香港大学结点建设

课题名称：*网格自适应服务技术研究*

2005年12月香港大学网格结点正式挂牌

狄　盛 (香港大学)
王卓立教授(香港大学)
依托单位：　清华大学

Dec. 3, 2007

Outline

- 总体目标
- 目前进度
- 主要研究进展

Objectives

- 高性能环境支持
  - 课题1:扩大和升级香港大学网格结点的资源
  - 课题2:实现自适应的Java计算环境
  - 课题3:基于SLIM-VM 的网格部署方案
  - 课题4:普适化智能网格接入
- 四大应用
  - 应用1:基于Tomcat的商业Web应用
  - 应用2:AlphaMiner (智能商业数据挖掘)
  - 应用3:物理网格(Physics Grid)
  - 应用4:并行图像渲染（风格化油画生成）

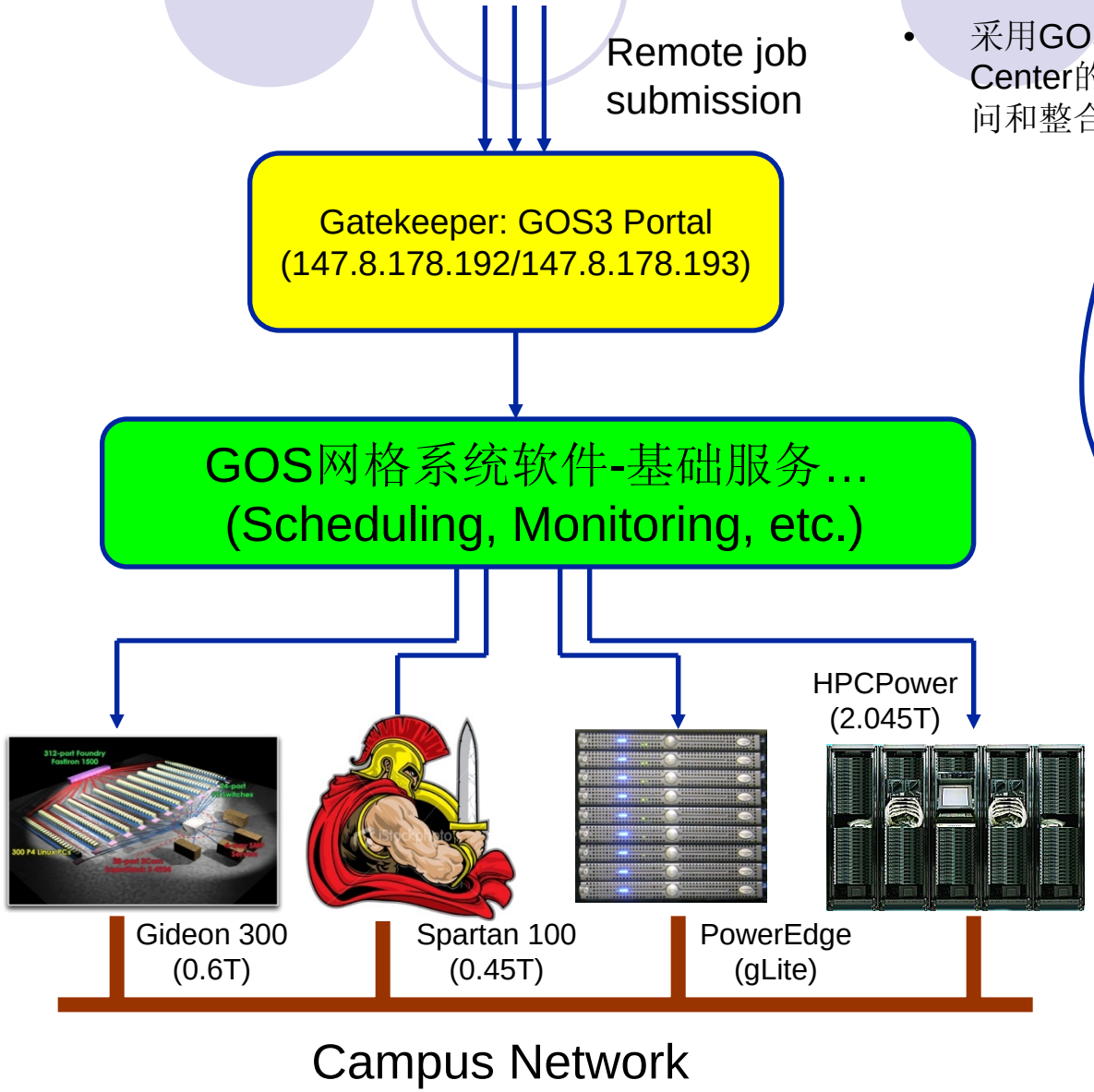
香港大学最新资源列表 (2007, Dec.1)

资源	单位	结点数目	系统配置	网络状况	理论计算峰值
HPC Power Cluster	HKUCC	178	356 Xeon processors,2GB RAM, 40GB IDE硬盘	Gigabit Ethernet (CISCO 4506 switch x 2)	2.045Tflops
Gideon 300 Cluster	HKUCS	300	Pentium 4@ 2.0 GHz, 512-2GB RAM, 40GB IDE 硬盘	Fast-Ethernet	0.6 Tflops
Dell PowerEdge	HKUCS	20	Intel P3 1400MHz , 512MB RAM /node, 37GB SCSI disk	Fast Ethernet x 4 per node	29.4 Gflops
Spartan 100 (2008年1月)	HKUCS	100	Intel P4 2.26GHz, 512MB RAM, 40GB hard disk	Gigabit Ethernet	452 Gflops

理论计算峰值: (2.045 + 0.6 + 0.0294 +0.452)Tflops = **3.126Tflops**
2008年底的理论计算峰值有希望再增加**2Tflops** (含200个结点),总的理论结算峰值有望达到: **5.126Tflops**

(课题重点1) 扩大和升级香港大学网格结点的资源

- 采用GOS3将CS Department和Computer Center的计算资源连接起来,实现资源透明化访问和整合. 开发支持GOS3的高性能网格应用.



