Nonvolatile Memory Technology for Future Computing

Latest Innovations: Device, process and system technologies

Jun. 27, 2019

Kazunari Ishimaru
Senior Fellow, IEEE Fellow

Institute of Memory Technology Research & Development
Toshiba Memory Corporation
Past : Digital Age

2005
The faithful gathered near St. Peter's Square at the Vatican, to witness Pope John Paul II's.

2013
St. Peter's Square at the Vatican, Pope Francis appearance on March 13, 2013.

Now : Smart Fab. (Yokkaichi Operation)

Promoting productivity improvement by using Big-data

Real-time Analysis/Control

Production M/C

Inspection M/C

Database

2Billion/day

Data analysis

Control

AI based analytical tools already introduced

Automated transport system
M/C requires precise control

Source : FMS Keynote 2018, Toshiba Memory
Info-plosion

Data generation exceeds 175ZB but stored <10ZB in 2025

Source: IDC’s Data Age 2025, April 2017

Where Data is Stored
[EB]

Source: IDC Global Datasphere, April 2017

Generated Data [ZB]

175 ZB

Connected Devices >50B@2020

CAGR +31%

Real-Time Data

Stored

Not Stored

0 100 180


Source: IDC’s Data Age 2025, April 2017

© 2019 Toshiba Memory Corporation
Issue: Energy Efficiency

Electricity usage of Data Center increases exponentially. Energy efficiency improvement of system is crucial.

Electricity usage (TWh) of Data Centers 2015-2025

- Typical Case 2021→2025 x4.2
- Best Case 2021→2025 x4.2

Traffic (ZB/y)

- 2015
- 2017
- 2019
- 2021
- 2023
- 2025

Source: 2017 paper “Total consumer power consumption forecast” by Anders Andrae

Power consumption by components (server)

- CPU: 61.0%
- DRAM: 18.0%
- STORAGE/Disk: 11.7%
- NETWORKING: 10.0%
- MISC: 2.0%

Consumption in France: 475TWh (2017)

System Bottleneck

Current system is not energy efficient, because…

What Processors are doing?

CPU time
Genomics: ~95%
Language Processing: ~80%

Talking to memories!

“Computing’s Energy Problem (and what we can do about it)”, M. Horowitz, ISSCC 2014
Requirement for Deep Learning (Inference)

Cloud, Fog, Edge require energy efficient system
Computing Systems

Each system architecture has trade-off between Efficiency and Flexibility

**von Neumann**
- PE
- Between PKG
- Memory

**Near Memory**
- PE
- in PKG
- Memory

**In Memory**
- PE + Memory
- Merger

Efficiency (Energy/Speed)

Flexibility (Memory size/Application)
System requirements

Memory space expansion, Storage latency reduction are required
Storage expansion: BiCS FLASH™ with QLC

Density trend continues by BiCS FLASH™ with QLC technology

Source: 13.1 ISSCC 2019, Toshiba Memory
Latency reduction: XL-FLASH™

BiCS FLASH™ based Low Latency SLC device (scalable)

Good for random IOPS and Better QoS at shallow QD in SSD

Source: FMS Keynote 2018, Toshiba Memory
BiCS FLASH™ with TSV Technology

Higher Data Rate
>1Gbps with 16 die/ch

Higher Density
1TB / package

Lower Power Consumption
~45% Power Reduction

Press release July 11, 2017

TSV : Through Silicon Via
Storage Class Memory

SCM will fill the “gap” between DRAM and NAND/SSD in the storage hierarchy
There are multiple ways to utilize SCM in a system
SCM devices

So far PCM is the only device on the market. Others may follow.

128Gbit 3D Xpoint™ (PCM)


32Gbit ReRAM

Figure 12.1.7: Die micrograph, cross-sectional view, and features. ISSCC 2013

4Gbit STT-MRAM

ISSCC 2017
System improvement: SCM/XL-FLASH™

SCM and XL-FLASH™+QLC SSD system expands memory/storage space and improves latency.
Proposed architecture

Each system architecture has trade-off between Efficiency and Flexibility

- **von Neumann**
  - Processor
  - Memory
  - Between PKG

- **Near Memory**
  - Processor
  - Memory
  - in PKG

- **In Memory**
  - Processor + Memory
  - Merger

Efficiency (Energy/Speed)

Flexibility (Memory size/Application)
Issue: In-memory type

- Utilization of crossbar array is only 7.4%
- Energy efficiency degrade ~x13
- SW/HW desired which applicable to all neural network

Source: 5.1 ASSCC 2018, Toshiba Memory
Experimental result ~execution cycles~

ResNet-50 ImageNet

×10^6

# Execution cycles

Source: 5.1 ASSCC 2018, Toshiba Memory
Work together

- Apps
- Algorithm
- SW to map DNNs onto HW

- Memory
- DNN accelerator
- GB/s

Application
Algorithm
System/SW

Architecture
Circuit
Device
Material/Process

Co-work/co-optimization is important
Conclusion

- In Info-plosion era, there are strong demands for storage and energy efficient system.
- BiCS FLASH™ is a key component and continuously grow GB/area benefit.
- XL-FLASH™+QLC-SSD improves system performance over the DRAM+HDD.
- PAM4 Multiplexing improves I/F BW with low power.
- Storage Class Memory is required to improve system performance.
- AI system needs SW/HW cooperation. Memory is a key component and BiCS FLASH™ based system can support.
Thank you!

TOSHIBA

BiCS FLASH™

All other company names, product names, and service names mentioned herein may be trademarks of their respective companies.