NUI MAYNOOTH
Ollscoil na hÉireann Má Nuad

embarkinitiative
Investing in People and Ideas

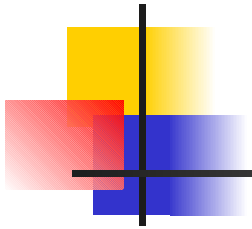
NDP
NATIONAL DEVELOPMENT PLAN

A High-throughput Bioinformatics Distributed Computing Platform

Presented by: Andrew J. Page

Thomas Keane[1], Andrew Page[2], James McInerney[1], Thomas Naughton[2]

[1]Department of Biology & [2]Department of Computer Science,
National University of Ireland, Maynooth,
Ireland



Outline

Introduce Parallel Computing

Distributed What?

Some Examples

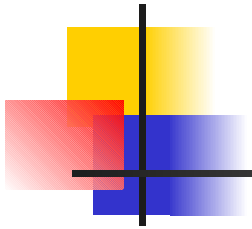
Our Distributed Computing Platform

DPRml

DSEARCH

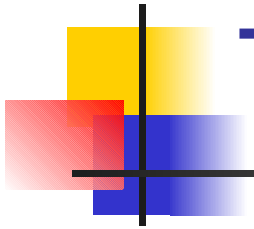
Speedup

Conclusions & Future Work



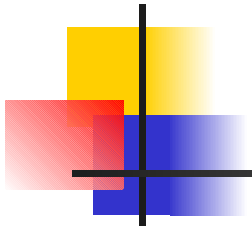
Introduction

- What is parallel computing?
  - Using multiple machines (CPU's) to solve computationally challenging problems
- Why parallel computing?
  - Many important scientific problems cannot be solved using a single machine
  - Multiple machines allows us to solve a problem in less time
  - Economics – buy a faster computer vs. efficiently use existing computer resources



Types of Problems

- Weather forecasting & modelling
- Economic Modelling
- Geological & seismic activity
- Computer-aided Design
- Searching for extra terrestrial life
- Bioinformatics
  - Drug design
  - Sequence alignment
  - Molecular evolution
  - Protein modelling
  - Etc….

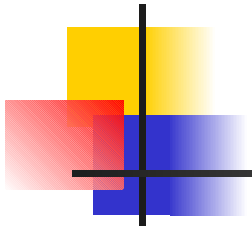


Distributed Computing

- Desktop PC's
  - Average computer processor idle 80% of day
  - Email, word processing, Internet browsing, etc.
  - Average size academic department 00's desktop PC's
- Bioinformatics
  - Massive accumulation genomic data
  - Many computationally intensive tasks
  - e.g. database searching, phylogenetic analysis
  - Conclusion: Large amounts of processing power needed
- Distributed Computing: Harness the 'idle' clock cycles of semi-idle machines

Existing Computing Power...

Do these computers look busy?!?



General Problem Suitability

Demonstrate characteristics such as the ability to be:

- Broken into similar discrete parts that can be solved independently
- Execute multiple program instructions at any moment
- Solved in less time using multiple processors

