

# Globus Toolkit & Grids for Synchrotrons

Advanced Photon Source  
Experiment Requirements  
+ others  
=>  
Grids

Gregor von Laszewski  
Rochester Institute of Technology  
laszewski@gmail.com

# Service Oriented Cyberinfrastructure Lab

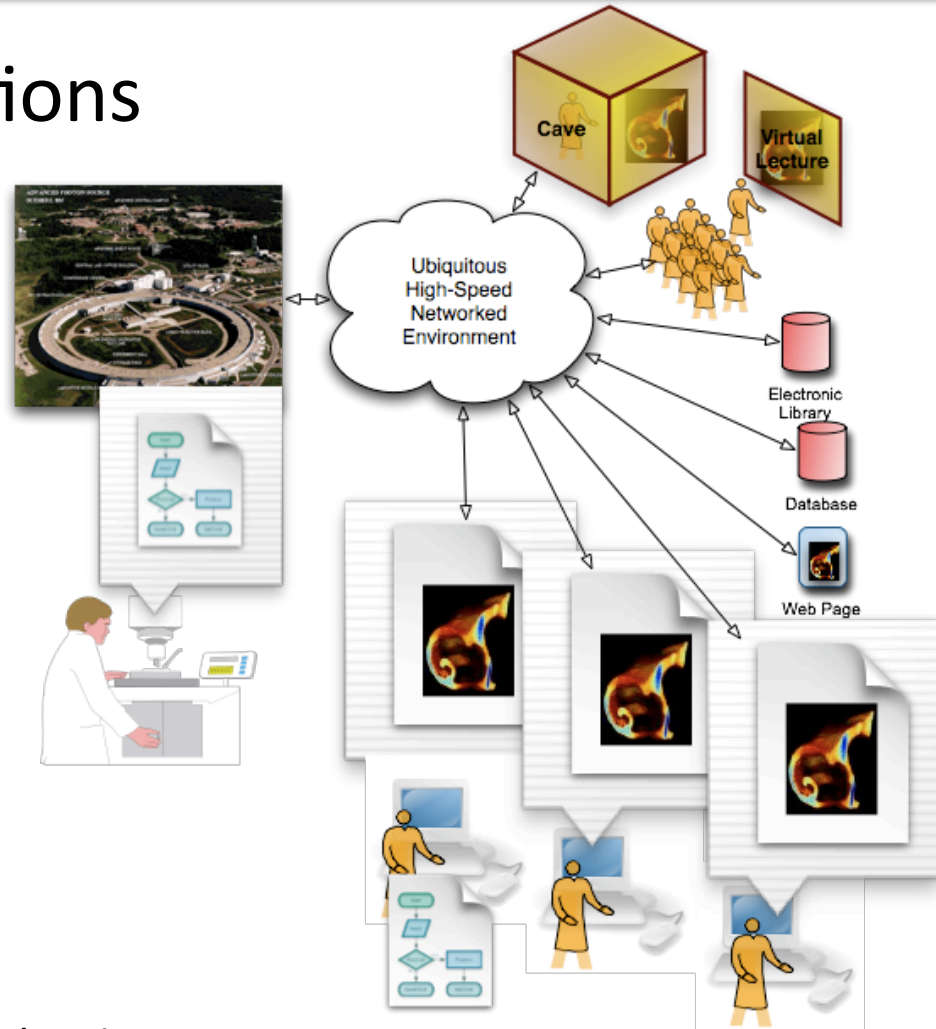
- Previous Tenured at ANL: 2-3 people
  - Possibly the 4<sup>th</sup> Grid person
  - What comes next?
  - I did research not production ...
- Small focused group
  - 1 Director
  - 1 Assoc. Director (soon)
  - 2 PhD (soon 3, come apply ;-)
  - 5 Masters Thesis
  - 15 students in Graduate Projects
  - 5 undergraduate student
- SOA, Clouds, Grid, Commodity Web 2.0 Grids, GPU (speedup of 70 is disruptive ;-))

# Outline

- Globus My View
- GT 4
- CoG Kits
- Cyberaide
  
- Administration Bottlenecks

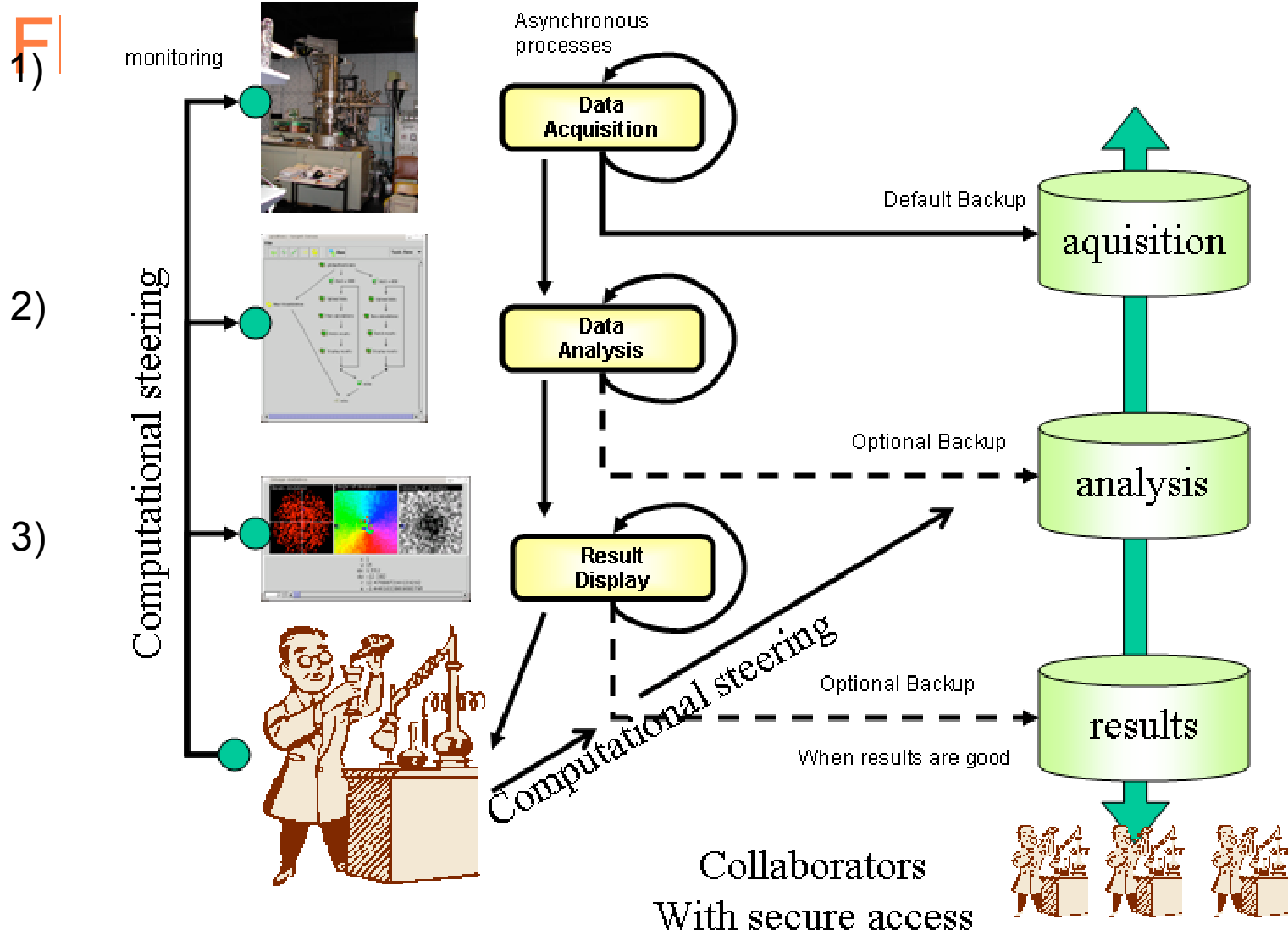
# Grid: My View

- Do real time calculations
- Queues are full
- Computers are remote
- People should not travel
- Role based Security model with ACL



