High Performance Computing(HPC) & Software Stack

January 30-31, 2012

Pidad D'Souza (pidsouza@in.ibm.com)
IBM, System & Technology Group

Agenda

- Parallel Computing
- Parallel Computer Architecture
- IBM Parallel Environment(PE)
- IBM PE Stack

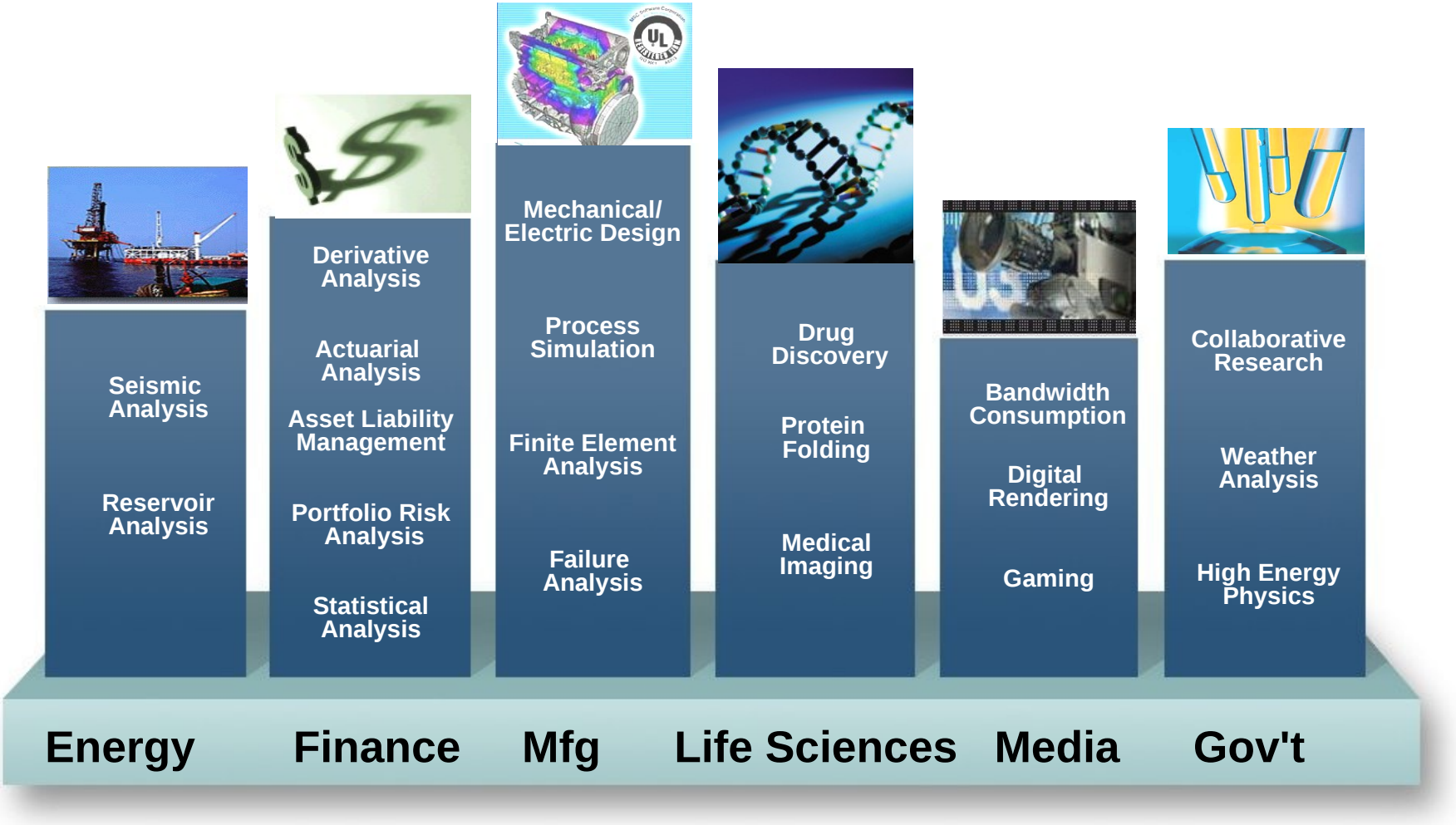
What is Parallel Computing?