Grid middleware

-**GOS 3.0**
-**gLite**

Job scheduling

-**基础服务:批作业服务**
-**OpenPBS**

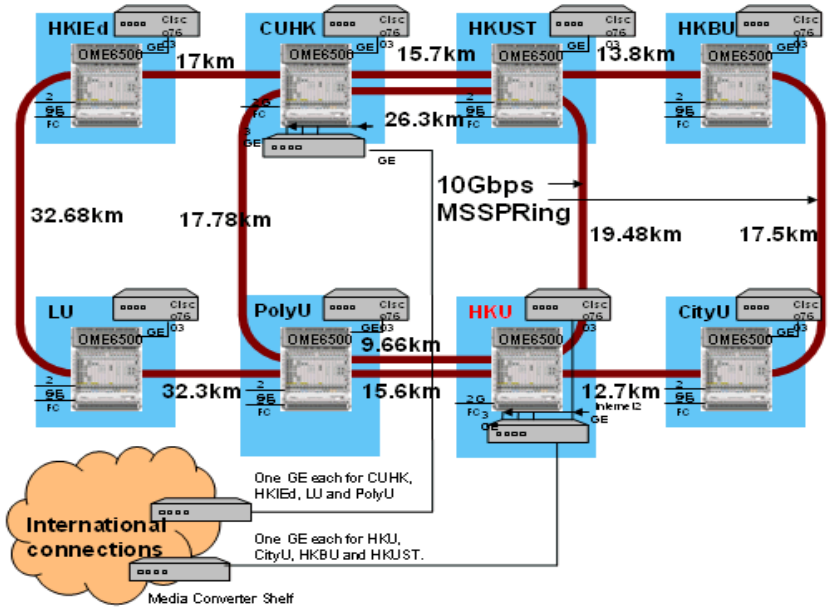
Programming

-**HPF, Fortran 90**
-**C, C++, Java with MPI**
-**JESSICA3 DJVM (HKU)**

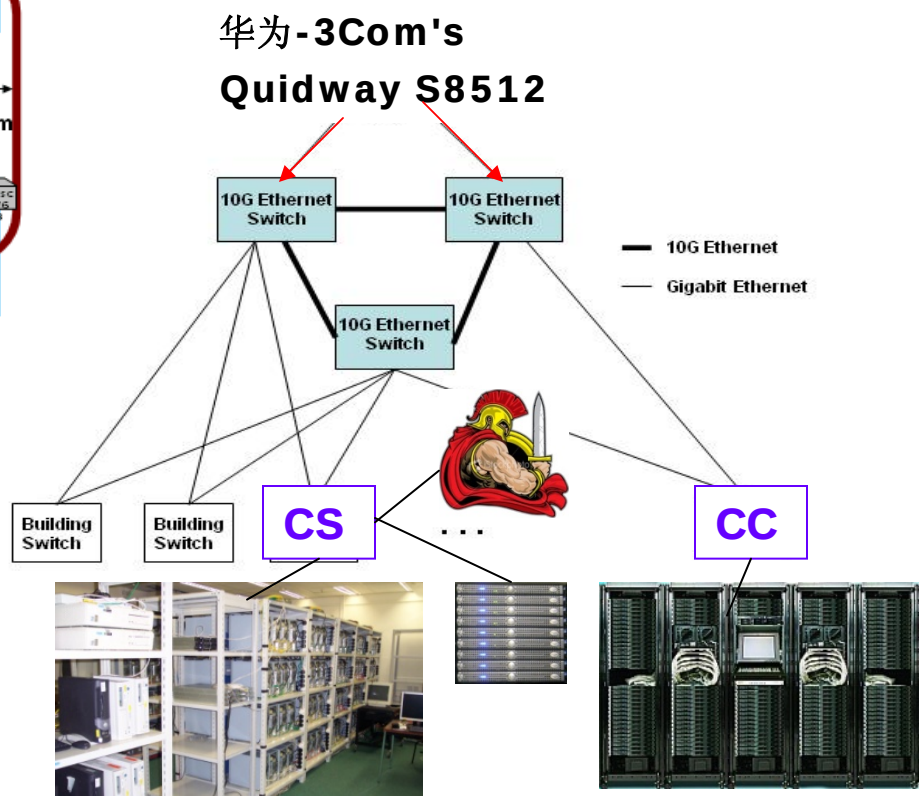
Communication Lib

-**MPICH-G2 1.2.3**

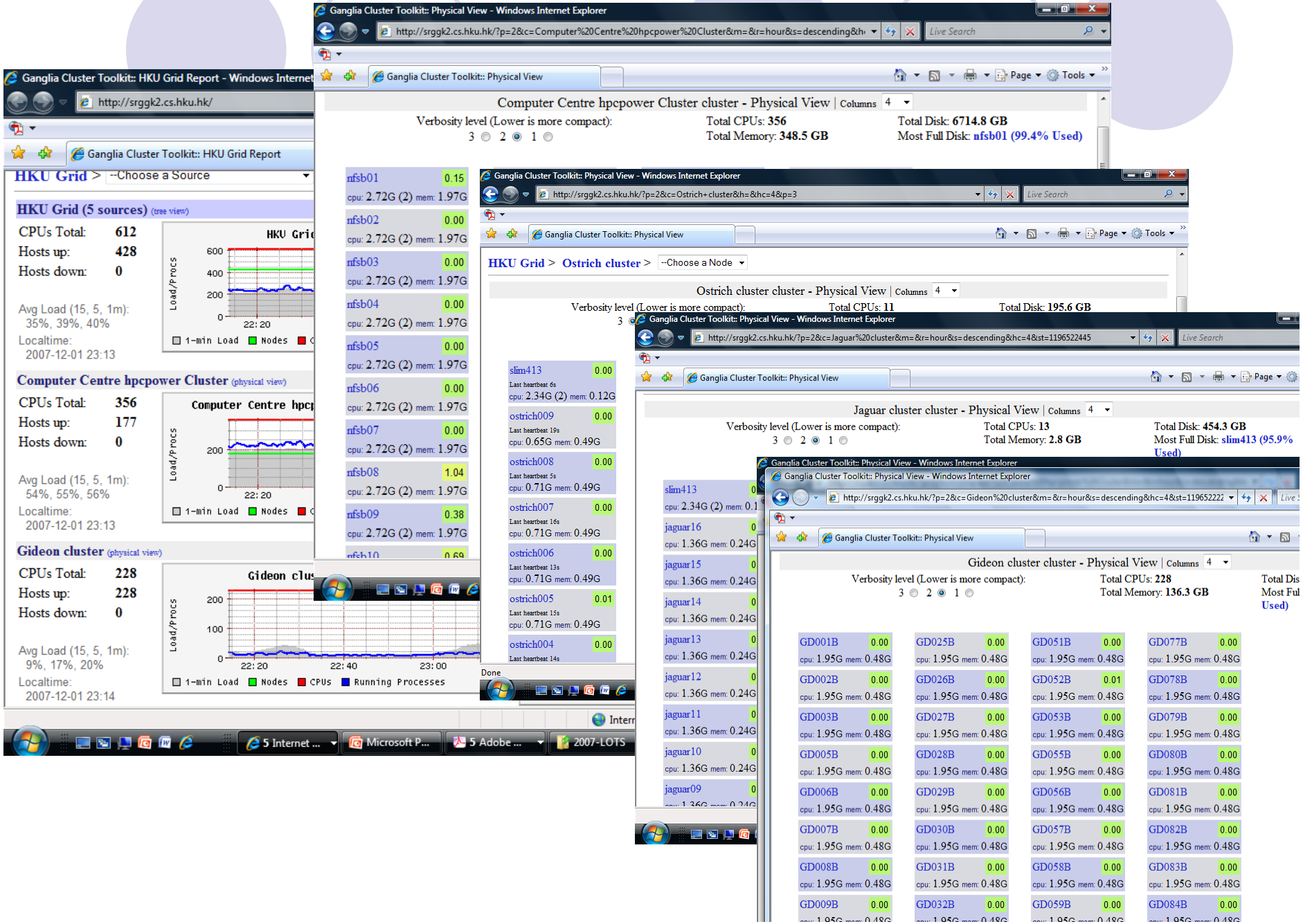
香港教育主干网：HARNET
(香港学术研究机构网)



