Memory Performance Tutorial
Hot Chips 16

J. Thomas Pawlowski
Sr. Dir. Architecture Development & Senior Fellow
NetCom Group, Micron Technology, Inc.

Agenda

- Objectives and background
- Soft Error Rate discussion
- Brief discussion of nonvolatile cells, devices, and device characteristics
- Volatile memory cell types, characteristics, general operation
- Specific device operation
  - Emphasis on factors for performance comparison
  - DRAM: DDR, DDR2
  - New DRAM: GDDR3, FCRAM II, RLDRAM II, RAMBUS XDR
- Performance comparison of most memory types
  - Many operational scenarios
- Analysis and conclusions
Disclaimers

- All information presented herein is from sources not requiring an NDA. More details are available from the individual manufacturers.
- All road maps shown use estimated dates only, subject to change, consult individual manufacturers for updates.
- Many devices mentioned have trademarked names:
  - QDR™ and derivatives are trademarked by Cypress, et al
  - FCRAM™ and derivatives are trademarked by Fujitsu
  - RLDRAM™ and derivatives are trademarked by Infineon
  - RAMBUS™ and RAMBUS XDR™ trademarked by Rambus
- No statements made herein are to be taken as design advice
  - Regard them as generalizations to assist your understanding
  - If you have specific design issues, contact your friendly Micron support personnel
- Do not design any systems based on this information. Consult manufacturer datasheets.

Soft Error Rate - Definitions

- Hard error
  - An error induced by a device fault
  - Data is lost and can no longer be stored at that location
- Soft error
  - A random error induced by an event which corrupts data.
  - The device is not damaged and can correctly store data when written again.
- Single event upset - SEU
- Multiple event upset - MEU
- Soft error rate - SER
- Failure in time - FIT
  - 1 failure per $10^9$ device operating hours
# Causes of SER

- **Radiation** (α particles, B10 fission, cosmic rays)
- **α particles** (helium nuclei)
  - Short mean path length in silicon
    - 4 MeV - 18µm
    - 9 MeV - 70µm
  - Was virtually a non-issue on Micron devices due to material elimination
  - Typical SRAM SER due to alpha is 5 FITs/ Mb at 1.8 V at Micron
- **B10 fission**
  - Eliminate or shield against BPSG
- **Cosmic rays**
  - > 99 percent neutrons
  - Use ECC or else errors will be observed

---

# Atmospheric Filtering and Secondary Effects

- **Primary particles interact in atmosphere**
  - Produce many generations of other particles
    - Neutrons, electrons, muons, protons (< 1 GeV)
    - Neutrons and protons have about 40x the impact in silicon
- **At sea level, fairly equal mix of the 4 particles**
- **Neutron flux increases with altitude**
  - \( F \sim A^N \) where \( A \approx 1.25 \) and \( N \) is altitude in 1000 feet (e.g. \( N = 3 \) @ 3,000 feet)
    - ~2x at 3,000 feet, 3x at 5,000 feet, 9x at 10,000 feet
Atmospheric Filtering and Secondary Effects (continued)

- 90 percent of humans live at 1,700 feet elevation or lower
- 95 percent live at 4,000 feet or lower
  - Luckily, most equipment operates in lower neutron flux
- Neutron flux is also a function of longitude/latitude
  - <2x difference based on geographic location (see JESD89 spec)
  - Singapore is sea level and has lowest flux based on location

Nonvolatile Memory – NOR Flash

- NOR Flash memory cells
  - Common gate and drain, separate source, field effect transistors (FETs) with a floating gate.
  - Floating gate holds charge
  - When everything is biased correctly, will see the 1 or 0 caused by the floating gate, hence see 1 bit per cell.
  - Typically $8.5F^2 \sim 15F^2$ cell size ($F =$ process minimum feature size)
Nonvolatile Memory – NOR Flash
(continued)

- NOR Flash memory multilevel cells
  - Same concept, but with tighter charge control and tighter sensing resolution, more overhead die area
  - 2 bits per cell, same cell size, equivalent result is $4.25F^2 \sim 7.5F^2$ cell size
  - Longer read and write times

Nonvolatile Memory – NAND Flash

- NAND Flash memory cells
  - Common gate, daisy-chained source to drain
  - Floating gate holds charge as with NOR
  - Read or write current passes through chain of devices
  - Needs to operate on larger data chunks
  - Typically $4F^2$ cell size, but 1 extra device per 16, result is equivalent to $4.25F^2$ cell size
Nonvolatile Memory – NAND Flash (continued)

- NAND Flash memory multilevel cells
  - Same concept, but with tighter charge control and tighter sensing resolution, more overhead die area
  - 2 bits per cell, same cell size, equivalent result is $2.13F^2$ cell size

---

Flash Memory Cell Comparison

<table>
<thead>
<tr>
<th></th>
<th>NAND</th>
<th>AND</th>
<th>NOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Array</td>
<td><img src="image1" alt="NAND Array" /></td>
<td><img src="image2" alt="AND Array" /></td>
<td><img src="image3" alt="NOR Array" /></td>
</tr>
<tr>
<td>Layout</td>
<td><img src="image4" alt="NAND Layout" /></td>
<td><img src="image5" alt="AND Layout" /></td>
<td><img src="image6" alt="NOR Layout" /></td>
</tr>
<tr>
<td>Cross-section</td>
<td><img src="image7" alt="NAND Cross-section" /></td>
<td><img src="image8" alt="AND Cross-section" /></td>
<td><img src="image9" alt="NOR Cross-section" /></td>
</tr>
<tr>
<td>Cell size</td>
<td>$4F^2$</td>
<td>$8F^2$</td>
<td>$10F^2$</td>
</tr>
</tbody>
</table>

- Cell size does not include overhead of extra “chain” device.
- All technologies capable of multiple bits per cell.
Basic NAND/ NOR Comparison

- **NAND**
  - **Advantages:**
    - Fast writes
    - Fast erases
  - **Disadvantages:**
    - Slow random access
    - Byte writes difficult
  - **Applications:**
    - File (disk) applications
    - Voice, data, video recorder
    - Any large sequential data

- **NOR**
  - **Advantages:**
    - Random access
    - Byte writes possible
  - **Disadvantages:**
    - Slow writes
    - Slow erase
  - **Applications**
    - Replacement of EPROM
    - Execute directly from nonvolatile memory

Generic Nonvolatile Memory Comparison (Today)

<table>
<thead>
<tr>
<th></th>
<th>NOR Flash</th>
<th>NAND Flash</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applications</td>
<td>Code, data</td>
<td>Mass storage</td>
</tr>
<tr>
<td>Future applications</td>
<td>MLC: mass storage</td>
<td>Code and data</td>
</tr>
<tr>
<td>Density range</td>
<td>Up to 512 Kb</td>
<td>Up to 4Gb</td>
</tr>
<tr>
<td>READ latency</td>
<td>60ns-120ns</td>
<td>25µs</td>
</tr>
<tr>
<td>Max Read bandwidth</td>
<td>41 MB/s-112 MB/s (16b)</td>
<td>40 MB/s (16 bus)</td>
</tr>
<tr>
<td>Max Write bandwidth</td>
<td>0.25 MB/s</td>
<td>5MB/s</td>
</tr>
<tr>
<td>Erase time</td>
<td>400ms (128KB blk)</td>
<td>2ms (128KB block)</td>
</tr>
<tr>
<td>Read device current</td>
<td>1.6x</td>
<td>1x</td>
</tr>
<tr>
<td>Write device current</td>
<td>3x</td>
<td>1x</td>
</tr>
</tbody>
</table>

Note that there is a wide variation among competing devices.
Micron Flash Memory Comparison

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>NAND Flash MT29F2G08/16A</th>
<th>NOR (Q-Flash) MT28F128J3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Access Read</td>
<td>25μs (first byte)</td>
<td>120ns</td>
</tr>
<tr>
<td></td>
<td>50ns for remaining 2111</td>
<td></td>
</tr>
<tr>
<td>Sustained Read Speed (sector basis)</td>
<td>16 MB/s (x8) or 32 MB/s (x16)</td>
<td>20.5 MB/s (x8) or 41 MB/s (x16)</td>
</tr>
<tr>
<td>Random Write Speed</td>
<td>300μs/2112 bytes</td>
<td>180μs/32 bytes</td>
</tr>
<tr>
<td>Sustained Write Speed (sector basis)</td>
<td>1.5MB/s</td>
<td>0.178MB/s</td>
</tr>
<tr>
<td>Erase block size</td>
<td>128KB</td>
<td>128KB</td>
</tr>
<tr>
<td>Erase time per block (typ)</td>
<td>2ms</td>
<td>750ms</td>
</tr>
</tbody>
</table>

- NOR Flash memory is ideal for direct code execution (boot code)
- NAND Flash memory is ideal for file storage (e.g. data or image files. If code is stored, it must be shadowed to RAM first, as in a PC).

Volatile Memory Cell Types

- 1 Transistor, 1 Capacitor (1T-1C)
  Dynamic random access memory (DRAM) cell
  - Used on all production DRAM
  - Used on SRAM replacement devices
    - e.g. Micron CellularRAM™

- 3T DRAM cell
  - Not used much
  - Easy way to make DRAM in logic process
  - Some activity for high-performance SRAM replacement

- 6T static random access memory (SRAM) cell
SRAM Cell

- Reads are not destructive
- Any lost charge is restored by the P-FET pull-up devices
- No refresh needed

SRAM cell layout requires 3 horizontal lines (VDD, Vss, word line) and 2 vertical lines (bit, bit#)

SRAM Cell Scaling

- Spacing between features increases on each process node
  - Required to deal with defects

Graphs showing size as square microns and size as multiple of $F^2$.
**SRAM Die Size**

- Depends on many factors, but...
- Can be calculated as \( n \times A / \text{eff} \)
  - \( n \) = bit count, \( A \) = area of 1 bit, \( \text{eff} \) = die efficiency
  - (e.g. 0.6 per unit die area is array bits, remainder is overhead)
- With all other factors remaining constant:
  - Die efficiency improves as density increases,
  - Worsens as I/O count increases,
  - Worsens as logic/complexity increases
- Any memory die size can be similarly calculated

**SRAM Soft Error Rate**

- Typical SRAM SER = 10K FITs / Mb
  - (Sea level, Singapore, 0.13 µm process, 1.8V core)
- Trend is that SER is growing with each new process node
  - Cell grows as function of minimum feature size
    - Target area to capacitance ratio increases
  - Voltage continues to reduce
    - Less cell charge
- Modeling suggests 20K-50K FITs/Mb on 90nm, 1.2V core
  - ~100K FITs for 36Mb SRAM on 90nm
- ECC is mandatory, or replace with DRAM
SRAM SER
18Mb 0.16μm QDR SER vs. Voltage

FITs
Extrapolated FITs

Voltage (V)

FITs normalized to Singapore

1T-1C DRAM

Reads are destructive
Bit value must be written back when read is done.
C will discharge through leakage paths
To restore charge perform read, elevate bit to full voltage and drive current into C

1T 1C DRAM cell layout requires
- 1 horizontal line (Word line)
- 1 vertical line (bit line)
- Vss (or a reference node)
1T-1C DRAM Charge Redistribution

- During read, charge from C is shared with $C_{\text{bit line}}$.
- If read begins with $V_{\text{bit line}} = V_{\text{DD}}/2$:
  - $V_{\text{bit line}} = (V_C - V_{\text{DD}}/2) \times \left[ C / (C + C_{\text{bit line}}) \right]$
  - $V_{\text{bit line}}$ is relatively small, e.g. 200mV
    - Need to amplify the voltage before sending the result off chip – sense amplifiers.

