Low-Power, High-Bandwidth, and Ultra-Small Memory Module Design

R. Jacob Baker
rjacobbaker@gmail.com

Overview: This work proposes a novel DRAM module and interconnect architectures in an attempt to improve computing energy use and performance. A low cost advanced packaging technology is used to propose an 8 die and 32-die memory module. The 32-die memory module measures less than 2 cm$^3$. The size and packaging technique allow the memory module to consume less power than conventional module designs. A 4 Gb DRAM architecture utilizing 64 data pins is proposed. The DRAM architecture is inline with ITRS roadmaps and can consume 50% less power while increasing bandwidth by 100%. The large number of data pins are supported by a low power capacitive-coupled interconnect. The receivers developed for the capacitive interface were fabricated in 0.5 µm and 65 nm CMOS technologies. The 0.5 µm design operated at 200 Mbps, used a coupling capacitor of 100 fF, and consumed less than 3 pJ/bit of energy. The 65 nm design operated at 4 Gbps, used a coupling capacitor of 15 fF, and consumed less than 15 fJ/bit and order of magnitude smaller consumptions than previously reported receiver designs.
Why is Power Such a Big Deal?

• Let’s say that at any given time there are, at least, 1,000,000 people playing World of Warcraft (WoW, a very reasonable assumption)

• Further let’s say that the power consumed by each of these players for: processor, memory (DRAM), computer fan, hard disk drive, monitor(s), modem, remote servers, communication channels (e.g., satellites links), cooling, etc. is 1000 Watts (again a very reasonable assumption)

• More than 1 GW of power is needed at any time for people to play WoW. This is the amount of power generated by a small nuclear power plant!
Mobile Platform

- Motorola Atrix (Front) found in the Google Droid

- Linear Power Amplifier
- RF Power
- DC Power
- Hard Drive
- Accelerometer

< 63.5 mm

< 117.75 mm
Mobile Platform

- Motorola Atrix (Back)
  - Memory (DSP)
  - Memory & CPU
  - HSPA+ DSP
  - 802.11n & Bluetooth
  - Compass
Server Platform

- Intel Server Board S5502UR

- Memory Slots
Organization

• Main Memory Limitations
• Nano-Module
• Wide I/O DRAM Architecture
• High Bandwidth Interconnect
• Hybrid Memory Cube
Main Memory Limitations

• Datacenter sparsity masked power limitations
  – Power trend: Energy consumption doubled every 5 years

• Historical server power
  – ~50 W in 2000
  – ~250 W in 2008

• Server power breakdown
  – CPU: 37%, Memory: 17%
  – Trend is Memory power > CPU power

• Main memory power
  – More die per module
  – Less modules per channel
  – Higher bandwidth
Main Memory Limitations
Main Memory Limitations

- CPU power wall
  - Voltage scaling reached its limit
  - Multiple cores supplement performance gains
  - No “multi-core” for DRAM
- DRAM voltage scaling reaching its limit
  - Current rate increase > voltage reduction rate
  - Power increasing
- DRAM pre-fetch
  - Memory core operates at slower frequency
  - High power I/O devices and data-path
Main Memory Limitations

![Graph showing the performance of main memory over time with labels indicating trends such as 25%/year, 52%/year, 20%/year, etc.](image-url)
Main Memory Limitations

• DRAM inefficiencies increase cost and power
  – Processor cache increasing
  – Intel Nehalem processor
  – DRAM would need to have L3 BW and latency
  – “…create the illusion of a large memory that we can access as fast as a very small memory.” – Patterson & Hennessy

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>RAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read BW [GB/s]</td>
<td>45.6</td>
<td>31.1</td>
<td>26.2</td>
<td>10.1</td>
</tr>
<tr>
<td>Write BW [GB/s]</td>
<td>45.6</td>
<td>28.8</td>
<td>19.9</td>
<td>8.4</td>
</tr>
<tr>
<td>Latency [ns] (cycles)</td>
<td>1.3 (4)</td>
<td>3.4 (10)</td>
<td>13.0 (38)</td>
<td>65.1 (191)</td>
</tr>
</tbody>
</table>
Main Memory Limitations

