# HYDRA: A&M's p5-575+ IBM Cluster

Spiros Vellas
Associate Director
Texas A&M Supercomputing Facility

# HYDRA: A&M's p5-575+ IBM Cluster Highlights in

- Architecture
- Acquisition & Installation
- Configuration

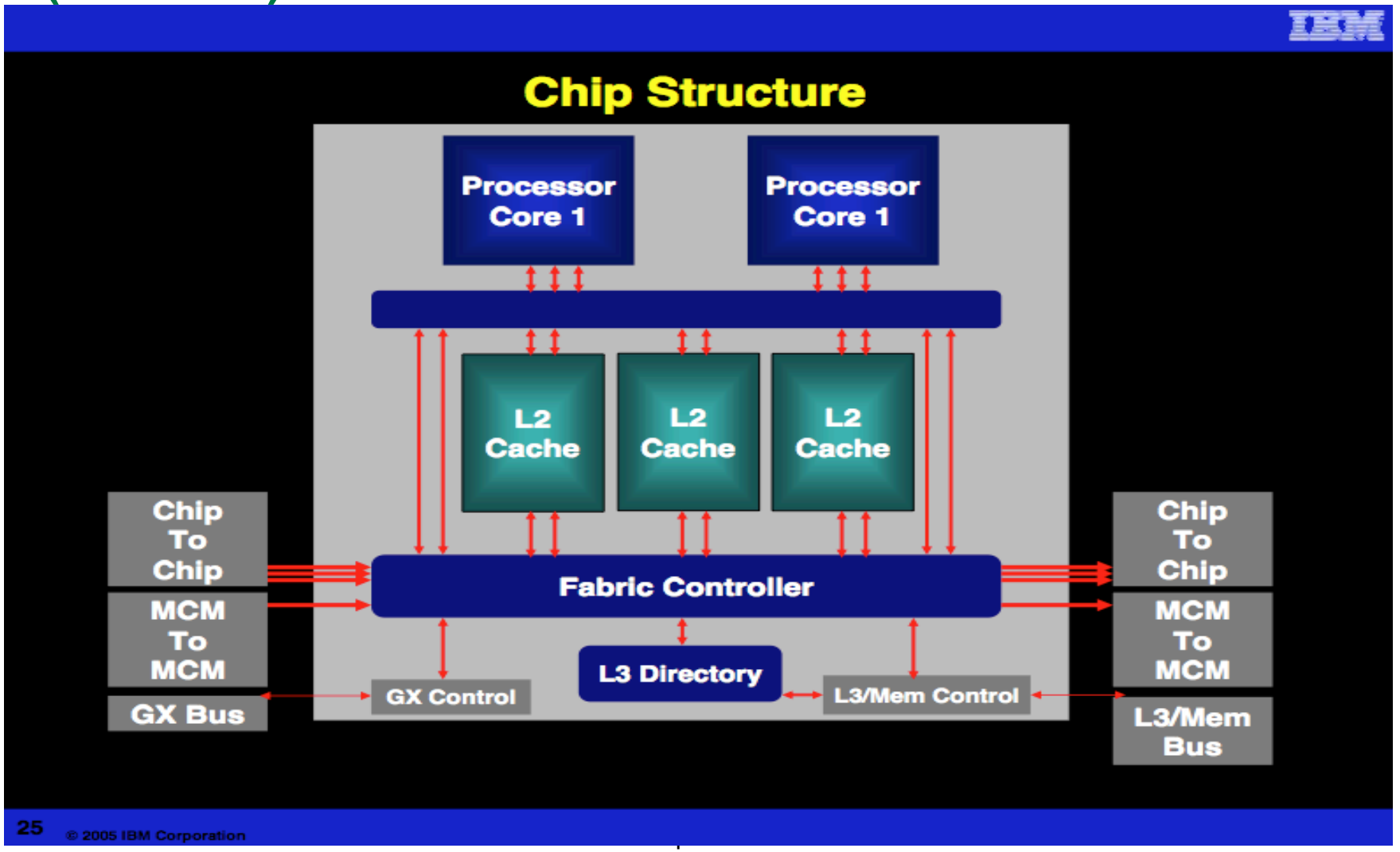
# Architecture

- The 1.9GHz Power5+ processor chip
- The 16-processor p5-575 SMP node
- The High Performance Switch (HPS)