Whole DRAM Array

- Example: 16M words, 16 bit wide bus
  - $n = \log_2 16M = 24$, $k = 8$, $m = \log_2 16 = 4$
DRAM Cell Size

- Commercial DRAM production uses two different cell types
  - Both are 1T-1C but the capacitor is formed differently
- Trench capacitor is formed first
  - Like digging a trench but with a high wall surface area
  - Implants added on top to form N and P FETs, contacts, metal lines, etc.
  - Infineon DRAM, IBM eDRAM are examples: cell size is 8F^2
- Stacked capacitor is formed with or after other chip elements
  - More vertical structure also with high wall surface area
  - Most other DRAM manufacturers do it this way, e.g. Micron, Samsung, Elpida
  - Cell size is 8F^2 for all except new Micron DRAMs

Industry Roadmap

DRAM Cell Size Scaling

4F^2 theoretical limit for single bit per cell

Note that there is a large difference between mass production starts versus "sweet spot" production.
Micron Jump to $6F^2$ Cell
Ahead of the Curve

- Reduces memory array size by 25 percent
- Same cell capacitance
- Shortens bit lines and/or word lines
  - Less line capacitance means lower energy requirement, faster speed

DRAM Soft Error Rates

- 1T-1C DRAM has small cell target area and high cell capacitance
- Target area with each new process node is smaller
  - Area trend reduces SER with each new process
- Voltage is reduced every few process nodes
  - Voltage reduction trend increases SER due to reduction in stored charge
- Capacitance tends to stay about the same with each new process node
- Net result is reduction in SER with each new process node
- Typical SER is less than 1 FIT/ Mb
  - e.g. 256Mb DRAM on 1.8V, 110nm process 100 FITs ~200 FITs
  - SER per component remains about the same even though density increases with time
  - Should expect lower SER on $6F^2$ than $8F^2$
  - No significant difference between trenched and stacked cell DRAMs for same cell size
DDR II and QDR II SRAM

Motivation
- Networking data structures
- Improved bus efficiency
- Improved pin bandwidth
- Increased request rates over previous SRAMs

Networking SRAM Architectures

- Double data rate (DDR) I/O interface
- 2-word burst and 4-word burst
- Quad data rate: QDR and QDR II
  - Separate DIN and QOUT buses
  - Separate/concurrent read and write ports
  - 2 reads (1 data pair) and 2 writes (1 data pair) per clock cycle
Networking SRAM Architectures (continued)

- Double date rate: DDR and DDR II common I/O (CIO)
- Common DQ bus, common R/W bus
  - 1 request per clock cycle, stall for read-to-write transitions
- Double data date: DDR2 separate I/O (2-word burst only)
  - Separate DIN and QOUT buses
  - Separate/ non-concurrent read and write ports
  - 1 request per clock cycle, no stalls
- All use HSTL I/O (1.4V - 1.9V)
**DDR II SIO R/W 2-Word Burst**

- **K**, **K#**, **R/W#**, **LD#**
- **Address**: A, B, C, D
- **Data-In**: D(A) +1, D(D) +1
- **Data-Out**: Q(B) +1, Q(C) +1

---

**Relative Bandwidth per Pin of All Network SRAMs at 250 MHz Clock**

- **READ:WRITE Ratio**
- **Relative Bandwidth per Pin**

*Note: SDR represents bandwidth per pin for NoBL, ZeroBL, NiRAM, and ZBT devices*
DDR II/ QDR II SRAM Performance Models

- DDR II CIO BL2 or BL4
  - R- R, W- W, W- R: BL/ 2 for next command
  - R- W: BL/ 2+ 1 cycle < 200 MHz, BL/ 2+ 2 > = 200 MHz
- DDR II SIO BL2
  - R- R, W- W, R- W, W- R: 1 cycle for next command
- QDR II BL2
  - Reads on K rising edges, writes on K falling edges, no restrictions
- QDR II BL4
  - R- W, W- R: 1 cycle
  - W- W, R- R: 2 cycles

Note: R = Read, W = Write, BL = Burst Length
No lower frequency limit

QDR III and DDR III

- Still being defined
- Target 250 MHz- 500+ MHz clock frequency
  - QDR III 2- word burst
  - QDR III 4- word burst
  - DDR III 2- word burst
- 1.2V core
QDR III and DDR III (continued)

- 18Mb - 288Mb densities
- 9 bits, 18 bits, and 36 bits wide
- Other design details available from suppliers under NDA
  - My guesses: 4- cycle latency, VTT mid- rail termination scheme like RLDRAM II, increased ratio of echo clocks to outputs, increased ratio of input clocks to inputs
  - For today’s performance analysis, latency guess is irrelevant; multiple identical addresses can be in flight simultaneously

---

QDR Comparison

<table>
<thead>
<tr>
<th>Feature</th>
<th>QDR</th>
<th>QDRII</th>
<th>QDR III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>BL2: 166 MHz, BL4: 200 MHz</td>
<td>BL2: 250 MHz, BL4: 333 MHz</td>
<td>BL2, BL4: 250-500 MHz</td>
</tr>
<tr>
<td>Data Valid</td>
<td>1.4ns @ 166MHz</td>
<td>1.9ns @ 166MHz</td>
<td>0.98ns @ 333MHz</td>
</tr>
<tr>
<td>Initial Latency</td>
<td>1.4 Cycles</td>
<td>1.6 Cycles</td>
<td>4? Cycles</td>
</tr>
<tr>
<td>Clocks</td>
<td>No Echo CLKs</td>
<td>Echo CLKs</td>
<td>Echo CLKs</td>
</tr>
<tr>
<td>Density</td>
<td>9/18/36Mb</td>
<td>18/36/72Mb+</td>
<td>36-144Mb+</td>
</tr>
<tr>
<td>Power Supply</td>
<td>2.5V</td>
<td>1.8V</td>
<td>1.2V</td>
</tr>
</tbody>
</table>
**DDR III/ QDR III SRAM Performance Models**

- DDR III CIO BL 2
  - R- R, W- W, W- R: BL/ 2 for next command
  - R- W: BL/ 2 + 2
- QDR III BL 2
  - Reads on K rising edges, Writes on K falling edges
  - Write address not = read address during same cycle
- QDR III BL4
  - R- W, W- R: 1 cycle
  - W- W, R- R: 2 cycles
- Note that this is speculative
  - Unknown if there are data dependencies in the final products

**QDR / DDR SRAM Summary**

- QDR SRAMs are optimized for systems with short-term, balanced READ and WRITE operations
- DDR CIO SRAMs are optimized for data streaming operations or READ/ WRITE unbalanced systems
- DDR SIO SRAMs are optimized for one address/ clock, 2-word burst systems
- High data availability, high cost, high SER
- Low density
- Extremely simple performance models
• Some notes on single data rate synchronous DRAM
• DDR SDRAM (2.5V core)
• DDR2 SDRAM (1.8V core)
• RLDRAM II (1.8V core)
• FCRAM I, II and II+ (2.5V and 1.8V cores)
• GDDR3 (1.8V core)
• Rambus XDR (1.8V core)

**Single Data Rate SDRAM**

<table>
<thead>
<tr>
<th>Density</th>
<th>Widths</th>
<th>Frequencies</th>
<th>Packages</th>
</tr>
</thead>
<tbody>
<tr>
<td>64 Mb</td>
<td>4, 8, 16</td>
<td>133-183 MHz</td>
<td>54 TSOP</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td></td>
<td>86 TSOP</td>
</tr>
<tr>
<td>128 Mb</td>
<td>4</td>
<td>133-166 MHz</td>
<td>54 TSOP</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>54 TSOP, 60 FBGA</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td></td>
<td>54 TSOP, 60 FBGA</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td></td>
<td>86 TSOP, 90 FBGA</td>
</tr>
<tr>
<td>256 Mb</td>
<td>4</td>
<td>133-166 MHz</td>
<td>54 TSOP</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>54 TSOP, 60 FBGA</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td></td>
<td>54 TSOP, 60 FBGA</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td></td>
<td>86 TSOP, 90 FBGA</td>
</tr>
<tr>
<td>512 Mb</td>
<td>4</td>
<td>133-183 MHz</td>
<td>54 TSOP</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>54 TSOP</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td></td>
<td>54 TSOP</td>
</tr>
</tbody>
</table>
SDRAM - Refresh

- The refresh rate is dependent on the number of rows in the device
  - Each refresh command refreshes a single row
  - Addressing is handled by an internal refresh controller; address bits are a “don’t care” during an auto-refresh command
  - $t\text{REF}$ represents the maximum time in which all rows must be refreshed at least once – typically 64ms for SDR and DDR SDRAM
  - $t\text{REFI}$ represents the average periodic refresh time – $t\text{REFI} / \#\text{of rows}$ (8K refresh: 64ms/8K = 7.8125µs, 4K refresh: 15.625µs)
  - $t\text{REFC}$ represents the absolute maximum time you can go without issuing any refresh commands – Micron DDR devices allow 9 $t\text{REFI}$ cycles (9 x $t\text{REFI}$), JEDEC standard is 8 $t\text{REFI}$ cycles
  - $t\text{RFC}$ represents the auto refresh command period. (CKE must remain HIGH during this time for all MICRON devices)

2.5V DDR SDRAM

- Motivation
  - Increase bandwidth per data pin without significant cost increase
  - 266 MB/s – 400 Mb/s data rates
- Inspired by original DDR SDRAM
Double Data Rate SDRAM

