Analysis of the MasPar MP-1 architecture
Bradley C. Kuszmaul
March 26, 1990

Please don't spread this document around outside of Thinking Machines Corporation. There may be mistakes, which are my fault, and if widely distributed, such mistakes might make Thinking Machines look bad.

This analysis is mostly based on a talk given at the MIT VLSI Seminar, March 13, 1990.
"VLSI Architecture for a Massively Parallel Computer"
Won S. Kim and John Zapisek
MasPar Computer Corporation
about the MasPar MP-1 computer.

Won S. Kim is the processor chip implementor.
John Zapisek is the router chip implementor.

I have tried to be clear about what is conjecture, and what is 'stated fact'. When I say that something is 'stated fact', I mean that Mr. Kim or Mr. Zapisek said it is true. When I say something is 'conjecture', I mean that I have reverse-engineered from the stated facts to hypothesize about their machine. When I do 'analysis', it is, unless otherwise stated, based on stated-fact. Unless indictated otherwise, the statements in this document are stated-fact. All of the analysis is based on the talk, rather than on other sources of information (i.e., this document is free from the effects of industrial espionage).

The processor chip architecture:
Each processor chip has 32 4-bit PE's.
There are 48 32-bit registers per PE (on chip memory)
There are 16K bytes DRAM/PE (off-chip) (64K bytes with 4Mbit DRAM)
The PE clock rate is 70ns (I saw 14Mhz elsewhere)
The memory is operated in fast-page mode at 80ns/byte
The processors are 'clustered' into 16 PE's per cluster.
(See Figure 1)

****

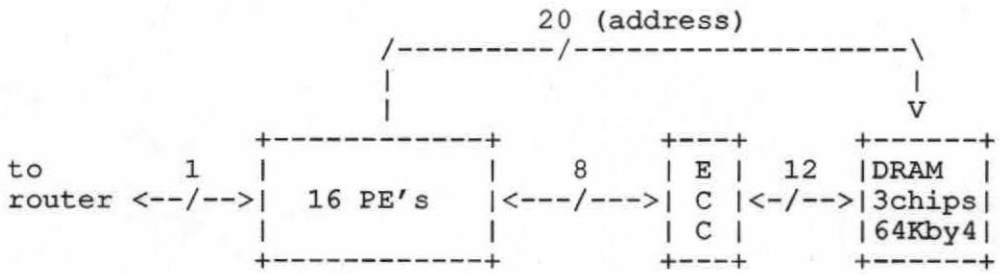

FIGURE 1: Cluster Architecture

****

For each cluster of 16 PE's, there is
* a 1-bit bidirectional connection to the router,
* a 20-bit address line to the DRAM chips (3 chips)
* an 8-bit bidirectional data-bus from the PE's to the ECC module
* a 12-bit bidirectional data-bus from the ECC module to the DRAM chips.

The memory is ECC protected on a per-byte basis. They need 4 bits of ECC to protect 8 bits of memory. (One can deduce the fact that you need at least 4 bits by an information-theoretic argument: To do one-bit correction double-bit detection, you need at least 10 states: 1 state says everything

is OK, 1 state says double-bit error, and 8 states to say which of the eight bits are broken when there is a single bit error.)

The DRAM refresh is controlled by a memory-controller inside the PE cluster. They run their DRAM in fast page mode. It looks as though all memory addresses originate from the PE cluster (so that if you are doing direct-addressing from the instruction-stream, the memory-address comes on-chip and then is routed off-chip. Note that this is a very different design assumption than was made on the CM1 chip, where we believed we could not afford to bring the memory address on-chip. Also note that the 20-bit address line length is hardwired into their PE chip, giving the MP-1 a 1Mbyte/Cluster (64Kbyte/processor) memory size.

The machine can do message routing or floating point operations at the same time as it is doing a memory operation.

Begin Conjecture-Mode:

I suspect that the DRAM controller is capable of quickly accepting a memory operation from all of the PE's (i.e., a single 'read/write' bit is provided to the DRAM controller, and also 16 20-bit addresses, and 16 1-bit context-bits, and 16 N-bit values are provided, where N is a multiple of 8, ranging from N=8 to N=64). The DRAM controller then cycles through all of the selected processors, performing the memory operation a byte at a time. When finished, the DRAM controller raises the global-and line to indicate that it is done. Thus, no DRAM bandwidth is used for unselected processors (however, there is still a 'worst-case' behavior: The slowest DRAM in the system determines the execution speed of the whole system.) The SIMD programming model looks like:

- -issue a memory operation
- -issue floating point stuff
- -issue more floating point stuff
- -wait for the global-and line to go true, indicating that the memory operation is finished.
- -now issue another memory operation

I don't understand the discrepency between the 70ns PE clock and the 80ns memory clock. It may be that the memory actually is clocked relatively asynchronously compared to the rest of the system.