- DRAM efficiencies increase performance
- Capacity versus Performance
- Capacity costs power
  - Multiple memory channels
  - Each additional module increases power
Main Memory Limitations

- Bandwidth versus performance
- Bandwidth costs power
  - Buffer on board
  - Multiple channels
Main Memory Limitations

- DRAM inefficiencies in practice
- Typical video/web server motherboard
  - 20+ layer PCB
  - 6 memory channels
- RDIMM
  - 10+ layer PCB
  - Maximum comp. count
Main Memory Limitations

• 12 RDIMM
  – Termination
    • 36 components per DIMM
    • 8 I/O per component
    • 2.7 W of termination power for a read/write per module
    • 32.4 W total termination power
  – Wordline firing
    • 100 ns activation rate
    • 8126 page size
    • 200 fF per bitline
    • 11.2 W total bitline sense amplifier power
• Sustaining performance gains through capacity and bandwidth increases power and cost – innovation required.
Nano-Module

• Goals
  – Purpose was to move labs into prototype generation
  – Required low cost, high bandwidth, and low power memory solution that can be used with capacitive coupled interconnects in advanced server architectures

• Module component count trends required a new approach

• Nano-module proposed
  – Low cost advanced packaging technology
  – Off-the-shelf memory components

• Results can be leveraged
  – NAND
  – Mobile
Nano-Module

- Literature review of high capacity memory stacks
- 1990’s
  - Multichip Modules
    - Realized planar space limitations
  - Val & Lemione
  - Irvine Sensors
- Solutions proposed in research
  - No industry due to memory hierarchy effectiveness
Nano-Module

- Memory stack technology gaining new attention
- Proposed in 2010 (more later about developments)
  - Samsung quad die with TSV
    - 80 μm pitch, 30 μm diameter, 300 TSV
    - $R_{TSV} = 5 \, \Omega$, $C_{TSV} = 300 \, fF$
- Pros:
  - Lower power, higher bandwidth
- Cons:
  - Cost, integration
Nano-Module

- Literature review revealed novel solutions
- Slant the die!
- Applicable to capacitive-coupled interconnects
Nano-Module

• Not the first to try it:
Nano-Module

• Controlled Impedance
  – All Signals 50 Ω controlled impedance
  – DQS and CLK 120 Ω differential impedance

• Trace Length Matching
  – All Data matched to worst case
  – All CLK matched to worst case
  – All Address/Command matched to worst case

\[
Z_0 = \frac{87}{\sqrt{\varepsilon_r} + 1.41} \ln \left( \frac{5.98H}{0.8W + T} \right)
\]

Microstrip
Nano-Module

• Size calculations

\[
\text{height} = t_{\text{sub thickness}} + t_{\text{connection}} + \frac{t_{\text{die thickness}}}{\cos \alpha} + \left( t_{\text{die width}} - t_{\text{die pad to edge}} \right) \sin \alpha
\]

\[
\text{width} = (\text{no. signals} - 1) t_{\text{con pitch}} + t_{\text{con diameter}} + 2 \left( t_{\text{die to pad edge}} + t_{wb_1} + t_{wb_2} \right)
\]

\[
\text{length} = 2 \left( t_{wb_1} + t_{wb_2} \right) + \sin \alpha \cdot t_{\text{die thickness}} + \frac{\left( \# \text{die} - 1 \right) t_{\text{die thickness}}}{\sin \alpha} + \cos \alpha \cdot t_{\text{die width}}
\]
Nano-Module

• Thermal option
  – Thermal conductivity
    • Silicon, Metals >> Mold Compound
    • Hot spots
    • Temperature gradient
Nano-Module