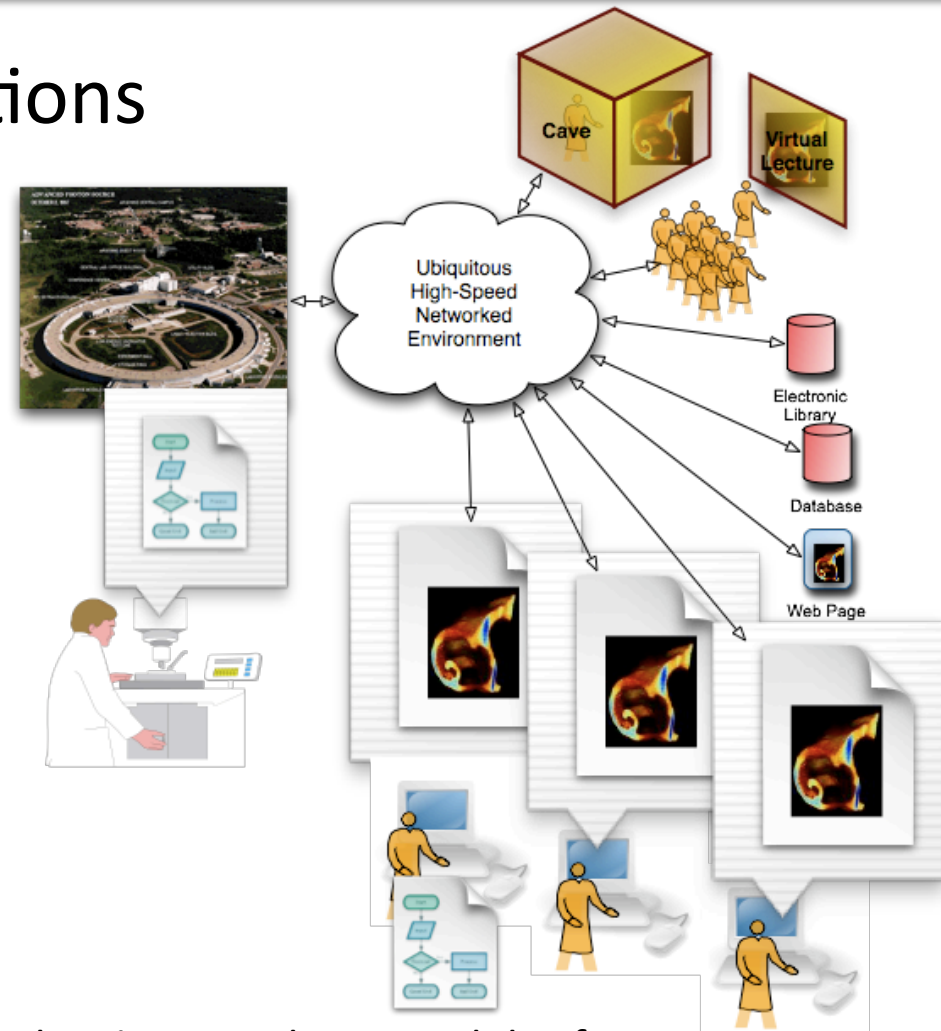
⇒ Most requirements came from this application.  
Used in DOE slides for years.

# Nanomaterials/Electron Microscope



# Realtime Data Analysis

- Do real time calculations
- Queues are full
- Computers are remote
- People should not travel
- Role based Security model with ACL



=> Most requirements came from this application. Used in DOE slides for years to come.

# Grids for Massive Data Analysis

Did not exist in 1996, 4 years later?

- Goal use (Grids and) remote Supercomputers for massive Crystallographic and Microtomographic Structural Problems and address
  - **Infrastructure:** develop new capabilities for the reconstruction of high-resolution tomographic data (including time resolved data) obtained from advanced X-ray sources such as the Advanced Photon Source (APS)
  - **Science:** develop new capabilities for the direct phase solution of atomic resolution, large-molecule crystallographic data obtained from APS detectors
  - **Modality:** demonstrate the tight coupling of Teraflop/sec-class supercomputers with APS detectors, to enable quasi-real-time analysis and visualization, and hence **interactive steering of experiments.**

# Requirements

- Many users
- Do my calculation quickly
- Give me my data
- Security, what are you talking about?
- Where is the GUI?
- Where is the script/commandline?
- Which Language
  - C/C++
  - FORTAN
  - Python (new)
- I do not want to care about
  - Metacomputing
  - Grids
  - Clouds

