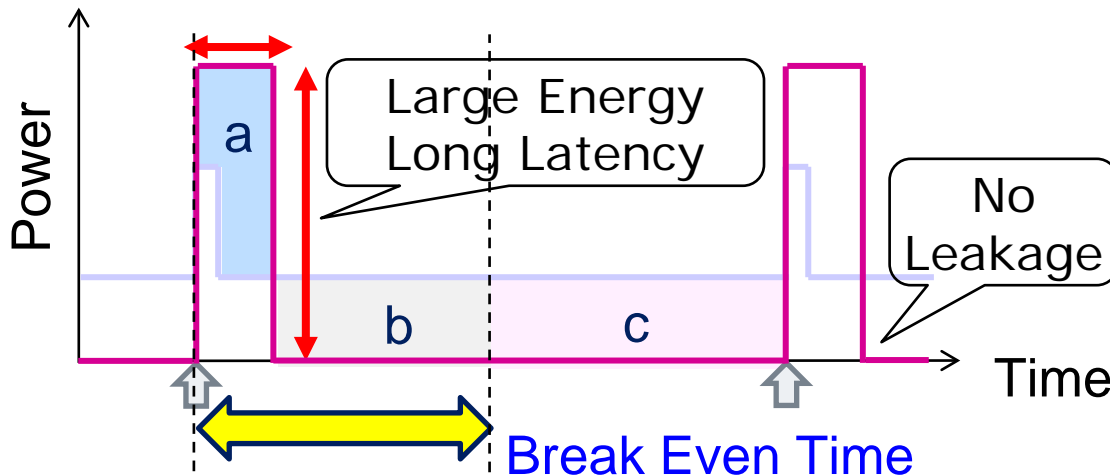
BUT,

- Too frequent power gating increases power consumption
 - Too frequent NV-RAM accesses lead to larger power consumption

Pitfall

- Replacing volatile RAM with NV-RAM always leads to power reduction \leftarrow *This is FALSE*
 - (Important) Access energy
Non-volatile RAM $>$ Volatile RAM
- Break Even Time of NV-RAM is important

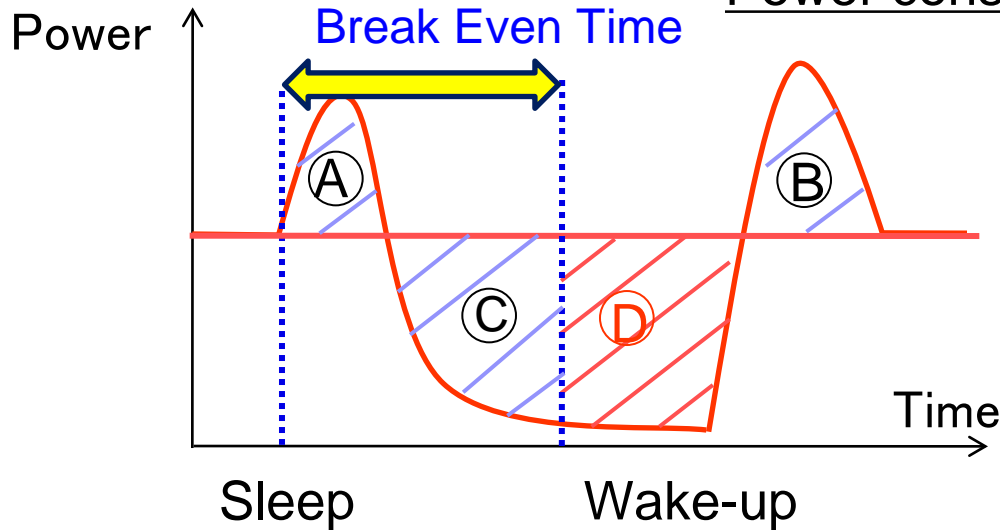
Power Consumption of Memory



a : extra access energy
b : reduced leakage energy
c : actual reduced energy
Break Even Time (BET)
Time when “a” = “b”

Temporal Granularity of Power Gating

Power consumption of power-gated (PG) logic



A+B: Energy overhead
C: Part of leakage saving
D: Net energy saving
Break Even Time (BET)
Time when $A+B=C$

Necessary condition for power reduction
Sleep period is longer than BET
→ ON/OFF oriented computing



Solution for Granularity Problem

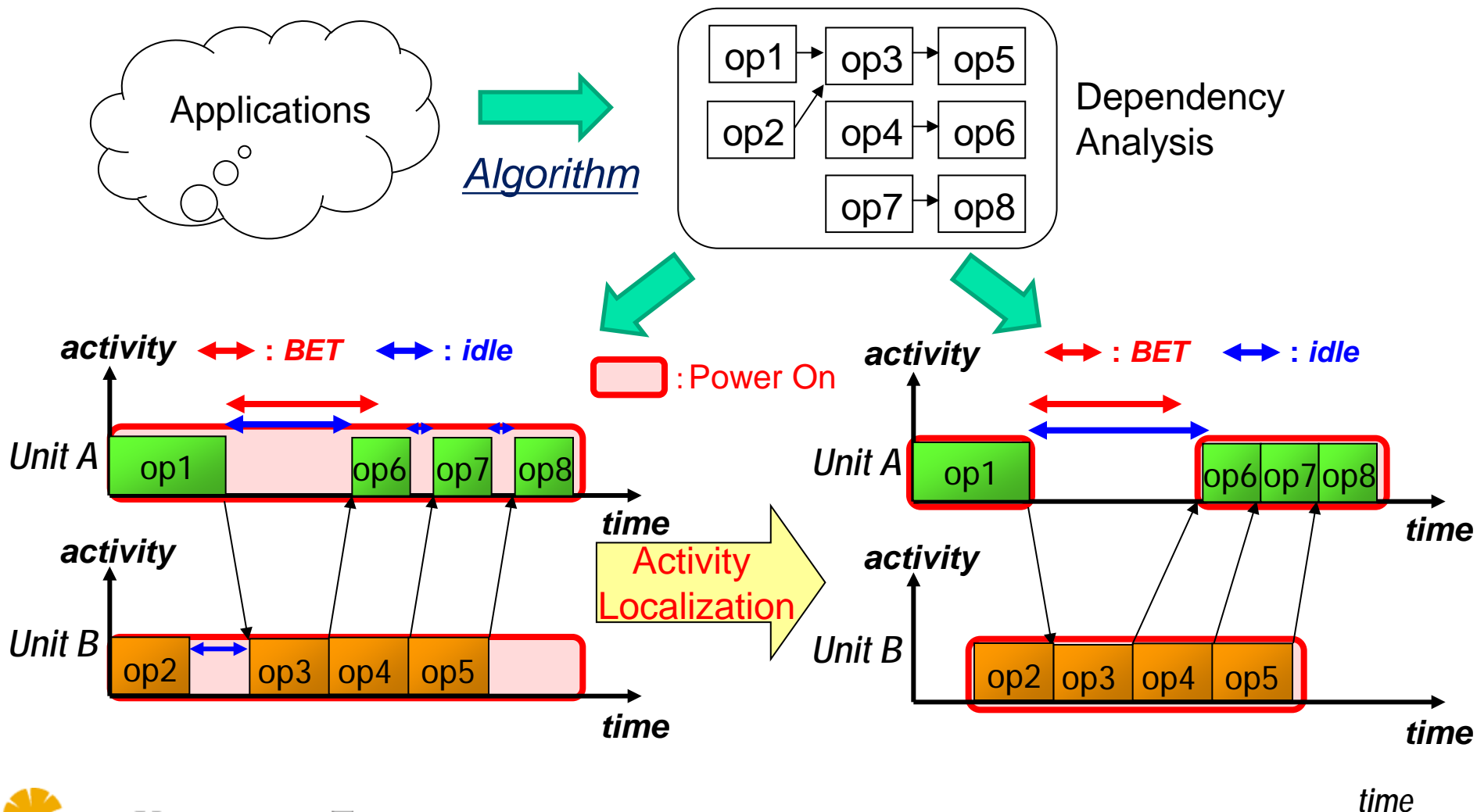
- How to Solve?

