

# Global Analysis of Arthropod Evolution

Craig A. Stewart, Richard Repasky, John Colbourne, David Hart, Donald K. Berry,  
Raymond Sheppard, Eric Wernert, Mary Papakhian, John N. Huffman

University Information Technology Services  
&  
Center for Genomics and Bioinformatics

In collaboration with  
High Performance Computing Center Stuttgart



# Outline

- The SCxy conference and the HPC Challenge
- The biological problem
- The software used
- The global grid
- The results!
- Acknowledgements

# The SCxy conference and the HPC Challenge



- Supercomputing Conference (sponsored by ACM and IEEE)
- High Performance Challenge – demonstrate new capabilities in advanced computing systems

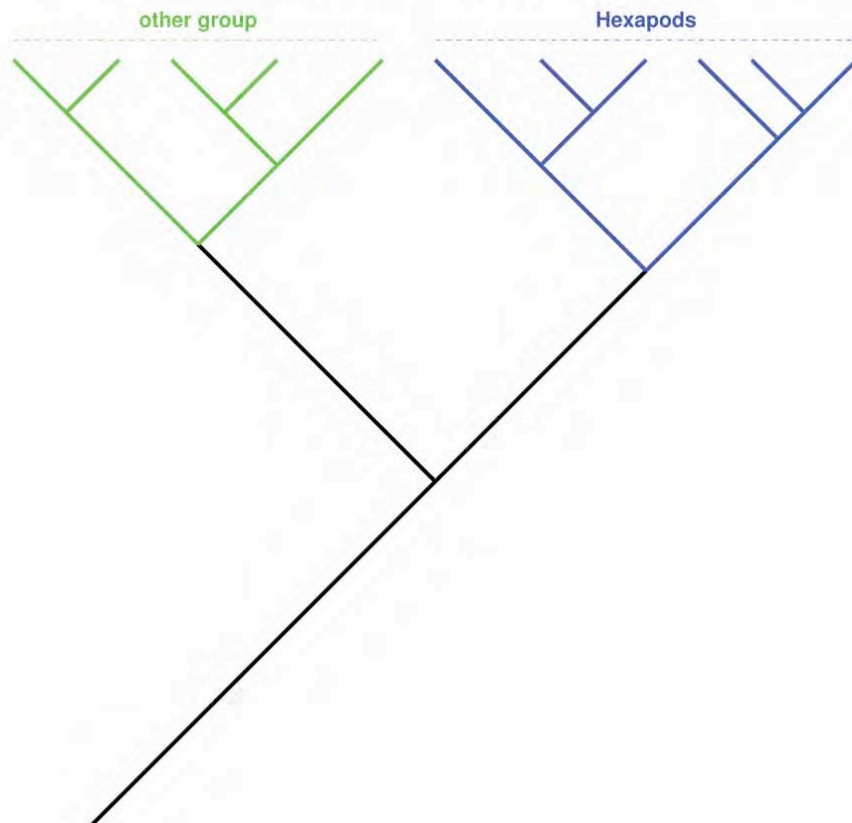


# Biological problem

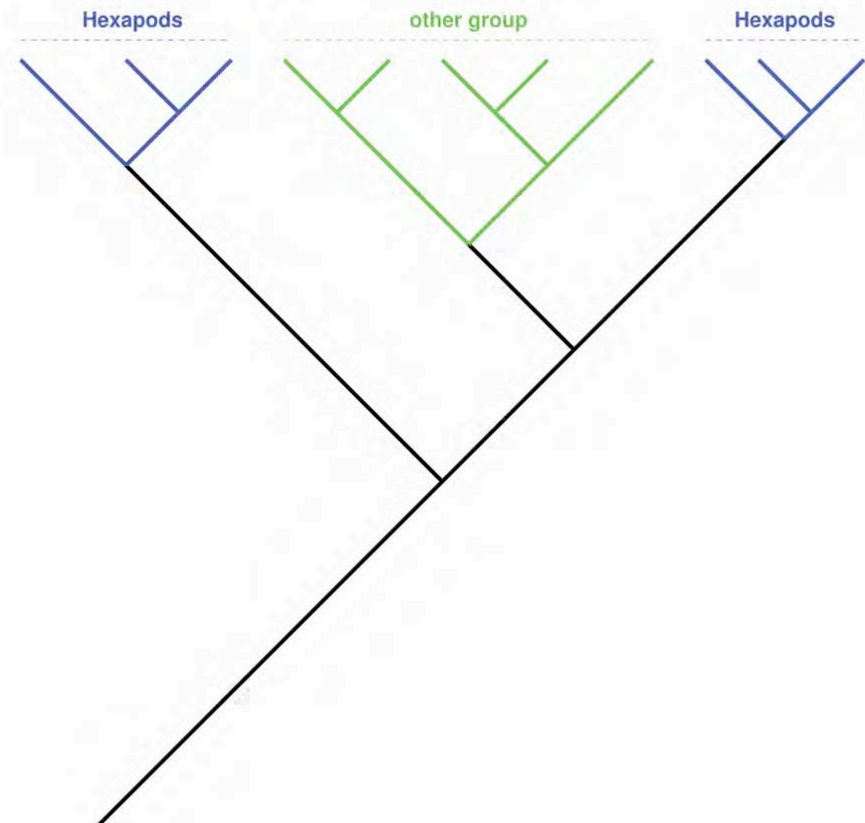


Are Hexapods a single evolutionary group? Are ecdysozoans a single evolutionary group?

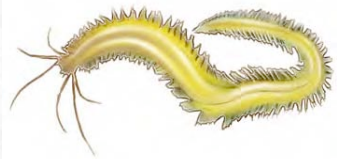
Hexapod monophyly



Hexapod paraphyly



# A partial bestiary



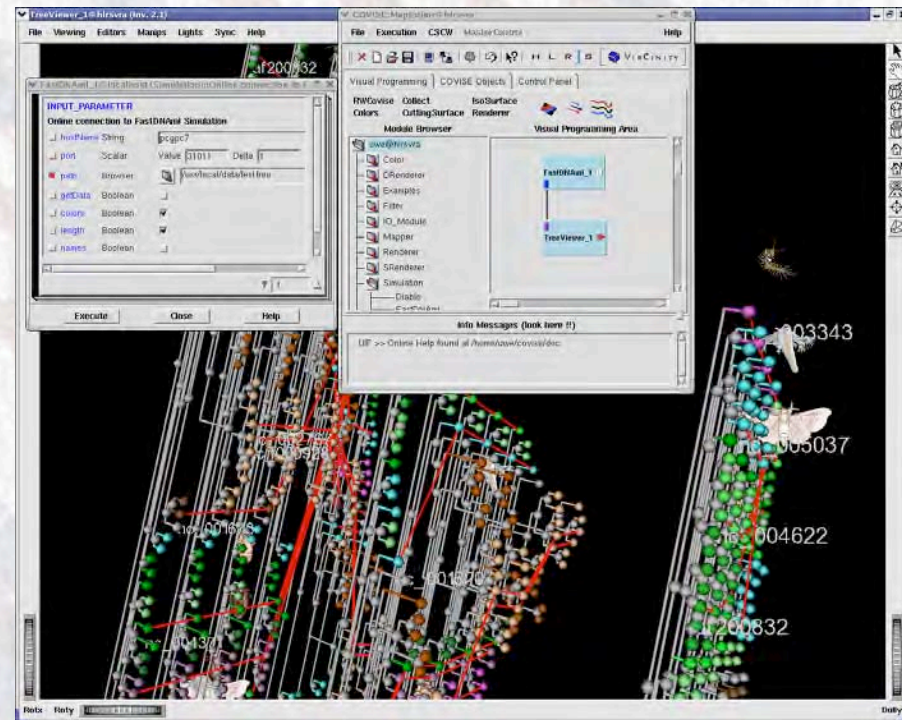
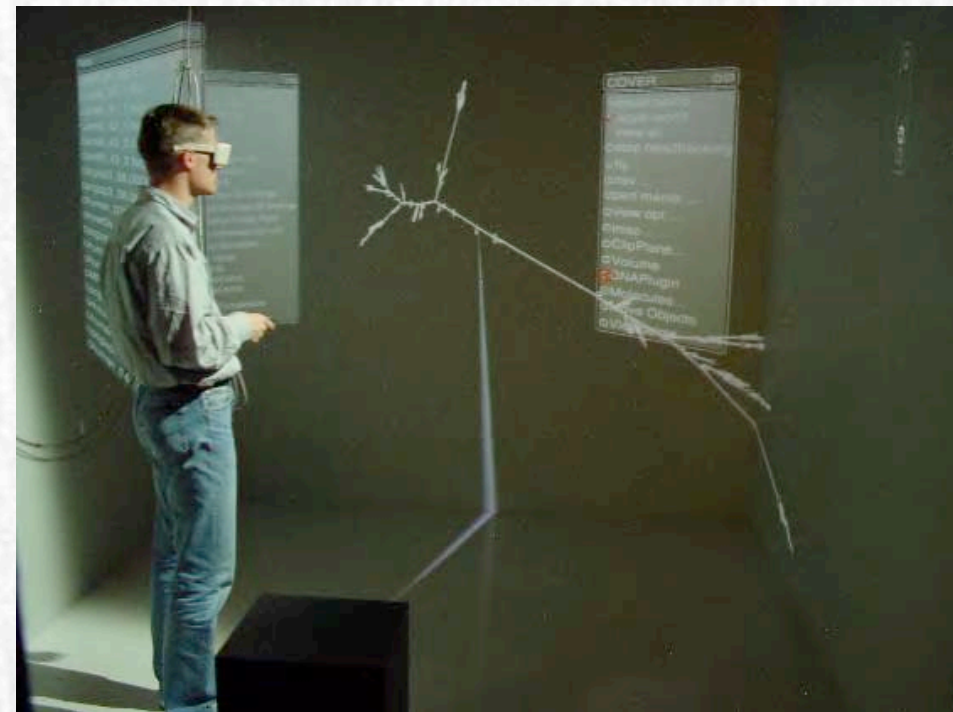
# Software and data analysis

