The IST Cluster: an integrated infrastructure for parallel applications in Physics and Engineering

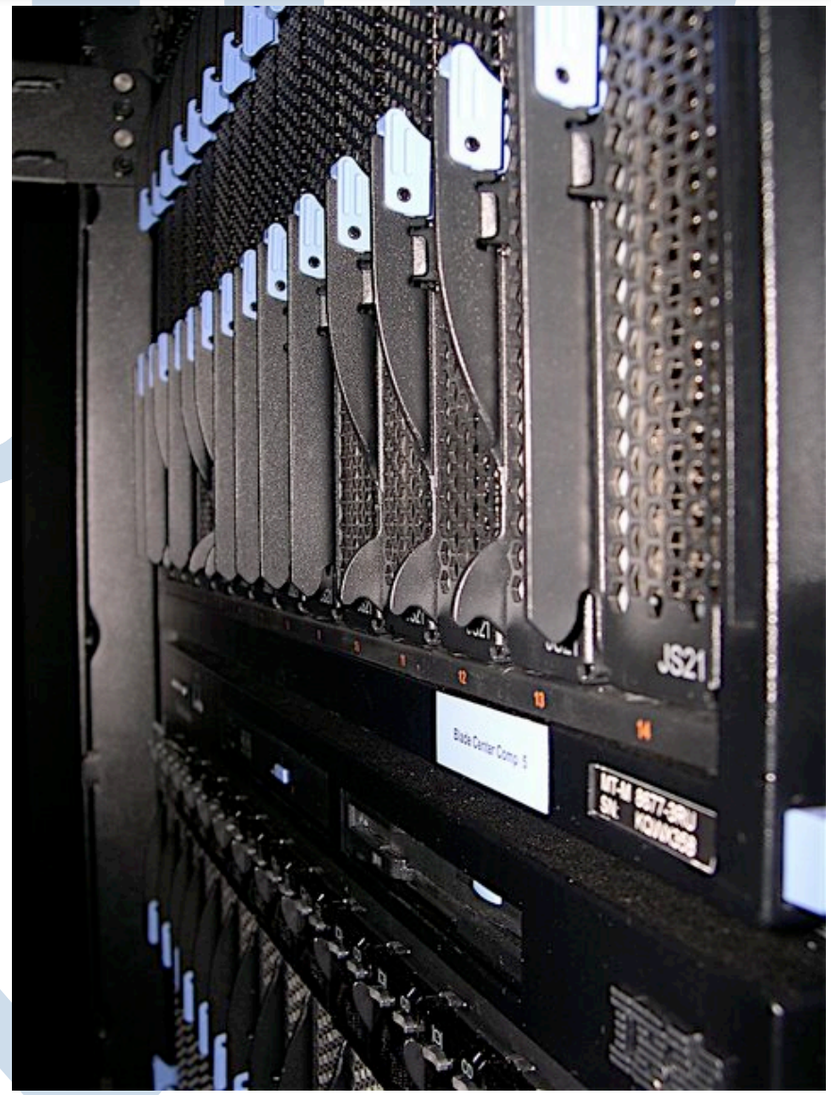
Michael Marti

