Grid Computing Research in Hong Kong

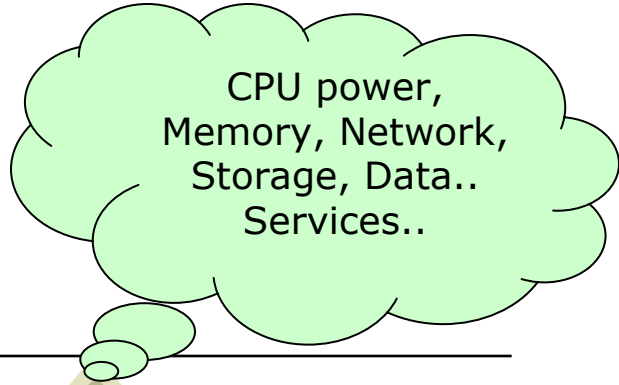
Cho-Li Wang (王卓立)
Systems Research Group (SRG)
Department of Computer Science
The University of Hong Kong
URL: http://www.cs.hku.hk/~clwang/

Outline

- Hong Kong Grid Status Report
  - Hong Kong Grid Initiatives
    - HKU CC, HKBU, HKU CS clusters
  - China National Grid Project
  - Asia Pacific Grid Project
- Grid Research Projects in HKU CS
  - SLIM and InstantGrid
  - JESSICA2
  - G-JavaMPI and G-PASS
- Summary and Conclusion

Hong Kong Grid

http://www.hkgrid.org/



Resource providers

End users

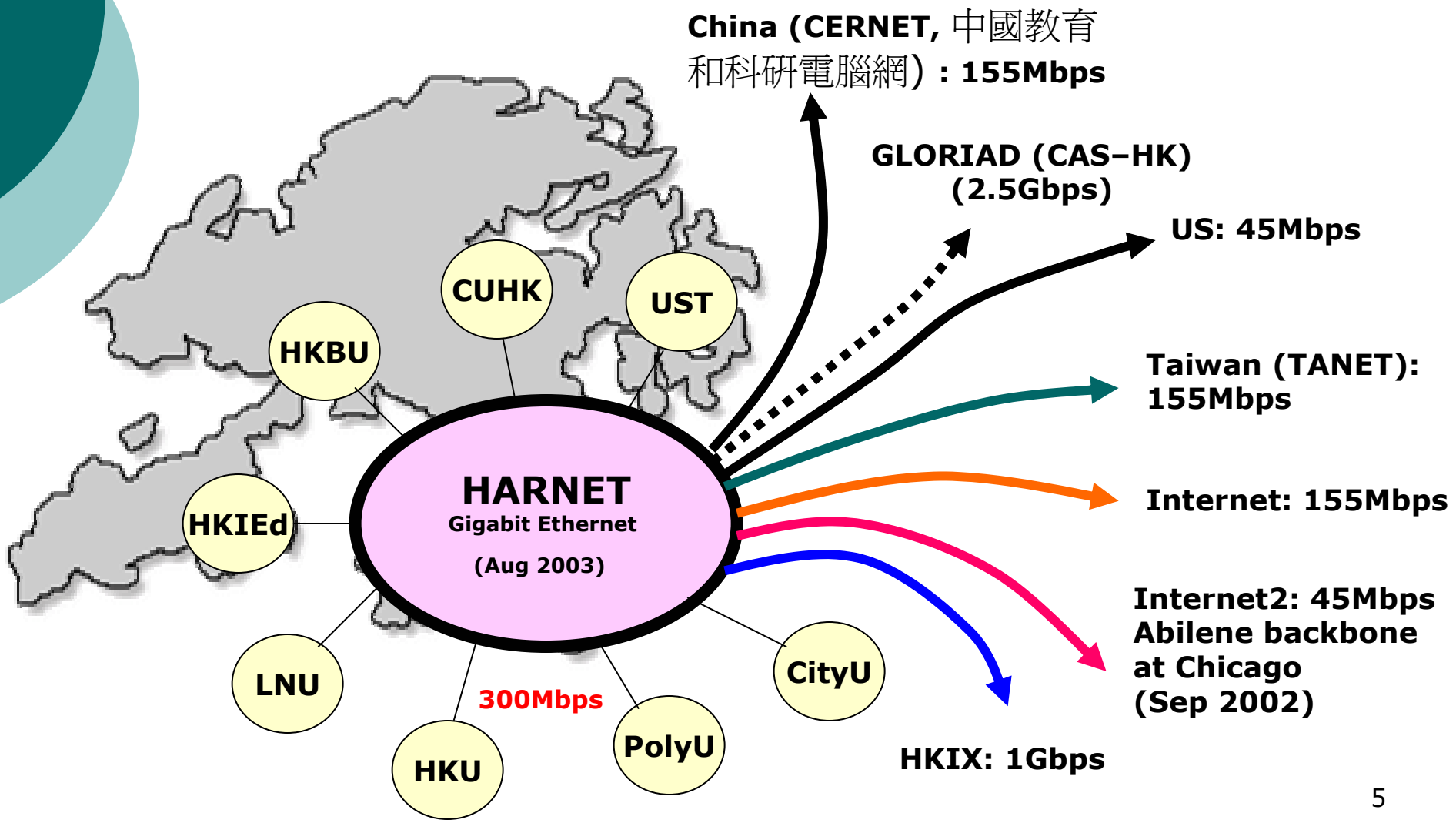
HKGrid Initiatives Launched in Cluster2003 (Dec. 2003)

HKGrid - Current Constituents

Institutions	Computing facilities
香港科技大學 (HKUST)	4-way SMP cluster
香港浸會大學 (HKBU)	2-way Xeon SMP x 64 (#300 in TOP500, 6/2003)
香港城市大學 (CityU)	1 2-way Xeon SMP Service gateway
香港高性能計算所 (HK HPC)	1 2-way Xeon SMP (Service gateway)
香港理工大學 (PolyU)	1 2-way Xeon SMP (Service gateway)
香港大學 (HKU/CC)	2-way Xeon SMP x 128 (#240 in TOP500, 11/2003)
香港大學 (HKU/CS)	Pentium 4 x 300 (#175 in TOP500, 11/2002)

Total computing power (theoretical maximum) = 4 Tflop/s

The Hong Kong Academic & Research Network: HARNET



Grid Research Projects in Hong Kong

- HKUST: Incentive scheduling, topology optimization
- HKBU: Knowledge grid, autonomous grid service composition
- CityU: Agent-based wireless grid computing
- PolyU: Peer-to-peer grid, meta-scheduling, fault tolerance
- HKU
  - CC: Scientific applications running across the ApGrid
  - CC: Biosupport project with HKU-Pasteur Research Centre
  - ETI: Modeling of Air Quality in Hong Kong (with the Environmental Protection Department, HKSAR)
  - ETI: RFID Grid
  - CS : China National Grid (CNGrid) project - HKU Grid Point
  - …..

HKU Computer Centre

hpcpower: 128 nodes (IBM x335) of dual Xeon 2.8GHz CPUs GigaEth connection (CISCO 4506), Linux OS

October 20, 2004 : Inaugural Ceremony of HPC Cluster on Windows Platform

Current Focus:

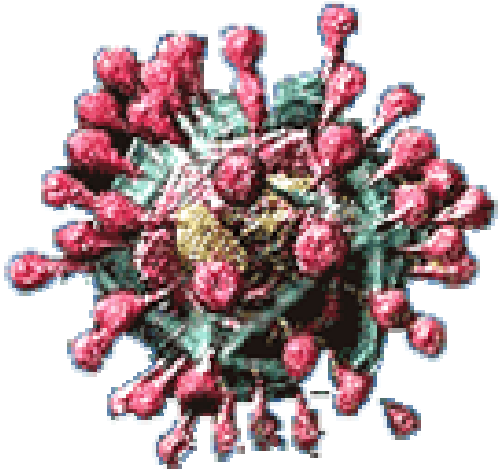
- Core member of HKGrid
- International collaboration supported by HARNET-Internet2 and HARNET-APAN connections
- More collaborations with Chinese institutions
- Exploring implementation of other forms of GRID computing for various purposes as viewed by different groups and companies.