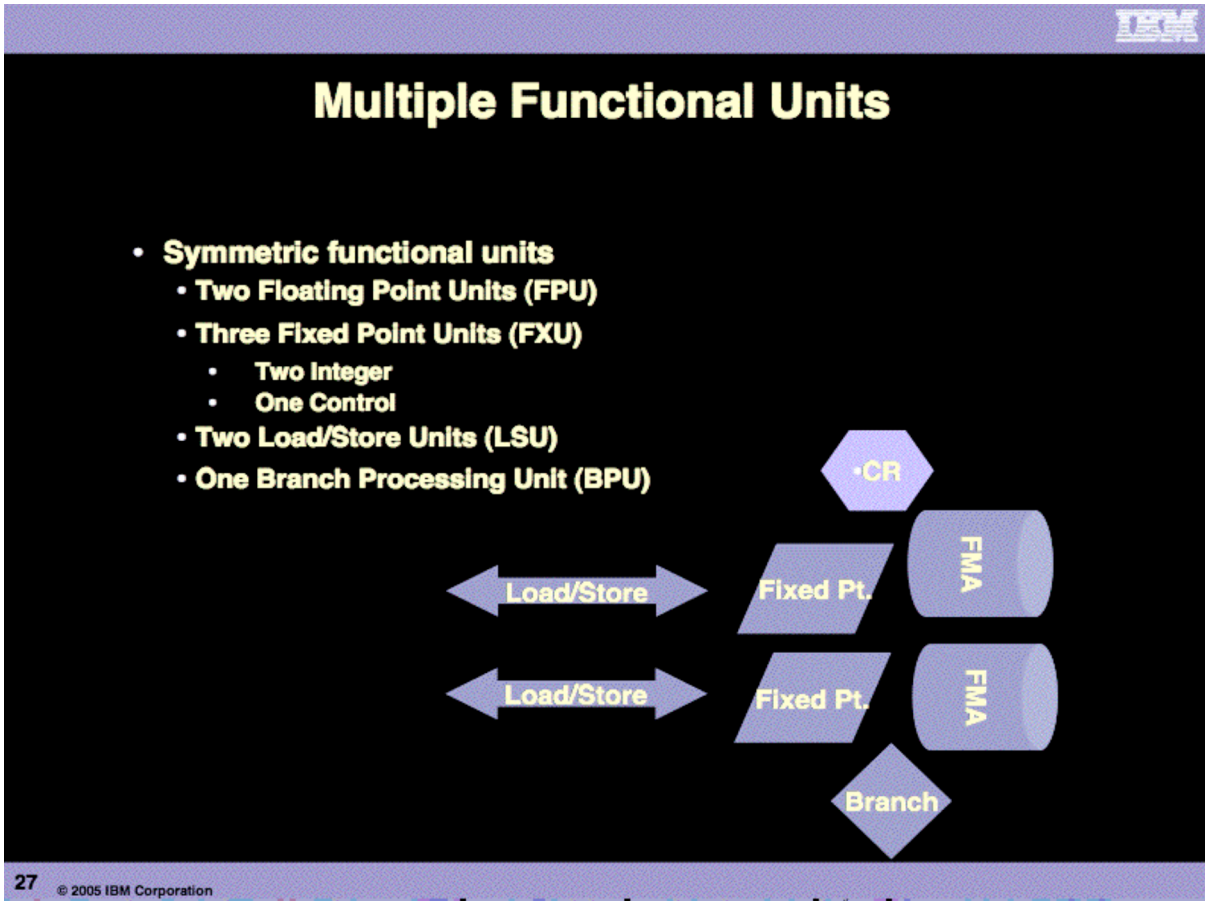
# Power5+ System Features

- Dual core (CPU) DCM/chip
- 1.9GHz CPU
- 7.6 Gflops per core
- Shared L2 cache
- Shared L3 cache
- Shared memory
- Multiple page size support
- Simultaneous Multi-Threading (SMT)

# Power5+ Dual Core Module (DCM)



# Power5+ : 8-way Superscalar




- speculative out-of-order execution
- up to 8 cache misses
- can issue up to 8 instructions per cycle
- Hw prefetching
- Branch prediction
- can have up to 200 instructions active

# Power5+ Memory System

- L1 Cache - 1 per core
  - 2 clock cycles latency
  - 32KB Data; 4-way set associative
  - 64KB Instruction; 2-way set associative
- L2 Cache - 1 per DCM
  - 13 clock cycles latency
  - 1.9MB unified; 3 independent 10-way set associative partitions (separate cache controllers)
  - 64KB Instruction; 2-way set associative
  - Full h/w coherence in SMP

# Power5+ Memory System

- L3 Cache - 1 per DCM
  - 87 clock cycles read latency
  - 36 MB unified; not on chip
  - 3 slices each 12-way set associative; separate controller (on chip)
  - L3 directory on chip
  - Operates at 1/2 of processor frequency
  - 30.4 GB/s between L3 and Power5+ chip
  - 243.2 GB/s aggregate per 16-way node

# POWER5+ Memory System

- Main Memory
  - 64 dimm slots (32GB memory per node @ tamu)
  - 1/2 GB DDR2 dimms @ 533MHz; maximizes memory throughput
  - 12.7 GB/s Bw per core

# Translation Lookaside Buffer (TLB)

- 2048 entries
- Page sizes
  - 4KB; default
  - 64KB; dynamically set
  - 16MB; kernel level setting required
  - 16GB; kernel level setting required
- Appropriate page size selection can substantially improve memory performance