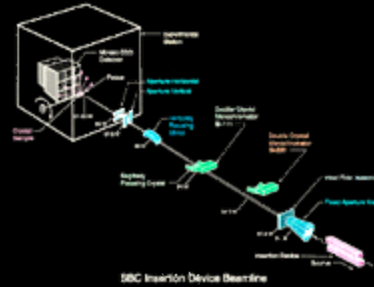


# Argonne National Laboratory Globus & Structural Biology

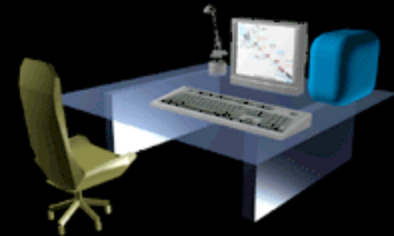
1997



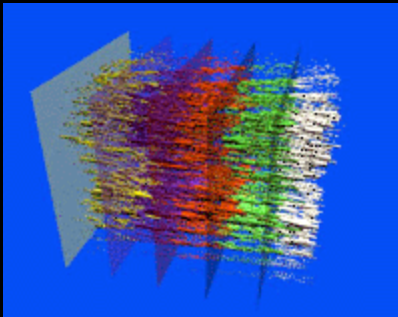
APS



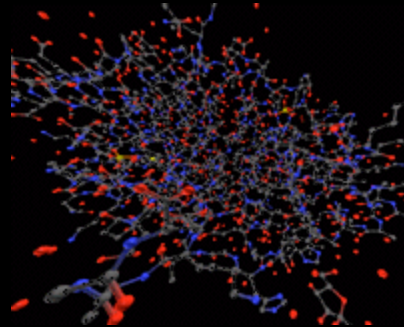
Experiment Setup



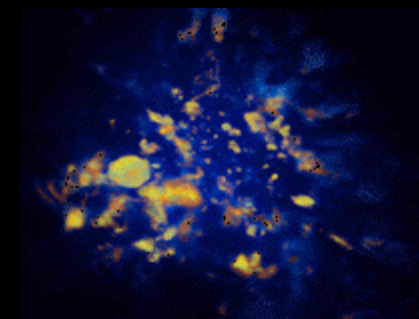
Globus Environment



Shoobox View



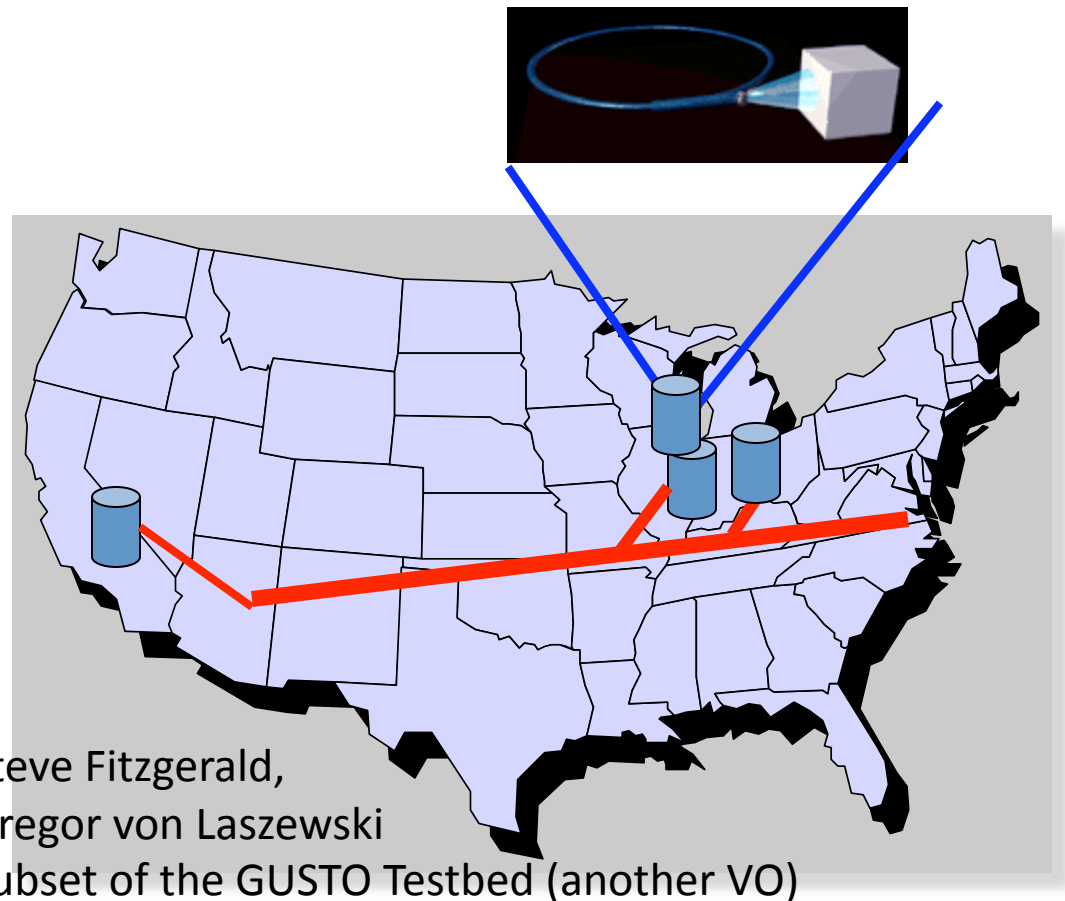
Molecule View



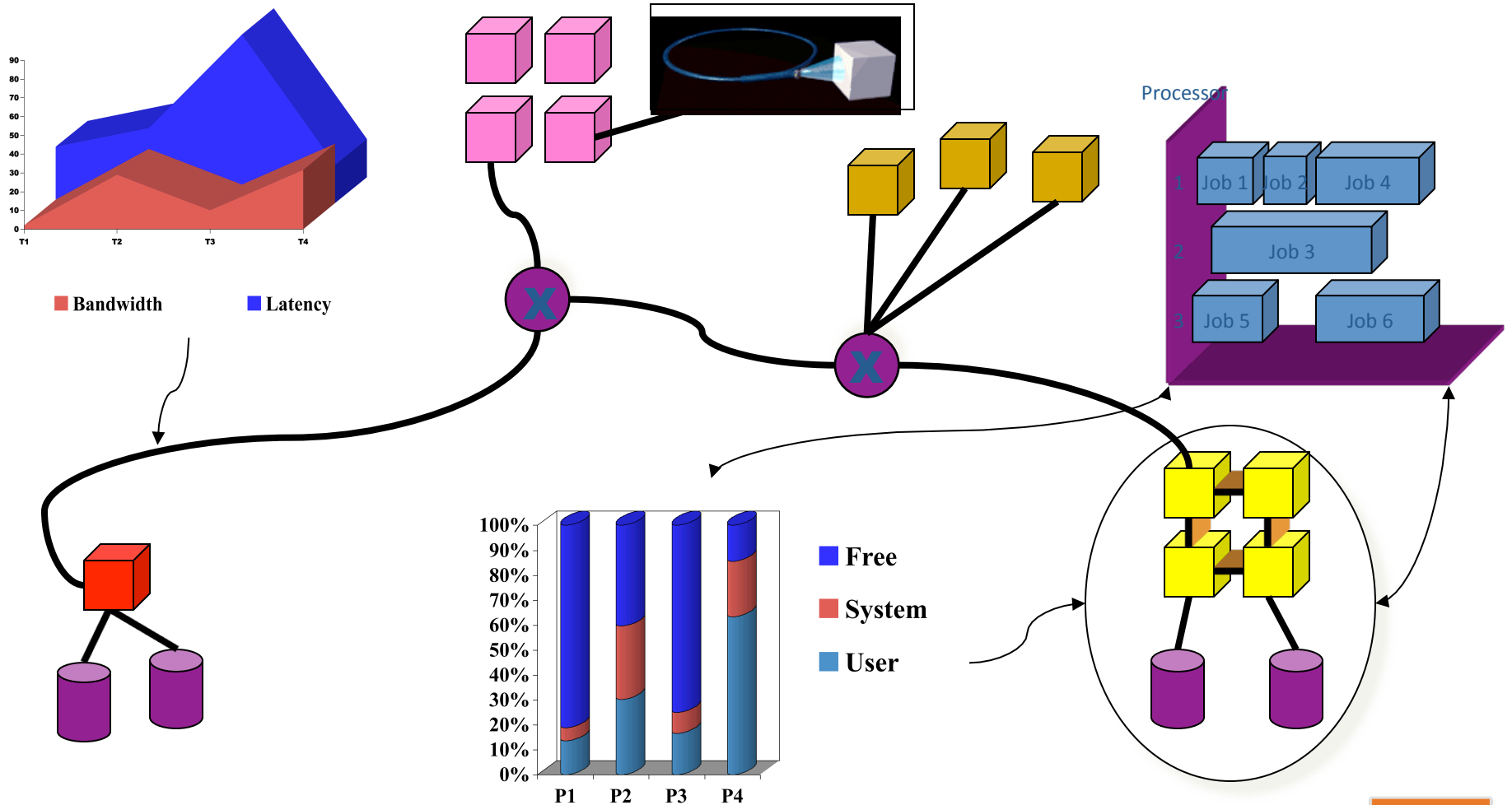
Density View

# The first “VO”: SC’97 APS - testbed

- (16) ANL, IL
- (8) ISI, CA
- (64) U Indiana
- (128) NCSA, IL
- Tomographic analysis
- One Grid (Foster) vs many Grids (Laszewski)



# State of the Network



Gregor von Laszewski, <http://www.mcs.anl.gov/gregor>

# 1997: von Laszewski Desktop Access To Remote Resources

**Heartbeat Monitor - Observer**

IP number	Hostname	Process	State	Status	Reachability	miss.	last	interval
38.245.76.14	maia.east.isi.edu	hbc_SunOS	●	●	●	0	36680 1997/05/08 14:07	
128.9.64.205	www.globus.org	/hbc_Solaris	●	●	●	0	35920 1997/05/08 17:24	
128.9.64.206	hammie.isi.edu	hbc_Solaris	●	●	●	0	39492 1997/05/08 00:47	
128.32.36.63	beefix.CS.Berkeley.EDU	hbc_Solaris	●	●	●	4	29967 1997/05/08 18:00	
128.120.56.128	suffix.cs.ucdavis.edu	hbc_DECstn	●	●	●	0	34500 1997/05/08 23:19	
128.169.92.51	dasher.cs.utk.edu	hbc_SunOS	●	●	●	5	24432 1997/05/08 15:58	
128.230.51.195	nova.npac.syr.edu	hbc_SunOS	●	●	●	0	35578 1997/05/08 18:43	
131.215.145.136	sampson	/hbc_SunOS	●	●	●	0	38651 1997/05/08 05:44	
132.239.51.91	perplex.ucsd.edu	hbc_Solaris	●	●	●	0	36590 1997/05/08 14:48	
140.221.3.110	tiamat.mcs.anl.gov	hbc_Solaris	●	●	●	0	39896 1997/05/08 00:50	

**Latency (ms)**

Source	Destination	Latency (ms)
ISI	bolan.isi.edu	1.177
ISI	tiamat.mcs.anl.gov	5.043
ISI	huntsman.isi.edu	32.673
ISI	dew.	44.221
ISI	dew.mcs.anl.gov	11.96
huntsman.isi.edu	dew.	108.584

**Throughput (Mbit/s)**

Source	Destination	Throughput (Mbit/s)
ISI	bolan.isi.edu	8.001
ISI	tiamat.mcs.anl.gov	0.753
ISI	huntsman.isi.edu	0.799
ISI	dew.	7.258
huntsman.isi.edu	dew.mcs.anl.gov	1.313

