Nonvolatile Erasable Flash NAND Memories

Raul Adrian Cernea
SanDisk Corporation, Milpitas, California
Outline

• NAND Flash and memory cards
• Basic operations and multi-level
• Performance trends and ABL
• 3X MLC program-throughput
• Die size challenge and solutions
• Accelerated data access rate
• 3 bits per cell
• Power and energy reduction
• Summary and conclusions
Mobile/CE/PC Driving NAND Future
Memory cards

Every card:

- One controller
- At least one memory chip
Erase and program

WORD LINE

FLOATING GATES
THIN OXIDE

P WELL

GROUNDED BIT LINE
FLOATING BIT LINE

E

P
Cross-section along the Word Line

WORD LINE

FLOATING GATES

NON-CONDUCTIVE NAND CHAIN

CONDUCTIVE NAND CHAIN
Cross-section along the Bit Line

CONDUCTIVE NAND CHAIN

NON-CONDUCTIVE NAND CHAIN
Program and verify

WORD LINE VOLTAGE

PROGRAM

WORD LINE VOLTAGE

PROGRAM

VERIFY

SLC

MLC

v_{TH}

v_{TH}

8
SLC and MLC

SINGLE LEVEL CELL

MULTI LEVEL CELLS

NUMBER OF CELLS

TWO BITS PER CELL

THREE BITS PER CELL

FOUR BITS PER CELL
Performance trends and All Bit Line Architecture
Program-throughput trends for NAND memories

- CURRENT 16Gb ABL
  - ISSCC 2008

- ABL MLC
  - FULL SEQUENCE

- ABL
  - SLC
  - WITHOUT HIGH SPEED I/O INTERFACE

- ABL MLC CONVENTIONAL

- 11%
  - VLSI 2007; 18.4

- MLC
  - ISSCC 2006; 7.7

- MLC
  - ISSCC 2005; 2.1

- SLC
  - ISSCC 2003; 16.7
  - ISSCC 2004; 2.7

- ABL MLC
  - WITHOUT HIGH SPEED I/O INTERFACE

- 10MB/s
  - 2003

- 20MB/s
  - 2004

- 30MB/s
  - 2005

- 40MB/s
  - 2006

- 50MB/s
  - 2007

- 60MB/s
  - 2008
Cell access

Y DECODER: ACCESS TO ALL BIT LINES

X DECODER: SELECTED WORD LINE

X DECODER: SELECTED BLOCK

X DECODER: UNSELECTED BLOCK(S)

SENSE AMPLIFIER

SENSE AMPLIFIER

SENSE AMPLIFIER

32 NAND CELLS
• Simultaneous access of two word lines of the same size
• ABL: double Column Logic, two sided
Differences between the two architectures

<table>
<thead>
<tr>
<th>Features</th>
<th>Conventional</th>
<th>All Bit Line Full-Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>8Gb</td>
<td>16Gb</td>
</tr>
<tr>
<td>Number of Blocks per Plane</td>
<td>1K</td>
<td>2K</td>
</tr>
<tr>
<td>Column Logic</td>
<td>One Sided</td>
<td>Two Sided</td>
</tr>
<tr>
<td>Number of Bit Lines per Sense Amplifier</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4KB Pages Programmed in Parallel</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Maximum MLC Program-Throughput</td>
<td>10MB/s</td>
<td>34MB/s</td>
</tr>
</tbody>
</table>
Maximum parallelism

- Conventional is an Even / Odd architecture
  - One Sense Amplifier handling two Bit Lines
  - Every other Bit Line shielded during sensing
- Full READ and PROGRAM capabilities for ABL
  - No shielding necessary (as in current mainstream NAND architecture)
Conventional column architecture

- Two Bit Lines for every read-program unit
- All similar column blocks
ABL: three level hierarchical architecture

- One Bit Line connected to one Sense Amplifier
- Sense Amplifiers, Data Latches, one Sequential Logic Unit and one Control Logic, all connected to the same bus
- Column Block Selection, Pointer, I/O and Local Drivers
3X MLC Program-Throughput
• There is a step up in programming time when one compares the lower page to the upper page (one state versus two states)
• The lower and upper pages together (three states) are the slowest to be programmed
Conventional, ABL and Full-Sequence