- Simultaneous use of multiple compute resources to solve a computational problem
  - The compute resource could be
    - a single computer with multiple processors
    - An arbitrary number of computers (nodes) connected by a network
    - A combinational of both

Why use Parallel Computing?

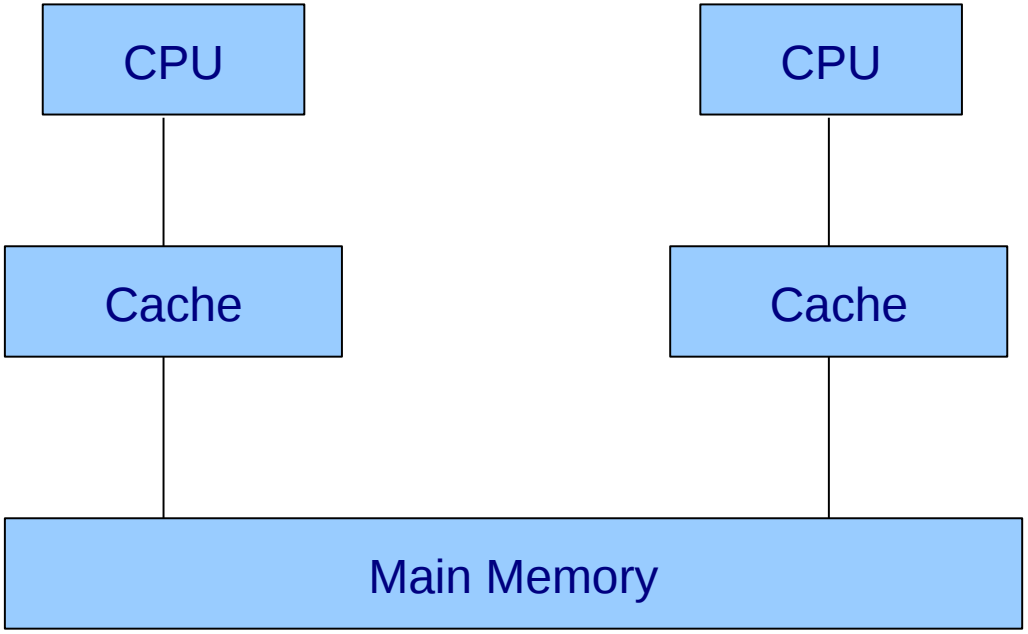
- Save time
- Solve larger problems
- Provide concurrency

Who's doing Parallel Computing?



