Washington University High Performance Cluster

IBM High Performance Computing
February 2010

Y. Joanna Wong, Ph.D.
yjw@us.ibm.com
HPC performance scaling

Scale Out
- distributed parallel
- node -> cluster (MPI)

Scale Up
- shared memory parallel
- smp -> ccNUMA (OpenMP)

Hybrid parallel

Enterprise eX4

iDataPlex

Consolidation, Virtualization

Scale, Power, Density, Optimization

Scale Out

Infrastructure Simplification, Application Serving

Scale Up

Enterprise Racks & Towers

BladeCenter
IBM System Cluster 1350

- Comprehensive HPC offering from IBM for Linux including
  - Servers: IBM System x (x3650 M2, x3950 M2), iDataPlex systems, BladeCenter
  - Storage:
    - Industry-leading OEM Interconnect
      - InfiniBand switches – Qlogic, Voltaire
      - Ethernet – Voltaire 10Gb Ethernet switch
      - Adapters – Mellanox HCA, QLogic InfiniBand QDR HCA, Chelsio 10Gb eEhernet adapters, Mellanox 10Gb Ethernet adapter
  - Software:
    - IBM General Parallel File System (GPFS)
    - Adaptive Computing MOAB Cluster Management Software
    - xCAT management software

- System is factory-integrated, thoroughly tested, assembled, and cabled for rapid deployment

- IBM is single point of contact for hardware support
IBM System Cluster 1350

Rapid Deployment, Integrated & Single Point of Contact

Transform “Roll Your Own” ..

Pre-integrated solution
• Servers, switches, PDU’s, cabling
• Ready to roll into place, apply power and networking – and get to work!
HPC Systems at Washington University

2 iDataplex Systems with 168 dx360 M2
(each with 2 QC 2.66GHz Intel Xeon X5550, 24GB DDR3-1333 mem, 250GB SATA HDD, Mellanox ConnectX dual-port 4X DDR HCA)

7 x3950 M2 (each with 16 QC 2.4GHz Intel Xeon E7440, 256GB DDR2 mem, 4 73GB SAS HDD, Mellanox ConnectX dual-port 4X DDR HCA)

4 BNT 8000F GigE leaf switches

2 BNT 8000R GigE aggregation switches

x3650 M2 management node
(2 QC 2.26GHz Intel Xeon L5520, 12GB DDR3-1333 mem, 2 300GB SAS HDD, Mellanox ConnectX dual-port 4X DDR HCA)

2 x3650 M2 login nodes
(2 QC 2.26GHz Intel Xeon L5520, 48GB DDR3-1333 mem, 2 300GB SAS HDD, Mellanox ConnectX dual-port 4X DDR HCA)

4 x3650 M2 gateway nodes
(2 QC 2.26GHz Intel Xeon L5520, 24GB DDR3-1333 mem, 2 73B SAS HDD, Mellanox ConnectX dual-port 4X DDR HCA)

DS4700 storage controller with 3 DS4000 EXP810 (48 146GB FC drives)

2 x3650 M2 GPFS NSD
(2 QC 2.26 GHz Intel Xeon L5520, 24GB DDR3-1333 mem, 2 146GB SAS HDD, Mellanox ConnectX dual-port 4X DDR HCA, Qlogic 4Gb FC dual-port HBA)

Qlogic 9420 DDR 288 port Infiniband Switch

External network

DS4700 storage controller with 3 DS4000 EXP810 (48 146GB FC drives)
iDataPlex – Designed for Data Center Flexibility

Broad portfolio of customizable components that adjust to your computing needs
iDataPlex dx360 M2 – High performance dual-socket

The dual-socket Nehalem-based server with leadership compute capacity in the iDataPlex data center solution

Highlights

Processor: Quad Core Intel Xeon 5500 series
Quick Path Interconnect up to 6.4 GT/s
Memory: 16 DIMM DDR3 - 128 GB max
Memory Speed: up to 1333 MHz
Chipset: Tylersburg-36D
PCle: 1 slot x16 electrical/ x16 mechanical (Gen2), 2 x8
Interconnect Fabric Adapters: Ethernet (1Gb, 10GbE), Fibre Channel HBA (4GB, 8 GB, dual-port), Infiniband (DDR, QDR, dual-port), Solid State Devices (SSD) PCIe
Internal storage: up to 5TB (2U Flex chassis) or 12TB (3U Storage chassis)
Power supply: 900W high efficiency (per two servers)
dx360 M2