- In ABL architecture, programming four pages in parallel is twice as fast as Conventional.
- MLC programming in Full-Sequence is about three times faster than Conventional.
Yupin Effect

- Additional negative voltage in a two-step sequence for Floating Gates programmed in step TWO (Yupin Effect*)

Ref.: SanDisk US Patent 5,867,429 (Feb. 2 1999)

- Minimized effect in one-step programming (Full-Sequence)
ABL: faster verify with current sensing

- Pre-charge (by Bit Line driver): non-linear, fast (both)
- Discharge (by NAND cell): linear, slow (Conventional)
- Faster verify operation in ABL mode (pre-charge only)
• A full Word Line distribution of random data, programmed in ABL, Full-Sequence at a 34MB/s program-throughput
Die Size Challenge and Solutions
Hierarchical architecture for area saving

- One SA per BL means *double column circuitry*
  - Area reduction is a must

- Hierarchical architecture: less repeating circuits
  - Only one Sequential Logic Unit and Control Logic per group
  - Only one Block Selection, Pointer and I/O per block

- Saved area allows for local drivers
Efficient pumps for area saving

- Every stage doubles the (regulated) input voltage
- Cross coupled transistors as “Dynamical Diodes” to prevent the loss of one diode per stage
Each stage doubles the input voltage
Accelerated Input / Output Access Rate
Fast internal data I/O stream

- Pointer selection
- Three stage pipeline
  1. Quad-split loading
  2. Four-point pick up
  3. Top / Bottom
Two step I/O redundancy access

- Regular clock for redundancy bytes at Data In (or Data Out)
  - **STEP 1:** Full data stream into the regular buffer
  - **STEP 2:** Data transfer between regular buffer and redundancy zone
- Reversed process for Data Out
- Minimal time penalty for the hidden data transfer
Three bits per cell
# Chip architecture comparison

<table>
<thead>
<tr>
<th></th>
<th>56nm 8G D2 (ISSCC2006)</th>
<th>56nm 16G D2 (ISSCC2008)</th>
<th>56nm 16G D3 (This Work)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chip size</strong></td>
<td>99mm²</td>
<td>182mm²</td>
<td>142.5mm² (this work) (22% saving)</td>
</tr>
<tr>
<td>ISSCC2006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Density</strong></td>
<td>8Gb (80.8Mb/mm²)</td>
<td>16Gb (82Mb/mm²)</td>
<td>16Gb (112Mb/mm²)</td>
</tr>
<tr>
<td><strong>Architecture</strong></td>
<td>4Gb / plane</td>
<td>8Gb / plane</td>
<td>8Gb / plane</td>
</tr>
<tr>
<td><strong>Bits/Cell</strong></td>
<td>2bits/cell</td>
<td>2bits/cell</td>
<td>3bits/cell</td>
</tr>
<tr>
<td><strong>Program/Sense</strong></td>
<td>Even or Odd</td>
<td>ABL</td>
<td>ABL</td>
</tr>
</tbody>
</table>
3 bit per cell programming

- 3 pages (Lower/Middle/Upper) on each Word Line
- Each page can be treated as an independent page
- The pages have to be programmed sequentially
3 bit per cell programming and verify

- 7 verify operations after each programming pulse
- Conversion from “page by page” to three page programming for speed improvement
- Program-throughput comparable to conventional (non-ABL 2 bit per cell)
7 state distribution
Power and Energy Reduction
Power reduction during Data Stream

• 2V internally down-regulated VDD (from 2.7~3.6V)
• Pointer selection to avoid global address bus switching
ABL for lower energy

- Even / Odd sensing in Conventional architecture
  - Double sensing operations per Word Line
  - Strong coupling between adjacent Bit Lines plus parasitic to Ground
- All Bit Line sensing in one operation
  - Bit Line coupling to ground only (20% of total parasitic)
  - Lower Bit Line voltages
- At least 10 times less charge wasted when using ABL
Summary and Conclusions
## Summary, 2 bit per cell