**Node Status Window (tiamat.mcs.anl.gov)**

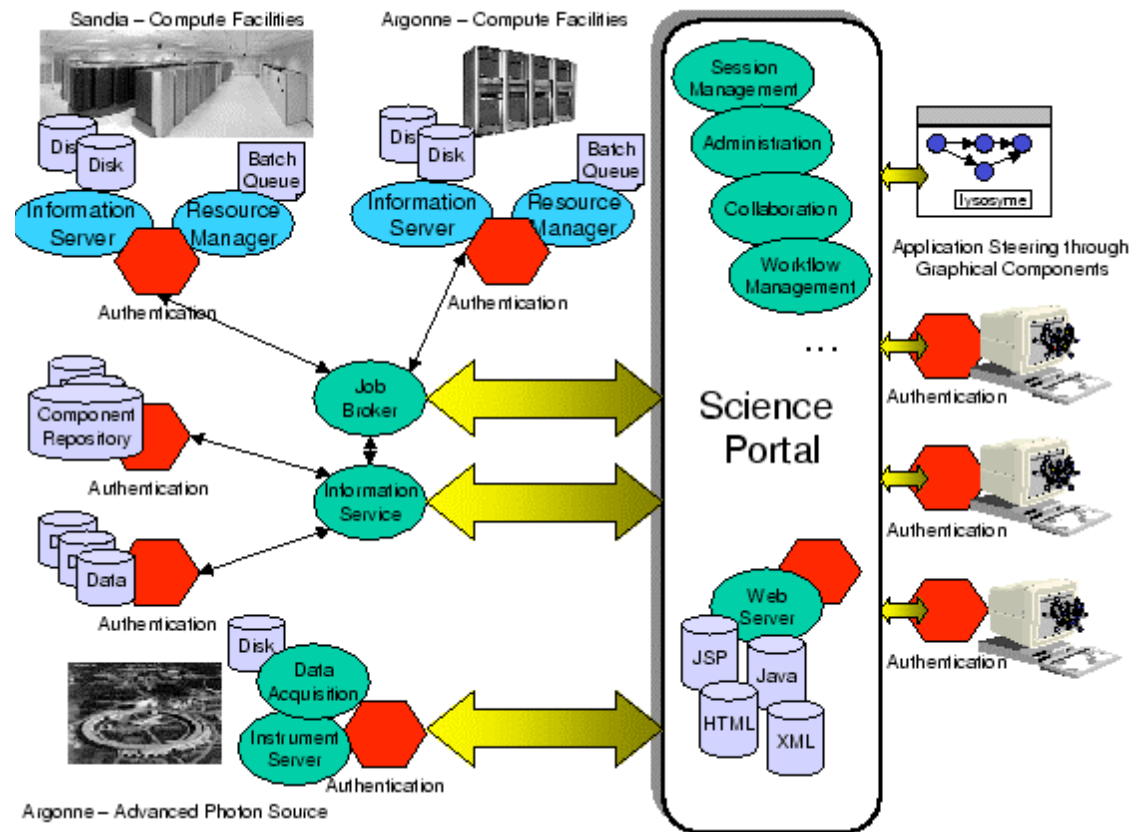
```

IP Number: 140.221.3.110
Hostname: tiamat.mcs.anl.gov
Process Name: hbc_Solaris
State: [X] alive
Status: [O] registered and alive
Registration Time: 1997/05/08 00:50:20 CMT
Seconds between heartbeat: 15
Number of last Heartbeat: 39896
Time of detected heartbeat: 1997/05/14 23:04:27 CMT
Number of missing heartbeats: 0
Reachability: [1] host reachable (based on HB heartbeat messages review)
    
```

Gregor von Laszewski, <http://www.mcs.anl.gov/gregor>

# The Portal is Important

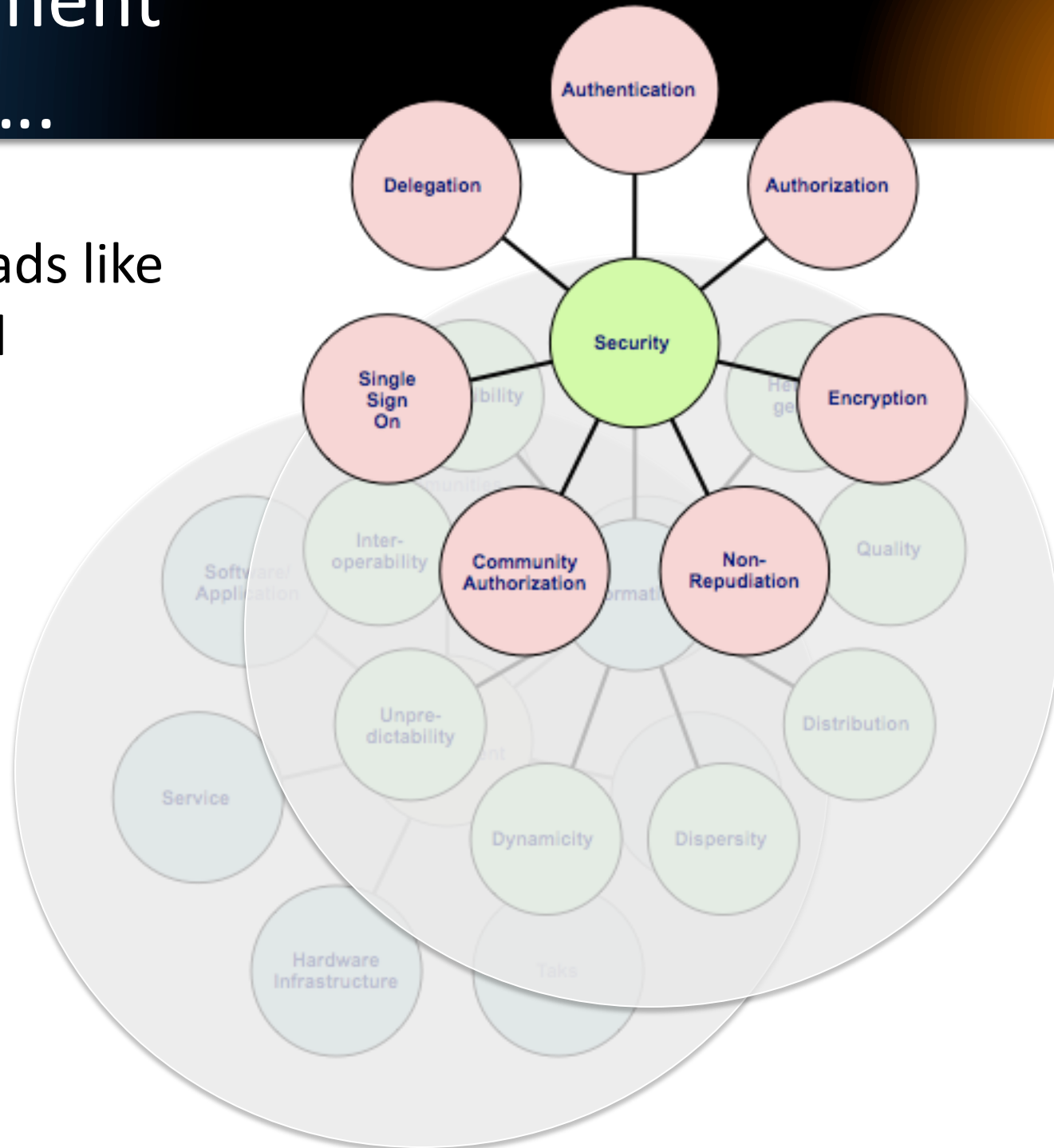
- ACM article



# Management aspects ....

1997

- It spreads like a weed



# Evolution of Globus

Cloud

CoG Kit workflow

Cyberaide

If you used the CoG Kit for on demand batch Processing and file transfer, little changes needed. CoG is above middleware.

Meta Computer

1995/6

Generation 0:  
Message Passing  
(Nexus, I-Way)

1998

Generation 1:  
On Demand Batch Computing, Runtime Library, Protocols, API, message passing (Gusto Testbed/ Metacomputing)

GvL: everything must be a service

1999

Generation 2:  
Addressing scalability  
removal of message passing, focus on Elementary functionality  
Bottom up design  
gridFTP  
GridForum

GvL: top down design

2003

Generation 3:  
Everything is a service  
Global Grid Forum  
Stateful Services  
Java  
Not the right way.