Onboard Ethernet

16 max. (8x per) DDR3 DIMMs (1GB, 2GB, 4GB & 8GB)

CPU 1 - Supports DIMM Bank 0 (Intel Nehalem EP)

DIMM Bank 2

CPU 0 - Supports DIMM Bank 1 (Intel Nehalem EP)

DIMM Bank 1

Hard drive carrier

Onboard Ethernet

Ethernet Ports & Management

Battery

16x PCIe Riser card slot

Virtual Media Key

Mini PCIe SAS Card Slot

SATA ports 6x
Nehalem-EP CPU Summary

**Performance/Features:**
- 4 cores
- 8M on-chip Shared Cache
- Simultaneous Multi-Threading capability (SMT)
- Intel® QuickPath Interconnect up to 6.4 GT/s, each direct. per link
- Integrated Memory Controller (DDR3)
- New instructions

**Power:**
- 95W, 80W, 60W

**Socket:**
- New LGA 1366 Socket

**Process Technology:**
- 45nm CPU

**Platform Compatibility**
- Tylersburg (TBG)
- ICH9/10

*Driving performance through Multi-Core Technology and platform enhancements*

Source: Intel Corporation
**Intel® Xeon™ 5500 Series (Nehalem-EP) Overview**

**IT Benefits**
- More application performance
- Improved energy efficiency
- End to end HW assist (virtualization technology improvements)
- Stable IT image
  - Software compatible
  - Live migration compatible with today’s dual and quad-core Intel® Core™ microarchitecture products using enabled virtualization software

**Key Technologies**
- New 45nm Intel® Microarchitecture
- New Intel® QuickPath Interconnect
- Integrated Memory Controller
- Next Generation Memory (DDR3)
- PCI Express Gen 2

Source: Intel Corporation
Energy Efficiency Enhancements
Intel® Intelligent Power Technologies

Integrated Power Gates\(^1\)

- Enables idle cores to go to near zero power independently

Automated Low Power States

- More & Lower CPU Power States
- Reduced latency during transitions
- Power management now on memory, I/O

Automatic operation or manual core disable\(^2\)

- Adjusts system power consumption based on real-time load

---

\(^1\) Integrated power gates (C6) requires OS support

\(^2\) Requires BIOS setting change and system reboot

Source: Intel Corporation
More Efficient Chipset and Memory

Memory Power Management

- DIMMs are automatically placed into a lower power state when not utilized\(^1\)
- DIMMs are automatically idled when all CPU cores in the system are idle\(^2\)

Chipset Power Management

- QPI links and PCIe lanes placed in power reduction states when not active\(^3\)
- Capable of placing PCIe* cards in the lowest power state possible\(^4\)

End-to-end platform power management

Source:
\(1\) Using DIMM CKE (Clock Enable)
\(2\) Using DIMM self refresh
\(3\) Using L0s and L1 states
\(4\) Using cards enabled with ASPM (Active State Power Management)
SciNet General Purpose Cluster Content

- **9 Scalable Units (5 racks or 420 nodes)**
  - 3780 total planars
  - 30,000 cores
  - 45 total racks
  - 16 GB DDR3-1333/node
  - 2-Intel Nahalem-EP Quad-Core @ 2.53GHz (80W)/node
  - 864 planars with Mellanox ConnectX Dual-Port 4X DDR IB Adapter

- **Infrastructure & Software**
  - Qlogic 4x DDR Infiniband
  - Myrinet 10 G & Force10 e300
  - BNT 1G switch
  - xCat Management nodes
  - MOAB
  - GPFS
  - Redhat Enterprise Linux 5.3
Compute systems

- IBM iDataPlex systems
  - 168 IBM dx360 M2 servers
    - node001 – node168
  - QLogic 9240 DDR InfiniBand Switch
    - 288 ports
  - Blade Network Technologies Ethernet Switch

- IBM dx360 M2
  - 2 quad-core Intel Xeon X5550 processors (8 x 2.66GHz Nehalem cores)
  - 24GB memory (6 x 4GB DDR3-1333 RDIMM)
  - Mellanox ConnectX dual-port 4X DDR HCA
  - 250GB 7200 rpm SATA HDD
IBM x3950 M2 enterprise server

Outstanding 4-socket Reliability
Investment projection with ability to upgrade to 2 node solution with ScaleXpander Option kit

ScaleXpander Option Kit
Includes ScaleXpander chip, scalability cable, cable management arm, and x3950 M2 Bezel

Superior Scalability
Superior 8+ socket performance
- Up to 4 node for a total of 96 cores and 1TB of memory
Delivering SMP capability at a fraction of the price.