港大校园主干网已提升为 **10Gbs** 以太网 **(Oct. 2006).**

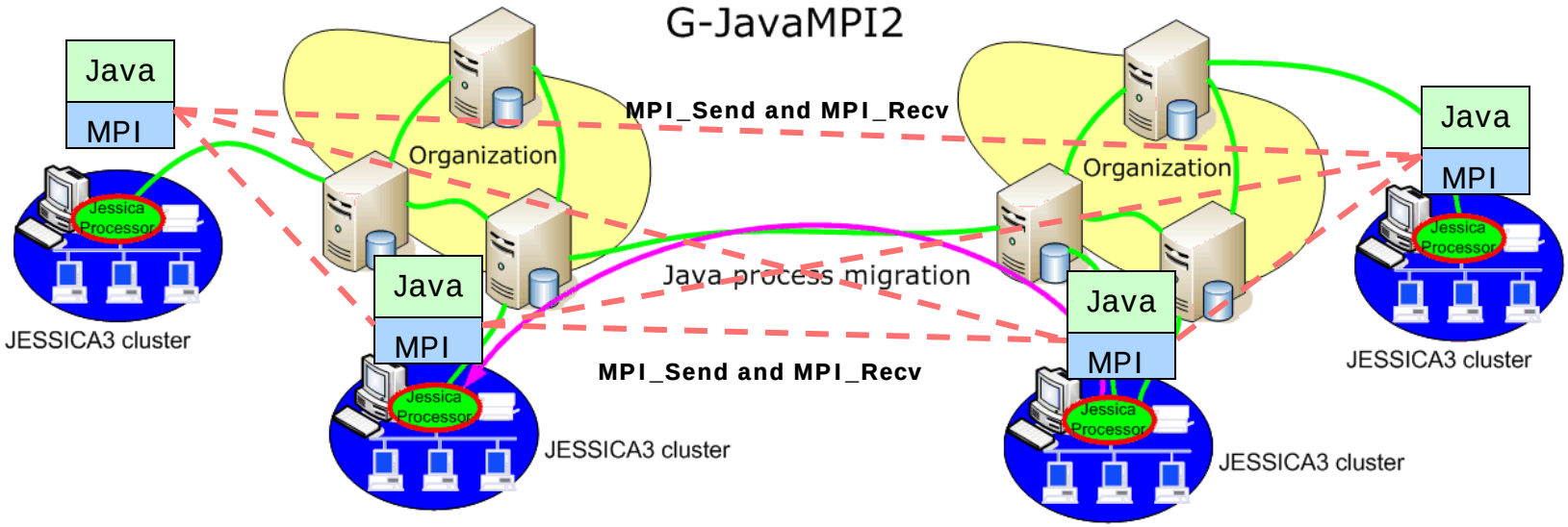


- HARNET 在香港的大学之间提供 10G 网络互联(Oct. 2006)
- 香港大学是香港教育主干网连接到国际网的交换结点。


Ganglia Cluster Toolkit:: HKU Grid Report - Windows Internet
http://srggk2.cs.hku.hk/
Ganglia Cluster Toolkit:: HKU Grid Report
HKU Grid > --Choose a Source
HKU Grid (5 sources) (tree view)
CPUs Total: 612
Hosts up: 428
Hosts down: 0
Avg Load (15, 5, 1m): 35%, 39%, 40%
Localtime: 2007-12-01 23:13
Computer Centre hpcpower Cluster (physical view)
CPUs Total: 356
Hosts up: 177
Hosts down: 0
Avg Load (15, 5, 1m): 54%, 55%, 56%
Localtime: 2007-12-01 23:13
Gideon cluster (physical view)
CPUs Total: 228
Hosts up: 228
Hosts down: 0
Avg Load (15, 5, 1m): 9%, 17%, 20%
Localtime: 2007-12-01 23:14
Load/Procs
1-min Load
Nodes
CPUs
Running Processes
Ganglia Cluster Toolkit:: Physical View - Windows Internet Explorer
http://srggk2.cs.hku.hk/?p=2&c=Computer%20Centre%20hpcpower%20Cluster&m=&r=hour&s=descending&h
Ganglia Cluster Toolkit:: Physical View
Computer Centre hpcpower Cluster cluster - Physical View | Columns 4
Verbosity level (Lower is more compact): 3 2 1
Total CPUs: 356
Total Memory: 348.5 GB
Total Disk: 6714.8 GB
Most Full Disk: nfsb01 (99.4% Used)
nfsb01 0.15 cpu: 2.72G (2) mem: 1.97G
nfsb02 0.00 cpu: 2.72G (2) mem: 1.97G
nfsb03 0.00 cpu: 2.72G (2) mem: 1.97G
nfsb04 0.00 cpu: 2.72G (2) mem: 1.97G
nfsb05 0.00 cpu: 2.72G (2) mem: 1.97G
nfsb06 0.00 cpu: 2.72G (2) mem: 1.97G
nfsb07 0.00 cpu: 2.72G (2) mem: 1.97G
nfsb08 1.04 cpu: 2.72G (2) mem: 1.97G
nfsb09 0.38 cpu: 2.72G (2) mem: 1.97G
Ganglia Cluster Toolkit:: Physical View - Windows Internet Explorer
http://srggk2.cs.hku.hk/?p=2&c=Ostrich+cluster&h=&hc=4&p=3
HKU Grid > Ostrich cluster > --Choose a Node
Ostrich cluster cluster - Physical View | Columns 4
Verbosity level (Lower is more compact): 3
Total CPUs: 11
Total Disk: 195.6 GB
slim413 0.00 Last heartbeat 6s cpu: 2.34G (2) mem: 0.12G
ostrich009 0.00 Last heartbeat 19s cpu: 0.65G mem: 0.49G
ostrich008 0.00 Last heartbeat 5s cpu: 0.71G mem: 0.49G
ostrich007 0.00 Last heartbeat 16s cpu: 0.71G mem: 0.49G
ostrich006 0.00 Last heartbeat 13s cpu: 0.71G mem: 0.49G
ostrich005 0.01 Last heartbeat 15s cpu: 0.71G mem: 0.49G
ostrich004 0.00 Last heartbeat 14s
