The Vector Coprocessor Unit (VU) for the CM-5

Jon Wade, Zahi Abuhamdeh, Dan Cassiday, Mahesh Ganmukhi, Jeff Hill, Bob Lordi, Margaret St.Pierre, Scott Smith, Monica Wong, Shaw Yang, Bob Zak: Thinking Machines Corporation


Hot Chips IV Symposium
August 11, 1992

Presentation Outline

- Design Objectives
- Architecture
- Implementation
- Performance
Design Objectives

- Maximum sustained MFLOPS/$
  * memory bandwidth
  * FLOPS (IEEE)
  * instruction bandwidth
  * low startup cost
- direct fast-page DRAM support, strided and indirect memory & register addressing
- high performance FPU
- vectorized instructions
- minimize bubbles in pipeline
- JTAG, full internal scan

Design Objectives (cont.)

- System Configurability
  * support future SPARC's
  * memory size
  * expandable
- use MBus interface
- 4 to 64 Mbit DRAMs
- 2 to 8 VUs per node

- Software
  * compilers & libraries
- support data parallel programming model

- Time to Market
  * existing technologies
  * existing designs
  * thorough verification
- TI EPIC-2 process, CPGA, DRAMs
- TI megacells, gate-array
- transistor, gate, arch & system level
CM-5 Processing Node with Vector Units

- SPARC Micro-processor
- CM-5 Network Interface
- 64-bit MBus
- Four vector units also serve as memory controllers
- DRAM

Execution Model

- Memory mapped into MBus address space
- Remote accesses
  - Memory
  - Vector Unit state (register file, control state)
- Instructions
  - May be issued to one, a pair or all VUs on a node (decoded from MBus address)
  - Load/store memory base address is decoded from MBus address
  - Instruction is decoded from MBus write data
## CM-5 Processing Node Address Mapping

<table>
<thead>
<tr>
<th>name</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>main memory</td>
<td>N 0000 0000</td>
<td>N 07FF FFFF</td>
</tr>
<tr>
<td>data registers</td>
<td>N 4000 0000</td>
<td>N 4000 01FF</td>
</tr>
<tr>
<td>instruction register</td>
<td>N 8000 0000</td>
<td>N 87FF FFFF</td>
</tr>
<tr>
<td>read only memory</td>
<td>F FN00 0000</td>
<td>F FN7F FFFF</td>
</tr>
<tr>
<td>control registers</td>
<td>F FN80 0000</td>
<td>F FNFF FFFF</td>
</tr>
<tr>
<td>N1 registers</td>
<td>0 0800 0000</td>
<td>0 080F FFFF</td>
</tr>
</tbody>
</table>

where N is the VU ID number (0 to 7, 8 for common space).
Vector Unit Instructions

- load/store architecture (load chaining)
- concurrent arithmetic & load/store operations
- single & double-precision FP, 32- & 64-bit integers
- floating-point (div, sqrt), integer & logical operations
- triadic multiply-adds (floating-point & integer)
- vectorized instructions with strided or indirect addressing of memory & register file, vector lengths from 1 to 16
- elemental masking of load/store & arithmetic functions

Vector Instruction Format: Short

```
<table>
<thead>
<tr>
<th>VS-S</th>
<th>LS-F</th>
<th>rLS</th>
<th>ALU-F</th>
<th>rS1</th>
<th>rS2</th>
<th>rD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>
```

- vector length and register stride select
- LOAD/STORE function
- LOAD/STORE register address
- ALU function
- ALU source 1 register address
- ALU source 2 register address
- ALU destination register address
VU Register File Addressing

- Vector Registers
  e.g. 4 vector registers: 16 x 64-bit, 16 x 32-bit elements
  8 vector registers: 8 x 64-bit, 16 x 32-bit elements
  16 vector registers: 4 x 64-bit, 8 x 32-bit elements

- Arbitrary Base Address & Striding
  wrap around ends of register file, negative striding

- Indirect Addressing for Operand-1
  arbitrary address sequence

Vector Instruction Format: Long

Long format extends the 32-bit short format with an extra 32 bits. This is further decoded to specify these operations and modes:

- immediate operands (integer or floating-point)
- arbitrary memory stride
- arbitrary register base addresses and strides
- indirect addressing of memory & register file
- change default vector length
- vector mask conditionalization control
- arithmetic result exchange between paired vector units
Technology Summary

Number of Transistors: 1.0M
Die Size: 14.7 x 15.8 mm²
Package: 319 Pin CPGA
Operating Frequency: 40MHz Worst Case
Power Dissipation: 5.0W (typical @ 40MHz)
Process Technology: 0.8um CMOS, metal pitch
1st, 2.0um; 2nd, 2.0um
Chip Features
(2 Vector Units)

- Peak FP performance: 80 MFLOPs (IEEE SP, DP)
- MBus bandwidth: 320 MB/s
- DRAM bandwidth: 320 MB/s
- Register file: 128 x 64-bit, 256 x 32-bit, 5-read, 3-write ports
- DRAM support: 16 MB to 256 MB

Vector Instruction Pipelining

The node microprocessor (μP) issues vector instructions. The timing need not be exact. The μP overlaps scalar control computations with vector processing.

Consecutive vector instructions overlap head-to-tail. After the first vector, vector startup time is zero when vector instructions are issued quickly enough.
Benchmark Test Results

This slide will list the performance of the vector unit for several numerical benchmarks.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>MFlop</th>
<th>32MFP</th>
<th>90 Pct.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LL #1</td>
<td>48</td>
<td>61</td>
<td>75%</td>
</tr>
<tr>
<td>MFLOP</td>
<td>82</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>MATMUL</td>
<td>58</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>MATMUL (V3)</td>
<td>58</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>FFT (N=512)</td>
<td>40</td>
<td>62</td>
<td></td>
</tr>
</tbody>
</table>

Thinking Machines Corporation

6.1.10