- Micron 2.5V devices available
  - 128Mb – 1Gb
  - 4-bit, 8-bit and 16-bit wide data bus
  - 66-pin TSOP on all devices, 60-ball FBGA on some devices
  - 133 MHz - 200 MHz clock, 266 Mb/s - 400 Mb/s per data pin
    - Check specific devices for frequency availability
    - Micron leads all suppliers in product breadth

DDR SDRAM Block Diagram
4 x 16 Meg x 8
Basic DDR SDRAM Commands

<table>
<thead>
<tr>
<th>NAME (FUNCTION)</th>
<th>CS#</th>
<th>RS#</th>
<th>CAS#</th>
<th>WE#</th>
<th>ADDR</th>
<th>NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>DESELECT (NOP)</td>
<td>H</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>9</td>
</tr>
<tr>
<td>NO OPERATION (NOP)</td>
<td>L</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>X</td>
<td>9</td>
</tr>
<tr>
<td>ACTIVE (Select bank and activate row)</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>H</td>
<td>Bank/Row</td>
<td>3</td>
</tr>
<tr>
<td>READ (Select bank and column, and start READ burst)</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>H</td>
<td>Bank/Col</td>
<td>4</td>
</tr>
<tr>
<td>WRITE (Select bank and column, and start WRITE burst)</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>L</td>
<td>Bank/Col</td>
<td>4</td>
</tr>
<tr>
<td>Burst TERMINATE</td>
<td>L</td>
<td>H</td>
<td>H</td>
<td>L</td>
<td>X</td>
<td>8</td>
</tr>
<tr>
<td>PRECHARGE (Reactivate row in bank or banks)</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>Code</td>
<td>5</td>
</tr>
<tr>
<td>AUTO REFRESH or SELF REFRESH</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>M</td>
<td>X</td>
<td>6, 7</td>
</tr>
<tr>
<td>LOAD MODE REGISTER</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>Op-Code</td>
<td>2</td>
</tr>
</tbody>
</table>

NOTE:
1. CKE is HIGH for all commands shown except SELF REFRESH.
2. BA0-BA1 select either the mode register or the extended mode register (BA0 = 0, BA1 = 0 select the mode register; BA0 = 1, BA1 = 0 select extended mode register; other combinations of BA0-BA1 are reserved). A0-A12 provide the op-code to be written to the selected mode register.
3. BA0-BA1 provide bank address and A0-A12 provide row address.
4. BA0-BA1 provide bank address; A0-A12 provide column address, where i=9 for x16, i=9,11 for x32, and i=9,11,12 for x48.
5. A10 HIGH enables the auto precharge feature (non-persistent), and A10 LOW disables the auto precharge feature.
6. A10 HIGH: all banks are precharged and BA0-BA1 are "Don't Care."
7. The command is AUTO REFRESH if CKE is HIGH; SELF REFRESH if CKE is LOW.
8. Internal refresh counter controls row addressing: for within the Self Refresh mode all inputs and i/0s are "Don't Care" except for CKE.
9. Applies only to read bursts with auto precharge disabled; this command is undefined (and should not be used) for read bursts with auto precharge enabled and for write bursts.
10. DESELECT and NOP are functionally interchangeable.

---

DDR: Read with Autoprecharge

---
DQSCK is the DQS output window relative to CK.

- DQs transitioning after DQS transition define DQSQ window.
- All DQs transition by DQSQ after DQS edge, regardless of AC.
- AC is the DQ output window relative to CK.
- LZ (MIN), AC (MIN) and HZ (MIN) are the first valid signal transition.
- LZ (MAX), AC (MAX) and HZ (MAX) are the latest valid signal transition.
- The DLL is used to realign the data strobe to the CK/CK# crossing.
- Data is edge aligned to the DRAM clock.

**DDR: DQS Preamble and Postamble**

- Preamble (RPRE) provides a timing window for the receiving device to enable its data capture circuitry while a known level is valid on the strobe signal.
  - avoids false triggers of the capture circuit
- Postamble (RPST) represents the DQS LOW time following the last transition.
DDR: Read Data Valid Window

- $t_{DQS}$ represents DQ to DQ skew in relation to data strobe.
- $t_{DQS}$ is derived at each DQS clock edge and is not cumulative over time.
- $t_{HP}$ is the lesser of $t_{CL}$ or $t_{CH}$ clock transition.
- The data valid window is derived for each DQS transitions and is defined as $t_{QH} - t_{DQS}$.

DDR: Write with Autoprecharge

- $t_{RCD}$ = Active to read or write delay.
- $t_{WR}$ = Write recovery.
- $t_{RP}$ = Precharge time.
- $t_{RAS}$ = Active to precharge.
- $t_{RC}$ = Active to active in same bank.
• DQSS = Write command to first DQS latching transition
• DSH (MIN) generally occurs during DQSS (MIN).
• DSS (MIN) generally occurs during DQSS (MAX).
• WRITE command issued at T0.

> Burst length = 4 in the cases shown
  - Applies for bursts of 8 and full page as well
  - If the burst length is 2, the BST command shown can be NOP.
> Shown with nominal tAC, tDQCK, and tDQS.
> tRAS (MIN) still applies
DDR: Write Followed by Read

- READ command can be earlier if to a different device
- If the READ command applies to a different row in the same bank the READ command would come after \(^1\)WR, \(^1\)RP, and \(^1\)RCD

DDR SDRAM Performance Model

<table>
<thead>
<tr>
<th>Clock Frequency</th>
<th>200 MHz</th>
<th>166 MHz</th>
<th>133 MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock period</td>
<td>T</td>
<td>5 ns, 7.5 max</td>
<td>6 ns, 13 max</td>
</tr>
<tr>
<td>Cas Latency</td>
<td>CL</td>
<td>3</td>
<td>2.5</td>
</tr>
<tr>
<td>ACT to same bk ACT</td>
<td>tRC</td>
<td>55 ns, 7.5 max</td>
<td>60 ns, 10 ns</td>
</tr>
<tr>
<td>ACT to R or W</td>
<td>tRCD</td>
<td>15 ns, 3 ns</td>
<td>15 ns, 3 ns</td>
</tr>
<tr>
<td>ACT to PRE</td>
<td>tRAS</td>
<td>40 ns, 8 ns</td>
<td>42 ns, 7 ns</td>
</tr>
<tr>
<td>PRE Period</td>
<td>tRP</td>
<td>15 ns, 3 ns</td>
<td>15 ns, 3 ns</td>
</tr>
<tr>
<td>W to PRE</td>
<td>tWR</td>
<td>15 ns, 3 ns</td>
<td>15 ns, 3 ns</td>
</tr>
<tr>
<td>W to R</td>
<td>tWTR</td>
<td>1 ns, 1 ns</td>
<td>1 ns, 1 ns</td>
</tr>
<tr>
<td>ACT to diff bk ACT</td>
<td>tRRD</td>
<td>10 ns, 2 ns</td>
<td>12 ns, 2 ns</td>
</tr>
<tr>
<td>Auto RFSH period</td>
<td>tRFC</td>
<td>70 ns, 14 ns</td>
<td>72 ns, 12 ns</td>
</tr>
</tbody>
</table>

- 7.8125\(\mu\)s refresh period used (512Mb device)
- Other CAS latencies are possible but aren’t used in performance model
- Some calculations require CL rounded up to next integer
- 4 banks
- T represents 1 clock cycle
- Clock period sweet spots: 5 ns, 6 ns, 7.5 ns
- Recall the disclaimer!
## DDR SDRAM Performance Model

<table>
<thead>
<tr>
<th>Command</th>
<th>To Same Bank</th>
<th>To Different Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>W to R</td>
<td>$1 + \frac{BL}{2} + T_{WTR}$</td>
<td>BL/2</td>
</tr>
<tr>
<td>W to W</td>
<td>BL/2</td>
<td>BL/2</td>
</tr>
<tr>
<td>W to PRE</td>
<td>$1 + \frac{BL}{2} + T_{WR}$</td>
<td>1</td>
</tr>
<tr>
<td>W to ACT</td>
<td>W to PRE + T_{RP}</td>
<td>1</td>
</tr>
<tr>
<td>R to R</td>
<td>BL/2</td>
<td>BL/2</td>
</tr>
<tr>
<td>R to W</td>
<td>CL(rounded up) + BL/2</td>
<td>CL(rounded up) + BL/2</td>
</tr>
<tr>
<td>R to PRE</td>
<td>BL/2</td>
<td>1</td>
</tr>
<tr>
<td>R to ACT</td>
<td>BL/2 + T_{RP}</td>
<td>1</td>
</tr>
<tr>
<td>ACT to ACT</td>
<td>TRC</td>
<td>T_{RDD}</td>
</tr>
<tr>
<td>ACT to R or W</td>
<td>TRCD</td>
<td>1</td>
</tr>
<tr>
<td>ACT to PRE</td>
<td>TRAS</td>
<td>1</td>
</tr>
</tbody>
</table>

?Recall the disclaimer: do not base controller design on this information, consult manufacturer data sheets for latest and most accurate information

---

### 1.8V DDR2 SDRAM

- **Motivation**
  - Increase frequency of DDR SDRAM without significant cost increase
  - 400 Mb/s – 800 Mb/s data rates
1.8V DDR2 SDRAM Availability

- Micron devices available
  - 256Mb – 2Gb
  - 4-bit, 8-bit and 16-bit wide data bus
- 256Mb: 4, 8 bit in 60 ball FBGA, x16 in 80 FBGA
- 512Mb – 2Gb in 92 ball FBGA
- 200 MHz- 333 MHz clock, 400 Mb/s- 667 Mb/s per data pin
  - Check specific devices for frequency availability
  - Micron leads all suppliers in product breadth

Basic Changes From DDR to DDR2

- Increased frequency spawns many changes
  - Lower core and I/O voltages
  - On-die termination (ODT)
  - Off-chip driver characteristics and calibration
  - 4n prefetch (instead of 2n, sets minimum burst length)
    - To keep column performance (cost) in check
    - Result is elimination of 2-word burst
  - Additive CAS latency
    - Since \( T_{\text{RR}} \) is larger, needed mechanism to clean up request issuing
  - Bank count increase, made possible by density, needed for performance
Feature Overview: DDR vs. DDR2

