EXPReS Update from Metsähovi

- 2008 Jan 29 in SURFnet, Utrecht

Ari Mujunen, Metsähovi Radio Observatory,
Helsinki University of Technology TKK

EXPReS is an Integrated Infrastructure Initiative (I3), funded under the European Commission's Sixth Framework Programme (FP6), contract number 026642 EXPReS.

Updates on WP1.1.1/2

- WP1.1 Data Acquisition...
  - .1 ...Architecture
  - .2 ...Prototype
- First: demo 0.5 Gbps over 1GE (using VSIBs; done)
- Then: demo n Gbps (4?) over 10GE (using iBOBs/iADCs and data acquisition PCs)

“Add-on” eVLBI

- Piggybacking VSIB+Tsunami onto regular Mark5A+FS experiments
  - Tested with regular IVS geo experiments (On, Mh, Bonn)
    - Mark5A takes disk backup, VSIB sends data from Mark5A outputs (via VSIC) in real-time /w Tsunami to correlator
  - Automated scripts to use regular FS schedule files
    - Demonstrates data acq control and FS compatibility
    - Used in all Mh IVS geo experiments recently (5)

./08jan_geo/recexpt_EURO91_Mh.sh
./07sep_geo/recexpt_EURO89_Mh.sh
./07jul_geo/recexpt_T2051_Mh.sh
./07may_geo/recexpt_EURO87_Mh.sh
./07may_geo/recexpt_T2050_Mh.sh

•Quick UT1 with eVLBI

- On, Mh, Tsukuba, and Kashima ”long-distance” transfers (256, 512 Mbps) for geodetic near-real-time UT1 observations
  - Tsunami to K5 software correlator, UT1 results out 30 minutes after transfer
  - Took part in 16..18-Jan-2008 JGN2 eVLBI demo (NICT/Ka)

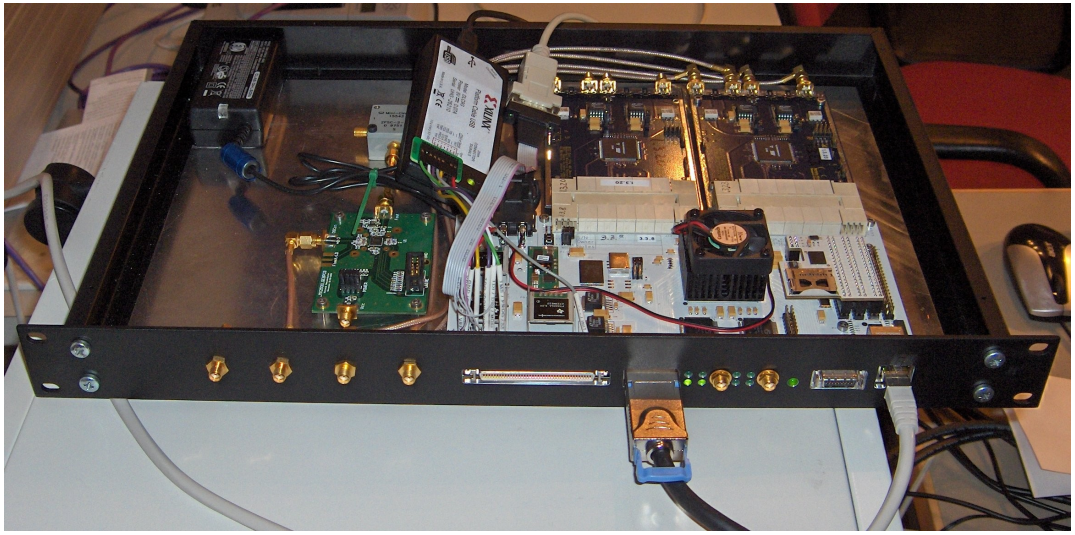
./08jan_geo_jpn/recexpt_u8016a_Mh.sh
./08jan_geo_jpn/recexpt_u8016b_Mh.sh
./08jan_geo_jpn/recexpt_u8016c_Mh.sh
./08jan_geo_jpn/recexpt_u8017a_Mh.sh
./08jan_geo_jpn/recexpt_u8017b_Mh.sh
./08jan_geo_jpn/recexpt_u8017c_Mh.sh
./07nov_geo_jpn/recexpt_u7326a_Mh.sh
./07nov_geo_jpn/recexpt_u7326b_Mh.sh
./07nov_geo_jpn/recexpt_u7326f_Mh.sh
./07oct_geo_jpn/recexpt_h07302_Mh.sh
./07oct_geo_jpn/recexpt_v07302_Mh.sh
./07oct_geo_jpn/recexpt_g07302_Mh.sh
./07oct_geo_jpn/recexpt_u7302_Mh.sh
./07sep_geo/recexpt_U07247_Mh.sh
./07sep_geo/recexpt_V07247_Mh.sh
./07sep_geo/recexpt_W07247_Mh.sh