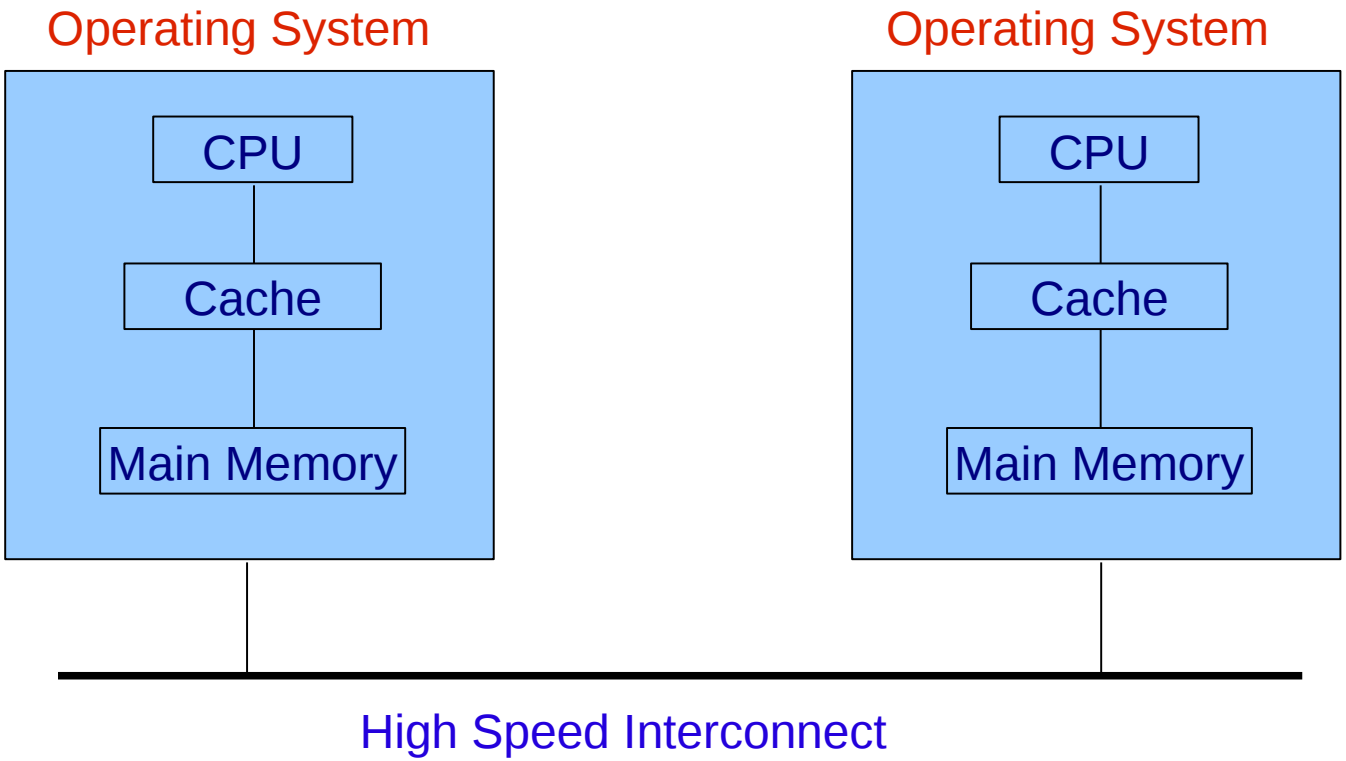
Parallel Computer Architectures

Shared Memory System



- Symmetric Multiprocessors (SMP)
  - ✓ SMP is a type of HPC architecture
  - ✓ multiple processors share the same memory
  - ✓ more expensive and less scalable than MPPs (massively parallel processors)

Distributed Memory System



- Clusters
  - ✓ Predominant type of HPC hardware
  - ✓ Processor in a cluster is referred as a node
  - ✓ Has its own CPU, memory, operating system and I/O subsystem
  - ✓ Capable of communicating with other nodes

Approaches to parallel programming

- **Distributed memory approach**
  - ✓The master node divides the work between several slave nodes.
  - ✓Slave nodes work on their respective tasks.
  - ✓Slave nodes intercommunicate among themselves if they have to.
  - ✓Slave nodes return back to the master.
  - ✓The master node assembles the results, further distributes work, and so on.

    - **Practical problems**
      - –Each node has access to only its own memory
      - –Data structures must be duplicated and sent over network if other nodes want to access them, leading to network problem

Approaches to parallel programming

- **Shared memory approach**
  - Memory is common to all processors
  - Programming is easier since all data is available to all processors