Co-Optimization of Memory and Computing

- Development of low access energy NV-RAM
→ shorter BET (Break Even Time)
- Normally-Off Oriented Computing
 - Architecture to prolong sleep state
 - Throttling: Run full speed & Sleep longer

N-OFF oriented computing

Collaboration between Computing and NV-Memory



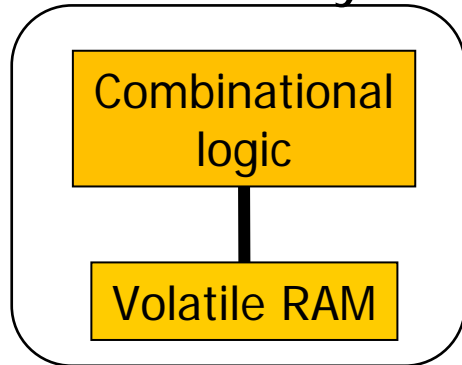


Outline of this presentation

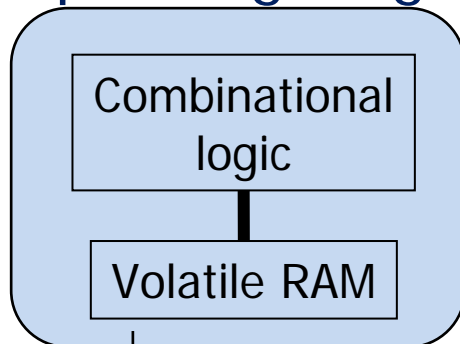
- Challenges of Normally-Off computing
- ➡ ■ Related Work: Low-power techniques
 - Fine-grained power gating processor
- Overview and Current Status of
'Normally-Off computing Project'

Goal of Normally-Off Computing

So Far: Always ON



Coarse-grained power gating



Long-term sleep

Non-volatile Storage

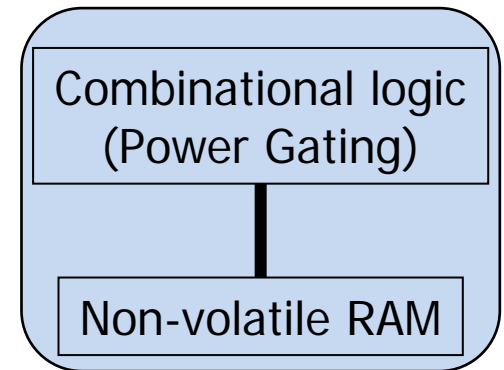
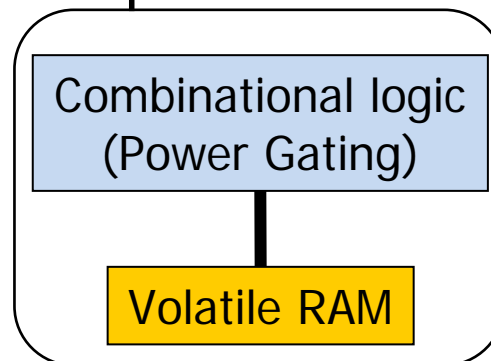
Power-off area

Power off as much as possible

Temporally and spatially fine-grained power gating



Related work:
Fine-grained PG
processor



Ideal
Normally-Off

Fine-grained Power Gating

- JST-CREST Project (2006.10-2012.3)
 - Innovative Power Control for Ultra Low-Power and High-Performance System LSIs

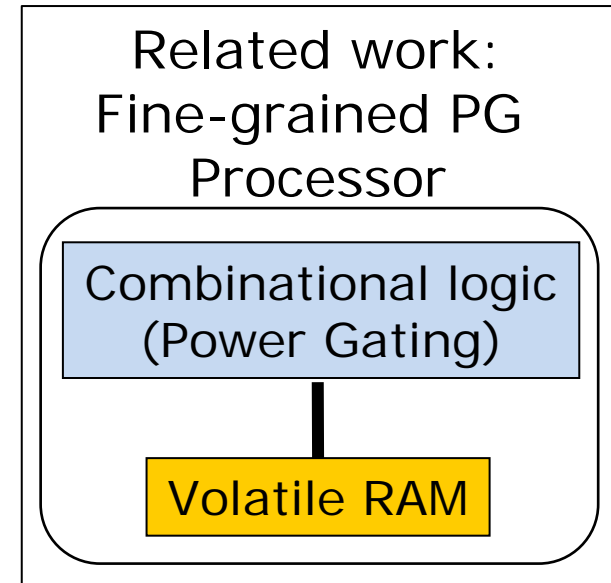
(JST : Japan Science and Technology Agency)

Leader: H. Nakamura (U. of Tokyo)

Co-PI: K.Usami, H.Amano, M.Kondo,
M.Namiki, T.Kuroda

(partly extended in)

- JSPS KAKENHI (S) (2013.4-2017.3)
(Japan Society for Promotion of Science)
 - Leader: H. Amano (Keio Univ.)
- http://www.am.ics.keio.ac.jp/kaken_s/eng