HKU-Pasteur Research Centre

Biosupport Project

- **Collaboration between HKUCC, HKU-Pasteur Research Centre and Centre de Ressources INFOBIOGEN (France).**

- **Bioinformatics Tools:** The **sequence analysis packages** installed include EMBOSS, NCBI tools, FASTA, STADEN, PHYLIP, READSEQ, ClustalW/ClustalX, DIALIGN2 and the PHRAP/PHRED/CONSED package. Some tools installed also have **on-line web interface**, such as JEMBOSS, EMBOSS-GUI, NCBI-BLAST, FASTA and GenoList



NCBI

EMBL-EBI European Bioinformatics

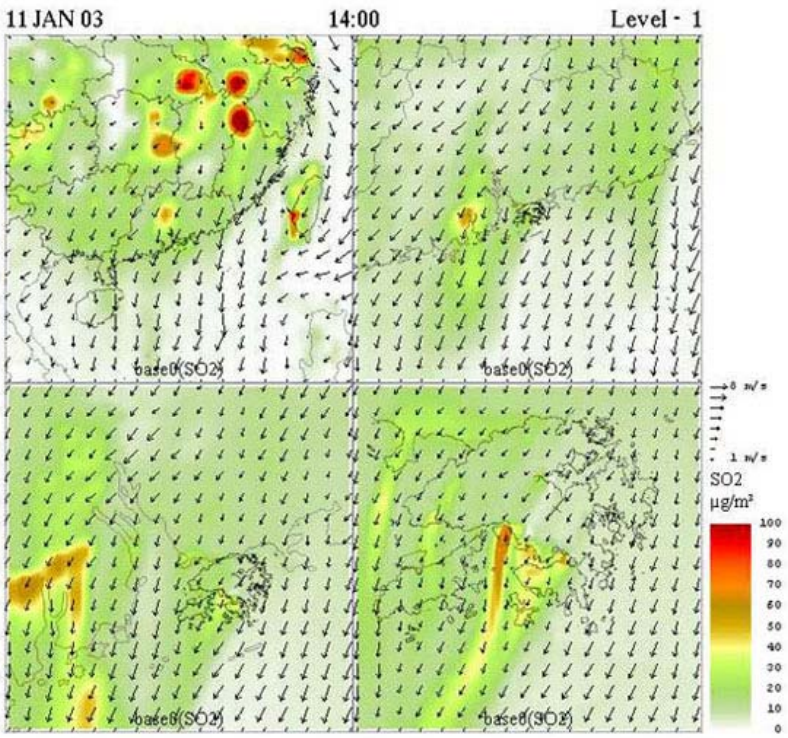
emboss

IUBio

http://www.hkupasteur.hku.hk/hkuip/Home_HKU_P_RCL.html

HKU ETI – EPD
Modeling of Air Quality in Hong Kong

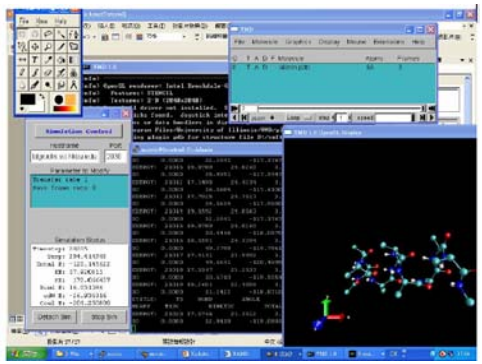
- Collaboration between HKU *E-business Technology Institute* (ETI) and the *Environmental Protection Department* (EPD), HKSAR
- Investigate the inter-connections of the air pollution mosaic through numerical simulation
- Government plans to harness grid technologies to utilize idle PCs during off-hours



Source: http://www.info.gov.hk/digital21/eng/knowledge/gripapp.html

Hong Kong Baptist University

High Performance Cluster Computing Centre



Quantum Chemistry

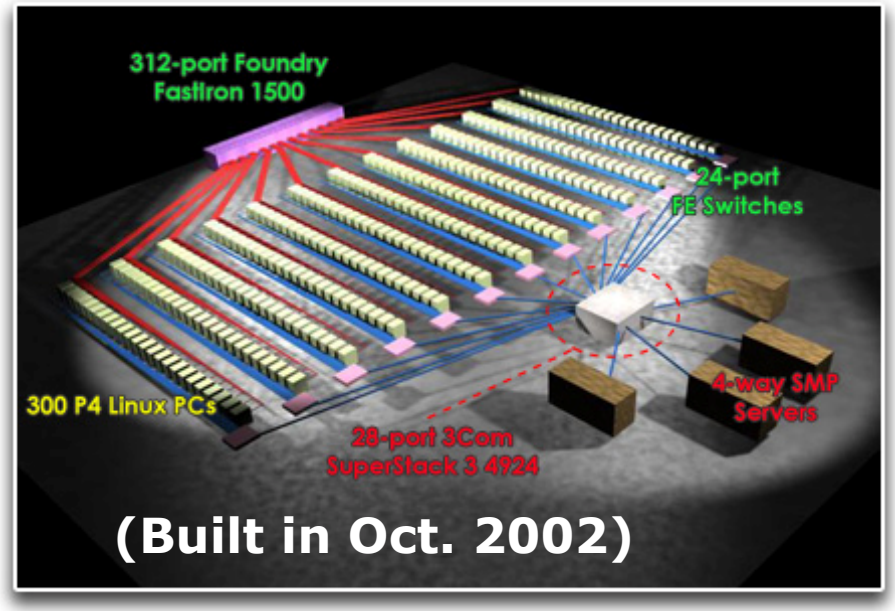
64 nodes (Dual Intel Xeon 2.8GHz with 2GB RAM), Network: 65-port Extreme BlackDiamond 6816 Gigabit Ethernet switch

- **Message Passing Interface**
  - MPICH, LAM/MPI
- **Mathematical:**
  - **fftw** (fast fourier transform)
  - **pblas** (parallel basic linear algebra software)
  - **atlas** (a collections of mathematical library)
  - **sprng** (scalable parallel random number generator)
  - **MPITB** -- MPI toolbox for MATLAB
- **Quantum Chemistry software**
  - gaussian, qchem
  - Molecular Dynamic solver
  - NAMD, gromacs, gamess
- **Weather modelling: MM5**

