A Crash course to (The) Bighouse CAEN Brown Bag

Brock Palen
brockp@umich.edu

CAEN Brown Bag, Oct 10th

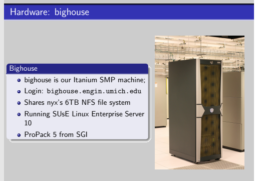
Outline

1. Resources
   - Configuration
   - Hardware
2. Architecture
   - ccNUMA
   - Altix 4700 Brick
   - Dual Fat Tree
   - cpu sets
   - NUMA Effects
3. Software Performance
   - MPI Code
   - OpenMP Code

Hardware: bighouse

Bighouse

- bighouse is our Itanium SMP machine;
- Login: `bighouse.engin.umich.edu`
- Shares nyx's 6TB NFS file system
- Running SUsE Linux Enterprise Server 10
- ProPack 5 from SGI



Hardware: bighouse

Bighouse

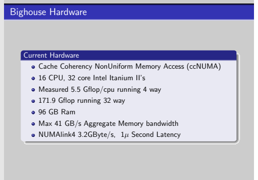
- bighouse is our Itanium SMP machine;
- Login: bighouse.engin.umich.edu
- Shares nyx's 6TB NFS file system
- Running SUsE Linux Enterprise Server 10
- ProPack 5 from SGI

ProPack: Provides performance tools, hardware tools and MPT(MPI) libraries

Bighouse Hardware

Current Hardware

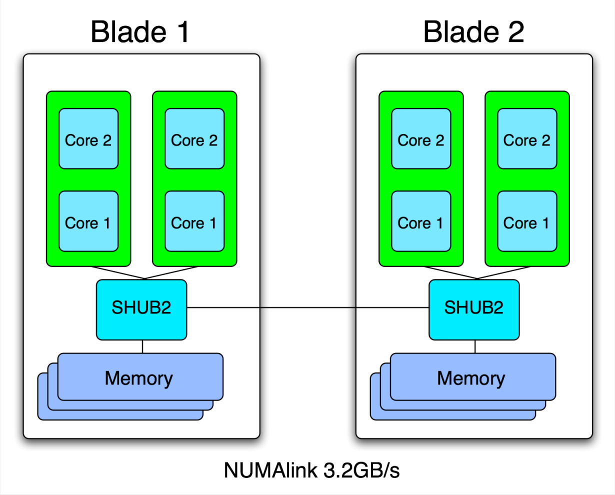
- Cache Coherency NonUniform Memory Access (ccNUMA)
- 16 CPU, 32 core Intel Itanium II's
- Measured 5.5 Gflop/cpu running 4 way
- 171.9 Gflop running 32 way
- 96 GB Ram
- Max 41 GB/s Aggregate Memory bandwidth
- NUMAlink4 3.2GByte/s, 1μ Second Latency

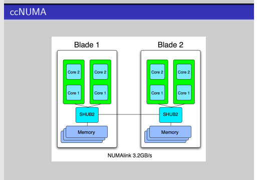


Bighouse Hardware
Current Hardware
- Cache Coherency NonUniform Memory Access (ccNUMA)
- 16 CPU, 32 core Intel Itanium II's
- Measured 5.5 Gflop/cpu running 4 way
- 171.9 Gflop running 32 way
- 96 GB Ram
- Max 41 GB/s Aggregate Memory bandwidth
- NUMAlink4 3.2GByte/s, 1μ Second Latency

1. Keeps cache lines in sync both good and bad
   Makes for easy programming, places upper limit vs. CRAY
2. HPL P=2 Q=2 N=20000, MKL no threads, MPT
3. HPL P=4 Q=8 N=20000, MKL no threads, MPT
4. 2 nodes have 24GB, 6 have 8GB