L. Gargaté, R. A. Fonseca

L. L. Alves, J. P. S. Bizarro, P. Fernandes,

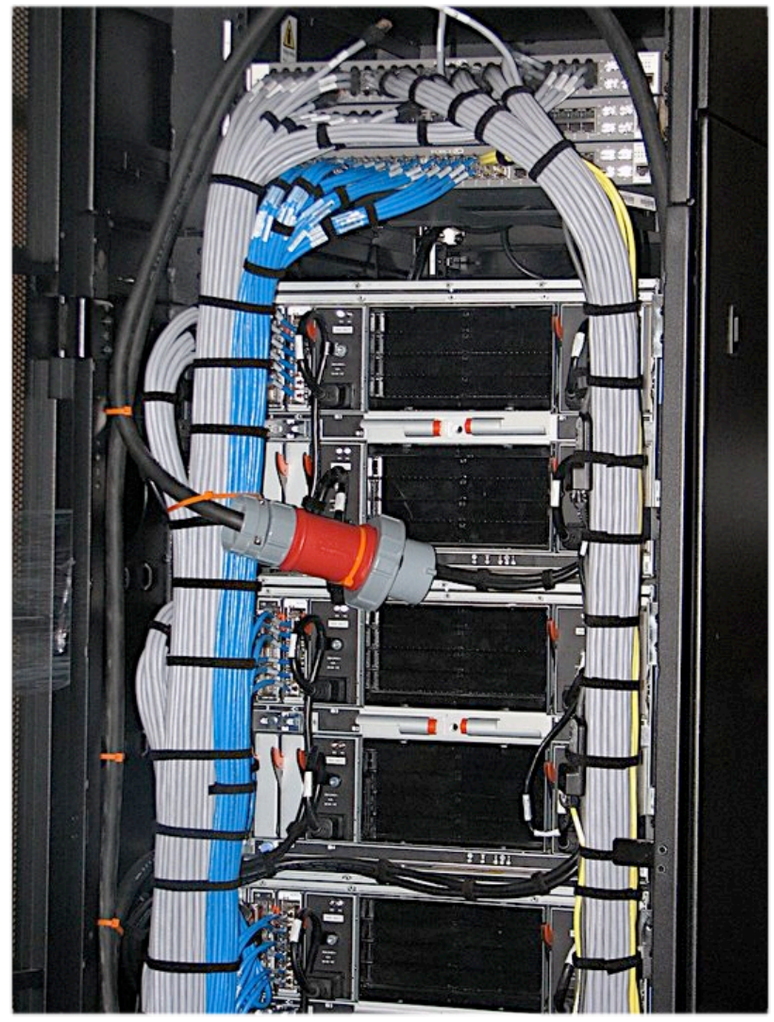
J. P. M. Almeida, H. Pina, F. M. Silva, L. O. Silva

