Enabling Grids for E-sciencE

Distributed Systems Laboratory

Technion

Computational Biology Laboratory

Superlink-online

A Distributed System For Genetic Linkage Analysis using EGEE and BOINC

Mark Silberstein, Artyom Sharov, Assaf Schuster, Dan Geiger

CS Department, Technion, Haifa

www.eu-egee.org

Information Society and Media

Genetic Linkage Analysis

- **Purpose**: to obtain crude chromosomal location of gene(s) associated with a phenotype of interest
  - examples: Cystic fibrosis (found), diabetes, Alzheimer, blood pressure
- We focus on parametric linkage analysis on pedigrees

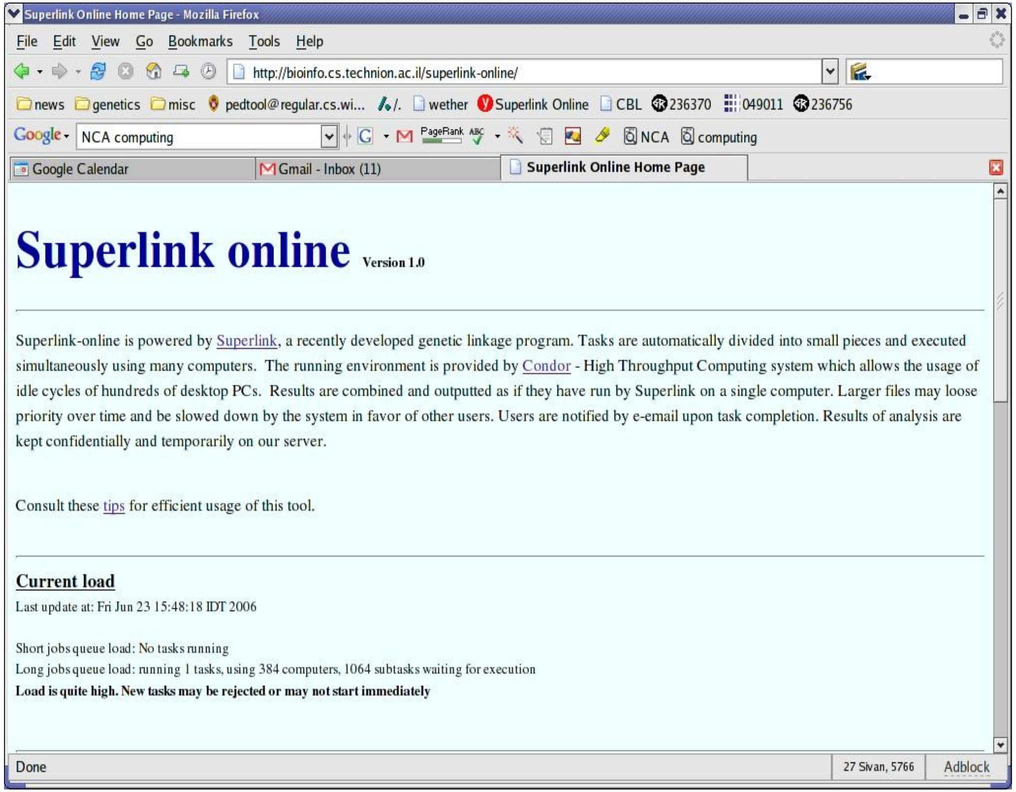
Main Problem

- **Problem:**
  - Many analyses are infeasible due to the high computational demands
- **Reason:**
  - Exponential nature due to inference in Bayesian networks
- **Solution:**
  - Split the task into substantially smaller subtasks
  - Execute subtasks on multiple CPUs

Interface to geneticists

http://bioinfo.cs.technion.ac.il/superlink-online

- **User submits his/her data for analysis**
  - No specification of running time
- **Secured user web interface**
  - Monitoring of partial results
  - Cancellation
- **E-mail notifications**


Superlink Online Home Page - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://bioinfo.cs.technion.ac.il/superlink-online/
news genetics misc pedtool@regular.cs.wi... /. wether Superlink Online CBL 236370 049011 236756
Google NCA computing PageRank NCA computing
Google Calendar
Gmail - Inbox (11)
Superlink Online Home Page

Superlink online Version 1.0

Superlink-online is powered by Superlink, a recently developed genetic linkage program. Tasks are automatically divided into small pieces and executed simultaneously using many computers. The running environment is provided by Condor - High Throughput Computing system which allows the usage of idle cycles of hundreds of desktop PCs. Results are combined and outputted as if they have run by Superlink on a single computer. Larger files may loose priority over time and be slowed down by the system in favor of other users. Users are notified by e-email upon task completion. Results of analysis are kept confidentially and temporarily on our server.

Consult these tips for efficient usage of this tool.

Current load

Last update at: Fri Jun 23 15:48:18 IDT 2006

Short jobs queue load: No tasks running

Long jobs queue load: running 1 tasks, using 384 computers, 1064 subtasks waiting for execution

Load is quite high. New tasks may be rejected or may not start immediately

Success stories

- ~14,000 tasks, > 250 CPU years utilized since 2006
- Over 30 citations in leading genetics journals
- Over 200 users from universities and research centers in US, France, Germany, UK, Italy, Austria, Spain, Taiwan, Australia, and others

Success stories

- ~14,000 tasks, > 250 CPU years utilized since 2006
- Over 30 citations in leading genetics journals
- Over 200 users from universities and research centers in US, France, Germany, UK, Italy, Austria, Spain, Taiwan, Australia, and others

Using over 3,500 computers in Israel and US
But need much more to allow comprehensive analysis of more complex data

Workload characteristics

- Few KB input/output
- **High RAM footprint**
- **Single job running time**
  - Seconds to hours (cannot be estimated exactly)
- **Single task**
  - up to **1M** jobs

Workload characteristics

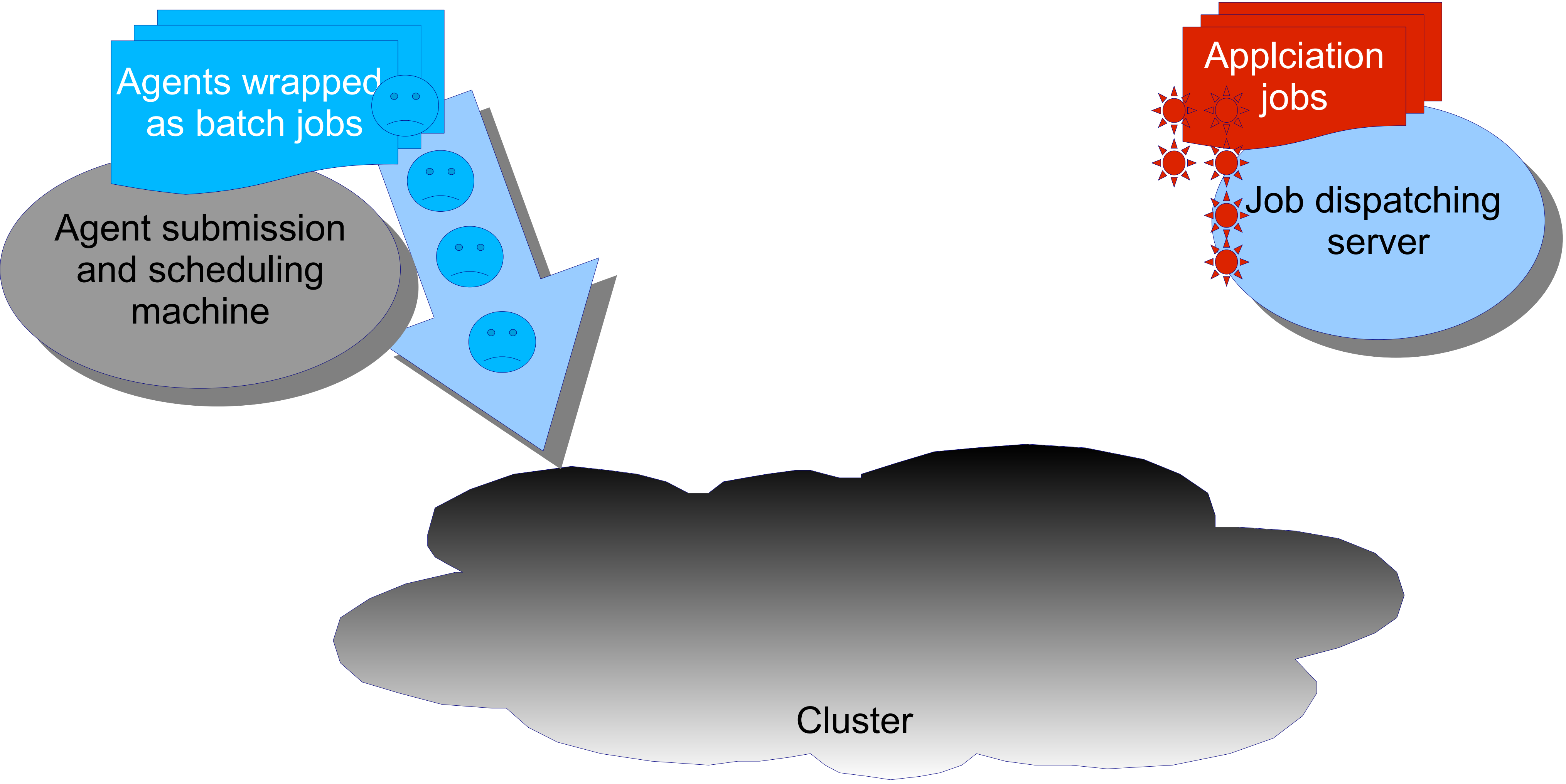
- Few KB input/output
- **High RAM footprint**
- **Single job running time**
  - Seconds to hours (cannot be estimated exactly)
- **Single task**
  - up to **1M** jobs