- Non-grid preparatory work
  - Download sequences from NCBI (67 Taxa, 12,162 bp, mitochondrial genes for 12 proteins)
  - Align sequences with Multi-Clustal
  - Determine rate parameters with TreePuzzle
- Grid preparatory work
  - Analyze performance of fastDNAmI with Vampir
  - Meetings via Access Grid & CoVise
- The grid software
  - PACXMPI – Grid/MPI middleware
  - Covise – Collaboration and visualization
  - fastDNAmI – Maximum Likelihood phylogenetics

# PACX-MPI

- A project of HLRS (High Performance Computing Center Stuttgart)
- PACX-MPI (PARallel Computer eXTension) enables seamlessly execution of MPI-conforming parallel applications on a Grid.
- Application recompiled and linked w. PACX-MPI.
- Communication between MPI processes internally is done with the vendor MPI, while communication to other parts of the Metacomputer is done via the connecting network.
- Key advantages:
  - Optimized vendor MPI library is used.
  - Two daemons (MPI processes) take care of communication between systems – allows bundling of communication.

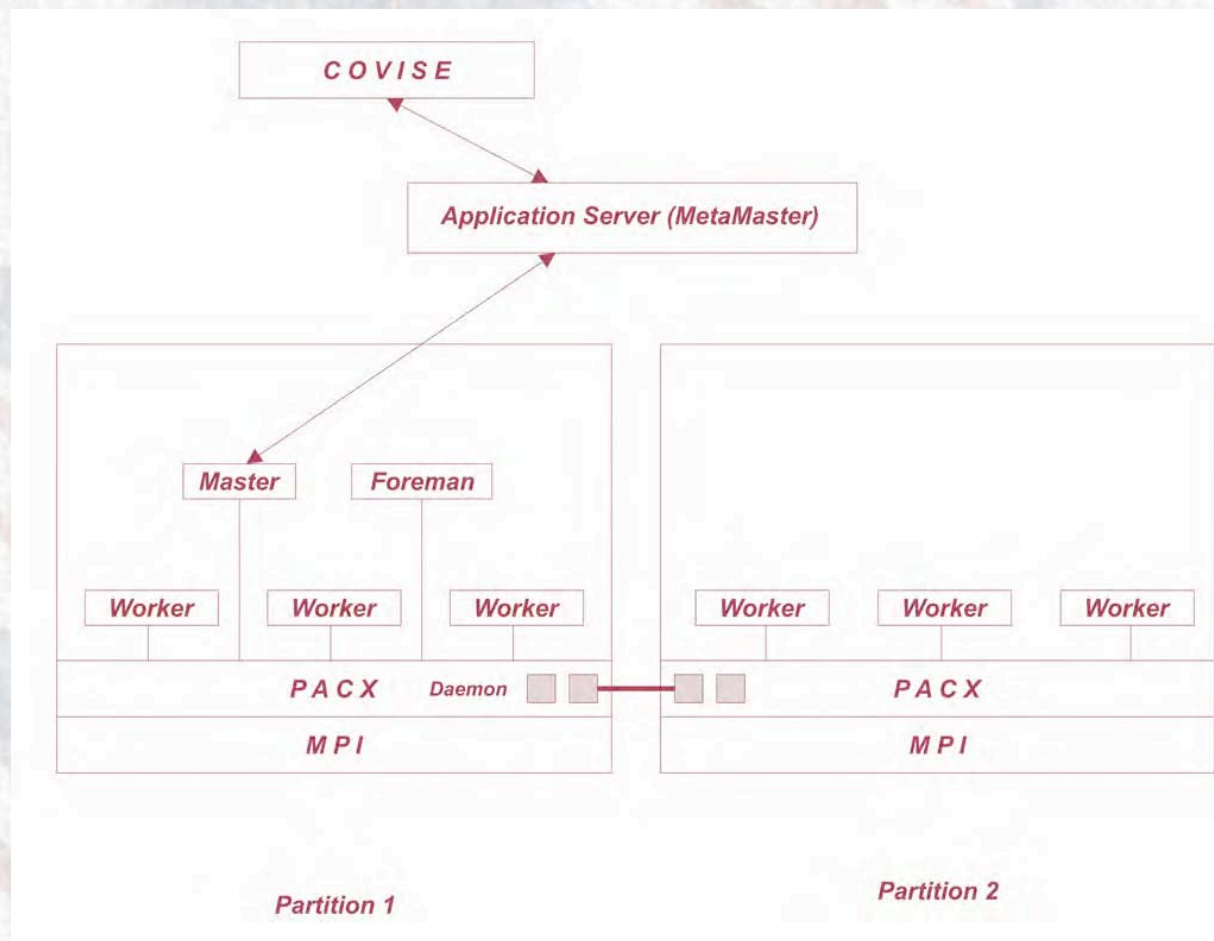
## COVISE



- Collaborative Visualization and Simulation Environment
- A project of HLRS (High Performance Computing Center Stuttgart)
- Focus on collaborative and interactive use of supercomputers
- Interactive startup of calculation on a Computational Grid
- Real-Time visualization of the results and the performance of computation.

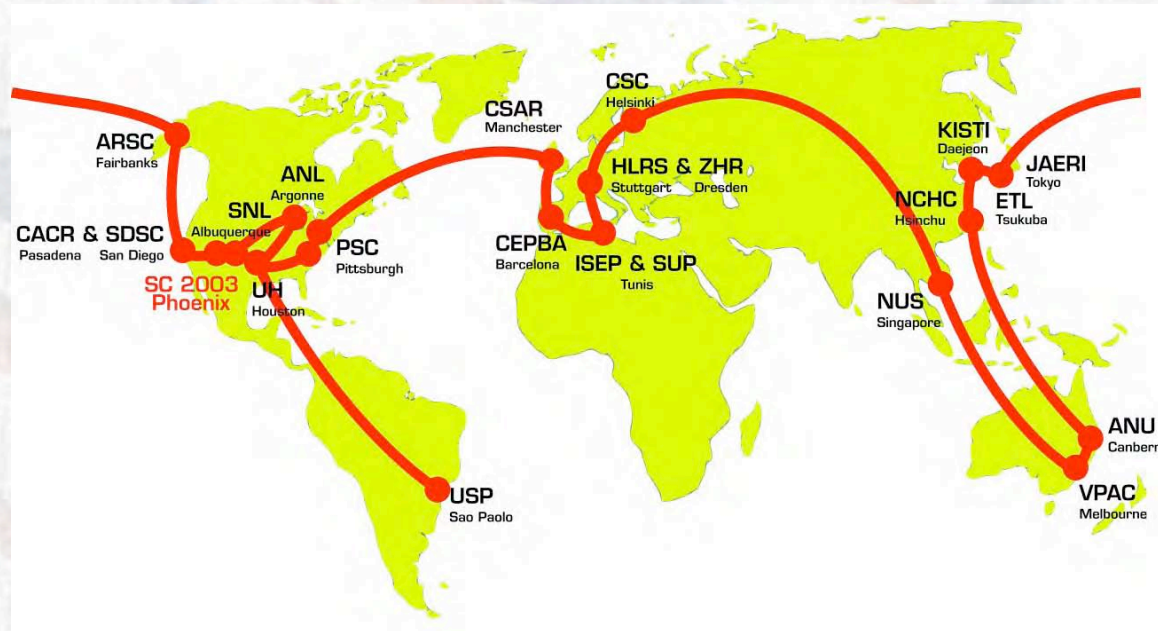
# fastDNAmI

- ML analysis of phylogenetic trees based on DNA sequences
- Foreman/worker MPI program
- Heuristic search for best trees
- For 67 taxa: 2.12  
~ $10^{109}$  trees
- Goal: 300 bootstraps, 10 jumbles per – 3000 executions (more than 3x typical!)



