Fusion Briefing

Ray Bair
Laboratory Computing Resource Center

October 29, 2009

Building Emergency Evacuation

- **There will be an emergency drill during this briefing**
  - We do not have to participate. ES&H permission granted.
  - Here is your training

- **In a real emergency requiring evacuation…**
  - Go up the stairs and out the main entry on the 2nd floor
  - Gather in 221/203 parking lot
  - Follow ES&H and emergency responder's instructions
  - Return to the building through the 2nd floor entrance

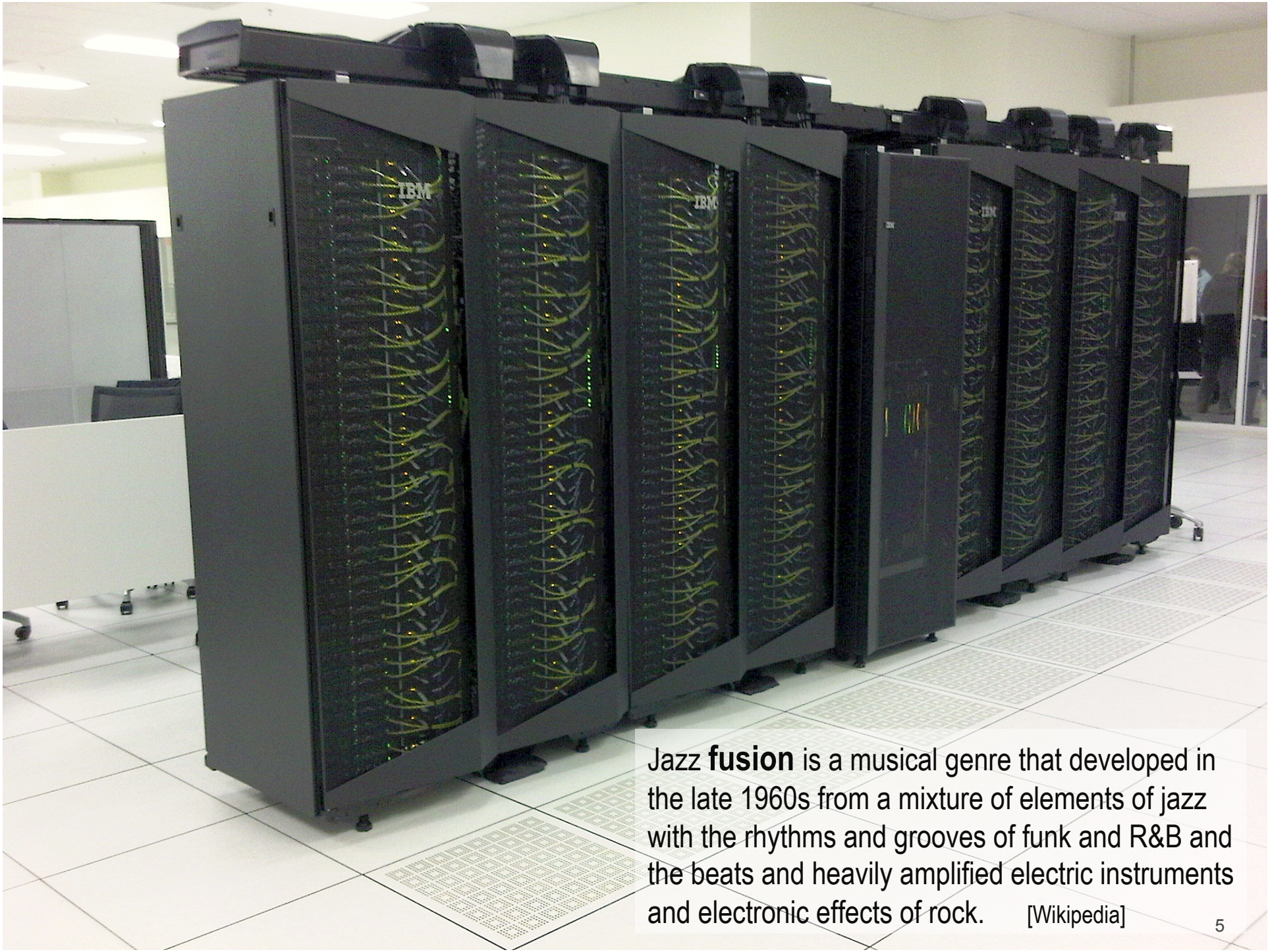
  - [Interim tornado shelters are in the stairwells.]

