Solid-State Drive (SSD) and Memory System Innovation

Ken Takeuchi
Dept. of Electrical Engineering and Information Systems
University of Tokyo
E-mail: takeuchi@lsi.t.u-tokyo.ac.jp
http://www.ls1.t.u-tokyo.ac.jp
Definition of SSD

- SSD: Solid State Drive
  Mass storage to replace HDD of PC/automobile application.
- SSD consists of NAND Flash Memory and NAND controller (+RAM)

All car manufactures are interested in SSD.

J. Elliott, WinHEC 2007, SS-S499b_WH07.
Reliability bottleneck for automobile application

Reliable Circuit

Reliable Device

Reliability

Reliable OS/Computer architecture

Need collaboration with IT/SW community
Key Challenge of SSD

- Need to improve **device** reliability such as endurance, data retention, and disturb.

- Require **co-design of NAND and NAND controller circuits** to best optimize both NAND and NAND controllers.

- **OS/Computer architecture innovation** essential.

---

Outline

- NAND Overview
- SSD Overview
- Operating System for SSD
- Green IT with SSD
- Summary
Outline

- NAND Overview
- SSD Overview
- Operating System for SSD
- Green IT with SSD
- Summary
NAND Flash Memory

43nm 16Gb NAND

NAND flash memory chip

Memory circuit

Memory cell: Floating Gate-FET
Memory cells are sandwiched by select gates.

Contactless structure: ideal $4F^2$ cell size

F. Masuoka, IEDM 1987, pp.552-555.
Top View of NAND Flash Cell Array

Simple structure: High scalability, High yield

MLC vs. SLC

- **SLC**: Single-level cell or 1bit/cell
- **MLC**: Multi-level cell or >2bit/cell
  - 2bit/cell: Long production record since 2001
  - 3bit/cell or 4bit/cell: Will be commercialized this year.
- Existing SSD uses SLC but MLC based SSD will be commercialized this year.

---

**SLC (Single-level cell)**

Number of memory cells

- "0"
- "1"

Vth

**MLC (Multi-level cell)**

Number of memory cells

- "0"
- "1"
- "2"
- "3"
NAND Density Trend

55% growth / year

ISSCC paper
MLC (Multi-level cell) NAND flash
SLC (Single-level cell) NAND flash

MLC performance is comparable with SLC.

Chip Architecture

- 56nm 8Gbit NAND Flash Memory

- 56nm CMOS process
- Cell size: 0.0075um²/bit
- Chip size: 98.8mm²
- Architecture
  - NAND string: 32 cells
  - Two 4Gb memory planes
  - # of blocks / plane: 1K
  - # of pages / block: 128
  - Min. page size: 4KB
  - Min. block size: 512KB

NAND Operation Principle

Read

After precharging, bit-lines are discharged through the memory cell.

- Unselected cells are biased to the pass voltage, Vread.
- Small cell read current (~1uA) → Slow random access (~50us)
- Serial access: 30-50ns → Fast read = 20-30MB/sec
NAND Operation Principle (Cont’)

■ Program : Electron injection

- 18V
- 0V
- 0V
- Channel-FN tunneling
- High reliability
- Low current consumption (~pA/cell)
- Page based parallel program
  Typical page size : 2-4kB

■ Erase : Electron ejection

- 20V
- 0V
- 20V

S. Aritome, IEDM 1990, pp.111-114.
Outline

- NAND Overview
- SSD Overview
- Operating System for SSD
- Green IT with SSD
- Summary
PC/automobile expected as the next killer application
Cost Trend of NAND and HDD

Analyst expectation

NAND will replace 2.5” HDD in 2010-2012 if the cost continues decreasing.

Unclear scaling scenario e.g. double exposure vs. EUV, floating gate vs. MONOS, and 2D vs. 3D cell.
SSD Reliability

- SSD is robust.
  - No mechanical parts.
- Need to be careful in PC/automobile application
  - Portable consumer electronics application
    (Digital still cameras, MP3 players, Camcorders)
    - Effective data retention time << 10 years
    - Data quickly transferred to PC or DVD through USB drive and memory cards.
    - Most probably data backup in PC
  - PC/automobile application
    - Higher reliability required w.o. backup
    - Need longer data retention time : 5-10 years
SSD Reliability (Cont’)

Failure mechanism of NAND

Program disturb
During programming, electrons are injected to unselected memory cells.

Read disturb
During read, electrons are injected to unselected memory cells.

Write/Erase endurance & Data retention
As the Write/Erase cycles increase, damage of the tunnel oxide causes a leakage of stored charge.
SSD Reliability (Cont’)

Program disturb

Both selected and unselected cells suffer from the disturb.

