

# Paracel HPCs for the GCP

## SP4 - Tools and Databases

### Task #25 Progress Report

Anthony Collins  
CIP

GCP/SP4 meeting in Brisbane, AUS, 2004

# HPC Task Goals

- Evaluate suppliers for global support & \$\$\$
- CIP, ICRISAT, IRRI to proceed to purchase
- ( Coordination with ILRI's CIDA project )
- Delivery and installation of 3 (+ILRI) systems
- Basic system training by supplier
- Configure global cluster/grid facilities
- Design interface for remote access
- Documentation & Training for GCP users

## Principal expectations for Cluster/Grid systems

- Run proprietary optimized applications
  - Decisive issue: vendor's BLAST => PARACEL !
- Develop applications structured for cluster systems using the MPI compiler (for example "R")
- Develop applications which decompose into independent executable modules for cluster/grid queuing (for example Structure )
  - Decisive issue: LSF queue license => PARACEL !

( **BUT any other ordinary application will only run on any ONE node of the cluster, and so CANNOT achieve cluster/grid performance gains !** )

# Summary of Objectives

## Achieved to date

- Select common HPC cluster system vendor
  - Decisive issue: global support capabilities => PARACEL !
- Pre-training in Pasadena for 9 X CGIAR attendees
- Install CIP, ICRISAT and IRRI systems for GCP,  
( ILRI installation by early October )
- Initial benchmark testing (later presentation)
- CIP developing a Web services wrapper:  
validation of users via CGIAR Active Directory

# Summary of Objectives

## Underway

- Configure CGIAR Web services wrapper, creating Global Group for CGIAR & GCP Bioinformatics users: simple for users and HPC administrators !
- [hpc.cip.cgiar.org](http://hpc.cip.cgiar.org)
- Paracel testing of CGIAR cluster/grid with CIP & IRRI, then ICRISAT plus ILRI systems

## Inauguration of Bioinformatics users group

- *With Paracel, test Cluster/Grid with CIP & IRRI*
- *Incorporate ICRISAT & ILRI in Cluster/Grid*
- *Team Review of GCP Web access interface*
- *Team development of User access policies*
- *Initiate CGIAR Bioinformatics Group for GCP*
- *Documentation and Training for CGIAR users*
- *Review performance monitoring and benefits*

## Advanced Research Networks – Internet2

\* optimize cluster/grid performance \*

- *IRRI online to APAN Internet2 since 1999*
- *CIP to join RAP Internet2 in Q4 2004 with ICT/KM ARN project support*
- *ILRI online via TAMU, but very slow ... upgrade for African continent pending ...*

# Paracel HPC for the GCP

- bioinformatics uses at CIP
- performance benchmarks

Reinhard Simon  
CIP

GCP/SP4 meeting in Brisbane, AUS, 2004

- Loosely coupled algorithms (no dependencies on intermediate results)
  - Scriptable applications
    - “Embarassingly parallel”\*\* applications
  - Segmentable algorithms/data sets, e.g.
    - Permutations, e.g. for empirical thresholds in statistical tests
    - Markov-Chains
    - Artificial intelligence
- Tightly coupled
  - *Ab initio* molecular modeling

\* modified after: Ahmar A (2004): Grid computing: A practical guide to Technology and Applications. Charles River Media

\*\* ibd.

# Types of uses - 1

- Loosely coupled algorithms
  - Scriptable applications
    - “Embarassingly parallel” applications
      - searches in huge databases, e.g. **BLAST**
      - Batch queues for parameter scans:  
e.g. **Structure (in progress using LSF)**