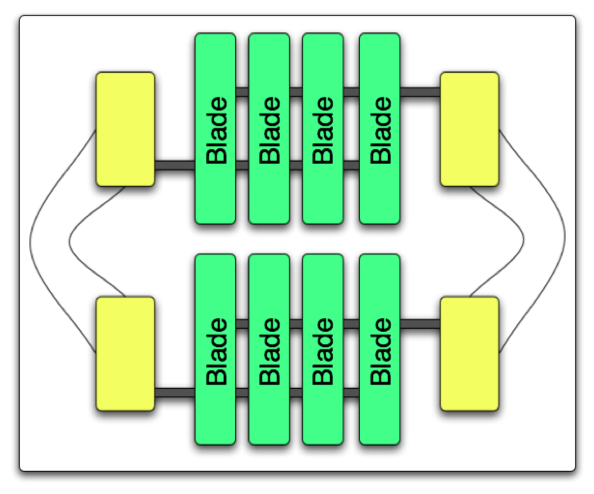
ccNUMA

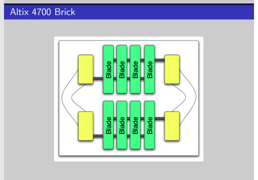




1. Itanium II's 9000's.
   L1 16k/d 16k/i
   L2 256k/d 1024/i
   L3 4MB
2. SHUB2 I FORGOT IT!
   It Sits between the cpus and memory Numa link connects to it.
   This is where the magic happens