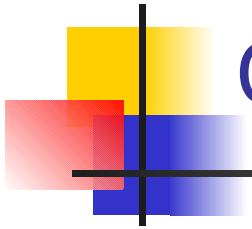
Some Examples

Condor
High Throughput Computing

Folding@home distributed computing

UNITED DEVICES™

the globus toolkit™
www.globustoolkit.org



Our Heterogeneous Distributed System

Heterogeneous distributed computing platform

General-purpose

Java based

- N.B. Strict Donor Machine Security

Execute several computations simultaneously

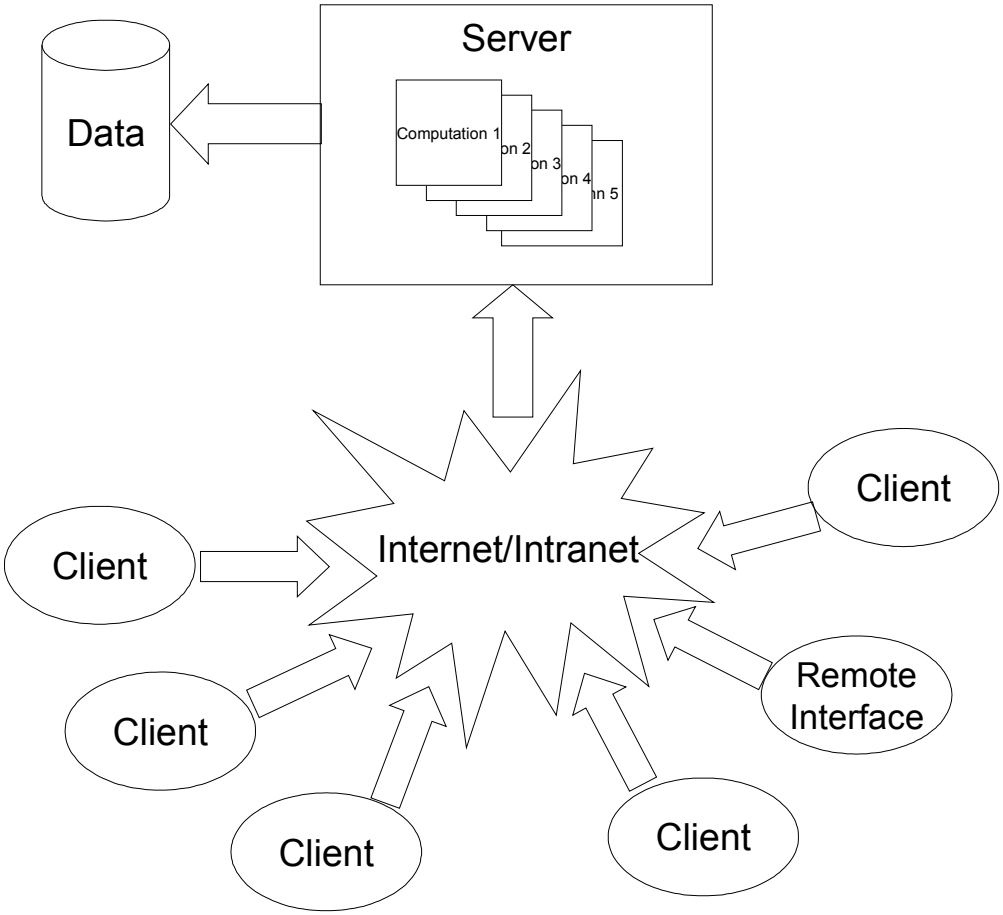
Completed Bioinformatics applications

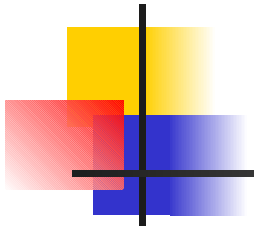
- DSEARCH (Keane and Naughton, *Bioinformatics*, 2005*)*
- DPRml (Keane *et al., Bioinformatics*, 2005)

Biomedical imaging – Light propagation through brain tissue

Cryptography – ElGamal, MD5

Our System


Server
Data
Computation 1
Internet/Intranet
Client
Client
Client
Client
Client
Remote Interface