<table>
<thead>
<tr>
<th>READ THROUGHPUT BY 8 I/O</th>
<th>50MB/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABL SLC PROGRAMMING (WITHOUT HIGH SPEED I/O INTERFACE)</td>
<td>60MB/s</td>
</tr>
<tr>
<td>ABL MLC PROGRAMMING NO FULL-SEQUENCE</td>
<td>24MB/s</td>
</tr>
<tr>
<td>ABL MLC PROGRAMMING FULL-SEQUENCE</td>
<td>34MB/s</td>
</tr>
<tr>
<td>16 Gb ABL DIE SIZE</td>
<td>182mm²</td>
</tr>
</tbody>
</table>
Summary, 3 bit per cell

<table>
<thead>
<tr>
<th>READ THROUGHPUT BY 8 I/O</th>
<th>40MB/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABL MLC PROGRAMMING FULL-SEQUENCE</td>
<td>8MB/s</td>
</tr>
<tr>
<td>16 Gb D3 DIE SIZE</td>
<td>142.5mm²</td>
</tr>
</tbody>
</table>
ABL is the architecture of the future

• The ABL architecture provides the fastest MLC program-throughput reported yet in 56nm and closes the gap between SLC and MLC

• By means of Full-Sequence programming the high voltage exposure is halved and the Bit Line interaction is minimized for maximum endurance

• The technology challenges of 43nm and beyond are well served by this architecture

• ABL is the architecture of choice for the next generations and 3 bit per cell encoding
Conclusions

• A new All Bit Line architecture boosts the MLC program-throughput of a 16Gb NAND by 240%

• With hierarchical column block architecture and a two-step redundancy access, a 50% faster internal I/O rate is achieved, even for lower power supply

• ABL provides an energy efficient system with an improved overall performance

• Lower die size when compared to a four plane “Conventional” chip of close performance
Acknowledgements

• The author wants to specially thank George Samachisa and Yan Li for their support

• The circuits presented here were made possible by the work of many engineers from SanDisk and ¹)Toshiba:

  Long Pham, Farookh Moogat, Siu Chan, Binh Le, Shouchang Tsao, Tai-Yuan Tseng, Khanh Nguyen, Jason Li, Jayson Hu, Cynthia Hsu, Fanglin Zhang, Teruhiko Kamei, Hiroaki Nasu, Phil Kliza, Khin Htoo, Jeffrey Lutze, Yingda Dong, Masaaki Higashitani, Junnhui Yang, Hung-Szu Lin, Vamshi Sakhamuri, Jong Hak Yuh, Alan Li, Feng Pan, Sridhar Yadala, Subodh Taigor, Kishan Pradhan, James Lan, James Chan, Takumi Abe¹, Yasuyuki Fukuda¹, Hideo Mukai¹, Koichi Kawakami¹, Connie Liang, Tommy Ip, Shu-Fen Chang, Jaggi Lakshmpati, Sharon Huynh, Dimitris Pantelakis, Seungpil Lee, Yupin Fong, Tien-Chien Kuo, Jong Park, Tapan Samaddar, Hao Nguyen, Man Mui, Emilio Yero, Gyuwan Kwon, Jun Wan, Tetsuya Kaneko¹, Hiroshi Maejima¹, Hitoshi Shiga¹, Makoto Hamada¹, Norihiro Fujita¹, Kazunori Kanebako¹, Eugene Tam, Anne Koh, Iris Lu, Calvin Kuo, Trung Pham, Jonathan Huynh, Qui Nguyen, Hardwell Chibvongodze, Mitsuyuki Watanabe, Ken Oowada, Grishma Shah, Byungki Woo, Ray Gao, Patrick Hong, Liping Peng, Debi Das, Dhritiman Ghosh, Vivek Kalluru, Sanjay Kulkarni, Mehrdad Mofidi, Khandker Quader, Chi-Ming Wang