Ganglia Cluster Toolkit:: Physical View - Windows Internet Explorer
http://srggk2.cs.hku.hk/?p=2&c=Jaguar%20cluster&m=&r=hour&s=descending&hc=4&st=1196522445
Jaguar cluster cluster - Physical View | Columns 4
Verbosity level (Lower is more compact): 3 2 1
Total CPUs: 13
Total Memory: 2.8 GB
Total Disk: 454.3 GB
Most Full Disk: slim413 (95.9% Used)
slim413 0 cpu: 2.34G (2) mem: 0.1
jaguar16 0 cpu: 1.36G mem: 0.24G
jaguar15 0 cpu: 1.36G mem: 0.24G
jaguar14 0 cpu: 1.36G mem: 0.24G
jaguar13 0 cpu: 1.36G mem: 0.24G
jaguar12 0 cpu: 1.36G mem: 0.24G
jaguar11 0 cpu: 1.36G mem: 0.24G
jaguar10 0 cpu: 1.36G mem: 0.24G
jaguar09 0
Ganglia Cluster Toolkit:: Physical View - Windows Internet Explorer
http://srggk2.cs.hku.hk/?p=2&c=Gideon%20cluster&m=&r=hour&s=descending&hc=4&st=119652222
Gideon cluster cluster - Physical View | Columns 4
Verbosity level (Lower is more compact): 3 2 1
Total CPUs: 228
Total Memory: 136.3 GB
GD001B 0.00 cpu: 1.95G mem: 0.48G
GD025B 0.00 cpu: 1.95G mem: 0.48G
GD051B 0.00 cpu: 1.95G mem: 0.48G
GD077B 0.00 cpu: 1.95G mem: 0.48G
GD002B 0.00 cpu: 1.95G mem: 0.48G
GD026B 0.00 cpu: 1.95G mem: 0.48G
GD052B 0.01 cpu: 1.95G mem: 0.48G
GD078B 0.00 cpu: 1.95G mem: 0.48G
GD003B 0.00 cpu: 1.95G mem: 0.48G
GD027B 0.00 cpu: 1.95G mem: 0.48G
GD053B 0.00 cpu: 1.95G mem: 0.48G
GD079B 0.00 cpu: 1.95G mem: 0.48G
GD005B 0.00 cpu: 1.95G mem: 0.48G
GD028B 0.00 cpu: 1.95G mem: 0.48G
GD055B 0.00 cpu: 1.95G mem: 0.48G
GD080B 0.00 cpu: 1.95G mem: 0.48G
GD006B 0.00 cpu: 1.95G mem: 0.48G
GD029B 0.00 cpu: 1.95G mem: 0.48G
GD056B 0.00 cpu: 1.95G mem: 0.48G
GD081B 0.00 cpu: 1.95G mem: 0.48G
GD007B 0.00 cpu: 1.95G mem: 0.48G
GD030B 0.00 cpu: 1.95G mem: 0.48G
GD057B 0.00 cpu: 1.95G mem: 0.48G
GD082B 0.00 cpu: 1.95G mem: 0.48G
GD008B 0.00 cpu: 1.95G mem: 0.48G
GD031B 0.00 cpu: 1.95G mem: 0.48G
GD058B 0.00 cpu: 1.95G mem: 0.48G
GD083B 0.00 cpu: 1.95G mem: 0.48G
GD009B 0.00
GD032B 0.00
GD059B 0.00
GD084B 0.00
Done
5 Internet ...
Microsoft P...
5 Adobe ...
2007-LOTS

(课题重点2)实现自适应的Java计算环境 G-JavaMPI Version 2.0

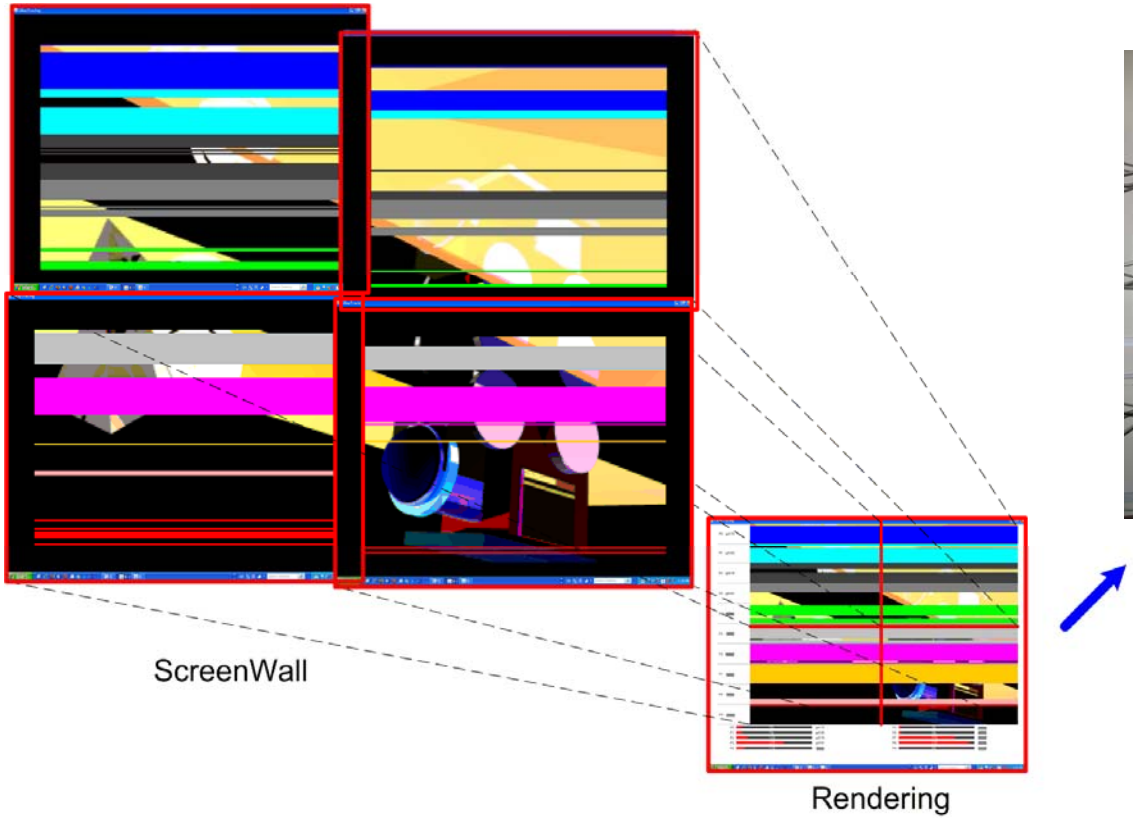
- 整合 G-JavaMPI 和 JESSICA3
- 轻量级Java进程/线程迁移
- 使用Java虚拟机工具接口java Virtual Machine Tool Interface (JVMTI)



迁移策略: 解决结点负载超载问题,解决不可避免的可能的不可预料的管理因素，为实现更优的资源调度，任务可能重新映射.

