*** TMC CONFIDENTIAL --- DO NOT DISTRIBUTE ***

Notes on the DR

HellHiYa 92

Steve Heller, Jeff Hill,and Shaw Yang
2 April 92
Last revised: not

Distribution:

Mark Bromley
Dick Clayton
Dave Douglas
Carl Feynman
Rolf Fiebrich
Steve Heller
Jeff Hill
Lennart Johnson
Charles Leiserson
Woody Lichtenstein
Bradley Kuszmaul
Cindy Spiller
Margaret St. Pierre
Ted Tabloski
Jon Wade
Shaw Yang

*** TMC CONFIDENTIAL --- DO NOT DISTRIBUTE ***

Cascading Degradation in the Data Router

Steve Heller, Jeff Hill, and Shaw Yang
2 April 92
Last revised: not

==========
Revision History:

==========

This document describes each current approach to improving router performance. Issues of resource requirements as well as system softawre and hardware impact are addressed.

Outline:

1. Background: General Issues of Various Approaches
2. The Approaches
3. Modelling and Simulation
4. Summary

1. Background: General Issues of Various Approaches

1.1. Split Level Solutions

There is a theory that the whole DR need not be replaced, but levels one and two can be left alone. This not only leaves most DR chips alone (most chips are in levels one and two), but leaves the processor boards alone. The following chart shows the distribution of DR chips across the DR by levels. "high chips" are from levels three and up.

	high chip count	total chip count	high chip %age
256 nodes	64	256	25%
1K nodes	384	1152	38%
4K nodes	2048	5120	40%
16K nodes	10240	22528	45%

It also may be possible to replace just some up the upper levels. The effectiveness of neither variation on any solution has been quantified.

The nice thing about split level solutions is that we can use all current chips it levels one and two in future machines.

Split level solutions cannot combine DR chips running at two different speeds.

1.2. Adding DR Chips

Several of the proposed solutions increase the number of DR chips. There is probably room on the DR boards (height three and up) to double the number of DR chips, but there probably isn't enough roomto quadruple the number of chips, so cabinets would need to change.

Each DR chip requires a separate clock signal, so increasing the number of chips means adding clock buffer boards. There is probably room for one more clock buffer board, but not three more, supporting doubling the number of DR chips, but not supporting quadrupling the number. Also, both height 3-4 and height 5-6 DR backplane will need to be changed.

DR chips take power and require cooling. These issues must be addressed in conjunction with any solution that increases the number of DR chips.

All the latencies (number of hops) will change.

1.3. Cranking the Clock

Increasing the clock speed has several effects.

The NI chip must be respun, as the NI assumes that the DR and CN use the same clock. We can also respin the CN chip (making the NI respin simpler --- two vs three clocks), or not respin the CN chip.

There is a constraint in the NI that the clock of the DR must be as fast as the PN clock, but no more than 1.6 times faster. In order to increase the DR clock to 60 MHz, the processors would need to be sped up to 37.5 MHz. In order to increase the DR clock to 50 MHz, the processors would need to run at 31.25 MHz, which we alreadt do.

The clock DN board would need to be respun if there are multiple network clocks.

The cable lengths may need to change as well as the number of bits "stored" on the wires.

Cranking the clock precludes split level solutions.

1.4 DR Chip Respinning

We can do more bit stacking to allow for longer wires than we currently support. This may simplify packaging for large machines (4K).

Any respin of the DR chip can also address a power dissipation weakness in several parts of the DR chip.

Also, the cost of the chip will likely go down by a factor of two due to newer technology.

1.5. Chip/Board Changes

Any chip or board changes would require modification to the diagnostics to support the new configurations.

Topological changes would also impact the boot, partitioning, and timesharing code --- the changes don't appear major, especially as we won't be doing them ourselves.

2. The Approaches

2.1. Software Solution

Description: The router does not suffer cascading degradation on all patterns, but on most. A notable exception is at least some of the NEWS patterns. The class of non-congestive patterns is somewhat understood. It may be able to make some random patterns look like a sequence of non-congestive patterns by sorting (a very simple one pass partial-sort will do it) the messages locally before sending them. The idea is to use injection order to mittigate the bad patterns. It is not clear how well this will work (if at all) or if there are large weaknesses. It is difficult to determine automatically (no current ideas) if this approach will help for a particular communication pattern, and there is an overhead to trying it. It might make things worse in some situations, so it is probably not safe to use as the default.

Effectiveness: not quantified

Software effort: It should take about two weeks of study to determine if this helps some situations, and if it hurts others. It will take another two weeks to have production base level software. It will require additional work to integrate into higher level interfaces, like the compilers. This is entirely a software solution, but has both limitations and potential large weaknesses.

Hardware effort: none

2.2. Repeater Mode

Description: By placing a DR chip in repeater mode directly above each DR chip, the length of the input fifos are effectively increased. This may reduce the cascading effect, but the effectiveness of the solution is not known.

Effectiveness: not quantified. 2-3 weeks simulation to study.

Software effort: Diagnostics, maybe partitioning and timesharing

Hardware effort: DR boards, more clocks

Hardware cost: double number of DR chips

2.3. Multi Chip Solution

Description: There is an approach that uses four DR chips to replace each DR chip. A PAL would also need to be designed. We have no idea of how effective this will be.

Effectiveness: not quantified. 3-4 weeks simulation to study.