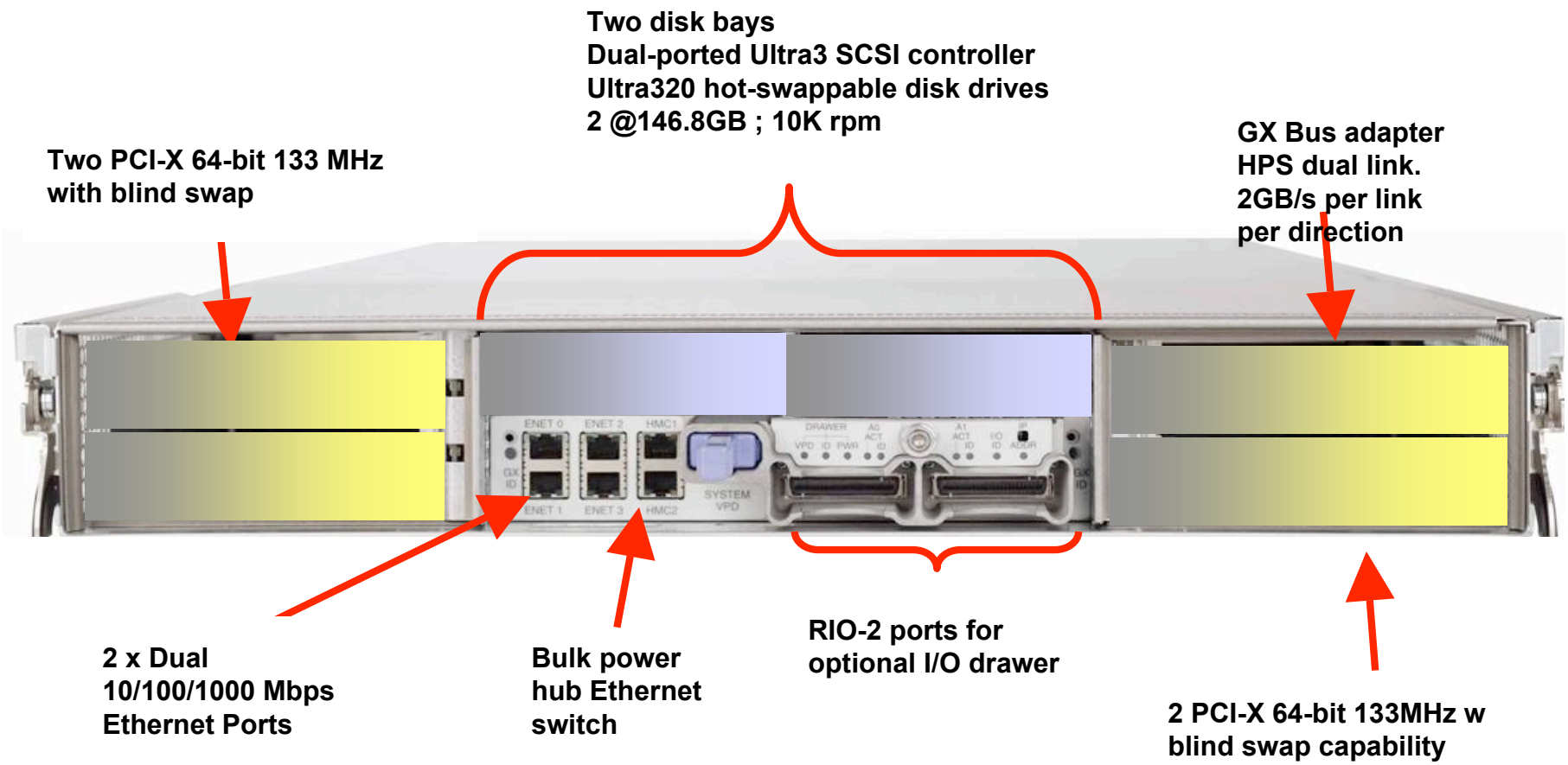
# Simultaneous Multi-threading (SMT)

- Each 2-core DCM appears as a 4-way SMP to AIX and applications
  - 2 execution threads per core/cpu
  - a thread: an independent sequence of instructions
- Suits naturally superscalar out-of-order execution feature
- Dynamic switching between single and multi- threaded mode
- Symmetric multiprocessing (SMP) programming model
- Better processor utilization. Utilizes unused cycles
- Improves the performance of many multi-threaded codes
- Some codes do not benefit

# P5-575+ Node Logical Configuration

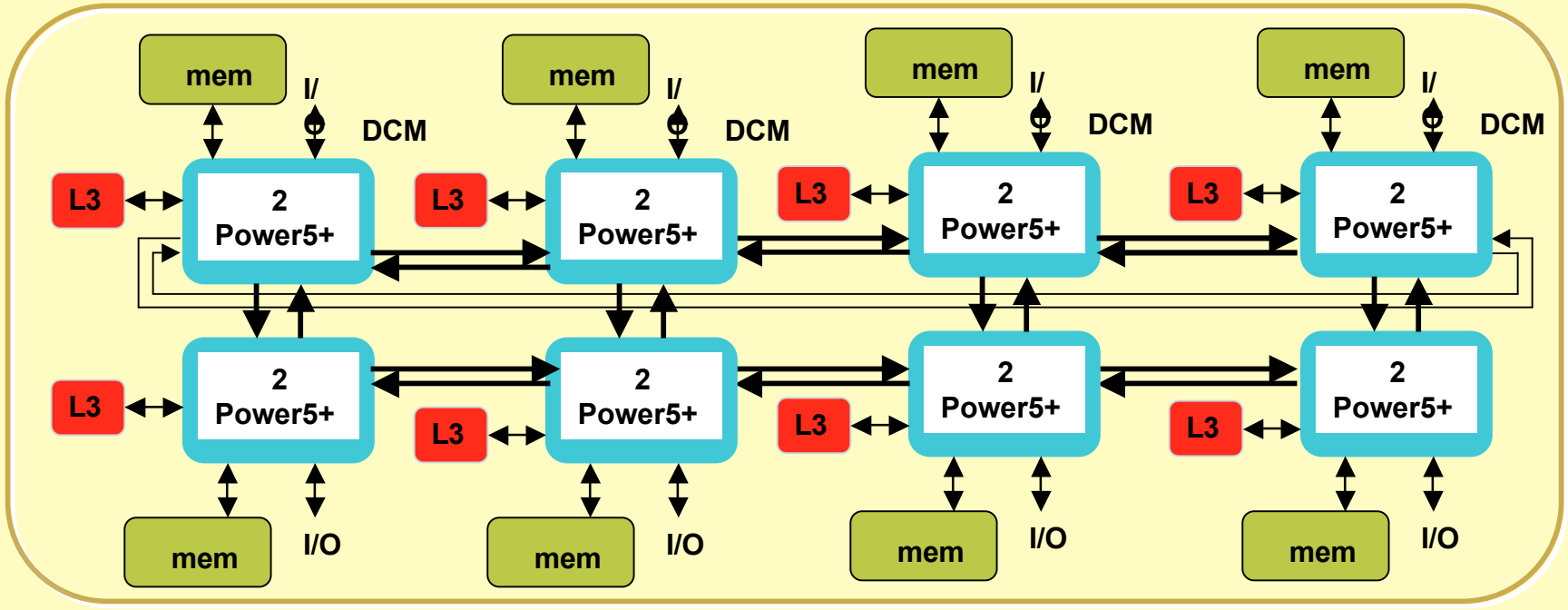
- 1 Logical Partition/node
- 16-way Power5+ @ 1.9GHz
- DIMMs (32 GB of DRAM
  - 64 x 1/2 GB DDR2 @ 533MHz)
- 2 internal 146.8GB scsi disks