<table>
<thead>
<tr>
<th>Feature/Option</th>
<th>DDR</th>
<th>DDR2</th>
<th>DDR2 Advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Package</td>
<td>TSOP (66-pin)</td>
<td>FBGA only</td>
<td>Better electrical performance and speed</td>
</tr>
<tr>
<td>Voltage</td>
<td>2.5V / 2.5V I/O</td>
<td>1.8V / 1.8V I/O</td>
<td>Reduces memory system power demand</td>
</tr>
<tr>
<td>Densities</td>
<td>64Mb – 1Gb</td>
<td>256Mb – 4Gb</td>
<td>High-density components enable large memory subsystems</td>
</tr>
<tr>
<td>Internal banks</td>
<td>4</td>
<td>4 and 8</td>
<td>1Gb and higher will have 8 banks for better performance</td>
</tr>
<tr>
<td>Pre-fetch (MIN write burst)</td>
<td>2</td>
<td>4</td>
<td>Provides reduced core speed dependency for better yields</td>
</tr>
<tr>
<td>Speed (data pin)</td>
<td>200, 266, 333, 400 Mb/s</td>
<td>400, 533, 667 Mb/s</td>
<td>Migration to higher speed I/O</td>
</tr>
<tr>
<td>Read Latency (CAS latency)</td>
<td>CL + AL</td>
<td>CL + AL = 3, 4, 5</td>
<td>Eliminating one half clock settings helps speed internal DRAM logic and improves yields</td>
</tr>
<tr>
<td>Additive Latency (Posted CAS)</td>
<td>N/A</td>
<td>AL options 0, 1, 2, 3, 4</td>
<td>Mainly used in server applications to improve command bus efficiency</td>
</tr>
</tbody>
</table>

Feature Overview (Continued)

<table>
<thead>
<tr>
<th>Feature/Option</th>
<th>DDR</th>
<th>DDR2</th>
<th>DDR2 Advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>WRITE Latency</td>
<td>1 clock</td>
<td>READ Latency - 1</td>
<td>Improves bus efficiencies</td>
</tr>
<tr>
<td>Termination</td>
<td>Motherboard parallel to VTT</td>
<td>DRAM on-die termination (ODT) optional on motherboard</td>
<td>ODT for both memory and controller improves signaling, and reduces system cost</td>
</tr>
<tr>
<td>Burst Lengths</td>
<td>2, 4, 8</td>
<td>2, 4, 8</td>
<td>Improved system timing margin by reduced strobe crosstalk</td>
</tr>
<tr>
<td>Data Strobes</td>
<td>Single ended</td>
<td>Differential or single ended</td>
<td>Modules are the same length, with added pins</td>
</tr>
<tr>
<td>Modules</td>
<td>184-pin unbuffered registered 200-pin SODIMM 172-pin MicroDIMM</td>
<td>240-pin unbuffered registered 200-pin SODIMM 244-pin MiniDIMM 214-pin MicroDIMM</td>
<td></td>
</tr>
</tbody>
</table>
**DDR- DDR2 Differences: Page Size**

- Page size is the minimum number of columns accessed with a single ACTIVATE command = # columns x bus width

<table>
<thead>
<tr>
<th>Density</th>
<th>DDR</th>
<th>DDR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>256Mb</td>
<td>1KB x 4</td>
<td>1KB x 4</td>
</tr>
<tr>
<td>512Mb</td>
<td>2KB x 4</td>
<td>1KB (x4, x8), 2KB (x16) x4</td>
</tr>
<tr>
<td>1Gb</td>
<td>2KB x 4</td>
<td>1KB (x4, x8), 2KB (x16) x8</td>
</tr>
<tr>
<td>2Gb</td>
<td>-</td>
<td>1KB (x4, x8), 2KB (x16) x8</td>
</tr>
</tbody>
</table>

**DDR2 Differences: Activate Commands**

- Activate for 4-bank DDR and DDR2

- Activate for 8-bank DDR2

Note: TRRD is 7.5ns on 1KB page size and 7.5ns on 2KB page size (two clock minimum).

Note: TPRE is 25.5ns on 1KB page size and 50ns on 2KB page size.
**DDR2 8-Bank Restrictions**

- **Bank ACTIVE restrictions**
  - No more than 4 banks may be activated in a rolling 4 * tRRD + 2 * tCK period
  - tRRD is now based on page size
    - 2KB page size = 10ns (x16 configuration on 1Gb and 2Gb)
    - 1KB page size = 7.5ns (x4, x8 configuration on 1Gb and 2Gb)
    - tRRD has 2 * tCK (MIN) at any tCK

- **Bank PRECHARGE restrictions**
  - Precharge(ALL) command timing equals tRP + 1 * tCK
    - Single-bank PRECHARGE = tRP
    - 8-bank PRECHARGE(ALL) = tRP + 1 * tCK

---

**DDR2 Read – No Additive Latency**

Desire to insert ACT here to keep data bus utilized
**Result: Improved Bus Utilization**

- Allows commands to be placed without conflict
- No gap in data
- Increased overall latency is a drawback in some applications
Latencies

- Only whole clock CAS latencies
- Blue indicates primary speed grades

<table>
<thead>
<tr>
<th>Speed Bin</th>
<th>DDR2-667</th>
<th>DDR2-533</th>
<th>DDR2-400</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>MIN</td>
<td>MIN</td>
<td>MIN</td>
<td>MIN</td>
</tr>
<tr>
<td>CAS Latency</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3 ¹ CK</td>
</tr>
<tr>
<td>tRCD</td>
<td>12</td>
<td>15</td>
<td>15</td>
<td>15 ns</td>
</tr>
<tr>
<td>tRP</td>
<td>12</td>
<td>15</td>
<td>15</td>
<td>15 ns</td>
</tr>
</tbody>
</table>

³-3-3 terminology means:
- 3-cycle CL (CAS latency)
- 3-cycle tRCD (ACTIVE to READ or WRITE delay, i.e. row time)
- 3-cycle tRP (precharge command period, i.e. row closing time)

- DDR2-800 will be 5-5-5, 12.5 ns
DDR2 Read - Precharge

- Read Latency = 4 x (AL = 1, CL = 3, BL = 4, TP = 2) clocks
- Shown with nominal TAC, TCK, and TDO

DDR2 Write to Precharge

- NOTE: 1. Burst length = 4, additive latency = 0, and WRITE latency = 2.
  2. Enable auto precharge.
  3. WR is programmed via BRM (11:10:0) and is calculated by dividing $\frac{BRM}{2}$ by $\frac{TCK}{2}$ and rounding up to the next integer value.
**DDR2 Read to Write**

**Figure 24: READ to WRITE**

- External refresh interval is maintained at 8K refresh per 64ms cycle
- DRAM is managing more internal row/bank refresh than external commands given
  - RFC time is increased for each density

**DDR2 Refresh**

<table>
<thead>
<tr>
<th>Density</th>
<th>256Mb</th>
<th>512Mb</th>
<th>1Gb</th>
<th>2Gb</th>
<th>4Gb</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refresh count (x4x8, x16)</td>
<td>8K,8K</td>
<td>16K,8K</td>
<td>16K,8K</td>
<td>32K,16K</td>
<td>TBD</td>
<td></td>
</tr>
<tr>
<td>Refresh cycle</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>ms</td>
</tr>
<tr>
<td>Refresh interval</td>
<td>7.8</td>
<td>7.8</td>
<td>7.8</td>
<td>7.8</td>
<td>7.8</td>
<td>μs</td>
</tr>
</tbody>
</table>

- *RFC*
On-Die Termination (ODT)

- ODT is a new termination scheme for DDR2 in which the controller and the DRAM have internal termination for the DQ, DQS/DQS# and DM signals.
- The DRAM termination is turned on and off by the controller depending on system loading and READ or WRITE operations.
- ODT allows the system bus to achieve improved signal integrity.

Off-Chip Driver (OCD)

- OCD calibration
  - Goal is 18 ohms ± 1.5 ohms
  - A new mode that allows the controller to measure and adjust the output driver strength of the DRAM.
  - System guarantees memory channel timing and margin.
  - Must recalibrate often for temperature variations.
- Using OCD default, DRAM parametrics are guaranteed over full process, voltage, and temperature.
### DDR2 SDRAM Performance Model

<table>
<thead>
<tr>
<th>Command</th>
<th>To Same Bank</th>
<th>To Different Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>W→R</td>
<td>(CL - 1) + BL/2 + TWTR</td>
<td>BL/2</td>
</tr>
<tr>
<td>W→W</td>
<td>BL/2</td>
<td>BL/2</td>
</tr>
<tr>
<td>W→PRE</td>
<td>(CL - 1) + BL/2 + WTR</td>
<td>1</td>
</tr>
<tr>
<td>W→ACT</td>
<td>W→PRE + TRP</td>
<td>1</td>
</tr>
<tr>
<td>R→R</td>
<td>BL/2</td>
<td>BL/2</td>
</tr>
<tr>
<td>R→W</td>
<td>BL/2 + 2</td>
<td>BL/2 + 2</td>
</tr>
<tr>
<td>R→PRE</td>
<td>AL + BL/2 + TRP - 2</td>
<td>1</td>
</tr>
<tr>
<td>R→ACT</td>
<td>R→PRE + TRP</td>
<td>1</td>
</tr>
<tr>
<td>ACT to ACT</td>
<td>TRC</td>
<td>TRPD</td>
</tr>
<tr>
<td>ACT to R or W</td>
<td>TRCD - AL</td>
<td>1</td>
</tr>
<tr>
<td>ACT to PRE</td>
<td>TRAS</td>
<td>1</td>
</tr>
</tbody>
</table>

- For optimal operation AL = TRCD – 1. Read latency = AL + CL.
- Write latency = read latency – 1. BL = 4 or 8.
- Auto precharge not used in current model.
- Recall the disclaimer.
Motivation for Faster \( t_{RC} \) Memories

- SRAM/DRAM size ratio growth makes future SRAM too expensive per bit
- Personal Computer industry is not demanding improvements
- Most companies not addressing the need for faster \( t_{RC} \), YET!
RLDRAM II

- RLDRAM = reduced latency DRAM
- Motivation
  - Address increasingly large percentage of systems that need:
    - Greater request rates
    - Short burst lengths
    - SRAM-like performance but with DRAM density and cost
    - Have low probability of using open rows
  - Maximize number of applications supported
- 2-533MHz clock frequencies

RLDRAM II Command/Response Examples

- Read Response
  - $^{t_{RC}} = 4$ cycles
  - $^{t_{RC}} = 6$ cycles
  - $^{t_{RC}} = 8$ cycles
- Write Data
  - $^{t_{RL}} = ^{t_{RC}}$
  - $^{t_{WL}} = ^{t_{RL}} + 1$

Note: R0 means read from bank 0. Q00 means first data from bank 0. W0 means write to bank 0. D01 means second data to bank 0.
**RLDRAM II Common I/O**

