Wide I/O DRAM Architecture Utilizing Proximity Communication

by
Qawi Harvard
Thesis Defense – October 8th, 2009
Introduction

Bandwidth and power consumption of dynamic random access memory stifles computer performance scaling

- Background
- Status of Proximity Communication
- DRAM Market Analysis
- 4 Gb DRAM Architecture
- Wide I/O DRAM Architecture Utilizing Proximity Communication
Background

- Memory Gap
  - Main memory does not scale with processor performance

- Power
  - Current consumption is rising
  - Bandwidth increases power
  - Voltage scaling masks the issue

- Density
  - Memory channel loading
  - Limits bandwidth

- Proximity Communication
  - Proposed by Ivan Sutherland – US Patent #6,500,696
  - Promises to reduce power and increase bandwidth
Capacitive Coupled Proximity Communication

- Top metal forms the parallel plates
- Chip-to-chip communication through coupling capacitor

Ref:[1]
Proximity Communication

- **Benefits**
  - Increased I/O density
  - Avoids on/off chip wires
  - Eases chip replacement at the system level
  - Enhances system level testability
  - Enables smaller chip sizes
  - Removes the need for ESD protection

- **Challenges**
  - Mechanical misalignment
  - Applying power to the chips
  - Thermal solution

Ref:[1-5]

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
Proximity Communication

- **Parallel Plate Capacitance**
  \[ C = \frac{\varepsilon_0 A}{d} \]
  \[ \varepsilon_0 = 8.9 \, \text{aF} / \mu\text{m} \]

- **10 pF/mm\(^2\)**
  - Chip-to-chip separation
  - \( d = 1 \, \mu\text{m} \)

- **One channel**
  - 50 fF
  - 200 signals/mm\(^2\)

Ref: [1] Qawi Harvard – Oct. 8\(^{th}\), 2009
Thesis Defense
Mechanical Misalignment

- Six axis
- Multiple sources

Ref:[5]
Qawi Harvard – Oct. 8th, 2009
Thesis Defense
Proximity Communication

- **Electronic Sensors**
  - Chip-to-chip separation sensors (0.2 µm resolution)
  - Vernier scale incorporated on chip (1.0 µm resolution)

- **Electrical Re-Alignment**
  - Receive array
  - Micro-transmit array
  - Electronic steering circuit

Ref: [1-5]

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
DRAM Market Analysis

- Revisit the Memory Gap
  - “Performance” becomes a relative term
  - Dichotomy in scaling
- Why Density?
- Why Not Latency?

Ref:[6]

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
DRAM Market Analysis

- Moore’s Law
  - 41% increase in transistor count per year
- Selling Price
  - 36% historic decline per year
- Putting it into Perspective
  - 1 Gb 2009 → $2.00
  - 2 Gb 2011 → $1.64
- Density or Bust!!

Historically the price per bit has declined by 9% every quarter (1974 – 2008).

Ref:[7-8]
Qawi Harvard – Oct. 8th, 2009
Thesis Defense
DRAM Market Analysis

- **Cost**
  - Low cost manufacturing process
    - 3 metal layers
      - Increased usage of each level
    - Small chip size
      - Limits I/O count

- **Moore’s Law**
  - 41% scaling per year
    - Wordline cross sectional area
    - Tight metal pitch
    - Contact resistance

- **Physics of Scaling**
  - Latency must increase

Ref:[7-8,13]

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
DRAM Market Analysis

- **Home PC**
  - Plugged into a wall

- **Mobile**
  - Battery life

- **Server**
  - Power consumption
  - Cooling

- **Trending Up**
  - Poor Efficiency
  - Bandwidth Driven

Ref:[14-17]

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
DRAM Market Analysis

- Interface versus Core
  - Interface bares the burden
  - Core cycles
- DRAM Pre-fetch
  - Doubles at each generation
- Density limited by bandwidth
  - SSTL loading in memory channel
  - Increase chip count per module

Ref: [14-17]
Qawi Harvard – Oct. 8th, 2009
Thesis Defense
Possible Architecture

✓ Compared to ITRS

2012 Production Release

74 mm²

56% Array Efficiency

40 nm

Ref: [10, 12, 20-24]

