A CMOS Vector Processor with a Custom Streaming Cache

Greg Faanes gjf@cray.com

Silicon Graphics
A CMOS Vector Processor With a Custom Streaming Cache

Overview:

- Goals and ground rules for the design
- CPU/Cache overview
- CPU chip
  - Scalar unit design
  - Vector unit design
- Streaming cache chip
  - Cache unit design
- Summary
Goals and Ground Rules for the Design

- Built to run large scale scientific applications
- Good vector performance
- High bandwidth to cache and main memory
  - unit stride
  - non-unit stride
  - gather/scatter
- YMP upward compatible
  - YMP ⇒ J90 ⇒ J90se ⇒ SV1 processor
- Few resources
  - 3 logic designers
  - “Off the shelf” technology
Not Your Ordinary Processor

- **Not** built to run SPECINT95 fast
- Different cache and system interface
  - 1 set of pins for both cache and memory
- High performance through multiple parallel and pipelined functional units
# SV1 Chipset

<table>
<thead>
<tr>
<th></th>
<th>CPU Chip</th>
<th>Cache Chip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock Frequency</td>
<td>250-300 Mhz</td>
<td>250-300 Mhz</td>
</tr>
<tr>
<td>Performance</td>
<td>1.0-1.2 GFLOP/sec</td>
<td>128 KB and 4.0-4.8 GB/sec</td>
</tr>
<tr>
<td>Die Size</td>
<td>14.6 mm x 14.6 mm = 213 mm²</td>
<td>13.7 mm x 13.7 mm = 188 mm²</td>
</tr>
<tr>
<td>Technology</td>
<td>CMOS, 5 layer metal, 2.6 v</td>
<td>CMOS, 5 layer metal, 2.6 v</td>
</tr>
<tr>
<td>Transistors</td>
<td>9.7 million</td>
<td>14.9 million</td>
</tr>
<tr>
<td>Power</td>
<td>12 watts</td>
<td>8 watts</td>
</tr>
<tr>
<td>Signal Pins</td>
<td>635</td>
<td>601</td>
</tr>
</tbody>
</table>
Scalar Unit

- Icache
  - 16-bit instruction parcels
  - 1/2/3 parcel instructions
  - 2 KByte size
  - 8 way associativity
  - 256 Byte lines
  - Branch prediction - always predict not taken
- Scalar
  - 8 32-bit Address Registers  \( \text{Areg} \)
  - 64 32-bit B Registers  \( \text{Breg} \)
  - 8 64-bit Scalar Registers  \( \text{Sreg} \)
  - 64 64-bit T Registers  \( \text{Treg} \)
  - Total of >1KB of scalar registers
  - 1 instruction per CP
  - No scalar cache
A CMOS Vector Processor with a Custom Streaming Cache

G. Faanes

Hot Chips 98
Vector Unit

- 8 64-bit Vector Registers \( V_{\text{reg}} \)
- 64 elements per \( V_{\text{reg}} \)
- Total of 4KB of vector registers
- 1 64-bit Vector Mask \( V_{\text{M}} \)
- 1 Vector Length Register \( V_{\text{L}} \)
- Full chaining and tailgating
- 1 instruction per CP
- 2 FP add units
- 2 FP multiply units
- 2 FP reciprocal units
- 2 integer add units
- 4 logical units
- 2 shift units
- 2 pop/parity/leading zero units
- 1 BMM unit - 2 results per cp
A CMOS Vector Processor with a Custom Streaming Cache
Hot Chips 98

G. Faanes
Streaming Cache Chip

- 128 KB
- Caches instructions, scalar data, and vector data
- 8 bytes per line
- 4-way associative
- 4.0-4.8 GB/sec bandwidth
- Write allocate/write through
- LRU replacement
- Scalar reference prefetch
- Software cache coherence
- Fast cache invalidation
- Parity protection for data and tags
- 192 outstanding references to memory
A CMOS Vector Processor with a Custom Streaming Cache

G. Faanes
Hot Chips 98

13
Summary

- Built to run large scale scientific applications
- YMP compatible
- Good vector performance
  - 1.0-1.2 GFLOP/sec and 3.0-3.6 GOP/sec
- High and flexible memory bandwidth
  - 8.0-9.6 GB/sec stride independent bandwidth from cache
  - 8.0-9.6 GB/sec stride independent bandwidth from memory
- “Off the shelf” technology
- 250-300 Mhz