**Cyclic Bank Switching**

- x9, x18, x36, 2-, 4- and 8-word bursts

<table>
<thead>
<tr>
<th>Burst Length</th>
<th>2</th>
<th>4</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Bus Utilization</td>
<td>100% less refresh</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>MIN Banks Utilized 400 MHz/300 MHz/200 MHz</td>
<td>8/6/4</td>
<td>4/3/2</td>
<td>2/1-2/1</td>
</tr>
</tbody>
</table>

- **Design tips:**
  - Organize data to minimize bank conflicts
  - Consider the bank signals as the lowest address bits

---

**RLDRAM II Minimized Bus Turnaround**

200 MHz, 2-Word Burst, CIO Data Bus

- Worst-case utilization of 66.7 percent at 2-word burst with no unidirectional data streaming
- True for ALL frequencies
Fast Bus Turnaround for High Utilization

- High bandwidth is essential in all systems
- Excessive latency can prevent a system from achieving rated bandwidth
- Bus turnaround and latency are reduced in RLDRAM for maximized usable bandwidth

<table>
<thead>
<tr>
<th>Burst Length</th>
<th>2</th>
<th>4</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Bus Utilization</td>
<td>2/3</td>
<td>4/5</td>
<td>8/9</td>
</tr>
</tbody>
</table>

RLDRAM II minimum bus utilization = \( BL/(BL + 1) \)

Eliminated Bus Turnaround
4-Word Burst, SIO Data Bus

- 100 percent data bus utilization with 4-word burst
- True for ALL frequencies
- Utilization reduced only by refresh requests
- 8-word burst also has 100 percent utilization, unaffected by refresh
RLDRAM II Separate I/ O

- Optimal packet buffer
- x9, x18 on each input and output data bus
- No turnaround cycles

<table>
<thead>
<tr>
<th>Burst Length</th>
<th>2</th>
<th>4</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Bus Utilization</td>
<td>50% less refresh*</td>
<td>100% less refresh</td>
<td>100%</td>
</tr>
<tr>
<td>MIN Banks Utilized 400 MHz/300 MHz/200 MHz</td>
<td>8/6/4</td>
<td>8/6/4</td>
<td>4/2~4/2</td>
</tr>
</tbody>
</table>

*Limited by address bandwidth.

Address Multiplexed Operation

- No performance penalty for 4- or 8-word burst
- Data bus utilization drops to half for 2-word burst

4-Word Burst Addresses

8-Word Burst Addresses

Available for refresh
**RLDRAM Advantages**

- Fastest DRAM tRC (15ns - 20ns)
- Highly tuned and flexible feature set
  - Excellent command/data bus utilization
  - Mode selectable 2, 4 and 8-word burst operation
  - Common I/O or Separate I/O data bus versions
  - 1.5V or 1.8V I/O for operation like SRAM or DDR2 DRAM
  - ODT for clean, high-frequency operation
  - Programmable, self-calibrating output impedance for bus tuning
  - Mode selectable tRC to match target frequency
- Low system power – fewer resources turn on
- Scalable
- SER is four orders of magnitude less than SRAM

**RLDRAM Public Road Map**

<table>
<thead>
<tr>
<th>Features</th>
<th>2024</th>
<th>2025</th>
<th>2026</th>
<th>2027</th>
<th>2028</th>
<th>2029</th>
<th>2030</th>
</tr>
</thead>
<tbody>
<tr>
<td>256Mb RLDRAM</td>
<td>144 FBGA x18, x32 CIO</td>
<td>300 MHz (286, 300, NC 25ns)</td>
<td>325 MHz (300, 320, NC 28ns)</td>
<td>350 MHz (320, 350, NC 30ns)</td>
<td>375 MHz (350, 375, NC 32ns)</td>
<td>400 MHz (375, 400, NC 34ns)</td>
<td>425 MHz (400, 425, NC 36ns)</td>
</tr>
<tr>
<td>256Mb RLDRAM II</td>
<td>144 FBGA x9, 18, x36 CIO, 90</td>
<td>533 MHz (500, 533, NC 35ns)</td>
<td>565 MHz (533, 565, NC 37ns)</td>
<td>590 MHz (565, 590, NC 39ns)</td>
<td>615 MHz (590, 615, NC 41ns)</td>
<td>640 MHz (615, 640, NC 43ns)</td>
<td>665 MHz (640, 665, NC 45ns)</td>
</tr>
<tr>
<td>576Mb RLDRAM II</td>
<td>144 FBGA x9, 18, x36 CIO, 90</td>
<td>533 MHz (500, 533, NC 35ns)</td>
<td>565 MHz (533, 565, NC 37ns)</td>
<td>590 MHz (565, 590, NC 39ns)</td>
<td>615 MHz (590, 615, NC 41ns)</td>
<td>640 MHz (615, 640, NC 43ns)</td>
<td>665 MHz (640, 665, NC 45ns)</td>
</tr>
<tr>
<td>1.125Gb RLDRAM II</td>
<td>144 FBGA x9, 18, x36 CIO, 90</td>
<td>750 MHz</td>
<td>780 MHz</td>
<td>810 MHz</td>
<td>840 MHz</td>
<td>870 MHz</td>
<td>900 MHz</td>
</tr>
<tr>
<td>2.25Gb RLDRAM II</td>
<td>144 FBGA x9, 18, x36 CIO, 90</td>
<td>1.25 Gbps (1.2)</td>
<td>1.30 Gbps (1.3)</td>
<td>1.38 Gbps (1.4)</td>
<td>1.46 Gbps (1.5)</td>
<td>1.54 Gbps (1.6)</td>
<td>1.62 Gbps (1.7)</td>
</tr>
</tbody>
</table>

Speeds are shown in clock rates. All information is subject to change without notice. Data are estimates only. Rev. 02/04. There are more devices on roadmap than shown. Contact Micron for further information.
RLDRAM II Performance Model

<table>
<thead>
<tr>
<th>Clock Frequency</th>
<th>533 MHz</th>
<th>400 MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock period</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.875 ns</td>
<td>2.5 ns</td>
</tr>
<tr>
<td>ACT to same bk ACT</td>
<td>tRC</td>
<td>15 ~8 20 ~8</td>
</tr>
<tr>
<td>Auto RFSH period</td>
<td>tRC</td>
<td>15 ~8 20 ~8</td>
</tr>
</tbody>
</table>

For lower frequencies, \( T = \frac{tRC}{tCK} \)

Examples:

<table>
<thead>
<tr>
<th>f</th>
<th>T RC</th>
<th>f</th>
<th>T RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>3</td>
<td>150</td>
<td>3</td>
</tr>
<tr>
<td>267</td>
<td>4</td>
<td>200</td>
<td>4</td>
</tr>
<tr>
<td>333</td>
<td>5</td>
<td>250</td>
<td>5</td>
</tr>
<tr>
<td>400</td>
<td>6</td>
<td>300</td>
<td>6</td>
</tr>
<tr>
<td>467</td>
<td>7</td>
<td>350</td>
<td>7</td>
</tr>
<tr>
<td>533</td>
<td>8</td>
<td>400</td>
<td>8</td>
</tr>
</tbody>
</table>

488.28125ns periodic refresh = 3.90625\(\mu\)s period for each bank

Minimum frequency = refresh rate = 2.048 MHz

8 banks

Recall the disclaimer!

RLDRAM Performance Model

<table>
<thead>
<tr>
<th>Command</th>
<th>To Same Bank</th>
<th>To Different Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common I/O Device</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R to R</td>
<td>TRC</td>
<td>BL/2</td>
</tr>
<tr>
<td>R to W</td>
<td>Max(BL/2, TRC)</td>
<td>BL/2</td>
</tr>
<tr>
<td>W to W</td>
<td>TRC</td>
<td>BL/2</td>
</tr>
<tr>
<td>W to R</td>
<td>Max(BL/2 + 1, TRC)</td>
<td>BL/2 + 1</td>
</tr>
</tbody>
</table>

Write latency = read latency + 1
BL = 2, 4 or 8-word burst length.
9, 18, 36b bus
Note that 4 cycles are required to transfer BL8 data, hence some bus conflict limitations apply when 1RC is very short.
Notice for total bus turnaround 1 cycle is lost regardless of BL.
Recall the disclaimer!
**RLDRAM Performance Model**

<table>
<thead>
<tr>
<th>Command</th>
<th>To Same Bank</th>
<th>To Different Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Separate I/O Device</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R to R</td>
<td>Max (BL/2, TRC)</td>
<td>BL/2</td>
</tr>
<tr>
<td>R to W</td>
<td>Max (BL/2, TRC)</td>
<td>1</td>
</tr>
<tr>
<td>W to W</td>
<td>Max (BL/2, TRC)</td>
<td>BL/2</td>
</tr>
<tr>
<td>W to R</td>
<td>Max (BL/2, TRC)</td>
<td>1</td>
</tr>
</tbody>
</table>

? Write latency = read latency + 1.
? BL = 2, 4 or 8-word burst length.
? 9, 18b buses (e.g. 18b D, 18b Q)
? Recall the disclaimer!

---

**FCRAM I, II, AND II+**

- FCRAM = fast cycle RAM
- **Motivation**
  - Address increasingly large percentage of systems that:
    - Need greater request rates
    - Require SRAM-like performance but with DRAM density
    - Have low probability of using open rows
  - Cost leveraging on DDR DRAM concepts

Note: most slide information is from Toshiba presentations at Denali MemCon, October 2003 and May 2004
# FCRAM I and II Overview

<table>
<thead>
<tr>
<th>Memory Density</th>
<th>Network FCRAM-I</th>
<th>Network FCRAM-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>( # of bank )</td>
<td>256M / 512Mb</td>
<td>288 / 576Mb</td>
</tr>
<tr>
<td></td>
<td>(4bank / 8bank)</td>
<td>(4bank / 8bank)</td>
</tr>
<tr>
<td>I/O Organization</td>
<td>x8, x16</td>
<td>x9, x18, x36</td>
</tr>
<tr>
<td>Clock Frequency</td>
<td>200MHz</td>
<td>266MHz+</td>
</tr>
<tr>
<td>Random Cycle time</td>
<td>25ns</td>
<td>22.5ns</td>
</tr>
<tr>
<td>Data Strobe</td>
<td>Bi-directional DQS</td>
<td>Uni-directional DQS</td>
</tr>
<tr>
<td>Vdd</td>
<td>2.5V</td>
<td>2.5V</td>
</tr>
<tr>
<td>VddQ</td>
<td>2.5V, 1.8~1.5V</td>
<td>1.8V ~ 1.5V</td>
</tr>
<tr>
<td>I/O Interface</td>
<td>SSTL-2</td>
<td>SSTL-1.8, HSTL</td>
</tr>
<tr>
<td>Package</td>
<td>64pin TSOP-II²</td>
<td>60ball mBGA</td>
</tr>
<tr>
<td></td>
<td>60ball mBGA</td>
<td>60ball mBGA( x18 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>144ball mBGA( x36 )</td>
</tr>
</tbody>
</table>