# Rear view of p5-575+ node With PCI-X and RIO-2



# P5-575+ node: A 16-way SMP

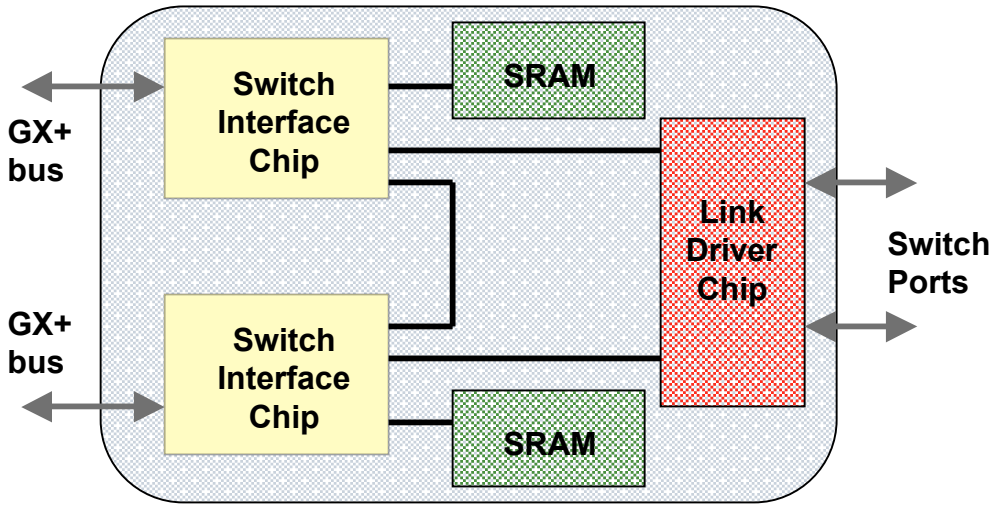
- P5-575+ node uses 8 interconnected dual-core modules (DCMs)
  - Intra-DCM fabric buses operate at processor speed
  - Inter-DCM fabric Buses operate at 1/2 processor speed
  - Fabric buses operate as coherent & distributed interconnect



# A Very brief look at Performance

| *ATTRIBUTES* | 16-way POWER5+ | 16-way Altix 4700 |
|---|---|---|
| Processor speed | 1.9 GHz | 1.6 GHz Montecito |
| Memory Bandwidth per CPU | ~12 GBps | ?? GBps |
| Peak GFLOPS/node | ~120 | ~102 |
| LINPACK GFLOPS/node | 111.4 | 96.2 |
| SPECint_rate<br>SPECfp_rate | 314<br>571 | 265<br>366 |
| STREAM (tuned) | 98,874 | 22,800 |

# HPS Switch Network Interface (SNI)

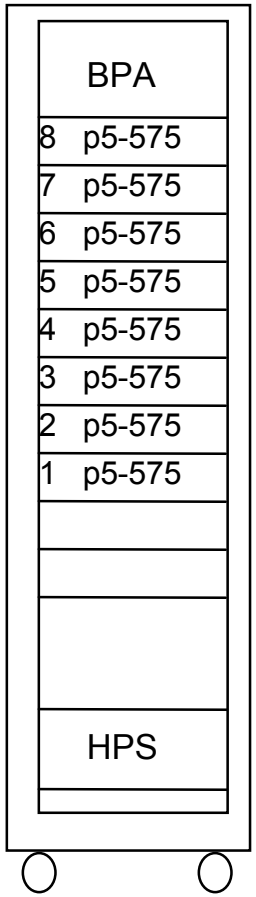



- SNI connects p5-575 node to HPS Switch
- SNI directly connects to 2 GX host buses
- 2 HPS ports per SNI
  - 2 full-duplex per SNI
  - 8GB/s aggregate BW