Source: http://www.sci.hkbu.edu.hk/hpccc/

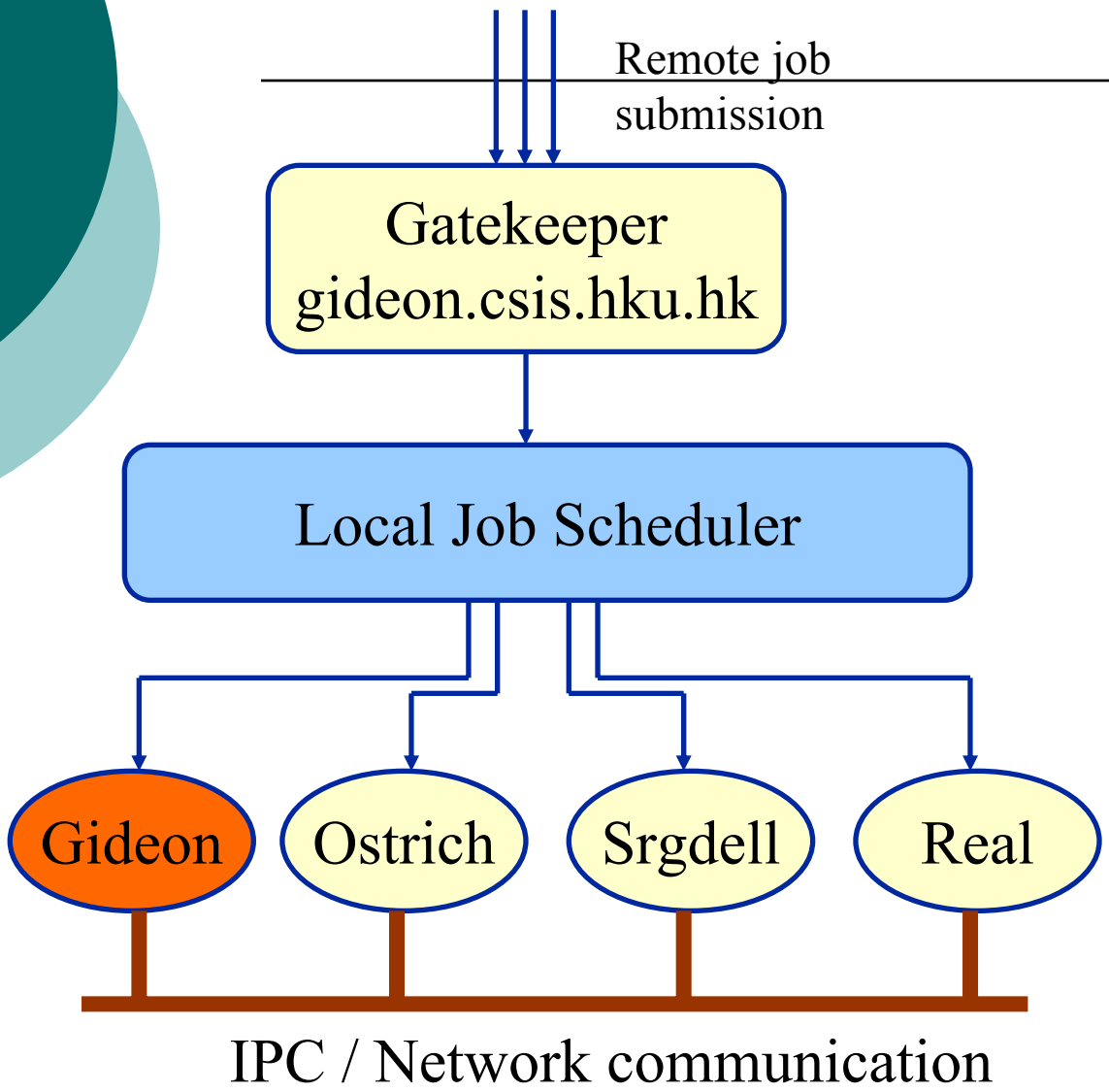
HKU Computer Science

"Self-Made" Gideon 300 Linux cluster



300 Pentium 4 PCs @355 Gflops; Ranked #175 in TOP500 (11/2002)

HKU CS Grid Point: Grid and Cluster Software



Grid middleware

- Globus Toolkit (GT) 2.0, 2.4, 3.0.1

Job scheduling

- OpenPBS 2.3.16
- Maui 3.2.5

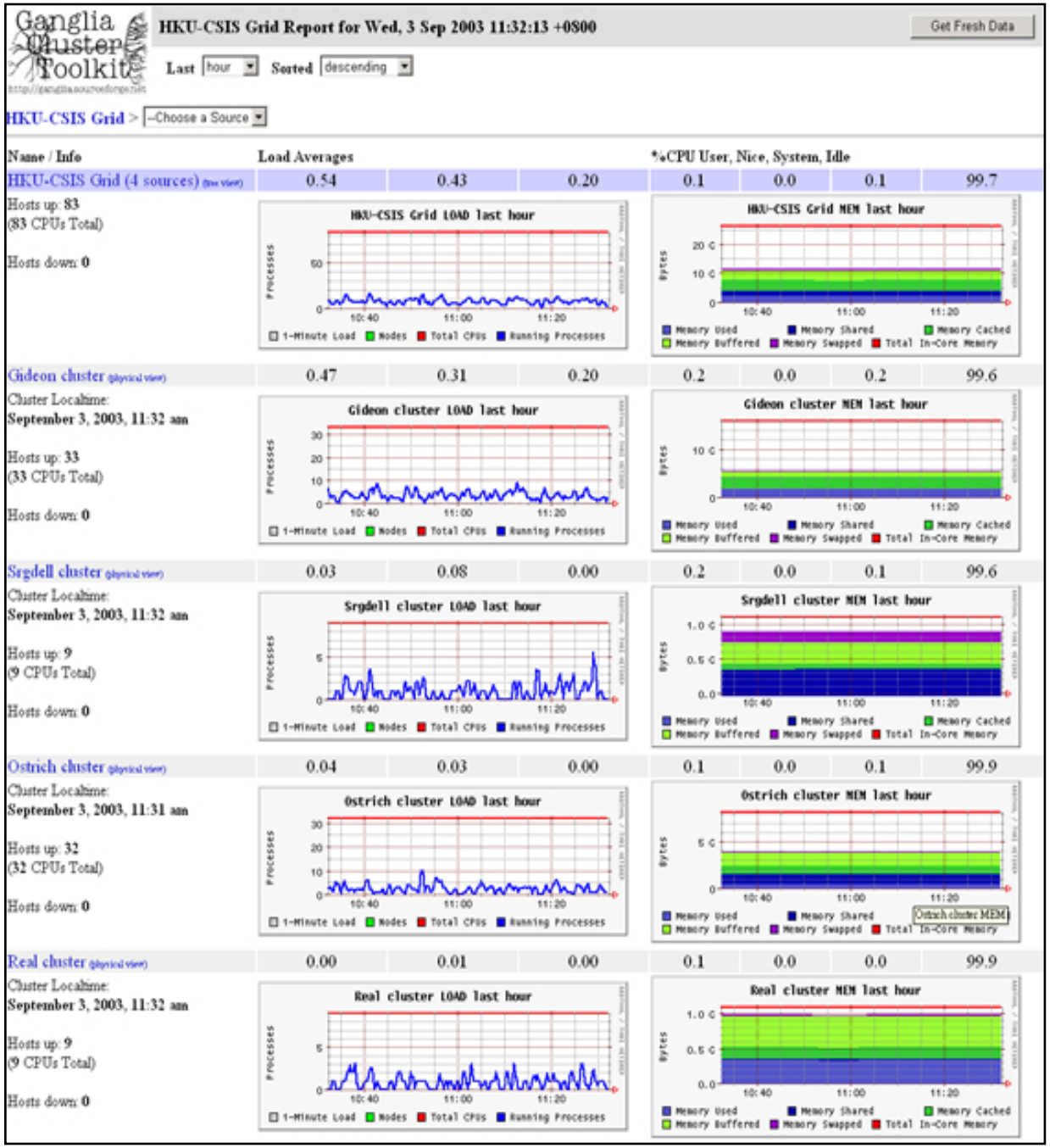
Programming

-HPF, Fortran 90
-C, C++, Java with MPI
-JESSICA2 (HKU)
-WireGL, MatlabMPI

Communication Lib

- MPICH-G2

Performance Monitoring with Ganglia



URL: http://gideon.cs.hku.hk/hkgrid/

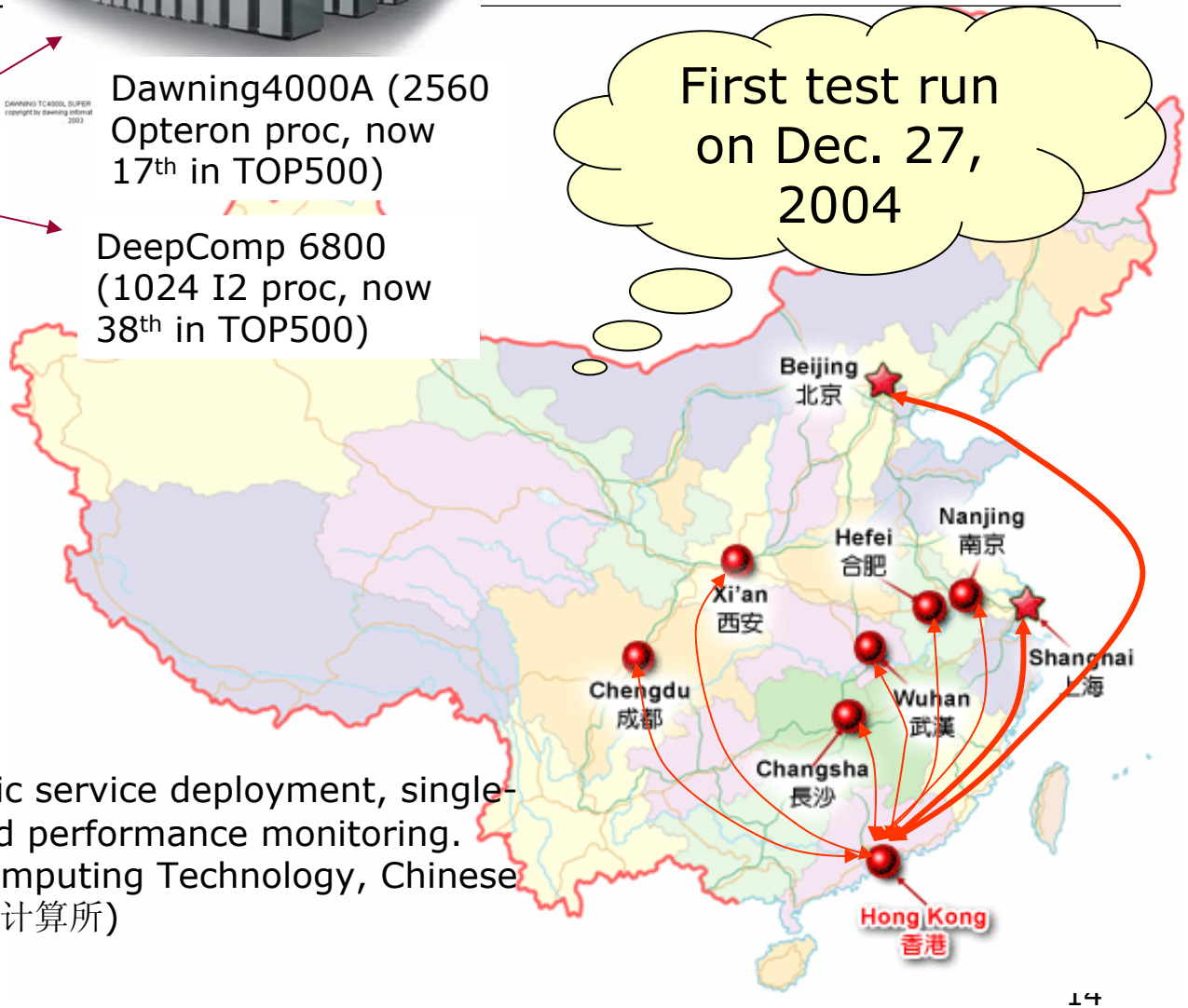
China National Grid : HKU Grid Point

Dawning4000A (2560 Opteron proc, now 17th in TOP500)