•Software Items

- DiFX on PS3 (directly ported, mainly PowerPC)
  - on hold
- Minicorrelator on PS3 with SPU
  - http://cellspe-tasklib.cvs.sourceforge.net/
    - Look for ”minicorrelator”..
- fuseMk5A
  - Makes Mark5 8-pack recordings readable as regular Linux files
  - http://fusemk5a.sourceforge.net/
- tsunamifs
  - A ”user space” file system layer being developed on top of connections to Tsunami server(s)
  - on hold

iBOB/iADC Status

- At end of October 2007 Digicom finally delivered the boards
  - On received 1 iBOB and 1 iADC
  - Mh received 3 iBOBs and 3 iADCs
  - Jb already has 10 iBOBs but no iADCs
- Tested 2048.0/1024.0MHz iADC clock synthesizer boards
  - Report by Guifré Molera at:
    - http://www.metsahovi.fi/en/vlbi/ibob/Comparing-ADI-NAT.pdf
- Boxed one iBOB with iADCs, synth board, and pwr



Current 10GE Networking

- Extended our 10 Gbps Funet connection (Extreme Summit X450 switch) with a second 10GE module (10GBASE-SR)
- Purchased one HP 6400cl switch with
  - 6 10GBASE-CX4 integral copper ports
  - 2 10GBASE-SR additional fiber modules (X2, not XFP...)
- Purchased two Myri-10GE boards
  - One CX4 and one XFP with SR module
- Purchased a pair of Chelsio 10GE-CX4 boards (”eval kit offer”)
- Purchased just one Asus L1N64-WS to see if it is any good
- €: Extr 2k, HP 3k, SR 5k, boards 2.5k, Asus 1.5k /wo disks

10GE Networking, Boards

- Recommendations from Jb/RHJ that Myrinet and Chelsio PCIe boards perform well; backed up with test results
  - They were not too expensive, either...