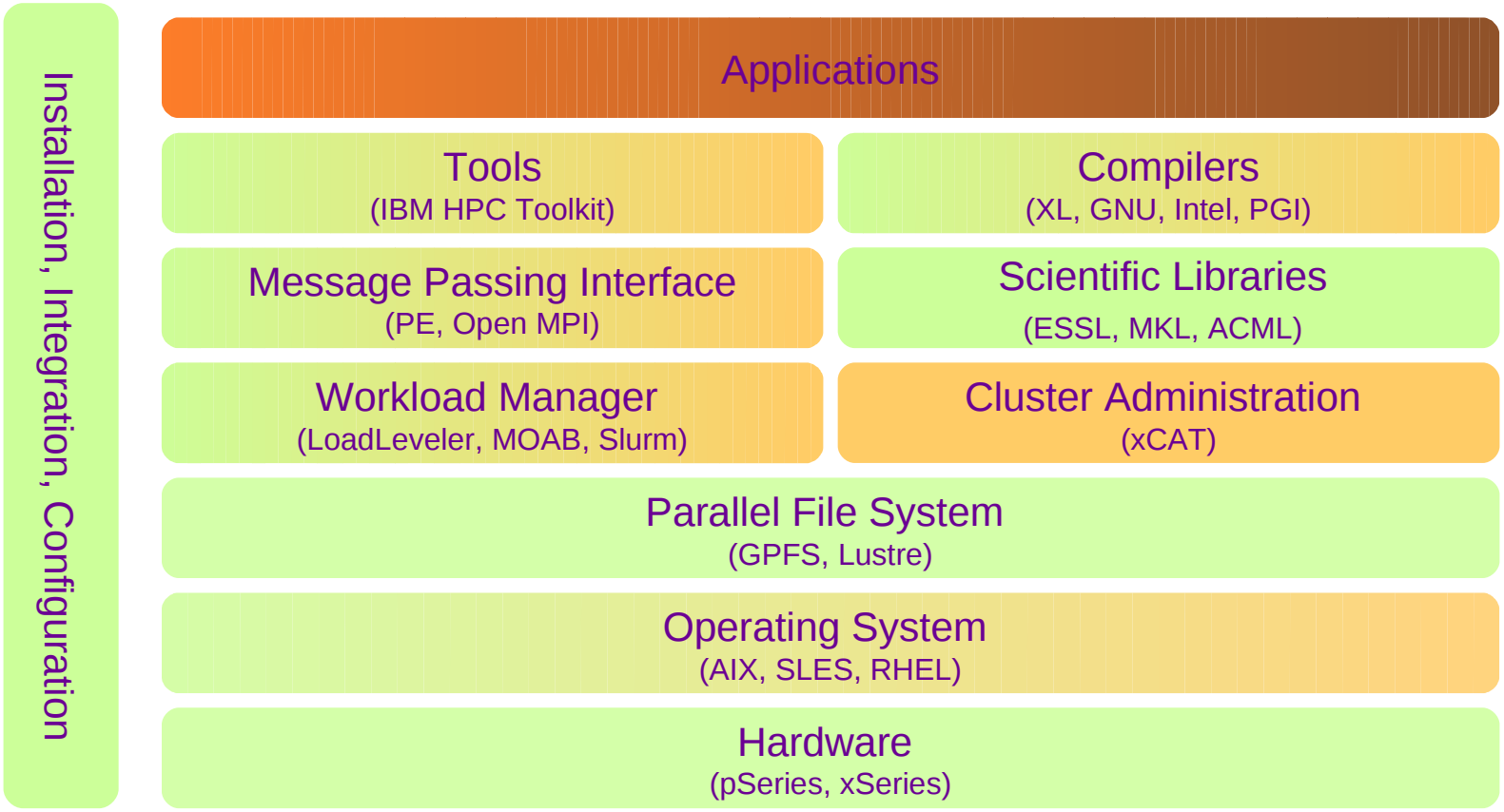
    - **Practical problems**
      - scalability

IBM Parallel Environment (PE)