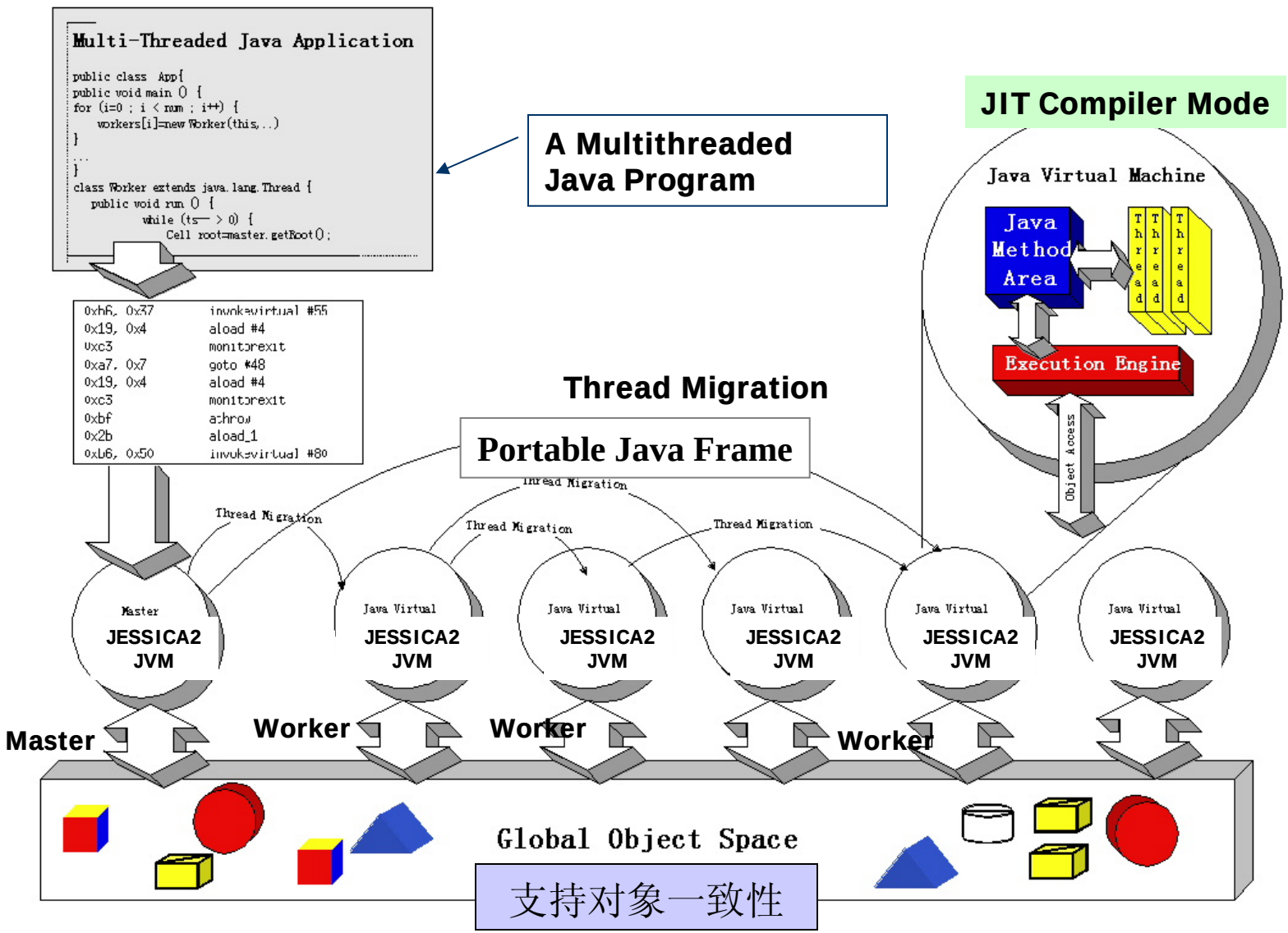
香港大学开放日——中国国家网格操作演示

HKU Open Day: G-JavaMPI Screen Wall演示，Oct. 27&28, 2007



Result

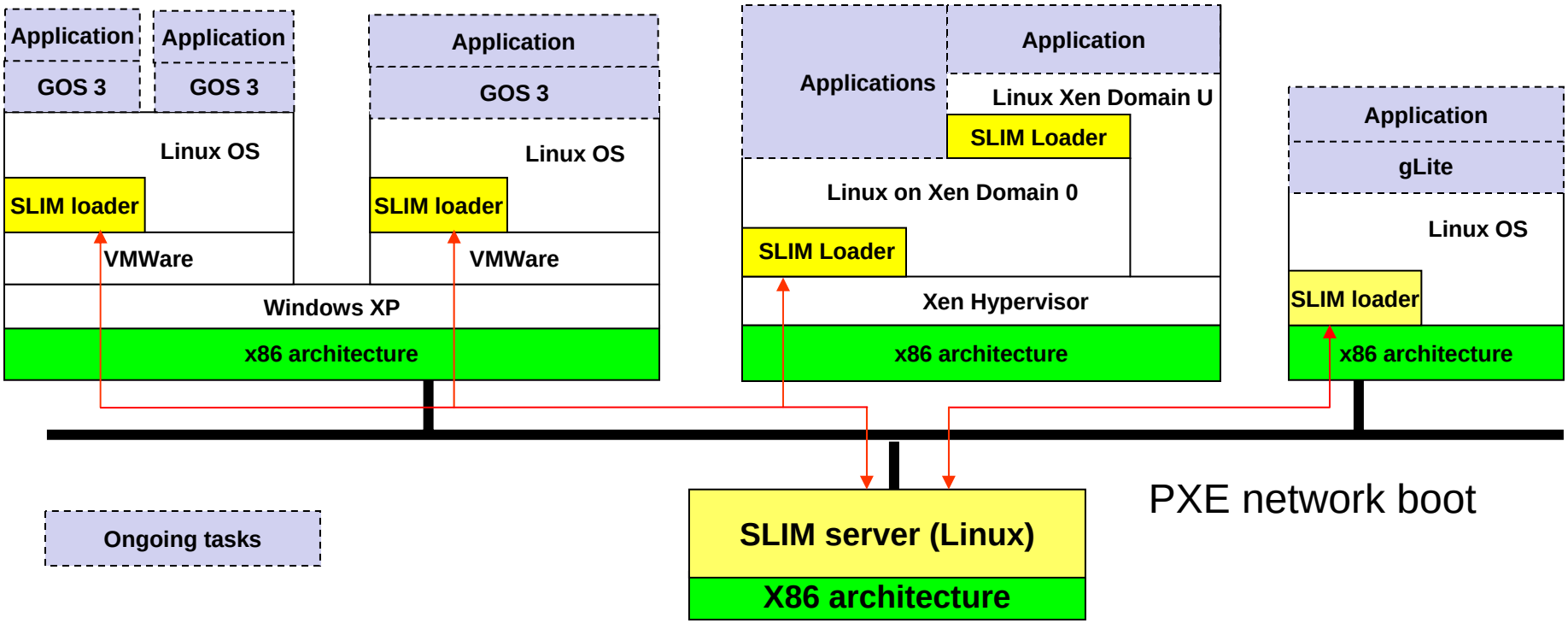
JESSICA2 体系结构



Java
Enabled
Single
System
Image
Computing
Architecture

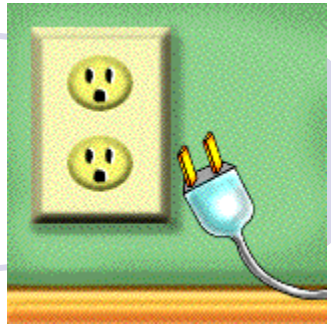
(课题重点3)基于SLIM-VM 网格部署

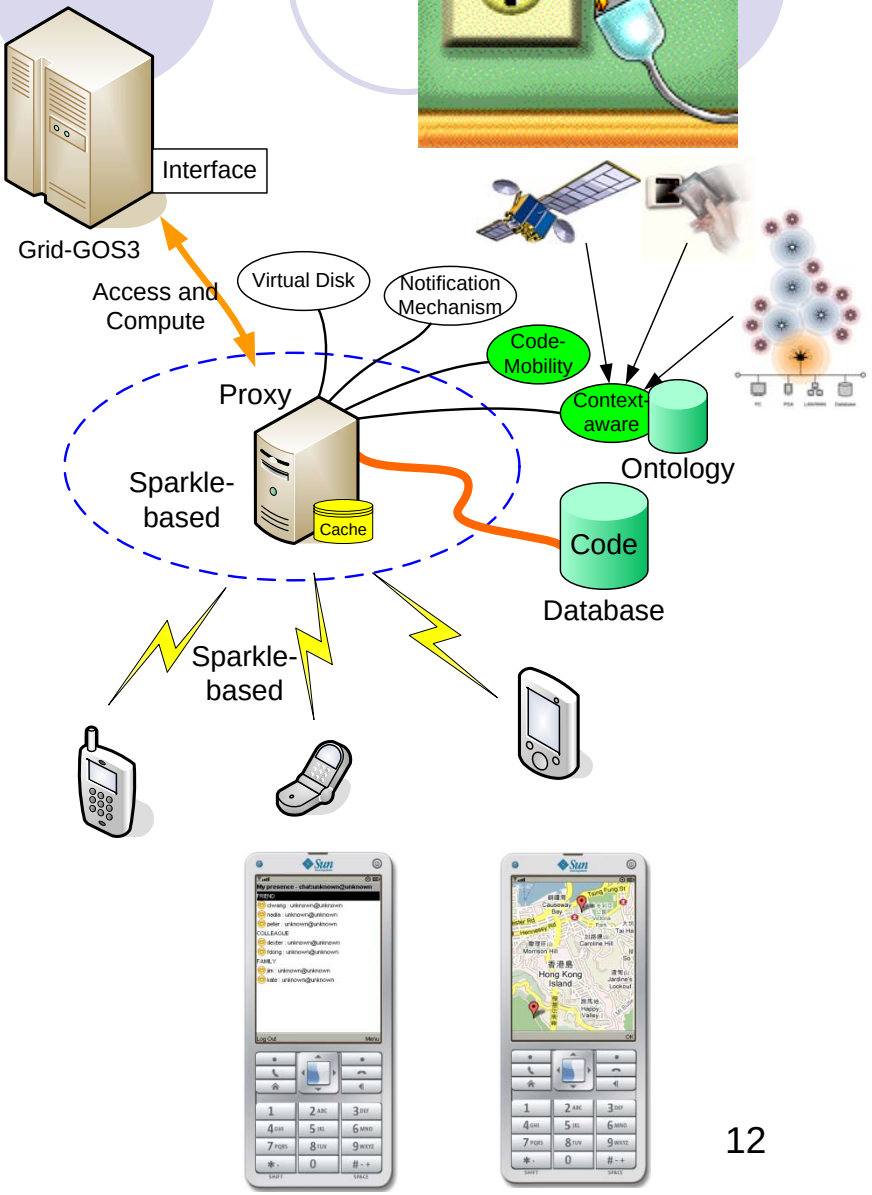
1. 快速安装操作系统镜像
2. 网格虚拟化: 在一个计算结点上同时执行各种不同网格的计算任务.在最大的范围内实现资源的高效共享



(课题重点4)普适化智能网格接入

- 越来越多的访问点 (WiFi) ➔用户在任何地点任合时刻都可以方便无缝的访问网格。
- 即插即用的网格服务。





普适化智能网格接入 (Grid Mobile Access)

Context Awareness:

- Ontology-based Context Modeling
- Location Awareness (GPS + RFID)

Mobile Code :

- Dynamic Grid stub downloading
- Intelligent cache proxy (HKU Sparkle project)

Other Features (not available yet)