# Application: *BLAST*

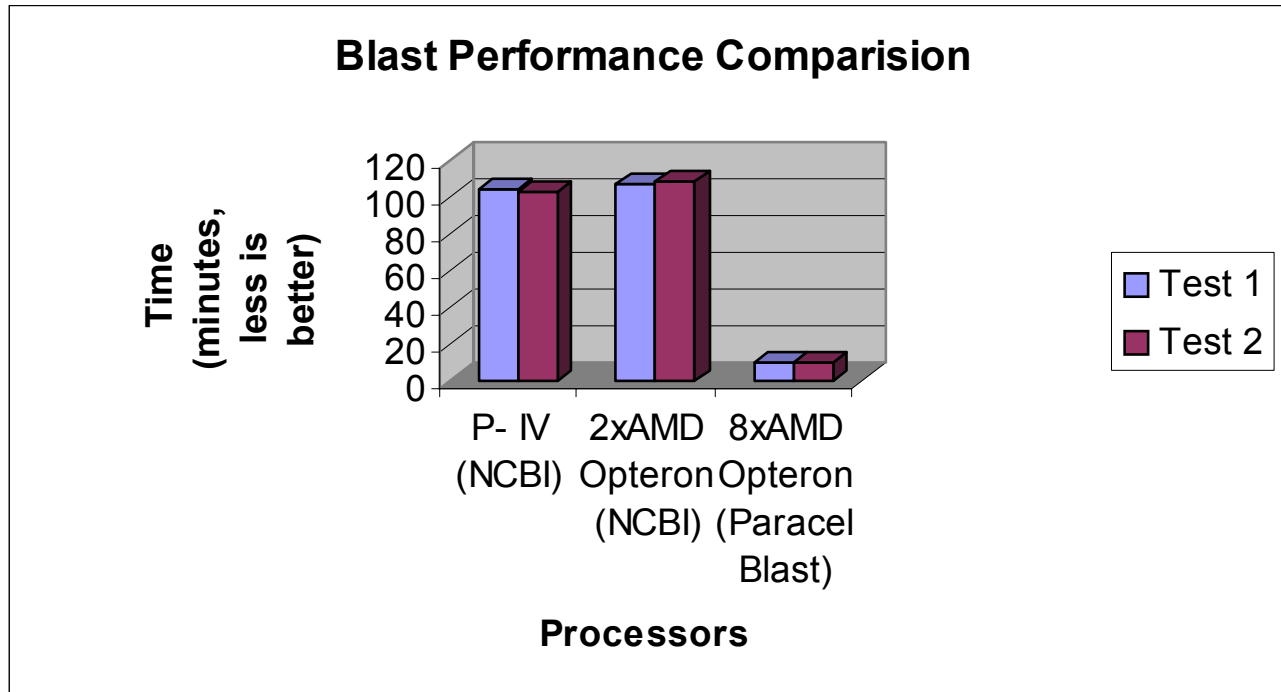
- Paracel *BLAST* performance

Dataset: Potato

Type	DNA sequences
Number of Sequences	32553
Total Length	26242629
Max Sequence Length	4968

