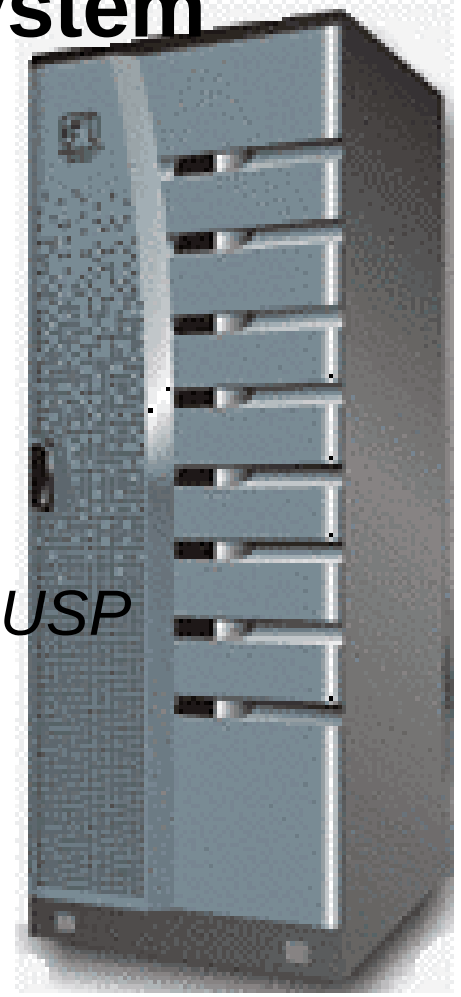
PAD Cluster:
An Open, Modular and Low Cost High Performance Computing System

Volnys Borges Bernal
Sergio Takeo Kofuji
Guilherme Matos Sipahi
Marcio Lobo Netto

Laboratório de Sistemas Integráveis, EPUSP

Alan G. Anderson

Elebra Defesa e Controles Ltda



Agenda

- Main Objectives
- PAD Cluster Environment
- PAD Cluster Architecture
- Communication Libraries
- System Administrator Tools
- Operator Tools
- User Tools
- Development Environment

PAD Cluster

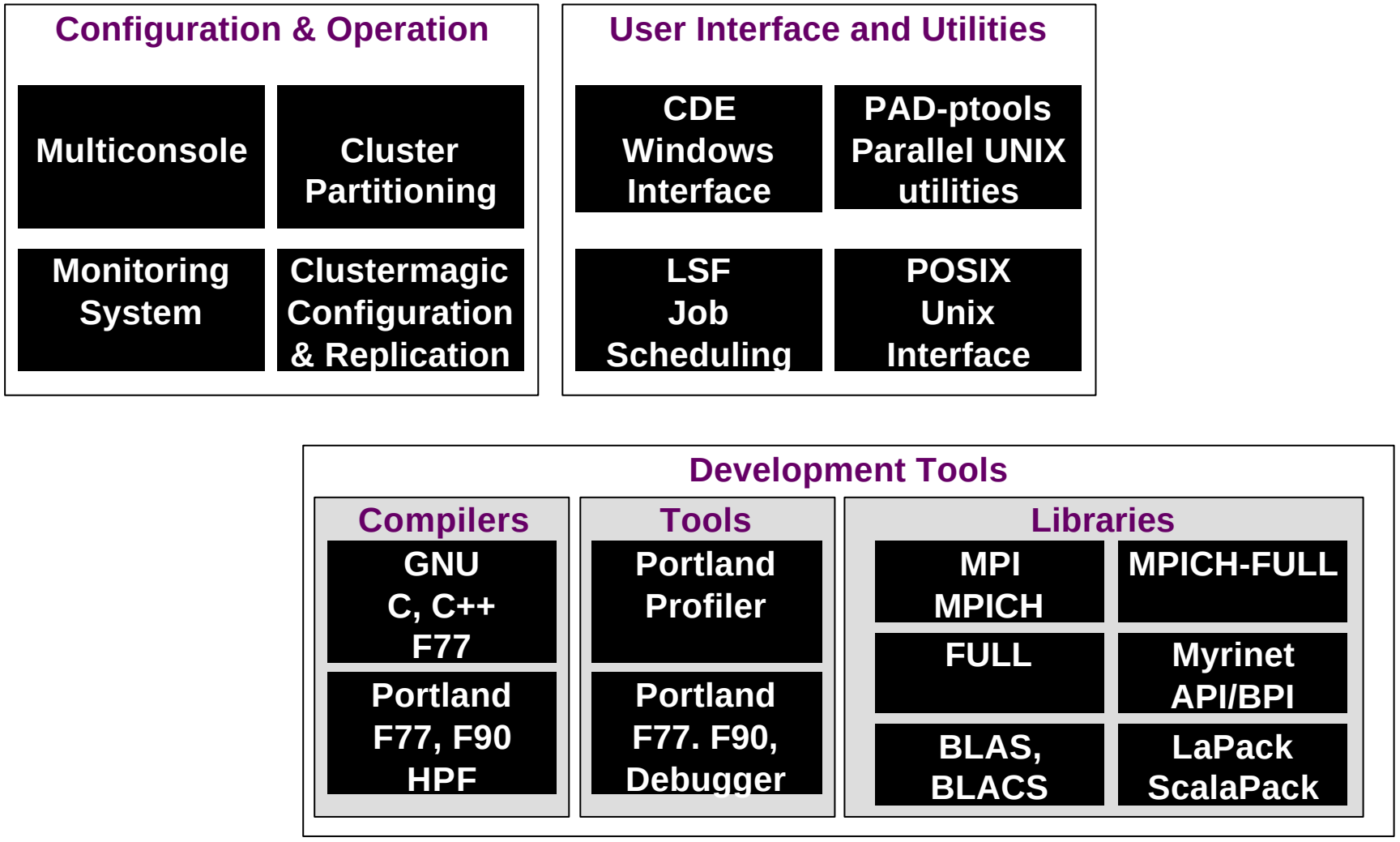
- Main goals
  - Parallel Cluster Based Computing Environment
    - Based on Commodity Components
    - High Performance Communication Medium
    - Development Environment for Fortran77, fortran90 & HPF
    - MPI Interface
    - IEEE POSIX UNIX Interface
    - X-Windows Interface
  - Initial Application:
    - RAMS (*Regional Atmospheric Modeling System*)
- Development: LSI-EPUSP + Elebra, FINEP support