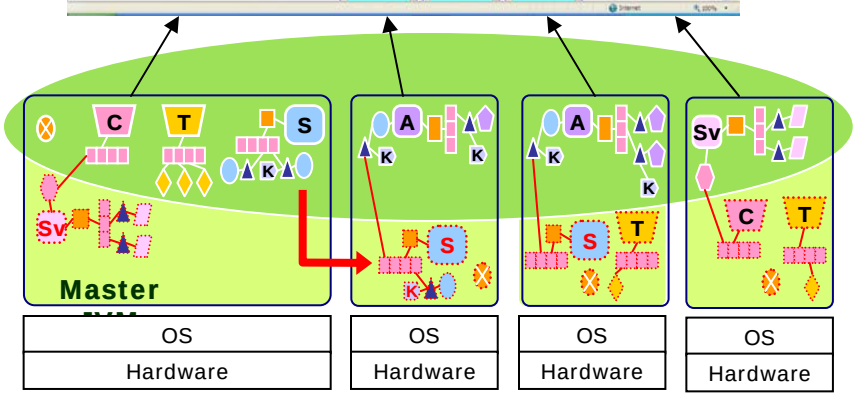
- Virtual Disk
- Notification Mechanism

目前已经利用Context-aware成功实现了一个动态访问移动计算Demo原型

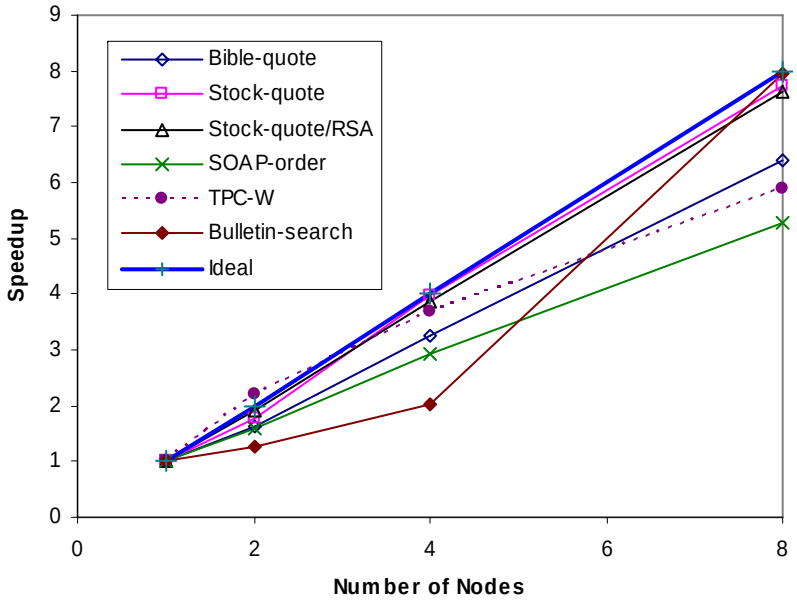
【应用1：基于Tomcat的商业Web应用】

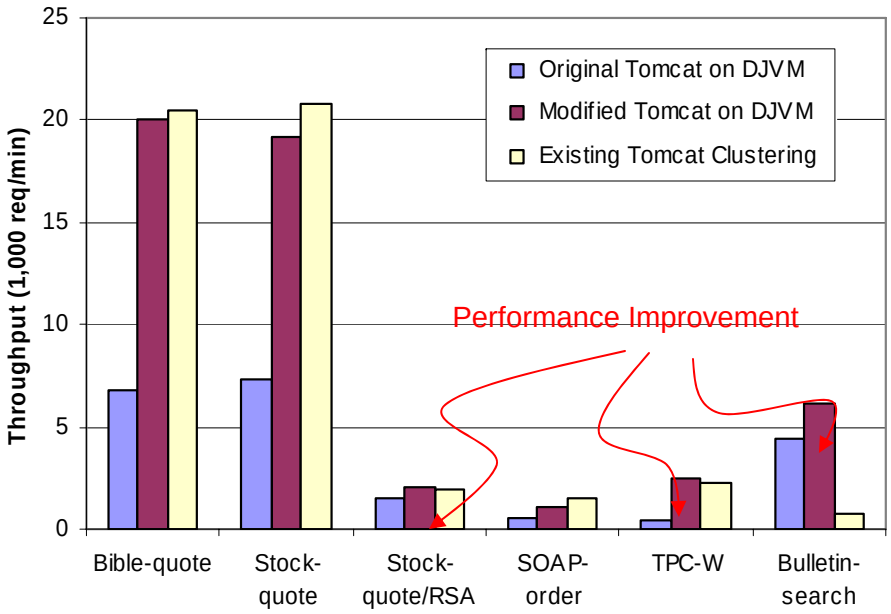
- 目标: 利用并行Tomcat及JESSICA2的细粒度线程技术实现高可靠性高性能的商业应用。
- 状态: 已经设计并开发了底层的并行分布式Tomcat环境,并实现的示范性商业应用包括实时股票报价(Stock Quotes)、电子书店(TPC-W bookstore)等。
- 演示:
  - 电子书店(TPC-W bookstore)
  - 股票报价(Stock Quotes)

电子书店



Tomcat on JESSICA2
性能评估





K. T. Lam, Y. Luo, and C. L. Wang. *"Clustering Web Application Servers on Distributed JVM: Challenges and Experiences*". In preparation.

Tomcat/JESSICA2股票报价

股票报价作业提交界

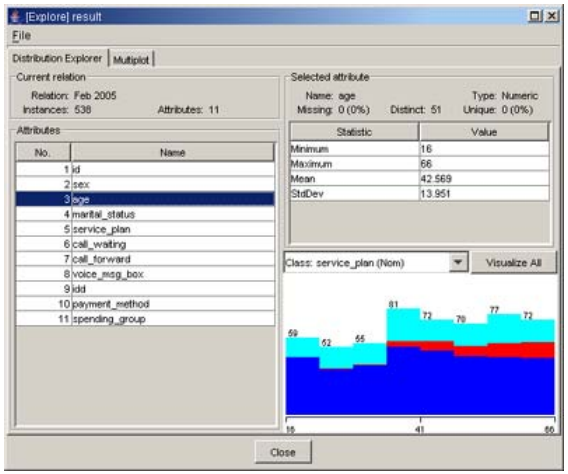
作业成功提交

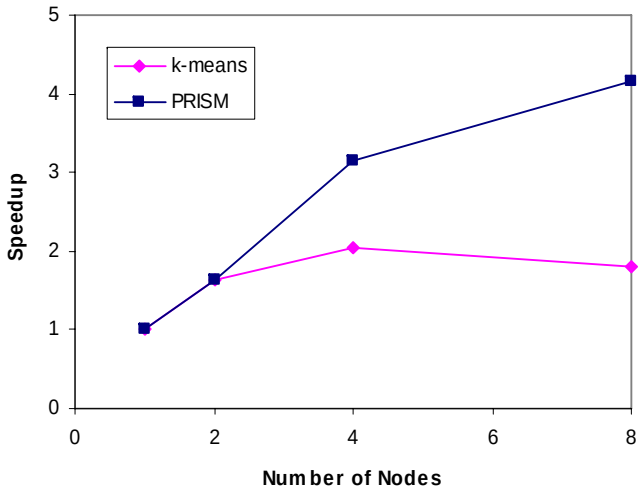
αMiner.【应用2：数据挖掘】

An Open Source Data Mining Platform

eti
e-business technology institute

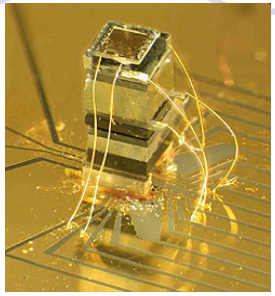
- 目标：该计划是一个科学计算数据挖掘应用，支持各种数据挖掘计算。
- 状态: 目前主要是基于Weka (Alpha Miner原始核心软件) 进行改进，实现对Decision Tree和K-means clustering 两种算法的分布式数据挖掘的分析和处理。