Arabidopsis

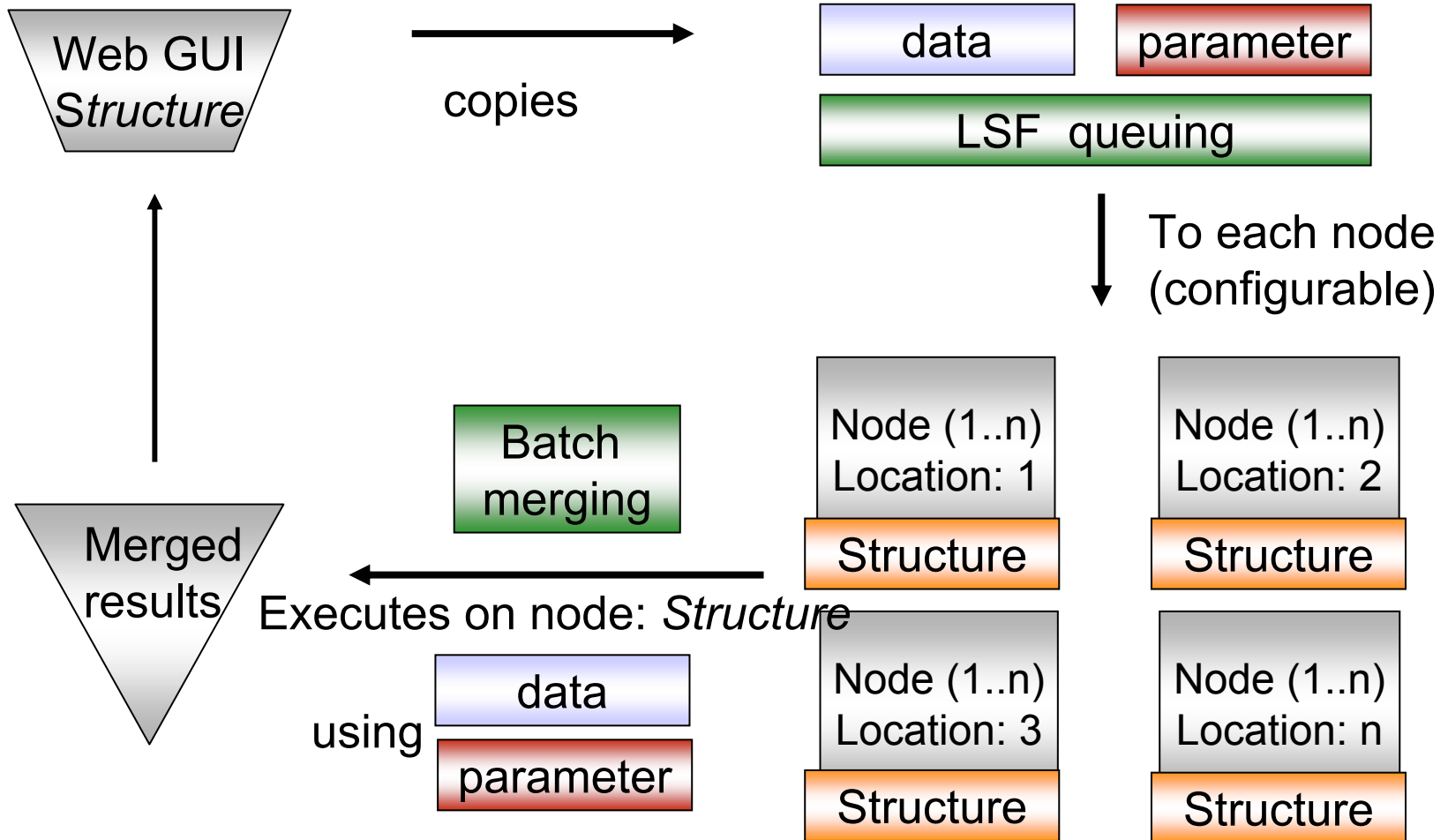
Type	DNA sequences
Number of Sequences	45683
Total Length	55938335
Max Sequence Length	16011



