Monolithic 3D DRAM Technology

Deepak C. Sekar, Brian Cronquist, Israel Beinglass, Paul Lim, Zvi Or-Bach

15th June 2011
Outline

- Status of the DRAM industry today
- Monolithic 3D DRAM
- Implications and risks of the technology
- Summary
Outline

- Status of the DRAM industry today
DRAM makers fall in the “endangered species” category

<table>
<thead>
<tr>
<th>1996 24 key DRAM players</th>
<th>2010 9 key DRAM players</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samsung</td>
<td>Samsung</td>
</tr>
<tr>
<td>Micron</td>
<td>Hynix</td>
</tr>
<tr>
<td>NEC</td>
<td>Micron</td>
</tr>
<tr>
<td>Mitsubishi</td>
<td>Elpida</td>
</tr>
<tr>
<td>Fujitsu</td>
<td>Nanya</td>
</tr>
<tr>
<td>TI-Acer</td>
<td>Inotera</td>
</tr>
<tr>
<td>Powerchip</td>
<td>Powerchip</td>
</tr>
<tr>
<td>Winbond</td>
<td>ProMOS</td>
</tr>
<tr>
<td>IBM</td>
<td>ProMOS</td>
</tr>
<tr>
<td>Motorola</td>
<td>ProMOS</td>
</tr>
<tr>
<td>Seiko Epson</td>
<td>Winbond</td>
</tr>
<tr>
<td>UMC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Why? What are the challenges? We’ll see in the next few slides
Reason 1: Profitability

- DRAM has not been a profitable business in the near past
- Balance sheets of most companies’ DRAM businesses similar to above 😞
Reason 2: Large fab cost for scaling-down

- Scaling-down → lower cost per bit → but huge litho and fab investment
- Hard for unprofitable companies to fund scaled-down fabs. But if they don’t fund new “scaled-down” fabs and others do → bad cost per bit → profitability even worse 😞

Today,
Litho Tool Cost = $42M
Etch, CVD, Implant, RTA tools each cost <$5M
Reason 3: Scaling-down the stacked capacitor challenging

Capacitance
Keep ~25fF

Memory Cell Transistor
Keep low leakage current

Al₂O₃ (90nm) → HfO₂ (80nm) → ZrO₂ (60nm) → ?

<table>
<thead>
<tr>
<th>Source: ITRS 2010</th>
<th>45nm</th>
<th>32nm</th>
<th>22nm</th>
<th>15nm</th>
<th>10nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dielectric constant</td>
<td>40</td>
<td>50</td>
<td>60</td>
<td>65</td>
<td>70</td>
</tr>
<tr>
<td>Aspect ratio</td>
<td>47:1</td>
<td>56:1</td>
<td>99:1</td>
<td>147:1</td>
<td>193:1</td>
</tr>
<tr>
<td>EOT</td>
<td>0.8nm</td>
<td>0.6nm</td>
<td>0.5nm</td>
<td>0.3nm</td>
<td>0.2nm</td>
</tr>
</tbody>
</table>

Requires >150:1 aspect ratios and exotic new high-k dielectrics!
Reason 4: The cell transistor needs major updates on scaling-down

Cell Transistor – Possible Scenario

A major new transistor every generation or two!
100nm Planar → 80nm RCAT → 60nm S-RCAT → 35nm Finfet (?) → 20nm Vertical (?)
To recap,
Things don’t look good for DRAM vendors because

(1) Low profitability
(2) Cost of scaled-down fabs
(3) Scaling-down stacked capacitor
(4) Cell transistor scaling-down

Related

Common theme ➔ Scaling-down

Is there an alternative way to reduce DRAM bit cost other than scaling-down?

Focus of this presentation
Outline

- Status of the DRAM industry today
- Monolithic 3D DRAM
- Implications of the technology
- Summary
Key technology direction for NAND flash: Monolithic 3D with shared litho steps for memory layers

To be viable for DRAM, we require

- Single-crystal silicon at low thermal budget → Charge leakage low
- Novel monolithic 3D DRAM architecture with shared litho steps
Single crystal Si at low thermal budget

- Obtained using the ion-cut process. It’s use for SOI shown above.
- Ion-cut used for high-volume manufacturing SOI wafers for 10+ years.

MonolithIC 3D Inc. Patents Pending
Double-gated floating body memory cell well-studied in Silicon (for 2D-DRAM)

Hynix + Innovative Silicon
VLSI 2010

- 0.5V, 55nm channel length
- 900ms retention
- Bipolar mode

Intel
IEDM 2006

- 2V, 85nm channel length
- 10ms retention
- MOSFET mode
Our novel DRAM architecture

Innovatively combines these well-studied technologies

- Monolithic 3D with litho steps shared among multiple memory layers
- Stacked Single crystal Si with ion-cut
- Double gate floating body RAM cell (below) with charge stored in body
Process Flow: Step 1
Fabricate peripheral circuits followed by silicon oxide layer
Process Flow: Step 2
Transfer p Si layer atop peripheral circuit layer

Top layer
- H implant
- Silicon Oxide
- p Silicon

Bottom layer
- Silicon Oxide
- Peripheral circuits

Flip top layer and bond to bottom layer

H implant
- p Silicon
- Silicon Oxide
- Silicon Oxide
- Peripheral circuits
Process Flow: Step 3
Cleave along H plane, then CMP