2006

Generation 4:  
WSRF  
Stateful Services  
Scalability Improvements  
TeraGrid  
Community Driven  
Open Grid Forum

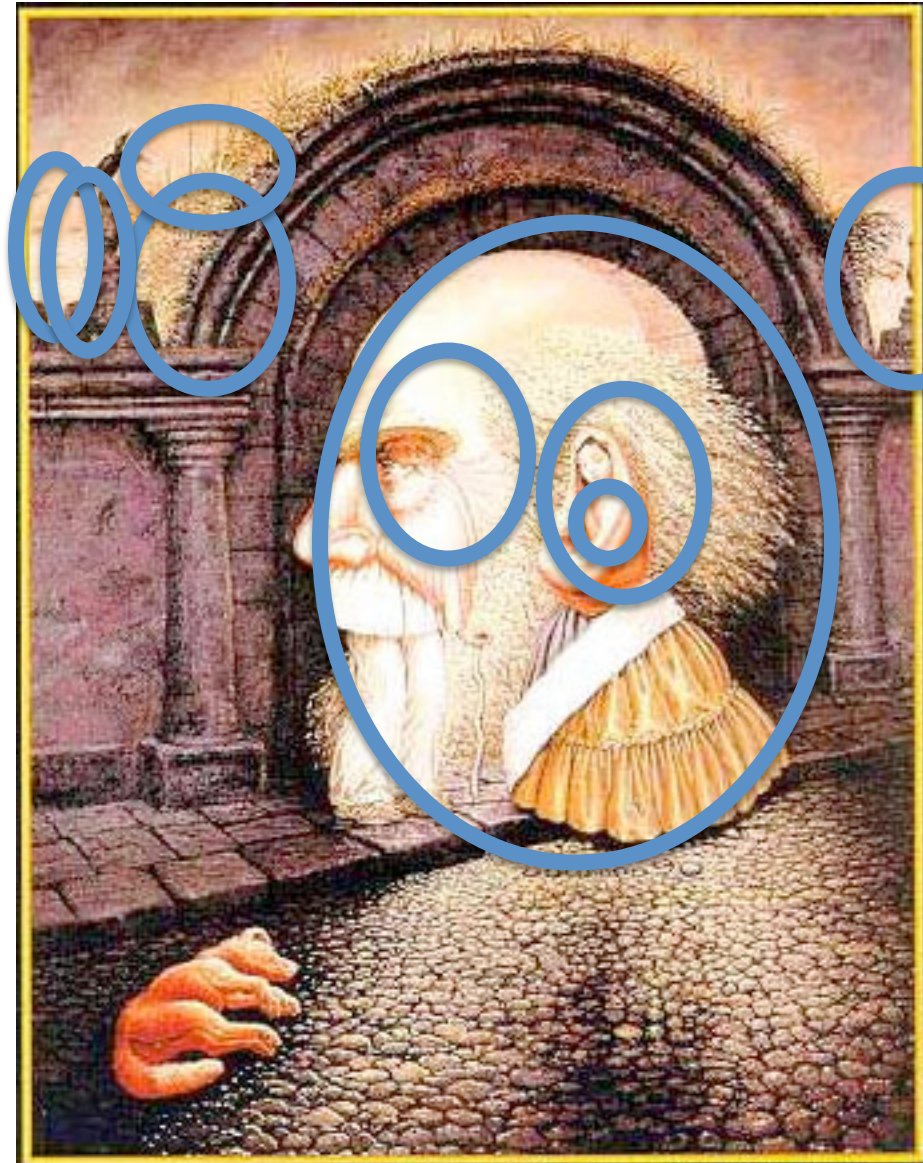
Note: This is the impression GvL Was with the GT Project from 1996-2007

Help me if the dates are wrong

# The wife and the mother-in-law



1. old man profile
2. Old man standing
3. Lady
4. face on the right side of the pillar (profile)
5. face on the left side of the pillar (profile)
6. another gestalt shape, at the opposite to the left pillars face.
7. a front facing extreme left in the sky.
8. face near a crow.
9. baby with the lady.





# Fast Forward to Today

# Today: Globus 4

- Contains pre-WS GT2, many claim they do GT4 but that may not mean they use web services
- Contains WSRF style web services
  - Improved job scalability
  - Improved file transfer
  - Many components added by the community
- Lots of useful components
- Lots of development of additional components

# dev.globus.org

## Globus Projects

- MPICH G2
- OGSA-DAI
- Incubation Mgmt

- Java Runtime
- C Runtime
- Python Runtime

- Delegation
- CAS
- C Sec

- MyProxy
- GSI-OpenSSH
- GridWay

- GRAM
- Data Rep
- GridFTP
- Reliable File Transfer

- ## Globus Toolkit
- Replica Location
  - MDS4
  - GT4 Docs

## Incubator Projects

- |           |          |        |         |          |           |      |           |
|-----------|----------|--------|---------|----------|-----------|------|-----------|
|           |          |        | Swift   | GEMLCA   | gRAVI     |      | MonMan    |
|           |          | GAARDS | MEDICUS | CoG WF   | Virt WkSp |      | NetLogger |
| GDTE      | GridShib | OGRO   | UGP     | Dyn Acct | Gavia JSC | DDM  | Metrics   |
| Introduce | PURSE    | HOC-SA | LRMA    | WEEP     | Gavia MS  | SGGC | ServMark  |

- |                |          |                |           |               |       |
|----------------|----------|----------------|-----------|---------------|-------|
| Common Runtime | Security | Execution Mgmt | Data Mgmt | Info Services | Other |
|----------------|----------|----------------|-----------|---------------|-------|



# Job Submission

# GRAM4 Scalability

- Job Submission
- Scalability a major focus of GRAM's design
  - GRAM4 can manage 32,000 active jobs
  - Ability to manage load on control node
  - GRAM4 can handle bursts of up to 50 job submissions
  - Each job requires ~2s to process
- Are the error conditions acceptable?
  - Job should be rejected or timeout before overloading the service container or service host

# Job Submission Features

## Globus has...

- Web service for job submission and control
- Cmd line and Java clients
- Data stage-in/stage-out
- Client notification support
- Plug-in interface for local resource managers
- Support for popular resource managers
- Plug-in interface for authorization decisions
- Advance reservation support (GARS)
- Standards compliance
- Supported on natl. grids

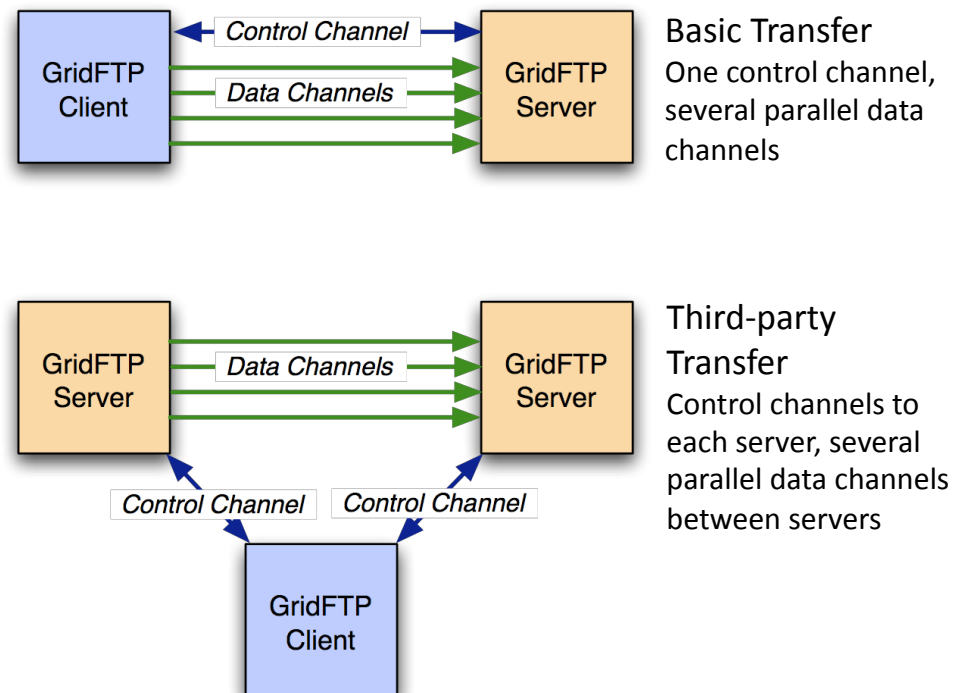
## Globus doesn't have...

