NAND Flash Memory: Basics, Key Scaling Challenges and Future Outlook

Pranav Kalavade
Intel Corporation
pranav.kalavade@intel.com

October 2012
Outline

• Flash Memory Product Trends
• Flash Memory Device Primer
• Flash Technology Scaling
• Key Scaling Challenges
• Summary
Stationary $\rightarrow$ Mobile $\rightarrow$ Digital = Solid State NVM
Flash Market

• Fast Read/Write performance, high endurance and reliability, inherent ruggedness combined with the continued price reduction through technology scaling has been enabling new Flash markets
  – Flash mostly used as BIOS in the early 90’s
  – Mobile Communication and Computing fueled the Flash growth through the mid-90’s
  – Data applications – Cameras and USB drives – continued the demand for higher densities
  – Density demand continued to grow with the use of Flash in MP3 players followed by video recording

• Today Flash is leading DRAM in driving the Semiconductor technology
SSDs Driving Future TAM Growth: ~3X growth Forecasted to become Largest Market Segment

### NAND Flash Outlook

A $49+ Billion Market in 2015

<table>
<thead>
<tr>
<th>END MARKET</th>
<th>TAM – 2011A*</th>
<th>TAM – 2015E*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile</td>
<td>$11.8</td>
<td>$19.5</td>
</tr>
<tr>
<td>Client SSD</td>
<td>$3.4</td>
<td>$9.2</td>
</tr>
<tr>
<td>Enterprise SSD</td>
<td>$2.9</td>
<td>$7.8</td>
</tr>
<tr>
<td>Consumer &amp; others</td>
<td>$9.2</td>
<td>$12.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>$27.2</strong></td>
<td><strong>$49</strong></td>
</tr>
</tbody>
</table>

*Source: Forward Insights, SSD Insights Q112, March 2012 and Forward Insights, NAND Insights Q112, March 2012*
Outline

- Flash Memory Product Trends
- Flash Memory Device Primer
- Flash Technology Scaling
- Key Scaling Challenges
- Summary
Floating Gate Flash Memory Device

- **Stacked Gate NMOS Transistor**
  - Poly1 Floating Gate for charge storage
  - Poly2 Control Gate for accessing the transistor
  - Silicon dioxide for Gate oxide (Tunnel oxide)
  - Oxide-Nitride-Oxide (ONO) for the inter Poly Dielectric
  - Source/Drain Junctions optimized for Program/Erase
• Floating Gate capacitively coupled to the Gate, Drain, Source, and Substrate

• Gate Coupling Ratio (GCR) = $\frac{C_{CG}}{C_{TOT}}$
Floating Gate Device

- **Threshold Voltage shift** = $\Delta Q_{FG}/C_{CG}$
- **Programming** = Electrons Stored on the FG = High Vt
- **Erasing** = Remove electrons from the FG = Low Vt
NOR & NAND Logic Circuit

1. Transistors directly connected to output & ground.
2. Current flows selectively through individual transistor

1. Transistors connected thru each other to output & ground.
2. Current flows simultaneously through all transistors
NAND & NOR Flash

Two of the most popular Flash Memory types:

- Both NAND and NOR have Dual Gate NMOS with charge storage in the Poly1 floating gate.

- NOR: Each cell is connected to the Bitline directly with the source being common to the whole block.

- NAND: A string of serially connected cells (typically 64) connected to the bitline through a drain select transistor and common source through a source select transistor.

NAND string typically has 64 cells.
• Lack of individual cell contacts in the NAND cell, makes NAND cell inherently smaller in size (NAND - typically 4-5$\lambda^2$; NOR typically 9-10$\lambda^2$)

• However, the Select Gates add additional area of about 20-30% for the case of NAND. Since the select gate devices do not scale very well, this overhead increases with scaling.
**NAND vs. NOR**

- NAND is the dominant technology for data storage since it provides both high programming throughput and high density at low cost. Not suitable for execute in place applications due to slow random access.

- NOR is the mainstream technology for code storage and execute in place applications since it provides fast random access. Less suitable for data storage due to the lower write throughput and higher cost.
NAND Flash

Layout

Equivalent Circuit

Select Transistors shown in the equivalent circuit, but not shown in the layout or cross-sections.