Today's Topics

- **Details of the new LCRC cluster - Fusion**
  - Hardware configuration
  - Software available and upgrades
  - Benchmark results
- **Details pertaining to usage of Fusion**
  - What is different about Fusion
  - Project allocations
  - Connecting to Fusion
  - Submitting jobs and getting help
- **Upcoming events**

Laboratory Computing Resource Center Staff

- Ray Bair – Chief Computational Scientist for CELS
- Shashi Aithal – Computational Scientist
- John Valdes – Sr. HPC Systems Administrator
- Jason Hedden – HPC Systems Administration Specialist



Jazz **fusion** is a musical genre that developed in the late 1960s from a mixture of elements of jazz with the rhythms and grooves of funk and R&B and the beats and heavily amplified electric instruments and electronic effects of rock. [Wikipedia]

Many benefits to Argonne research

- More compute power (~16 times as much)
- More memory (~25 times more RAM)
- Faster processors and interconnect
- Newer software versions
- Minimal change in the way users use LCRC

Compute Hardware Overview

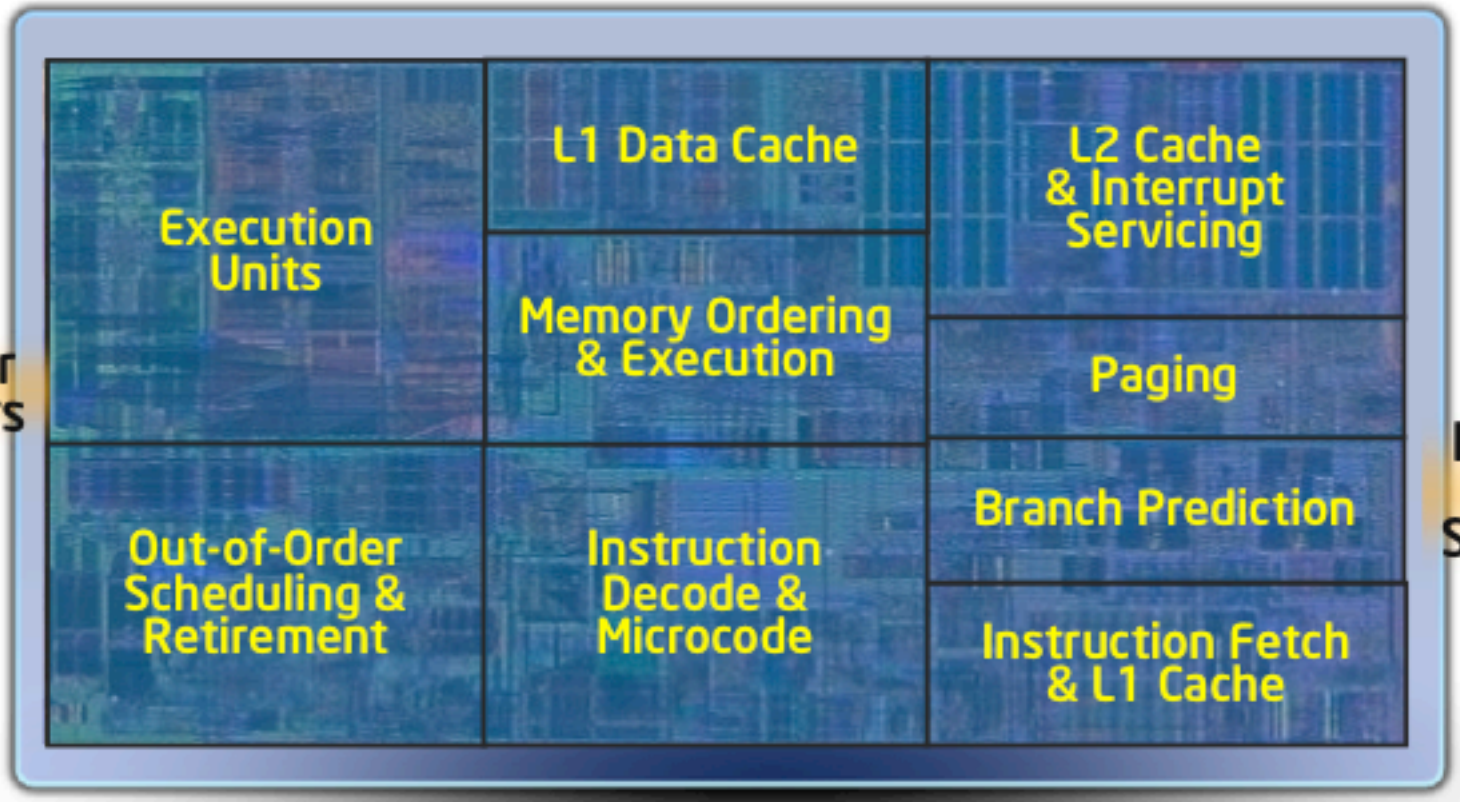
- **320 compute nodes**
  - Intel Nehalem series
  - dual-socket, quad-core 2.53 GHz
- **2560 compute processors (320 x 8) total**
- **12.5 TB memory**
  - 304 Regular Nodes: 36 GB each
  - 16 Fat Nodes: 96 GB each (in 2010; currently 64 GB)
- **Peak performance = 25.9 Tflops (25.9 * 1012 floating point operations per second)**
  - Linpack Benchmark = 21.5 Tflops
- **High Speed Interconnect**
  - InfiniBand QDR 4GB/s per link, per direction
- **IBM iDataPlex – 2 Nodes in 2U**
  - 250 GB disk
  - Integrated Gigabit Ethernet