Feature List
4U Rack-optimized server six-core Intel Xeon 7400 Series Processors
eX4 chip set with FSB 1066 MHz
32 DIMM Slots, DDR2/533MHz
1/2/4 & 8GB DIMM support
7 PCI-e 8x half-length slots

2 hot-swap
SAS Storage Controller (RAID 0/1 STANDARD; RAID 5 optional)
4 2/5” hot-swap HDDs
Dual port Gb Ethernet w/TOE, iSCSI & RDMA
On-board management controller delivers full remote presence capability
X-Architecture features

- **Fourth Generation of Snoop Filter**
  - Highly efficient, directory based cache coherency

- **Scalability Ports**
  - Modular
  - NUMA Architecture (4th Generation)
  - Up to four servers
    - 16 CPUs (Dual or Quad Core)
    - 1TB PC2-5300 DDR II Memory
    - 28 expansion slots

- **Active Memory**
  - Chipkill & ECC
  - Memory ProteXion – Redundant Bit Steering
  - Mirroring - Hot-swap / hot-add

- **Expansion Slots**
  - PCI-E – 7 slots per chassis

- **Fourth Generation of EXA**

  - Supporting the latest generation of Intel six-, quad-, and dual-core processors!
Shared memory compute systems

- **IBM x3950 M2**
  - IBM’s uniquely scalable System x architecture based on eX4 technology
    - 16 quad-core Intel Xeon E7440 processors (4 2.40 GHz Dunnington cores, 16MB L3 cache, 90W)
    - 256GB memory (64 x 4GB PC2-5300 DDR2 RDIMM)
    - Mellanox ConnectX dual-port 4X DDR HCA
    - 4 x 73GB 10k rpm SAS HDD
- 7 systems: smp001 – smp007
Storage solution

- General Parallel File System (GPFS) / Clustered NFS (CNFS) servers
  - 2 x3650 M2 GPFS NSD servers / Clustered NFS servers
    - 2 quad-core Intel Xeon L5520 processors (8 x 2.26GHz Nehalem cores)
    - 24GB memory (6 x 4GB DDR3-1333 RDIMM)
    - QLogic 4GB FC dual-port PEIe HBA
    - Mellanox ConnectX dual-port 4X DDR HCA
    - 2 x 146GB 10k rpm SAS HDD
  - DS4700 storage solution
    - DS4700 storage controller with 16 146 GB drivers
    - 3 DS4000 EXP810 storage expansions, with 48 146GB FC drives
- GPFS version 3.2.1
- Filesystems: /home, /scratch, /export
HPC cluster system environment

**Network setup**
- 4 private subnets within the cluster and one external subnet
- 172.20.0.0/16 on eth0 of most nodes
  - hostname of compute clusters (node001, …, smp001, …) on this subnet
- 192.168.0.0/16 on ib0 on all nodes
  - IPoIB subnet of compute clusters (node001-ib0, …, smp001-ib0, …)
- 2 management subnets for system administration of cluster
- 10.28.56.0/24 – the external subnet

**Login nodes**
- login1.wustl.edu
  - eth0: 172.20.4.1, login001 (hostname)
  - ib0: 192.168.4.1, login001-ib0
- login2.wustl.edu
  - eth0: 172.20.4.2, login002 (hostname)
  - ib0: 192.168.4.2, login002-ib0
HPC cluster system environment

- Operating Systems: RHEL 5.4
  - `cat /etc/redhat-release`
  - kernel: Linux version 2.6.18-164.el5
- OFED 1.4.0
- MOAB 5.3.4
- Torque 2.4.2

- Compilers
  - GNU compiler 4.1.2 (gcc, g++, gfortran)
    - `/usr/bin/gcc --version`
      gcc version 4.1.2 20080704 (Red Hat 4.1.2-46)
  - Intel Compiler Suite (C/C++, Fortran) for Linux, v 11.1.064
    - `/export/intel/Compiler`
- **MPI libraries**
  - bundled with OFED 1.4.0
    - Compiled with GNU compilers
    - Open MPI 1.2.8
    - mvapich 1.1.0
    - mvapich2 1.2p1
    - MPI tests 3.1
  - Open MPI 1.4.1, built with Intel compiler
    - `/export/openmpi-1.4.1`