P&S Modern SSDs

Basics of NAND Flash-Based SSDs

Dr. Mohammad Sadrosadati
Prof. Onur Mutlu

ETH Zürich
Spring 2024
26 March 2024
Today’s Agenda

- SSD Organization & Request Handling
- NAND Flash Organization
- NAND Flash Operations
A modern SSD is a complicated system that consists of multiple cores, HW controllers, DRAM, and NAND flash memory packages.
Another Overview

Host Interface Layer (HIL)

Flash Translation Layer (FTL)
- Data Cache Management
- Address Translation
- GC/WL/Refresh/...

Flash Controller
- ECC
- Randomizer

DRAM
- Host Request Queue
- Write Buffer
- Logical-to-Physical Mappings
- Metadata (e.g., P/E Cycles)

NAND Flash Package

Logical-to-Physical Mappings

Metadata (e.g., P/E Cycles)
Request Handling: Write

- Communication with the host operating system (receives & returns requests)
  - Via a certain interface (SATA or NVMe)

- A host I/O request includes
  - Request direction (read or write)
  - Offset (start sector address)
  - Size (number of sectors)
  - Typically aligned by 4 KiB
Request Handling: Write

- **Host Interface Layer (HIL)**
- **Flash Translation Layer (FTL)**
  - Data Cache Management
  - Address Translation
  - GC/WL/Refresh/…

**Flash Controller**
- ECC
- Randomizer

**DRAM**
- Host Request Queue
- Write Buffer
- Logical-to-Physical Mappings
- Metadata (e.g., P/E Cycles)

- **Buffering data to write (read from NAND flash memory)**
  - Essential to reducing write latency
  - Enables flexible I/O scheduling
  - Helpful for improving lifetime (not so likely)

- **Limited size (e.g., tens of MBs)**
  - Needs to ensure data integrity even under sudden power-off
  - Most DRAM capacity is used for L2P mappings
Request Handling: Write

**Host Interface Layer (HIL)**

**Flash Translation Layer (FTL)**
- Data Cache Management
- Address Translation
- GC/WL/Refresh/...

**Flash Controller**
- ECC
- Randomizer

**DRAM**
- Host Request Queue
- Write Buffer
- Logical-to-Physical Mappings
- Metadata (e.g., P/E Cycles)

- Core functionality for out-of-place writes
  - To hide the erase-before-write property
- Needs to maintain L2P mappings
  - Logical Page Address (LPA) → Physical Page Address (PPA)
- Mapping granularity: 4 KiB
  - 4 Bytes for 4 KiB → 0.1% of SSD capacity
Request Handling: Write

Host Interface Layer (HIL)

Flash Translation Layer (FTL)
- Data Cache Management
- Address Translation
- GC/WL/Refresh/...

Flash Controller
- ECC
- Randomizer
- NAND Flash Package
- NAND Flash Package
- NAND Flash Package

DRAM
- Host Request Queue
- Write Buffer
- Logical-to-Physical Mappings
- Metadata (e.g., P/E Cycles)

- Garbage collection (GC)
  - Reclams free pages
  - Selects a victim block → copies all valid pages → erase the victim block

- Wear-leveling (WL)
  - Evenly distributes P/E cycles across NAND flash blocks
  - Hot/cold swapping

- Data refresh
  - Refresh pages with long retention ages
Request Handling: Write

### Host Interface Layer (HIL)
- Data Cache Management
- Address Translation
- GC/WL/Refresh/

### Flash Translation Layer (FTL)
- Write Buffer
- Logical-to-Physical Mappings
- Metadata (e.g., P/E Cycles)

### Flash Controller
- ECC
- Randomizer

- NAND Flash Package
- NAND Flash Package
- NAND Flash Package

### DRAM
- Host Request Queue
- Error-correcting codes (ECC)
  - Can detect/correct errors: e.g., 72 bits/1 KiB error-correction capability
  - Stores additional parity information together with raw data
- Randomizer
  - Scrambling data to write
  - To avoid worst-case data patterns that can lead to significant errors
- Issues NAND flash commands
Request Handling: Read

- Host Interface Layer (HIL)

- Flash Translation Layer (FTL)
  - Data Cache Management
  - Address Translation
  - GC/WL/Refresh/...

