Wednesday, July 19. 2006

Do-it-yourself X4500 ?

I hate the big discussion forums at IT news sites. They are full of trolls. My favourite trollishness at the moment is the comment "The X4500 is fucking expensive. I can build one of my own for 10 grands". Okay, let´s talk about this statement.

Well, surely you can build one, perhaps much cheaper than Sun. But is it really the same? Or is it only a cheap ripoff of an X4500 that has only the capacity in common.

For an outsider the X4500 looks like an normal opteron server with 48 disks. But it starts with the mainboard. It isn´t a standard mainboard. Consider the follwing things:

In a Thumper all disks are directly coupled onto an disc controller. There are no SATA2-Port extenders in the way. At the end there are 6 seperate SATA2 controllers with 8 SATA-2-Ports each. We´ve used the Marvell 88SX60xx-series for this task.The widespread suggestion to use some Areca Cards to mimick an Thumper isn´t an option as well. It´s senseless to plug 24 discs on one controller that uses only one PCI-Bridge with roundabout 1.08 GByte per seond. Only a quick calculation: A Hitachi Deskstar was benched several times with a little bit more than 60 MBytes per Seconds. 24 of this discs results in a datastream a little bit short of 1500 Mbytes per second. So you can get only two thirds of the maximum performance through the PCI-X slots. And this is calculated with the peak max performance of PCI-X. So you have to use at three of this cards. I would even say, you need four or five to have headroom for the next generation high capacity harddrives (perpendicular recording accelerates the drives as well it increases their capacity)Now you have a different problem. Find a PCI-X board with 3 logical seperate PCI-X busses. Every configuration sharing busses will see the bottleneck described . So you have to find a bord with two AMD8132 Bridges, something i doubt you find on el-cheapo mainboards.The Thumper uses three HT-Channels to directly connect this dics controllers. 16 Ports shares one HT-Channel with 8 GBytes per second, as two Marvell-Chips are coupled to one HT-Channel. 8 discs with 60 MBytes per seconds gives you 480 MByte/s. Well inside your 1.06 GByte/s limit to have some headroom for the future. Two of this 960 MByte/s. Okay, well inside the limit of 8 MBytes/s per HT-Channel.

You should look at this diagram to get an overview of the special design of this board. It´s designed specifically for it´s task:

I havn´t even mentioned RAS-Feature, Lights-out-management, Airflow-Design, Serviceability (eg. changing the CPU-Unit).

Before somebody tell me again, that he can build a X4500 for 10 grands, please read the architecture white paper of the X4500. Until then, youshould think twice before making bold statements. And to all customers of such systems: When a competitor shows you a similar system, you should look with care on the internal architecture. Perhaps they will sell you the cheap ripoff.

Posted by Joerg Moellenkamp in English, Oracle at 13:23

4500 looks lice, but I think you also have to consider what happens when you have the disks hung off couple of PCIe cards, and whether, actually, many customers who want the capacity will also care so much about the peak performance.
Anonymous on Jul 27 2006, 12:46

Real question is if home-made 4-CPU box with 2-core 885 Opterons and 4 GbE and same 48 drives connected to 3-4 PCI-X controllers would not be both significantly (like 2x) cheaper AND faster. Power consumption would be higher only by about 20-25%
Anonymous on Jul 27 2006, 19:15

Okay, you can´t compare a 2 CPU to an 4 CPU-BOX, but this is only a side note.

But:
1. 20-25% can annihilate any cost advantage of a 3 year operation..

2. I don´t think it would be faster at I/O as general purpose boards have not the connectivity to do such stuff. You have to use an mainboard with 4 independent PCIX-busses and you have to be sure, that the onboard cpu and cache poses no bottleneck. Do you know a board with 4 independent busses ?
Anonymous on Jul 27 2006, 19:39

Don't get me wrong, I think very highly of Thumper from a technical point of view. Price is another matter.
1. Extra power consumption + cooling is about 400-500W, which results in extra 10-12 KWh/day - cost about $2 or less; so in 3 years it's about $2,000 extra or so, far smaller than $15-25,000 price difference we're talking about
2. 2 PCI-X busses should be able to provide about 1.5GB/s (2 PCI-X 100 controllers on each bus connected to 12 drives each - 2x2x12). Are you sure Thumper can actually provide more than that? and where this data will be going if there are only 2 GbE links?
Anonymous on Jul 27 2006, 20:20

We were able to get out 2.5 GBytes per second in peak. of of this system. So 1.5 Gigabytes per Second would not sufficient. And this was done with 60 MByte/s-Disks. Now imagine perpendicular recording discs, that have a even higher bandwidth. With 100 MByte/s or 120 MByte/s per disk the problem gets even worse. The Thumper has the headroom for such disks. A PCIX solution with would is already maxed ... with the slow discs.
Anonymous on Jul 27 2006, 20:48

What kind of program/command were you running that was reading (or writing) data FROM DISKS at 2.5GB/s? Are you sure this data was NOT cached? 2.5GB/s from cache seems plausible, from disks - almost too good to be true. But again, I've never seen Thumper "in real life" and you apparently did.
Anonymous on Jul 27 2006, 21:11

Several mkfiles in parallel. To be honest, it were 2.1 to 2.2 average.

Oh yeah, i have one at the office. Would be nice at home for video and mp3
Anonymous on Jul 27 2006, 21:38

Tyan's got a motherboard that would gets you close as far as I can tell.

ftp://ftp.tyan.com/manuals/m_s3892_110.pdf

Sorry that file is huge, but it has the architecture picture in it. If you squint hard at the grainy image it looks like you get 2 x16 PCIe channels and 2 PCI-X 133 channels.

Although that probably saturates the HyperTransport link to the processors. Of course, your 2.2GB/s sounds like HT's saturation point anyway.
Anonymous on Aug 8 2006, 08:54

Disregard my HT saturation comment on the X4500, obviously I didn't squint at the architecture pic you conveniently included.

Also, the fact that Sun has uplinked some of the PCI-X I/O to CPU1 is sweet. Unlike the Tyan that routes everything through CPU0.

Now if I can just convince the wife to tolerate the noise and keep driving the junker for another 4-6 years maybe I can get some of that classy Sun kit.

Anyway, that's the closest I've seen 'commodity' get, and obviously Sun is out ahead quite a ways.
Anonymous on Aug 8 2006, 09:04

SuperMicro has a very nice 4-way board with 2 PCI-X busses. It's using the same AMD8132 chipset as the Thumper. Also the same Intel NICs. This would make a nice 'lil' Thumper if you put it in a 24-bay chassis. It won't have some of the Thumper's capabilities, but at
Anonymous on Sep 1 2006, 14:39