# Acquisition & Installation

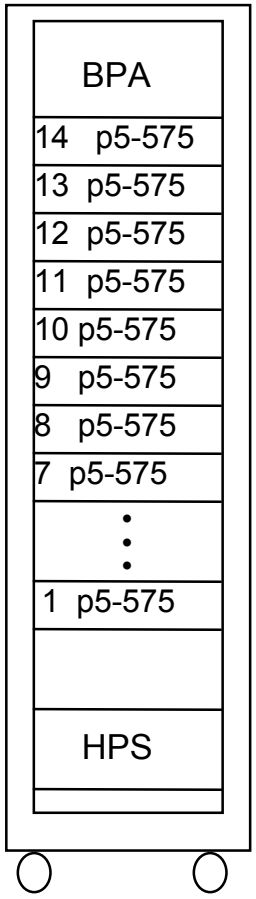
- 2nd half of 2005 SC looking for new system
- Jan 2006 APIT & VPR offices join efforts to consider participation in the proposed SURA-IBM agreement (SURA BOD approved on May 02)
- Mid March 2006 SURA-IBM system offerings are revealed
- Many technical exchanges follow ….

# SURA-IBM Config 1: 8-node frame




- 8 16-way p5-575 nodes: 128 1.9GHz Power5+'s
- Single-plane High-Performance Switch (HPS) connects 8 nodes
- 19" management rack: p520 workstation for cluster mgt, 1 HMC console
- Peak HPS bandwidth: 32 GB/s
- Peak Network (Ethernet) bandwidth: 32 Gbits/s
- Total system memory: 128GB or 256GB
- Total local (SCSI) disk: 2.34TB
- Cumulative SpecFP_rate: 4576
- Price: $410K (256GB memory)
- Price per additional node: $67,433+

# SURA-IBM Config 2: 14-node frame




- 14 16-way p5-575 nodes: 224 1.9GHz Power5+'s
- Single-plane High-Performance Switch (HPS) connects 14 nodes
- 19" management rack: 1 p520 workstation for cluster mgt, 1 HMC console
- Peak HPS bandwidth: 56 GB/s
- Peak Network (Ethernet) bandwidth: 56 Gbits/s
- Total system memory: 224GB or 448GB
- Total local (SCSI) disk: 4.11TB
- Cumulative SpecFP_rate: 8008
- Price: $660K (448GB memory)
- Price per additional node: $67,433+

# Where is the money …

- By May 2006 APIT & VPR commit
- By Aug 17, College Geoscience & Computer Science Dept also commit
- Aug 18, a 40-node p5-575 cluster + DDN Raid is agreed upon by all parties
- September 29, A&M issues purchase order
- Oct 6, IBM delivers system in 12 crates