Challenges:
Automatic, reliable, efficient execution
Acquiring many resources without VO coordination

Our EGEE statistics

- **A typical run:**
  - 0.5M jobs from several seconds to 1 hour (~20 minutes on average)
  - accomplished within 10 days
  - up to 2000 (1300 on average) concurrently executing clients in BIOMED VO
  - **600 PFLOPs** consumed
  - **fully automated**
  - **no prior coordination** with EGEE admins
  - 19 CPU years within 10 days utilized
    - Note: WISDOM project's recent **prioritized** data challenge utilized 144.21 years within 37 days

Generic concept: virtual cluster



Agents wrapped as batch jobs

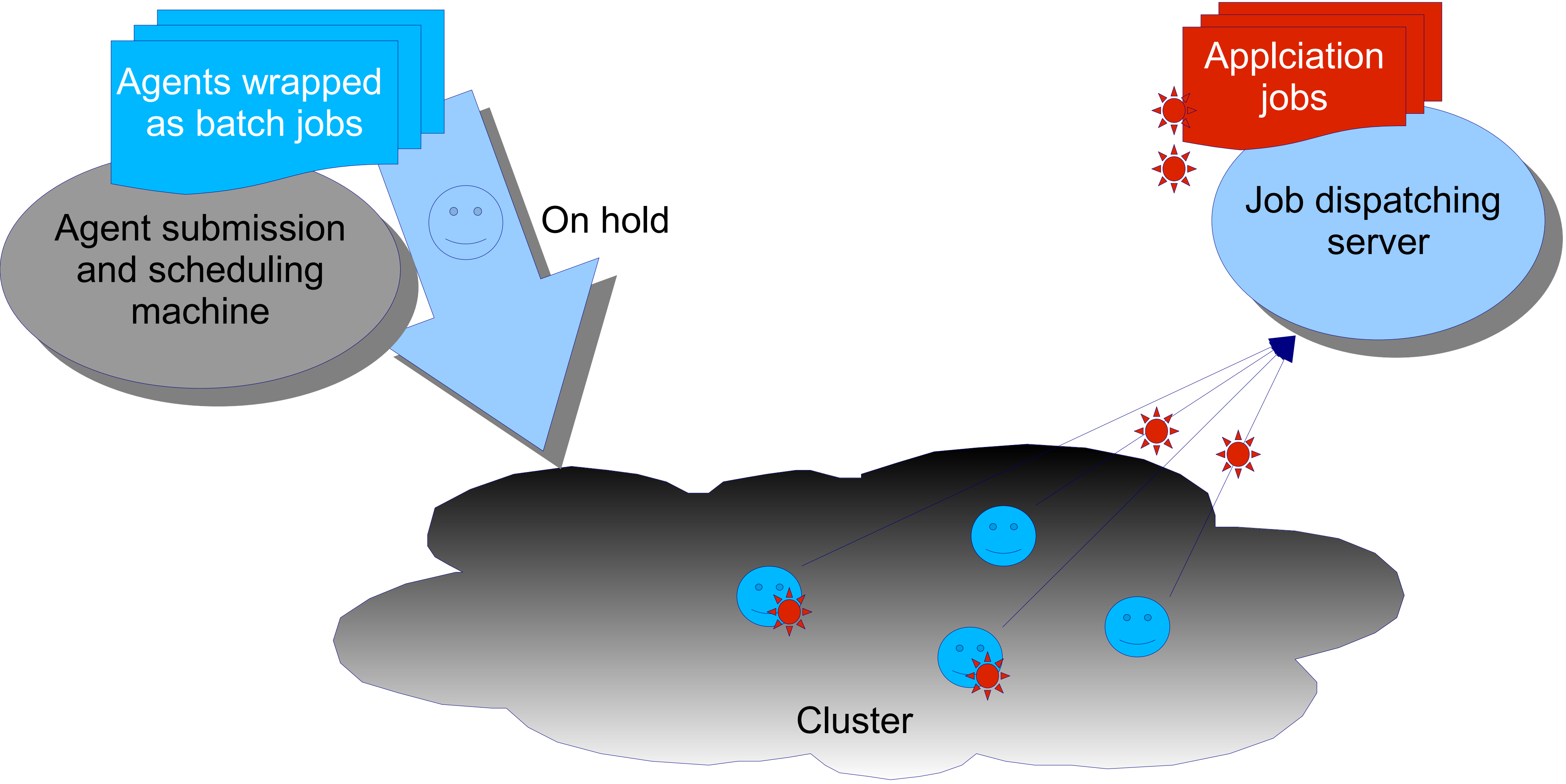
Agent submission and scheduling machine

Applciation jobs

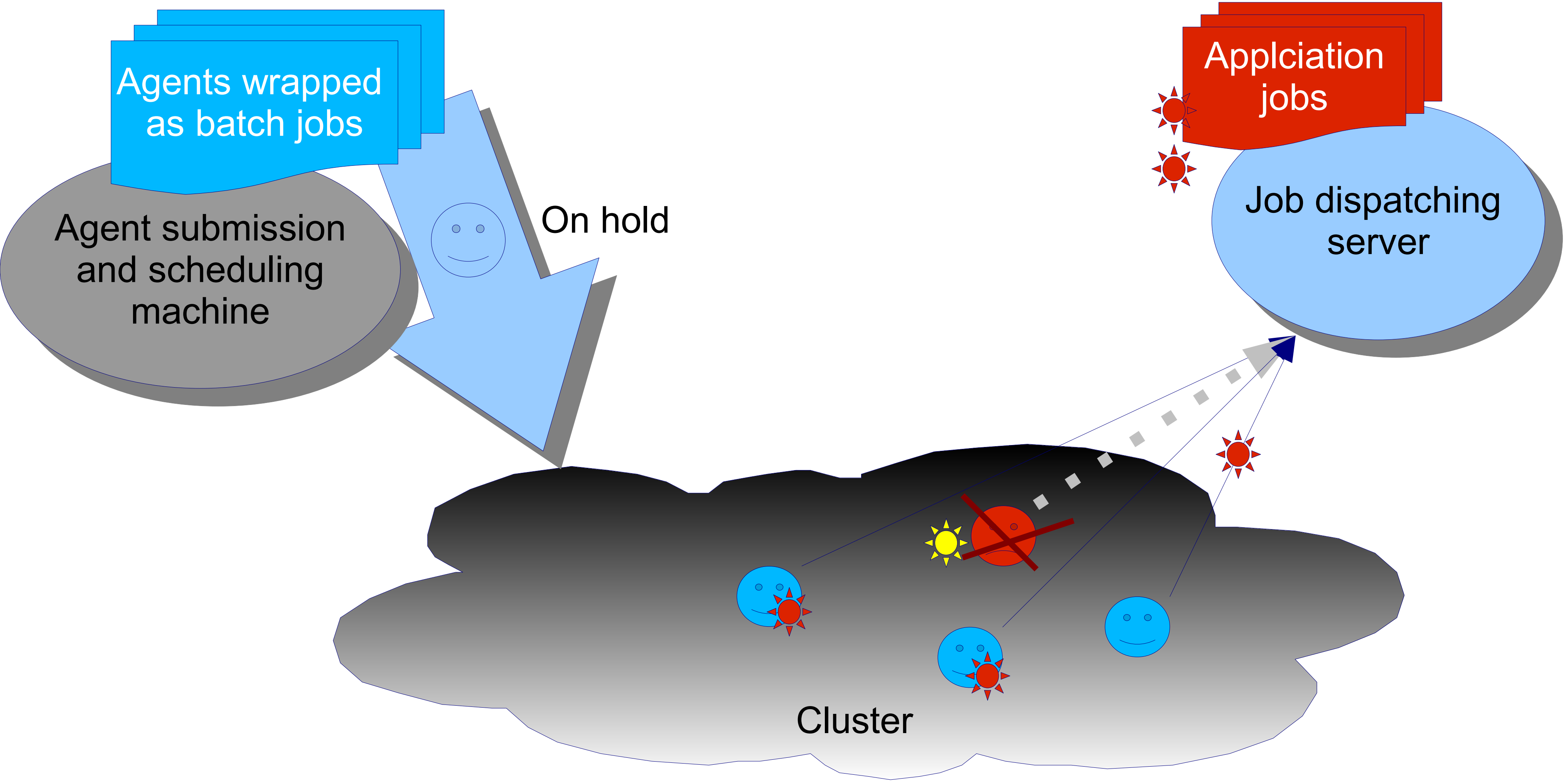
Job dispatching server

Cluster

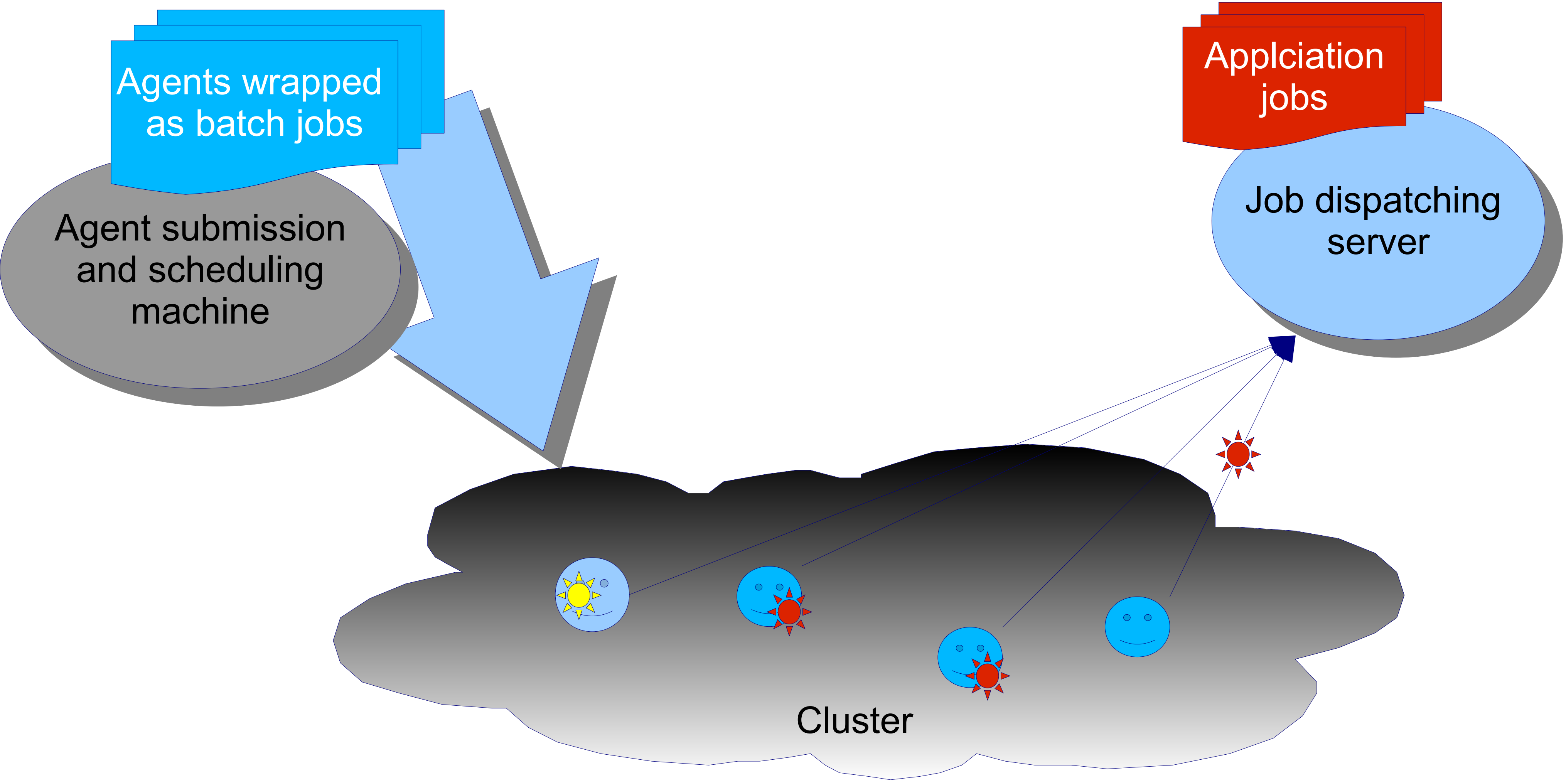
Concept: virtual cluster


Agents wrapped as batch jobs
Agent submission and scheduling machine
On hold
Applciation jobs
Job dispatching server
Cluster