DeepComp 6800 (1024 I2 proc, now 38th in TOP500)

- 上海超级计算中心
- 中科院计算所
- 香港大学 (Gideon300)
- 西安交通大学
- 中国科技大学
- 国防科技大学
- 中科院应用物理所
- 清华大学

First test run on Dec. 27, 2004



Supporting software:

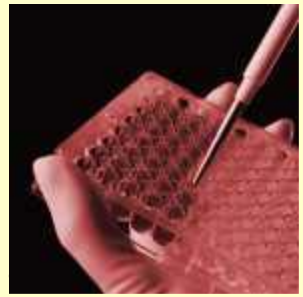
Vega (织女星) GOS: dynamic service deployment, single-sign-on, data replication, and performance monitoring. Developed by Institute of Computing Technology, Chinese Academy of Sciences (中科院计算所)

(2004. Nov. 28) : HKU supports G-JavaMPI, JESSICA2, WireGL, MatlabMPI

Drug Discovery Grid (DDGrid)

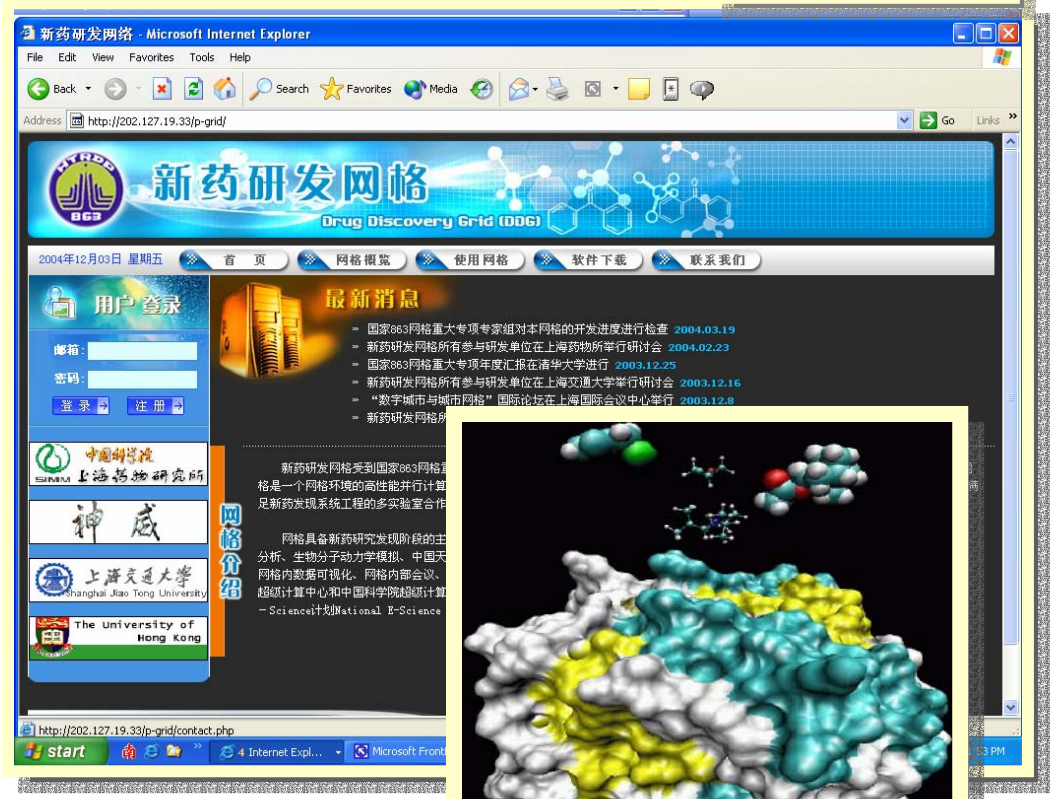
新药研发网格

http://202.127.19.33/



- **Shanghai Institute of Materia Medica** (上海药物所)
- **Shanghai Jiao Tong University** (上海交通大学)
- 江南计算技术研究所
- **University of Hong Kong** (香港大学)

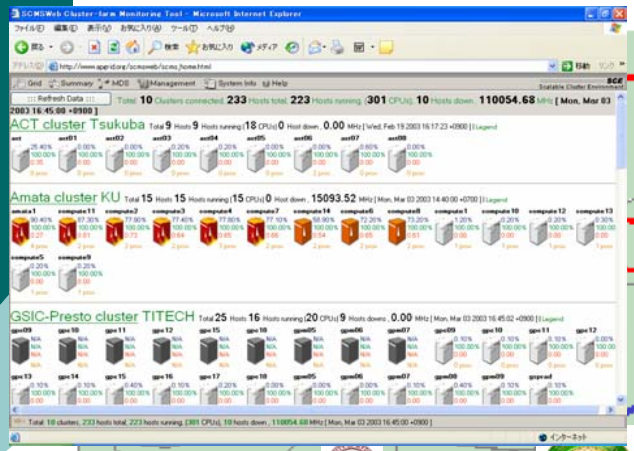
Database: 中国天然产物（中草药）分子数据库、合成化合物分子数据库，化合物毒性数据库、



化合物数据库筛选

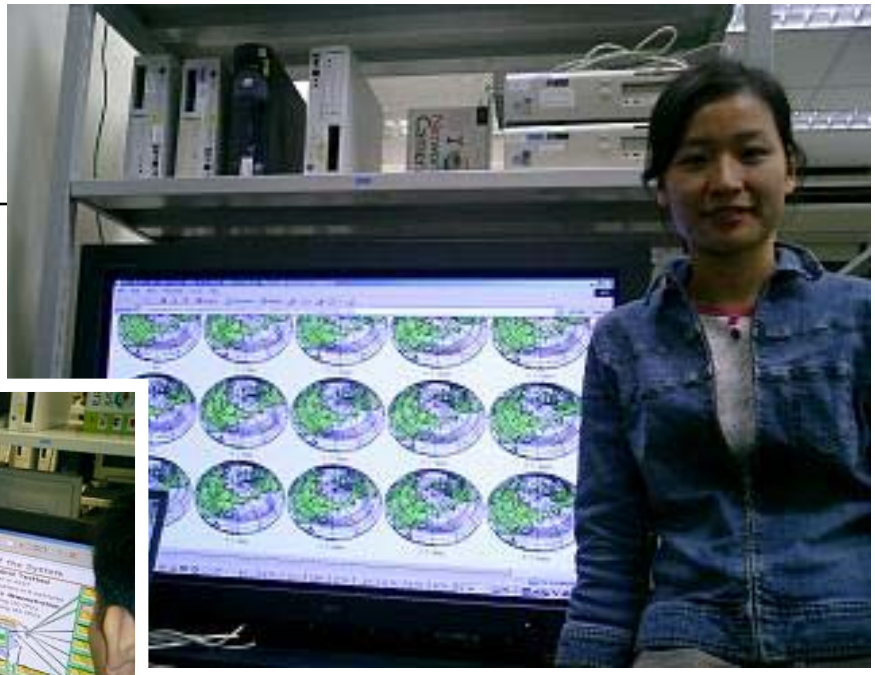
Computing Resources: 上海药物所神威32A集群、北京军事医学科学院神威256P集群、香港大学Gideon 300集群、上海超级计算中心神威64P集群、曙光4000A、大连理工大学等多个网格结点

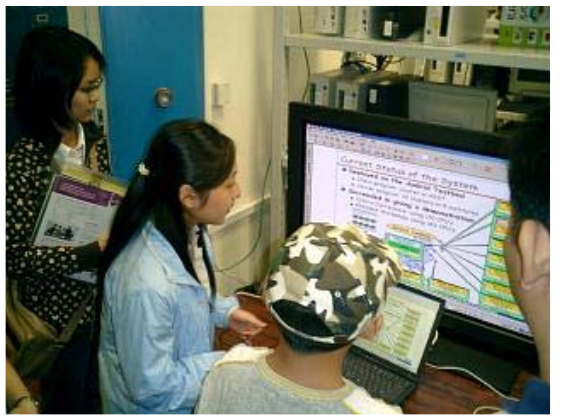
Asia Pacific Grid (APGrid)