PAD Cluster

- Characteristics
  - Use of High Performance Commodities Components
  - Linux Operating System

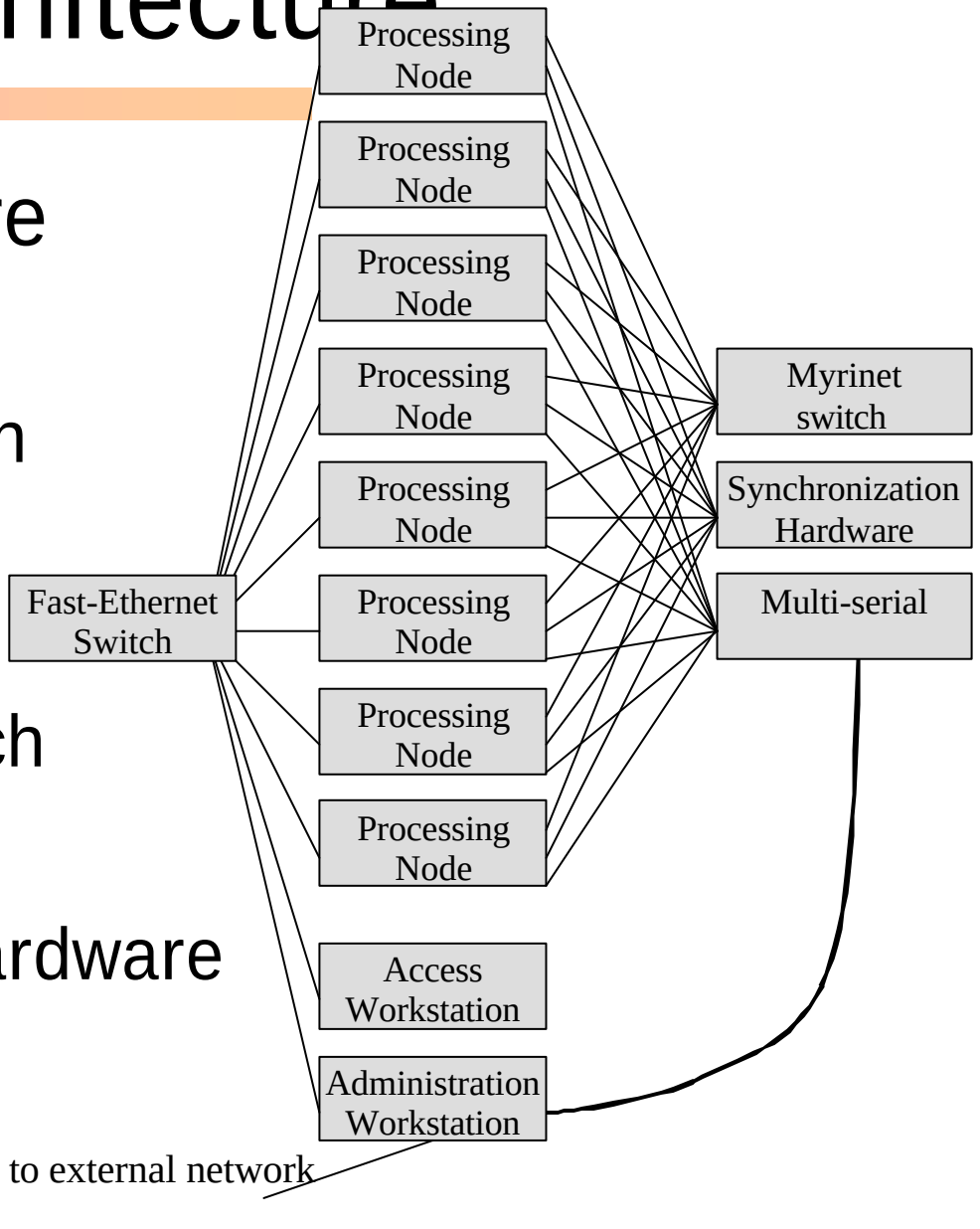
- Important:
  - Integration
    - Hardware components
    - Software subsystems

