A 3bit/Cell 32Gb NAND Flash Memory at 34nm with 6MB/s Program Throughput and with Dynamic 2b/Cell Blocks Configuration Mode for a Program Throughput Increase up to 13MB/s

Speaker:
Naso Giovanni – Micron Flash Design Center Avezzano Italy

ISSCC 2010 paper 24.7
Design team

**Micron Flash Design Center Avezzano (Italy)**:
L. Botticchio, C. Cerafogli, P. Conenna, A. D’Alessandro,
D. Di Cicco, L. De Santis, W. Di Francesco, M.L. Gallese,
G. Imondi, M. Incarnati, C. Lattaro, A. Macerola, G. G. Marotta,
V. Moschiano, C. Musilli, G. Naso, D. Orlandi, F. Paolini,
S. Perugini, L. Pilolli, G. Santin, F. Rori, M. Rossini, E. Sirizotti,
M. Tiburzi, A. Torsi, T. Vali

**Micron Flash Design Center S. Jose’ (California)**:
R. Ghodsi, F. Roohparvar

**Micron Product Engineering Boise (Idaho)**:
D. Rivers

**Intel Flash Design Center Folsom (California)**:
M. Goldman, C. Haid
Agenda

Device features
Dynamic bits per cell configuration
Device architecture and organization
Sensing
Data flow
VT placement
Throughput
Source bias
Erase ramp
Analog system
Temperature compensation
# Device features

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology</strong></td>
<td>34nm triple-well 3 metals</td>
</tr>
<tr>
<td><strong>Cell size</strong></td>
<td>0.0054 um² (select gates included)</td>
</tr>
<tr>
<td></td>
<td>64 WLs per string</td>
</tr>
<tr>
<td><strong>Chip size</strong></td>
<td>126 mm²</td>
</tr>
<tr>
<td><strong>Organization</strong></td>
<td>4096 Bytes per page x 384 pages x 684 blocks x 4 planes x 8 IO</td>
</tr>
<tr>
<td><strong>ECC per 4KB page</strong></td>
<td>224 Bytes</td>
</tr>
<tr>
<td><strong>Array read time</strong></td>
<td>60 us typ – 100 us max</td>
</tr>
<tr>
<td><strong>Program time</strong></td>
<td>2 ms typ – 10 ms max</td>
</tr>
<tr>
<td><strong>Erase time</strong></td>
<td>10 ms typ – 30 ms max</td>
</tr>
<tr>
<td><strong>Clock cycle time</strong></td>
<td>12 ns</td>
</tr>
<tr>
<td><strong>Supported multiple bit/cell</strong></td>
<td>1 bc – 2 bc – 3 bc (dynamic configuration)</td>
</tr>
</tbody>
</table>
Dynamic bits per cell configuration

Number of bits per cell can be dynamically set by the user to 1, 2, 3 through the set feature command.

Program and read algo are optimized for the different bit per cell configurations to achieve maximum performance and margins.

<table>
<thead>
<tr>
<th></th>
<th>3 bits per cell</th>
<th>2 bits per cell</th>
<th>1 bit per cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>tBERS (ms)</td>
<td>10 - 30</td>
<td>10 - 30</td>
<td>10 - 30</td>
</tr>
<tr>
<td>tPROG (ms)</td>
<td>2 - 10</td>
<td>1 - 2.2</td>
<td>0.25 - 0.9</td>
</tr>
<tr>
<td>tR (us)</td>
<td>60 - 100</td>
<td>50</td>
<td>30</td>
</tr>
</tbody>
</table>

Pages addressing is automatically adjusted based on bits per cell configuration.
Device architecture

- Plane 0
- Plane 1
- Plane 2
- Plane 3

- Cache
- Redundancy & drivers

- Control logic
- Datapath
- Analog

- Pads

32 Gbit at 3 bits per cell

4 planes

126 mm²

8 I/O
Device photo
Plane organization

- 684 blocks
- 64 active WL per block
- 384 pages at 3 bc
- 4096 bytes
- 224 bytes ECC
- 8 Gbit at 3bc
- Cache
Array string structure

66 physical WLs with different options to manage ‘hot carrier disturb’ at the edge

<table>
<thead>
<tr>
<th>WL</th>
<th>option1</th>
<th>option2</th>
<th>option3</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>1 bc</td>
<td>dummy</td>
<td>1 bc</td>
</tr>
<tr>
<td>64</td>
<td>2 bc</td>
<td>3 bc</td>
<td>1 bc</td>
</tr>
<tr>
<td>63</td>
<td>2 bc</td>
<td>3 bc</td>
<td>3 bc</td>
</tr>
<tr>
<td>62</td>
<td>3 bc</td>
<td>3 bc</td>
<td>3 bc</td>
</tr>
<tr>
<td>2</td>
<td>3 bc</td>
<td>3 bc</td>
<td>3 bc</td>
</tr>
<tr>
<td>1</td>
<td>3 bc</td>
<td>3 bc</td>
<td>3 bc</td>
</tr>
<tr>
<td>0</td>
<td>1 bc</td>
<td>dummy</td>
<td>1 bc</td>
</tr>
<tr>
<td>even/odd pages</td>
<td>384</td>
<td>384</td>
<td>384</td>
</tr>
</tbody>
</table>
Edge WL ‘hot carrier disturb’