- Flash Controller
  - ECC
  - Randomizer

- NAND Flash Package

- NAND Flash Package

- NAND Flash Package

- NAND Flash Package

- NAND Flash Package

- DRAM
  - Host Request Queue
  - Write Buffer
  - Logical-to-Physical Mappings
  - Metadata (e.g., P/E Cycles)

- First checks if the request data exists in the write buffer
  - If so, returns the corresponding request immediately with the data

- A host read request can be involved with several pages
  - Such a request can be returned only after all the requested data is ready
Request Handling: Read

- **Host Interface Layer (HIL)**
- **Flash Translation Layer (FTL)**
  - Data Cache Management
  - Address Translation
  - GC/WL/Refresh/…
- **Flash Controller**
  - ECC
  - Randomizer
- **DRAM**
  - Host Request Queue
  - Write Buffer
  - Logical-to-Physical Mappings
  - Metadata (e.g., P/E Cycles)

- Finds the PPA where the request data is stored from the L2P mapping table
### Request Handling: Read

<table>
<thead>
<tr>
<th>Host Interface Layer (HIL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flash Translation Layer (FTL)</td>
</tr>
<tr>
<td>- Data Cache Management</td>
</tr>
<tr>
<td>- Address Translation</td>
</tr>
<tr>
<td>- GC/WL/Refresh/...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Host Request Queue</td>
</tr>
<tr>
<td>- Write Buffer</td>
</tr>
<tr>
<td>- Logical-to-Physical Mappings</td>
</tr>
<tr>
<td>- Metadata (e.g., P/E Cycles)</td>
</tr>
</tbody>
</table>

#### Flash Controller

- ECC
- Randomizer

#### NAND Flash Packages

- NAND Flash Package
- NAND Flash Package
- NAND Flash Package

<table>
<thead>
<tr>
<th>CTRL</th>
</tr>
</thead>
</table>

- First reads the raw data from the flash chip
- Performs ECC decoding
- Derandomizes the raw data
- ECC decoding can fail
  - Retries reading of the page w/ adjusted $V_{\text{REF}}$
Today’s Agenda

- SSD Organization & Request Handling
- NAND Flash Organization
- NAND Flash Operation
A Flash Cell

- Basically, it is a transistor
A Flash Cell

- Basically, it is a transistor
  - w/ a special material: Floating gate (2D) or Charge trap (3D)
A Flash Cell

- Basically, it is a transistor
  - w/ a special material: Floating gate (2D) or Charge trap (3D)
  - Can hold electrons in a non-volatile manner

\[ V_{PGM} = 20 \, V \]

\[ \text{FG (Floating Gate)} \]

\[ \text{G (Control Gate)} \]

\[ \text{S (Source)} \]

\[ \text{D (Drain)} \]

\[ \text{GND (Substrate)} \]

\[ V_{TH} \]

\[ V_{GS} \]

\[ \text{Tunneling} \]
A Flash Cell

- Basically, it is a transistor
  - w/ a special material: Floating gate (2D) or Charge trap (3D)
  - Can hold electrons in a non-volatile manner
  - Changes the cell’s threshold voltage ($V_{TH}$)

![Flash Cell Diagram]

- $S$ (Source)
- $D$ (Drain)
- $G$ (Control Gate)
- $GND$
- $FG$ (Floating Gate)
- $Tunneling$
- $20 \, V$
- $V_{GS}$
- $I_D$
- $V_{TH}$
- $V_{TH}'$
- $V_{REF}$
- $V_{TH} < V_{REF}$
- $V_{TH} > V_{REF}$
Flash Cell Characteristics

- **Multi-leveling:** A flash cell can store multiple bits

  - *Program:* Inject electrons
  - *Erase:* Eject electrons

- **Retention loss:** A cell leaks electrons over time

  - 1 year
  - 10 years

- **Limited lifetime:** A cell wears out after P/E cycling

  - 1 year @ 1K P/E cycles
  - 1 year @ 10K P/E cycles
  - Retention error!
A NAND String

- Multiple (e.g., 128) flash cells are serially connected

$$V_{\text{PASS}} (> 6 \text{V})$$

Target Cell

$$V_{\text{PASS}}$$

$$V_{\text{PASS}}$$

NAND String
Pages and Blocks

- A large number (> 100,000) of cells operate concurrently

Page = 16 + α KiB

Block = {(# of WL) × (# of bits per cell)} pages
Pages and Blocks (Continued)

- Program and erase: Unidirectional
  - Programming a cell → Increasing the cell’s $V_{TH}$
  - Erasing a cell → Decreasing the cell’s $V_{TH}$

- Programming a page cannot change ‘0’ cells to ‘1’ cells → Erase-before-write property

- Erase unit: Block
  - Increase erase bandwidth
  - Makes in-place write on a page very inefficient → Out-of-place write & GC
A large number (> 1,000) of blocks share bitlines in a plane.
A large number (> 1,000) of blocks share bitlines in a plane.
Planes and Dies

- A die contains multiple (e.g., 2 – 4) planes

![Diagram of planes and blocks]

- Planes share decoders: limits internal parallelism (only operations @ the same WL offset)
Today’s Agenda

- SSD Organization & Request Handling
- NAND Flash Organization
- NAND Flash Operation
Threshold Voltage Distribution