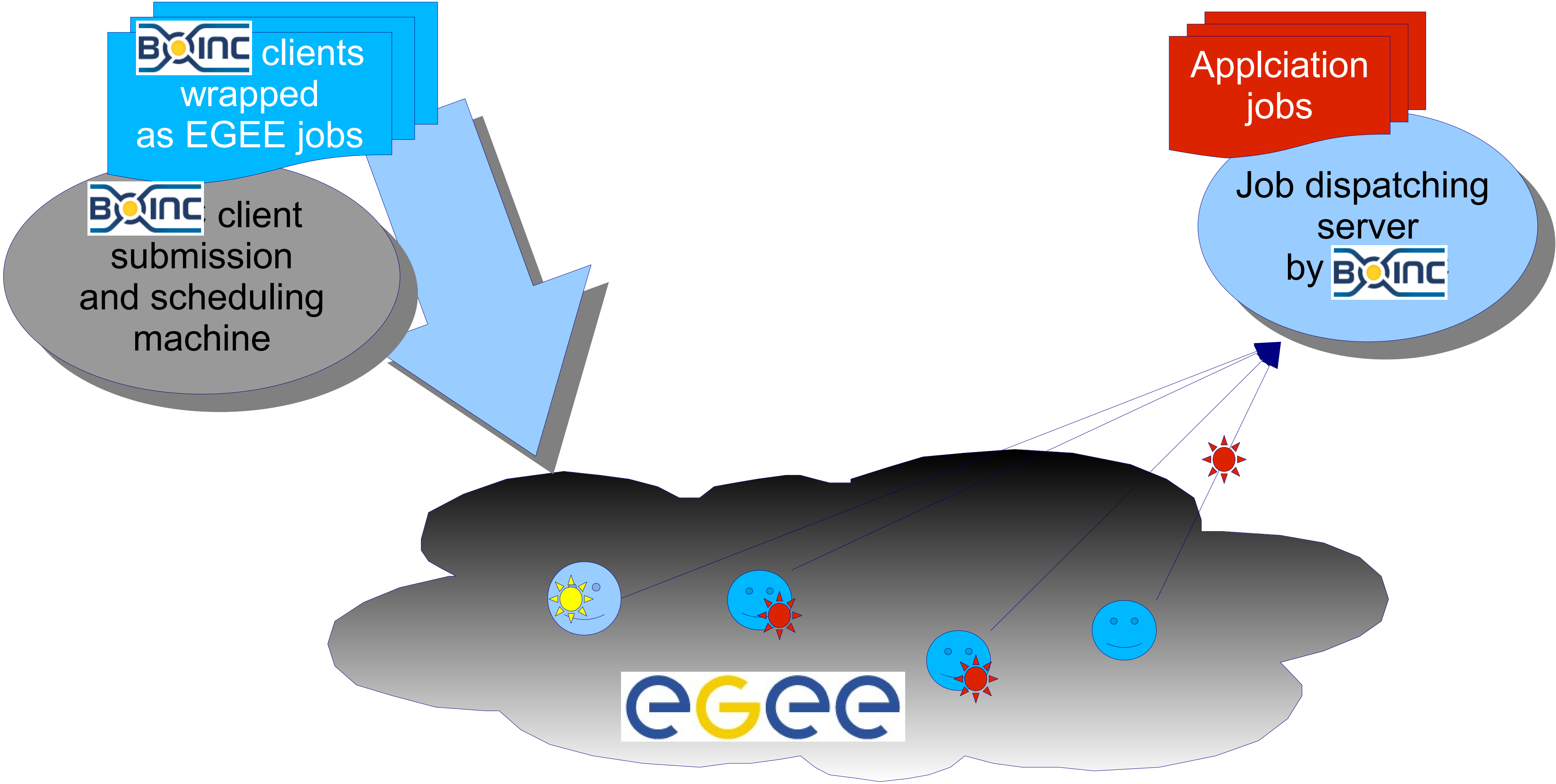
Concept: virtual cluster


Agents wrapped as batch jobs
Agent submission and scheduling machine
On hold
Applciation jobs
Job dispatching server
Cluster

Concept: virtual cluster


Agents wrapped as batch jobs
Agent submission and scheduling machine
Applciation jobs
Job dispatching server
Cluster

BOINC backend



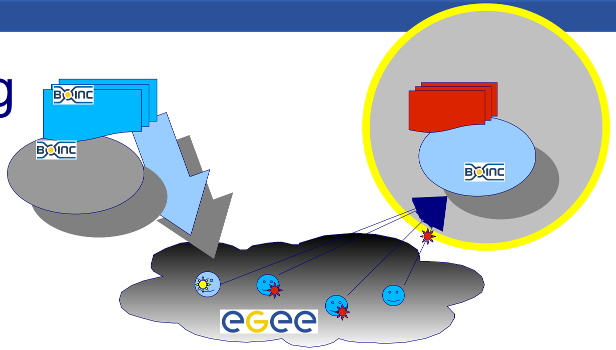
BOINC clients wrapped as EGEE jobs

BOINC client submission and scheduling machine

Applciation jobs

Job dispatching server by BOINC

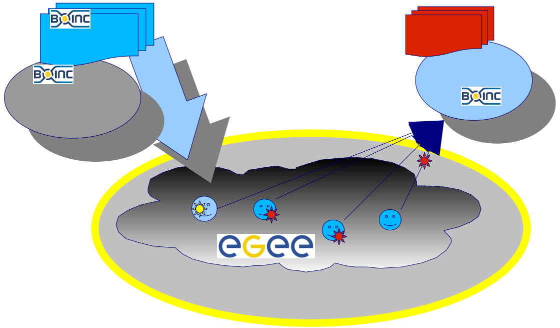
BOINC server



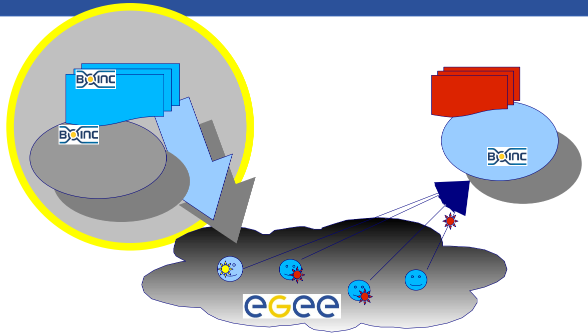
- Berkeley Open Infrastructure for Network Computing
- Out-of-the-box solution with 9 years of reputation
  - Backend of SETI@HOME
- Scalability: up to 2M hosts, billions of jobs in the queue
- Advanced scheduling
- Fault-tolerant
  - built for opportunistic environments
- Firewall-friendly
  - Clients pull jobs via HTTP
- Built in mechanisms to verify integrity and validity of results
- Built in accounting and statistics

BOINC clients

- Submitted as an ordinary EGEE job
- Runs as long as there are jobs on the server, self-terminates if idle
- Restricted to 1 core to comply with batch system allocation (policy can be adjusted)
- Performance benefits:
  - run jobs back-to-back, caching the executable and constant data
  - no batch system scheduling overhead
  - beneficial even for seconds-long jobs