Granularity Issue for Leakage Power

■ Circuit technology

■ Power Gating

- Insert sleep transistor between target circuit and GND

■ Dynamic Body Biasing

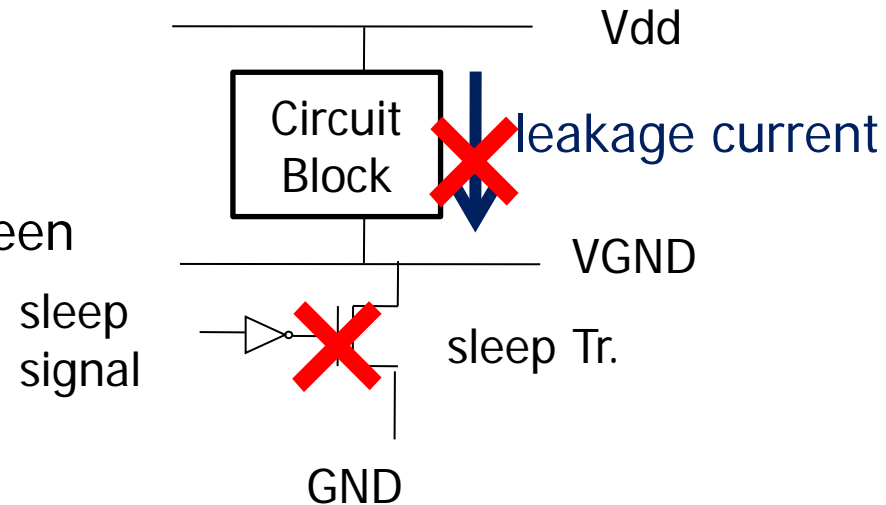
- Adjust Body Voltage:
Low V_{th} for operation, High V_{th} for idling

→ Coarse grain in time

→ Coarse grain in space

■ Hard to control from architecture..

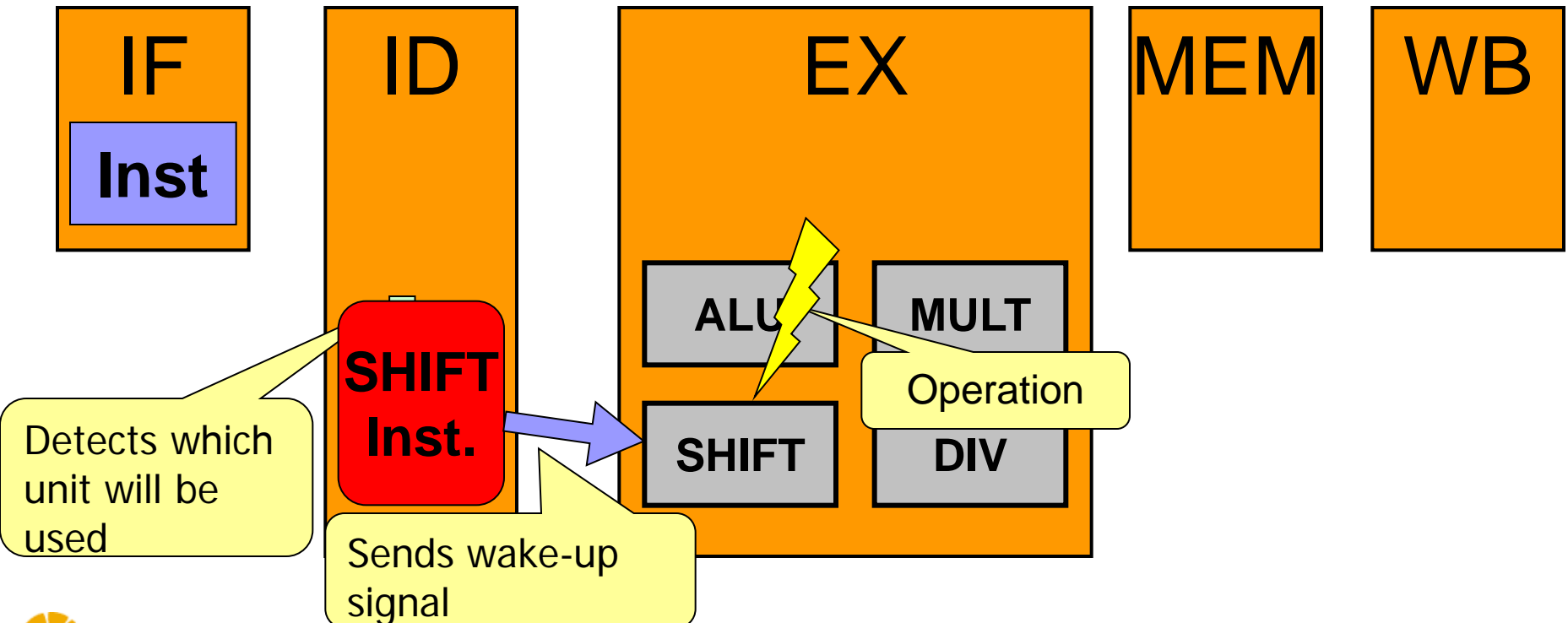
- Higher-density integration, higher frequency →
Distribution of idle parts: **Finer grain** in time and space



disparity in granularity

Geyser: Fine-grained power gating

- MIPS compatible processor with 5-stage pipeline
- Straightforward Fine Grain Run Time Power Gating
 - Turn EX-units into active mode only if necessary
 - Ex-unit gets active when an affecting instruction enters the IF stage
 - The activated EX-unit returns to sleep mode after execution



