eDRAM to the Rescue

- Why eDRAM
  - 1/3 Area
  - 1/5 Power
  - SER 2-3 Fit/Mbit vs 2k-5k for SRAM
  - Smaller is faster

What’s Next?
Integrating DRAM and Logic

- Integrate with Logic without impacting logic Performance, Reliability or Yields

  - The Deep Trench process is intrinsically logic friendly
    - Capacitor fabricated first
    - No perturbation of the remaining logic flow

- Significant Process & Test know-how in DRAMs using deep trench technology

Over six generations, embedded DRAM has adapted to logic technology resulting in simpler processes and significantly higher cell performance
Technology Innovation – Development of SOI DRAM cell

- Technology:
  Use the Buried oxide to simplify the process & reduce parasitics – half the cost of bulk eDRAM
  Scale the pass transistor for higher performance

- Design
  Address retention through concurrent refresh
  Ultra short bitlines with direct sense architecture
eDRAM Performance Advantage ?
Itanium® 2 Processor 9M Highlights

- 592M transistors
- 432mm² die size
- 9MB on-die L3 cache
- 1.7GHz at 1.35V
- 6.4GB/s 400MT/s 4-way bus interface
- Plug-in compatible with existing platforms
- Extensive RAS, DFT and DFM features

Largest microprocessor transistor count and on-die cache

ISSCC 04 – Paper 27.3
ISSCC’04 CPU

CPU with DRAM
(4x denser/2x slower)

chip size

432mm² → 245mm²
(43% smaller)

*39% decrease in wire delay to furthest L3-cache subarray

L3 Latency (normalized to 20 cycles)

<table>
<thead>
<tr>
<th></th>
<th>L3-Tag</th>
<th>L3-Cache</th>
<th>Wire Delay</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRAM</td>
<td>5 cycles</td>
<td>5 cycles</td>
<td>10 cycles</td>
<td>20 cycles</td>
</tr>
<tr>
<td>eDRAM</td>
<td>5 cycles</td>
<td>10 cycles</td>
<td>6 cycles</td>
<td>21 cycles</td>
</tr>
</tbody>
</table>
**eDRAM Size/Latency Advantage**

### 45nm eDRAM vs. SRAM Latency

<table>
<thead>
<tr>
<th>Memory Block Size</th>
<th>eDRAM Total Latency</th>
<th>SRAM Total Latency</th>
<th>eDRAM Wire/Repeater Delay</th>
<th>SRAM Wire/Repeater Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>1Mb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4Mb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8Mb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16Mb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32Mb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64Mb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- eDRAM has lower latency!
eDRAM in IBM systems

- eDRAM used on the MCM till 65 nm in p systems
- eDRAM integrated with power PC in BlueGene usic ASICs flow
- In 45 nm we integrated eDRAM with the processor in SOI
Integrating large amounts of low-latency memory is a huge challenge for modern multi-core processor design.

- 567mm² Technology: 45nm lithography, Cu, SOI, eDRAM
- 1.2B transistors
  - Equivalent function of 2.7B
  - eDRAM efficiency
- Eight processor cores
  - 12 execution units per core
  - 4 Way SMT per core
  - 32 Threads per chip
  - 256KB L2 per core
- 32MB on chip eDRAM shared L3
- Dual DDR3 Memory Controllers
  - 100GB/s Memory bandwidth per chip sustained
- Scalability up to 32 Sockets
  - 360GB/s SMP bandwidth/chip
  - 20,000 coherent operations in flight
- Advanced pre-fetching Data and Instruction
- Binary Compatibility with POWER6
Deep Trench Capacitor Advantage

- DT decoupling capacitance: 200fF/μm²
  - Total cap of 300pF
  - 20X efficiency of MOSCAP
- DT connected to VCS grid
  - Grid on Metal 4 (low R)
- No performance impact
  - 200mV charge up in < 150ps
Stacked DRAMs to increase capacity without increasing power

- I/O ckts drive considerable power requirements
- These do not need to be duplicated on multiple DRAMs
- Master – Slave approach to share I/Os, PLLs etc
- Enabled by TSVs

- Must compete with conventional scaled 3D chip

Kang et al, ISSCC 2009
Getting Power in and out of high-end Processors is the challenge

- 150 – 200 W versus a few watts
- Need to deliver about 200 A to die uniformly
  - Perimeter Wire Bond Inadequate
  - High Power Die needs uniform power delivery across the Die (Grid)
- Need to heat sink the die
Structural Schematic of 3D integration of cache and processor – Why TSVs

P die: 11 level build
Die-die interconnect @~50 \( \mu \text{m} \) pitch
M die: 6LM, TSVs, lots of decaps

C4s @186 \( \mu \text{m} \) pitch

Heat Sink

Processor Die

Thinned memory Die with TSVs

TSVs

Laminate

TSV: Through Silicon Via

C4s: bumps