PAD Cluster Environment



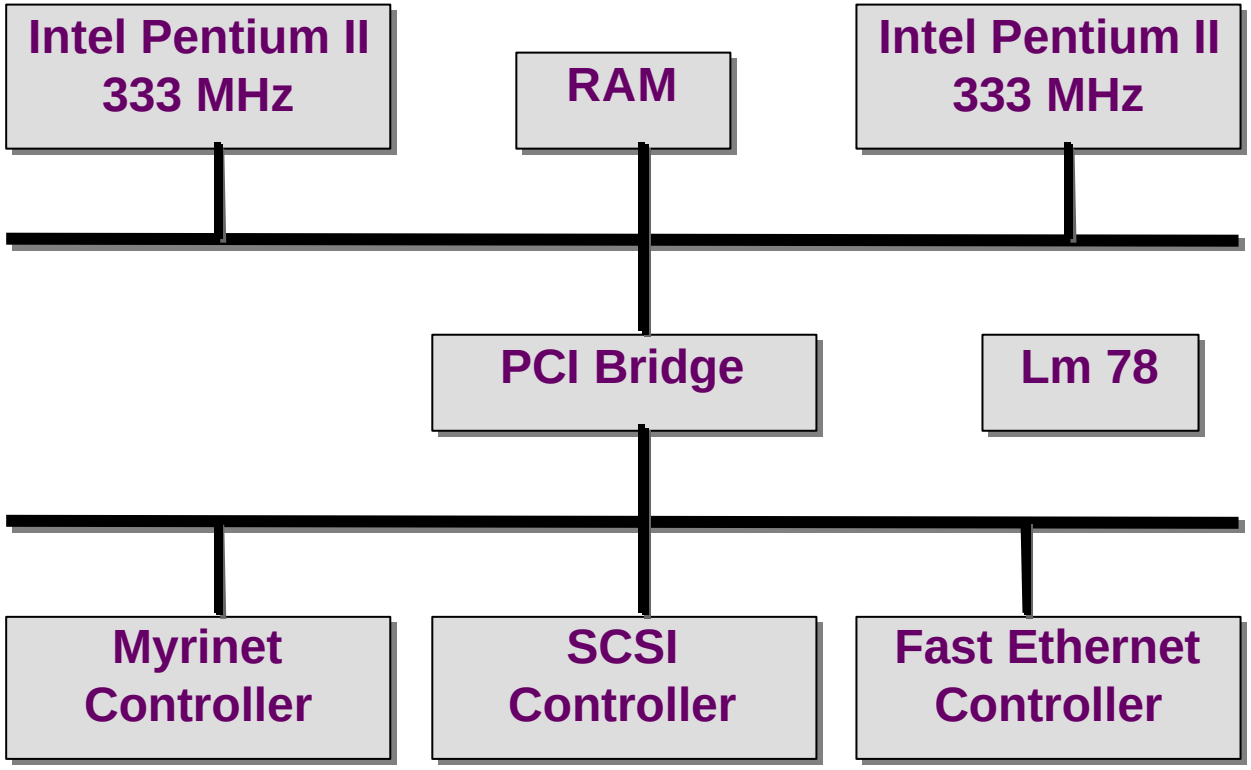
PAD Cluster Architecture

- System Architecture
  - Processing nodes
  - Access Workstation
  - Administration Workstation
  - Fast-ethernet switch
  - Myrinet Switch
  - Synchronization Hardware



PAD Cluster Architecture

- Node Architecture



Communication Infrastructure

- Primary Network
  - Fast-Ethernet
  - General purpose network
    - For traditional network services (NFS, DNS, SNMP, XNTP, ...)
  - Operating System TCP/IP Stack

Communication Infrastructure

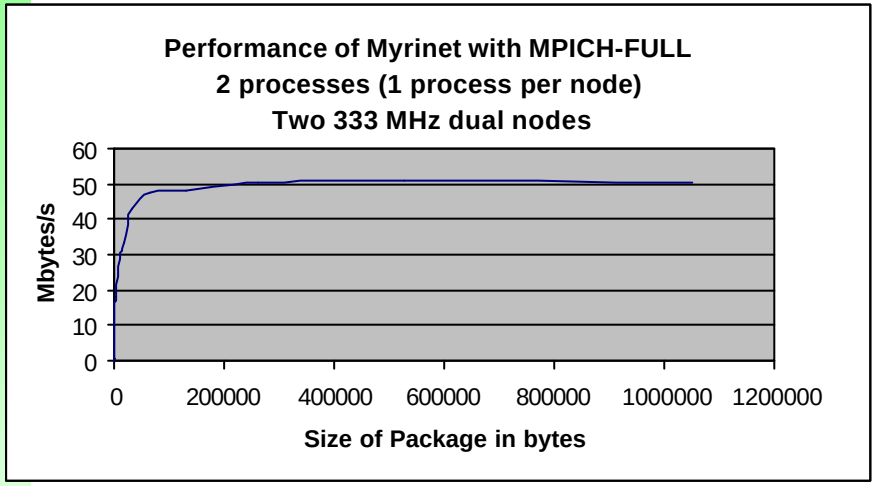
- High Performance Network
  - Myrinet
  - For application data
  - Communication Libraries:
    - MPICH over Operating System TCP/IP Stack
    - FULL user level interface library
    - MPICH-FULL user level interface library