Instituto Superior Técnico (IST)
Lisbon, Portugal
http://istcluster3.ist.utl.pt

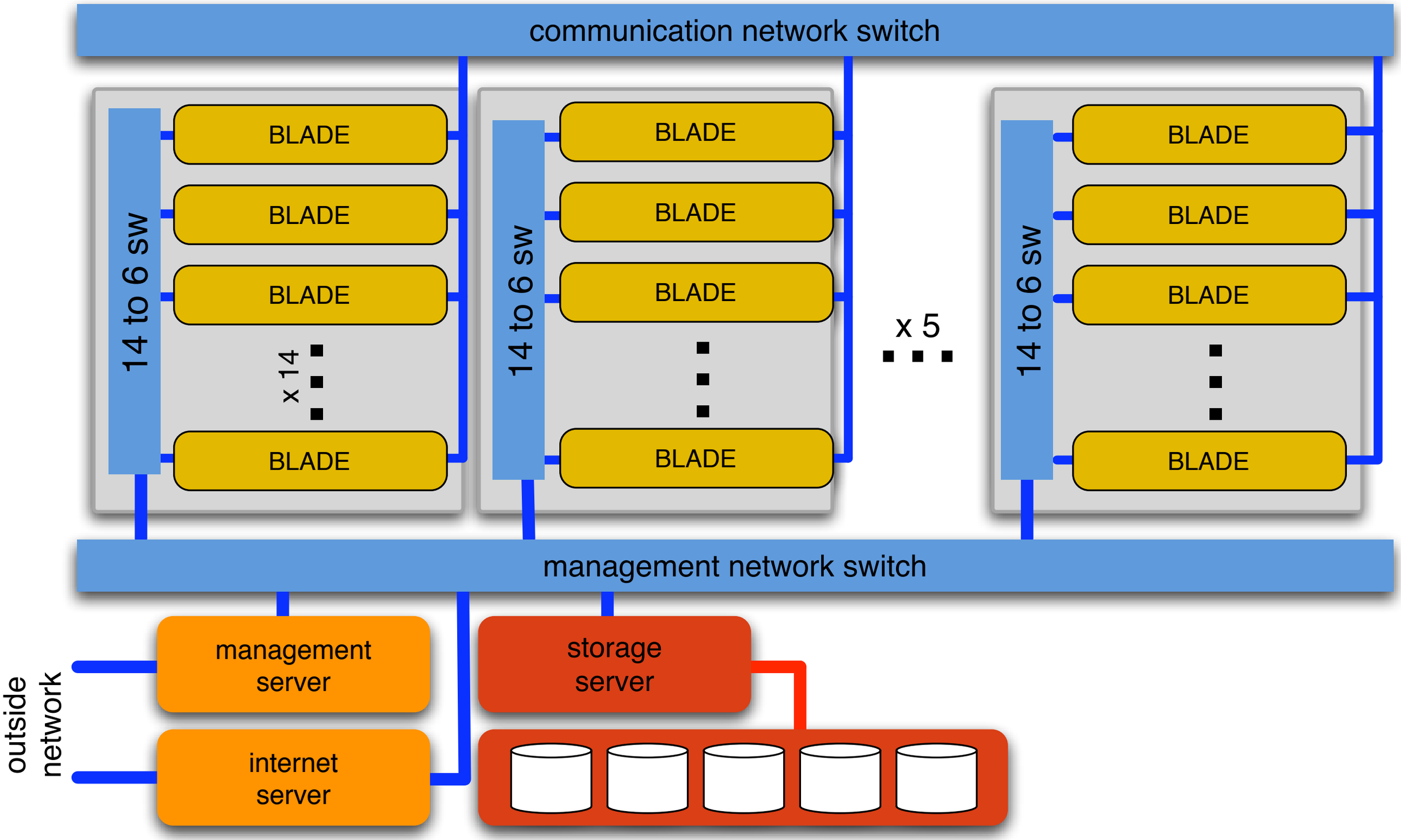


Outline

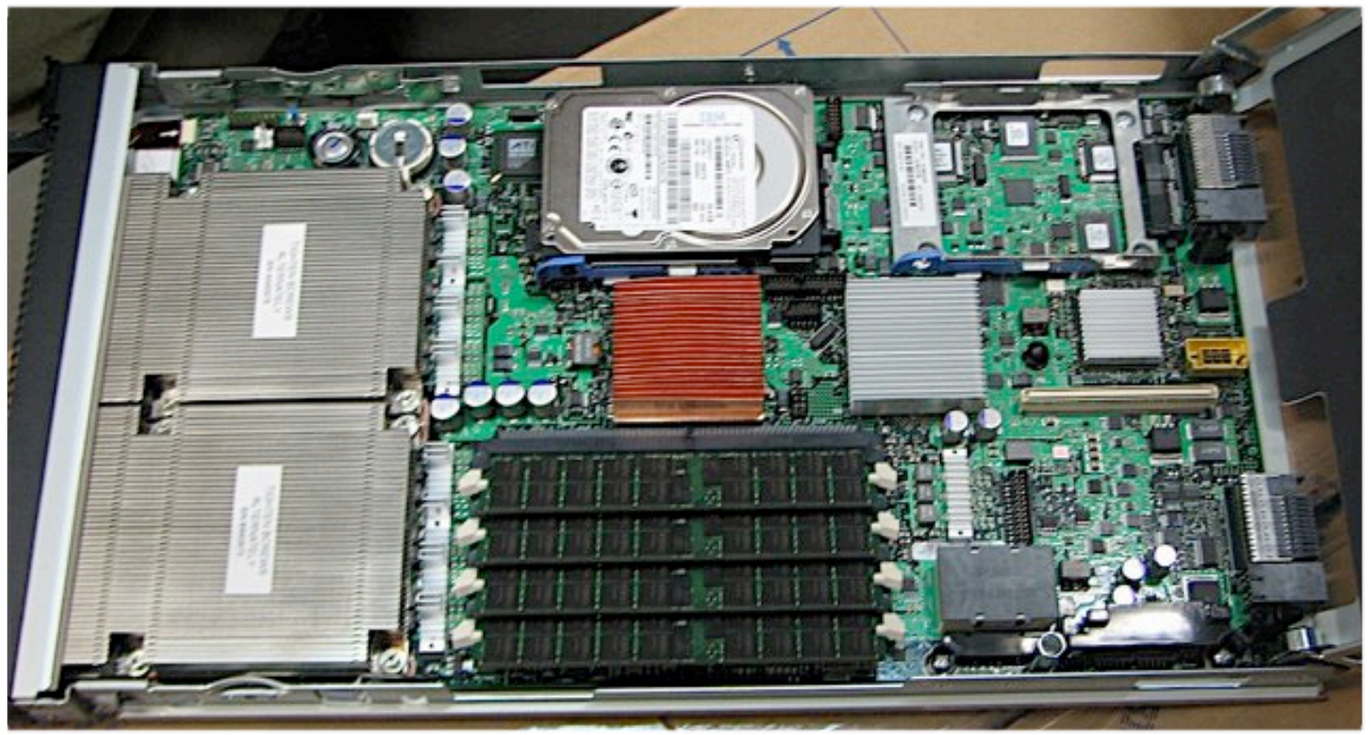
- **Hardware**
  - node and cluster architecture overview
- **Software**
  - operational and cluster level software
- **Benchmarks**
- **Administrative Model**
  - authentication, administrative roles, queuing system
- **Results**
- **Summary**