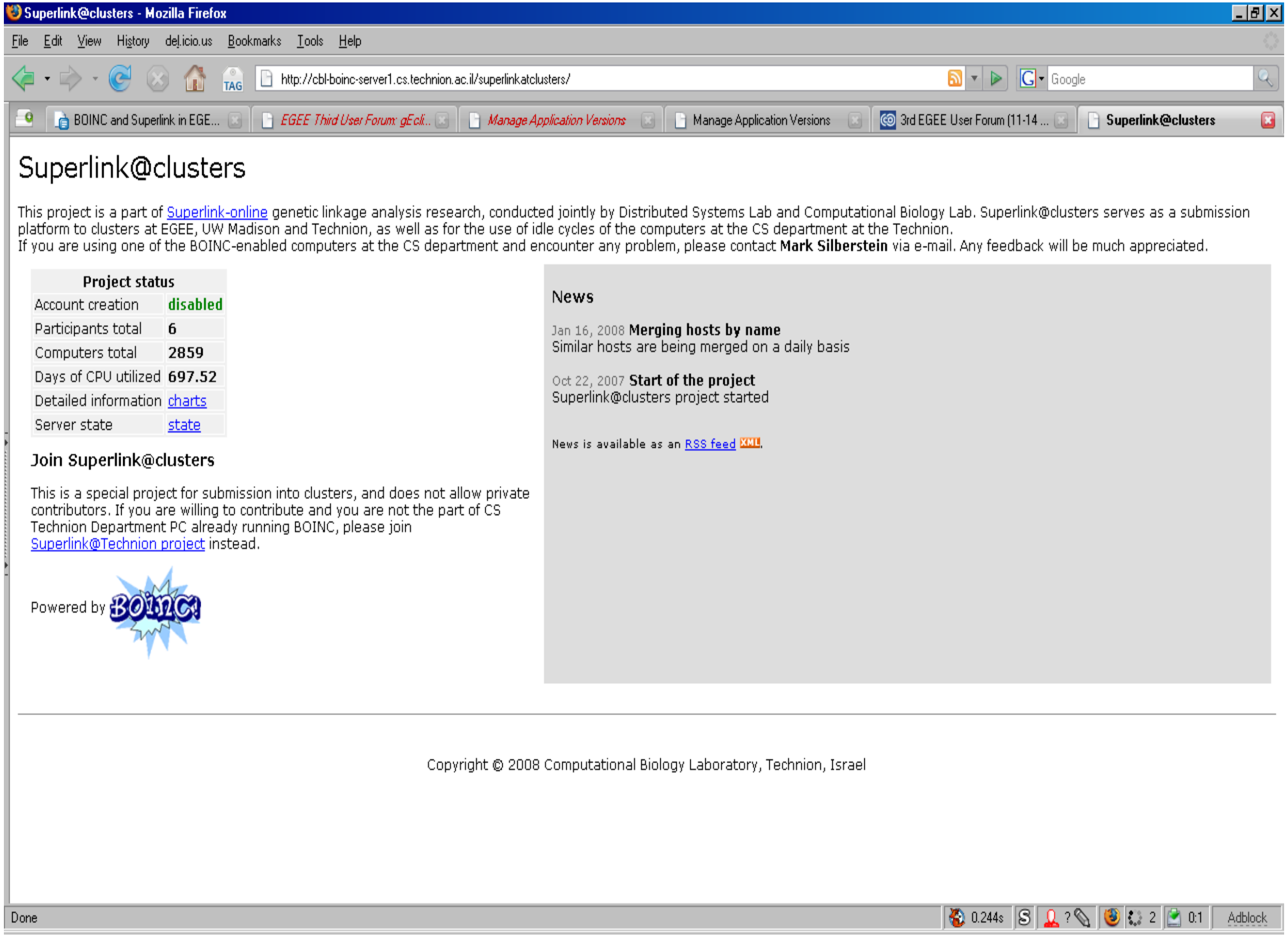
Submission machine



- Injects clients into the batch system
- Keeps track of the running clients
  - finalizes output of finished clients
  - kills long-waiting clients
  - maintains *virtual cluster:* the required number of running clients
- Avoids Resource Broker overload
  - Use of multiple resource brokers

Interface to administrators

- Centralized database for client policy
- Accounting and statistics
- Jobs results


Superlink@clusters - Mozilla Firefox
File Edit View History del.icio.us Bookmarks Tools Help
http://cbl-boinc-server1.cs.technion.ac.il/superlinkatclusters/
Google
BOINC and Superlink in EGE...
EGEE Third User Forum: gEcli...
Manage Application Versions
Manage Application Versions
3rd EGEE User Forum (11-14 ...
Superlink@clusters
Done
0.244s
2
0:1
Adblock

Superlink@clusters

This project is a part of Superlink-online genetic linkage analysis research, conducted jointly by Distributed Systems Lab and Computational Biology Lab. Superlink@clusters serves as a submission platform to clusters at EGEE, UW Madison and Technion, as well as for the use of idle cycles of the computers at the CS department at the Technion.
If you are using one of the BOINC-enabled computers at the CS department and encounter any problem, please contact **Mark Silberstein** via e-mail. Any feedback will be much appreciated.

Project status	
Account creation	disabled
Participants total	6
Computers total	2859
Days of CPU utilized	697.52
Detailed information	charts
Server state	state

Join Superlink@clusters

This is a special project for submission into clusters, and does not allow private contributors. If you are willing to contribute and you are not the part of CS Technion Department PC already running BOINC, please join Superlink@Technion project instead.

Powered by BOINC!

News

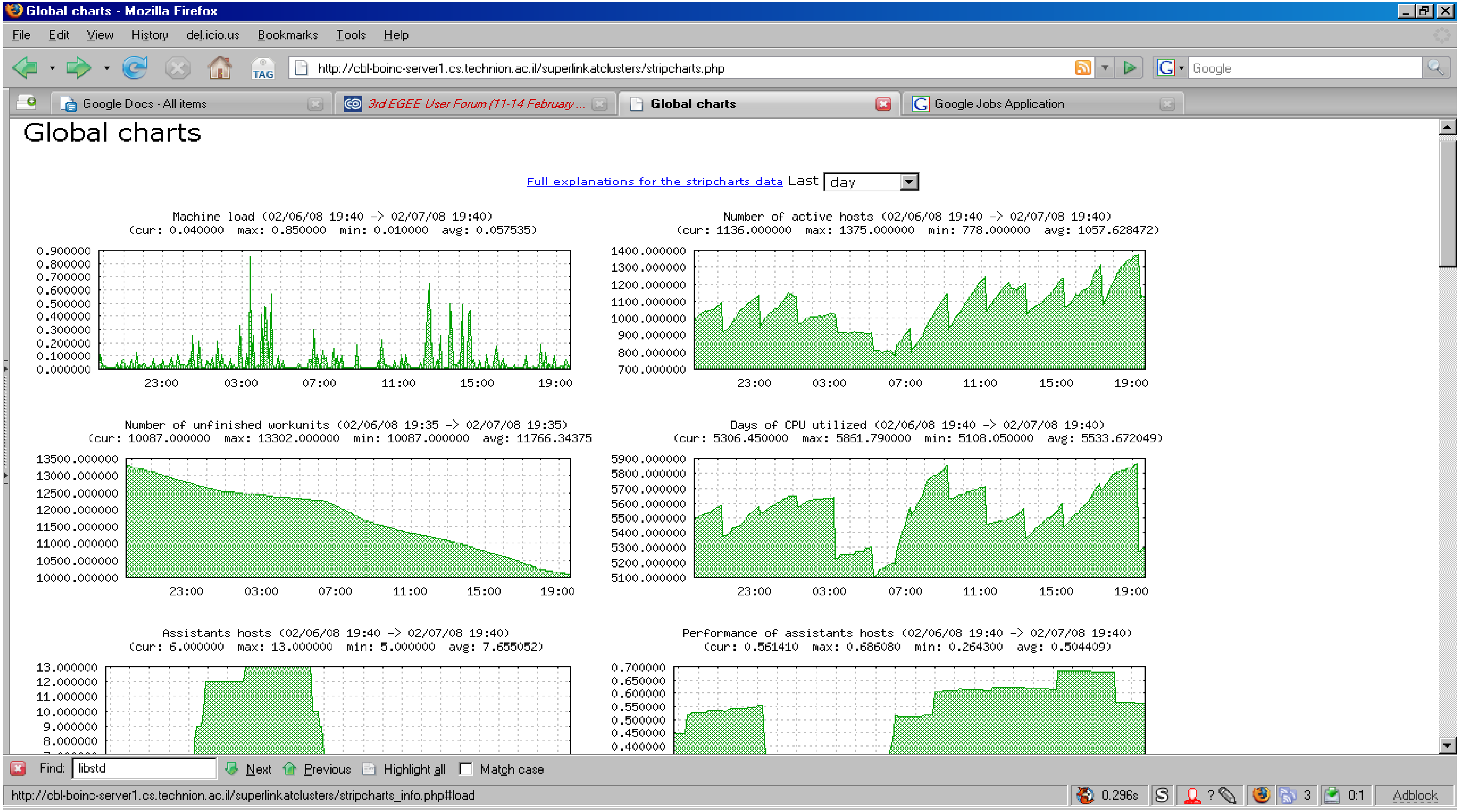
Jan 16, 2008 **Merging hosts by name**
Similar hosts are being merged on a daily basis

Oct 22, 2007 **Start of the project**
Superlink@clusters project started

News is available as an RSS feed XML.