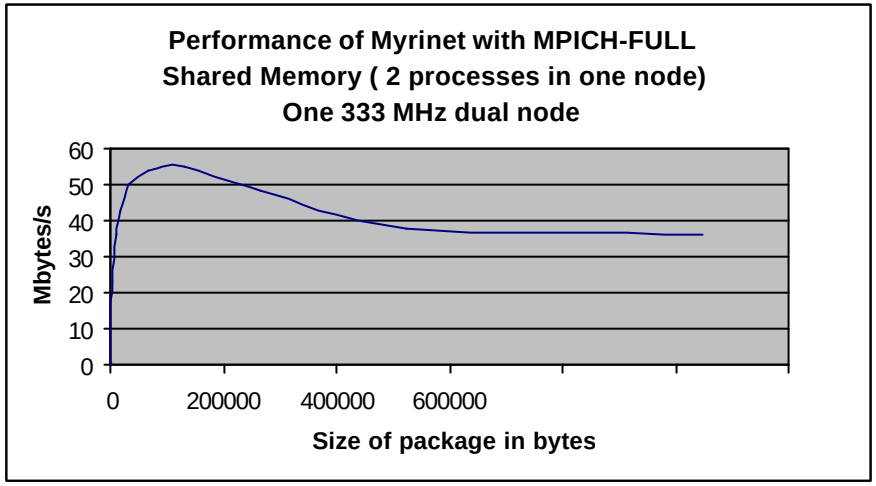
Communication Libraries

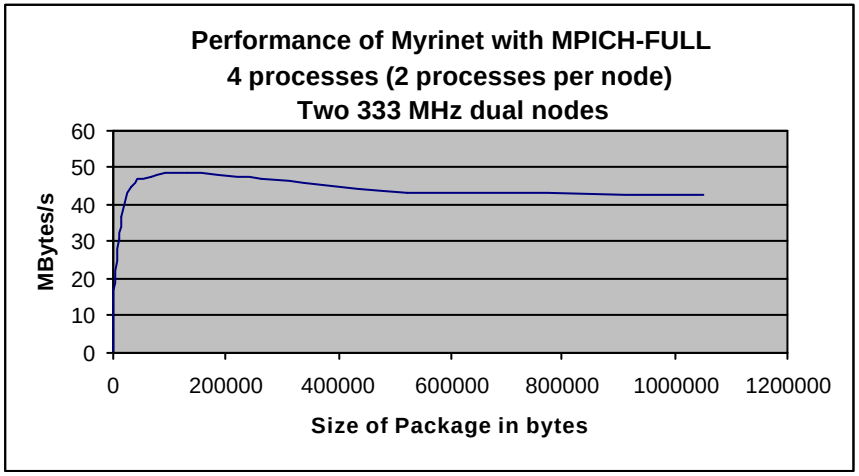
- MPICH Library
  - MPI over TCP/IP stack
- FULL Library
  - User level communication library
  - Developed in LSI-EPUSP in 1998
  - Implementation Based on Cornell's UNET
- MPICH-FULL Library
  - User level communication library
  - Internode communication: MPICH + FULL
  - Intranode communication: MPICH + Shared Memory

Communication Libraries

- MPI-FULL performance

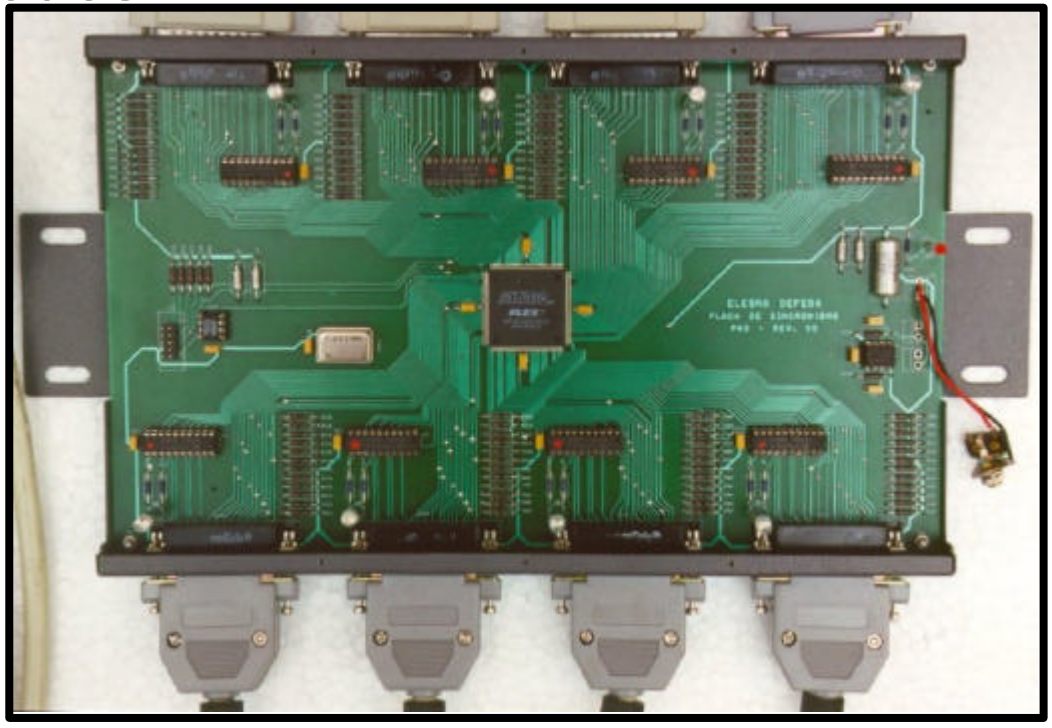






Communication Infrastructure

- Synchronization Hardware
  - Support for collective MPI operations
  - Implemented in FPGA
  - Interfaces for 8 nodes
  - Based on PAPERS
  - Operations
    - barrier
    - broadcast
    - allgather
    - allreduce
  - Global Wall Clock