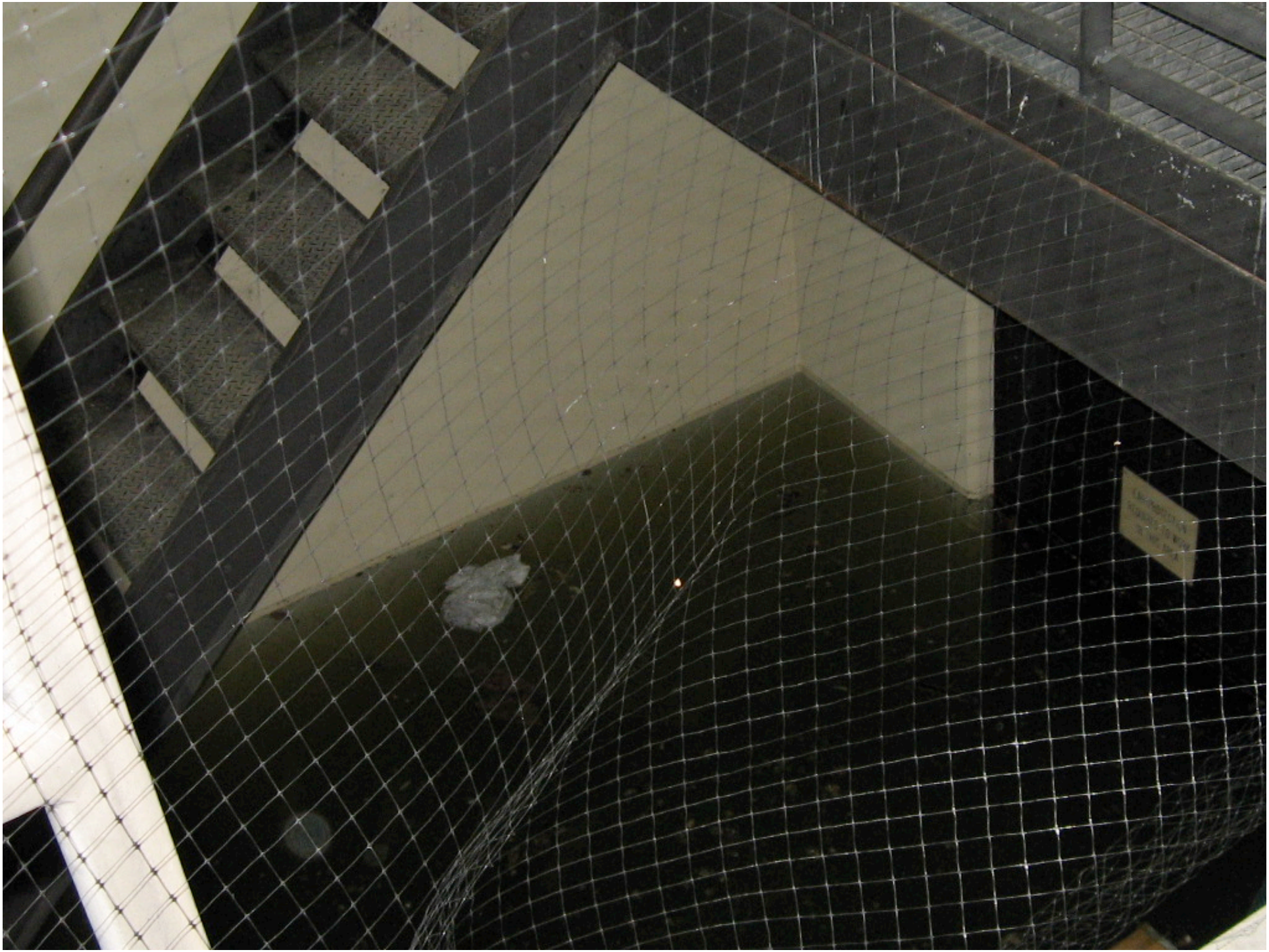
# Installation: eventfull …

- Machine room required major power & AC upgrades: Add ~200kilo-watts & 30-ton cooling unit
- Liebert AC units not available soon enough
- Physical Plant proceedures & contractor scheduling took time …
- Nov 10, IBM power-tests frames individually via temporary arrangements
- Nov 20, Liebert AC 30-ton unit arrives
- Nov 26, Disaster: flooding in machine room basement destroys transformer
- Nov 27, 2006 DDN raid arrives
- Nov 29-30, Basic AIX install complete on all nodes

# Installation: the end of the tunnel

- Dec 8, power upgrade almost complete & Liebert AC unit powered on permanently
- Dec 11, IBM power tests whole system & HPS verification test complete-without DDN connection
- Dec 21, heat exchanger doors in each frame connected to water supply
- Jan 4-5 2007, DNN raid installed & GPFS file systems /work and /scratch configured
- Late Jan 2007, allow friendly use
- Mar 1 2007, HYDRA in production use

# TOTAL COST: ~$2.9M

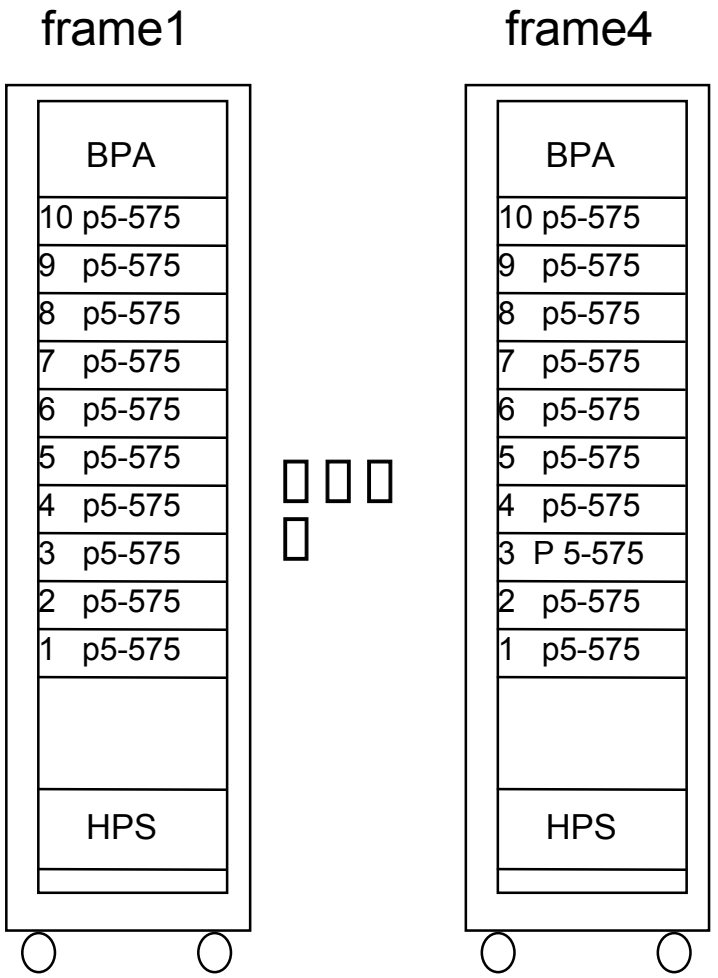


# Credits: Many People … lots of hours

- Dr. Cantrell; Dr. Ewing; Prof. Taylor; Prof. Kjerfve; Mr. Flournoy (A&M's SURA rep)
- SC Steering Committee: Prof's Panetta (Chair); Adelson; Gagin; Hall; Rauchwerger; Roschke; Dr. Johnson; Mr. Putnam; Spiros
- Dr. Thomadakis; Messrs. Dang, Jackson
- Messrs. Johny Rauser & Myron Walden
- Messrs. Willis, Chris Pruitt, Josh Bartosh

# A&M p5-575+ Frame Features

- 4 racks (40 p5-575+ nodes 32GB memory per node)
- 10 p5-575+ nodes per rack
- Interconnected w IBM's dual-plane HPS switch
- All racks connected to CSM (Cluster System Mgt) server
- All nodes HPS-connected
- All nodes GigE-connected
- Local SCSI storage/rack ~2.9TB
- A water-cooled door per rack