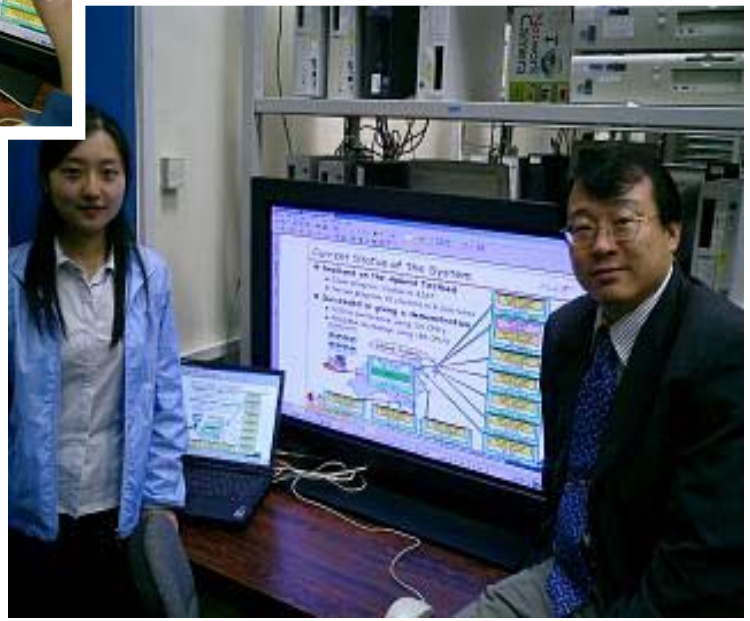


10 countries
21 organizations
22 clusters
853 CPUs

Weather Forecast Demonstration on HKU Open Day – (Oct 2003)







Grid Research at HKU SRG

Selected Projects

- SLIM + InstantGrid
- JESSICA2
- G-JavaMPI + G-PASS

Acknowledgement

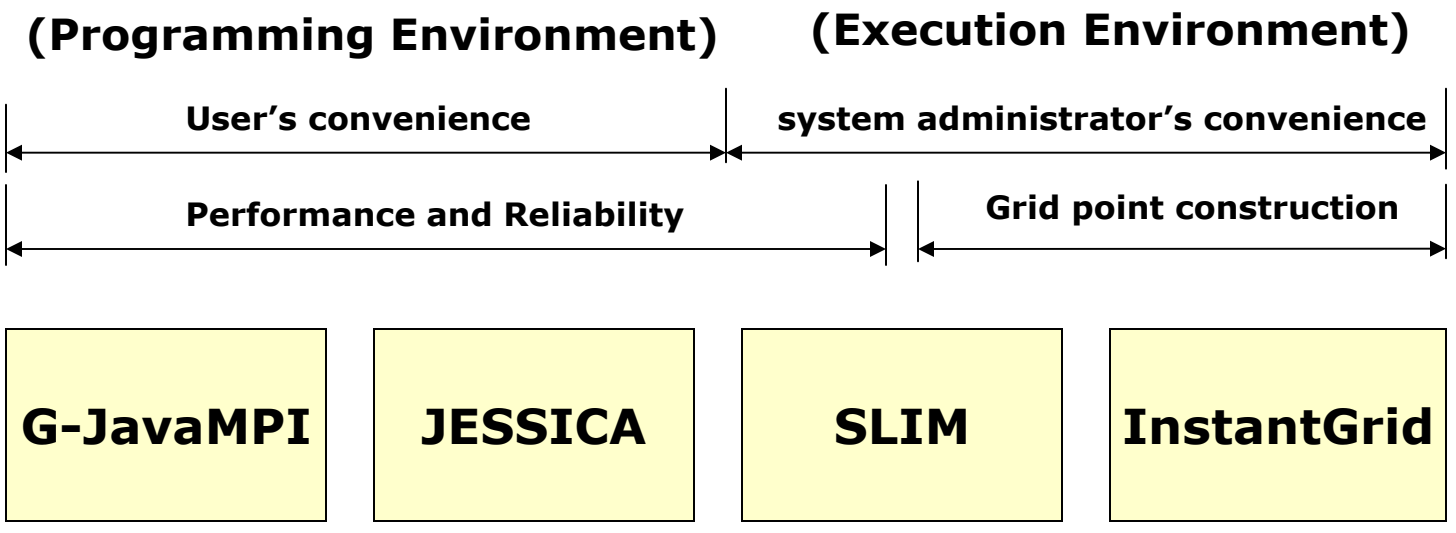
HKU Systems Research Group (SRG)

Our Goal

To construct an advanced grid computing platform to accommodate utility-like computing via traditional and "pervasive" means

- Utility computing: to aggregate and make use of distributed computing resources transparently
- Traditional means: to utilize the dedicated HPC facilities distributed across institutions
  - Performance and reliability are key
- Pervasive means: any user can be resource provider (e.g., idle PCs, etc.) or consumer, or both
  - Convenience and security are key

An Advanced Grid Computing Platform



SLIM

Single
Linux
Image
Management

URL: http://slim.cs.hku.hk/

SLIM 專案的兩位發明人，分別是香港大學計算機科學及資訊系統系電算師孔慶輝(右)及助理電算師李俊暉(左)。

香港大學作為本地歷史最攸久的專上學府，一直以培育世界級的科研、人文人才為使命，在全球開放源碼運動上，他們即將有一項震撼世界的貢獻，Linux Pilot讀者將可率先了解這項劃時代的開放源碼專案SLIM，將如何在Linux的教育、科研及企業應用上發揮巨大影響力。

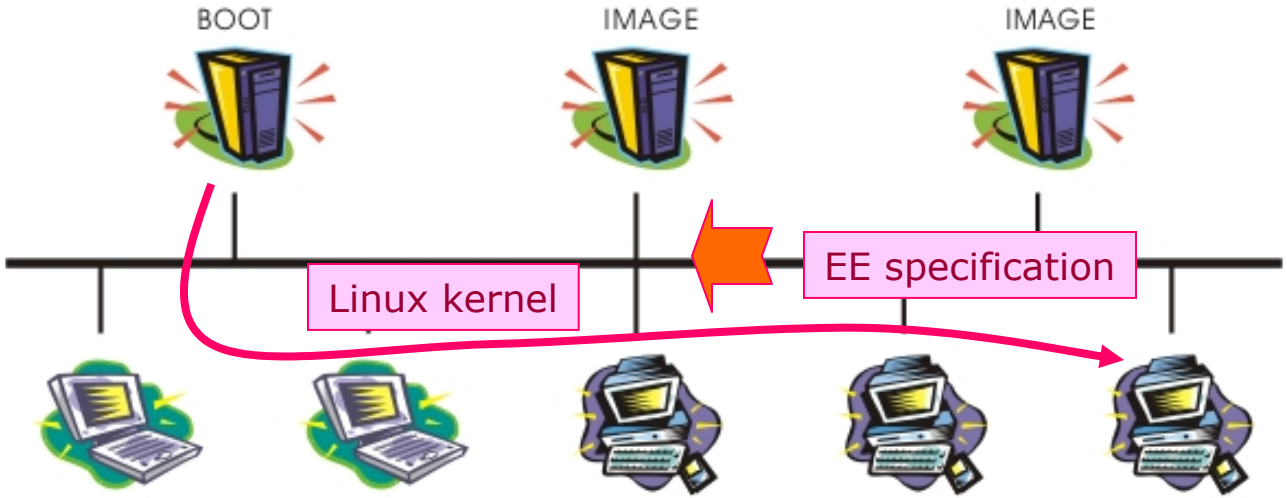
On-demand construction of customized execution environments

(LinuxPilot 2004/04)

SLIM

- Utility computing decouples computing platforms (resources) and computing logic (applications)
- I.e., a single platform can run completely different applications
- Problem: different applications demand different execution environments (OS, shared libraries, supporting apps, etc.)
- SLIM is a network service for managing and constructing EE's, and disseminating them to remote computing platforms