- Thermal option
  - Heat plate

Heat Removal Plate

High Thermal Conductive Material
Wide I/O DRAM Architecture

- 4 Gb DRAM
  - Meets ITRS predictions
- Edge aligned pads
- Page size reduction
- Low cost process
  - < 4 levels of metal
  - No impact to die size
  - No impact to array efficiency
- Move to 64 data pins
  - Report challenges
  - Propose innovations

<table>
<thead>
<tr>
<th>256M Array</th>
<th>256M Array</th>
<th>256M Array</th>
<th>256M Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row</td>
<td>Column</td>
<td>Column</td>
<td>Column</td>
</tr>
</tbody>
</table>

Chip Size = 71.4 mm²
Array Efficiency = 57.7%

7.0 mm

10.2 mm

0.4 mm
Wide I/O DRAM Architecture

- 4 Gb Edge DRAM
  - Centralized Row and Column
  - Smaller die
  - Higher efficiency
  - < 4 levels of metal

Chip Size = 68.88 mm²
Array Efficiency = 59.9%
Wide I/O DRAM Architecture

- **Challenges**
  - Number of metal layers
  - Global data routing
  - Local data routing

- **Proposals**
  - Split bank structure
  - Data-path design
  - Through bitline routing
  - SLICE architecture
  - Capacitive-coupled I/O
High Bandwidth Interconnect

- Capacitive-coupling
  - Increased bandwidth
    - Reduced ESD capacitance
    - Smaller I/O channel = more I/O
    - Removal of inductive channel
  - Low power
    - Reduced ESD capacitance
    - Low power Tx & Rx
  - Low cost
    - Simple
  - Alignment required
- Literature review
  - Revealed inefficiencies and lack of application
High Bandwidth Interconnect

- Proposed receiver design
  - Extreme low power
  - ~1 gate delay latency
  - ‘DC’ transmission
  - RTZ $\rightarrow$ NRZ
High Bandwidth Interconnect

- 0.5 µm CMOS design (proof of concept)
  - 5.0 V process
  - 50 fF poly-poly capacitor
  - 200 Mbps
  - 3 – 8 pJ/bit
  - 325 Gb/mm²

Unlabeled NMOS are 1.5/0.6
Unlabeled PMOS are 2.1/0.6
Units in µm
High Bandwidth Interconnect

- Chip micrograph
  - 1.5 mm x 1.5 mm
  - 9 structures

- Experimental results
  - Operate at $V_{TX} = 2.0$ V
  - 3 pJ/bit at 200 Mbps
High Bandwidth Interconnect

- 65 nm CMOS design (proof of scalability)
  - 1.2 V process
  - 15 fF metal-metal capacitor
  - 4 Gbps
  - 17 µm²
  - 227 Tbps/mm²
High Bandwidth Interconnect

- Die micrograph
  - 2 mm x 2 mm
- Experimental results
  - 2 Gbps @ 0.9V
  - 50 fF coupling capacitor

