PicoServer: Using 3D Stacking Technology To Enable A Compact Energy Efficient Chip Multiprocessor

Taeho Kgil, Shaun D’Souza, Ali Saidi, Nathan Binkert, Ronald Dreslinski, Steve Reinhardt, Krisztian Flautner, Trevor Mudge

Advanced Computer Architecture Lab, University of Michigan, In collaboration with HP Labs, Reservoir Labs, ARM
Motivation

- Vast amounts of servers required
  - AOL, Google, Yahoo maintain large datacenters
  - General purpose processors not efficient to handle server workloads

- Opportunities with 3D stacking technology
  - Extreme integration
  - Improved throughput and latency

- Leverage 3D IC to build energy efficient Tier 1 servers
  - Tier 1 workloads require high memory throughput and modest ILP
  - CPU, Memory Controller, NIC, on-chip DRAM altogether in a single package
Outline

- **Background**
  - Server platform
  - Advances in technology - 3D stacking & DRAM

- PicoServer Architecture

- Methodology

- Results

- Conclusions and Future Work
3 Tier Architecture

HTTP Request → Tier 1

Tier 1 handles HTTP

HTTP Response

Invoke Component

Invoke Query

Response

Internet

IP Services

Application Services

DataBase Services
## Behavior of Commercial Server Workloads

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Web99</th>
<th>SAP 2T</th>
<th>TPC-H</th>
<th>TPC-C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Application Category</strong></td>
<td></td>
<td>ERP</td>
<td>DSS</td>
<td>OLTP</td>
</tr>
<tr>
<td>Web Server</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ILP</strong></td>
<td>Low</td>
<td>Med</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td><strong>TLP</strong></td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Working-set size</td>
<td>Large</td>
<td>Med</td>
<td>Large</td>
<td>Large</td>
</tr>
<tr>
<td>Data-sharing</td>
<td>Low</td>
<td>Med</td>
<td>Med</td>
<td>High</td>
</tr>
</tbody>
</table>

From S.R Kunkel et al, IBM J. R&D vol. 44 no.6, 2000
What is 3D stacking technology? – using 3D vias to connect multiple dies

Face to Back

- Bulk Si
- Active

Face to Face

- Bulk Si
- Active

- TSV vias: 20~100µm
- F2F vias: 5~30µm
3D stacking pros and cons

- High bandwidth (throughput)
  - Millions of die to die connections

- Reduces interconnect length
  - Interconnect becoming a problem as feature sizes shrink

- Extreme integration of components manufactured from different process technology
  - DRAM, Flash Memory, Analog, RF circuits etc

- Thermal problems
  - Power density limits the number of stacks

- Chip verification & Yield
  - Verification at the die, wafer and post-package level is necessary
  - Overall Yield is a product of individual die yield and 3D stacking yield
### Roadmap for 3D stacking and DRAM - Where are we?

<table>
<thead>
<tr>
<th></th>
<th>2005</th>
<th>2007</th>
<th>2009</th>
<th>2011</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of stack max. for low-cost / handheld - 3W power budget</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>Number of stack max. for high performance</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Cell Density of SRAM MBytes / cm²</td>
<td>11</td>
<td>17</td>
<td>28</td>
<td>46</td>
<td>74</td>
</tr>
<tr>
<td>Cell Density of DRAM MBytes / cm²</td>
<td>153</td>
<td>243</td>
<td>458</td>
<td>728</td>
<td>1,154</td>
</tr>
</tbody>
</table>

From ITRS 2005 Roadmap
Outline

- Background
- **PicoServer Architecture**
  - Overall Architecture
  - Architecture of logic components
  - Architecture of interconnect
  - Role of on-chip memory
- Methodology
- Results
- Conclusions and Future Work
PicoServer Architecture – Using simple cores with simple interconnect

3D vias

Main memory Including primary file cache

System Level components

Heat sink

Logic to Memory – F2F via, Memory to Memory – TSV via
Extreme integration and NUMA

CMP with 3D stacking
PicoServer and 3D stacking

- No need for L2 cache
  - Access latency and bandwidth of on-chip DRAM similar to a L2 cache
  - Additional cores can replace the L2 cache