*1: support by 110nm
*2: only for 256Mb

Source: Ohshima-san, Toshiba

## 288Mb Network FCRAM-II “D”-die Features

- **Process:** 130nm
- **Organization:**
  - x18: 4M words x 4 Banks x 18 bits
  - x36: 2M words x 4 Banks x 36 bits
- **Maximum Clock Frequency:** 333MHz (666Mbps)
- **Pipeline Architecture:** Fast Core cycle, tRC=20ns
- **Function:**
  - Read /CAS Latency(CL) : 4.5 and 6
  - Write /CAS Latency : CL-1
  - Burst Length : 2 and 4
  - Double Data Rate
  - Uni-directional Data Strobe (DS, QS)
  - Interface(VddQ): SSTL-1.8 (Half-strength)&HSTL(1.8~1.5V)
- **Power Supply Voltage:** 2.5V
- **Package:**
  - x18: 60ball(=4x15) mBGA, 1.0 x 1.0mm Ball pitch
  - x36: 144ball(=8x18) mBGA, 0.8 x 1.0mm Ball pitch, JTAG
- **Schedule:**
  - x18: Now Mass production
  - x36: Now CS Sampling

![288Mb Chip & Package Photos](image-url)

**In Production**

288Mb(x18)

9.0mm

16.5 mm

90 mm

18.5 mm

**New Products**

288Mb(x36)

11.0 mm

9.0 mm

18.5 mm
512Mb Network FCRAM1 “A”-die Features

- Process: 130nm
- Organization:
  - 8,000,000 words x 8 Banks x 8 bits
  - 16,000,000 words x 8 Banks x 16 bits
- Pipeline Architecture: Fast Core cycle, tRC=22.5ns
- Maximum Clock Frequency: 266MHz (533Mbps)
- Function:
  - Read/CAS Latency(CL): 3.4 and 5
  - Write/CAS Latency: CL-1
  - Burst Length: 2 and 4
  - Double Data Rate
  - Data Strobe:
    - x8....Bi-directional DQS, /DQS
    - x16....Bi-directional LDQS, UDQS
- Power Supply Voltage: 2.5V
- Interface (VddQ): two version products
  1) SSTL1.8/HSTL (1.8~1.5V), 2) SSTL2 (2.5V)
- Package: 60ball mBGA, 1.0x1.0mm ball pitch
- Schedule:
  1) SSTL1.8/HSTL I/F: CS Sampling, VP in 3Q’04
  2) SSTL2 I/F: CS in June’04, VP in 3Q’04

576Mb Future FC2+ spec ~ now finalized ~

- Number of banks: 8 bank
- I/O Organization : x9, x18, x36 with one chip solution
- Clock Freq & tRC: 400MHz+(800+Mbps) & tRC=<20ns
- Function:
  - Read/CAS Latency(CL): 5, 6 and 7 (8)
  - Write/CAS Latency(WL): CL-1
  - Burst Length: 2, 4 and 8
  - Data Strobe: Uni-directional Differential DS, QS
  - New Features
    1) ZQ Type OCD, 2) ODT, 3) QVLD
    4) Multi Bank Write, 5) External Bank Refresh
- Power Supply Voltage: 1.8V (1.5V capability considered)
- Interface (VddQ): SSTL-1.8 & HSTL (1.8~1.5V)
- Package: 144ball mBGA, 0.8 x 1.0mm ball pitch with JTAG
- Schedule: ES in 2Q’05, CS in 3Q’05, VP in 4Q’05
FCRAM II+ Operation

Ext. Bank Refresh with single refresh counter set

<table>
<thead>
<tr>
<th>Refresh Counter Set</th>
<th>Disturb Rate @ BL4</th>
<th>Disturb Rate @ BL8</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC2+(New) 1</td>
<td>1.024%(Best) or 0.512%*</td>
<td>0(Best)</td>
</tr>
</tbody>
</table>

Multi Bank Write (MBW) Performance and Features

- Features
  - Selectable by EMRS2 set command (A7, A6)
  - Write cycle: Same data write to two banks (ignored BA2 address)
  - Read cycle: Bank Interleave for BA2=0/1
  → Multi Bank Write feature is best fit for higher read duty
    (~100% read) applications such as look up table memory.

- Performance Comparison

<table>
<thead>
<tr>
<th>Parameter</th>
<th>BL4</th>
<th>BL8</th>
</tr>
</thead>
<tbody>
<tr>
<td>tRC @ CL4</td>
<td>5 CLK (CL+1)</td>
<td>7 CLK (CL+3)</td>
</tr>
<tr>
<td>Bus Efficiency @ Read</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o MBW</td>
<td>40%</td>
<td>57%</td>
</tr>
<tr>
<td>with MBW</td>
<td>80%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Multi Bank Write and Bank Interleave Read operation

Read Operation with Multi Bank Write @BL4

<table>
<thead>
<tr>
<th>RIC(CL=1)</th>
<th>CMD</th>
<th>Bank</th>
<th>Address</th>
<th>Do</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Read Operation with Multi Bank Write @BL8

<table>
<thead>
<tr>
<th>RIC(CL=3)</th>
<th>CMD</th>
<th>Bank</th>
<th>Address</th>
<th>Do</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Network FCRAM Roadmap

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Network FCRAM1</td>
<td>64pmx</td>
<td>64pmx</td>
<td>512Mb (x8) x16)</td>
<td>256Mb (x8) x16)</td>
<td>256Mb (x8) x16)</td>
<td>256Mb (x8) x16)</td>
<td>256Mb (x8) x16)</td>
<td></td>
</tr>
<tr>
<td>Network FCRAM2 &amp; Network FCRAM3</td>
<td>64pmx</td>
<td>64pmx</td>
<td>1Gb (x8) x16)</td>
<td>512Mb (x8) x16)</td>
<td>512Mb (x8) x16)</td>
<td>512Mb (x8) x16)</td>
<td>512Mb (x8) x16)</td>
<td></td>
</tr>
</tbody>
</table>

mBGA is gold( bare chip ) BGA.


- September 04

103

104
### FCRAM II Performance Model

#### Clock Frequency

<table>
<thead>
<tr>
<th>Clock Frequency</th>
<th>-30</th>
<th>333 MHz</th>
<th>285.7 MHz</th>
<th>250 MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock period</td>
<td>T</td>
<td>3 ns</td>
<td>3.5 ns</td>
<td>4 ns</td>
</tr>
<tr>
<td>Cas Latency</td>
<td>CL</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>tRC</td>
<td>-21</td>
<td>7</td>
<td>-21</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>ns</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

#### Clock Frequency

<table>
<thead>
<tr>
<th>Clock Frequency</th>
<th>-33</th>
<th>300 MHz</th>
<th>266.7 MHz</th>
<th>222.2 MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock period</td>
<td>T</td>
<td>3.33 ns</td>
<td>3.75 ns</td>
<td>4.5 ns</td>
</tr>
<tr>
<td>Cas Latency</td>
<td>CL</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>tRC</td>
<td>-23.3</td>
<td>7</td>
<td>-22.5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>ns</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

#### Clock Frequency

<table>
<thead>
<tr>
<th>Clock Frequency</th>
<th>-40</th>
<th>250 MHz</th>
<th>222.2 MHz</th>
<th>200 MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock period</td>
<td>T</td>
<td>4 ns</td>
<td>4.5 ns</td>
<td>5 ns</td>
</tr>
<tr>
<td>Cas Latency</td>
<td>CL</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>tRC</td>
<td>-28</td>
<td>7</td>
<td>-27</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>4- bank refresh, all devices tRFC</td>
<td>25</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>ns</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

**3.9 µs period for 4-bank refresh**

**Maximum clock period is 7.5 ns on all devices**

**Recall the disclaimer!**

---

### FCRAM II Performance Model

#### Command To Same Bank To Different Bank

<table>
<thead>
<tr>
<th>Command</th>
<th>To Same Bank</th>
<th>To Different Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common I/O Device</td>
<td>TRC</td>
<td>BL/2 + 2</td>
</tr>
<tr>
<td>R to R</td>
<td>TRC</td>
<td>2</td>
</tr>
<tr>
<td>R to W</td>
<td>TRC</td>
<td>2</td>
</tr>
<tr>
<td>W to W</td>
<td>TRC</td>
<td>2</td>
</tr>
<tr>
<td>W to R</td>
<td>TRC</td>
<td>2</td>
</tr>
</tbody>
</table>

**TRC = CL + 1**

**Write latency = CL – 1**

**BL = 2 or 4-word burst length.**

**Notice for total bus turnaround 2 cycles are lost regardless of BL**

**Recall the disclaimer!**
FCRAM II+ Performance Model

<table>
<thead>
<tr>
<th>Clock Frequency</th>
<th>? MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock period</td>
<td>T</td>
</tr>
<tr>
<td>Cas Latency</td>
<td>CL</td>
</tr>
<tr>
<td>tRC</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
</tbody>
</table>

? tRC range is 6-9 cycles, calculate as tRC/tCK.
? This is speculative, based only on presentations, not Toshiba data sheets. tRC might be less than 20ns
? Refresh takes tRC cycles, required every 488ns, 8 bank device
These assumptions may be incorrect.
? Recall the disclaimer!