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
4 Gb DRAM Architecture

- 6F² Memory Cell
- Feature Size = 40 nm
- Cell Area = 0.0096 µm²
- 3F Pitch per Wordline
- 2F Pitch per Bitline
- 256 kb Array Macro
  - Core Array
    - 512 bitlines ≈ 43.6 µm
    - 512 wordlines ≈ 65.4 µm
- Periphery Circuitry
  - 4 µm space allocated

Ref:[24]
Qawi Harvard – Oct. 8 th, 2009
Thesis Defense
4 Gb DRAM Architecture

Ref: [25]

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
4 Gb DRAM Architecture

- 256 Mb Array
  - ✔ 32 x 32 256 kb macros
  - \[32 \cdot (43.6 \, \mu m + 4 \, \mu m) = 1.532 \, mm\]

- 1 Gb Array
  - ✔ Multiple implementations
4 Gb DRAM Architecture

- **ITRS**
  - 74 mm²
  - 56% Array Efficiency

- **Wide I/O Architecture**
  - Moving the pads
  - Centralized Row
  - Centralized Column

Chip Size = 71.4 mm²
Array Efficiency = 57.7%

Ref: [29]
Qawi Harvard – Oct. 8th, 2009
Thesis Defense
Wide I/O Chip Architecture

- 64 Bytes per Chip
- 6.2% Chip Size Reduction
- 6.6% Increase in Array Efficiency
- Challenges
  - Routing from the edge
  - Array I/O route increase
    - 2.3 mm → 4.6 mm
  - Additional row decode
- Create Eight Internal Banks

Chip Size = 67.0 mm²
Array Efficiency = 61.5%
4 Gb DRAM Architecture

- 1 mm Allocation for Proximity Channel
- Buffers at the Center
  - Increase global I/O metal usage
- Array I/O Routing Reduced to 2.3 mm
- Architecture NOT Efficient for Proximity Communication
  - 6.7 mm versus 10.4 mm
  - Buffers required
  - Large metal usage

Chip Size = 69.68 mm²
Array Efficiency = 59.2%

Proximity Interface

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
4 Gb DRAM Architecture

- Multiple Bank Architectures
- Page Size
  - Standard size = 8k
  - Energy efficiency
- Global Row Routing
  - ~20 ns latency
- Global Column Routing
  - ~5 ns latency
- D – Architecture
  - Page decode

Ref: [13, 16, 27, 30,]

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
Wide I/O Chip Architecture

- Chip Size = 68.88 mm², Array Efficiency = 59.9%
  - Centralized row & column
  - Buffers not required
  - 12.3 mm for proximity communication
  - Enables two levels of metal
Wide I/O DRAM Architecture

- **Split Bank Architecture**
  - 64 bytes = 512 signals
  - 6 mm / 256 ≈ 43 µm per signal
  - 0.4 µm pitch < 1 % metal usage

<table>
<thead>
<tr>
<th>RAS &amp; Address</th>
<th>Local Wordline</th>
<th>Global Wordline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half-Bank&lt;7&gt;</td>
<td>ROW</td>
<td>Half-Bank&lt;7&gt;</td>
</tr>
<tr>
<td>Half-Bank&lt;6&gt;</td>
<td>ROW</td>
<td>Half-Bank&lt;6&gt;</td>
</tr>
<tr>
<td>Half-Bank&lt;5&gt;</td>
<td>ROW</td>
<td>Half-Bank&lt;5&gt;</td>
</tr>
<tr>
<td>Half-Bank&lt;4&gt;</td>
<td>ROW</td>
<td>Half-Bank&lt;4&gt;</td>
</tr>
<tr>
<td></td>
<td>COLUMN</td>
<td></td>
</tr>
<tr>
<td>Half-Bank&lt;3&gt;</td>
<td>ROW</td>
<td>Half-Bank&lt;3&gt;</td>
</tr>
<tr>
<td>Half-Bank&lt;2&gt;</td>
<td>ROW</td>
<td>Half-Bank&lt;2&gt;</td>
</tr>
<tr>
<td></td>
<td>COLUMN</td>
<td></td>
</tr>
<tr>
<td>Half-Bank&lt;0&gt;</td>
<td></td>
<td>Half-Bank&lt;0&gt;</td>
</tr>
</tbody>
</table>

Proximity Interface
Wide I/O DRAM Architecture

- Split Page Architecture
  - 8k page keeps current relative
  - Page decode required
  - 32 differential signals per macro

- Local I/O Routing
  - Space limited
  - Increase space?
  - Increase page size?