End Conjecture-Mode.

It takes about 200 clock cycles to do a 64-bit floating-point add.
It takes about 110 clock cycles to do a 32-bit floating-point add.
Floating point format: VAX (Format G?)

Mr. Kim believes that the processor chip price is very important. I don't know how important it really is, but it is interesting to note that adding more pins to support certain features was considered 'too expensive'

PE chip Implementation technology:

- 1.6u N-well CMOS, double metal
- 450,000 transistors
- the die is 415^2 square mils. (Not square, but the area of a 415mil by 415 mil square.)
- Power: 3/4 watts per chip
- Speed: Not critical path
- Vendor: Sierra
- Registers account for 2/3 of the transistors and 1/3 of the area
- floating point account for about 1/2 of the remaining transistors
- Clock: about 70ns
- Package: 164 pin quad flat pack

PCB: 10 layers

Analysis: The memory-bandwidth to floating-point bandwidth is pretty well matched. If they spend almost all of their time IN fast-page mode, they can do 1.5 memory operations per floating-point operation (which is a similar to the CM2, which can do 1 memory operation per floating-point operation). I am not convinced that the MP-1 PE is really the most effective way to get lots of floating-point operations. In 200 clock cycles, they can do 32 floating point operations per chip. This is 6.25 clocks per floating-point operation. Given the large number of registers available on their chip, it seems as though there might be some benefit to getting more raw floating-point speed. It seems as though they ought to have been able to get that number down better than 3 clocks per floating-point operation by building a single 64-bit floating-point unit, and time-sharing that among the PE's.

Router:
The router has a peak throughput of 1.6Gbyte/second
and an sustained throughput of 1.3Gbyte/second.

The router is organized as a 3-stage circuit-switched benes network. (It was refered to as a 'Ma Bell' network.)
There are three phases to each 'petit-cycle (my choice of terminology)

1) open the connection
2) send data
3) get an acknowledgement (with a parity or checksum)

Any sender whose message is blocked or garbled does not get an acknowledgement and the sender tries again. (Note that this is different than the CM2, because in the CM2 messages which are blocked are allowed to 'retain' the progress they have made towards their destination, while the MP-1 router forces messages to start over from their original source). Steps (1) and (3) account for around 24 clock cycles of overhead. (Note that this is a little more than the overhead for a CM2 petit-cycle, which is about 14 clock cycles of overhead.)

Analysis: The router bandwidth is about 1/10 of the memory bandwidth (even when operated in fast-page mode). This is substantially better than for the CM2 (where the router bandwidth is an order of magnitude smaller compared to the memory bandwidth).

News Network: The MP-1 has a full 2-dimensional news network (with something like 24 pins of their chip devoted to the 2-dimensional news network). It looks like the news network can do a good job at SCAN.

Performance
1K by 1K linpack in 2.3 seconds (about 300Megaflops) on a 16KPE MP-1

Mr. Kim stated that he feels that the MP-1 has low floating-point performance. He wanted to know about the teraops project, but did not want to ask me any 'sensitive questions'. I told him that I didn't mind, since I believe it is the responsibility of the question-asker to ask all the right questions and it is the responsiblity of the question-answerer to not let secrets leak out. Neither Mr. Kim nor Mr. Zapisek seemed to be prepared for the 'what are you going to do next?' question. They did not know that one should say something like "we are always thinking about the future, and I can't tell you the specifics". Instead, I found out

- Neither Mr. Kim nor Mr. Zapisek has any real ideas about how to change the architecture. Mr. Kim talked about using 8-bit processors. Mr. Zapizek is worried that he can't wire-up a 64KPE router, even though he can use the same chip to build a 64KPE router. Mr. Kim said something like ''We have thought about the next machine a little bit, but not much.''

Analysis: I believe that these two guys were being totally candid. The fact that neither of the chip implementors has much idea about what to do next does not really mean much. Much more significant is that they don't worry about keeping secrets, and don't know what secrets they should keep.

I suspect that this is because most people are more interested in Thinking Machine's next product than that of Maspar.