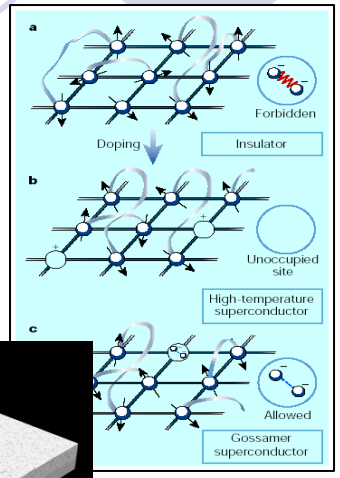


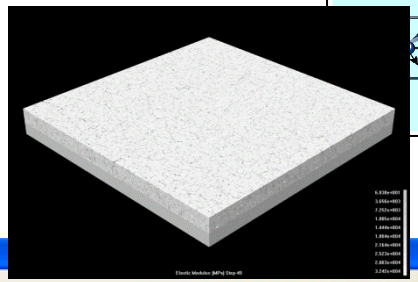


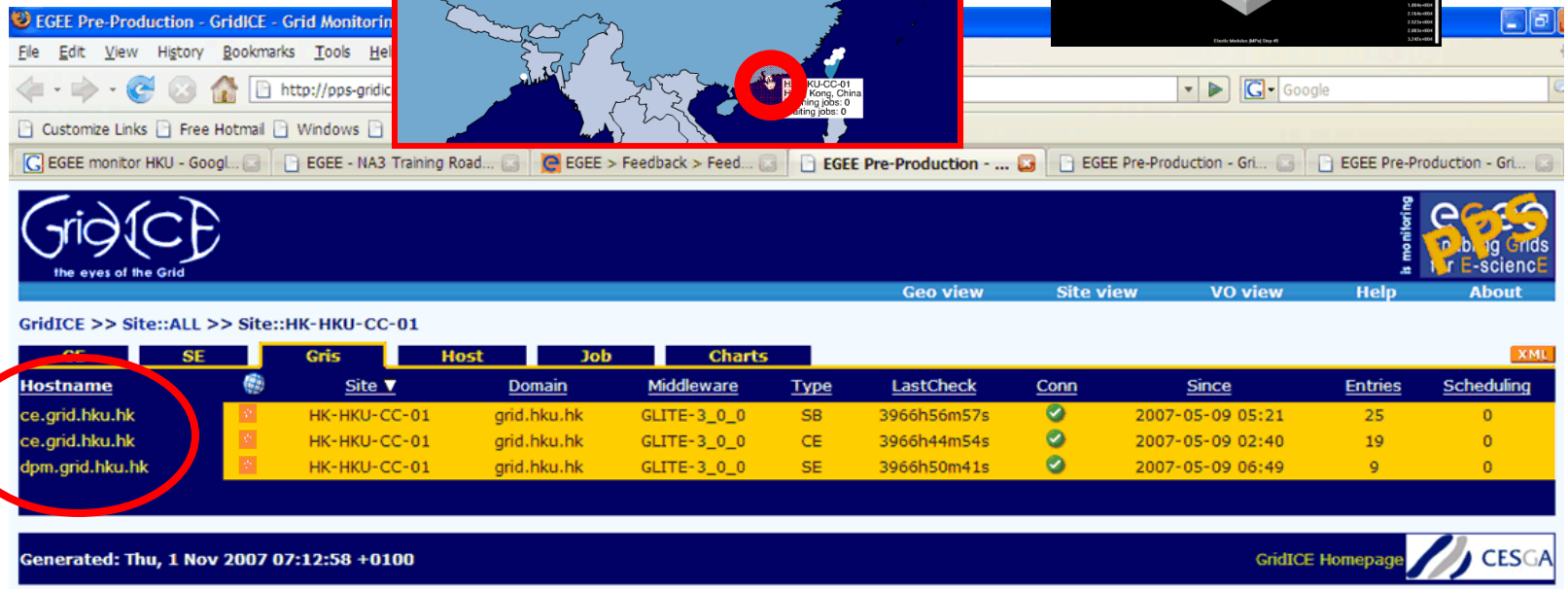
应用3：校园物理网格 (Physics Grid)

状态：计算机中心已经通过TW-CA的认证接入EGEE(gLite)。
未来计划：08年1月10号前配合OMII-CNGrid联合培训完成gLite教学平台。







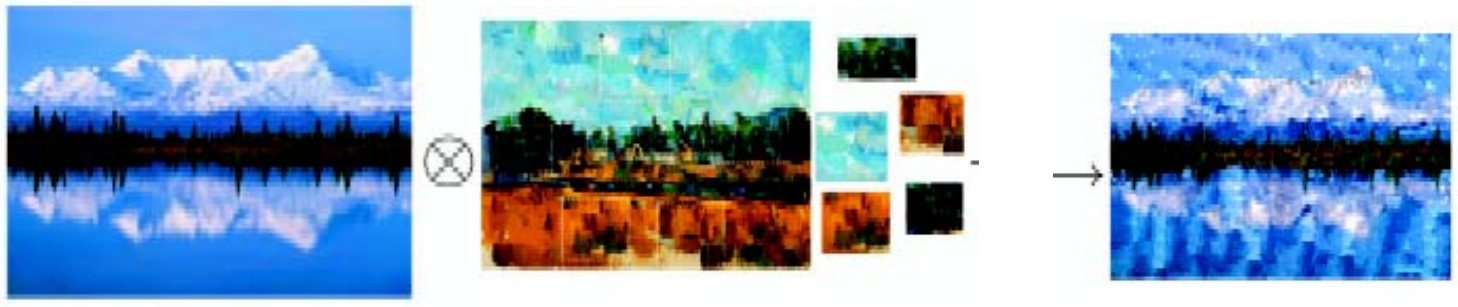


17

【应用4：风格化油画生成】
(painting synthesis with style）

面向普通用户推广发布一些有娱乐性网格应用。

可将任意的数码图像，进行重新渲染，生成具各种风格的油画版本。



通过开发高效的并行算法，让用户能够通过网格服务的形式参与，对网格计算有亲身的体会，起到良好的推广与示范作用。

人人是画家 - 我是畢卡索

主要研究进展

Thanks!

1) 实现网格环境下负载均衡的进程迁移调度策略
2) 实现Group-based的大规模信息传递系统的Checkpoint/Restart机制
3) 实现轻量级进程迁移和内存预取（Memory page prefetching）机制
4) 实现并行环境下的进程漫步（Process Roaming）机制
5) 在JESSICA2 Java虚拟机上实现Tomcat Web应用服务的集群应用环境

Publications:

1) Lin Chen, Cho-Li Wang, Francis Lau, Process Reassignment with Reduced Migration Cost in Grid Load Rebalancing, to appear in The Seventeenth International Heterogeneity in Computing Workshop (HCW), 2008.
2) Justin Ho, Cho-Li Wang, Francis Lau, Scalable Group-based Checkpoint/Restart for Large-Scale Message-passing Systems, to appear in IPDPS'08.
3) Roy Ho, Cho-Li Wang, Francis Lau, Lightweight Process Migration and Memory Prefetching on openMosix, to appear in IPDPS'08.
4) Roy Ho, Cho-Li Wang, Francis Lau, Process Roaming: Design and implementation, submitted to ICDCS'08.
5) K. T. Lam, Y. Luo, and C. L. Wang. Clustering Web Application Servers on Distributed JVM: Challenges and Experiences. In preparation.