January 30-31, 2012

Pidad D'Souza (pidsouza@in.ibm.com)
IBM, System & Technology Group

HPC Stack

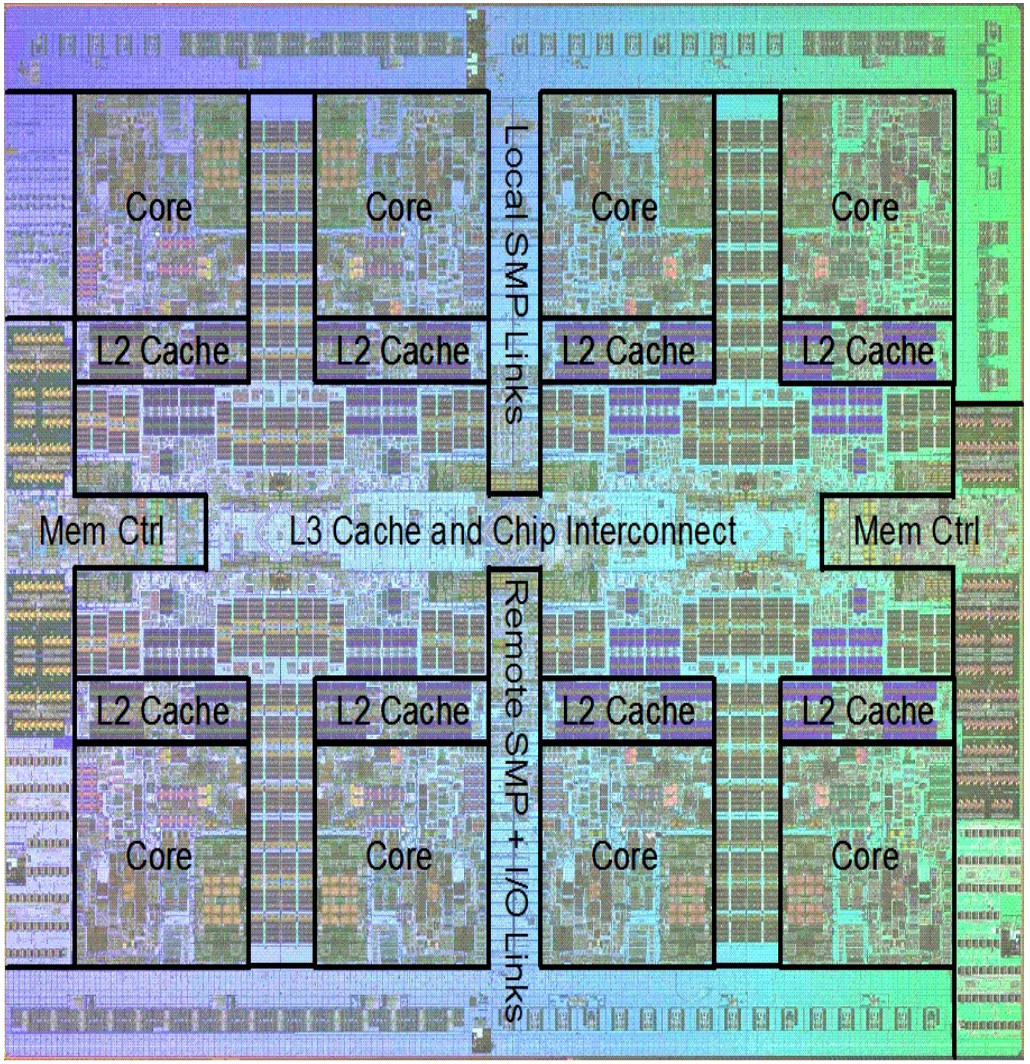

Installation, Integration, Configuration
Applications
Tools
(IBM HPC Toolkit)
Compilers
(XL, GNU, Intel, PGI)
Message Passing Interface
(PE, Open MPI)
Scientific Libraries
(ESSL, MKL, ACML)
Workload Manager
(LoadLeveler, MOAB, Slurm)
Cluster Administration
(xCAT)
Parallel File System
(GPFS, Lustre)
Operating System
(AIX, SLES, RHEL)
Hardware
(pSeries, xSeries)

HPC Stack

- Hardware
- Operating System
- Parallel File System
  - Provides concurrent high-speed file access to applications executing on multiple nodes of clusters
- Cluster Administration
  - Install/manage OS, setup HPC stack, create, manage clusters
- Scientific Libraries
  - Enginering and scientific subroutine libraries
- Workload manager – Job management system
  - Allows users to run more jobs in less time by matching the jobs' processing needs with the available resources
  - Schedules jobs, and provides functions for building, submitting, and processing jobs quickly and efficiently in a dynamic environment
- Message Passing Interface(MPI)
- Compiler
- Tools
  - Profiling tools