- $V_{TH}$ distribution of cells in a programmed page/block/chip

There are $y$ cells whose $V_{TH} = xV$

Why distribution? Variations across the cells
- Some cells are more easily programmed or erased
Multi-level cell (MLC) technique

- \(2^m V_{TH}\) states required to store \(m\) bits in a single flash cell

Limited width of the \(V_{TH}\) window: Need to

- Make each \(V_{TH}\) state narrow
- Guarantee sufficient margins b/w adjacent \(V_{TH}\) states
**V_{TH} Distribution of MLC NAND Flash**

- **Multi-level cell (MLC) technique**
  - $2^m V_{TH}$ states required to store $m$ bits in a single flash cell

- **Limited width** of the $V_{TH}$ window: Need to
  - Make each $V_{TH}$ state narrow
  - Guarantee sufficient margins b/w adjacent $V_{TH}$ states
  - $V_{TH}$ changes over time after programmed
  - Narrower margins $\rightarrow$ Lower reliability
  - More bits per cell $\rightarrow$ higher density but lower reliability
Basic Operation: Page Program
Basic Operation: Page Program

- **WL control** – All other cells operate as a **resistance**
Basic Operation: Page Program

- BL control – **Inhibits cells** to not be programmed

```
V_{PROG} \quad WL_k
```

```
program

0 1 0 1 0 ...

BL_0  BL_1  BL_2  BL_3  BL_{132,095}

inhibit
```
Basic Operation: Page Program

- **BL control** – **Inhibits cells** to not be programmed

Diagram:

- **V<sub>PROG</sub>**
- **WL<sub>k</sub>**
- **BL<sub>0</sub>**
- **BL<sub>1</sub>**
- **BL<sub>2</sub>**
- **BL<sub>3</sub>**
- **BL<sub>132,095</sub>**

Connections:

- **To GND**
- **To V<sub>CC</sub>**
Basic Operation: Page Program

- **V\text{PROG}** \(\rightarrow\) WL\(_k\)
- **BL\(_0\)**: Program
- **BL\(_1\)**: Inhibit
- **BL\(_2\)**
- **BL\(_3\)**
- **BL\(_{132,095}\)**

- \# of cells
- Erased (E)

Threshold voltage (\(V_{TH}\))

\(V_{REF}\)
Basic Operation: Page Program

V_{PROG} \rightarrow WL_k

BL_0: 0 (program)
BL_1: 1 (inhibit)
BL_2: 0
BL_3: 1
BL_{132,095}: 0

To GND
To V_{CC}
To GND
To V_{CC}
To GND

# of cells

Threshold voltage (V_{TH})

Inhibited cells
Programmed cells

Erased (E)
Programmed

1
0

V_{REF}
Basic Operation: Page Program

Program and inhibit cells are shown in the diagram.

- Program cells: $V_{prog}$ applied to WL$_k$.
- Inhibit cells: $V_{cc}$ applied to BL.

$V_{th}$: Threshold voltage for cells.

- Erased (E): $V_{ref}$.
- Inhibited cells: Between $V_{ref}$.

Cells to program:
- Hard-to-program cells: $V_{th}$ close to $V_{ref}$.
- Easy-to-program cells: $V_{th}$ far from $V_{ref}$.

Number of cells:
- BL$_0$, BL$_1$, BL$_2$, BL$_3$, BL$_{132,095}$.
Basic Operation: Page Program

- Incremental Step-Pulse Programming (ISPP)

**Diagram:**
- **V_{PROG0}** and **WL_k**
- **BL_0** programmed
- **BL_1** inhibited
- **BL_2** and **BL_3** programmed
- **BL_{132,095}**

**Legend:**
- **# of cells**
- **Threshold voltage (V_{TH})**
- **Inhibited cells**
- **Erased (E)**
- **Cells to program**
- **V_{REF}**

**Flow Chart:**
- **Program**
- **Inhibit**
- **To GND**
- **To V_{CC}**
- **Verified as programmed**
Basic Operation: Page Program

- Incremental Step-Pulse Programming (ISPP)

- **Program**
  - WL_k
  - BL_0
  - BL_1
  - BL_2
  - BL_3
  - BL_{132,095}

- **Inhibit**
  - To GND
  - To V_{CC}

- **Cells to Program**
  - Inhibited cells
  - Erased (E)

- **Threshold voltage (V_{TH})**
  - V_{REF}
  - # of cells

- **Inhibited cells**
  - 1

- **Cells to program**
  - 1
Basic Operation: Page Program

- Incremental Step-Pulse Programming (ISPP)

![Diagram]