Schematic overview



Blade JS 21



- 2x dual core dual-core PowerPC 970
  2.5 GHz core clock
  32 KB (data) L1 cache
  64 KB (instructions) L1 cache
  2 MB L2 cache
  FSB 2 x 32bit @1,25 GHz
- 2x 400 MHz DDR2 memory channels
- 8 GB PC2-3200 CL3 ECC DDR2
- 73 GB SAS disk
- 2x gigabit ethernet
- expansion slot for 2nd disk
  or high performance interconnect

Cluster

blade technology

- 5 blade centers with 14 blades each
- roughly double density compared to 1HE servers
- total: 280 cores, 560 GB mem, 5.1 TB disk (local)

network

- communication: gigabit, passthrough
- management and storage: gigabit with 14 to 6 switches per bladecenter (2 level)
- kvm over ethernet

auxiliary

- frontend node: login, ldap, queuing system ...
- storage node: file systems, quotas, nfs gpfs ...
- web server: webpage, ganglia, moab portal ...
- storage DS4700: 8TB, FSC homes, scratch, pio ....

Operational software

clustering

- OS on local disk of nodes
- IBM network installation manager (NIM)
- IBM cluster system manager (CSM)

OS

- AIX 5L
- SUSE Linux Enterprise Server 10 (SLES)

compilers

- IBM xl compiler suite
  xlc V8.0: c / c++,
  xlf V10.1: fortran 77 / 90 / 95
- gnu compiler suite (4.2.3)
  c/c++ java objc obj-c++

Cluster level software

CLUSTER RESOURCES

IBM

MPICH2

queue

- TORQUE resource manager
- MOAB scheduler

parallel

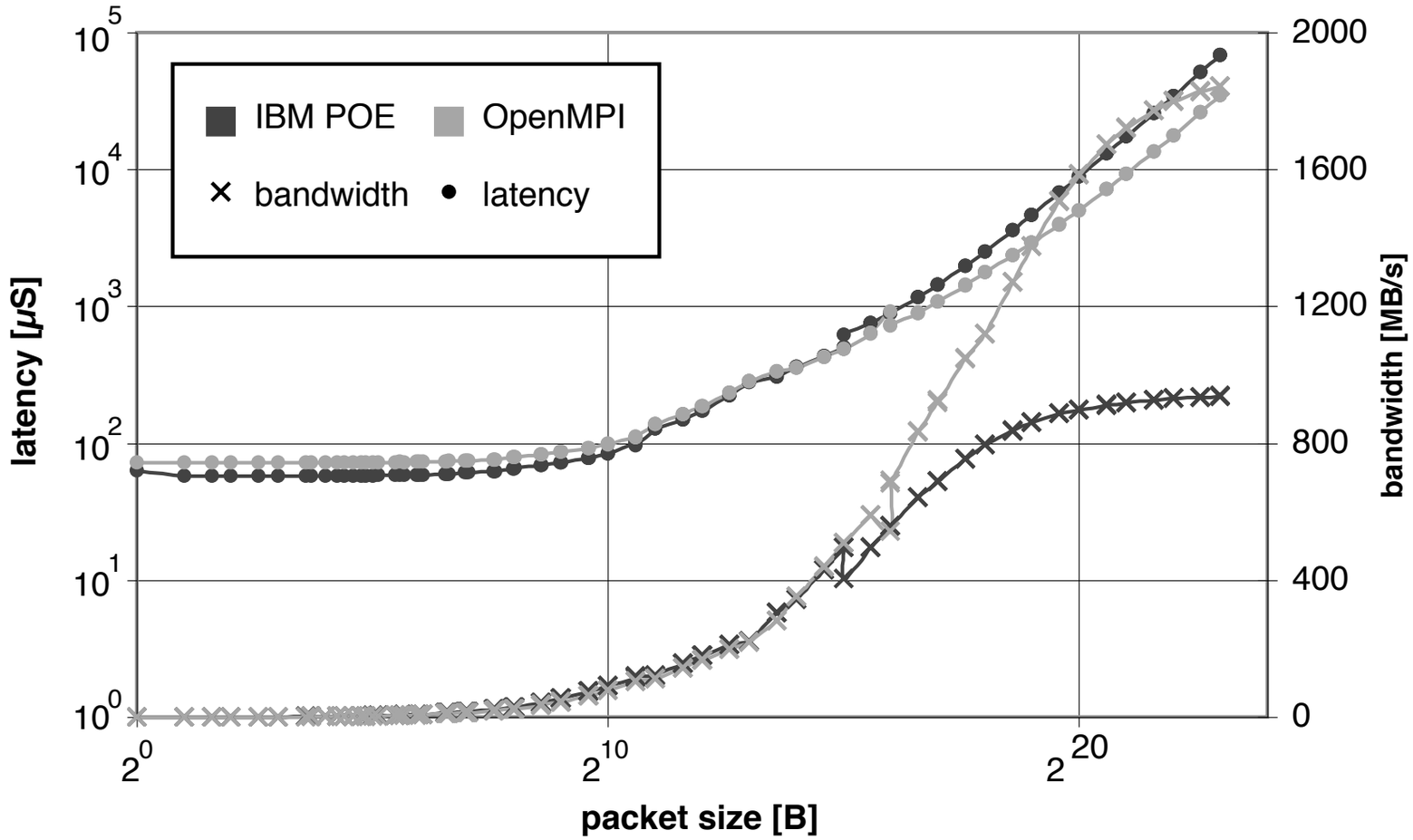
- LAM / MPI 7.1.4
- MPICH2 1.0.6p1
- OpenMPI 1.2
- IBM POE 4.3
- IBM gpfs

