Cluster Versions of CrystalWave and OmniSim

Hardware Recommendations

❑ Introduction

If you do not yet have a cluster, then this document contains suggestions on how to choose a cluster which will run optimally on our CrystalWave and OmniSim Cluster Editions. Certain high end features available with some clusters may provide no benefit to you and waste your money. On the other hand some relatively cheep additions may substantially enhance your attained performance.

This document relates entirely to the clustering of the FDTD engines in CrystalWave and OmniSim.

❑ Compute Nodes: factors affecting FDTD cluster speed

Memory Bandwidth

FDTD is a memory intensive application. Therefore it is important to get as much memory bandwidth in the whole cluster as possible. Total memory bandwidth is given simply by:

(Memory bandwidth of one node) x (number of nodes)

You may see terms like PC2100 or PC2700 which means a memory bandwidth of 2100MB/s or 2700MB/s.

Below is a table showing the total memory bandwidth of different memory modules. Note that the actual speed at which the memory is run is also a function of the chipset.

				System Memory Bandwidth (max theoretical)		
Technology	Type	Alt Name	Nominal frequency	Single Channel	Dual Channel	Quad Channel
DDR-2	DDR2-400	PC2-3200	400MHz	3.2 GB/s	6.4 GB/s	12.8 GB/s
DDR-2	DDR2-533	PC2-4200	533MHz	4.2 GB/s	8.4GB/s	16.8GB/s
DDR-2	DDR2-667	PC2-5300	667MHz	5.3 GB/s	10.6 GB/s	21.2 GB/s
DDR-2	DDR2-800	PC2-6400	800MHz	6.4 GB/s	12.8 GB/s	25.6 GB/s
DDR-3	DDR3-1066		1066MHz	8.5 GB/s	17 GB/s	34 GB/s
DDR-3	DDR3-1333	PC3-10664	1333MHz	10.6 GB/s	21.2 GB/s	42.4 GB/s
DDR-3	DDR3-1600	PC3-12800	1600MHz	12.7 GB/s	25.4 GB/s	50.9 GB/s
DDR-3	DDR3-2000	PC3-16000	2000MHz	15.9 GB/s	31.8 GB/s	63.6 GB/s
RDRAM		PC1066	1066MHz	2.1 GB/s	4.2 GB/s	8.4 GB/s
XDR/XDIMM	XDR 3.2GHz		3.2GHz	6.4 GB/s	12.8GB/s	25.6GB/s

Table 1: Memory Bandwidths by Technology. N.b. the quoted nominal frequencies are not comparable – different technologies define the frequency in different ways.

Chipsets

The table below indicate the maximum memory bandwidth of common chipsets. Note that the AMD Opteron and Athlon and newer Intel CPUs have integrated memory controllers so we list the CPU rather than a chipset.

Chipset/CPU Model	Num Channels	Fastest Mem	Max Theoretical Memory Bandwidth
Intel E5-26xx-v2 (Dual socket capable)	8 (2xCPUs)	DDR3-1866	2x59.7 GB/s
Intel Core-i7 35xx, 36xx, 37xx, 47xx	2	DDR3-1600	25.4 GB/s
Intel X58/Core i7 9xx series (LGA-1366 socket CPUs)	3	DDR3-1066	25.6 GB/s
Intel Core i7 860, 870 (LGA-1156 socket CPUs)	2	DDR3-1333	21.2 GB/s
Intel Core i7 26xx, series	2	DDR3-1333	21.2 GB/s
Intel 5000X (for Xeon 50xx series)	4	FBDIMM-667MHz	21.2GB/s
Xeon 55xx and 56xx series CPUs	6 (2xCPUs)	DDR3-1333	64 GB/s [1]
Intel P965 (PentiumD or Core2Duo)	2	DDR2-800	12.8GB/s
Intel P45/P43/G45/G43 (Core2Duo)	2	DDR3-1066	17 GB/s
Intel X38/X48 (Core2Duo)	2	DDR3-1333	21.2 GB/s
Opteron “Istanbul” series 24xx series	4 (2xCPUs)	DDR2-800	25.6 GB/s
Opteron “Magny-Cours” 61xx series (G34 socket)	8 (2xCPUs)	DDR3-1333	85 GB/s [2]
Athlon 64 X2 (AM2 socket)	2	DDR2-800	12.8GB/s
AMD Phenom II (AM3 socket)	2	DDR3-1333	21.2 GB /s

Notes:
[1] some Xeon 55xx and 56xx motherboards have only 4 memory channels – reducing bandwidth by 33%
[2] some sources say this system will reach only 57GB/s due to “Northbridge limits”

CPU Speed
Obviously more is better but at some point it wont speed up your simulations if the memory cannot supply data fast enough to the CPU. This is often the case.