Communication Infrastructure

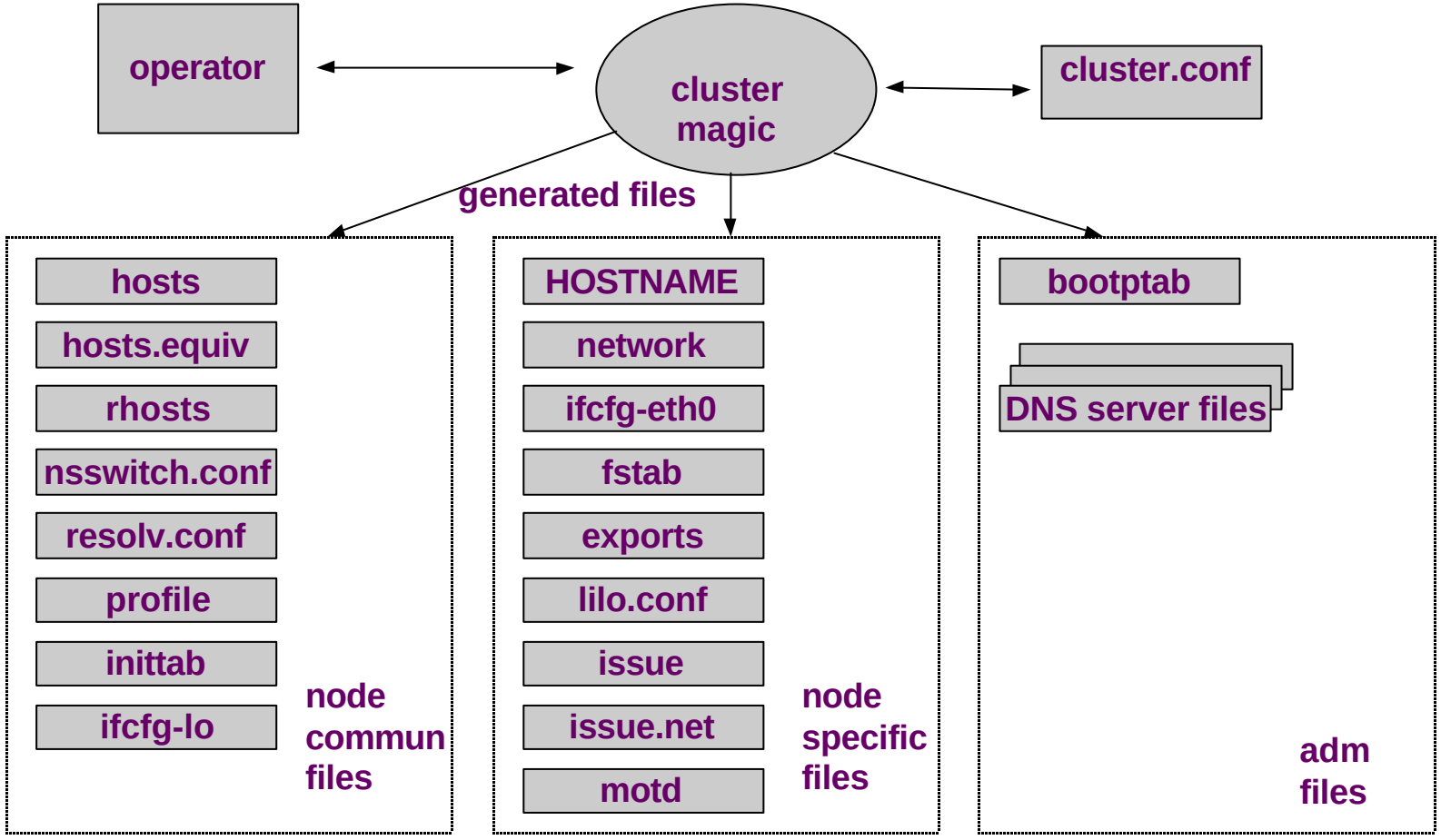
- Serial Lines
  - Connects each node to the administration workstation
  - Allows remote console on the administration workstation

System Administrator Tools

- ClusterMagic
  - Two main funcions:
    - Cluster Configuration
    - Node Replication
  - Advantages
    - Easy configuration / reconfiguration
    - Assure uniformity
    - Fast node replication