- Fancy submission tools (see Condor-G)
- Portlets (see CoG, OGCE)
- Scheduling/queuing (see GridWay, PBS, LSF...)
- Co-scheduling (see GARS, HARC, GUR)
- Support for every resource manager
- Support for complicated authorization decisions (see VOMS, CAS, GridShib)
- CoG Kit Karajan Workflow

# Filetransfer

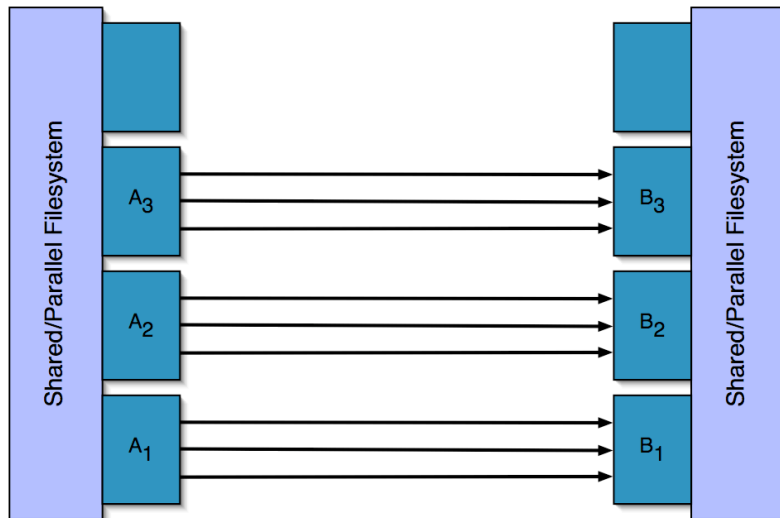
# GridFTP Features

- A high-performance, secure data transfer service optimized for high-bandwidth wide-area networks
  - ◆ FTP with extensions
  - ◆ Uses basic Grid security (control and data channels)
  - ◆ Multiple data channels for parallel transfers
  - ◆ Partial file transfers
  - ◆ Third-party (direct server-to-server) transfers





# Striped GridFTP



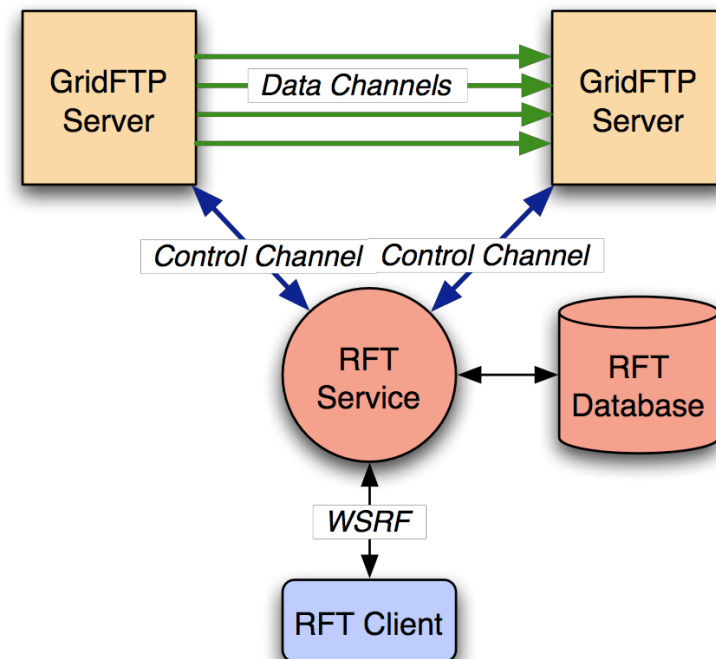
- GridFTP supports a striped (multi-node) configuration
  - ◆ Establish control channel with one node
  - ◆ Coordinate data channels on multiple nodes
  - ◆ Allows use of many NICs in a single transfer
- Requires shared/parallel filesystem on all nodes
  - ◆ On high-performance WANs, aggregate performance is limited by filesystem data rates

# globus-url-copy

- Command-line client for GridFTP servers
  - Text interface
  - No “interactive shell” (single command per invocation)
- Many features
  - Grid security, including data channel(s)
  - HTTP, FTP, GridFTP
  - Server-to-server transfers
  - Subdirectory transfers and lists of transfers
  - Multiple parallel data channels
  - TCP tuning parameters
  - Retry parameters
  - Transfer status output

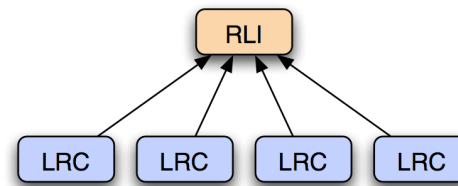
# RFT – Reliabal File Transfer Service

- A WSRF service for queuing file transfer requests
  - ◆ Server-to-server transfers
  - ◆ Checkpointing for restarts
  - ◆ Database back-end for failovers
- Allows clients to request transfers and then “disappear”
  - ◆ No need to manage the transfer
  - ◆ Status monitoring available if desired

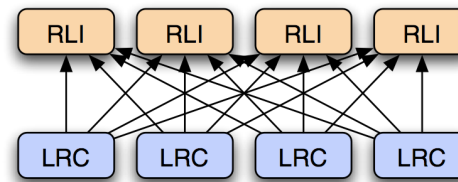


# RLS - Replica Location Service

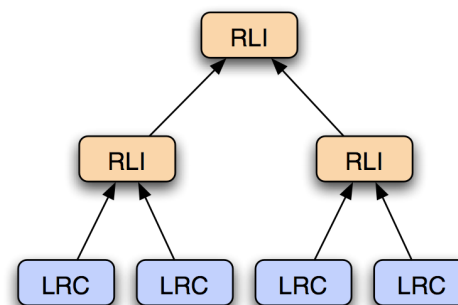
- A distributed system for tracking replicated data
  - Consistent local state maintained in Local Replica Catalogs (LRCs)
  - Collective state with relaxed consistency maintained in Replica Location Indices (RLIs)
- Performance features
  - Soft state maintenance of RLI state
  - Compression of state updates
  - Membership and partitioning information maintenance



Simple Hierarchy  
The most basic  
deployment of RLS



Fully Connected  
High availability of  
the data at all sites



Tiered Hierarchy  
For very large  
systems and/or very  
Large collections

# pyGlobus\*

- High-level, object-oriented interface in Python to GT4 Pre-WS APIs.
  - GSI security
  - GridFTP
  - GRAM
  - XIO
  - GASS
  - MyProxy
  - RLS
- Also includes tools and services
  - GridFTP server
  - GridFTP GUI client
  - Other GT4 clients

# Globus Toolkit Features

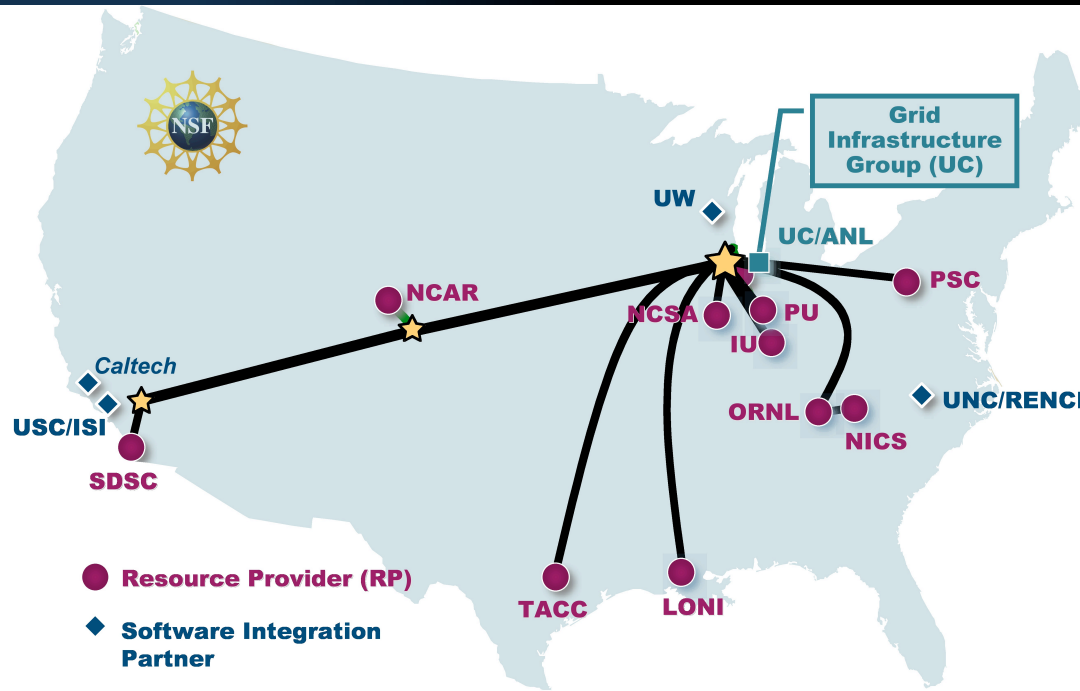
## Globus has...