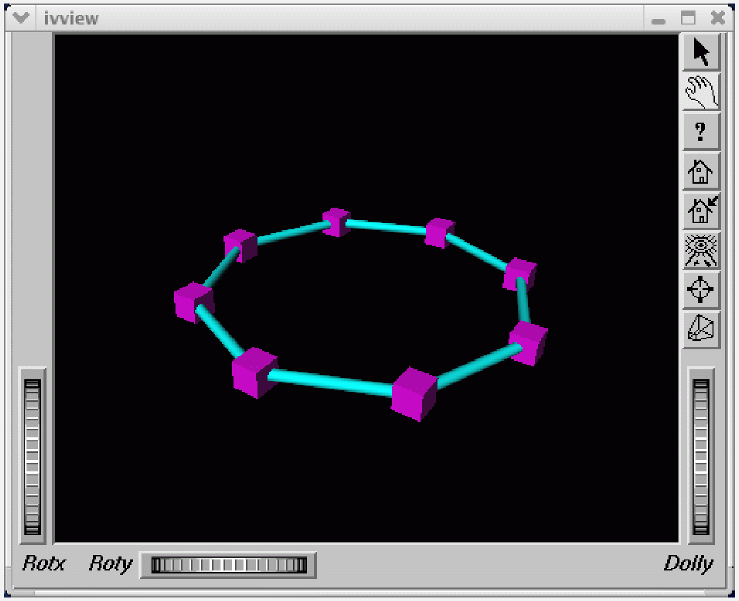
Altix 4700 Brick

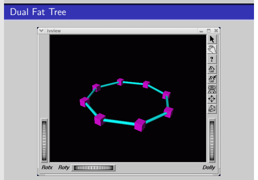



Altix 4700 Brick

1. Each blade has 2 NUMAlink connections, each goes to a differnt router, each router has a 200 nanoSec pass time.

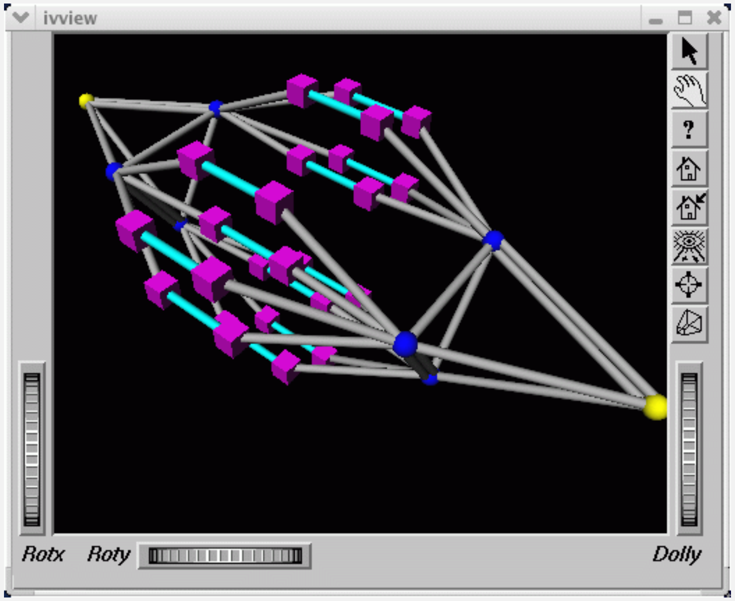
Dual Fat Tree

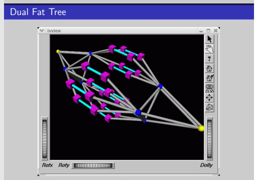



Dual Fat Tree

1. This would be our layout but turns out its not this would apply to the 450 if we had it.

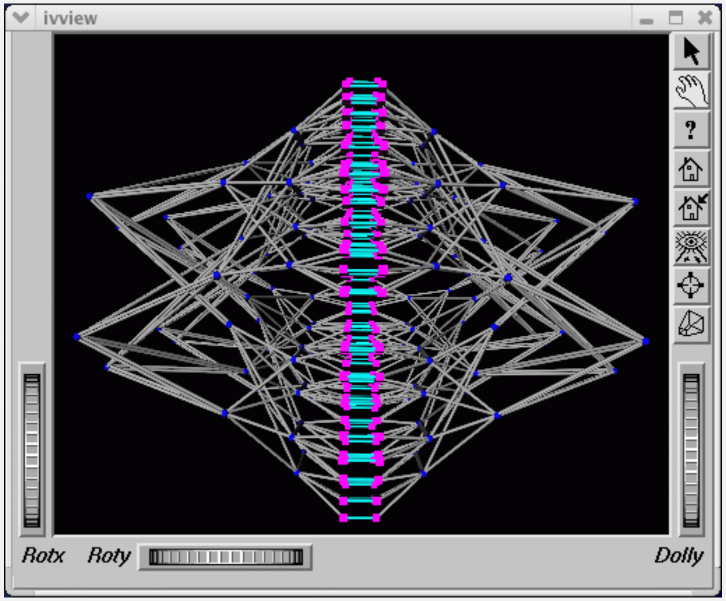
Dual Fat Tree

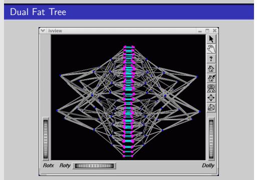




1. this is our layout (at 8 blades), We only have half the ring though, max number of hops will equal up to 16 blades 64 cores

Dual Fat Tree





1. Provides 1024 cores 512 sockets
   This is the max supported config from SGI, system can add one more router out for 1024 sockets 2048 cores, MTTF is to high though

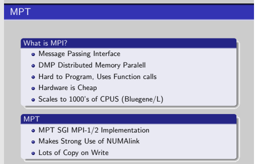
MPT

What is MPI?

- Message Passing Interface
- DMP Distributed Memory Paralell
- Hard to Program, Uses Function calls
- Hardware is Cheap
- Scales to 1000's of CPUS (Bluegene/L)

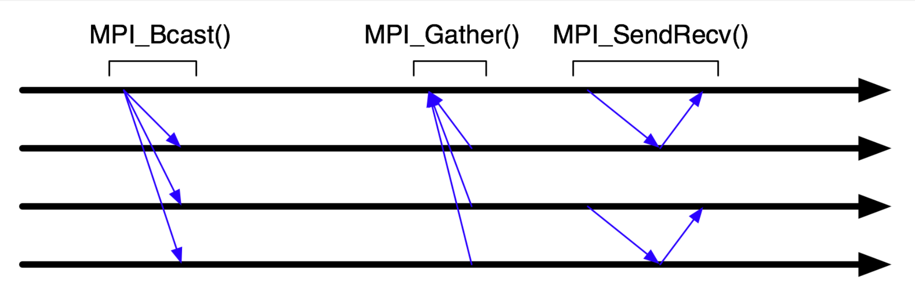
MPT

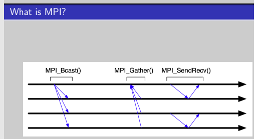
- MPT SGI MPI-1/2 Implementation
- Makes Strong Use of NUMAlink
- Lots of Copy on Write