Ref: [12,20-24,26,28]

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
**Wide I/O DRAM Architecture**

- **New Column Routing**
  - Global I/O operates at higher frequency
  - Protocol allows for insertion of data

<table>
<thead>
<tr>
<th>Half-Bank&lt;7&gt;</th>
<th>BUSY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half-Bank&lt;6&gt;</td>
<td>FREE</td>
</tr>
<tr>
<td>Half-Bank&lt;5&gt;</td>
<td>FREE</td>
</tr>
<tr>
<td>Half-Bank&lt;4&gt;</td>
<td>BUSY</td>
</tr>
<tr>
<td>Half-Bank&lt;3&gt;</td>
<td>FREE</td>
</tr>
<tr>
<td>Half-Bank&lt;2&gt;</td>
<td>FREE</td>
</tr>
<tr>
<td>Half-Bank&lt;1&gt;</td>
<td>FREE</td>
</tr>
<tr>
<td>Half-Bank&lt;0&gt;</td>
<td>FREE</td>
</tr>
</tbody>
</table>

- **Proximity Interface**
  - Wordline Fires
  - Data Latched & Inserted on Global I/O Bus
  - Next Wordline Fires
  - Data Latched & Inserted on First Available Slot of the Global I/O Bus
  - Global I/O Bus
Wide I/O DRAM Architecture

- **Slice Architecture**
  - Ease of design
    - Uniformity, speed, verification

<table>
<thead>
<tr>
<th>DATA SLICE</th>
<th>DATA SLICE</th>
<th>CONTROL SLICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half-Bank&lt;0&gt;</td>
<td>ROW</td>
<td></td>
</tr>
<tr>
<td>Half-Bank&lt;6&gt;</td>
<td>ROW</td>
<td></td>
</tr>
<tr>
<td>Half-Bank&lt;5&gt;</td>
<td>ROW</td>
<td></td>
</tr>
<tr>
<td>Half-Bank&lt;4&gt;</td>
<td>ROW</td>
<td></td>
</tr>
<tr>
<td>COLUMN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-Bank&lt;3&gt;</td>
<td>ROW</td>
<td></td>
</tr>
<tr>
<td>Half-Bank&lt;2&gt;</td>
<td>ROW</td>
<td></td>
</tr>
<tr>
<td>Half-Bank&lt;1&gt;</td>
<td>ROW</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Half-Bank&lt;0&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proximity Interface</td>
</tr>
</tbody>
</table>

50 µm
Serves 4 DQ

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
Wide I/O DRAM Architecture

- **64 Bytes per Chip**
  - Significant bandwidth increase

- **Power Consumption**
  - Standard 8k page size
  - Split bank, split page

- **Cost Performance**
  - Two metals enabled for 4 Gb
  - Smaller chip size, higher array efficiency
Wide I/O DRAM Architecture

Power Consumption

- DDR3
- PxCDRAM x16
- PxCDRAM x32
- PxCDRAM x64

Relative Energy [%]

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
Wide I/O DRAM Architecture

- Bandwidth

<table>
<thead>
<tr>
<th>Year</th>
<th>Module Bandwidth [GB/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDR3</td>
<td>25</td>
</tr>
<tr>
<td>DDR4</td>
<td>100</td>
</tr>
<tr>
<td>DDR5</td>
<td>225</td>
</tr>
<tr>
<td>PxCDRAM x16</td>
<td>50</td>
</tr>
<tr>
<td>PxCDRAM x32</td>
<td>125</td>
</tr>
<tr>
<td>PxCDRAM x64</td>
<td>200</td>
</tr>
</tbody>
</table>

Qawi Harvard – Oct. 8th, 2009
Thesis Defense
Future Work

- Applying Proximity Communication to New Memory Technologies
  - “High” density
  - Chalcogenide
  - Slice architecture
  - Circuit design techniques

- Local I/O Routing
  - New column global I/O structure
  - Through bitline routing
  - Novel local I/O latch
Acknowledgments

- Dr. Jake Baker
- Dr. Kris Campbell
- Dr. Robert Drost
- Dr. Sin Ming Loo
- Dr. Thad Welch
- Ms. Donna Welch
- Family support

Questions?
References


Rambus, “Challenges and Solutions for Future Main Memory,”


Samsung Semiconductor Inc. Various Datasheets:

J. Handy, “Where Silicon is Headed and Why You Need to Know,” Objective Analysis: http://www.media-tech.net/usa-09.html

References