- sgs/0V
- WL0 24V
- WL1 10V

Inhibited string during program

a = channel de-boost due to GIDL
b = hot electrons due to WL0-sgs E field

Hot carrier disturb depends on:
- sgs/WL0 dimension
- SGS bias
- WL0 voltage during program. In case 3bc WL0 is high (24V) in case 1bc WL0 is lower (15V).
- Hot carrier disturb applies also to WL65
Edge WL ‘hot carrier disturb’

Hot carrier disturb has been widely investigated in IEEE NVSMW 2006 pag. 31
Sensing (erased and programmed cells)

- Vi – Vf accounts for BL leakage
- Unsel BLs are used as gnd-shield

Diagram showing the timing and circuit elements:
- VCC, blclamp, sel BL, TDC, sgd/WL, sgs, pre,Cb, Vi, Vf, strobe time
Quad planes

Quad planes architecture with reduced bitlines RC and alternate bitlines program with reduced bitlines interference (the unselected are used as shield) allow fast program operation and high program sustainable throughput.
Byte data flow
Low-Mid-Upper page VT placement (LMU) for FG-FG reduction

When user programs a page, all the non programmed surrounding lower order pages can be programmed by an internal algorithm to minimize the FG-FG interference.
Programming time

Mean
Tprog = 2.46 msec
3bits/cell throughput

\[ \text{CKtr} = \frac{\text{Data DDR throughput}}{12 \text{ nsec}} = \frac{2 \text{ Bytes}}{12 \text{ nsec}} = 166 \text{ MB/s} \]

Sustain din throughput = \[\min(\text{CKtr}; \frac{4096 \text{ Bytes} \times 4}{2.46 \text{ msec}})\] = 6.6 MB/s (quad)

Sustain dout throughput = \[\min(\text{CKtr}; \frac{4096 \text{ Bytes} \times 4}{70 \text{ usec}})\] = 166 MB/s (quad)
Source bias:
total VT budget increase

Purpose of the Source bias technique is to ‘apparently’ move down all the distributions of about 1.5V and expand the negative VT region without the use of negative voltages on the word lines.

In this manner the total VT budget is expanded.

The source bias technique is performed during read and program verify.
Source bias

\[ V_{gs} = (WL - SRC) = VT \]
\[ V_{gs} = WL - (SRC + d) = VT - d \]

Conventional

gnd pass1 pass WL pass pass1 vclamp

Source bias

d pass1+d pass+d WL pass+d pass1+d vclamp+d

Vgs = (WL – SRC) = VT  \quad Vgs = WL – (SRC+d) = VT - d
- Ramping first erase pulse decreases the Fowler-Nordheim peak current.
- Ramp applies only to the first pulse because charged floating gates prevent further peak currents.
- Pulse ramp technique is beneficial also in program but it is not realistic for performance reasons.
Effect of erase ramp on Threshold VT placement

VT shift due to cycling and bake

Cycles (AU)

VT shift (AU)

0us

1ms

No ramp

1ms ramp

x1

x2

baked
cycled

22
Analog system

Resistor ladder provides a maximum number of 1024 analog voltages to 24 DACs: each one of the DACs has a max resolution of 10 bits.

Regulators (reg) are used to generate voltage for the core.

Regulators are divided in two categories: 5 on-off regulators to enable pumps and 19 linear regulators to directly provide regulated voltage to the core.

Linear regulators can be powered by auxiliary pumps to provide medium range voltages to the core or can be powered directly by VCC to provide voltages to the core that are lower than VCC.
Temperature effect on VT

US pat 7630266

% of cells

VT (AU)

Different target VT associated to the same target WL in prog verify

WL temperature compensated distributions

Uncompensated distributions

30C

50C

70C
Temperature detection

$V = hT$

$V_{ref} = k \times 1.25V$

bandgap and regulators

termolev

termoref

MUX

128 res

termobits [6:0]

To temperature compensation

US pat 7630265

SAR
Temperature compensation

register → nominal → DAC → reg

thermometer → table → nominal + compensated → DAC → reg
Temperature compensation features

7 bits digital resolution
-40°C to 90°C temperature range
1°C temperature resolution
Conversion time includes amplifier offset cancellation

6 different regulated voltages can be compensated:
  - wordline read/vfy
  - vpass read/pgm
  - SGS, SGD read/pgm
Typical uncycled distributions

sigma

VT (arbitrary unit)
Summary

A 3bit/cell 32Gb NAND Flash at 34nm was presented.

It has high sustainable write/read throughput obtained with the use of quad planes architecture and alternate BL shield in program and sensing operations.

Design techniques related to:
- LMU sequence
- Edge WL configuration
- Source bias
- Temperature compensation
- First erase pulse shaping

were used to increase VT distribution margins at time 0 and with age.