MPT
What is MPI?
- Message Passing Interface
- DMP Distributed Memory Paralell
- Hard to Program, Uses Function calls
- Hardware is Cheap
- Scales to 1000's of CPUS (Bluegene/L)
MPT
- MPT SGI MPI-1/2 Implementation
- Makes Strong Use of NUMAlink
- Lots of Copy on Write

1. www.mpi-forum.org
2. We have similar SM ability on nyx though OpenMPI

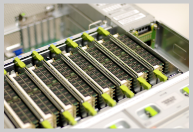
What is MPI?




What is MPI?

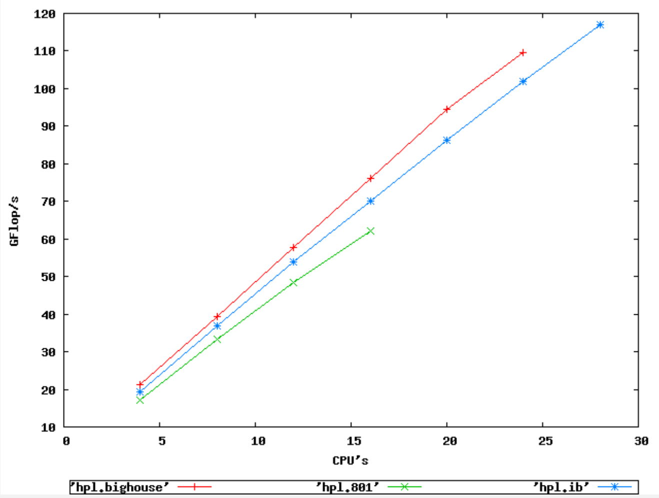
1. Duplicates allot of data between processes
2. nothing shared unless given

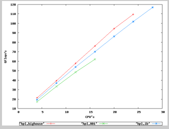
The Challenger



1. `nyx801` Owned by Dr. J Norman MD, PHD.
   64 GB ram, on 8 sockets, dual core 8218's

MPI Perfromance/HPL





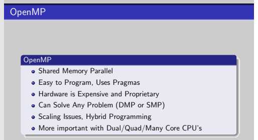
1. **Hardware:**
   'hpl.bighouse' is bighouse
   - mpt
   - mkl no thread

   'hpl.801' is nyx801
   'hpl.ib' is EMike Nodes, dual core dual socket opt2220, 16 GB ram, DDR Infiniband 20Gbit/s $< 4\mu$ Sec. Latency
   - openmpi-1.2-pgi, OFED
   - goto-blas
2. point out gapping as number of CPUS increase
   Why Bighouse is surperior, but not at this size and price

OpenMP

OpenMP

- Shared Memory Parallel
- Easy to Program, Uses Pragmas
- Hardware is Expensive and Proprietary
- Can Solve Any Problem (DMP or SMP)
- Scaling Issues, Hybrid Programming
- More important with Dual/Quad/Many Core CPU's