- Modular architecture
- APIs
- Embeddable libraries
- Web service interfaces
- Globus-enabled frameworks for MPI, RPC, parallel jobs, etc.
- Globus support on national infrastructure

## Globus doesn't have...

- Your application already Grid-enabled
- A tool to automatically adapt your code
- Domain-specific frameworks

# Deployment: NSF TeraGrid



## ● TeraGrid DEEP: Integrating NSF's most powerful computers (60+ TF)

- ◆ 2+ PB Online Data Storage
- ◆ National data visualization facilities
- ◆ World's most powerful network (national footprint)

## ● TeraGrid WIDE Science Gateways: Engaging Scientific Communities

- ◆ 90+ Community Data Collections
- ◆ Growing set of community partnerships spanning the science community.
- ◆ Leveraging NSF ITR, NIH, DOE and other science community projects.
- ◆ Engaging peer Grid projects such as Open Science Grid in the U.S. as peer Grids in Europe and Asia-Pacific.

## ● Base TeraGrid Cyberinfrastructure: Persistent, Reliable, National

- ◆ Coordinated distributed computing and information environment
- ◆ Coherent User Outreach, Training, and Support
- ◆ Common, open infrastructure services

## A National Science Foundation Investment in Cyberinfrastructure

**\$100M 3-year construction (2001-2004)**

**\$150M 5-year operation & enhancement (2005-2009)**

\* Slide courtesy of Ray Bair, Argonne National Laboratory

# Security

I am leaving this out as this will fill another hour, day,  
week, ...

GSI authentication:  
delegation, time restricted proxys



# Which Services Do You need?

- Authentication
- (Authorization)
- File Transfer
- Job Management
- Workflow Management
- Probably more ...

CoG Kit

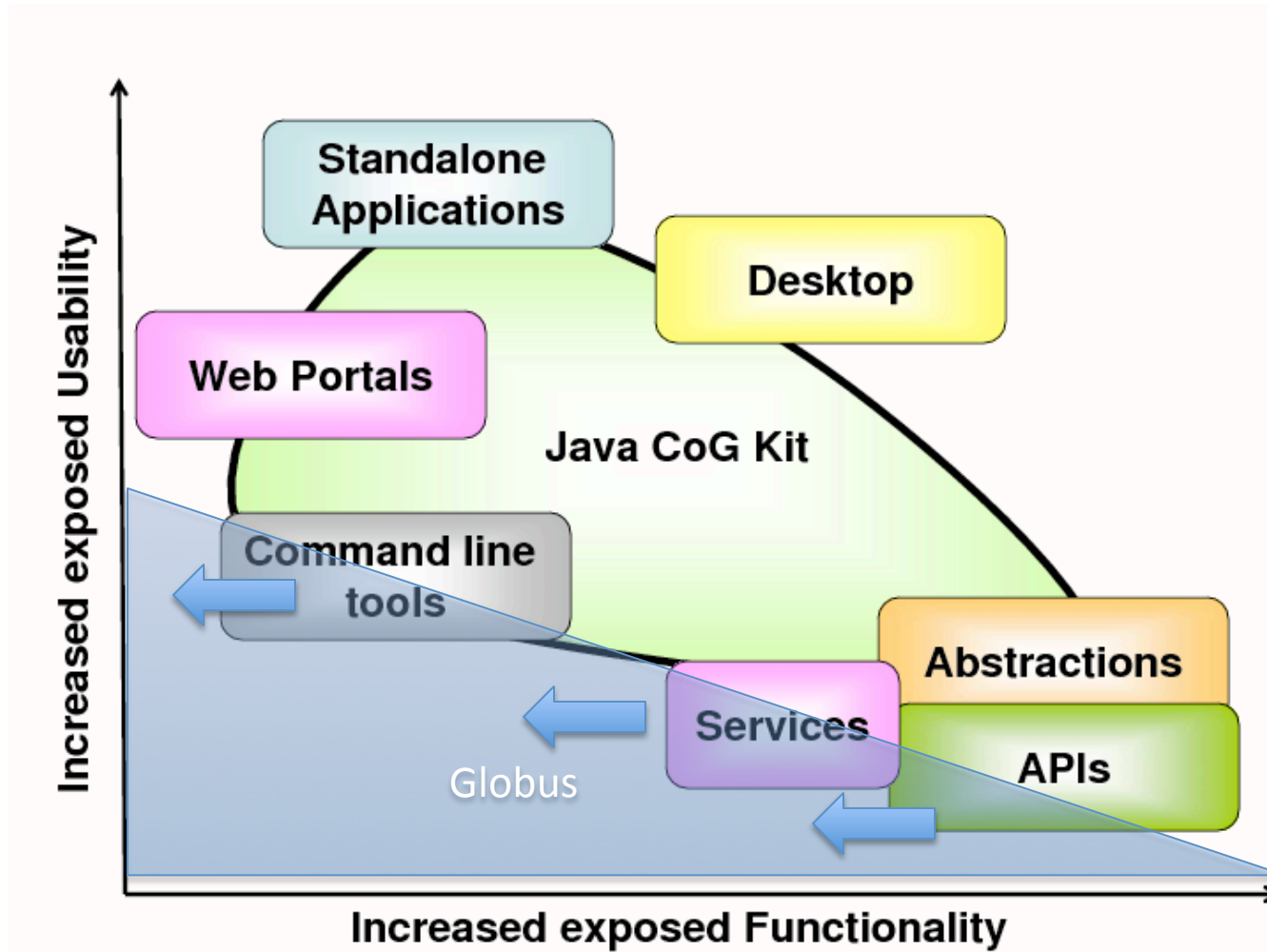
Condor

Globus

Others

?

# CoG Kit



# Visual Interfaces

The screenshot shows a desktop environment with several windows and icons. Annotations point to various elements:

- System Icons:** A red circle highlights the system tray icons on the left side of the desktop.
- Active Grid Icons:** A red circle highlights the 'Job Subc17' icon on the desktop.
- Native Icons:** A red circle highlights the 'Microsoft Word' icon on the desktop.
- Toolbar:** A red circle highlights the toolbar on the left side of the desktop.
- Log:** A red circle highlights the 'CoG Log' window at the bottom of the desktop.

The desktop background features the text 'Java CoG Kit' and a gear icon. A context menu is open over the 'Job Subc17' icon, showing options: 'Add Icon', 'Arrange Icons', 'Pin', 'JOB\_SUBMISSION', 'JOB\_SPECIFICATION', 'SERVICE', and 'Native Icon'.

The 'CoG Log' window displays the following text:

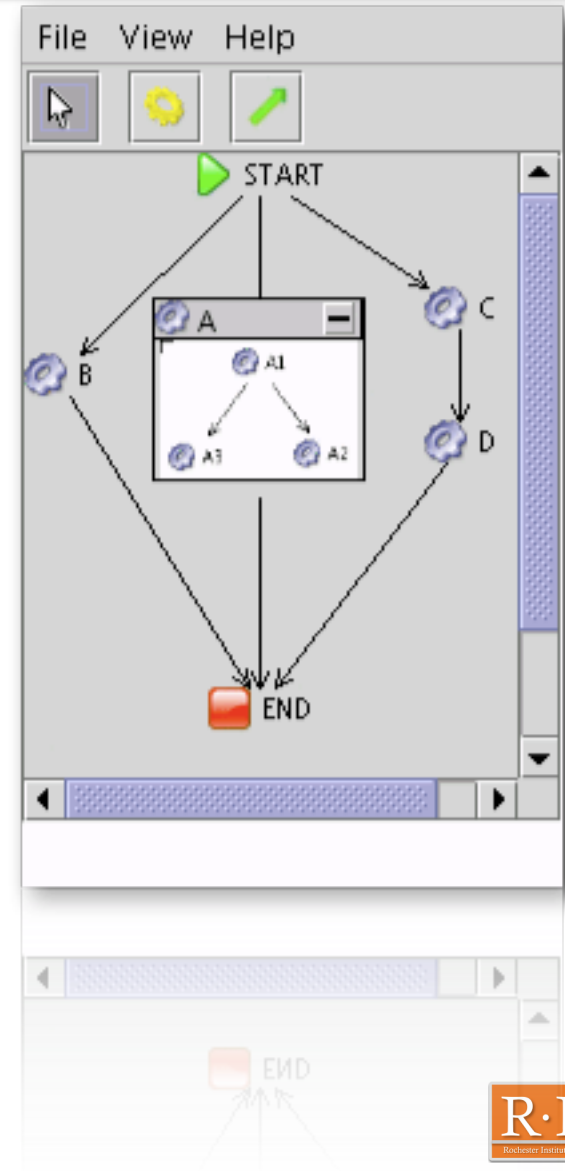
```

CoG Log
-----
Actions: Help
Info: Log started at 2004-10-20 14:13:53
Info: Status changed for command: umc-coq-1000001195810 to Submitted
Info: Status changed for command: umc-coq-1000001195810 to Active
Info: Status changed for command: umc-coq-1000001195810 to Completed
    
```