Copyright © 2008 Computational Biology Laboratory, Technion, Israel

Accounting and statistics


Global charts - Mozilla Firefox
File Edit View History del.icio.us Bookmarks Tools Help
http://cbl-boinc-server1.cs.technion.ac.il/superlinkatclusters/stripcharts.php
Google
Google Docs - All items
3rd EGEE User Forum (11-14 February ...
Global charts
Google Jobs Application

Global charts

Full explanations for the stripcharts data Last day

Machine load (02/06/08 19:40 -> 02/07/08 19:40)
(cur: 0.040000 max: 0.850000 min: 0.010000 avg: 0.057535)

Number of active hosts (02/06/08 19:40 -> 02/07/08 19:40)
(cur: 1136.000000 max: 1375.000000 min: 778.000000 avg: 1057.628472)

Number of unfinished workunits (02/06/08 19:35 -> 02/07/08 19:35)
(cur: 10087.000000 max: 13302.000000 min: 10087.000000 avg: 11766.34375

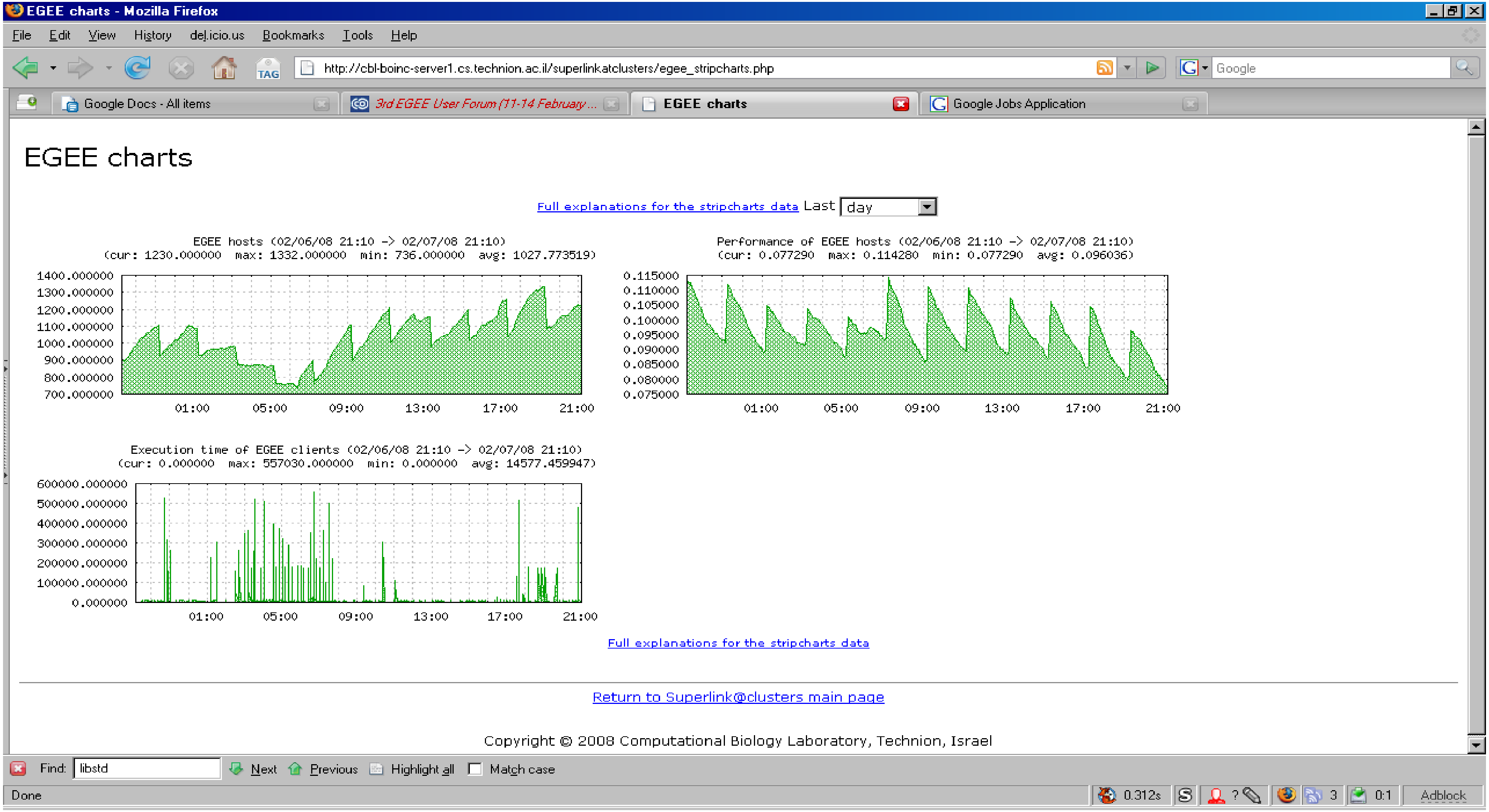
Days of CPU utilized (02/06/08 19:40 -> 02/07/08 19:40)
(cur: 5306.450000 max: 5861.790000 min: 5108.050000 avg: 5533.672049)

Assistants hosts (02/06/08 19:40 -> 02/07/08 19:40)
(cur: 6.000000 max: 13.000000 min: 5.000000 avg: 7.655052)

Performance of assistants hosts (02/06/08 19:40 -> 02/07/08 19:40)
(cur: 0.561410 max: 0.686080 min: 0.264300 avg: 0.504409)

http://cbl-boinc-server1.cs.technion.ac.il/superlinkatclusters/stripcharts_info.php#load

Accounting and statistics (cont.)



http://cbl-boinc-server1.cs.technion.ac.il/superlinkatclusters/egee_stripcharts.php

EGEE charts

Full explanations for the stripcharts data Last day

EGEE hosts (02/06/08 21:10 -> 02/07/08 21:10)
(cur: 1230.000000 max: 1332.000000 min: 736.000000 avg: 1027.773519)

Performance of EGEE hosts (02/06/08 21:10 -> 02/07/08 21:10)
(cur: 0.077290 max: 0.114280 min: 0.077290 avg: 0.096036)

Execution time of EGEE clients (02/06/08 21:10 -> 02/07/08 21:10)
(cur: 0.000000 max: 557030.000000 min: 0.000000 avg: 14577.459947)