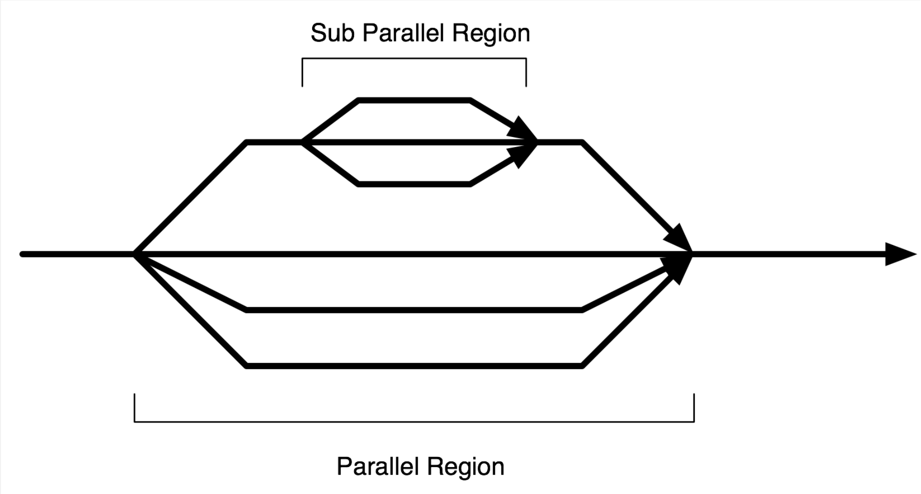


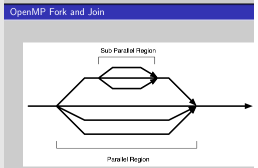
OpenMP

OpenMP
- Shared Memory Parallel
- Easy to Program, Uses Pragmas
- Hardware is Expensive and Proprietary
- Can Solve Any Problem (DMP or SMP)
- Scaling Issues, Hybrid Programming
- More important with Dual/Quad/Many Core CPU's

1. Thread sync issues, implemented with libpthread normally
2. in GCC 4.1
3. Can FLOOD bus/interconect because of cache sync issues

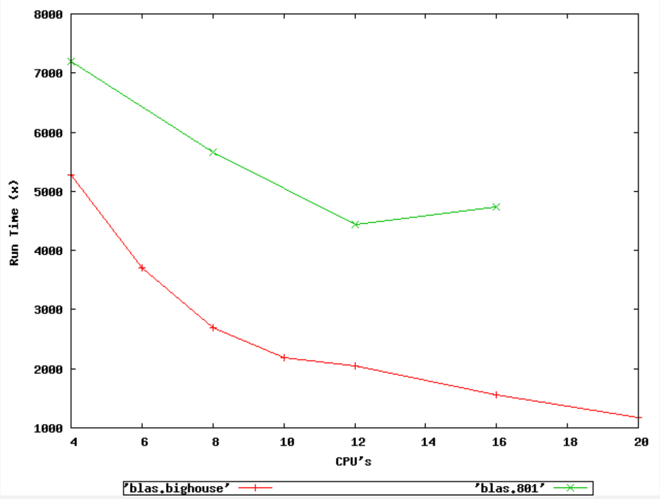
OpenMP Fork and Join





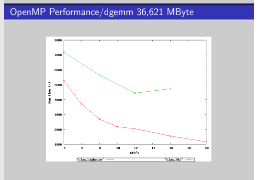
1. This is what allot of Direct CAE apps use (Nastran Abaqus)
   Most interative solvers are dense matrix solvers in DMP (LS-DYNA)
2. STRESS This is bighouse's benefit, it can the ram and SMP ability to run these codes at a speed a regular cluster could never do

OpenMP Performance/dgemm 36,621 MByte



2009-11-10

Bighouse Crash
└Software Performance
 └OpenMP Code
  └OpenMP Performance/dgemm 36,621 MByte



1. `www.netlib.org/blas`
   Bighouse uses MKL
   nyx801 Uses ACML-pgi-mp
   2 equal square matrix's of random nubmers with a dim of: 40,000 Doubles.
   This is 3,200,000,000 (3.2 billion numbers)
   Same building block used in hpl

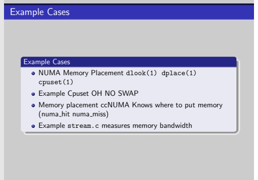
Example Cases

Example Cases

- NUMA Memory Placement `dlook(1) dplace(1) cpuset(1)`
- Example Cpuset OH NO SWAP
- Memory placement ccNUMA Knows where to put memory (numa_hit numa_miss)
- Example `stream.c` measures memory bandwidth

2009-11-10

Bighouse Crash
└Software Performance
 └OpenMP Code
  └Example Cases



- **Example 1**, cpu sets
- cpuset -c brockp -f brockp/cpuset.conf
- echo $$ >> /dev/cpusets/brockp/tasks
- **Example 2**, link speeds
- `linkstat -A`
- `pmchart` numa.mem.util.used
- `pmchart` numa.link.send_bytes
- `run stream.c`

Questions

Questions?

Questions?

`http://cac.engin.umich.edu/resources/bighouse.html`

`cac-support@umich.edu`