1.0 Gbps – 1.2 V  
2.0 Gbps – 0.9 V
High Bandwidth Interconnect

<table>
<thead>
<tr>
<th>Work</th>
<th>Process</th>
<th>Supply (V)</th>
<th>Data Rate (Gbps)</th>
<th>Coupling (fF)</th>
<th>Bandwidth (Gbps/mm²)</th>
<th>Energy (pJ/bit)</th>
<th>Requires CLK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kanda 2003 [31]</td>
<td>0.35 µm</td>
<td>3.3</td>
<td>1.27</td>
<td>10</td>
<td>2.117 × 10³</td>
<td>2.4</td>
<td>Yes</td>
</tr>
<tr>
<td>Franzon 2006 [27]</td>
<td>0.18 µm</td>
<td>1.8</td>
<td>3</td>
<td>150</td>
<td>5.55 × 10²</td>
<td>5.0</td>
<td>No</td>
</tr>
<tr>
<td>Fazzi 2007 [32]</td>
<td>0.13 µm</td>
<td>1.2</td>
<td>1.23</td>
<td>10</td>
<td>1.922 × 10⁴</td>
<td>0.14</td>
<td>Yes</td>
</tr>
<tr>
<td>Kim 2009 [33]</td>
<td>0.18 µm</td>
<td>1.8</td>
<td>2</td>
<td>600</td>
<td>6.90 × 10²</td>
<td>0.8</td>
<td>Yes</td>
</tr>
<tr>
<td>This Work</td>
<td>0.5 µm</td>
<td>5.0</td>
<td>0.2</td>
<td>50</td>
<td>3.25 × 10²</td>
<td>8</td>
<td>No</td>
</tr>
<tr>
<td>This Work</td>
<td>65 nm</td>
<td>1.2</td>
<td>4</td>
<td>15</td>
<td>2.268 × 10⁵</td>
<td>0.015</td>
<td>No</td>
</tr>
</tbody>
</table>
So What is the Industry Moving Towards?

- Hybrid Memory Cube

Hybrid Memory Cube (HMC)

Notes: Tb/s = Terabits / second
HMC height is exaggerated
Hybrid Memory Cube

HMC Near Memory – MCM Configuration

- All links are between host CPU and HMC logic layer
- Maximum bandwidth per GB capacity

Notes: MCM = multi-chip module
Illustrative purposes only; height is exaggerated
HMC “Far” Memory

- **Far memory**
  - Some HMC links connect to host, some to other cubes
  - Serial links form networks of cubes
    - the memory = the network
  - Scalable to meet system requirements
  - Can be in module form or soldered-down
  - Can form a variety of topologies e.g., tree, ring, double-ring, mesh

- **Future interfaces**
  - Higher speed electrical (SERDES)
  - Optical
  - Whatever the most appropriate interface for the job
HMC<sub>Gen1</sub>: Technology Comparison

**Generation 1 (4 + 1 memory configuration)**

<table>
<thead>
<tr>
<th>Technology</th>
<th>VDD</th>
<th>IDD</th>
<th>BW GB/s</th>
<th>Power (W)</th>
<th>mW/GB/s</th>
<th>pj/ bit</th>
<th>real pj/ bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDRAM PC133 1GB Module</td>
<td>3.3</td>
<td>1.50</td>
<td>1.06</td>
<td>4.96</td>
<td>4664.97</td>
<td>583.12</td>
<td>762</td>
</tr>
<tr>
<td>DDR-333 1GB Module</td>
<td>2.5</td>
<td>2.19</td>
<td>2.66</td>
<td>5.48</td>
<td>2057.06</td>
<td>257.13</td>
<td>245</td>
</tr>
<tr>
<td>DDRII-667 2GB Module</td>
<td>1.8</td>
<td>2.88</td>
<td>5.34</td>
<td>5.18</td>
<td>971.51</td>
<td>121.44</td>
<td>139</td>
</tr>
<tr>
<td>DDR3-1333 2GB Module</td>
<td>1.5</td>
<td>3.68</td>
<td>10.66</td>
<td>5.52</td>
<td>517.63</td>
<td>64.70</td>
<td>52</td>
</tr>
<tr>
<td>DDR4-2667 4GB Module</td>
<td>1.2</td>
<td>5.50</td>
<td>21.34</td>
<td>6.60</td>
<td>309.34</td>
<td>38.67</td>
<td>39</td>
</tr>
<tr>
<td>HMC, 4 DRAM w/ Logic</td>
<td>1.2</td>
<td>9.23</td>
<td>128.00</td>
<td>11.08</td>
<td>86.53</td>
<td>10.82</td>
<td>13.7</td>
</tr>
</tbody>
</table>

Simple calculation from IDD7 (SDRAM IDD4)

Real system, some with lower density modules

- 1Gb 50nm DRAM Array
- 90nm prototype logic
- 512MB total DRAM cube
- 128GB/s Bandwidth
- 27mm x 27mm prototype
- Functional demonstrations!
- Reduced host CPU energy