monitor

- IBM director
- ganglia
- MOAB access portal

Benchmarks

Network Benchmark (NetPipe)



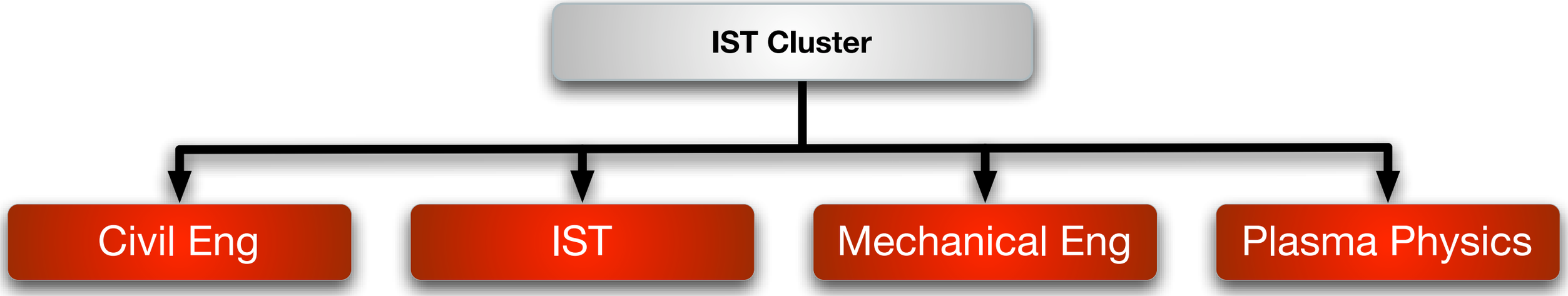
HPL

1.6 Tflops
(15.7 W/Gflops)

Parallel FS

aggregate throughput:
240MB/s

Administrative model



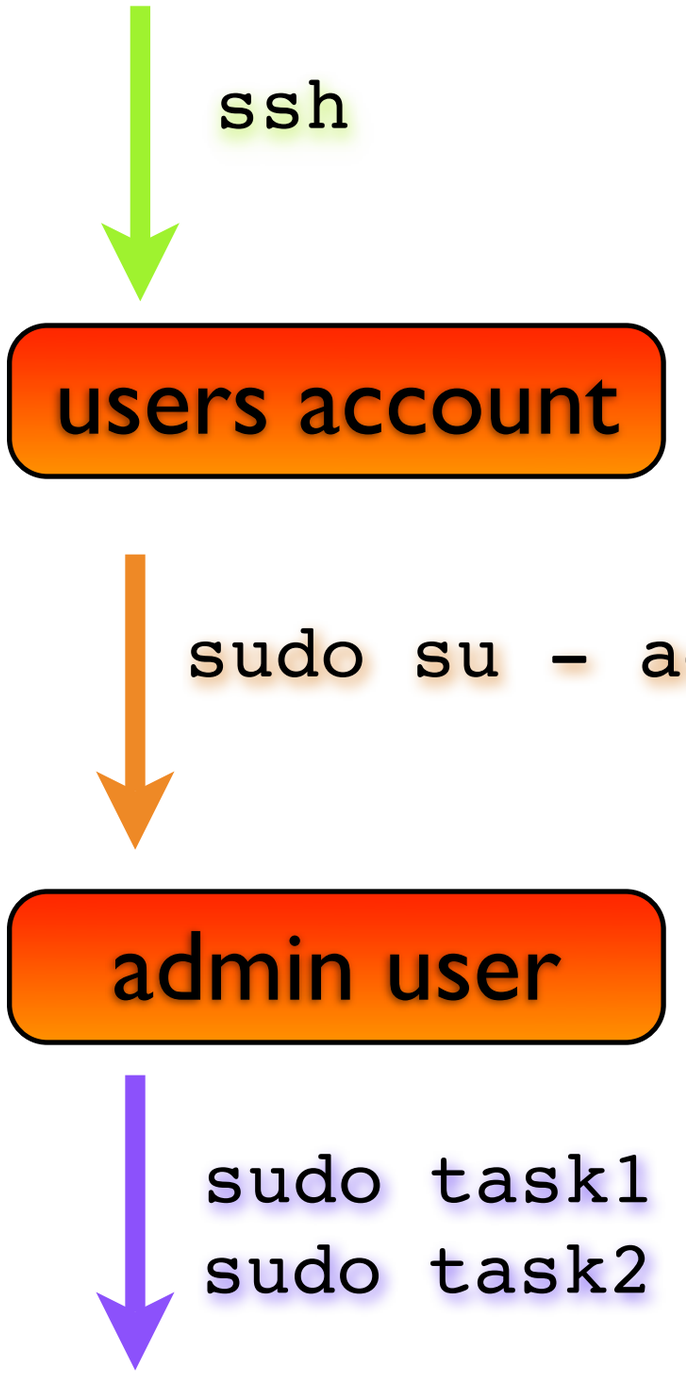
administrative structure

- groups of users are non hierarchical and independent
- special administrative model to address this ecosystem
- a non centralized system administration is required

set of administrative tools

- queuing system
- group management
- software maintenance
- remaining minimal amount of centralized task

Two level authentication



administrator authenticates to his normal *user account* - only privilege: *su - adm* without password

administrator switches to *adm account*

adm account is privileged to do certain, restricted operations as root.

Administrative roles

	group administrator	software administrator
admin home	**support files for group**	**actual files belonging to software package**
available scripts	• create user • delete user • modify user • check user quota • move/delete users files • manage users jobs	• add user to software group • remove user from software group • kill users process (of that software)

Queuing system administration

cluster

creates / manages groups

- node hours
- group manager

groups

creates / manages projects

- node hours
- project leaders

projects

assigns / manages project members

distribution of node hours available

users

selects project upon job submission

consumes node hours

fairness within project is assumed

moab (Cluster Resources)