FCRAM II+ Performance Model

<table>
<thead>
<tr>
<th>Command</th>
<th>To Same Bank</th>
<th>To Different Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common I/O Device</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R to R</td>
<td>TRC</td>
<td>max(BL/2 , 2)</td>
</tr>
<tr>
<td>R to W</td>
<td>TRC</td>
<td>BL/2 + 2</td>
</tr>
<tr>
<td>W to W</td>
<td>TRC or TRC+ 2 if BL8 MBW</td>
<td>max(BL/2 , 2)</td>
</tr>
<tr>
<td>W to R</td>
<td>TRC or TRC+ 2 if BL8 MBW</td>
<td>max(BL/2 , 2)</td>
</tr>
</tbody>
</table>

? This is speculative, based only on presentations, not Toshiba data sheets.
? Write latency = CL – 1
? BL = 2, 4 or 8-word burst length.
? MBW is Multi Bank Write, i.e. dual-bank write
not used in today’s performance scenarios
? Notice for total bus turnaround 2 cycles are lost regardless of BL
? Recall the disclaimer!
GDDR3

- GDDR3 = graphics double data rate DRAM version 3
- Motivation
  - Provide vastly higher pin bandwidth for wide-bus graphic applications
  - Simplify DRAM commands
  - Optimize die size

GDDR3 Features

- Clock frequency of 500 MHz- 700 MHz
- Single-ended read and write strobes
- RDQS and WDQS per byte
- On-die termination
- Dynamic programmable impedance output driver
- Duty cycle correction on clock input
- 1.8V core
- 4x 2M x 32 (256Mb)
GDDR3 Features
(continued)

- 1.8V pseudo-open drain I/O
- tRAS lockout
- Concurrent AUTO PRECHARGE
- 4 banks
- 4K refresh
- Burst length 4
  - 8 is defined but not supported by all manufacturers
- Sequential burst type only

High-Speed Advantages

- Dynamically controlled impedance output driver
- The DRAM controls the on-die termination for the reads
- On-die termination on all address and control pins
- Single-ended read and write strobes
- Reduced WRITE latency
- Designed specifically for high-speed, point-to-point applications
Key Timing Parameters

- \( t_{\text{CK}} = 500 \text{ MHz to 700 MHz} \)
- \( t_{\text{RAS}} = 30\text{ns} \)
- \( t_{\text{RC}} = 60\text{ns} \)
- \( t_{\text{RCD}} = 16\text{ns} \)
- \( t_{\text{RP}} = 13\text{ns} \)
- \( t_{\text{RRD}} = 6\text{ns} \)
- \( t_{\text{WR}} = 4 \cdot t_{\text{CK}} \)
- WRITE latency = 1 to 4 clocks
  - Set in mode register
    - 3 to 4 clocks for low-power operation
    - 1 to 2 clocks with increased operating power

Read Data

NOTE: 1. DO n = data out from column n.
2. Skew length = \( \Delta \).
3. Three subsequent elements of data out appear in the programmed order following DO n.
4. Shown with masked \( \overline{WE} \), and \( \overline{RAS} \).
### GDDR3 SDRAM Performance Model

#### Clock Frequency

<table>
<thead>
<tr>
<th>Clock Frequency</th>
<th>700 MHz</th>
<th>600 MHz</th>
<th>500 MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock period</td>
<td>T</td>
<td>1.429 ns</td>
<td>1.667 ns</td>
</tr>
<tr>
<td>Cas Latency</td>
<td>CL</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ns</th>
<th>T</th>
<th>ns</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT to same bk ACT</td>
<td>tRC</td>
<td>31</td>
<td>27</td>
<td>21</td>
</tr>
<tr>
<td>ACT to R</td>
<td>tRCDR</td>
<td>10</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>ACT to W</td>
<td>tRCDW</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>ACT to PRE</td>
<td>tRAS</td>
<td>22</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>PRE Period</td>
<td>tRP</td>
<td>9</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>W to PRE</td>
<td>tWR</td>
<td>9</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Internal W to R</td>
<td>tWTR</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>ACT to diff bk ACT</td>
<td>tRRD</td>
<td>8</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Auto RFSH period</td>
<td>tRFC</td>
<td>39</td>
<td>33</td>
<td>27</td>
</tr>
<tr>
<td>Last data in to R</td>
<td>tCDLR</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Write Latency</td>
<td>WL</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

- WL is programmable; value shown is used in model
- T represents 1 clock cycle
- Maximum clock period is 3.33 ns
- Recall the disclaimer!

---

### GDDR3 SDRAM Performance Model

#### Command To Same Bank To Different Bank

<table>
<thead>
<tr>
<th>Command</th>
<th>To Same Bank</th>
<th>To Different Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>W to R</td>
<td>WL + BL/2 + T WTR</td>
<td>1</td>
</tr>
<tr>
<td>W to W</td>
<td>BL/2</td>
<td>BL/2</td>
</tr>
<tr>
<td>W to PRE</td>
<td>WL + BL/2 + TWR</td>
<td>1</td>
</tr>
<tr>
<td>W to ACT</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>R to R</td>
<td>BL/2</td>
<td>BL/2</td>
</tr>
<tr>
<td>R to W</td>
<td>CL + BL/2</td>
<td>CL + BL/2 + 1 - WL</td>
</tr>
<tr>
<td>R to PRE</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>R to ACT</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ACT to ACT</td>
<td>tRC</td>
<td>tRRD</td>
</tr>
<tr>
<td>ACT to R</td>
<td>tRCDR</td>
<td>1</td>
</tr>
<tr>
<td>ACT to W</td>
<td>tRCDW</td>
<td>1</td>
</tr>
<tr>
<td>ACT to PRE</td>
<td>TRAS</td>
<td>1</td>
</tr>
</tbody>
</table>

- Refresh is 32ms, 4K = 7.8125 µs periodic.
- BL = 4 banks.
- AUTO precharge not enabled.
- Recall the disclaimer.
RAMBUS XDR™

Rambus XDR

XDR DRAM Features

- Process: 130 nm / 110 nm
- Organization:
  - 512Mb... 4M words x 8Banks x 16bits
  - 256Mb... 2M words x 8Banks x 16bits
- Function:
  - 1clock Request Packet
  - 2clock Data Packet (16 bit burst)
  - Page Size: 2K Byte
  - tREFI: 0.49us (= 16ms/32K)
  - Bank Interleaved Refresh
  - Frequency (CLK): 300MHz / 400MHz / 500MHz
  - Frequency (DQ): 2.4GHz / 3.2GHz / 4.0GHz
  - Octal Data Rate (ODR)
  - Write Mask
  - Early Read After Write
  - Powerdown Self-refresh support
  - Dynamic Width Control: x16/x8/x4

Source: Ohshima-san, Toshiba
XDR DRAM Package Information

108ball molded BGA (Lead free)

---

XDR DRAM Application-1

Dynamic Width Control allows system MB scaling without impacting performance

- XDR Solution realizes 4.8~6.4GB/s with 90 controller pins (incl. power/ground)
  - Solves the procurement nightmare, minimizes inventory loss, saves pin count, and saves development cost of having multiple DRAMs and controllers.
  - XDR DRAM will realize 2 to 3 chip system solution with significant BW for Digital Consumer Systems in 2005 timeframe.
### XDR Performance Model

<table>
<thead>
<tr>
<th>Clock Frequency</th>
<th>400MHz (2.5ns)</th>
<th>300MHz (3.3ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Device A or B</td>
<td>Device A</td>
</tr>
<tr>
<td>Clock period</td>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>ACT to ACT</td>
<td>tRC / tRR</td>
<td>16</td>
</tr>
<tr>
<td>ACT to R</td>
<td>tRCD - R</td>
<td>5</td>
</tr>
<tr>
<td>ACT to W</td>
<td>tRCD - W</td>
<td>1</td>
</tr>
<tr>
<td>ACT to PRE</td>
<td>tRAS</td>
<td>10</td>
</tr>
<tr>
<td>PRE Period (P to A)</td>
<td>tRP</td>
<td>6</td>
</tr>
<tr>
<td>R to Q</td>
<td>tCAC</td>
<td>6</td>
</tr>
<tr>
<td>W to Q</td>
<td>tCWD</td>
<td>3</td>
</tr>
<tr>
<td>R to R or W to W</td>
<td>tCC</td>
<td>2</td>
</tr>
<tr>
<td>R to W</td>
<td>tdRW</td>
<td>8</td>
</tr>
<tr>
<td>W to R</td>
<td>tdWR</td>
<td>9</td>
</tr>
<tr>
<td>W to PRE</td>
<td>tWRP</td>
<td>10</td>
</tr>
<tr>
<td>R to PRE</td>
<td>tRDP</td>
<td>3</td>
</tr>
<tr>
<td>Auto RFSH period</td>
<td>tRFC</td>
<td>16</td>
</tr>
<tr>
<td>Refresh interval</td>
<td>488ns</td>
<td>488ns</td>
</tr>
</tbody>
</table>

- T represents 1 clock cycle
- Maximum clock period is 3.83ns (261 MHz)
- Note: 500 MHz spec (XDR4000Mb/ s) is excluded from performance analysis – too far out! Recall the disclaimer!
XDR Examples

- 300 MHz Device A: 53.3 ns tRC
- 400 MHz Device B: 50 ns tRC
- 400 MHz Device A: 40 ns tRC
- Future
  - 500 MHz Device B: 40 ns tRC
  - 500 MHz Device A: 32 ns tRC

Performance Analysis Methodology

- Proprietary software written in Visual Basic includes
  - User-defined advanced address generation
    - Unlimited multiple-thread capability
    - Unique request definition for each thread
    - Address generation per probability inputs
  - Memory behavioral model
    - The “Performance Model” for each memory type
  - Controller behavioral model optimized for each memory device
Performance Analysis Methodology (continued)

- Objective is to compare all memory devices fairly
- Expose each one to the identical request steam
- Controllers independently allowed to optimize request stream limited by common constraints
  - e.g. allowed to switch threads if stalled
  - None of today's examples use multi-threading
- Frequency sweep of all devices for each scenario
  - Include "sweet spot" for each memory

Component Choices

- 512-576Mb device wherever possible, otherwise nearest lower density
- DDR SDRAM:
  - 76.9-133.3 MHz CL2, to 166.7 MHz CL2.5, to 200 MHz CL3
- DDR2 SDRAM:
  - 125-200 MHz CL3, to 266.7 MHz CL4, to 333.3 MHz CL5
- RLDRAM II:
  - 15 ns 1RC for ClO and SIO: 2-200 MHz 3T, to 266.7 MHz 4T, to 333.3 MHz 5T, to 400 MHz 6T, to 466.7 MHz 7T, to 533.3 MHz 8T
- FCRAM2: -30 (22.5 ns 1RC): 133.3-250 MHz CL4, to 285.7 MHz CL5, to 333.3 MHz CL6 (note RC = CL+1)
- FCRAM2+: -25 (20 ns 1RC): 200-300 MHz CL5, to 350 MHz CL6, to 400 MHz CL7, to 450 MHz CL8
- QDR2 SRAM: 0-250 MHz
- DDR2 SRAM: 0-333.3 MHz
- QDR3 & DDR3 SRAM: 250-500 MHz
Frequencies Analyzed