System Administrator Tools

- Cluster Magic: Cluster Configuration



System Administrator Tools

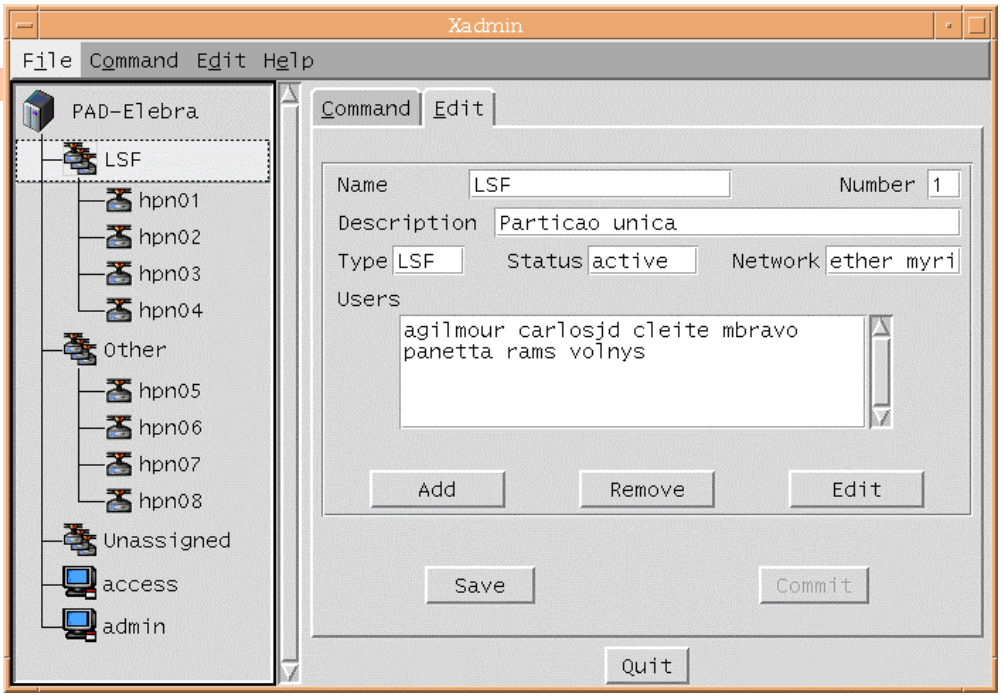
- Cluster Magic: Node Replication
  - Node installation based on the replication of a "Womb Node"
  - ClusterMagic replication diskette:
    - boots a small Linux System
    - disk partitioning
    - womb image copying
    - configuration files instalation
    - Boot sector initialization
  - Automatic process
  - Takes about 12 minutes

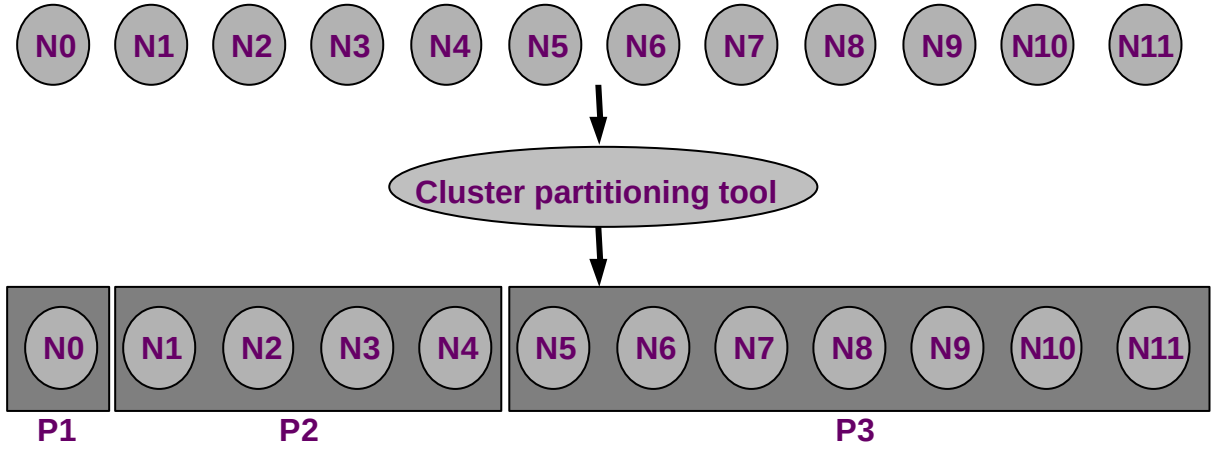
Operator Tools

- Xadmin
  - Cluster Partitioning
  - Remote Commands
- Multiconsole
  - Node console access
- Job Scheduling
  - Job submission
  - LSF integrated with Cluster Partitioning
- Cluster Monitoring