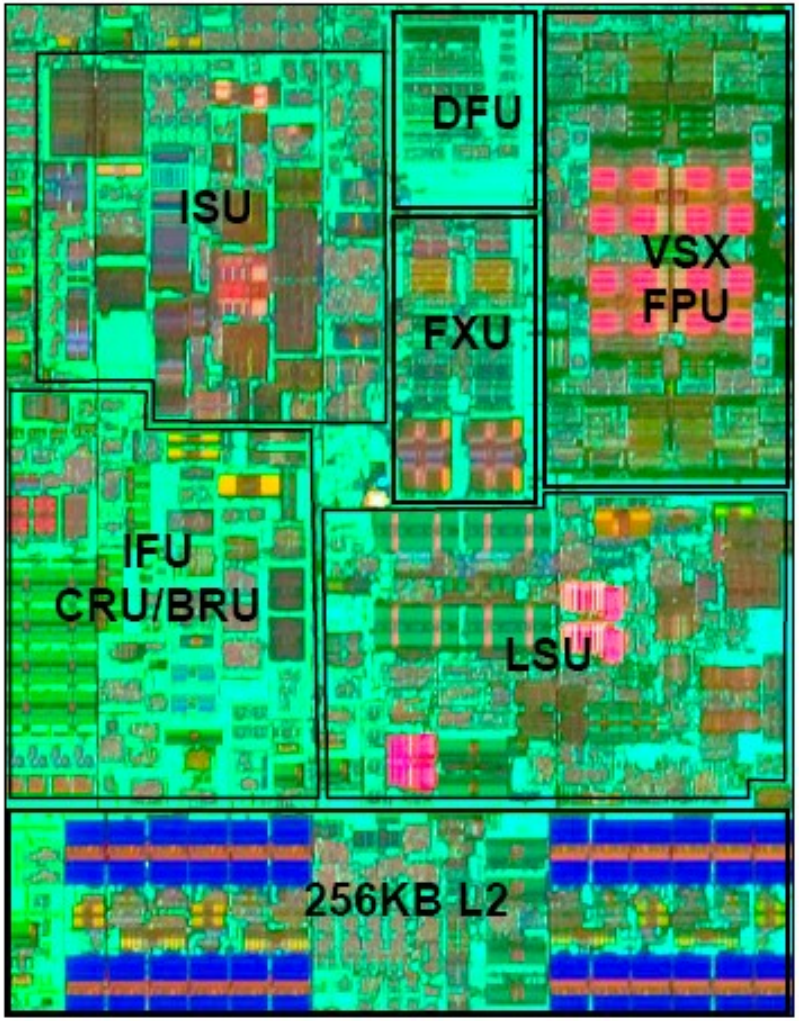
POWER 7 Die Overview

- 64-bit 8 Core/chip
- Die size – 567 mm2
- Fabrication -
  - 45Nm
  - Copper interconnect
  - Silicon-on-insulator
  - eDRAM
- Max execution threads core/chip –
  - 4/32
- On chip L3 Cache core/chip –
  - 4MB /32MB
- L2 Cache core/chip – 256 KB/2MB
- DDR3 Memory controller – 2
- Compatibility –
  - With prior generation of POWER



One POWER7 Core
Overview

- Execution Units
  - 2 Fixed point units
  - 2 Load Store units
  - **4 Double Precision floating point**
  - 1 Branch
  - 1 Condition register
  - **1 Vector unit**
  - 1 Decimal floating point unit
  - 6 Wide dispatch
- Recovery Function Distributed
- **1,2,4 Way SMT** Support
- Out of Order Execution
- 32 KB I-Cache
- 32KB D-Cache
- 256KB L2 cache
  - Tightly coupled to core