<table>
<thead>
<tr>
<th>Clock Frequency (MHz)</th>
<th>Clock Period (ns)</th>
<th>Frequencies of Interest Especially Pertain To:</th>
</tr>
</thead>
<tbody>
<tr>
<td>76.9</td>
<td>13</td>
<td>DDR SDRAM</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>8</td>
<td>DDR2 SDRAM</td>
</tr>
<tr>
<td>133.3</td>
<td>7.5</td>
<td>DDR SDRAM, FCRAM2</td>
</tr>
<tr>
<td>150</td>
<td>6.667</td>
<td>RLDRAM2</td>
</tr>
<tr>
<td>166.7</td>
<td>6</td>
<td>DDR SDRAM</td>
</tr>
<tr>
<td>200</td>
<td>5</td>
<td>DDR SDRAM, DDR2 SDRAM, RLDRAM2, FCRAM2</td>
</tr>
<tr>
<td>222.2</td>
<td>4.5</td>
<td>FCRAM2</td>
</tr>
<tr>
<td>250</td>
<td>4</td>
<td>RLDRAM2, FCRAM2, ALL SRAM</td>
</tr>
<tr>
<td>261.1</td>
<td>3.83</td>
<td>XOR</td>
</tr>
<tr>
<td>266.7</td>
<td>3.75</td>
<td>DDR2 SDRAM, RLDRAM2, FCRAM2</td>
</tr>
<tr>
<td>286.7</td>
<td>3.5</td>
<td>FCRAM2</td>
</tr>
<tr>
<td>300</td>
<td>3.33</td>
<td>RLDRAM2, FCRAM2, FCRAM2, GDDR1, XDR</td>
</tr>
<tr>
<td>333.3</td>
<td>3</td>
<td>DDR2 SDRAM, RLDRAM2, FCRAM2, GDDR2/DDR2 SRAM</td>
</tr>
<tr>
<td>350</td>
<td>2.857</td>
<td>RLDRAM2, FCRAM2</td>
</tr>
<tr>
<td>400</td>
<td>2.5</td>
<td>RLDRAM2, FCRAM2, XOR</td>
</tr>
<tr>
<td>450</td>
<td>2.222</td>
<td>FCRAM2</td>
</tr>
<tr>
<td>466.7</td>
<td>2.143</td>
<td>RLDRAM2</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>GDDR3, QDR3, DDR3, DDR3 SRAM</td>
</tr>
<tr>
<td>533.3</td>
<td>1.876</td>
<td>RLDRAM2, FCRAM2</td>
</tr>
<tr>
<td>600</td>
<td>1.667</td>
<td>GDDR3</td>
</tr>
<tr>
<td>700</td>
<td>1.429</td>
<td>GDDR3</td>
</tr>
</tbody>
</table>

Shaded areas signify precise frequency or period

Performing a Fair Comparison

- Must have
  - Same number of data bus signals
  - Same data size response per request

- Examples
  - Differential data devices have double the pin count
    - Compare x16 diff I/O device with x32 single-ended
  - Separate I/O devices have two data buses
    - Compare x8 SIO (2 buses, x8 each) with x16 CIO
### Fair Comparison Chart
32b Base Data Size (16b x 2 Word Burst)

<table>
<thead>
<tr>
<th>Device</th>
<th>BL2</th>
<th>BL4</th>
<th>BL8</th>
<th>BL16</th>
<th>BL32</th>
<th>BL64</th>
</tr>
</thead>
<tbody>
<tr>
<td>16b DDR SDRAM, 16(18)b RLDRAM2ClO &amp; FCRAM2+</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>2x8</td>
<td>4x8</td>
<td>8x8</td>
</tr>
<tr>
<td>16b DDR2 SDRAM</td>
<td>-</td>
<td>4</td>
<td>8</td>
<td>2x8</td>
<td>4x8</td>
<td>8x8</td>
</tr>
<tr>
<td>16b FCRAM2, 16(18)b DDR2 SRAM</td>
<td>2</td>
<td>4</td>
<td>2x4</td>
<td>4x4</td>
<td>8x4</td>
<td>16x4</td>
</tr>
<tr>
<td>8(9)b RLDRAM2SIO</td>
<td>4</td>
<td>8</td>
<td>2x8</td>
<td>4x8</td>
<td>16x8</td>
<td>32x8</td>
</tr>
<tr>
<td>8b XDR</td>
<td>-</td>
<td>-</td>
<td>16</td>
<td>2x16</td>
<td>4x16</td>
<td>8x16</td>
</tr>
<tr>
<td>8(9)b QDR2 &amp; QDR3 SRAMs</td>
<td>4</td>
<td>2x4</td>
<td>4x4</td>
<td>8x4</td>
<td>16x4</td>
<td>32x4</td>
</tr>
<tr>
<td>16(18)b DDR3 SRAMs</td>
<td>2</td>
<td>2x2</td>
<td>4x2</td>
<td>8x2</td>
<td>16x2</td>
<td>32x2</td>
</tr>
<tr>
<td>8(9)b DDR2SIO SRAM</td>
<td>2x2</td>
<td>4x2</td>
<td>8x2</td>
<td>16x2</td>
<td>32x2</td>
<td>64x2</td>
</tr>
</tbody>
</table>

### Fair Comparison Chart
64b Base Data Size (32b x 2 Word Burst)

<table>
<thead>
<tr>
<th>Device</th>
<th>BL2</th>
<th>BL4</th>
<th>BL8</th>
<th>BL16</th>
<th>BL32</th>
<th>BL64</th>
<th>BL128</th>
</tr>
</thead>
<tbody>
<tr>
<td>32b GDDR3 SDRAM</td>
<td>-</td>
<td>4</td>
<td>8</td>
<td>2x8</td>
<td>4x8</td>
<td>8x8</td>
<td>16x8</td>
</tr>
<tr>
<td>16b XDR</td>
<td>-</td>
<td>-</td>
<td>16</td>
<td>2x16</td>
<td>4x16</td>
<td>8x16</td>
<td>16x16</td>
</tr>
<tr>
<td>32(36)b RLDRAM2ClO &amp; FCRAM2+</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>2x8</td>
<td>4x8</td>
<td>8x8</td>
<td>16x8</td>
</tr>
<tr>
<td>16(18)b RLDRAM2SIO</td>
<td>4</td>
<td>8</td>
<td>2x8</td>
<td>4x8</td>
<td>8x8</td>
<td>16x8</td>
<td>32x8</td>
</tr>
<tr>
<td>32b FCRAM2</td>
<td>2</td>
<td>4</td>
<td>2x4</td>
<td>4x4</td>
<td>8x4</td>
<td>16x4</td>
<td>32x4</td>
</tr>
<tr>
<td>16(18)b QDR2, DDR2SIO, QDR3 SRAMs</td>
<td>4</td>
<td>2x4</td>
<td>4x4</td>
<td>8x4</td>
<td>16x4</td>
<td>32x4</td>
<td>64x4</td>
</tr>
<tr>
<td>32(36)b DDR2 &amp; DDR3 SRAMs</td>
<td>2</td>
<td>4</td>
<td>2x4</td>
<td>4x4</td>
<td>8x4</td>
<td>16x4</td>
<td>32x4</td>
</tr>
</tbody>
</table>
Latest Results

- All latest scenarios use “Even Distribution” of requests
  - E.g. 1R 1W means 1 read from a random address, then 1 write from a random address
  - Random address also means random bank #
- If Burst Length exceeds device burst length
  - Additional data is from open row (e.g. DDR, DDR2, GDDR3, XDR DRAMs)
  - Additional data is from next adjacent bank for devices with no open rows (e.g. FCRAM, RLDRAM)
  - Additional data is from next address for SRAMs

Corrected Scenario A – Random BL2

Scenario Bus Utilization

- Frequency (MHz)
- Average Data Bus Utilization

Graph showing data utilization for various memory types at different frequencies.
Corrected Scenario B – Random BL4

Scenario Bus Utilization

Corrected G – Random 9R 1W BL2, Locality*

Scenario Bus Utilization

*Locality means 25% of device accesses are directed to an open resource, applies only to devices with ACT command.
Evenly Distributed 1R 1W BL2

Scenario Elapsed Time (usec) vs. Clock Period (ns)

Evenly Distributed 1R 1W BL4

Scenario Bus Utilization
Evenly Distributed 1R 1W BL4

ScenarioElapsed Time (usec) vs. Clock Period (ns)

Evenly Distributed 1R 1W BL8

Scenario Bus Utilization

September 04
Evenly Distributed 1R 1W BL8

Scenario Elapsed Time (usec) vs. Clock Period (ns)

Evenly Distributed 1R 1W BL16

Scenario Bus Utilization

September 04
Evenly Distributed 1R 1W BL32

Scenario Elapsed Time (usec) vs. Clock Period (ns)

Evenly Distributed 1R 1W BL64

Scenario Bus Utilization
Evenly Distributed 3R 1W BL8

Performance Results Notes

- Long latency QDR3 and DDR3 performs like their predecessors
  - Differences will be seen when pipeline stalls are modeled
  - Thread stalls model system waiting for dependent data
  - Can’t issue new command until data is received and manipulated
  - Longer latency devices affected most by this

- Observed influence of controller design on results
  - Any knowledge of what comes next helps
    - Controller can add latency to gain this knowledge
    - Done only to a very limited degree in this analysis
      - E.g. QDR complex controllers stall R issue until W available but will not reorder operations otherwise
      - E.g. Low latency DRAMs don’t issue refresh to a bank with a pending request
Conclusions

- “Commodity” DRAMs improve performance with
  - Greater locality
    - Fewer bank misses
  - Fewer bus turnarounds
    - Longer burst lengths and greater R:W ratios
- XDR outperforms all other memories once data chunking gets extreme
  - E.g. Some specialized graphics applications
- RLDRAM II outperforms other DRAMs in every other scenario
  - Rivals SRAM performance
- SRAM still outperforms other memories for 2-word burst
  - Has the edge at 4-word burst above 300–400MHz but not less than that
  - No longer has much advantage otherwise except for determinism

Conclusions (continued)

- It is easy to predict SRAM behavior
- Not so easy to predict DRAM behavior
  - Probabilistic resource availability
  - Require appropriate controller and device model to assess
    - Similar controller intelligence is essential when comparing memory devices
- Future: continued DRAM improvements
  - Eventual disappearance of discrete SRAM
    - Perhaps this decade!
Acknowledgements

- **Micron**
  - Jeff Janzen, Chris Johnson, Computing & Consumer Group
  - Todd Dinkelman, Jim Cooke, Negin Roohi, NetCom Group
  - Jim Lane (Technical Writer)
- **Toshiba**
  - Shigeo Ohshima
- **QDR Co-development Group (Cypress, IDT, NEC, Renasas, Samsung)**
  - Mike Pearson
- **RAMBUS**
  - XDR data sheets
- **Stanford University**
  - Christos Kozyrakis for his valuable review and suggestions

Micron Technology