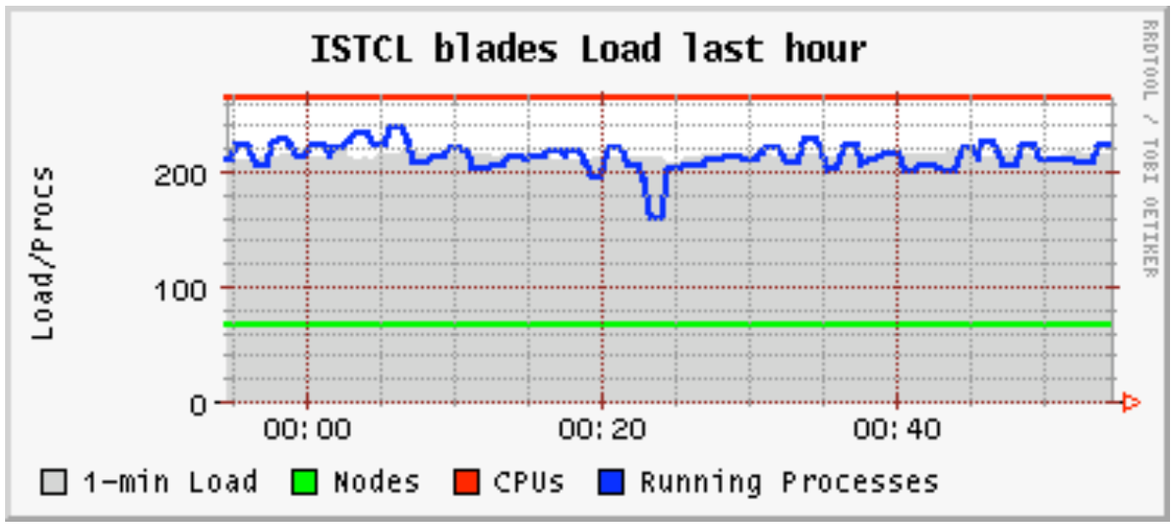
- implementation of project entity
- interfaces: command line tools and files included from moab.cfg

administration level (IST)

- database for storage of entities (groups projects, credits etc)
- set of scripts (sudo) to modify objects in database

Output of IST Cluster

In production since March 2007



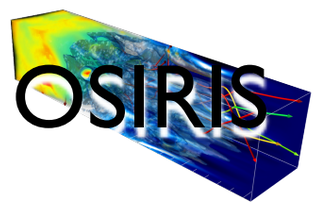
55 **users** active, including users from 4 institutions outside IST

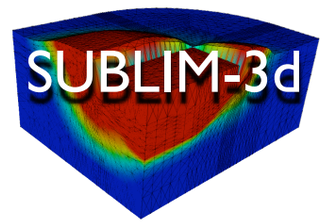
over 2.5 million **cpu hours** consumed

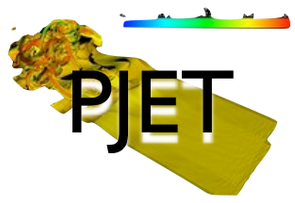
average **system load** 60-80%

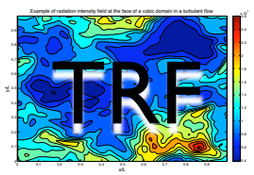
Codes running in production

ANSYS









CD-adapco

Scientific Output

- 7 papers in international refereed journals
- 3 thesis
- 20 contributions to conferences and workshops
- 2 prizes awarded

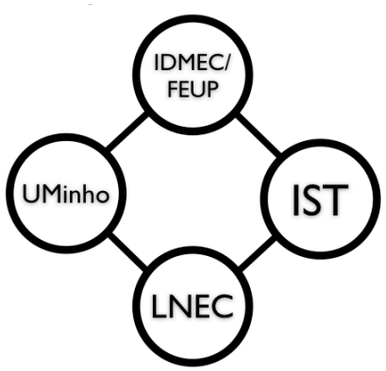
5th Oscar Buneman Award

Luís Gargaté
5th Oscar Buneman Award
Best Animation
20th International Conference on Numerical Simulation of Plasmas
Austin, Texas, 2007

RNCA grid integration

Rede Nacional de Computação Avançada

http://www.rnca.org.pt/



Instituto de Engenharia Mecânica - Pólo FEUP

Instituto Superior Técnico

Laboratório Nacional de Engenharia Civil

Universidade do Minho

Moab based integration of the 4 RNCA nodes

- transparent view of the grid and local node
- integration of ldap domains
- user mapping
- resource mapping
- staging in / out in background
- possibility to integrate with GLOBUS

Summary

- Medium sized cluster successfully deployed
  - cluster hardware
  - system software
  - preproduction since late January 2007
  - production since March 2007
- Integration of the RNCA grid in progress

Next steps

- increase number of cpus to 392
- High speed interconnect
- fully automatic dual boot for nodes, integrated with queuing system
- test of different grid middleware