Designed for Performance

New SSE4.2 Instructions

Improved Lock Support

Additional Caching Hierarchy



Deeper Buffers

Improved Loop Streaming

Simultaneous Multi-Threading

Faster Virtualization

Better Branch Prediction

Compute Hardware – Comparison with Jazz

	Fusion	Jazz
Number of nodes	320	350
Login Nodes	4	4
Sys Mgt Nodes	4	8
Cores/node (processors)	8 (dual quad-core)	1
Chip	Intel® Xeon quad core Nehalem @ 2.53 GHz	Intel Pentium IV Xeon @ 2.4GHz
Memory	12.5 TB 304 Regular Node: 36 GB 16 Fat Node: 96 GB/node	0.5 TB 175 Nodes 1GB 175 Nodes 2GB
Local scratch disk/node	250 GB	80GB

Network – Comparison with Jazz

	Fusion	Jazz
High Performance Interconnect	• InfiniBand QDR @ 4 GB/s per link, 2 µsec latency	• Myrinet 2000 to all systems @ 0.25 GB/s, 6-8 µsec latency
Ethernet Infrastructure	• Gigabit Ethernet to nodes (1000 gbps) • 10 GigE infrastructure to Lab	• Fast Ethernet to nodes (100 gbps) • 1 GigE infrastructure to Lab

Parallel Home File System

- **General Parallel File System (GPFS)**
  - GPFS provides concurrent high-speed file access to applications executing on multiple nodes of clusters
  - Your home directory is /home/<username>
  - Available on all compute nodes
  - 90 TB space
  - Backed up nightly to CELS new tape library
  - Higher performance and more robust shared file access than NFS.
  - Other features such as high availability, disaster recovery etc.

Large Parallel Scratch File System

- **Parallel Virtual File System (PVFS2)**
  - Developed in MCS
  - 320 TB of space
  - High performance parallel access to large scratch storage
  - MPI-IO (ROMIO) support over InfiniBand
  - Mounted at /pvfs/scratch/<username>
  - NOT backed up, data can be lost
  - Cannot run executables from this space
  - No symbolic link support

Software – Compiler versions

Package	Fusion (version)	Jazz (version)
Intel Compilers and MKL •C/C++/F77/F90	**11.1**	**8.1**
PGI Compilers •C/C++/F77/F90	**9.0**	**5.2**
Absoft Fortran compilers •F77/F90	**10.2**	**9.0**
NAG F95 Compiler	**5.2**	**5.0**
Gnu gcc	**4.1**	**3.2**
MPI	**MPICH-2**	**MPICH-1**

- **Fusion default MPI is MPICH-2 over InfiniBand (MVAPICH-2)**
- **Always compile applications on Fusion**
  - Never build binaries on another machine

Software – Commercial software

Package	Fusion (version)	Jazz (version)
Red Hat Enterprise Linux	5.4	3.x
Mathematica (serial)	7	5.0
Matlab (serial)	R2009a	R13
IDL	7.1	6.2
NCAR Graphics	5.0	4.4.1
STAR-CD	4.08	4.08
GAUSSIAN/LINDA	G09	G03
VASP	4.6.x	4.6.x