Operating System Support

- AIX
  5.3, 6.1, 7.1
- IBM i
  6.1, 7.1
- Linux
  SLES 11, RHEL 6.1

Parallel File System - GPFS

- Scalable high-performance parallel file system for AIX 5LTM and Linux® clusters
- New information lifecycle management features simplify data management and enhance administrative control
- Capable of supporting multi-terabytes of storage and over 1000 disks within a single file system
- Shared-disk file system can provide every cluster node with concurrent read/write access to a single file
- High reliability/availability through redundant pathing and automatic recovery from node and disk failures
- Powers many of the world's largest supercomputers distributed memory message passing system

GPFS Featuers

- Scalable
- Parallel access from multiple nodes
- Distributed locking – allows for parallelism and Consistency
- Striping implementation within file system
- Portable – Provides POSIX interface to file system
- High availability & fault tolerance
- Provides deep prefetching of data
- Simplified storage management

xCAT

- **Extreme Cluster (Cloud) Administration Toolkit**
  - Open source (Eclipse Public License) cluster management solution
  - Configuration database – a relational DB with a simple shell
  - Distributed network services management and shell commands
  - Framework for alerts and alert management
  - Hardware management – control, monitoring, etc.
  - Software provisioning and maintenance
- **Design Goals**
  - Build on the work of others – encourage community participation
  - Use Best Practices – borrow concepts not code
  - Scripts only (no compiled code) portability – key to customization!
  - Customer requirement driven
  - Provide a flexible, extensible framework
  - Ability to scale “beyond your budget”

Loadleveler

- Job management system
- Match job's processing needs with available resources
- Schedule, build, submit and proces the jobs

IBM Parallel Program(PE)

- Environment designed for Developing and executing parallel Fortran, C, or C++ programs
- A distributed memory message passing system (LAPI/MAPI)
- Consists of Components and tools for developing, executing, debugging, profiling and tuning parallel programs
- Parallel programs are run as a number of individual, but related, parallel tasks on a number of your system's processor nodes
- Processor nodes are connected on same network

Scientific Libraries

- MASS
  - Mathematical Acceleration sub-system
  - High performance mathematical functions (accuracy and exception handling not necessarily the same as standard math library)
  - Scalar(libmass.a) and vector(libmassv.a) versions available
  - Common architecture and machine-specific vector library provided
  - Thread Safe
- ESSL
  - Engineering and Scientific Subroutine Library
  - Matrix computations (linear algebra, linear equations, eigensystems) for dense and sparse matrices (BLAS, LAPACK)
  - Signal-processing computations (Fourier transforms, convolutions and correlations)
  - Sorting and searching
  - Interpolation (polynomial, cubic spline)
  - Numerical quadrature
  - Random-number generation
- PESSL (Parallel ESSL)
  - Matrix computations (linear algebra, linear equations, eigensystems) for dense and sparse matrices (BLACS, PBLAS, ScaLAPACK)
  - Fourier transforms
  - Random-number generation

Compilers

- XL C & C++ Compiler
  - High performance optimizing compiler
  - Designed to exploit power processor
  - Enables development of parallel applications
  - Leverage multi-core and vector features
- XL Fortran Compiler
  - Supports extensive numerical, scientific and high-performance computing
  - Leverage Power systems hardware advancements
  - Supports Fortran 95, Fortran 90, Fortran 77
- Others
  - GNU, Intel, PGI

Performance Tools

- HPC Toolkit
  - Collection of tools analyze performance of parallel and serial applications
  - Supports C, C++ and Fortran
  - OS: AIX, Linux (pSeries, xSeries)
  - MPI performance and communication patterns
  - Hardware performance counters
  - OpenMP performance
  - Analyzing I/O patterns
  - Helps identify hotspots and locating relationship between functions

References

- IBM Cluster Product Information
  http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp

- xCAT
http://sourceforge.net/apps/mediawiki/xcat/index.php?title=Main_Page

- XL Compilers
  http://publib.boulder.ibm.com/infocenter/lnxpcomp/v111v131/index.jsp

Thank You