SSD Reliability (Cont’)

- Read disturb

- Weak program bias condition
- Unselected word-lines suffer from the read disturb.
Program disturb and read disturb summary

Program disturb and read disturb is a “bit error” not a “burst error”.
- Two bits in MLC are assigned to different pages.
- Even if one MLC cell fails, one bit in two pages fails.

ECC (Error correcting code) effectively corrects the bit error.
- Existing ECC corrects 8-12 bit errors per 512 Byte sector.

Page assignment of MLC

SSD Reliability (Cont’)

Write/Erase Endurance & Data Retention

- Endurance: how many times data are written
- Data retention: how long the data remains valid
- Clear correlation between endurance and data retention

Damages to the tunnel oxide during write and erase cause the data retention problems.
- Traps are generated during write and erase.
- The unlucky cell with traps results in a leakage path, causing the charge transfer.
- The leakage current is called SILC (Stress Induced Leakage Current).

To guarantee data retention, Write/Erase cycles are limited to 100K (SLC) or 10K (MLC).

K. Prall, NVSMW 2007, pp. 5-10.
SSD Reliability (Cont’)

100K (SLC) or 10K (MLC) W/E cycles acceptable?

W/E cycles estimation for PC

- 32GB SSD
- Usage scenario: 2~5GB/day (#)
- Service for 5 years
- 100% efficient wear leveling

\[(365 \text{ days/year}) \times 5 \text{ years} / (32GB / 2\sim5GB/\text{day})\]

= 114~285 W/E cycles

114~285 cycles are far below the NAND limitation of 100K for SLC or 10K for MLC.

Actual W/E cycles are higher for the file management such as garbage collection.

(#) W.Akin, IDF 2007_4, MEMS003.
Y.Kim, Flash Memory Summit 2007.
High Reliability Technology

Wear-leveling

Problem

Write/Erase cycle of NAND is limited to 100K for SLC and 10K for MLC.

Solution

- Write data to be evenly distributed over the entire storage.
- Count # of Write/Erase cycles of each NAND block.
- Based on the Write/Erase count, NAND controller re-map the logical address to the different physical address.
- Wear-leveling is done by the NAND controller (FTL), not by the host system.

![Bitline Diagram](image-url)
Example of wear-leveling

- If the block is occupied with old data, data is programmed to a new block.
- If there is no free block, the invalid block are erased.

![Diagram showing wear-leveling process]

- Old file
  - Block 1
  - Block 2
  - Block 3
  - Block 4
  - Block 5
  - Block 6
  - Block 7
  - Block 8
  - Block 9

- New File
  - Block 4 ➔ Invalid
  - Block 1
  - Block 2
  - Block 3
  - Block 4
  - Block 5
  - Block 6
  - Block 7
  - Block 8
  - Block 9

- Empty block
- Rewrite old file
- Write new file to an empty block
High Reliability Technology (Cont’)

- **Static data**
  Data that does not change such as system data (OS, application SW).

- **Dynamic data**
  Data that are rewritten often such as user data.

- **Dynamic wear-leveling**
  Wear-level only over empty and dynamic data.

- **Static wear-leveling**
  Wear-level over all data including static data.
Dynamic wear-leveling

- Block with static data is NOT used for wear-leveling.
- Write and erase concentrate on the dynamic data block.

Red: Static data such as system data.
Blue: Dynamic data such as user data

N.Balan, MEMCON2007.
SiliconSystems, SSWP02.
High Reliability Technology (Cont’)

Static wear-leveling

- **Write/Erase count**
  - Red: Static data such as system data.
  - Blue: Dynamic data such as user data

- Physical block address

- **Wear-level more effectively than dynamic wear-leveling.**
- Search for the least used physical block and write the data to the location. If that location is empty, the write occurs normally.
- Contains static data, the static data moves to a heavily used block and then the new data is written.

High Reliability Technology (Cont’)

- OS support required for effective wear-leveling
- SSD does NOT know the contents of the file.

Write/Erase count

Red: Static data such as system data.
Blue: Dynamic data such as user data

Physical block address

0 1 2 3 4 5 6 7 8 9 10 11 12 - - N

N.Balan, MEMCON2007.
SiliconSystems, SSWP02.
Random access

<table>
<thead>
<tr>
<th></th>
<th>Read</th>
<th>Write</th>
<th>Erase</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAND (SLC)</td>
<td>25us</td>
<td>300us</td>
<td>1ms</td>
</tr>
<tr>
<td>NAND (MLC)</td>
<td>50us</td>
<td>800us</td>
<td>1ms</td>
</tr>
<tr>
<td>HDD</td>
<td>3ms</td>
<td>3ms</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

Erase are hidden by operating the erase during the idle period.