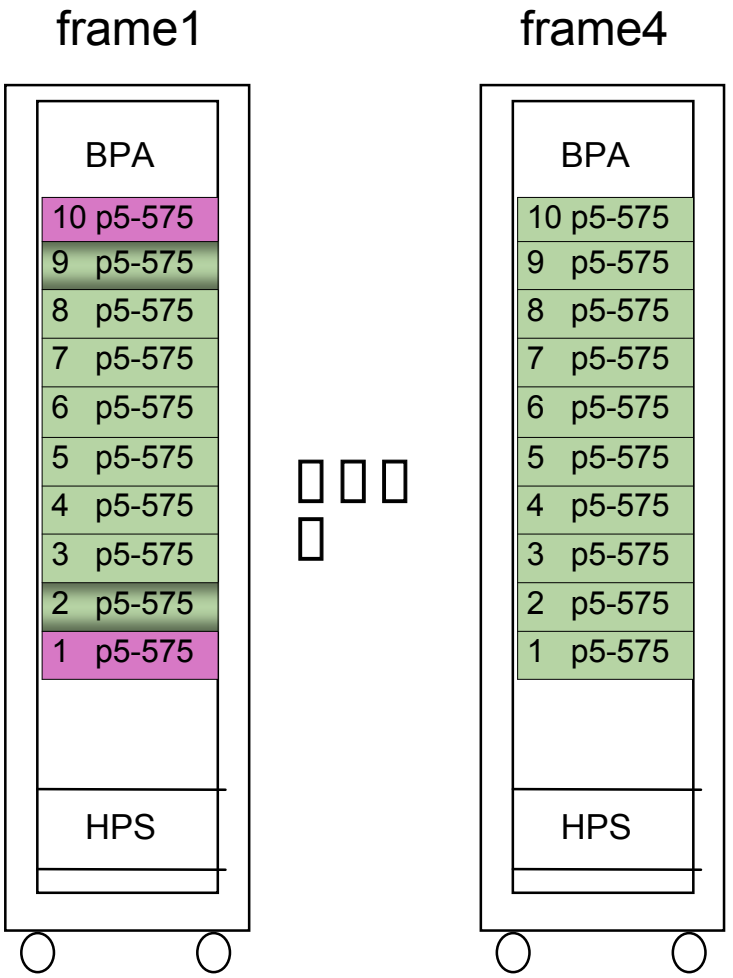


# Configuration

- A Balanced System for a Varied Job Profile: Powerfull smp nodes, Powerfull interconnect, Powerfull I/O subsystem
  - Distribute dual-plane HPS bandwidth evenly across all nodes
  - Commit sufficient nodes for I/O to leverage DDN's High Performance
  - Balanced division of nodes
    - 38 compute nodes
    - 4 I/O nodes
    - 2 nodes for login and interactive processing
  - Configure 10% all memory to be in large pages, 16MB
  - Enable simultaneous multithreading (SMT) when appropriate
  - Adopt IBM's Load Leveler as Batch facility
  - Batch queues: meet needs of contributors and satisfy the requirements of general jobs

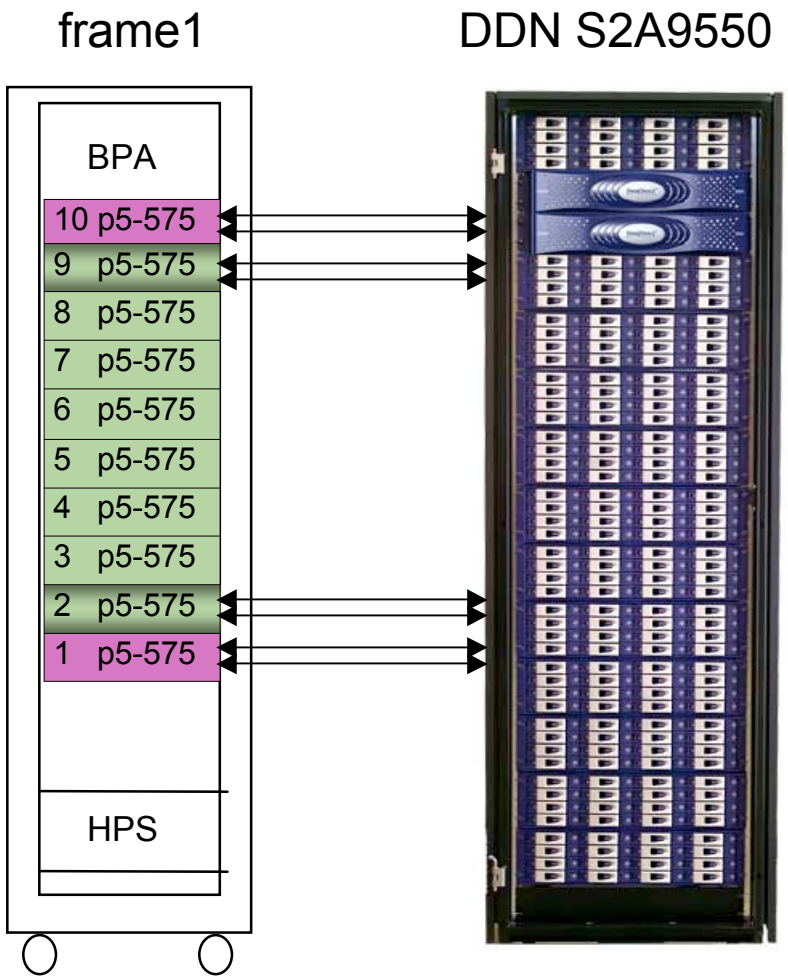
# Node Role

- Login nodes: f1n1, f1n10
- I/O nodes: f1n1, f1n2, f1n9, f1n10
- I/O nodes connect to DDN raid array directly
- 38 Compute nodes (includes 50% of f1n2 & 50% f1n9)
- Acess to compute nodes by batch only
- Login nodes not available for batch processing