- p Silicon
- Silicon Oxide
- Silicon Oxide
- Peripheral circuits
Process Flow: Step 4
Using a litho step, form n+ regions using implant
Process Flow: Step 5
Deposit oxide layer
Process Flow: Step 6
Using methods similar to Steps 2-5, form multiple Si/SiO$_2$ layers, RTA
Process Flow: Step 7
Use lithography and etch to define Silicon regions

This n+ Si region will act as wiring for the array… details later
Process Flow: Step 8
Deposit gate dielectric, gate electrode materials, CMP, litho and etch

Symbols

- n+ Silicon
- Silicon oxide
- Gate electrode
- Gate dielectric

MonolithIC 3D
Process Flow: Step 9
Deposit oxide, CMP. Oxide shown transparent for clarity.
Process Flow: Step 10
Make Bit Line (BL) contacts that are shared among various layers.
Process Flow: Step 11
Construct BLs, then contacts to BLs, WLs and SLs at edges of memory array using methods in [Tanaka, et al., VLSI 2007]
Some cross-sectional views for clarity. Each floating-body cell has unique combination of BL, WL, SL.
A different implementation:
With independent double gates
Outline

- Status of the DRAM industry today
- Monolithic 3D DRAM
- Implications and risks of the technology
- Summary

➢ Implications and risks of the technology
## Density estimation

<table>
<thead>
<tr>
<th></th>
<th>Conventional stacked capacitor DRAM</th>
<th>Monolithic 3D DRAM with 4 memory layers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell size</strong></td>
<td>$6F^2$</td>
<td>Since non self-aligned, $7.2F^2$</td>
</tr>
<tr>
<td><strong>Density</strong></td>
<td>$x$</td>
<td>$3.3x$</td>
</tr>
<tr>
<td><strong>Number of litho steps</strong></td>
<td>26 (with 3 stacked cap. masks)</td>
<td>~26 (3 extra masks for memory layers, but no stacked cap. masks)</td>
</tr>
</tbody>
</table>

### 3.3x improvement in density vs. standard DRAM, but similar number of critical litho steps!!!

Negligible prior work in Monolithic 3D DRAM with shared litho steps, poly Si 3D doesn’t work for DRAM (unlike NAND flash) due to leakage
Scalability

- Multiple generations of cost per bit improvement possible
  (eg) 22nm 2D →
  22nm 3D 2 layers →
  22nm 3D 4 layers → ...

- Use same 22nm litho tools for 6+ years above. Tool value goes down 50% every 2 years → Cheap 😊

- Avoids cost + risk of next-gen litho
Reduces or avoids some difficulties with scaling-down

EUV delays and risk

(EETimes 2002)
"EUV to be in production in 2007"

(EETimes 2003)
"EUV to be leading candidate for the 32nm node in 2009"

(EETimes 2004)
"EUV to be pushed out to 2013"

(EETimes 2010)
"EUV late for 10nm node milestone in 2015"

Capacitor manufacturing

<table>
<thead>
<tr>
<th>ε</th>
<th>45 nm</th>
<th>32 nm</th>
<th>22 nm</th>
<th>15 nm</th>
<th>10 nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>47</td>
<td>56</td>
<td>99</td>
<td>147</td>
<td>193</td>
</tr>
</tbody>
</table>

Continuous transistor updates

Planar → RCAT → S-RCAT → Finfet → Vertical devices
Risks

- **Floating-body RAM**
  Retention, reliability, smaller-size devices, etc

- **Cost of ion-cut**
  Supposed to be <$50-75 per layer since one implant, bond, cleave, CMP step.
  But might require optimization to reach this value.
Summary of Monolithic 3D DRAM Technology

- 3.3x density of conventional DRAM, but similar number of litho steps
- Scalable (eg) 22nm 2D → 22nm 3D 2 layers → 22nm 3D 4 layers → ...
- Cheap depreciated tools, less litho cost + risk, avoids many cap. & transistor upgrades
- Risks = Floating body RAM, ion-cut cost

Monolithic 3D with shared litho steps
Single crystal Si
Floating body RAM
*Under development...*
Backup slides
A note on overlay

- Implant n+ in p Si regions layer-by-layer, then form gate \( \rightarrow \) non self-aligned process

- ITRS <20% overlay requirement
  ASML 1950i = 3.5nm overlay for 38nm printing. <10% overlay.

- So, gate length = 1.2F. Penalty of 0.2F for non-self-aligned process
Bias schemes for floating body RAM

Bipolar Mode [S. Alam, et al, TED 2010]

MOS Mode [Intel, IEDM 2006]

<table>
<thead>
<tr>
<th></th>
<th>$V_D$</th>
<th>$V_{FG}$</th>
<th>$V_{BG}$</th>
<th>$V_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>2</td>
<td>1.5</td>
<td>-1.5</td>
<td>0</td>
</tr>
<tr>
<td>Erase</td>
<td>-1</td>
<td>0.7</td>
<td>-1.5</td>
<td>0</td>
</tr>
<tr>
<td>Read</td>
<td>0.2</td>
<td>0.7</td>
<td>-1.5</td>
<td>0</td>
</tr>
<tr>
<td>Hold</td>
<td>0</td>
<td>-1</td>
<td>-1.5</td>
<td>0</td>
</tr>
</tbody>
</table>
Contact processing with shared litho steps

- Similar to Toshiba BiCS scheme [VLSI 2007]

Fig. 3 (a) Birds-eye view of BiCS flash memory, (b) Top down view of BiCS flash memory array.