Multi-Core CPUs and multi-CPU nodes
These architectures increase CPU compute speed but do not generally increase the total memory bandwidth of the node (but see Opteron below). On many machines our FDTD engine uses most of the machine’s memory bandwidth. Therefore running two FDTD engine nodes on a dual CPU (or dual-core) machine you will not get the speed doubling might expect – typically compute speed will increase by from 1.2x to 1.6x.
Opteron systems work very differently from traditional Intel systems – each Opteron has a separate memory bus called HyperTransport. So if you have a dual-Opteron machine (not just a dual-core Opteron!) then you get double the total memory bandwidth and you will get close to the 2x speed-up.
The latest Intel **Xeon** CPUs now have an equivalent technology Intel calls QuickPath which is used by the Nehalem and Westmere range of Xeons and newer (Xeon 55xx and 56xx series and later) – see Intel 5500 chipset in table above. The Core-i7 CPU also uses QuickPath but you can’t put two Core-i7s in the same PC.

Network Performance
The faster the interconnect between nodes the smaller the simulation you can do efficiently. If you are doing a problem using 100MB or more per node, then your network is unlikely to be slowing your simulations down at all. However if you have a problem that is using only 500kB per node then network latency and bandwidth will likely be significant – the faster the network the better in that case.

From these points we can offer the following guidelines:

- Choose a compute node with a high memory bandwidth. This is a function of a) the memory speed in MHz and b) the chipset design - the chip that sits between the CPU and memory. Refer to Table 1 above.
- As of writing the majority of modern systems can use “DDR3“ memory. DDR3 can run at higher data frequencies than the older DDR2 – see Table 1 above.
- A “Dual Channel DDR3” design is faster than a standard DDR3 design. Basically “dual channel” can in principle *double* your memory bandwidth so worth looking out for! Similarly “Quad-Channel” will double the bandwidth again.

- Look also at the memory frequency supported by the system. 1333MHz is common now and some systems have 1600MHz memory frequencies. Again refer to the table above.
- Older Intel systems use a front side bus (FSB). Look at the FSB frequency. This controls the speed at which the CPU talks to the chipset. The faster the better. 800MHz is common now and some systems have 1024MHz.
- CPU frequency. Don’t pay for the fastest CPUs – they tend to be very expensive. Two nodes of 2.8GHz CPUs each with 2GB of memory will probably run almost twice the speed of one node with a 3.6GHz CPU and 4GB of memory.
- Level-2 and Level-3 cache sizes. A bigger cache may substantially speed up certain simulations and make little difference to others. This will be determined by the dimensions of your simulations – the number of FDTD grid cells in each direction.
- Hyperthreading. Currently we do not take advantage of hyperthreading. To do so would require more than one cluster node running on the cpu-core – it is better to have each node running on its own core.
- Network: we currently support only TCP/IP cluster interconnect and recommend a Gigabit Ethernet fabric. Older 100MB/s Ethernet links slow your simulations down in certain circumstances. We do not currently support Infiniband or Myrinet interconnects and in any case our view is that they are unlikely to speed up your simulations at all, except possibly in very unusual cases.
- Ethernet switches: buy a Gigabit Ethernet switch but do not spend money on the fastest low-latency switches – it won’t speed up your FDTD simulations except in very special circumstances. The important thing is for the total bandwidth of your switch to allow all nodes to communicate at or close to 1Gb/s at the same time. So if you have a 16 port switch it should have a bandwidth of 8Gb/s to 16Gb/s. Be careful of a switch topology with two layers of switches, such that some nodes are “closer” to each other than others. For example imagine a cluster of 64 nodes where groups of 16 nodes are connected to 4 switches and the 4 switches are then connected to each other. This may potentially create a bottleneck if e.g. 8 nodes connected to switch-1 wanted to simultaneously talk to 8 nodes connected to switch-2 – all 8 nodes would have to share a 1Gb/s link.

  Some switch vendors offer stacked switches where the interconnect bandwidth between switches in the stack is much higher than 1Gb/s. Anything above 8Gb/s is likely to be adequate for most applications.
- Hard disk: not important.

❑ Head Node (Linux)

- If the Head Node PC (Linux cluster) is not also hosting a Compute Node, then it will not need a lot of CPU power. However all the communication between the Compute Nodes and the Controller PC must pass through the Head Node so it will want good network connections – at least a 100MB/s Ethernet interface.
- It is recommended that the Head Node and the Cluster Nodes are all connected by an Ethernet Switch that does not host general network traffic. Even better would be for the Compute Nodes to be on a separate LAN and the Head Node have two Ethernet interfaces – one for the Cluster LAN and another for the general Office LAN that connects to the Controller (see diagram in fdtd_cluster_specs.pdf)