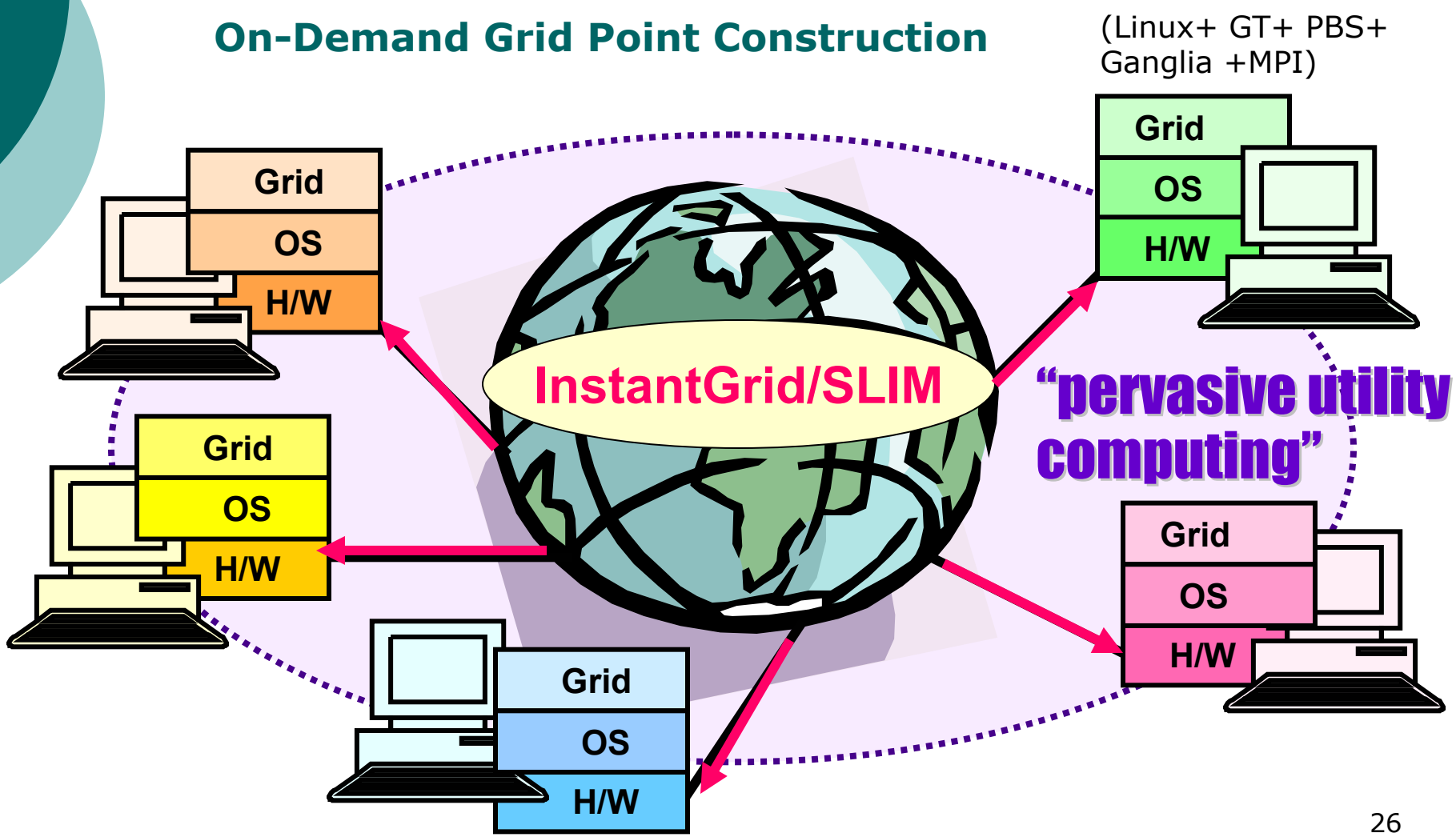
SLIM – System design



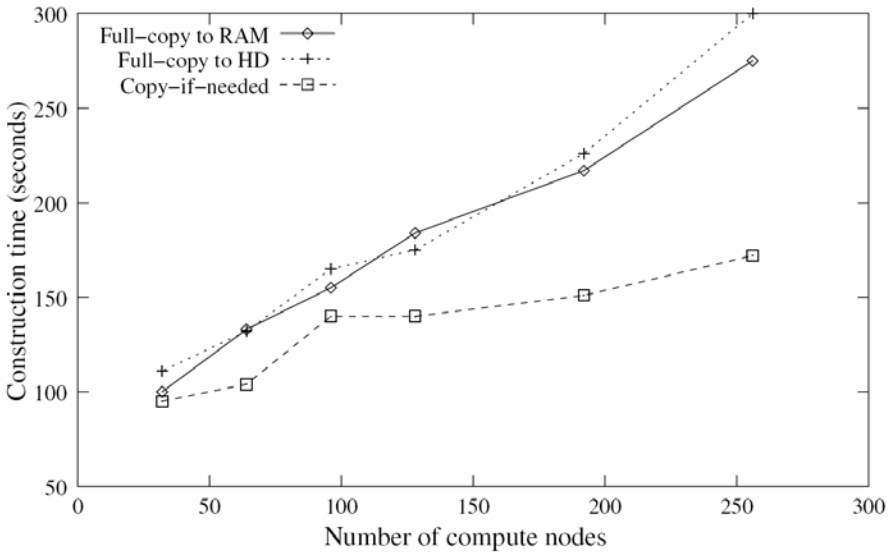
How it works?

- A node sends a EE specification across the network to find the Boot server
- Boot server delivers the requested Linux kernel
- Image server constructs an EE by collecting shared libraries, user data, etc.
- Linux kernel boots, and contacts the Image Server to "mount" the EE via a file synchronization protocol such as NFS
- Aggressive caching techniques are deployed to optimize performance

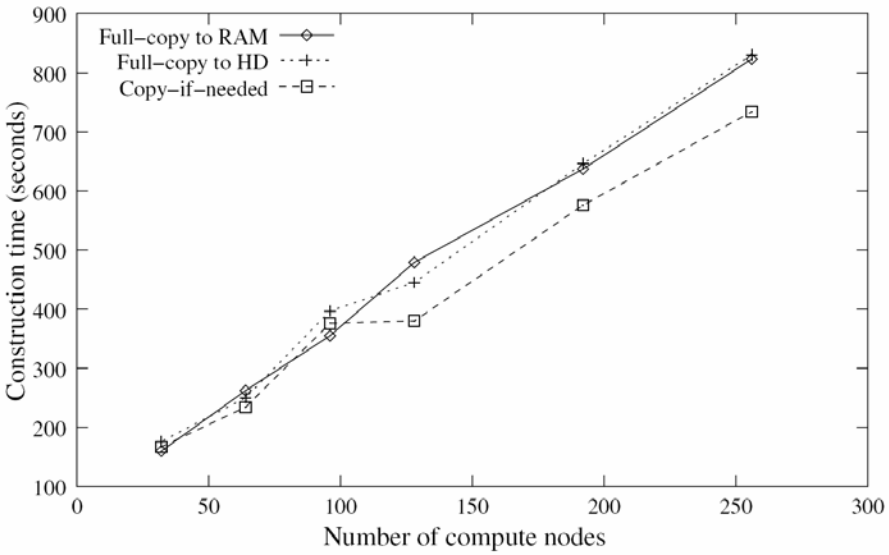
InstantGrid on SLIM



InstantGrid Performance



(a) A cluster–based grid point



(b) Standalone grid points

- Construct a 256-node grid point from scratch (PXE enabled) through Fast Ethernet in **three** (copy-if-needed) to **five** (full-copy to hard disk) minutes using one SLIM server
- Construct 256 standalone grid points take longer time to construct. The overhead mainly lies on the process to generate host certificates

SLIM – Ongoing and future work

- SLIM has been managing:
  - the HKU CS grid point (350 nodes) for various grid research projects
  - an addition 300+ lab machines for teaching purpose (different courses have different requirements)
- Future work
  - To overcome the challenges in deploying SLIM over broadband links for realizing the "pervasive utility computing"

SLIM/InstantGrid – Key references

- R.S.C. Ho, C.M. Lee, D.H.F. Hung, C.L. Wang, and F.C.M. Lau, "Managing Execution Environments for Utility Computing," *Network Research Workshop, APAN 2004*, July, 2004

- R.S.C. Ho, K.K. Yin, C.L. Wang, and F.C.M. Lau, "InstantGrid: A Framework for Automatic Grid Point Construction," The International Workshop on Grid and Cooperative Computing (GCC 2004), Oct 21-24, 2004, Wuhan, China.