- **Read**: SSD with SLC and MLD has a great advantage over HDD.
- **Write**: SSD still has a performance advantage. Write performance can be an issue in the future if the NAND performance degrades by scaling the memory cell or increasing the number of bits per cell.
SSD Performance (Cont’)

### Sequential access

<table>
<thead>
<tr>
<th></th>
<th>NAND : Single chip operation</th>
<th>NAND : 4 chip interleaving</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read</td>
<td>25MB/sec</td>
<td>100MB/sec</td>
</tr>
<tr>
<td>Write</td>
<td>20MB/sec</td>
<td>80MB/sec</td>
</tr>
<tr>
<td>NAND (SLC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read</td>
<td>20MB/sec</td>
<td>80MB/sec</td>
</tr>
<tr>
<td>Write</td>
<td>10MB/sec</td>
<td>40MB/sec</td>
</tr>
<tr>
<td>NAND (MLC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read</td>
<td>80MB/sec</td>
<td>-</td>
</tr>
<tr>
<td>Write</td>
<td>80MB/sec</td>
<td>-</td>
</tr>
<tr>
<td>HDD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Block diagram of SSD w. interleaving function]

- **SSD (SLC)**: Comparable read and write performance with HDD.
- **SSD (MLC)**: Comparable read performance. By introducing 8chip interleaving, the write performance can be comparable with HDD.

SSD Power Consumption

Power consumption

<table>
<thead>
<tr>
<th></th>
<th>NAND : Single chip operation</th>
<th>NAND : 4 chip interleaving</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Read</td>
<td>Write</td>
</tr>
<tr>
<td>NAND (SLC)</td>
<td>20mA</td>
<td>20mA</td>
</tr>
<tr>
<td>NAND (MLC)</td>
<td>20mA</td>
<td>20mA</td>
</tr>
<tr>
<td>HDD</td>
<td>&gt;300mA</td>
<td>&gt;300mA</td>
</tr>
</tbody>
</table>

In SSD, additional current (~100mA) are consumed in the NAND controller, RAM and IO.

Actual Power Consumption

<table>
<thead>
<tr>
<th></th>
<th>Active</th>
<th>Idle</th>
<th>Standby</th>
<th>Sleep</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDD</td>
<td>360mA</td>
<td>83mA</td>
<td>38mA</td>
<td>19mA</td>
</tr>
<tr>
<td>NSSD</td>
<td>152mA</td>
<td>52mA</td>
<td>17mA</td>
<td>15mA</td>
</tr>
</tbody>
</table>

In all modes, the power consumption of SSD is smaller than HDD.

Outline

- NAND Overview
- SSD Overview
- Operating System for SSD
- Green IT with SSD
- Summary
Operating System for SSD

- SLC/MLC Combo

- Performance Optimization
  - Sector Size Optimization

- Reliability Optimization
  - EWF (Enhanced Write Filter)
  - SMART (Self-Monitoring, Analysis and Reporting Technology)
Why OS?

Motivation

- Existing OS is optimized for HDD.
- Current SSD based computers uses the conventional OS and just replace HDD with SSD.
- To achieve the best performance and reliability of SSD, OS especially file system should be optimized.
New Memory System: NAND/HDD Combo

- **NAND as a cache**
  - Intel Robson
  - Microsoft Ready Boost

- **Multi-drive of NAND/HDD**
  - SanDisk Vaulter Disk
  - NAND: OS data
  - HDD: User data

Temporary solution until NAND cost becomes comparable with HDD cost.


New Memory System: MLC/SLC Combo

Future Direction: Hybrid SSD with SLC and MLC

- Concept: Right device for the right use.
- Enjoy the Benefit of both SLC and MLC.
- SLC: Fast and highly reliable but low capacity. Use SLC as a cache or system data storage.
- MLC: Large capacity but slow. Use MLC as user data storage.
- OS support essential: SSD does NOT know the contents of the file.


Toshiba LBA-NAND
http://www1.toshiba.com/taec/index.jsp

Spansion MirroBit Eclipse
http://www.spansion.com/products/MirrorBit_Eclipse.html

MLC (Multi Level Cell)

Combi (SLC+MLC)

SLC (Single Level Cell)

R/W Speed: 2006 2007 2008 2009 2010

57/32 64/45 100/80 160/160 800/800 1300/1300
Performance Optimization

**Sector size optimization**

- Minimum write/read unit of NAND is a page.
- Typical page size is 4-8KByte.
- A page is written only ONCE to avoid the program disturbance.
- With current OS having 512Byte sector, one sector write wastes >80% of data in a page.