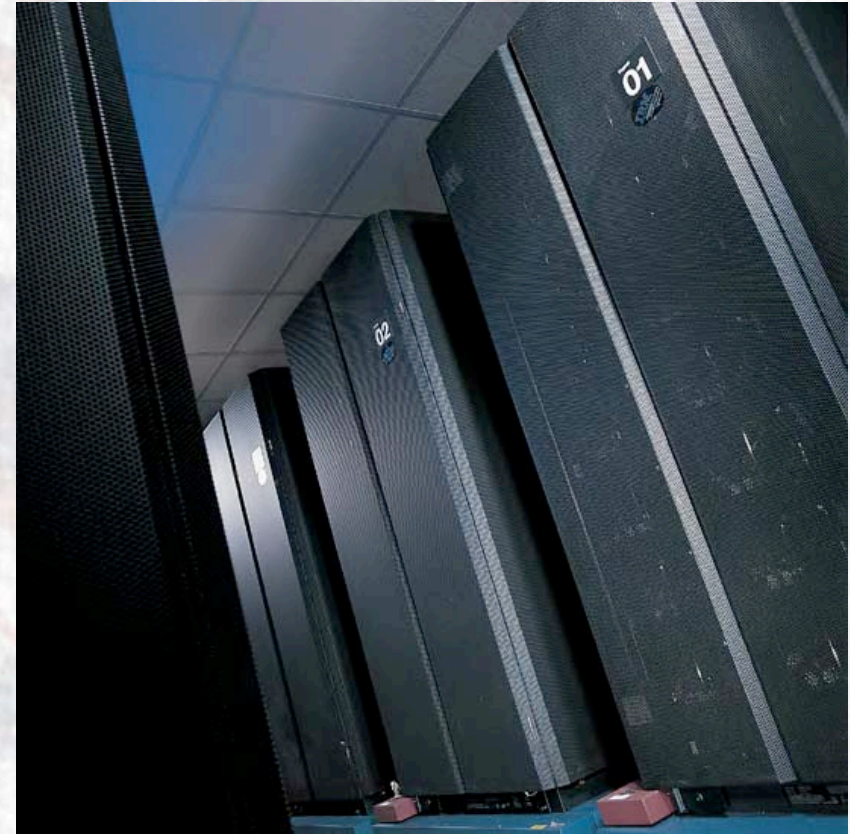
# Why this project on a grid?

- Important & time-sensitive biological question requiring massive computer resources
- A biologically-oriented code that scales well
- Grid middleware environment & collaboration tool well suited to the task at hand
- Opportunity to create a grid spanning every continent on earth (except Antarctica)



# IBM Research SP (Aries/Orion Complex)

- 1.005 TeraFLOPS. 1st University-owned supercomputer in US to exceed 1 TFLOPS peak theoretical processing capacity.
- Geographically distributed at IUB and IUPUI
- Initially 50th, now 302nd in Top 500 supercomputer list
- An enabler of collaborative research using very large scale computations



# AVIDD



- Analysis and Visualization of Instrument-Driven Data
- Distributed Linux cluster. Three locations: IUN, IUPUI, IUB
- 2.164 TFLOPS, 0.5 TB RAM, 10 TB Disk
- First distributed Linux cluster to achieve more than 1 TFLOPS on Linpack benchmark
- Originally 50<sup>th</sup>, currently 114<sup>th</sup> on Top500 list

# The metacomputers

- One
 

Origin 2000	32	Spain
Linux cluster	64	Japan
Linux cluster	12	Australia
IBM SP	32	US
- Two
 

T3E	128	Germany
IBM SP	64	US
Dec Alpha	4	Brazil
Sun fire 6800	16	Singapore
- Three
 

Hitachi SR8000	32	Germany
Cray T3E	128	UK
Cray T3E	32	US
IBM SP (Blue Horiz)	32	US
- Four
 

Dec Alpha (Lemieux)	64	US
---------------------	----	----
- Five
 

Linux system	1	Tunisia
--------------	---	---------

*Five functional units; 8 types of systems (several on Top500 list); 6+ vendors; 641 processors; 9 countries, 6 continents*