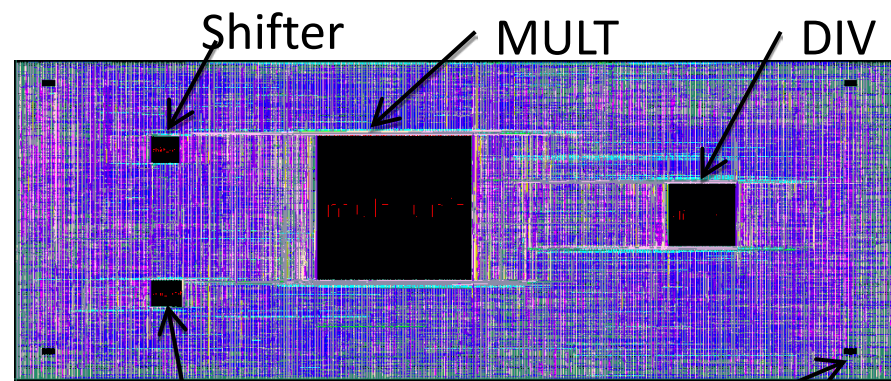
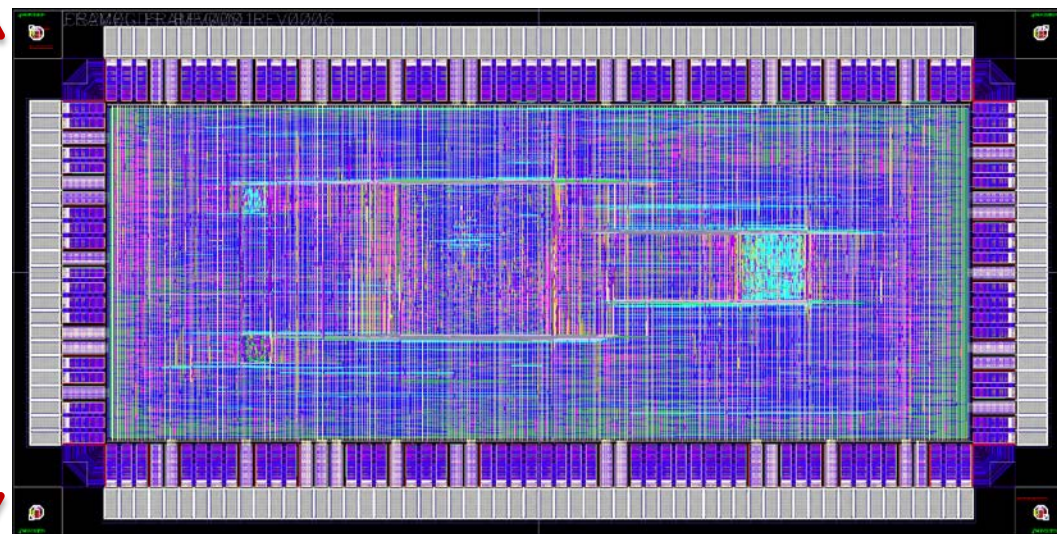
Prototype CPU : Geyser-1

[Ikebuchi et. al. ASSCC '09]

- MIPS R3000
 - Fujitsu e-shuttle 65nm
 - Vdd=1.2V
- successfully in operation
 - the first successful cycle by cycle power gating

2.1 mm

4.2 mm



ALU

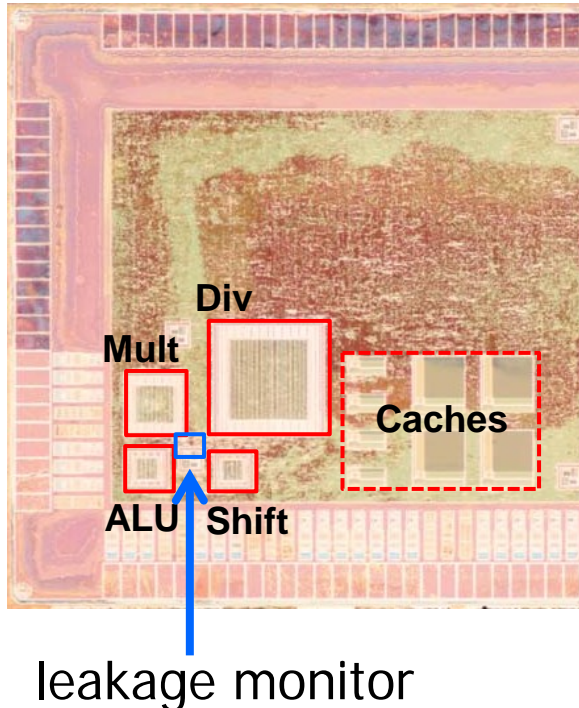
leakage monitor

ASP-DAC 2014

17



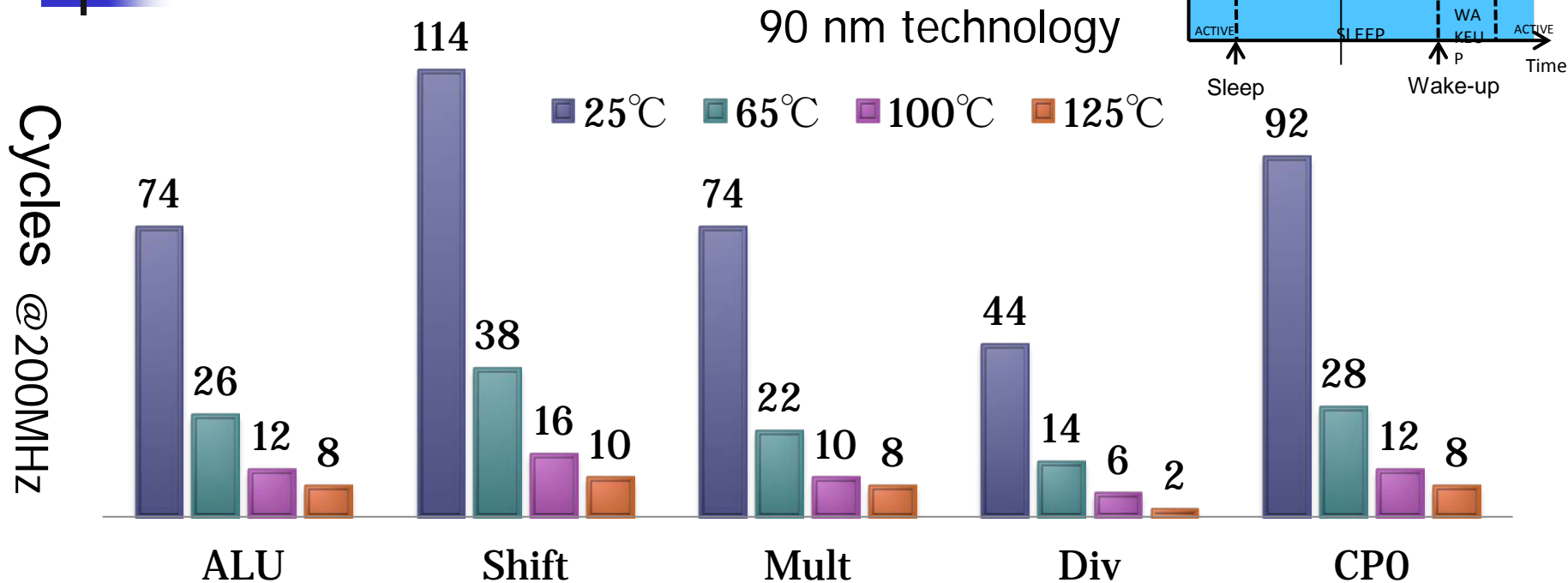
The latest Geyser Prototype



Process Technology	Fujitsu e-shuttle CMOS 65nm, 12 metal layers
Chip Area [um]	ALU: 121.4 x 113.4 Shift: 116.4 x 114.8 Mult: 199.4 x 199.4 Div: 369.0 x 368.6 Total: 1610 x 1443
Vdd	1.2 [V]
L1 cache	L1-I: 8KB, 64B-line, 2way L1-D: 8KB, 64B-line, 2way
Synthesis	Synopsys Design Compiler
Layout	Synopsys ICC UPF