- High performance low power interconnect
  - High bandwidth memory to core interface
  - The added degree of freedom reduces interconnect length

- Multicores clocked at modest frequency (500MHz)
  - Tier 1 server workloads are not computationally intensive
  - TLP more of an issue

- On-chip memory
  - Server applications → on-chip DRAM
  - Hundreds of MB of DRAM can be integrated on-chip
    - Additional memory can be available externally
Using Scalar Cores and Intelligent NICs

- Core and NIC

- Simple 5 stage pipeline clocked at low frequency – 500MHz
  - Maintain a reasonable power density to stack many die layers.
  - Opportunities to use low power process technology and DVS

- Standard branch predictor
  - 90 ~ 95% branch prediction

- ISA support for multicores

- Integrated DRAM controller per core to interface with on-chip memory

- Intelligent NICs are required to do load balancing
  - Load balancing achieved with Microsoft RSS like methods
Shared simple interconnect

- More than 70% of interconnect traffic is due to cache misses
  - Interconnect should handle **cache miss traffic** better than other types of traffic.
- Low frequency wide bus provide high throughput & low transfer latency
  - 3D stacking enables high throughput low frequency interconnect to on-chip DRAM
  - Simulations suggested a wide shared bus produced sufficient performance
    - Minimal queue delay in wide shared bus
The role of on-chip DRAM

- Niagara unloaded L2 cache access latency : 19ns
- Xeon unloaded L2 cache access latency : 8ns
The role of on-chip DRAM (cont.)

- A large portion of main memory is used as disk cache
  - Less than 64MB occupied by application, OS
  - Similar memory usage also reported in many server applications
- 100’s of MB of on-chip DRAM is enough to hold code & data and a portion of disk cache
Outline

- Background
- PicoServer Architecture
- **Methodology**
- Results
- Conclusions and Future Work
Methodology

- Full-system simulator M5
  - Models client-server connection
  - Generated client requests that saturate processor utilization in the server
- SURGE (static web), SpecWeb99 (dynamic web), Fenice (video streaming) and dbench (file serving) for Tier 1 server workloads
- Relied on empirical measurements from ISSCC, IEDM papers and datasheets to estimate power
- Calibrate empirical measurements with ITRS roadmap predictions, scaling rules and analytical FO4 model (for processor)
  - Overestimate most values to be on the safe side
Outline

- Background
- PicoServer Architecture
- Methodology
- Results
  - Overall Network Bandwidth - Mbps
  - Overall Estimated Total Power
  - Energy Efficiency
- Conclusions and Future Work
Additional cores yield improvement in Network Performance while operating at half the frequency.
Overall Estimated Total Power

PicoServers consume 2~3× less power

Similar die area

Similar die area
Energy Efficiency Pareto Chart

10x more energy efficient than OO4-large

PicoServers with similar die area are 2~3x more energy efficient than conventional CMP
Outline

- Background
- PicoServer Architecture
- Methodology
- Results
- Conclusions and Future Work
Conclusions & Future Work

- 3D stacking complements Tier 1 server workloads
  - High throughput memory bandwidth
  - More Processing Elements on die
  - Extreme integration for small form factors

- Simple multicores generate acceptable network bandwidth while consuming low power
  - For a 3W budget, 0.6~1.4Gbps network bandwidth

- Future Work
  - Investigate core architecture for computation intensive server workloads
  - Investigate energy efficient NUMA architectures for datacenter platforms
Questions???

???

???
System Level Power consumption

SunFire T2000 Power running SpecJBB

Power-wise
- Processor power 25% of total power
- Memory power 22% of total power
- I/O power (Mainboard, Gigabit Ethernet NICs, I/O pad, PCB interconnect) 22% of total power
- Misc. power (Fans + power supply) 25% of total power