# The results

- Hundreds of trees were analyzed during the course of the week
- The biological results are still being analyzed
- Our HPC challenge project was awarded the prize for “Most geographically distributed application”



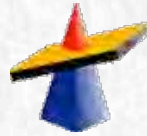
## For further information

- fastDNAMl: <http://www.indiana.edu/~rac/hpc/fastDNAMl/>
- PACXMPI: [www.hlrs.de/organization/pds/projects/pacx-mpi](http://www.hlrs.de/organization/pds/projects/pacx-mpi)
- COVISE: [www.hlrs.de/organization/vis/covise](http://www.hlrs.de/organization/vis/covise)
- HLRS: [www.hlrs.de](http://www.hlrs.de)
- UITS: [uits.iu.edu](http://uits.iu.edu)
- Center for Genomics and Bioinformatics: [www.cgb.indiana.edu](http://www.cgb.indiana.edu)
- SCxy: [www.supercomp.org](http://www.supercomp.org)
- [about.uits.iu.edu/divisions/rac/index.html](http://about.uits.iu.edu/divisions/rac/index.html)
- [about.uits.iu.edu/divisions/rac/pubsstaff.html](http://about.uits.iu.edu/divisions/rac/pubsstaff.html)
- [ingen.iu.edu](http://ingen.iu.edu)
- [it.iu.edu](http://it.iu.edu)

# Acknowledgments

- This research was supported in part by the Indiana Genomics Initiative. The Indiana Genomics Initiative of Indiana University is supported in part by Lilly Endowment Inc.
- This work was supported in part by Shared University Research grants from IBM, Inc. to Indiana University.
- This material is based upon work supported by the National Science Foundation under Grant No. 0116050 and Grant No. CDA-9601632. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).
- Assistance with this presentation: John Herrin, Malinda Lingwall, W. Les Teach
- Thanks to the SciNet team and SC2003 organizers!

# Our partners



独立行政法人 産業技術総合研究所  
先端情報計算センター  
Tsukuba Advanced Computing Center



UNIVERSIDADE DE SÃO PAULO  
CENTRO DE COMPUTAÇÃO ELETRÔNICA



Rainer Keller, Matthias Hess	HLRS, University of Stuttgart
Richard Repasky	UIITS, Indiana University
John Colbourne	Center for Genomics and Informatics, Indiana University
Craig Stewart, David Hart	UIITS, Indiana University
Jennifer Steinbachs	Center for Genomics and Bioinformatics, Indiana University
Uwe Woessner	HLRS, University of Stuttgart
Donald Berry	UIITS, Indiana University
Matthias Mueller	HLRS, University of Stuttgart
Huian Li	UIITS, Indiana University
Gary W. Stuart	Center for Genomics and Bioinformatics, Indiana University
Michael Resch	HLRS, University of Stuttgart
Eric Wernert	UIITS, Indiana University
Martin Aumüller, Ulrich Lang	HLRS, University of Stuttgart
Markus Buchhorn	Australia National University
Hiroshi Takemiya	National Institute of Advanced Industrial Science & Technology, Japan
Rim Belhaj	ISET'Com, Tunisia
Wolfgang E. Nagel	ZHR, Technical University of Dresden
Sergui Sanielevici	Pittsburgh Supercomputing Center
Sergio takeo Kofuji	LCCA/CCE-USP
David Bannon	Victorian Partnership for Advanced Computing, Australia
Norihiro Nakajima	Japan Atomic Energy Research Institute
Rosa Badia	CEPBA-IBM Research Institute
Mark A. Miller	San Diego Supercomputer Center
Hyungwoo Park	Korea Institute of Science and Technology Information
Rick Stevens	Argonne National Laboratory
Fang-Pang Lin	National Center for High Performance Computing
John Brooke	Manchester Computing
David Moffett	Purdue University
Tan Tin Wee	National University of Singapore
Greg Newby	Arctic Region Supercomputer Center
J.C.T. Poole	CACR, Cal-Tech
Ramched Hamza	Sup'com, Tunisia
Mary Papakhian, John N. Huffman	UIITS, Indiana University
Leigh Grundhoeffer	UIITS, Indiana University
Ray Sheppard	UIITS, Indiana University
Peter Cherbas	Center for Genomics and Bioinformatics, Indiana U.
Stephen Pickles, Neil Stringfellow	CSAR, University of Manchester

Thank you!

Questions?