X Cross-section

Y Cross-section
NAND Flash Cell Read

- Erased Cell Vt: $< -1V$
- Programmed Cell Vt: $1-3V$
- To Read a Cell:
  - Bitline is pre-charged to $\sim 1.0V$
  - Vcc is applied to the select gates of the string (block) to be selected
  - All the deselected WL on this string (block) are biased at Vpass ($\sim 5V$) which has to be higher than the highest program Vt
  - Selected WL is held at 0V
  - If the selected cell Vt $< 0V$, the string will conduct and the bitline is discharged and Sense Amp reads the data as “1”
  - If the selected cell Vt $> 0V$, the string will not conduct and the bitline stays at $\sim 1.0V$ and Sense Amp reads the data as “0”
Multi-level Cell

MLC = 2X the bits in 1X the area
3b/c = 3X the bits in 1X the area
4b/c = 4X the bits in 1X the area

Use of multi-level cells offers additional scaling. However, makes technology requirements more stringent

M. Goldman, IMW 2011, “25nm 3b/c NAND Flash”
NAND Flash Erase is by FN Tunneling.

- Tunnel Erase using Channel Erase with high positive voltage applied to the Well with gate grounded.
- Erase time ~1ms. Block Size ~ 2-4MByte

Y. S. Yim, et al. IEDM 2003
NAND Flash Programming - FN Tunneling

• Tunnel Programming from channel by biasing the Top Gate positive with respect to the channel and S/D.

• Actual Program Time (cumulative pulse width) ~100us. Program current ~ Displacement and Tunneling current. Low current allows large parallelism. Page size: 8-16KBytes.
Program Inhibit

- Uses the capacitive coupling between the Control Gate and the Channel to boost the channel to high enough a voltage to inhibit programming
Program/Inhibit Disturb

- There are two types of disturbs during programming:
  - Cells on the same WL see program disturb. If the Programming WL voltage is high and boosted channel voltage is not high enough these cells can get disturbed up.
  - Cells on the selected bitline see Inhibit disturb. Here the channel is grounded and all the deselected WL are sitting at Inhibit gate voltage. If the Inhibit gate voltage is high, these cells can get disturbed up.

- Program Disturb happens when the cells on that WL are programmed.
- Inhibit disturb happens every time any other WL in that block is programmed.
Outline

• Flash Memory Product Trends
• Flash Memory Device Primer
• Flash Technology Scaling
• Key Scaling Challenges
• Summary
Flash Scaling Constraints

- Data retention limits the Tunnel-oxide scaling to ~ 70A
- ONO leakage limits the ONO scaling to ~120 - 140A
- Programming/Erase requires 20V between the Control gate and channel

- Limited voltage and oxide scaling precludes traditional CMOS type scaling
Technology Scaling

- Yet Cell Area has scaled ~2X every full generation
- However, scaling challenges increase every generation
Outline

• Flash Memory Product Trends
• Flash Memory Device Primer
• Flash Technology Scaling
• Key Scaling Challenges
• Summary
Future Scaling Challenges & Trends

• Cell to Cell interference
• Floating Gate Cell Scaling Limitations
• Few Electron Effects
  – Random Telegraph Signal (1/f) Noise
  – Programming Fluctuations
• Voltage Scaling Issues
• 3-Dimensional approaches
• With scaling, the influence of the neighboring cells becomes important as well
• A more thorough capacitive analysis should consider the capacitive coupling of the floating gate to other nodes as well – such as to the neighboring floating gates, wordlines and even the adjacent channels.
FG-Channel Coupling

- Floating gate potential of cell1 affects the inverted channel of cell2 by capacitive coupling.
- Effectively, if cell1 is programmed, a higher gate potential is required to achieve the same amount of inversion layer charge density of cell2 than if cell1 is erased.
- This effect adds to cell to cell coupling.
At historic rate of increase, floating gate interference would have become ~50% by the 20nm node
Interference Reduction