# Powerful I/O w DataDirect RAID

- 8 400MB/s per direction FC direct links
- ~22TB of raw FC disk storage capacity
- DDN raid hosts parallel (GPFS) & global files systems: /work and /scratch
- Sustain 2 concurrent disk failures without drop in performance



# Simultaneous MultiThreading (SMT)

- Power5+ processor can concurrently dispatch instructions from 2 independent execution threads, thus functioning and appearing to the OS as 2 virtual processors
- Can improve program performance & system throughput by an avg of 30%
- Currently NOT enabled because of incorrect scheduling of threads by AIX. Requires general OS and LoadLever upgrade. Will rectify in near future

# Batch Queues

```
> qlimit
Per job step limits per class:

Class              Walltime    CPU time   Max   Max   RA  Description
                                          Tasks CCpus
smp_normal         24:00:00   200:00:00    16    16      shared-memory
smp_long           96:00:00   400:00:00     1     2      shared-memory
mpi32              24:00:00   800:00:00    32           Max 32-proc
mpi64              12:00:00   800:00:00    64           Max 64-proc
mpi128             06:00:00   800:00:00   128        Y  Max 128-proc
mpi256             04:00:00  1000:00:00   256        Y  Max 256-proc
cs_group           24:00:00   200:00:00   128        Y  CS Dept Excl
geo_group          24:00:00   400:00:00    32        Y  Geoscience Excl

  RA = Restricted Access Class
  CCpus = ConsumableCpus
```

# Batch Queues … in action

```
# listjobs
Step Id                  Owner       Class      Queue Date  Disp. Date  Ndes ST Type Cpus
------------------------ ----------- ---------- ----------- ----------- ---- -- ---- ----
f1n2.3858.0              m0u1586     mpi32      04/24 18:00 04/24 21:30 2    R  PAR    32
f1n9.3758.0              m0u1586     mpi32      04/24 17:57 04/24 21:20 2    R  PAR    32
f1n2.3797.0              yubofan     mpi32      04/24 08:18 04/24 08:18 8    R  PAR    32
f1n9.3755.0              yubofan     mpi32      04/24 17:28 04/24 17:28 4    R  PAR    32
f1n2.3868.0              giese       mpi64      04/24 20:51 04/24 20:51 6    R  PAR    48
f1n9.3723.0              yubofan     mpi64      04/24 13:27 04/24 13:27 8    R  PAR    64
f1n2.3740.0              hong        smp_long   04/23 09:21 04/23 09:22 1    R  PAR     4
f1n9.3713.0              yangxz      smp_long   04/24 11:18 04/24 11:18 1    R  PAR     4
           □ □ □
f1n9.3727.0              yangxz      smp_long   04/24 13:35 04/24 13:35 1    R  PAR     4
f1n2.3810.0              das018a     smp_normal 04/24 10:46 04/24 10:46 1    R  PAR     8
f1n2.3820.0              hong        smp_normal 04/24 13:19 04/24 16:01 1    R  PAR    16
f1n2.3849.0              yangxz      smp_normal 04/24 15:43 04/24 21:29 1    R  PAR    16
f1n9.3716.0              yangxz      smp_normal 04/24 13:16 04/24 13:16 1    R  PAR    16
f1n2.3875.0              yubofan     smp_normal 04/24 22:21 04/24 22:21 1    R  PAR    12
f1n9.3776.0              yubofan     smp_normal 04/24 22:21 04/24 22:21 1    R  PAR    12
f1n2.3769.0              s0v3474     cs_group   04/23 16:13                  I  PAR
f1n2.3859.0              m0u1586     mpi32      04/24 18:01                  I  PAR
           □ □ □
f1n2.3872.0              m0u1586     mpi32      04/24 21:44             0    NQ PAR
f1n2.3873.0              m0u1586     mpi32      04/24 21:48             0    NQ PAR
f1n2.3874.0              m0u1586     mpi32      04/24 21:51             0    NQ PAR
f1n9.3760.0              m0u1586     mpi32      04/24 18:03                  I  PAR
           □ □ □
f1n2.3865.0              rrl581a     mpi32      04/24 19:58                  I  PAR
f1n2.3866.0              rrl581a     mpi32      04/24 19:58                  I  PAR

54 job step(s) in queue, 15 waiting, 0 pending, 30 running, 9 held, 0 preempted
542 cpu(s) assigned to running jobs.      91.5% (542.00/592.00) loaded
```

# Sample User research

- Prof. Girimaji's group (AERO). 3D numeric simulations of strongly compressible turbulence. User Dr. Johannes Kerimo. 32p MPI runs
- Prof. Chang's group (OCEAN). Seasonal-to-interannual climate simulation & prediction. Coupled Community Atmospheric (CAM3) and GFDL Modular Ocean Model (MOM3). User Dr. Link Ji. 32p MPI runs
- Prof. Gagin's group (CHEN). Design configurations of new ferroelectric materials for critical performance properties using Ab-inition methods. User Dr. Uludogan. 32p-128p runs
- G. Creager (VPR sponsored research). Hurricane prediction, weather and environmental research. User Gerry Creager. 32p, 64p, 128p - 256p runs