- LBD (Long Block Data) sector standard (Windows Vista): 4KByte sector size fits better with SSD.

- As the page size increases as NAND is shrinking, larger sector size such as 64KByte or 128KByte is required.
Enhanced Write Filter (Windows Embedded)

- Decrease write/erase cycles of NAND, extending the NAND lifetime.
- Control the file allocation to store frequently rewritten file in DRAM and not to access NAND.
- Enhanced Write Filter (EWF) is located between file system and low level driver interfacing with SSD.
- OS/Application SW support essential: Again, SSD does NOT know the contents of the file.

Reliability Optimization (Cont’)

SMART
(Self-Monitoring, Analysis and Reporting Technology)

- Monitor the storage and report/predict the failure.
- SMART for HDD is NOT smart because it is very difficult to predict the mechanical failure.
- SMART for SSD can be really smart.
- Product lifetime can be predicted because the failure rate is highly correlated with the write/erase cycles.
- Predict the SSD lifetime by monitoring the write/erase cycles and replace SSD before the fatal failure occurs.

http://www.tdk.co.jp/tefe02/ew_007.pdf
Outline

- NAND Overview
- SSD Overview
- Operating System for SSD
- Green IT with SSD
- Summary
Green IT: Power Crisis of Data Center

- Data through internet is increasing drastically.
- In the U.S., power consumption at the data center doubled during the last 5 years. (5 nuclear power plants!)
- In 2025, the data increases by 200 times and the power consumption increases by 12 times.

Ken Takeuchi
Univ. of Tokyo - INRIA - Ecole des Mines Paris - INRETS Joint Symposium 2008.7.7

Data Center

Power increase of HDD
Replace HDD with SSD

SSD (NAND Flash)
Problems of SSD

- Low capacity
  - Currently 2GByte, >TByte required

- Low write/erase cycles
  - Currently 10K cycles, >100K required

- Power increase because of scaling

---

Scaling limit of NAND

NAND will face a scaling limit around 10-20nm.
Scaling limit of NAND (Cont’)

Reduction of electrons in a floating-gate

![Graph showing number of electrons vs. design rule (gate length)]

- Stored electrons @ $\Delta V_{th}=4.0\,\text{V}$
- Charge loss tolerance @ $\Delta V_{th}=0.2\,\text{V}$

Enhanced capacitive-coupling between memory cells

- FG-FG coupling shifts the Vth of a memory cell as the neighboring cell are programmed.
- The Vth modulation becomes significant as the memory cell is scaled down.

J.D. Lee, EDL 2002, pp. 264-266.
M. Ichige, Symp. on VLSI Technologies 2003, pp.89-90.
Fe(Ferroelectric)-NAND Flash Memory

- NAND Flash Memory w. Ferroelectric Transistor
  - Scalable below 20nm
  - Low voltage/power operation: 20V → 5V
  - Write/Erase cycles: 10K cycles → 100M cycles
  → Most suitable for data server application

MFIS Structure
(Metal-Ferroelectric-\underline{Insulator-Semiconductor})

S. Sakai, NVSMW 2008, pp.103-104.
Operation Principle of Fe-NAND Flash

- Low voltage operation

![Diagram of Fe-NAND Flash operation](image)

Ken Takeuchi
Univ. of Tokyo - INRIA - Ecole des Mines Paris - INRETS Joint Symposium 2008.7.7

S. Sakai, NVSMW 2008, pp.103-104.
Scalable below 20nm

Ferro electricity is maintained in 20nm size.

SrBi$_2$Ta$_2$O$_9$ (SBT)

 TEM Photograph

a = 0.5506 nm  
b = 0.5534 nm  
c = 2.498 nm

SrBi$_2$Ta$_2$O$_9$  
$\sim 400$nm

Hf-Al-O  
$\sim 10$nm

IL

Si

IL: Interfacial layer  
major component – SiO$_2$

S. Sakai, NVSMW 2008, pp.103-104.
10-year Data Retention

On states

Off states

Drain Current, $I_d$ (A)

Time, $t$ (s)

1st
2nd
3rd
4th

10 years
37.0 days
33.5 days

MFIS-Type FeFET

Buffer layer improved Si interface.

S. Sakai, NVSMW 2008, pp.103-104.
Excellent W/E Cycles up to 100M

Fe-NAND

NAND

S. Sakai, NVSMW 2008, pp.103-104.

SSD will replace HDD in PC/automobile in 2010-2015 if reliability bottleneck is solved.

Device, circuit and OS innovation essential to realize a highly reliable, low power, high speed and high capacity SSD.

SSD expected to save power at a data center.

- Scaling limit of NAND: 10-20nm
- Fe-NAND for low voltage/power high capacity SSD.