# Application: *Structure*

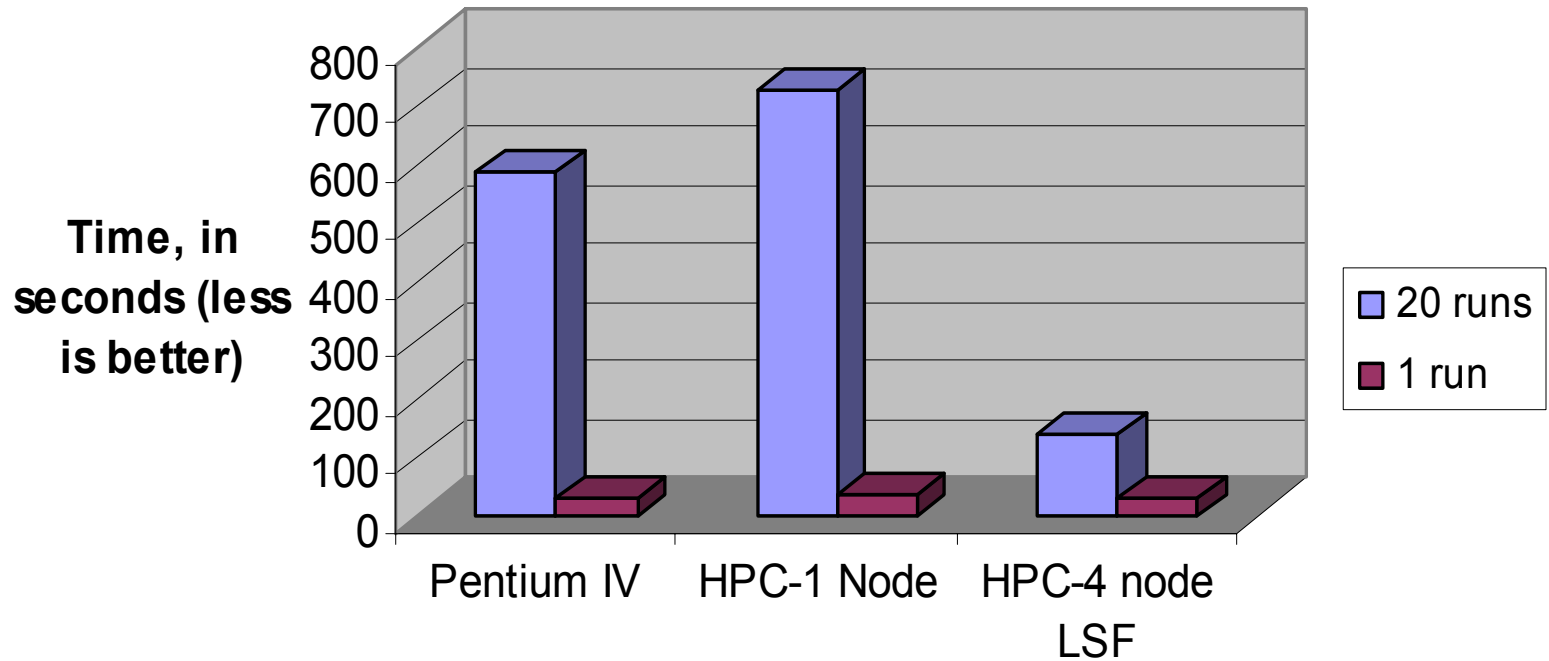
- Under development:**

parallel parameter scan architecture for **HPC grid**



# Application: *Structure*

## Structure execution Time Comparison



# Types of uses - 2

- Loosely coupled algorithms
  - Segmentable algorithms/data sets
    - Permutations, e.g. for empirical thresholds
      - in standard statistical applications, e.g. **ANOVA using R** (underway)
      - For scriptable applications: validation of genetic maps (**MapMaker: under development**) and QTL (planned)

# Application testing: *R*

R:

- is an open source statistical package
- has several parallel execution environments

Status:

- 1<sup>st</sup> parallel version on Paracel HPC  
was configured and installed at CIP
- configuration protocol shared
- currently limited to 2 processing nodes

# Types of uses - 3

## Next round?

- Tightly coupled algorithms
  - Segmentable algorithms/data sets
    - Markov Chains
      - Markov-Chain Monte-Carlo (MCMC), e.g for
        - » Large-scale Marker-Trait association (beyond **structure**)
        - » QTL validation
      - Hidden Markov Chains (HMM)
        - » DNA/Protein sequence alignments (pairwise, multiple)
        - » Sequence profiles for gene family characterization
        - » Phylogenies
    - Artificial intelligence
      - Genetic algorithms (GA): e.g. QTL discovery
      - Ant Colony optimization (ACO): e.g. genetic maps

# Research Informatics Team

- Reinhard Simon (Head, molecular bioinformatics)
- Felipe Mendiburu (Statistician)
- Edwin Rojas (Senior systems analyst)
- Henry Juarez (GIS assistant, modeling)
- Miguel Blancas (Statistician, Informatics)
- Enver Tarazona (Statistician, Informatics)
- Luis Avila (Informatics, HPC)
- Magna Schmitt (Informatics)
- Sara Villanueva (Informatics)
- Kumari Gurusamy (GIS specialist)
- Darwin Gomez (GIS consultant)
- Kelly Theisen (GIS consultant)
- Sara Moreno (GIS trainee)
- Melina Perez (IT consultant)
- Christian Solis (biologist, bioinformatics trainee)

# Information Technology Team

- Anthony Collins (Head, molecular bioinformatics)
- Roberto del Villar (Network Manager)
- Dante Palacios (HPC Systems Support)
- Peter Valdivieso (LINUX Systems Support)
- Luis Avila (Informatics, HPC Administrator)

# Paracel HPC for the GCP

- bioinformatics uses at IRRI

# IRRI uses – Overview

- Biomirror (approx 160Gb and being updated weekly)
- BioPerl
- EMBOSS
- Java SDK (non-64 bit, Sun now only supports Itanium2)
- PostgreSQL
- Tomcat
- R (parallel version in collaboration with CIP)
- BGI
- TIGR
- pseudomonas
- Genesis Server