- Linux successfully operates @ 190MHz

BET of Each Unit



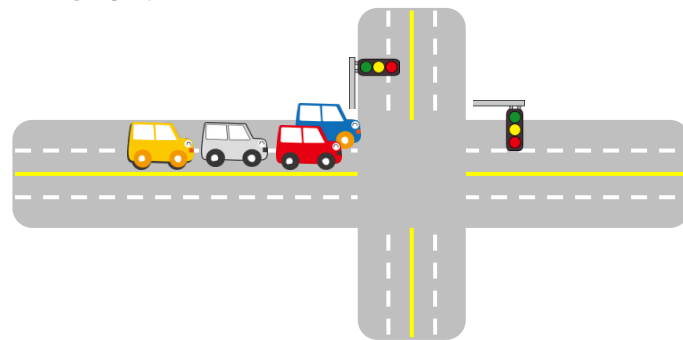
- BET is shortened when the chip temperature climbs up
 - Leakage current depends on temperature heavily
- We need Novel PG strategies taking BET into account

PG Strategies:

compared to idle reduction



- Idle reduction: very useful
 - Quick start (in a few seconds): less than one car
 - Power gating: 10~100 cycles
 - Restart penalty: 10~100 instructions stall...
 - Sophisticated strategy is required!
 - **Depend on a situation**
 - ☺: Signal has just turned to red
 - ☹: Traffic jam: Unpredictable
 - ☺?: Traffic jam: You can see 10 (100) cars ahead
- OS/Architecture know the situation, not circuit/device.





Adaptive Profile-based Power Gating

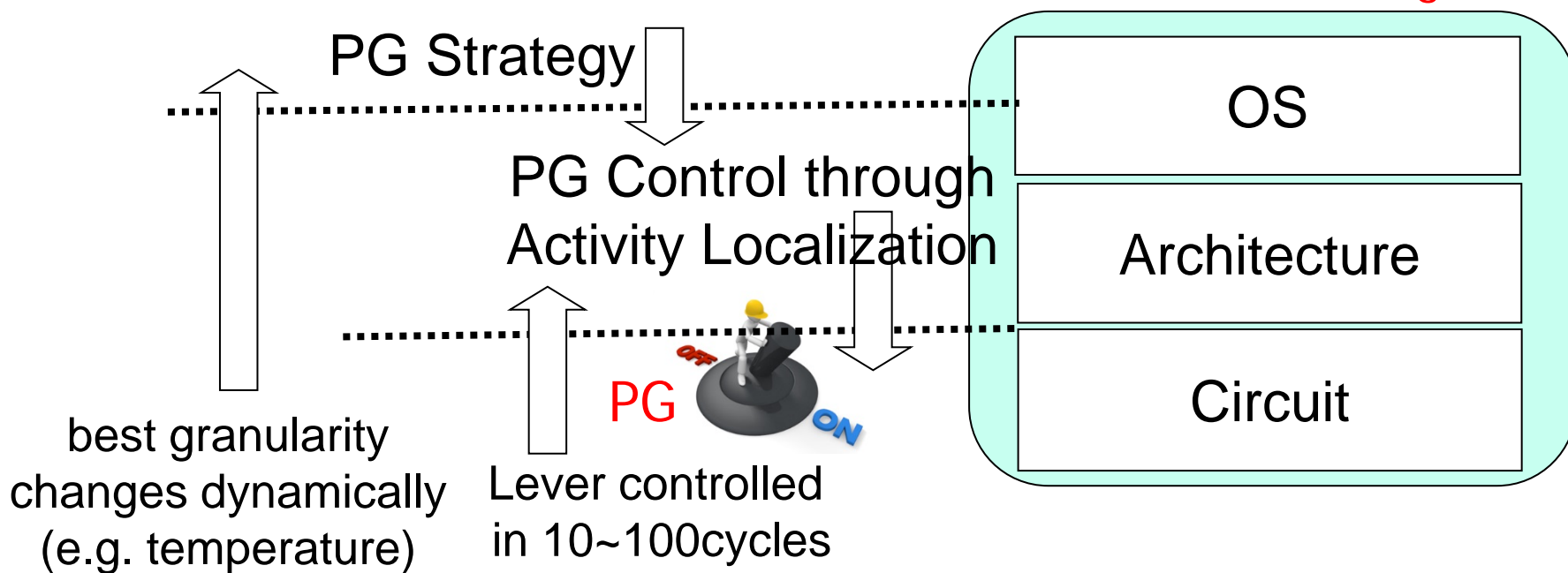
- Two power gating Policies for functional units
 - always put into sleep, if not used
 - do not sleep, even if not used
 - Profile-based:
 - Average Sleep Time of each functional unit by **code profiling**
 - Relationship between Temperature and BET by **on-chip leakage monitor**
 - By observing the operating temperature, OS selects the best policy
- details will be presented at 9C-1 on Thursday

Co-Optimization of Throttle Lever Control in Fine-grained PG



Innovative power control by cooperation with
Circuit, Architecture, System software

Control
granularity



- Who should be responsible for PG Control

→ depends on granularity of Control

- PG control granularity (BET) : 10 ~ 100 cycles
- best granularity of control changes every msec



Outline of this presentation

- Challenges of Normally-Off computing
- Related Work: Low-power techniques
 - Fine-grained power gating processor
- ➡ ■ Overview and Current Status of
'Normally-Off computing Project'

Recap: Goal of N-Off Computing

So Far:

Combinational logic

Volatile RAM

Always ON

Power-off area

Power off as much as possible

Temporally and spatially fine-grained power gating

Coarse-grained power gating

Combinational logic

Volatile RAM

long time for data save

Non-volatile Storage

Characteristics of NV-RAM

- Zero Stand-by Power 😊
- Slow speed ☹️
- Higher write Power ☹️

Combinational logic (Power Gating)