Full explanations for the stripcharts data

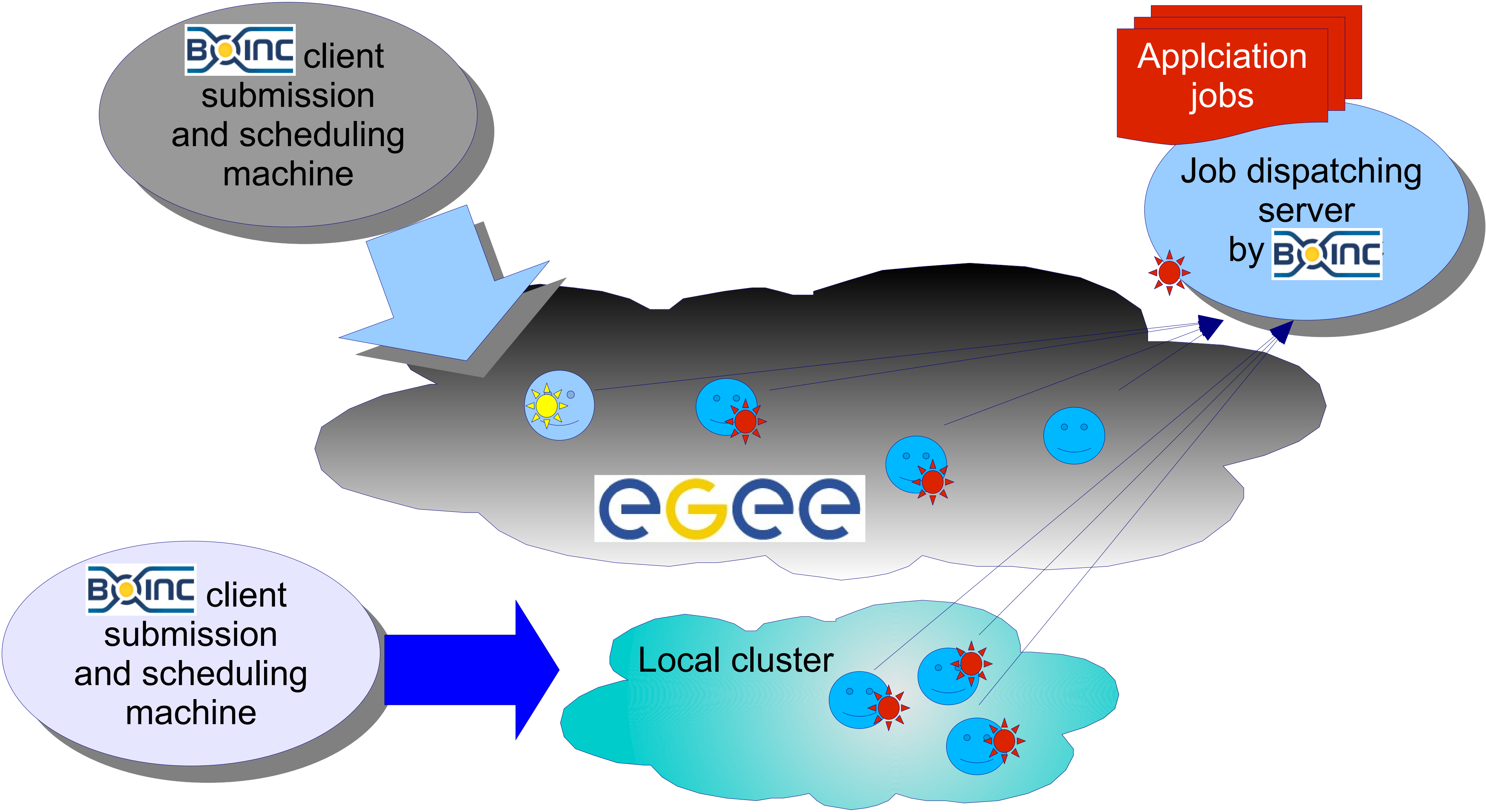
Return to Superlink@clusters main page

Copyright © 2008 Computational Biology Laboratory, Technion, Israel

Find: libstd

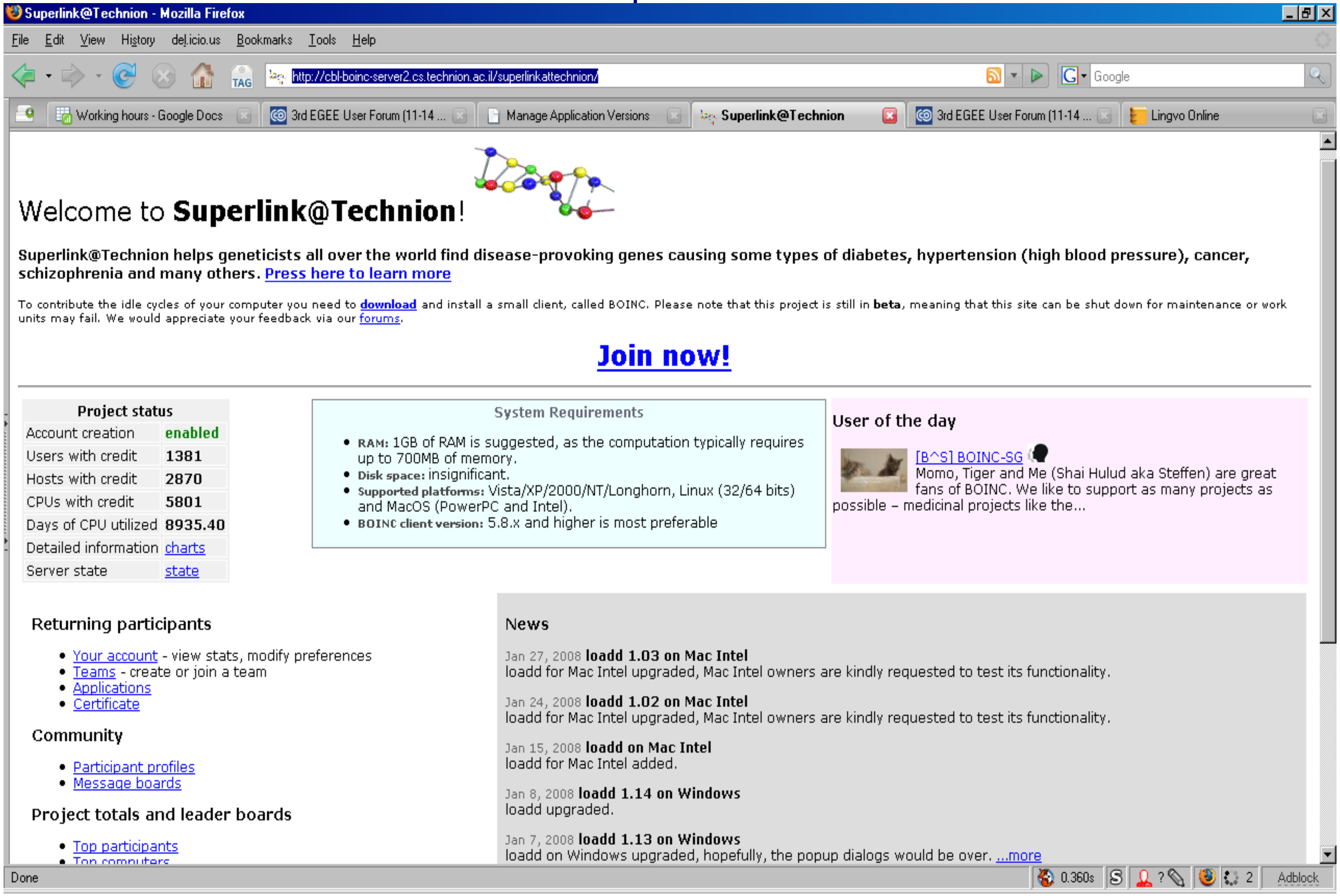
Done

Generalized solution
multiple resource pools



Community computing - same framework Superlink@Technion

- http://cbl-boinc-server2.cs.technion.ac.il/superlinkattechnion


Superlink@Technion - Mozilla Firefox
File Edit View History del.icio.us Bookmarks Tools Help
TAG
http://cbl-boinc-server2.cs.technion.ac.il/superlinkattechnion/
Google
Working hours - Google Docs
3rd EGEE User Forum (11-14 ...
Manage Application Versions
Superlink@Technion
3rd EGEE User Forum (11-14 ...
Lingvo Online
Done
0.360s
?
2
Adblock

Welcome to **Superlink@Technion**!

Superlink@Technion helps geneticists all over the world find disease-provoking genes causing some types of diabetes, hypertension (high blood pressure), cancer, schizophrenia and many others. Press here to learn more

To contribute the idle cycles of your computer you need to **download** and install a small client, called BOINC. Please note that this project is still in **beta**, meaning that this site can be shut down for maintenance or work units may fail. We would appreciate your feedback via our forums.

Join now!

Project status	
Account creation	**enabled**
Users with credit	**1381**
Hosts with credit	**2870**
CPUs with credit	**5801**
Days of CPU utilized	**8935.40**
Detailed information	charts
Server state	state

System Requirements

- **RAM:** 1GB of RAM is suggested, as the computation typically requires up to 700MB of memory.
- **Disk space:** insignificant.
- **Supported platforms:** Vista/XP/2000/NT/Longhorn, Linux (32/64 bits) and MacOS (PowerPC and Intel).
- **BOINC client version:** 5.8.x and higher is most preferable

User of the day

[B^S] BOINC-SG
Momo, Tiger and Me (Shai Hulud aka Steffen) are great fans of BOINC. We like to support as many projects as possible – medicinal projects like the...