• Many techniques will be tried to reduce the interference
  – Air gap between the Floating Gates (reduce K)
  – Floating Gate thickness scaling (Reduce Coupling Area)
  – Programming algorithms to account for this
High-K for IPD in Flash

- Not enough space between Floating Gate to accommodate Poly2 and ONO at 20nm node
- Potentially High K material as replacement of the inter-poly dielectric ONO to provide good control gate to floating gate coupling
  - Interface/mobility requirements of the Inter-poly dielectric requirements less stringent than those for the gate oxide.
Intel–Micron’s 20nm Cell

• Intel’s 20nm Flash Cell is a planar cell with high-K dielectric
• The planar cell is the best solution for 20nm and beyond
Few Electron Effects

• As the cell dimensions scale:
  – Cell capacitance scales. Less number of electron stored per unit shift in the threshold voltage
    ➢ Higher impact of single electron events
  – The channel area scales. With gate-oxide not scaling, less number of electrons in the channel
    ➢ Higher impact of 1/f noise
  – Less number of dopant atoms in the channel
    ➢ Larger spread in Vt due to doping fluctuations
Injection Statistics

- Statistical distribution in the number of injected electrons during programming introduces fluctuation in the cell programming rate
- The fluctuation increases as the memory cell size scales down

C. M. Compagnoni, et al. IEDM 2007
Random Telegraph Signal (RTS)

- Channel electron trapping and de-trapping during read causes current to fluctuate in discrete steps, like a random telegraph key being hit -> RTS noise
- The longer you look (many reads), the bigger the events you’ll see...
- The more number of cells you look, the bigger the events you’ll see
Random Telegraph Signal (RTS)

- Large-amplitude-noise tail portion of the noise distribution produced by the current percolation effect
- The noise amplitude is found to follow $1/(L_{\text{eff}}W_{\text{eff}})^n$ making it worse as the cell is scaled

3-D Approach

• Beyond the planar cell, NAND scaling will continue with 3D NAND
  ➢ Go vertical by stacking cells instead of scaling horizontally.
Various 3D architectures

<table>
<thead>
<tr>
<th></th>
<th>P-BiCS</th>
<th>TCAT</th>
<th>VSAT</th>
<th>VG</th>
</tr>
</thead>
<tbody>
<tr>
<td>String</td>
<td><img src="image1" alt="Diagram" /></td>
<td><img src="image2" alt="Diagram" /></td>
<td><img src="image3" alt="Diagram" /></td>
<td><img src="image4" alt="Diagram" /></td>
</tr>
<tr>
<td>Cell Shape</td>
<td><img src="image5" alt="Diagram" /></td>
<td><img src="image6" alt="Diagram" /></td>
<td><img src="image7" alt="Diagram" /></td>
<td><img src="image8" alt="Diagram" /></td>
</tr>
<tr>
<td>Cell Size in X, Y</td>
<td>6F^2 (3F*2F)</td>
<td>6F^2 (3F*2F)</td>
<td>6F^2 (3F*2F)</td>
<td>4F^2 (2F*2F)</td>
</tr>
<tr>
<td>Gate Process</td>
<td>Gate first</td>
<td>Gate Last</td>
<td>Gate First</td>
<td>Gate Last</td>
</tr>
<tr>
<td>Current Flow</td>
<td>U-turn</td>
<td>Vertical</td>
<td>Multi-U-turn</td>
<td>Horizontal</td>
</tr>
</tbody>
</table>

Y. H. Hsiao et. al. 2010 IMW

- Several approaches are being explored in the industry, which can be broadly classified based on current flow directions: vertical vs horizontal
3D: Similarity and differences

• For a small effective cell foot-print, the physical cell size is quite a bit larger than planar 2D, due to stacking
  – This providing relief to the scaling issues causing state-width degradations
  – However, 3D process integration could introduce its own degradations

• One new “feature” of 3D approaches appear to be the use of a poly-Si body / channel
Conclusions

• Flash Memory is playing an important role in the way we work and live
• This was enabled through continued technology scaling over the past 25 years
• Path ahead for scaling is more challenging than the path already traversed
• There are a large number of possible technical paths
  – With planar and 3D structures, we see continued scaling for the next several generations