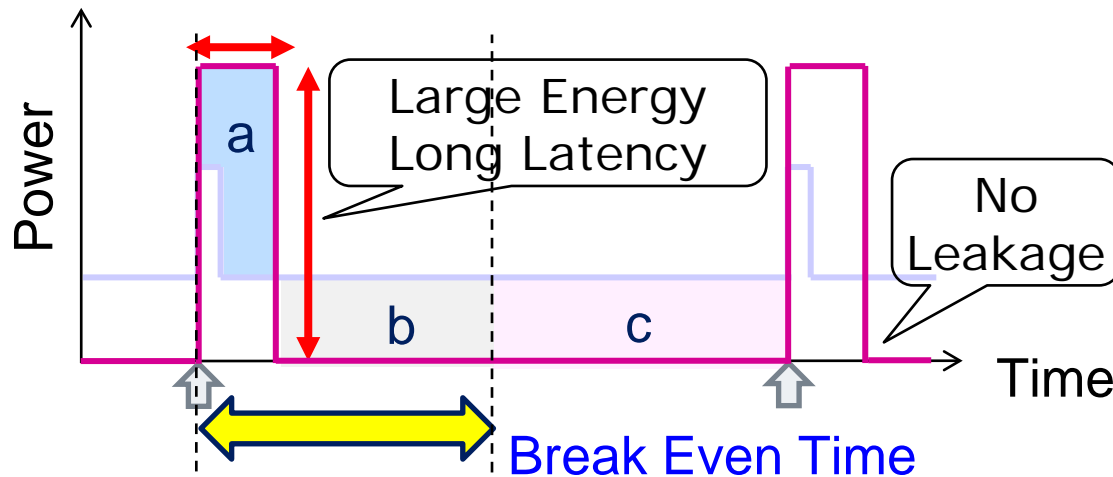
Non-volatile RAM

Ideal
Normally-Off

Challenge is “Granularity”

Recap: Break Even Time of Non-volatile RAM

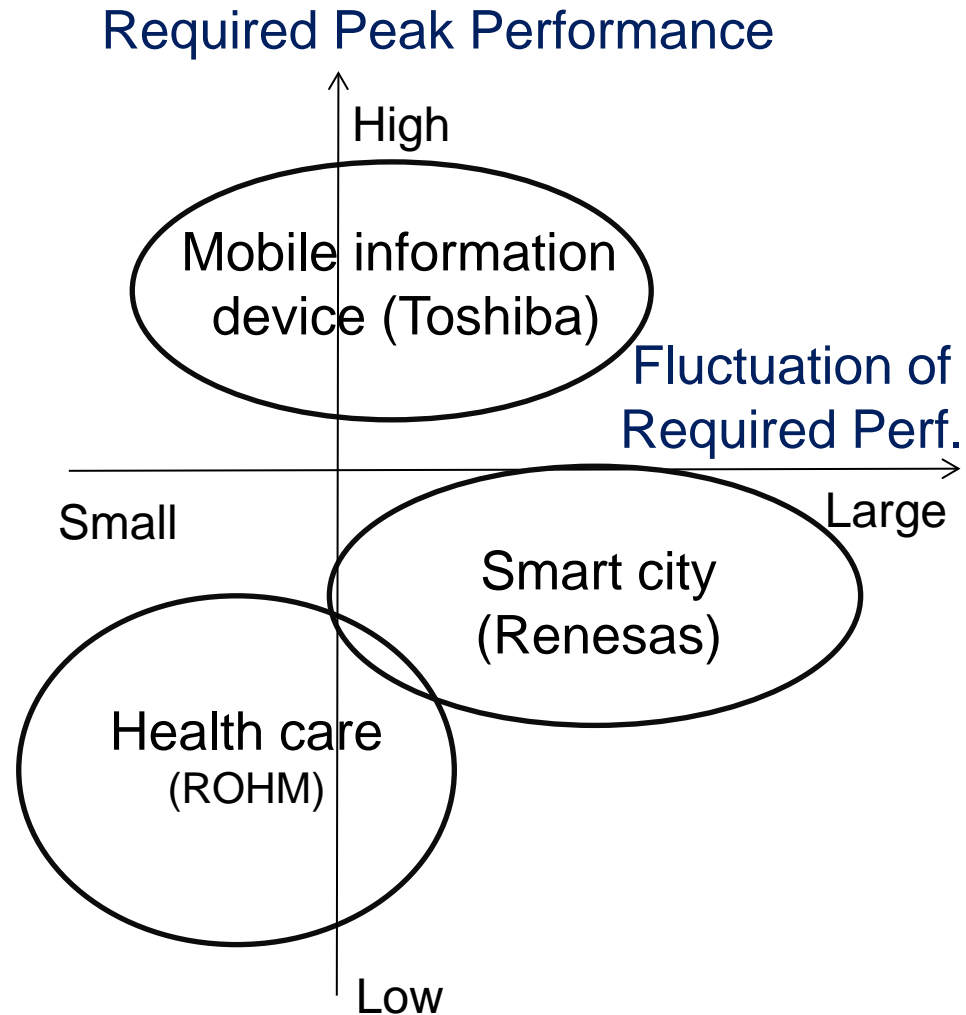
Power Consumption of Memory



a : extra access energy
b : reduced leakage energy
c : actual reduced energy
Break Even Time (BET)
Time when “a” = “b”

Target and Strategy of N-OFF Project

- Ideal N-OFF computing depends on applications
- Distributed Lab. :
Establish normally-off computing in each area of applications
→ Industries initiate and apply N-OFF computing to real society
- Central Lab. :
Generalize the above technologies
→ Collaboration of academia and industries
→ Establish N-OFF computing applicable to low power society in the future



NEDO Project Organization

Research Topic (2) "Research on technology to realize innovative normally-off computing for future sustainable social infrastructure"

Central Lab.