- Download: http://slim.cs.hku.hk/

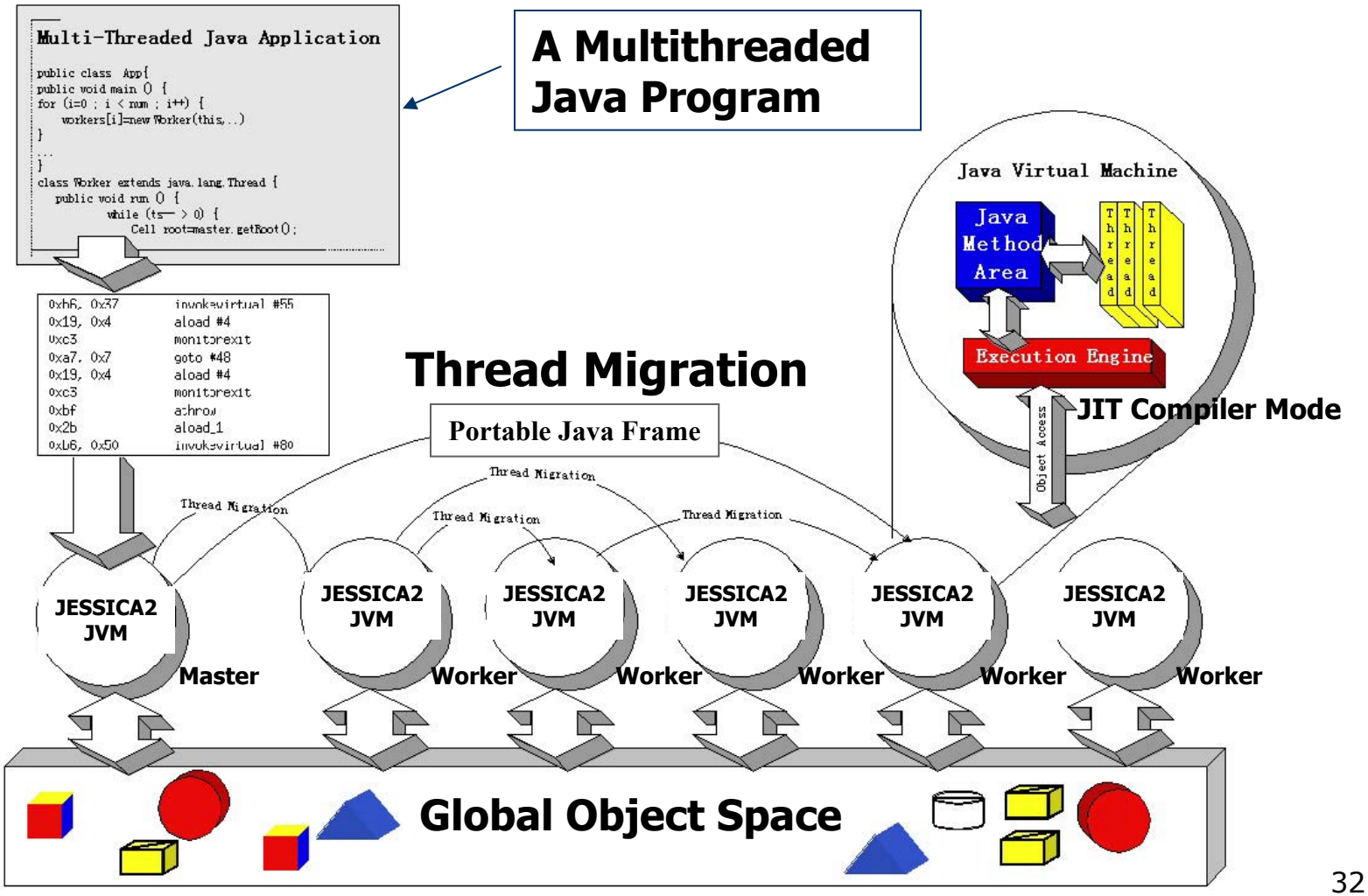
JESSICA2

"Java-Enabled Single-System-Image Computing Architecture", project started in 1996. First version (JESSICA1) in 1999

JESSICA2

- **JESSICA2 is a *distributed Java Virtual Machine* (DJVM) which consists of a group of extended JVMs running on a distributed environment to support true parallel execution of a multithreaded Java application.**
- **Java threads can freely move across node boundaries and execute in parallel to achieve more scalable high-performance computing.**
- **The JESSICA2 DJVM provides standard JVM services, that are compliant with the Java language specification, as if running on a single machine – Single System Image (SSI).**

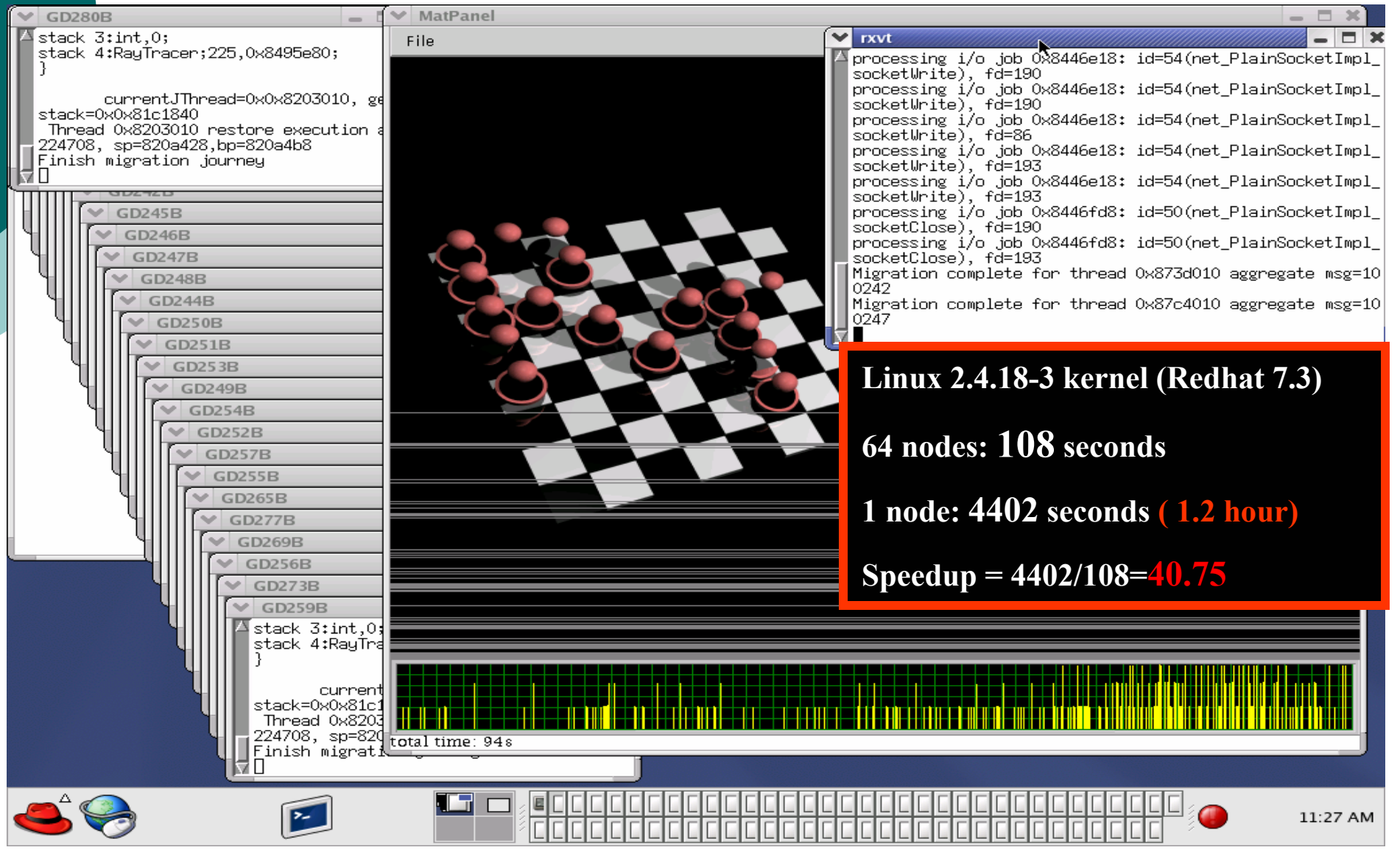
JESSICA2 Architecture



JESSICA2 Main Features

- **Transparent Java thread migration**
  - Runtime capturing and restoring of thread execution context.
  - No source code modification; no bytecode instrumentation (preprocessing); no new API introduced
  - Enable dynamic load balancing
- **Full Speed Computation**
  - JITEE: cluster-aware bytecode execution engine
  - Operated in Just-In-Time (JIT) compilation mode
  - Zero cost if no migration
- **Transparent Remote Object Access**
  - Global Object Space : A shared global heap spanning all running nodes
  - Adaptive migrating home protocol for memory consistency + various optimizing schemes.
  - I/O redirection

Ray Tracing on JESSICA2 (64 PCs)



Linux 2.4.18-3 kernel (Redhat 7.3)

64 nodes: 108 seconds

1 node: 4402 seconds (1.2 hour)