CMOSedu.com

Nano Memory Module – by R. Jacob Baker
Conclusions

• Nano-Module
  – Developed a new research direction for industry research labs
  – Developed initial motivation
  – Developed initial prototype

• DRAM Architecture
  – Demonstrated benefits of wide I/O topologies
  – Proposed several low power innovations
  – Provided application for novel interconnect technologies

• Capacitive-Coupled Receiver
  – Demonstrated low power receiver designs
  – Achieved 2 Gbps at < 15 fJ/bit in 65 nm

• Summarized industry direction – Hybrid Memory Cube
Questions
Appendix - PLL

• 65 nm test chip
  – PLL
  – PRBS generator
Appendix - PLL

• PLL
  – Phase detector
Appendix - PLL

- Charge pump

\[
A_{PD} = \frac{I_{PUMP}}{2\pi}
\]

\[
A_F = \frac{1 + sR_1C_1}{sC_1}
\]
Appendix - PLL

- Voltage controlled oscillator

\[ A_{VCO} = 2\pi \cdot \frac{f_{\text{MAX}} - f_{\text{MIN}}}{V_{\text{MAX}} - V_{\text{MIN}}} \]
Appendix - PLL

\[
\phi_{CLK_{OUT}} = V_{INVCO} \cdot \frac{A_{VCO}}{s}
\]

\[
\phi_{CLK_{D}} = \frac{1}{N} \cdot \phi_{CLK_{OUT}} = \beta \cdot \phi_{CLK_{OUT}}
\]

\[
A_{OL} = A_{PD}A_{F}A_{VCO}
\]

\[
H(s) = \frac{\phi_{CLK_{IN}}}{\phi_{CLK_{OUT}}} = \frac{A_{PD}A_{F} \frac{A_{VCO}}{s}}{1 + \beta A_{PD}A_{F} \frac{A_{VCO}}{s}} = \frac{A_{PD}A_{F}A_{VCO}}{s + \beta A_{PD}A_{F}A_{VCO}}
\]

\[
A_{F} = \frac{1 + sR_{1}C_{1}}{sC_{1}}
\]

\[
H(s) = \frac{\phi_{CLK_{IN}}}{\phi_{CLK_{OUT}}} = \frac{A_{PD}A_{VCO} \left(1 + sR_{1}C_{1}\right)}{s^{2} + s \left(A_{PD}A_{VCO}R_{1}\frac{1}{N}\right) + \left(A_{PD}A_{VCO}NC_{1}\right)}
\]

\[
\omega_{n} = \sqrt{\frac{A_{PD}A_{VCO}}{NC_{1}}}
\]

\[
\zeta = \frac{\omega_{n}}{2R_{1}C_{1}}
\]
Appendix - PLL

- PLL at lock
Appendix - PLL

• PLL layout
Appendix - PLL

- PRBS generator
Appendix - PCB

- PCB test board
Appendix – Dead Bug
Appendix – Dead Bug

[Graphs of waveforms]
Appendix – 65 nm Chip
References


[3] Uksong Kang; Hoe-Ju Chung; Seongmoo Heo; Duk-Ha Park; Hoon Lee; Jin Ho Kim; Soon-Hong Ahn; Soo-Ho Cha; Jaesung Ahn; DukMin Kwon; Jae-Wook Lee; Han-Sung Joo; Woo-Seop Kim; Dong Hyeon Jang; Nam Seog Kim; Jung-Hwan Choi; Tae-Gyeong Chung; Jei-Hwan Yoo; Joo Sun Choi; Changhyun Kim; Young-Hyun Jun; , "8 Gb 3-D DDR3 DRAM Using Through-Silicon-Via Technology," *Solid-State Circuits, IEEE Journal of*, vol.45, no.1, pp.111-119, Jan. 2010


References