U-Tokyo, Renesas, Toshiba, Rohm

Distributed Lab

Topic (1)-1
Mobile
Device

Toshiba

Topic (1)-2
Smart
City

Renesas

Topic (1)-3
Health
Care

Rohm

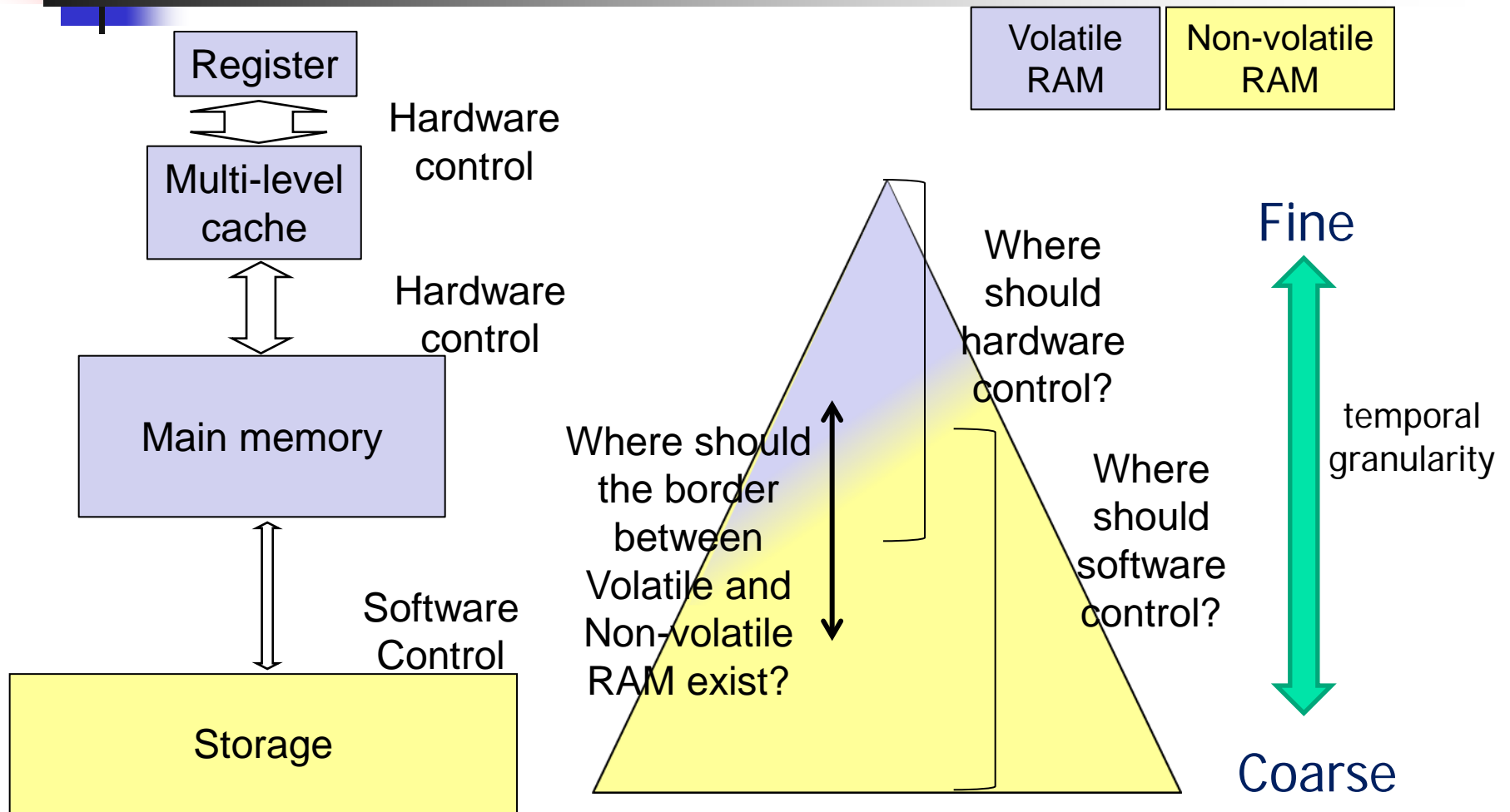
General Methodology
on N-OFF Computing

Application Specific
Leading-Edge
N-OFF Computing

Research Topic (1) "Development of power management techniques by using next generation non-volatile device"



Memory Hierarchy for Normally-Off



Conventional
memory hierarchy

memory hierarchy for
Normally-Off



Toshiba (Mobile Info. Device)

【Central Lab.】

Architectural methodology
for N-OFF processor by
making use of STT-MRAM

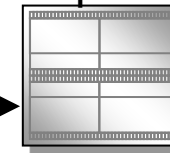
【Toshiba】 Development of novel ultra fast STT-MRAM

- More than 10 times
faster (SRAM comparable)
- Minimize active energy

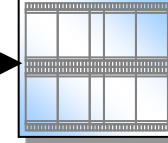
【Toshiba】 Development of
controller and bus interface for
normally-off cache

D-MRAM Cache: Enhancing Energy Efficiency with
3T-1MTJ DRAM/MRAM Hybrid Memory (DATE 2013)

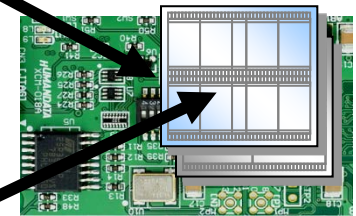
Normally-off
support processor



Magnetic
cache
memory



Mixed or 3D stack

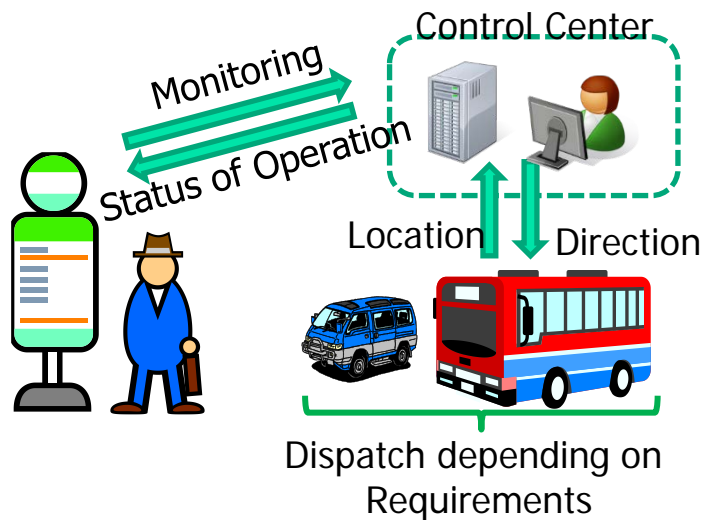


【Goal Product】
Normally-off processor

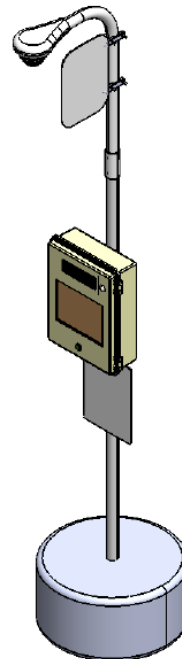
Compared to
conventional processor
**Power/Performance:
less than 1/10**

Renesas electronics (smart city)

- [Central Lab.]: Development of **Task-level scheduling**, Noff-API, as a part of **N-OFF software technology**
 - [Renesas]: Establish normally-off power management technology for sensor network system and demonstrate effectiveness and adaptability of normally-off micro computer system.
- Demand transportation system** with N-OFF sensors are under development as a demonstration with Future University Hakodate.