Total Power 271W

From Sun talk given by James Laudon
### 3D via parameters

<table>
<thead>
<tr>
<th></th>
<th>Tezzaron 2(^{nd}) Generation</th>
<th>Tezzron Face to Face</th>
<th>RPI</th>
<th>MIT 3D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td>1.2µ x 1.2µ</td>
<td>1.7µ x 1.7µ</td>
<td>2µ x 2µ</td>
<td>1µ x 1µ</td>
</tr>
<tr>
<td><strong>Minimum Pitch</strong></td>
<td>N / A</td>
<td>0.4µ</td>
<td>N / A</td>
<td>N / A</td>
</tr>
<tr>
<td><strong>Through Capacitance</strong></td>
<td>2~3fF</td>
<td>&lt;&lt;</td>
<td>&lt;&lt;</td>
<td>2.7fF</td>
</tr>
<tr>
<td><strong>Series Resistance</strong></td>
<td>&lt;0.35Ω</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
</tbody>
</table>

A 3D via delivers minimal delay overhead & about the size of a 90nm 6T SRAM cell. Via density exceeds 14,000/mm\(^2\)

Numbers from Tezzaron Semiconductor, RPI, MIT
Evaluation of a Wide shared Bus

- Cacheline size = bus width

Increasing cacheline size reduced overall cache miss rates

- A data bus width of 1024 bits produced optimal results
The role of on-chip DRAM

- Niagara unloaded L2 cache access latency: 19ns
- Xeon unloaded L2 cache access latency: 8ns

R. Matick IBM
Improving word line delay

- Word line delay depends on the resulting RC caused by the large number of gates.
- One solution in reducing RC delay is by dividing the word line into smaller sections and to add buffers. However, additional drivers and buffers add area.
- Another solution is to route the word lines in metal rather than polysilicon or silicide. Independent studies show that aluminum word lines reduce wordline delay by 3x [Tanabe92].
- On-chip DRAM enables one to reallocate die area that was previously assigned to I/O & address multiplexing to improving word line delay with the above solutions.
Example timing diagram – DRAM read

$\text{t}_{\text{RC}} = 5 \text{ cycles}$
## Commonly used configurations

<table>
<thead>
<tr>
<th>Syntax</th>
<th>General Purpose Processor</th>
<th>PicoServer</th>
<th>Conventional CMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>OO4-&lt;small,large&gt; w/ w/o 3D stacking</td>
<td>Pico MP&lt;# of cores&gt; – &lt;freq&gt;</td>
<td>MP &lt;# of cores&gt; w/o 3D stacking</td>
</tr>
<tr>
<td>Operating Frequency</td>
<td>4GHz</td>
<td>500MHz / 1GHz</td>
<td>1GHz</td>
</tr>
<tr>
<td>Number of Processors</td>
<td>1</td>
<td>4, 8, 12</td>
<td>4, 8</td>
</tr>
<tr>
<td>Processor Type</td>
<td>Out-of-Order</td>
<td>In-order</td>
<td>In-order</td>
</tr>
<tr>
<td>Issue width per core</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>L1 cache size</td>
<td>2 way 16KB or 128KB</td>
<td>4 way 16KB</td>
<td>4 way 16KB</td>
</tr>
<tr>
<td>L2 cache size</td>
<td>8 way 256KB or 2MB 25 cycle hit latency</td>
<td>N/A</td>
<td>8 way 2MB 16 cycle hit latency</td>
</tr>
<tr>
<td>Memory bus width</td>
<td>64 bit @ 400MHz / 1024 bit 250MHz</td>
<td>1024 bit 250MHz</td>
<td>64 bit @ 333MHz</td>
</tr>
<tr>
<td>NIC location</td>
<td>PCIBus</td>
<td>Memory Bus</td>
<td>Memory Bus</td>
</tr>
</tbody>
</table>
Overall Network Bandwidth – Mbps

- w/o L2 cache & w/o 3D stacking
- impact of L2 cache
- impact of 3D stacking

- 33% better performance
- 11% better performance

Similar die area

Specweb99
Energy Efficiency Pareto Chart

![Diagram showing energy efficiency comparison between different server models (Conventional, CMP, PicoServer) with Specweb99 benchmark results.]

- **Conventional**
  - MP8 w/o 3D
  - MP 4 w/o 3D

- **CMP**
  - OO4-large
  - 4000M

- **Optimal**
  - Pico MP12 500M
  - Pico MP8 500M

Axes:
- Y-axis: Mbps (throughput)
- X-axis: Mb/J (energy efficiency)