- **Program** (BL0)
  - 0
  - To GND

- **Inhibit** (BL1)
  - 1
  - To Vcc

- (BL2)
  - 0
  - To Vcc

- (BL3)
  - 1
  - To Vcc

- (BL132,095)
  - 0
  - To GND

`V_{PROG}` **WL**

- # of cells
- Erased (E)
- Inhibited cells
- VREF
- Cells to program

Threshold voltage (V_{TH})
Basic Operation: Page Program

- Incremental Step-Pulse Programming (ISPP)

![Diagram showing page programming with control signals and voltage levels for inhibit and program]
Basic Operation: Page Read

- WL control – All other cells operate as a resistance

![Diagram showing WL control and cell operation](image-url)
Basic Operation: Page Read

- **BL control** – Charge all BLs

![Diagram showing BL control and Vcc connections](image_url)

- **# of cells**
  - **Erased (E)**
  - **Programmed**

- **Threshold voltage (V_{TH})**
  - **V_{REF}**
Basic Operation: Page Read

- Sensing the current through BLs

\[ V_{\text{REF}} \quad \text{WL}_k \]

\[ \begin{array}{c}
\text{BL}_0 \\
\text{0} \\
0 \\
(\text{No current}) \\
(\text{Current}) \\
\text{1} \\
\text{BL}_1 \\
\text{BL}_2 \\
\text{BL}_3 \\
\end{array} \]

\[ \begin{array}{c}
\text{0} \\
\text{1} \\
\text{0} \\
\text{1} \\
\text{BL}_{132,095} \\
\text{0} \\
\end{array} \]

- Threshold voltage (V\(_{\text{TH}}\))

\[ V_{\text{TH}} < V_{\text{REF}} \quad \text{Erased (E)} \]

\[ V_{\text{TH}} > V_{\text{REF}} \quad \text{Programmed} \]
Basic Operation: Page Read - MLC

- Sensing the current through BLs

[Diagram showing the sensing process through BLs]
Basic Operation: Page Read - MLC

- Sensing the current through BLs

![Diagram showing sensing the current through BLs with WL_k, BL_0 to BL_132,095, and cells E, P1 to P7 with reference voltages VREF0 to VREF6.](image-url)
Basic Operation: Page Read - MLC

- Sensing the current through BLs

![Diagram of Basic Operation: Page Read - MLC]

- Sensing the current through BLs

![Diagram of Basic Operation: Page Read - MLC]

- Sensing the current through BLs

![Diagram of Basic Operation: Page Read - MLC]
Basic Operation: Page Read - MLC

- Sensing the current through BLs
Basic Operation: Page Read - MLC

- Sensing the current through BLs

![Diagram showing sensing the current through BLs](image_url)
Basic Operation: Page Read - MLC

- Sensing the current through BLs

...
Basic Operation: Page Read - MLC

- Sensing the current through BLs

Sensing the current through BLs

**V_{TH} < V_{REF}**

- **V_{TH} < V_{REF}**
- **V_{TH} < V_{REF}**
- **V_{TH} < V_{REF}**
- **V_{TH} < V_{REF}**

**CSB**

**V_{REF0}**

**V_{REF1}**

**V_{REF2}**

**V_{REF3}**

**V_{REF4}**

**V_{REF5}**

**V_{REF6}**

**P4**

**P5**

**P6**

**P7**

**V_{TH}**
Basic Operation: Page Read – Takeaways

- MLC NAND flash memory requires an **on-chip XOR logic**
- Bit-encoding affects the read latency!
  - Compare # of sensing for LSB

![Diagram of V_{TH} and V_{REF} levels with binary values for MSB and LSB]
Basic Operation: Page Read – Takeaways

- MLC NAND flash memory requires an on-chip XOR logic
- Bit-encoding affects the read latency!
  - Compare # of sensing for LSB
Basic Operation: Page Read – Takeaways

- MLC NAND flash memory requires an on-chip XOR logic
- Bit-encoding affects the read latency!
  - Compare # of sensing for LSB

![Diagram showing V_{TH} for different states of cells](image-url)
Basic Operation: Page Read – Takeaways

- MLC NAND flash memory requires an on-chip XOR logic
- Bit-encoding affects the read latency!
  - Compare # of sensing for LSB

![Diagram of Page Read with V_{TH} and V_{REF} levels]

- # of cells
  - MSB
  - LSB
  - CSB
- E
  - P1
  - P2
  - P3
  - P4
  - P5
  - P6
  - P7
- V_{REF0}
  - V_{REF1}
  - V_{REF2}
  - V_{REF3}
  - V_{REF4}
  - V_{REF5}
  - V_{REF6}
P&S Modern SSDs
Basics of NAND Flash-Based SSDs

Dr. Mohammad Sadrosadati
Prof. Onur Mutlu
ETH Zürich
Spring 2024
26 March 2024