

SciDAC Prototypes

Don Holmgren
Lattice QCD Project Review
Cambridge, MA
May 24, 2005

Outline

- Prototyping activities
- SciDAC clusters

Prototyping Activities

- Single node investigations:
 - Processors
 - Intel and AMD ia32, x86_64
 - Intel ia64
 - PPC970/G5
 - Motherboards/chipsets
 - PCI, PCI-X, PCI-E buses
 - RAMBUS, DDR, DDR2
- Small clusters
 - Networks
 - Gigabit ethernet (various vendors)
 - Myrinet
 - Infiniband
- Remote benchmarking

SciDAC Clusters

- Funding began in 2001
 - 3 years + 2 year extension
 - Effort (software development, machine design and deployment) + equipment
 - JLab and FNAL contributions
 - Facilities (space, power, cooling, mass storage)
 - Equipment (\$ roughly matching SciDAC FTE costs)
 - Coordination of prototyping efforts (FNAL/JLab)
 - Staggered acquisitions to sample market
 - JLab: single CPU's, DWF oriented, gigE meshes
 - FNAL: SMP, asqtad oriented, switched networks
 - Four years of operational experience
 - Manpower needs are well established

Prototype Clusters

Name (FNAL/JLab)	QCD80	2M	NQCD	W
Speed Processor (Processor Count)	700 MHz P-III (160)	2.0 GHz Xeon (128)	2.0 GHz Xeon (96)	2.4 GHz Xeon (256)
Memory Bus Speed	100 MHz	400 MHz	400 MHz	400 MHz
Single or Dual CPU	Dual	Single	Single	Dual
Interconnect Fabric	Myrinet	Myrinet	Myrinet	Myrinet
Performance	10 Gflop/s asqtad	100 Gflop/s DWF	40 Gflop/s asqtad	130 Gflop/s asqtad
Date in Production	3/2001	7/2002	7/2002	1/2003

Prototype Clusters

Name (FNAL/JLab)	3G	QCD	NQCD (reuse)	4G	PION
Speed Processor (Processor Count)	2.67 GHz Xeon (256)	2.8 GHz P4E (128)	2.0 GHz Xeon (64)	2.8 GHz Xeon (384)	3.2 GHz P4- 640 (260) → (520)
Memory Bus Speed	533 MHz	800 MHz	400 MHz	800 MHz	800 MHz
Single or Dual CPU	Single	Single	Dual	Single	Single
Interconnect Fabric	3D GigE Mesh	Myrinet	Infiniband	5D GigE Mesh	Infiniband
Performance	350 Gflop/s DWF	140 Gflop/s asqtad	30 Gflop/s asqtad	700 Gflop/s DWF	370 Gflop/s asqtad → 740
Date in Production	3/2004	6/2004	7/2004	1/2005	6/2005 (9/2005)

Lessons

- Operating clusters with minimal manpower
 - Network booting (2000)
 - PXE
 - OS installs
 - BIOS/firmware updates
 - Remote management
 - IPMI
 - Local + serial + IPMI-over-LAN
 - Out of band reset, power control
 - Health monitoring

Lessons

- Maximizing utilization
 - Nannys
 - Monitor and correct key processes and parameters:
 - PBS clients (“moms”)
 - Disk utilization
 - Network health (for example, Myrinet GM troubles)
- Importance of small prototyping
 - Delay of initial Xeon acquisitions because of poor PCI (Intel) and memory (ServerWorks) implementations
 - Faulty Intel PCI-E BIOS implementation
- Cost advantages of gigE meshes
 - Network costs cut in half compared with Myrinet
 - Loss of flexibility in job sizes

Future Prototyping

- Processors
 - Intel: multi-core, faster FSB
 - ia32/x86_64
 - ia64?
 - AMD
 - Athlon64/FX
 - Opteron – duals, quads, 8-way
 - Power
 - PPC970 successors
- Memory
 - 1066 FSB (and faster)
 - Fully buffered DIMMs

Future Prototyping

- Networks
 - Faster, wider Infiniband
 - New Myrinet
 - faster physical layer
 - MX instead of GM
 - PathScale Infinipath
 - Infiniband physical layer
 - Hypertransport connection on motherboard
 - Better latency
 - Proprietary software

Questions?