Software effort: Diagnostics, partitioning and timesharing

Hardware effort: DR boards, PAL design, more clocks

Hardware cost: quadruple number of DR chips, and a PAL. This won't fit in the current cabinets.

2.4. Cranking the Clock

Description: The chip can be reimplemented in .8 micron technology (currently in 1.0 micron), and the clock could be sped up at the same time to between 50 and 60 MHz. A new cell library as well as some custom work (phase locked loops) would be needed. As with any DR respin, power consumption could decrease by fixing power lossage.

Effectiveness: should improve the weak patterns, but won't improve the patterns that currently have no problem, as we are close to running up against the bandwidth limitations of the NI. A better NI will fix this, though.

Software effort: diags

Hardware effort: NI and possibly CN respin; maybe faster PNs; clock DN board (if no CN rework); cable lengths.

Hardware cost: new chips (cheaper)

2.5. Bigger Fifos

Increasing the on chip fifo sizes may mitigate the problem. If simply doubling the fifos does the trick, the repeater mode solution may be better. We believe that there is an interaction between the buffer sizes and the ability to, in the future, have NI. Without bigger buffers, a new NI probably cannot send large messages.

Effectiveness: not quantified. 2-3 weeks simulation study.

Software effort: diags

Hardware effort: DR chip

Hardware cost: new chips (cheaper)

2.6. Adding Buffers

This is the non-topological chip change that is felt to have the most potential.

Effectiveness: not quantified. 3-4 weeks simulaiton study.

Software effort: diags

Hardware effort: DR chip

Hardware cost: new chips (cheaper)

2.7. Fifos and Buffers

If we change the chip but don't go with a topological change, we probably want to add both longer fifos and buffers. The incremental cost of the different parts is small.

Effectiveness: not quantified. 3-4 weeks simulation study.

Software effort: diags

Hardware effort: DR chip

Hardware cost: new chips (cheaper)

2.8. Dialated Tree

A relatively small change in the chip would allow a different topology. This change would double the number of chips on a DR board, but effects would be kept on board. If we made this change, we'd probably also want to expand fifos (to accomodate future NI or NI/DMA), and we may wish to add buffers as well, if that helps performance (not known).

Effectiveness: not quantified. 2-3 weeks simulation study.

Software effort: Diagnostics, partitioning and timesharing

Hardware effort: DR boards and backplane, more clocks

Hardware cost: double chips

2.9. Multi Tree

This is the topology that is felt to have the most theoretical potential for performance. It also has the most pervasive system repercussions. A small change in the chip logic and a doubling of the number of DR chips, accompanied by on board changes and cable changes (basically a new DR network) should do much better in terms of utilization. It was thought that no cable changes would be necessary, but there is little to no flexibility in the backplane. As with the Dialated tree, we'd probably also want to expand fifos (to accomodate future NI or NI/DMA), and we may wish to add buffers as well, if that helps performance (not known).

Effectiveness: not quantified. 2-3 week simulation study.

Software effort: Diagnostics, partitioning and timesharing

Hardware effort: DR boards and backplane, more clocks, recable wires

Hardware cost: double chips

3. Modelling and Simulation

All the above approaches cannot be quantified well enough to make an intelligent decision without simulation. The initial abstract models predicted that buffers should help quite a bit, but they can only indicate trends.

Mark Bromley has already implemented the greater part of a packet-level simulator that can yield some more information on some approaches. We should be able to study some buffer and fifo approaches, but the results are still abstract and, while they can indicate trends, they cannot quantify the effectiveness. Initial (*very* preliminary) results concerning the effectiveness of some solutions are not encouraging --- there are likely still bugs.

Mark has also implemented a good bit of a flit-level simulator. It is felt that this level of simulation should be able to model the current machine well enough to predict actual performance figures, and should be able to project reasonably accurate figures for all of the above approaches.

Both of the above simulators are CnC codes that run on a CM5.

Simulation of all the options should be achievable on two staff months. Mark cannot spend additional time on this without significantly impacting his support of the DASH run-time. Carlf may be available --- and is at a breaking point in his current project. He seems an excellent candidate for the job. Margaret has the appropriate background for this kind of stuff, and this seems to tie in well with her current project.

4. Sumamry

Approach	chip design (staff months)	system HW (staff months)	time to alpha (mo)	system SW (mo)	model (weeks)	guess at effeciveness
1. Software	-		2	1	2	weak
2. Repeater	-		board	little	2-3	weak
3. Multi-Chip	-		PAL board	clock	3-4	
4. Clock	12 DR 16 NI	PN bd DR bd Clk DN	12	-	-	25%
CN	12 DR 12 NI 12 CN	-	10	-	-	25%
5. Fifos 1.0	10 DR	-	8	diag	2-3	weak
0.8	14 DR		10			
0.8/clock	16 DR 16 NI	see clock	13			
6. Buffers 1.0	14 DR	-	10	diag	3-4	good
0.8	16 DR		11			
0.8/clock	18 DR 16 NI	see clock	14			
7. F & B 1.0	16 DR	-	11	diag	3-4	good+
0.8	18 DR		12			
0.8/clock	20 DR 16 NI	see clock	15			
8. Dialated	10	DR	8	diag	2-3	good-
fifo	12	clocks	9	part		
buffers	12		9	TS		
F & B	14		10			
0.8						
0.8/clock						
9. Multi	10	DR	8	diag	2-3	good++
fifo	12	clocks	9	part		
buffers	12	cables	9	TS		
F & B	14	...	10			
0.8						
0.8/clock						