Returning participants

- Your account - view stats, modify preferences
- Teams - create or join a team
- Applications
- Certificate

Community

- Participant profiles
- Message boards

Project totals and leader boards

- Top participants
- Top computers

News

Jan 27, 2008 **loadd 1.03 on Mac Intel**
loadd for Mac Intel upgraded, Mac Intel owners are kindly requested to test its functionality.

Jan 24, 2008 **loadd 1.02 on Mac Intel**
loadd for Mac Intel upgraded, Mac Intel owners are kindly requested to test its functionality.

Jan 15, 2008 **loadd on Mac Intel**
loadd for Mac Intel added.

Jan 8, 2008 **loadd 1.14 on Windows**
loadd upgraded.

Jan 7, 2008 **loadd 1.13 on Windows**
loadd on Windows upgraded, hopefully, the popup dialogs would be over. ...more

More info

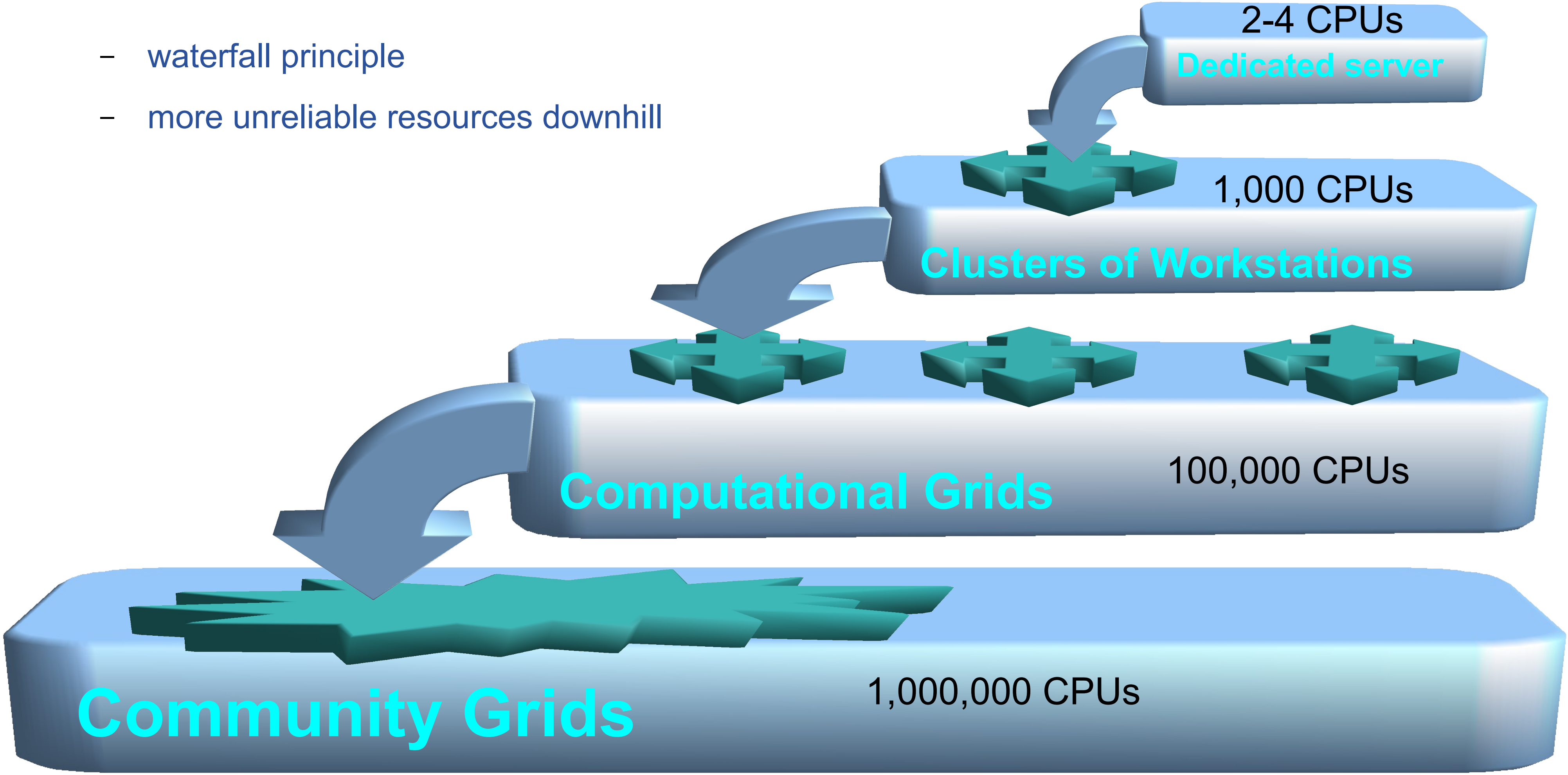
- Superlink-online genetic analysis portal
  - http://bioinfo.cs.technion.ac.il/superlink-online
- Superlink@Technion Community computing backend
  - http://cbl-boinc-server2.cs.technion.ac.il/superlinkattechnion
- Superlink@clusters EGEE and other clusters integration backend
  - http://cbl-boinc-server1.cs.technion.ac.il/superlinkatclusters
- Contact us:
  - Mark Silberstein - marks@cs.technion.ac.il
  - Artyom Sharov - sharov@cs.technion.ac.il

Questions

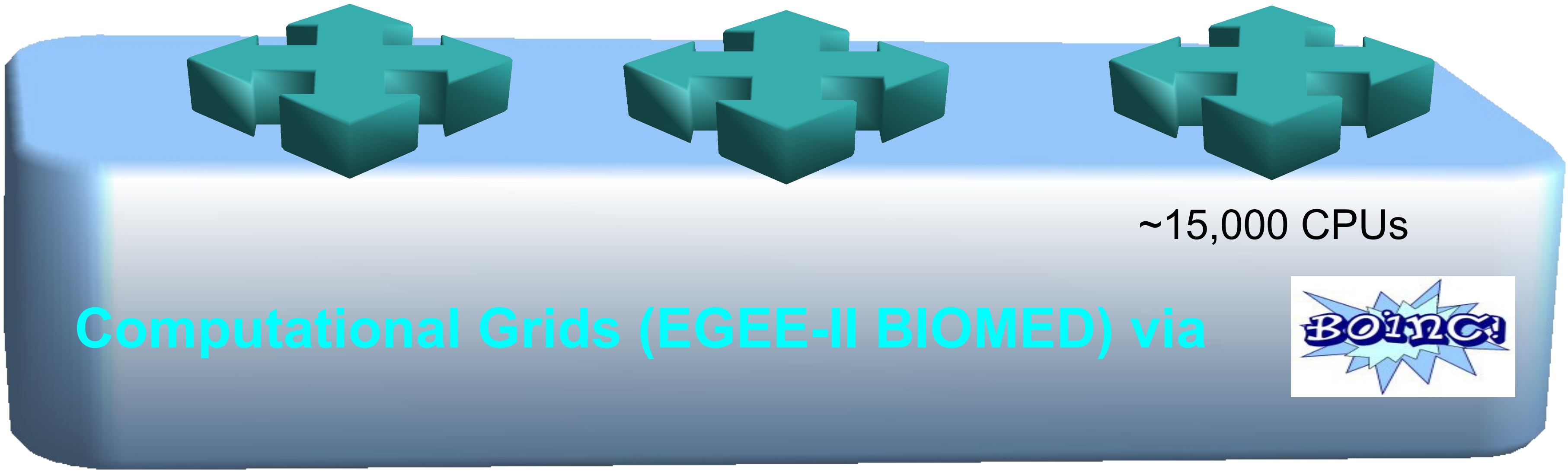
?

Solution – big picture

- Execution hierarchy:
  - waterfall principle
  - more unreliable resources downhill



Solution – zoom in



~15,000 CPUs

Computational Grids (EGEE-II BIOMED) via