Backup (Details)

Production Prototypes

- QCD80 (FNAL)
 - Pre-SciDAC (supplemental DOE + base)
 - 80 dual 700 MHz Pentium III
 - Switched fast ethernet (Dec 2000)
 - Myrinet 2000 (Mar 2001)
 - 9 nodes also on SCI 2-D mesh
 - ~ 10 Gflop/s sustained (asqtad)
- Lessons
 - Unattended OS installs (PXE network boots)
 - Unattended BIOS and BMC firmware flashing
 - Myrinet software woes (NCSA VMI fixes)
 - IPMI software development
- Now used for automated perturbation theory
 - Switched gigE

Production Prototypes

- 2M (JLAB)
 - 128 single 2.0 Ghz Xeon, 400 Mhz FSB
 - Entered production July 2002
 - Myrinet 2000
 - O(100 Gflop/s) sustained on DWF
- Lessons
 - Big performance boost from quad-pumped bus
 - Long delay waiting for good PCI-X
 - Intel i850 had great memory BW (RAMBUS) but terrible PCI implementation
 - Serverworks-GC had great PCI but poor memory BW
 - Finally, Intel E7500 delivered very good PCI and acceptable memory BW

Production Prototypes

- **NQCD (FNAL)**
 - 48 dual 2.0 Ghz Xeon, 400 Mhz FSB
 - Entered production July 2002
 - Myrinet 2000 (filled last 48 ports of 128-port switch)
 - Also, 9 node 2-D gigE mesh
 - ~ 40 Gflop/s asqtad
- **Lessons**
 - Early gigE mesh testing
 - Explored GAMMA software
 - Reproduced Fodor's results
 - Learned how to operate a heterogeneous cluster under PBS/Maui
 - More IPMI
 - Supermicro BMC add-in cards
 - IPMI over LAN software

Production Prototypes

- **W** (FNAL)
 - 128 dual 2.4 Ghz Xeon, 400 Mhz FSB
 - Entered production January 2003
 - Myrinet 2000 (second 128-port switch)
 - ~ 130 Gflop/s asqtad
- Lessons
 - The need for high performance file I/O
 - How to install with minimal manpower
 - Establish strong relationships with white box vendors
 - The joy of optical links
 - Fiber is cheap and easy
 - Beware of laser reliability (and Myrinet software)

Production Prototypes

- **3G (JLAB)**
 - 256 single 2.67 Ghz Xeon, 533 Mhz FSB
 - Entered production March 2004
 - 3D gigE mesh, 4x8x8 or 4x4x16
 - 3 dual gigE cards + onboard gigE
 - QMP software implementation using VIA drivers
 - ~ 350 Gflop/s DWF
- **Lessons**
 - Large production gigE meshes are viable
 - Reduce network cost fraction to 25%

Production Prototypes

- QCD (FNAL)
 - 128 single 2.8 Ghz P4-E (“Prescott”), 800 Mhz FSB
 - Entered production June 2004
 - Reused Myrinet 2000 from QCD80/NQCD
 - ~ 140 Gflop/s asqtad
- Lessons
 - Single Pentium 4's are good and bad
 - Only 800 Mhz FSB at the time
 - Inexpensive systems (\$900 each)
 - First PCI-X for single processors (E7510)
 - Poor implementation
 - Still, most cost effective at the time
 - IPMI 1.5 software implementation

Production Prototypes

- **NQCD-2 (FNAL)**
 - Reuse 32 dual 2.0 Ghz Xeons, 400 Mhz FSB
 - Entered testing July 2004
 - Infiniband (PCI-X) – two 24-port switches
 - ~ 30 Gflop/s asqtad
- **Lessons**
 - First Infiniband experience
 - Fewer software problems than Myrinet
 - Some MPI grief (MVAPICH and MPICH-VMI)
 - LQCD code will tolerate significant oversubscription

Production Prototypes

- 4G (JLAB)
 - 384 single 2.8 Ghz Xeon, 800 Mhz FSB
 - Entered production January 2006
 - 3 dual gigE cards + dual on-board
 - 5D mesh
 - $6 \times 8 \times 2^3$
 - 700 Gflop/s DWF
- Lessons
 - Duals looked promising, but singles still more cost effective
 - Careful testing of dual Xeon and dual Opteron

Production Prototypes

- **Pion (FNAL)**
 - 260 single Pentium 640, 800 Mhz FSB (\$1000 each)
 - Will enter production June 2005
 - E7221 chipset (PCI-E)
 - Infiniband (\$870/node)
 - ~ 375 Gflop/s asqtad
- **Lessons to learn**
 - Large scale Infiniband use
 - IPoIB and SDP robustness for file I/O
 - Stability and scaling
 - Establish oversubscription tolerance
 - Expansion in late FY05
 - 260 additional processors
 - Pentium 640 or dual Opteron

Future Prototyping

- Speculative work
 - GPU's
 - Nvidia, ATI, streams programming
 - Also, Cell processor
- Software
 - QMP over VAPI
 - OpenIB Gen2 stack
 - x86_64 optimizations
 - Rewrite SSE QLA to exploit more registers
 - Multithreading on SMP (multicore)