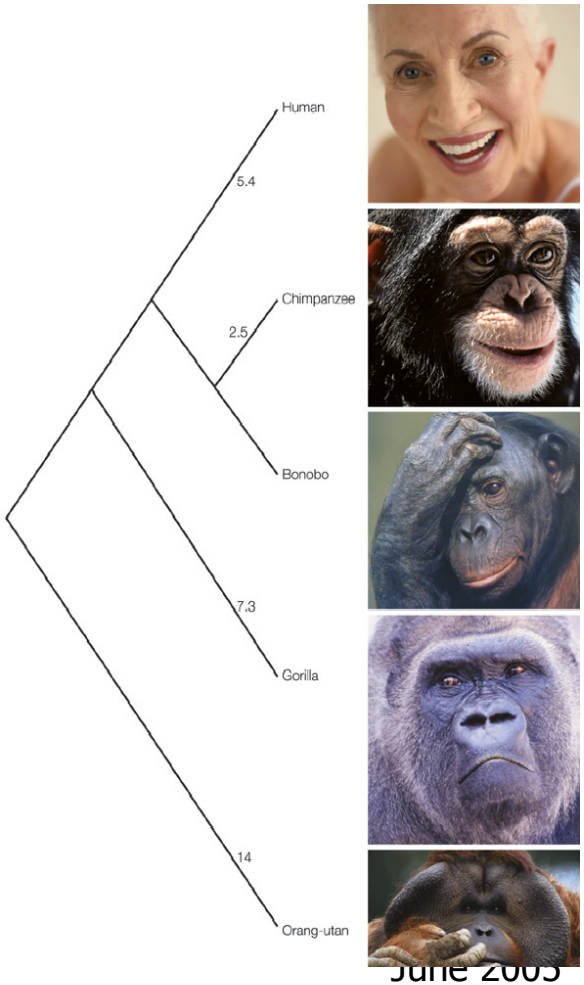


Current Deployment

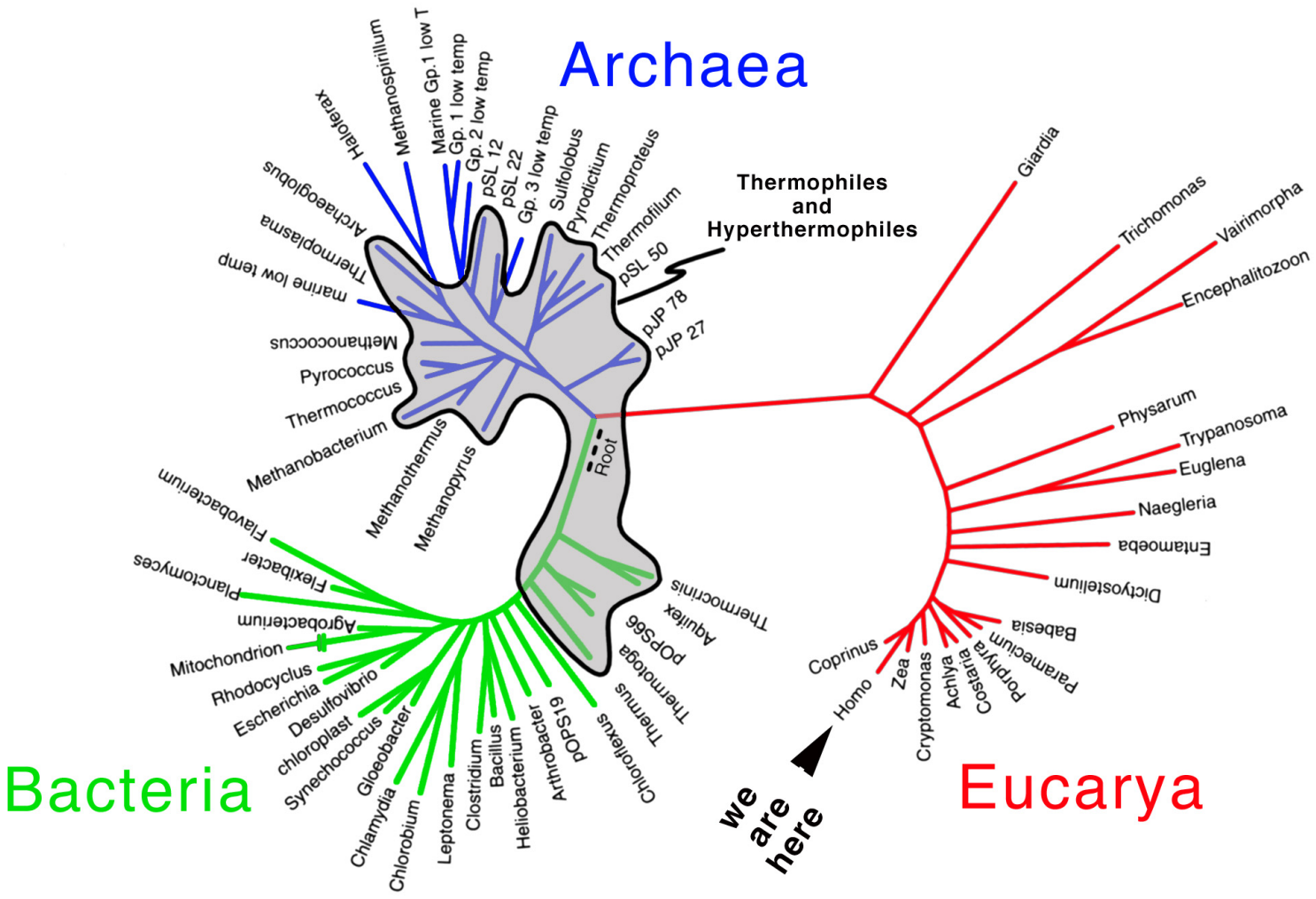
- 200+ Desktop PC's
  - Computer Science, Biology & Elec. Engineering
  - Win 9X/NT/2000/XP
  - Linux – Debian, Redhat, Gentoo
  - Solaris
  - Mac OSX
  - HP PA-RISC
- NUIM Bio Linux Cluster (32 dual processor nodes)
- Low Priority Background Service (running 24/7)
- Free Supercomputing
- ~20 GigaFLOPS