Myri 10G-PCIE-8A-C+E

$695

Myri 10G-PCIE-8A-C+E
plus SR XFP

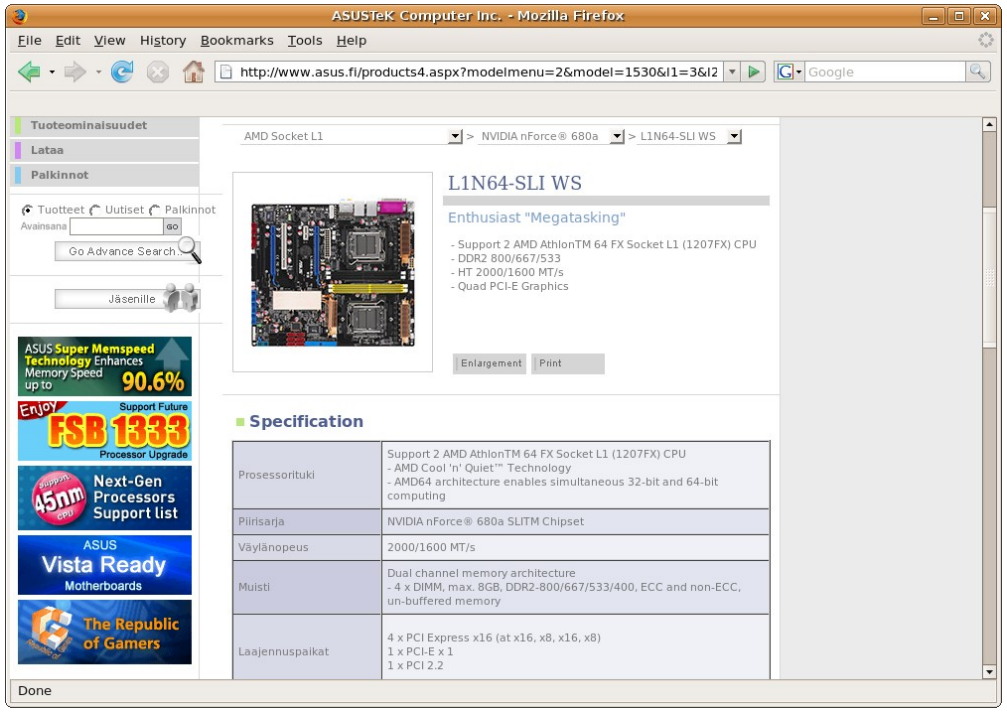
$795 + $500

Chelsio N310E-CX

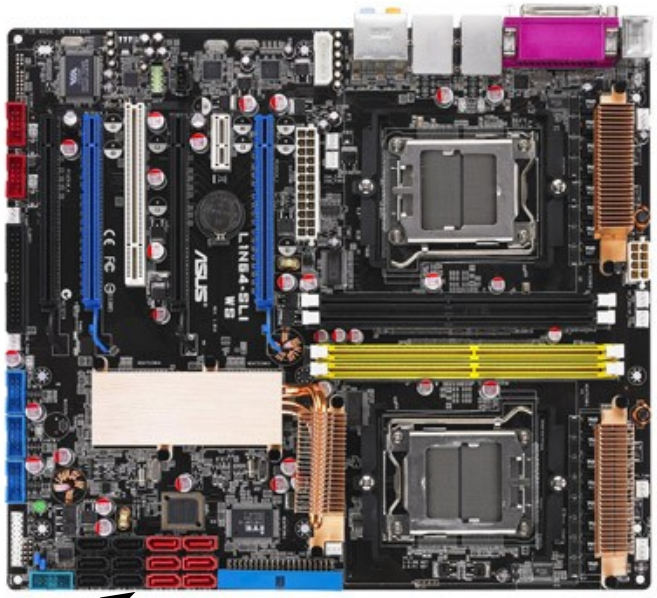
$695 (evkit $995 per pair)

PCIe x8 capable PCs?

- 10GE board is not of much use unless there are PCs which can pump data at >>4Gbps around
- Just starting to appear; first ones might not be too good in practice
- For instance Asus L1N64-WS:



L1N64-SLI WS



12 SATA ports!

© 2007 ASUSTeK Computer Inc. All rights reserved.

”abidal.kurp.hut.fi”

- Test notes at http://www.metsahovi.fi/en/vlbi/10gbps/10gbNetworkTests.html

”abidal” Performance

- Out-of-box 4470 or 9000 MTU
  - ~9 Gbps TCP, 8 Gbps UDP(!), 1.5Gbps Tsunami(?)
- Some TCP tweaking: TCP ~9.5Gbps
- Both Chelsio and Myrinet benefit a lot from switching off UDP data payload checksums (SO_NO_CHECK)
  - Checksummed iperf: 4.9 Gbps / 1 core (can run two)
  - No checksums: 9.9Gbps / 1 core
- Tsunami: a new v1.2 ”Petabit Tsunami” at
  - http://tsunami-udp.sf.net/
  - 7 Gbps memory-to-memory, 3.5 Gbps memory-to-raid-XFS
  - Original protocol parameters 32-bit, not enough for 10Gbps!
  - Interpacket timing had ~50usec resolution, now ~1usec
- Reasonable performance for an entry-level PCIe x8 PC
  - Net *10, disks *4, CPU *2, memory bandwidth *~1.5 when compared to the previous generation

Performance Lessons

- 10GE is really ***ten*** times faster than 1GE...
  - Where 3 old SATA disks could do 1Gbps...
  - ...you would need 30 old or faster new SATA disks
    - Faster SATA disks are coming, e.g. Samsung F1 series
  - Extra memory copies did not cause too much CPU consumption in 1GE...
  - ...whereas those hurt 10x times more with 10Gbps rate
    - Would benefit from zero-copy drivers and user sw
      - Net driver, disk drivers, filesystem drivers, userland..
  - It is easier to hit main memory bandwidth bottleneck
  - It is quite easy to consume the CPU power of e.g. four cores
    - Luckily more cores are coming...

Remoting Disk Accesses

- Juha Aatrokoski tested direct disk access network protocols in ”abidal”--”juliano”--PS3 environment with 1&10GE
  - AoE (ATA-over-Ethernet)
  - NBD (Network Block Device)
- As an alternative means to make data available to grid correlator nodes (ix86/Cell/...)
- Test report in http://www.metsahovi.fi/en/vlbi/vsib-tests/aoe-nbd/index.html
- Not spectacularly fast, except NBD gets data to PS3 ~0.5Gbps, faster than e.g. NFS

iBOB Development

- Xilinx ISE +EDK/XPS + Matlab Simulink + Xilinx SysGen pretty finicky to install and maintain
  - Old versions needed for CASPER MSSGE flow
  - With help from Jb, now finally being able to build MSSGE-based designs
- Linux expansion board for iBOB (memory + SD flash)
  - Yet to find out how useful this is...
- Next, trying to feed the data acquisition PC (with udp2raid software) with iBOB

So, what next?

- Get iBOB-based UDP transmitters (Jb+Mh; deploy at On, too) and receivers (data acq PC @Mh, iBOB @Jb) running
- Add real iADC data into UDP packets (Jb, Mh)
  - Take a look if there is any useful channelization firmware available that can be added
    - Mileura 1024-ch? Haystack DBE?
- Add more advanced protocols to iBOB (Jb)?
- See what we can and want to demo at 4(?)Gbps