Speedup = 4402/108=40.75

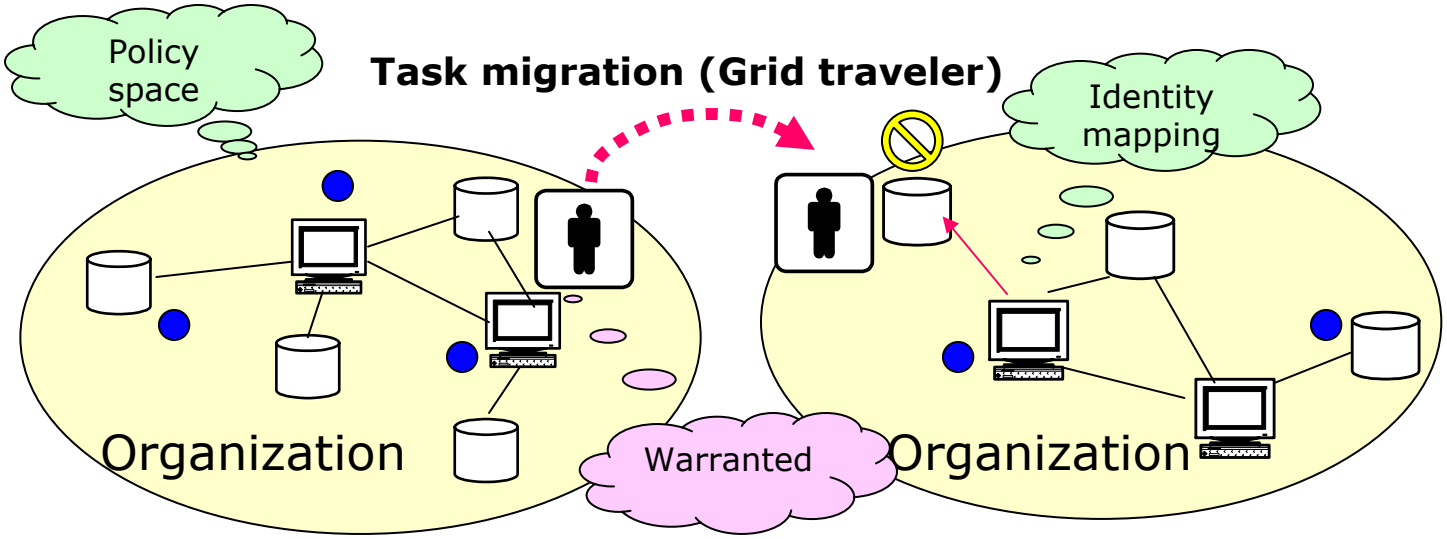
JESSICA – Key references

- Wenzhang Zhu, Weijian Fang, Cho-Li Wang, and Francis C.M. Lau, ``High Performance Computing on Clusters : The Distributed JVM Approach,'' to appear in *High Performance Computing: Paradigm and Infrastructure*, John Wiley & Sons, Inc. 2004.
- W.Z. Zhu , C.L. Wang, and F.C.M. Lau "A Lightweight Solution for Transparent Java Thread Migration in Just-in-Time Compilers," *The 2003 International Conference on Parallel Processing (ICPP-2003)*, pp. 465-472, Taiwan, Oct. 6-10, 2003
- W.Z. Zhu, C.L. Wang and F.C.M. Lau, "JESSICA2: A Distributed Java Virtual Machine with Transparent Thread Migration Support," *IEEE Fourth International Conference on Cluster Computing (CLUSTER 2002)*, Chicago, USA, September 23-26, 2002, pp. 381-388.
- M.J.M. Ma, C.L. Wang, F.C.M. Lau. "JESSICA: Java-Enabled Single-System-Image Computing Architecture," *Journal of Parallel and Distributed Computing*, Vol. 60, No. 10, October 2000, pp. 1194-1222.

JESSICA2 in CNGrid : http://147.8.179.124:8080

G-JavaMPI

A grid-enabled Java-MPI system with dynamic load-balancing via process migration



G-JavaMPI

- **A grid middleware that supports portable messaging-passing programming for achieving dynamic load-balancing and non-stop parallel computing in grid.**
- **Special feature: Transparent Java process migration**
  - State capturing and restoration through JVM Debugger Interface (JVMDI). No modification of JVM
  - Facilitates more flexible task scheduling and more effective resource sharing. Avoid running hotspots.
- **G-PASS: security enhancement for G-JavaMPI**
  - Perform identity mapping and access control while Java processes move across multiple grid points that are under different control policies. Avoid chain-delegation.
- **Migration policies :**
  - Grid point CPU and network workload
  - Application's communication pattern
  - Scheduled down time
  - Data location

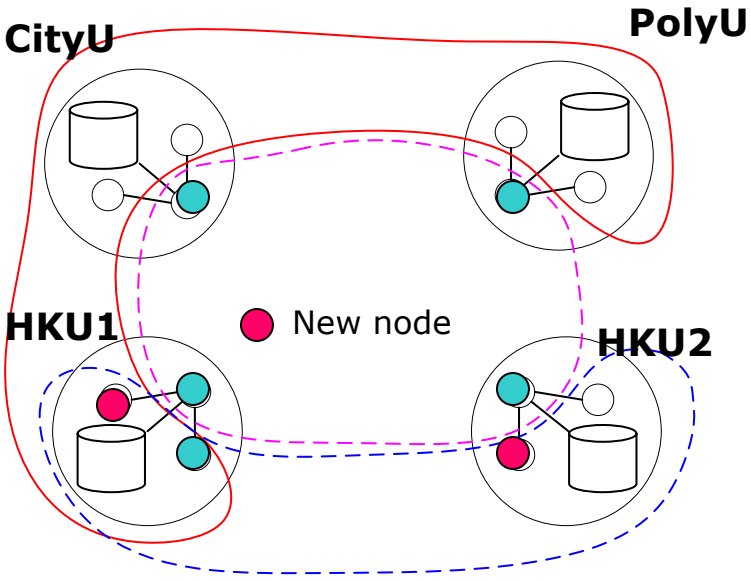
Preliminary Results at HKGrid

- **Parallel BLAST program implemented by G-JavaMPI**
  - Three universities sharing CPU cycles and local bio-databases
  - Executing 3 Blastp programs concurrently, total 18 processes
  - Original no. of nodes: 5; 2 nodes join then 2 nodes quit
- The size of the migrated execution context is about 2.1 Kbytes.
- Total execution time : 566~911 seconds under different scheduling policies.

Migration Overhead Analysis

	HKU-PolyU	PolyU-CityU	HKU-CityU
G-PASS	1.21s	0.51s	0.43s
Migration	1.90s	1.67s	0.46s
Total	3.112	2.18s	0.89s

Single process migration is **less than 0.5%** of the total execution time under different CPU load.



G-JavaMPI – Key references

- Lin Chen, Tianchi Ma, Cho-Li Wang, Francis C.M. Lau, and Shanping Li, ``G-JavaMPI: A Grid Middleware for Transparent MPI Task Migration,'' to appear in *Engineering the Grid: Status and Perspective*, Nova Science Publisher.
- Tianchi Ma, Lin Chen, Cho-Li Wang, and Francis C.M. Lau, ``G-PASS: Security Infrastructure for Grid Travelers, The International Workshop on Grid and Cooperative Computing (GCC 2004), pp. 301-308, Oct 21-24, 2004, Wuhan, China.
- L. Chen, C.L. Wang, and F.C.M. Lau, "A Grid Middleware for Distributed Java Computing with MPI Binding and Process Migration Supports," Journal of Computer Science and Technology (China), Vol. 18, No. 4, July 2003, pp. 505-514.
- Ricky K. K. Ma, Cho-Li Wang, and Francis C. M. Lau, ``M-JavaMPI : A Java-MPI Binding with Process Migration Support,'' The Second IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2002), Berlin, Germany.

Summary

- **Performance**
  - SLIM and InstantGrid: for high-speed construction of Grid computing environment, establish extensible grid platforms
  - G-JavaMPI and JESSICA : Process/thread migration enables performance optimization and load balancing
- **Reliability**
  - Java checkpointing (G-JavaMPI and JESSICA)
  - SLIM helps construct platforms for failover
- **Convenience**
  - G-JavaMPI and JESSICA enable users to utilize HPC facilities distributed across institutions via traditional means (e.g., message passing, Java)
  - SLIM and InstantGrid fulfill on-demand Grid point construction, and simplify Grid point management.

Conclusion

- Grid/utility computing are relatively new paradigms that deserve further investigation
- We address the performance, reliability, and user convenience issues in grid/utility computing
- Our advanced grid computing platform (consisting of G-JavaMPI/G-PASS, JESSICA2, and SLIM/InstantGrid) is geared to deploy in the HKGrid for easy adoption of Grid technologies.

Thanks!

For more information:

The HKU Systems Research Group
http://www.srg.csis.hku.hk/

Hong Kong Grid
http://www.hkgrid.org/

Grid Computing Research Portal
http://grid.csis.hku.hk/