Operator Tools

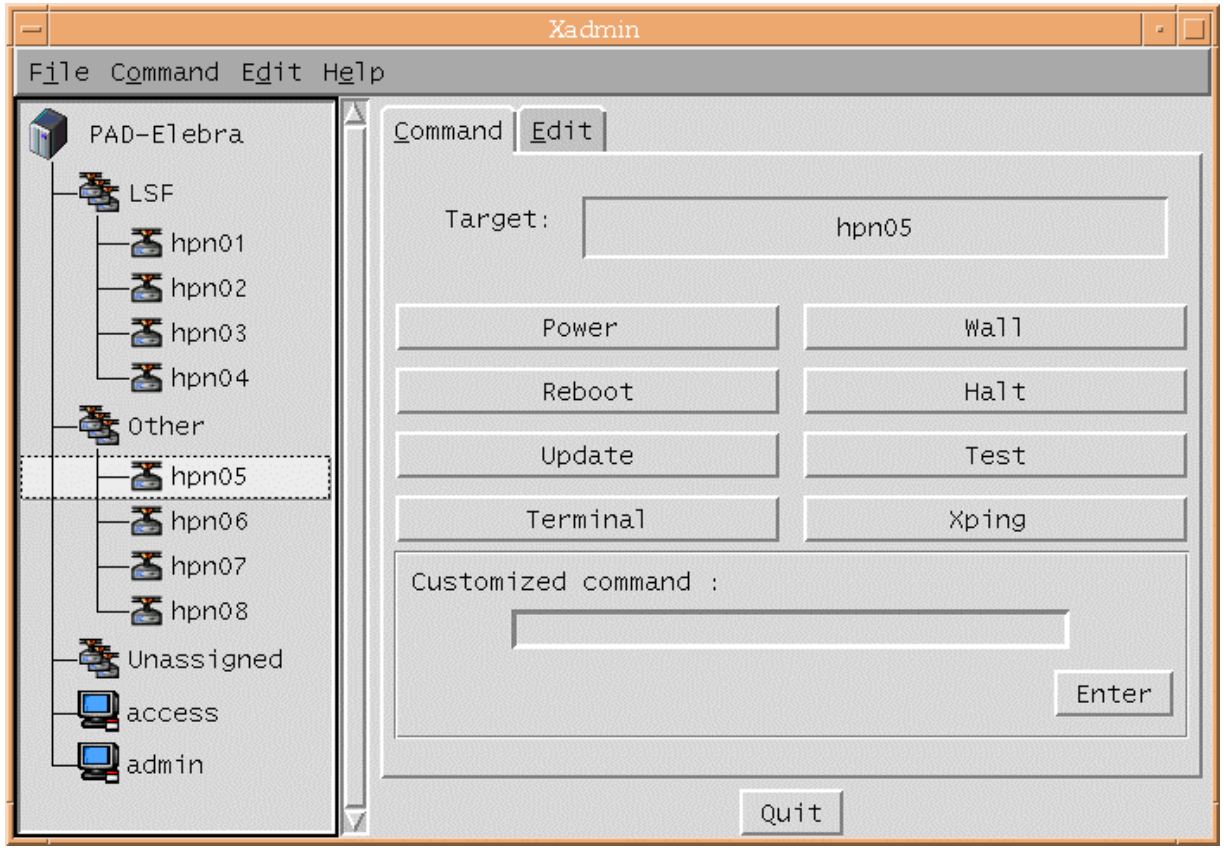
- Xadmin
  - Node partitioning





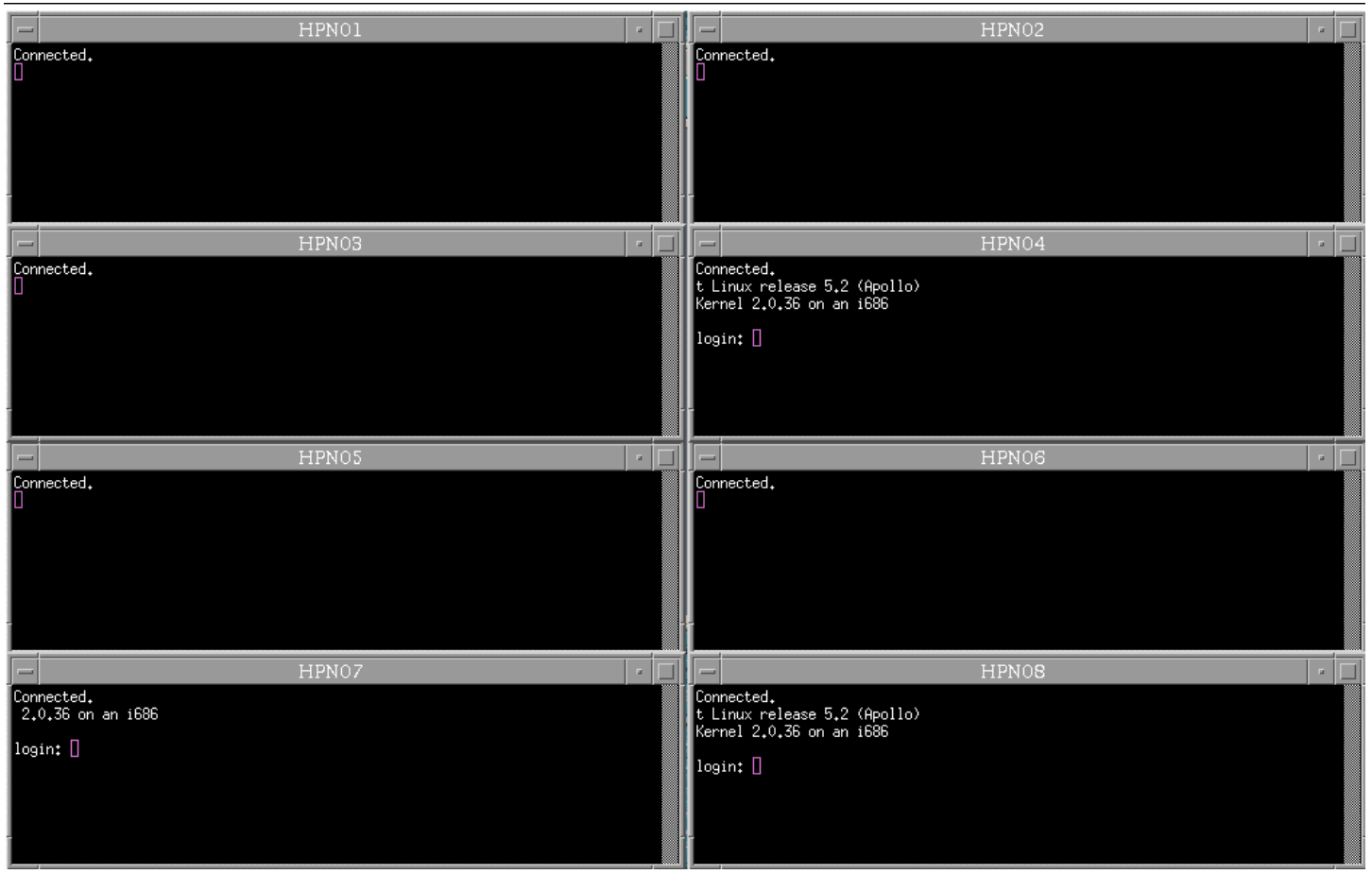
Operator Tools

- Xadmin
  - Remote Commands



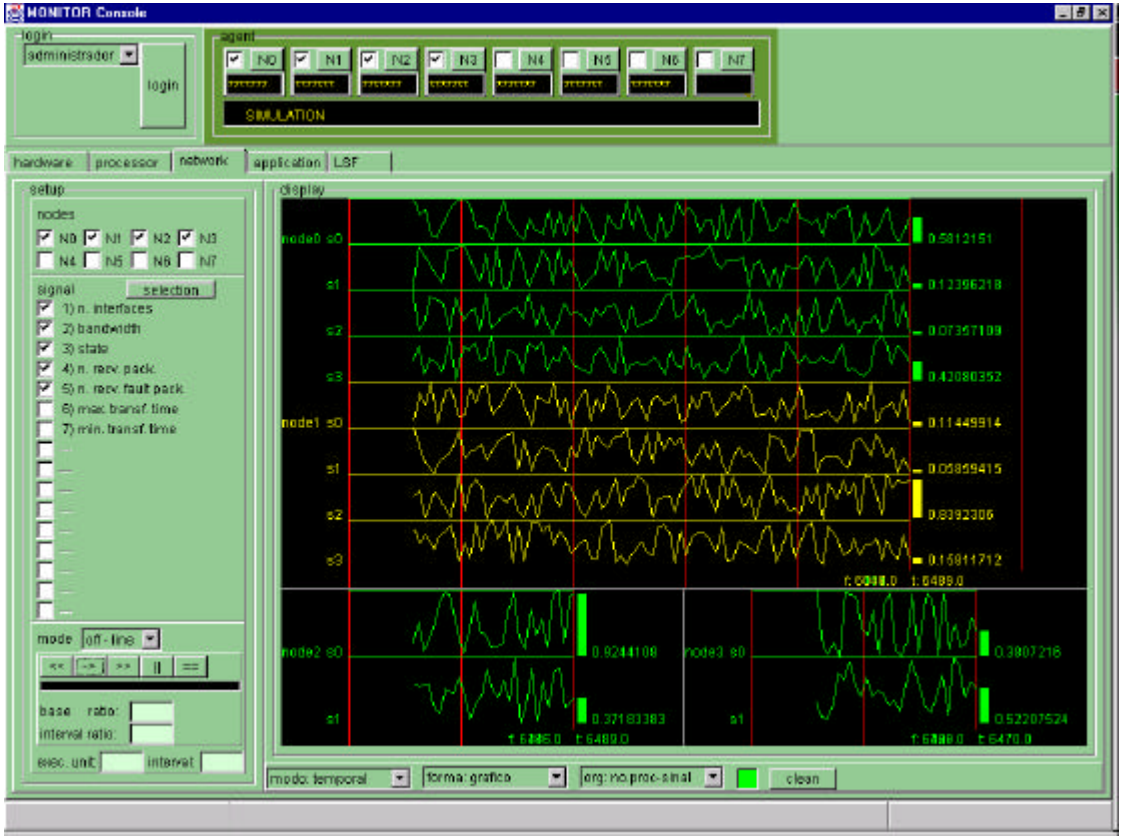
Operator Tools

- Multiconsole



Operator Tools

- Cluster Monitoring
  - Java + SNMP agents



User Tools

- PAD-ptools
  - Parallel versions of UNIX utilities
  - pcp, pls, pcat, ...
  - Integratded with cluster partitioning
- LSF
  - Job submission and control
- mpirun
  - MPICH, MPI-FULL

Development Environment

- Portland
  - Fortran77
  - Fortran90
  - HPF
  - Profiler
  - Debugger
- Libraries
  - BLAS, BLACS, LaPack, ScaLaPack
- TotalView debbuger
- VAMPIR profiler

Conclusions

- Complete product system:
  - Elebra Vortix Cluster (PAD Cluster)
    - www.elebra.com.br/aero
- Several Developments:
  - Hardware
    - Collective operations, Synchronization and Global Clock
  - Software
    - Communication Libraries
    - Cluster Tools
    - Communication Drivers

Future Works

- University of São Paulo + Purdue University + University of Pittsburg
  - Hardware for collective operations and synchronization with PCI 64 bits Interface
- University of São Paulo + ICS-FOTH (Greece)
  - ATM Like Switch on 2.4 Gbps/s
- University of São Paulo
  - New cluster administration, management and secure tools
  - High Availability Data Base applications

Acknowledgments

- FINEP
- LSI-EPUSP Development Team
- Elebra Development Team