Software – Profilers and Debuggers

- **TotalView (Version 8.6.2)**
  - Allows parallel debugging of MPI programs
  - Has nice graphical user interface
  - 32 process license (1 user with 32 processes, 2 users with 16 processes etc.)
- **Intel Debugger**
  - Trace analyzer and collector
  - Intel Cluster Toolkit
- **GNU Debugger – GDB**
- **JumpShot**
  - Developed in MCS
  - Automates visualization of MPI calls
- **Allinea DDT (new)**
  - Advanced source code browser shows the state of the processes within a parallel job
  - Simplifies the task of debugging large numbers of simultaneous processes
  - Allows debugging problems with deadlock and memory leaks
  - Fusion has a 32 process Allinea license (1 user with 32 processes, 2 users with 16 processes etc.)
- We are interested in your feedback on both TotalView and Allinea DDT

Benchmarks – NEK5000

- MCS Computational Fluid Dynamics – spectral element solver

Total Cores (processors)	Cores/ Node	Nodes	Fusion (sec)	Jazz (sec)	Speed-up/Core
8	1	8	810	3819	4.7
8	8	1	1193	N/A	3.2
128	1	128	81.9	282	3.4
128	8	16	70.7	N/A	4.0

Benchmarks - Stream

- **Sustainable memory bandwidth test**

	1 thread/node MB/s (copy)	8 threads/node MB/s (copy)
Fusion	8,425	24,227
Jazz	1,193	1,171
Speed-up	7.1	20.7

- For memory intensive applications, up to 20x faster per node when using all 8 cores.
- Worst case is 2.5x faster (8 independent Jazz nodes vs. 8 cores on 1 Fusion node)

Benchmarks – Intel MPI Benchmark

- Measures inter-node communications performance

	Ping Pong latency (2 nodes)	All reduce – 128 nodes (4MB packets)
Fusion	2.0 µsec	0.0107 sec
Jazz	8.7 µsec	0.116 sec
Speed-up	4.4	10.8

Access to Fusion

- **Startup Allocations (initial account request)**
  - ANL employees receive 8,000 core-hours (1,000 node hours)
  - Provides resources for getting familiar with Fusion
  - Go to https://accounts.lcrc.anl.gov/
    - Click: Account Request Page
- **Next step is to request a project (due Friday 11/6)**
  - Go to https://accounts.lcrc.anl.gov/
    - Click: Project Request Page – requires an Argonne PI
  - Big projects are 300,000 core-hours or larger
- **Friendly user period: November-December**
  - Limited pre-production access
  - Spectrum of applications; large jobs
  - Send email to support@lcrc.anl.gov

Connection to Fusion

- **Same approach on Fusion as on Jazz**
  - Log in using SSH with Private/Public key authentication
- **Login nodes have 8 cores, 64 GB memory**
  - ssh to fusion.lcrc.anl.gov
  - DNS round robin to flogin*.lcrc.anl.gov
  - Normal UNIX shells (tcsh/bash)
- **Use login nodes for**
  - Code development
  - Submit jobs, monitor jobs
  - Run visualization apps e.g. Jumpshot, Totalview
  - Debugging short jobs

Job submission

- **Torque/MAUI (Jazz used PBSPro)**
  - Handles scheduling, starting, stopping of jobs
  - Commands are mostly the same
  - Highly configurable by the user
    - Email notification when jobs finish
    - Interactive job submission
  - Works with the account system (qbank) to track hours
  - Some additional options for multicore machines (webpage will be updated with new instructions)

Job queues

- **Starting with familiar queues and policies**
  - Tweaking to optimize research throughput, support HPC
- **Batch queue**
  - Default queue; intended for most work;
  - Allocated and charged for all 8 cores/node
  - Includes most cluster nodes (nominally 300)
- **Bigmem queue**
  - 16 96-GB nodes
  - Supports jobs requiring a lot of shared memory
- **Shared queue**
  - 4 nodes always available; allocated/charged by individual cores
  - Multiple users can run on these 4 nodes
  - Useful for debugging

Getting help using LCRC

- **Check the LCRC web pages**
  - Using Jazz – http://www.lcrc.anl.gov/jazz/Documentation/
  - Fusion pages are under construction – January rollout
- **Email support@lcrc.anl.gov for**
  - System problems (nodes down, scheduler, interconnect)
  - Reservation requests
  - Help compiling code and installing new software
  - Job submission scripts (examples available on webpage)
  - Performance improvement consultation

Upcoming Events

Events	Nov. 09	Dec. 09	Jan. 10
Friendly users	→		
Fusion project requests due	Nov. 6		
Transition from Jazz to Fusion		→	
Jazz shut down		Dec. 31	
Fusion production target			Jan.1
Introduction to Fusion and MPI			Jan. 21

Questions and Comments

2002 Supercomputer. No rust. Free to good home. Shipping not included. Call Ray at 2-5751.