Demand transportation system
based on Intelligent bus stop

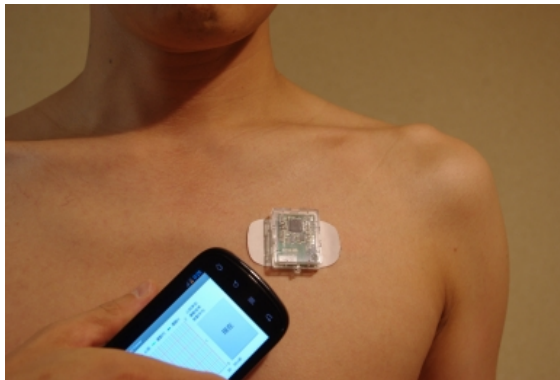


Intelligent Bus
Stop System
with N-OFF

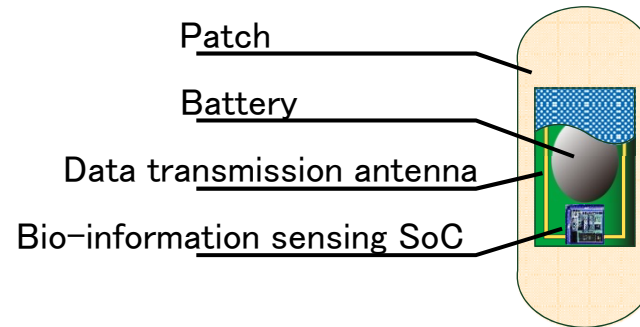
Rohm (Health care)

Rohm Co., + Omron healthcare Co., Kobe U.,

● 1st gen. Bio-information sensor



● Image of goal product



Measure	Heartbeat, 3-axis acceleration
Size	22mm x 30mm
Weight	About 4g (w/battery, w/o case)
Data Transmission	NFC (near field communication) (a.k.a. Wallet Mobile)

Low-power by
Normally-Off Computing

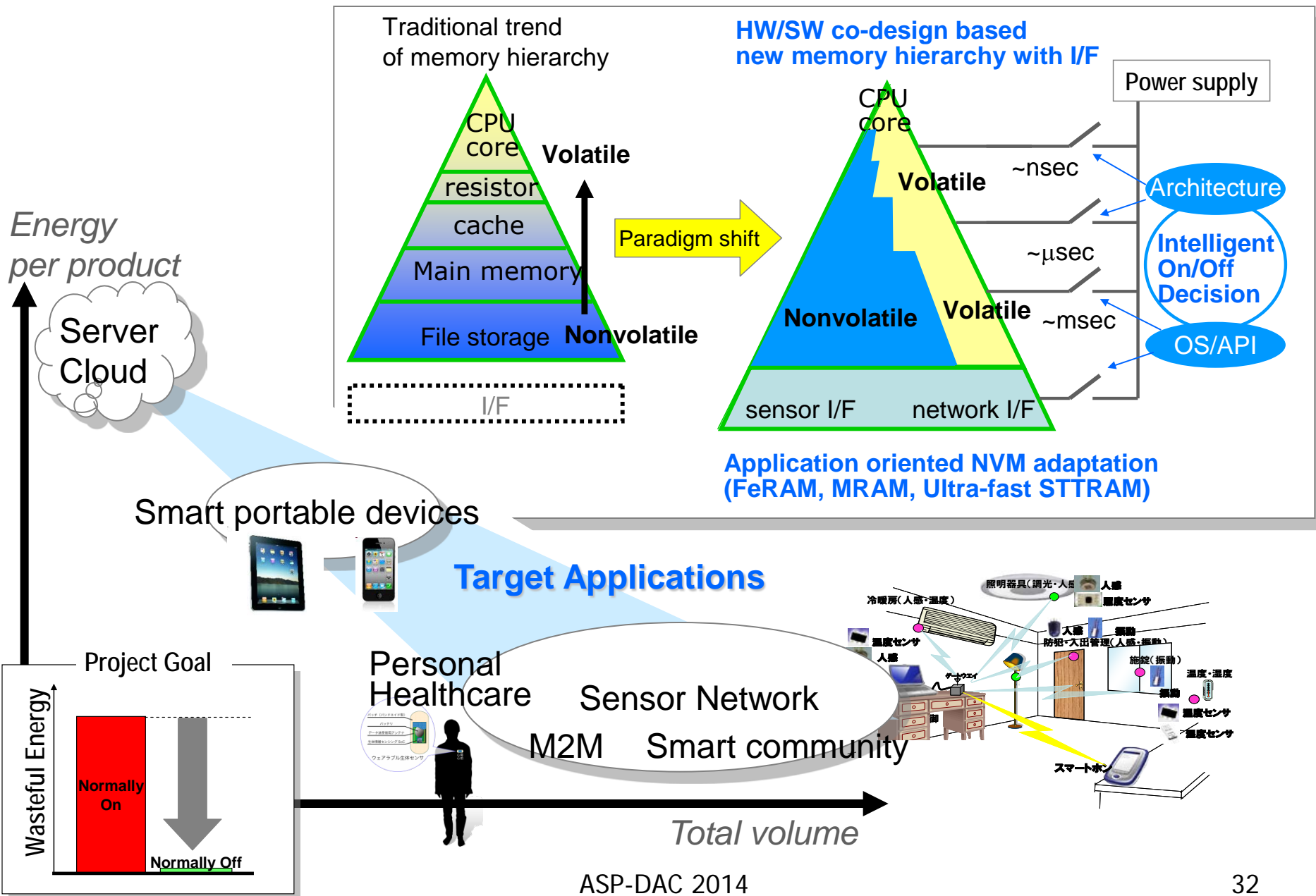


Realize wearable measurement
by light battery and device



Prevention of lifestyle disease

Normally-Off Computing Project (2011 ~ 2015)





Concluding Remarks: Challenges and Opportunities

- Opportunities

- Non-volatile memory: Potential is extremely high: fast, large capacity, and low power

- Challenges:

How to make use of this attractive memory?

→ Co-Optimization of Memory & Computing

- NV-RAM: BET is the most important
- Responsibility of Computing :
Optimize memory accesses dedicated for NV-RAMs

→ Realize Normally-Off Computing through
collaborative optimization of Application ~
Algorithm ~ OS ~ Architecture ~ Circuit ~ Device