Phylogenetic Analysis

- Molecular Evolution
- Phylogenetic analysis
  - Study of the evolution of life forms
- Phylogenetic tree
  - Tree of several life forms and their relations
- Internal Node
  - Two species diverged from a common ancestral species
- Diversity of life on earth



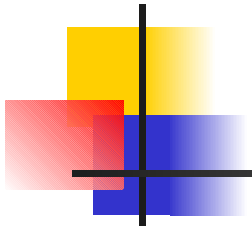
The Tree of Life



http://www.astro.washington.edu/endsofworld

Andrew Page, NUIM

June 2005



DPRml-II

Distributed Phylogeny Reconstruction by maximum likelihood

Batch alignment upload (00's or 000's simultaneously)

Protein and nucleotide sequences

80 protein and 56 nucleotide models

Model selection

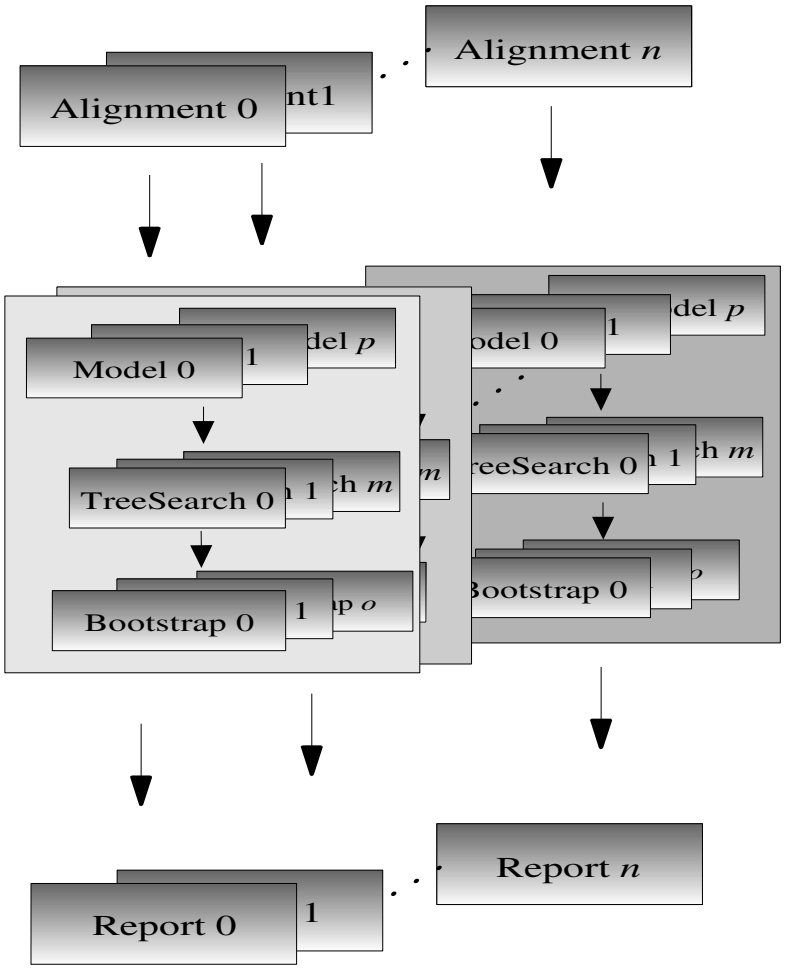
