



**gelato**  
CONF

Gelato Federation Meeting, May 24-26, Urbana-Champaign-IL

## **Heterogeneous mid-size clusters: a new research focus for the cluster track**

*Prof. Dr. César De Rose  
PUCRS - Porto Alegre, Brazil*



**gelato**  
CONF

## **Summary**

- **Mission/Objectives**
- **Motivation**
- **Areas of interest**
- **Current state**
- **Open issues**
- **Next Steps**

Gelato Federation Meeting - May 24-26, Urbana-Champaign, IL - 2/12



## Mission/Objectives

- The Heterogeneous mid-size cluster (HMC) focus group comprises people and organizations drawn from the Gelato federation interested in heterogeneous clusters that include IPF nodes,
- Mission
  - Find ways to better exploit the potential of IPF nodes in mid-size heterogeneous clusters
- Objectives
  - Identify and solve open issues
  - Develop open source tools
  - Share information, tools and resources

Gelato Federation Meeting - May 24-26, Urbana-Champaign, IL - 3/12



## Motivation

- Cluster track initially with two research focus
  - Cluster scalability
  - Cluster performance
- Since difference was not clear research focuses were merged
  - Large scale homogeneous IPF clusters
    - Very specific scalability problems
    - Small group of high performance research labs
- University members demonstrated interest in widening the scope of the cluster track
  - Idea is to include some IPF nodes in actual clusters to improve their performance
- Result is an heterogeneous cluster
  - Test bed to identify the potential of IPF nodes in PP
  - Increase the number of clusters with IPF nodes (lower cost)

Gelato Federation Meeting - May 24-26, Urbana-Champaign, IL - 4/12



## Areas of interest

- Installation and Maintenance
  - Automatic tools
- Node interoperability
  - Network compatibility
  - Resource management
    - Quantitative → qualitative
  - Cross-compilation
    - Automatic binary generation
- Performance
  - Performance evaluation
  - Monitoring
  - Load balancing
    - Static x dynamic
    - System x application level
  - Application development

Gelato Federation Meeting - May 24-26, Urbana-Champaign, IL - 5/12



## Current state

- Test bed running for 6 months
- Simple qualitative resource management in production
- Automatic cross compilation for the gcc compiler
- Initial performance comparison among x86 and Itanium 2 nodes (benchmarks)
- Static adaptive load balancing for SPMD applications being tested

Gelato Federation Meeting - May 24-26, Urbana-Champaign, IL - 6/12



## Current state: test bed

- 26 nodes cluster (50 processors)
  - 16 HP E-60 servers / Dual Pentium III 550 MHz (2-way SMP) 256M RAM
  - 8 HP E-800 servers / Dual Pentium III 1GHz (2-way SMP) 256M RAM
  - 2 Itanium 2 Workstations / 900 MHz 1G Ram
- Running Linux (Debian)
- *Myrinet* as primary network and a switched *Fast-Ethernet* as secondary network
- Host machine
  - Compilation
  - Resource Allocation

Gelato Federation Meeting - May 24-26, Urbana-Champaign, IL - 7/12



## Amazônia Cluster



- Software stack
  - Crono resource manager (in-house solution)
    - Design to be simple
    - Space and time sharing
    - Few configuration scripts
    - Open Source
  - Heterogeneous version of Mpich
    - Fast-Ethernet support

Gelato Federation Meeting - May 24-26, Urbana-Champaign, IL - 8/12



## Current state: resource management

- Crono is being improved to support IPF nodes
  - Node types for qualitative allocation
  - Automatic cross compilation if needed
    - If allocated partition has IPF nodes (host is 32 bit)
    - Compilation generates several binaries that are loaded by demand
  - Option for static load balancing when loading processes to nodes
- Beta version already available
  - [www.cpad.pucrs.br/crono](http://www.cpad.pucrs.br/crono)

Gelato Federation Meeting - May 24-26, Urbana-Champaign, IL - 9/12



## Current state: load balancing

- Static adaptive load balancing for SPMD applications being tested
- Done at resource manager level
- Default performance values are calculated for each node type (number of processes)
- Default values are used to dispatch MPI processes in each node type
- Idle nodes are used to improve values for specific applications
- System keeps track of the best combinations for each application in each set of nodes to use in the next run

Gelato Federation Meeting - May 24-26, Urbana-Champaign, IL - 10/12



## Open issues

- Automatic cross compilation with Intel compiler
  - Intel Linux compiler not up to date with x86 version
- Myrinet node interoperability
  - No support for x86 and IPF communication
  - Myricom has no intention to include this feature due to performance loss ☹
- More flexible load balancing
  - Ex: for MPMD applications

Gelato Federation Meeting - May 24-26, Urbana-Champaign, IL - 11/12



## Next Steps

- Bring more members to HMC
  - PUCRS (focus team lead)
  - UPRM
  - ...
- Define members expertise and interests
- Include new areas of interest based on members feedback
- Define members commitment to open issues
- Create own E-Mail distribution list
- Create research focus homepage
- Replicate test system in members institutions

Gelato Federation Meeting - May 24-26, Urbana-Champaign, IL - 12/12