Treesearch

- NNI branch swapping, SPR rearrangements

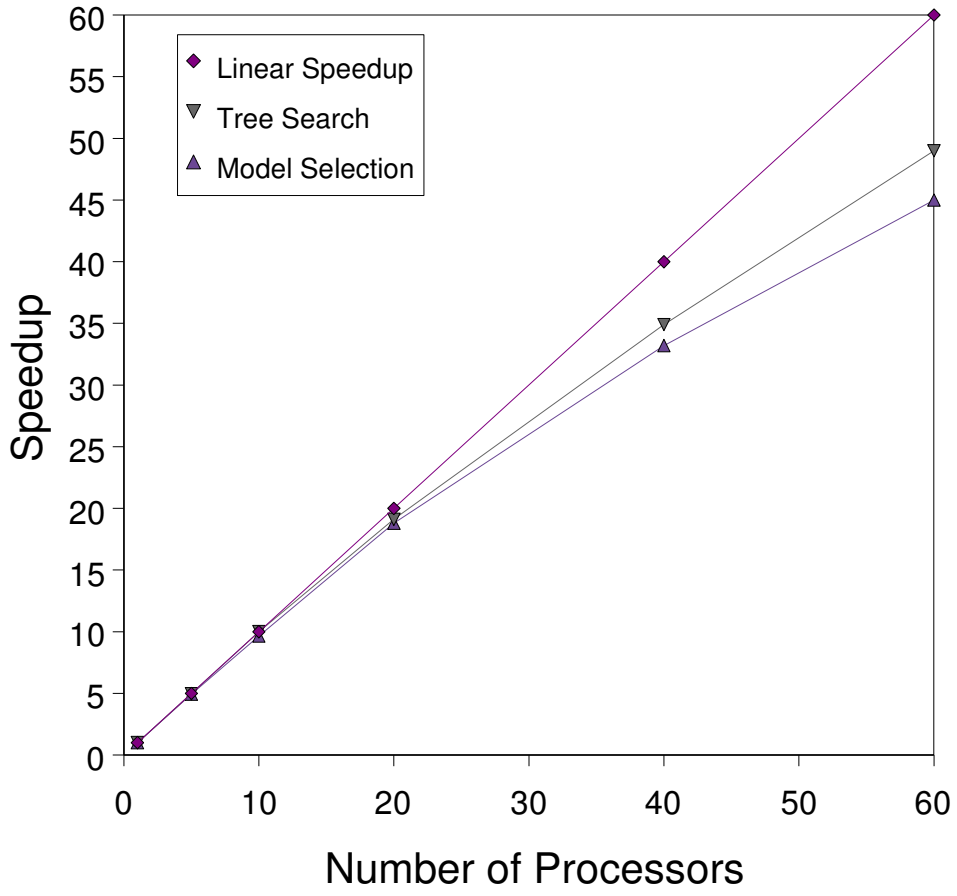
Parallel Bootstrapping

Consensus Tree

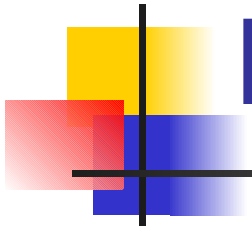
DPRml-II


Alignment n
Alignment 0
Model 0
TreeSearch 0
Bootstrap 0
Report n
Report 0

DPRml-II Parallel Speedup

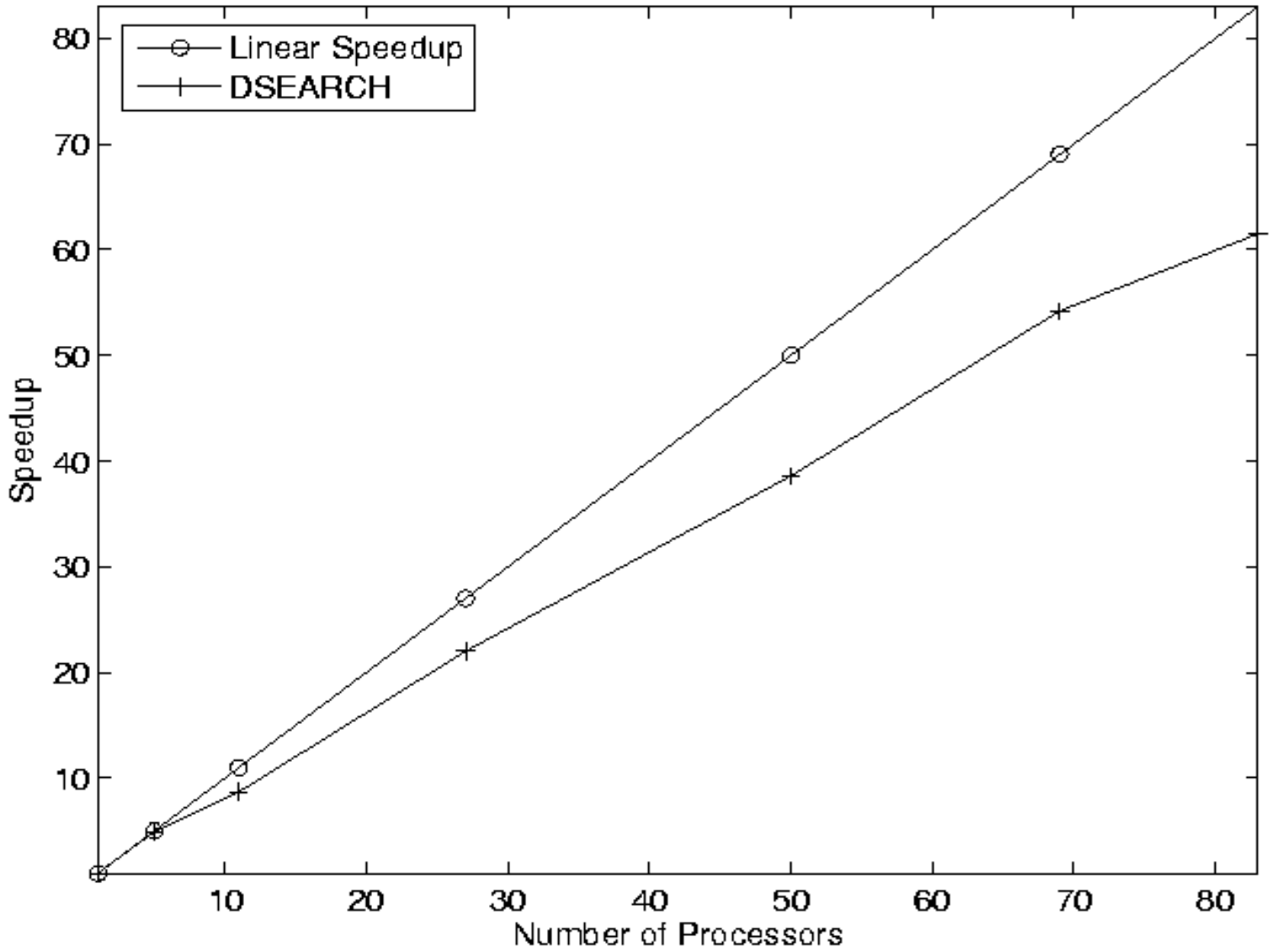


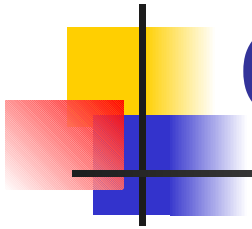
100 alignments
9-63 taxa per alignment
Maximum Likelihood:
Model Selection
Tree Searching



DSEARCH

- Database searching is a fundamental task in bioinformatics
- Large databases and long searches
- Heuristics reduce search sensitivity
- DSEARCH
  - Split database into chunks
  - Parallel Granularity controlled by issuing varying sized groups of query sequences
  - Search database remotely on donor machines


Linear Speedup
DSEARCH
Speedup
Number of Processors
10
20
30
40
50
60
70
80



Conclusions and Future Work

Conclusions

- Many computationally intensive tasks in bioinformatics
- Developed a general-purpose distributed computing platform
- Focus on bioinformatics applications

Future

- Multi-user accounts
- Efficieny Scheduling Algorithms
- Establish an International Phylogenetic Supercomputer

Contact

Homepage

http://www.cs.nuim.ie/distributed

People

Andrew Page - apage@cs.nuim.ie

Thomas Keane - tkeane@cs.nuim.ie

James McInerney

Thomas Naughton