Below the desktop screenshot is a 'Java CoG Kit - Qtzr Worker' window showing a table of job monitoring data:

Monitor	JID	JIDTime	Monitor	Running	Nodes	State	Location	Mail	Process	Queue	Status
✓	51631	-	50:00:00	N/A	256	queued	N/A	on	512	batch	N/A
✓	51632	-	50:00:00	N/A	256	queued	N/A	on	512	batch	N/A
✓	51734	-	20:00:00	15:55:26	256	running	R032_L	on	256	default	07:06:...
✓	51741	-	40:00:00	N/A	128	queued	N/A	on	128	default	N/A
✓	51723	-	20:00:00	N/A	128	queued	N/A	on	256	default	N/A
✓	51831	-	30:00:00	17:20:26	128	running	R032_L	on	256	batch	07:06:...
✓	51834	-	40:00:00	17:14:52	128	running	R031_L	on	256	batch	07:06:...
✓	51834	-	30:00:00	17:11:50	128	running	R031_L	on	256	batch	07:06:...
✓	51835	-	30:00:00	17:56:45	128	running	R031_L	on	256	batch	07:06:...
✓	51842	-	20:00:00	N/A	32	queued	N/A	on	64	default	N/A
✓	51843	-	20:00:00	N/A	32	queued	N/A	on	1	batch	N/A
✓	51844	seconds...	20:00:00	N/A	32	queued	N/A	on	12	batch	N/A
✓	51845	-	20:00:00	N/A	128	queued	N/A	on	256	default	N/A
✓	51846	-	20:00:00	20:11:19	128	running	R032_L	on	256	batch	07:07:...
✓	51847	-	20:00:00	N/A	128	queued	N/A	on	256	batch	N/A

At the bottom of the window, it says: 'Job Monitoring by j.mca and g.pv. Refresh rate: 1 minute(s)'



# CoG Karajan Workflow

```
<project>
  <include file="cogkit.xml"/>
  <execute executable="/bin/date"
    stdout="thedata"
    host="hot.mcs.anl.gov" provider="GT2"/>
  <echo message="Job completed. Transferring the output"/>
  <transfer srchost="hot.mcs.anl.gov" srcfile="thedata"
    desthost="localhost" provider="gridftp"/>
  <echo message="Transfer complete"/>
  <set name="date">
    <readFile file="thedata"/>
  </set>
```



GT4

Swift

Swift

Software People do  
not talk about  
There are bugs

CoG Kit  
Workflow

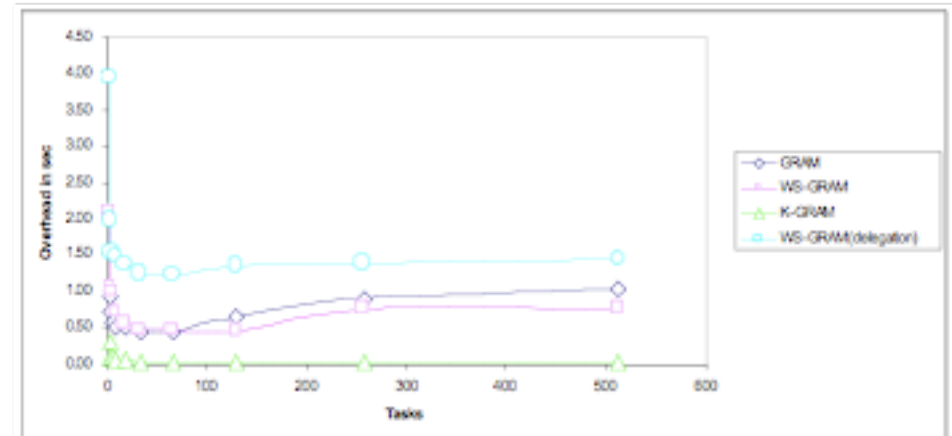
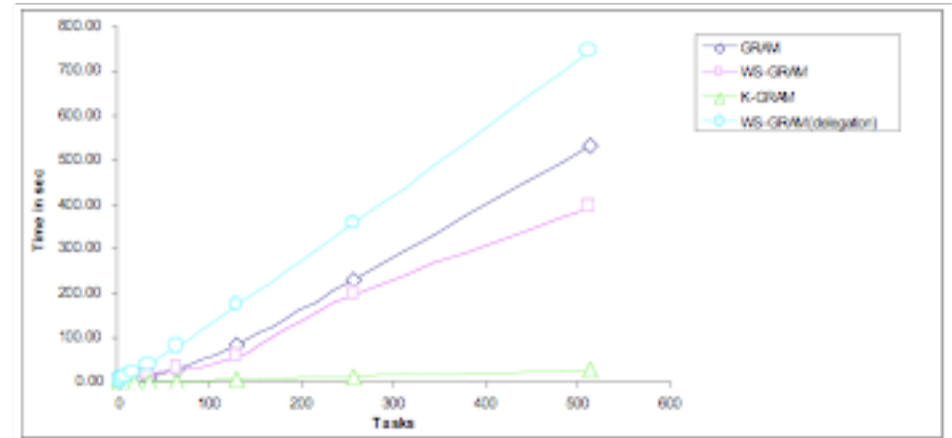
Globus

# Task vs Job

- We have developed frameworks doing millions of tasks before
- Because the Grid is available, some now try to use the Grid to map a task 1-to-1 to jobs
- Some call this Many-Task-Computing
  - This is not new
  - The basic construct of the CoG Kit is a “Task” not a job
- What is new:
  - Definition of a term
  - Wrong approach to Get Things Done
  - Use CoG Kits binning ability
  - CoG Kit Castor may replace Falcon

# Job Binning

- Example: Executing many jobs in parallel, a typical case in parameter studies
- CoG is 30 times faster than WS-GRAM
- For 512 0.05 seconds instead of 1.46 per job
- GRAM
  - WS-GRAM performance in second range per job
  - CoG Execution service for multiple jobs: milliseconds per job (scales well through light weight threading)



# Real Time Data Analysis

**Ubiquitous High-Speed Networked Environment**

**Heartbeat Monitor**

IP number	Hostname
68.245.76.14	maia.east.isi.edu
128.9.64.205	www.globus.org
128.9.64.206	hammie.isi.edu
128.32.36.63	beefix.CS.Berkeley.EDU
128.120.56.128	suffix.cs.ucdavis.edu
128.169.92.51	dasher.cs.utk.edu
128.230.51.195	nova.npac.syr.edu
131.215.145.136	sampson
132.239.51.91	perplex.ucsd.edu
140.221.3.110	tiamat.mcs.anl.gov

**Latency**

ISI, huntsman.isi.edu, bolas.isi.edu, tiamat.mcs.anl.gov

Latency values (ms): 1.177, 32.673, 44.221, 108.584, 0.043

**Throughput**

throughput Mbit/s

ISI, ANL, huntsman.isi.edu, dew.mcs.anl.gov, tiamat.mcs.anl.gov

Throughput values (Mbit/s): 6.001, 0.753, 0.799, 7.258, 1.313

Gregor von Laszewski, <http://www.mcs.anl.gov/gregor>



# If we only knew than ...

- what we know now ...
  - The Future
  - We are working on some of the components as we speak, while others are already available in prototype form.

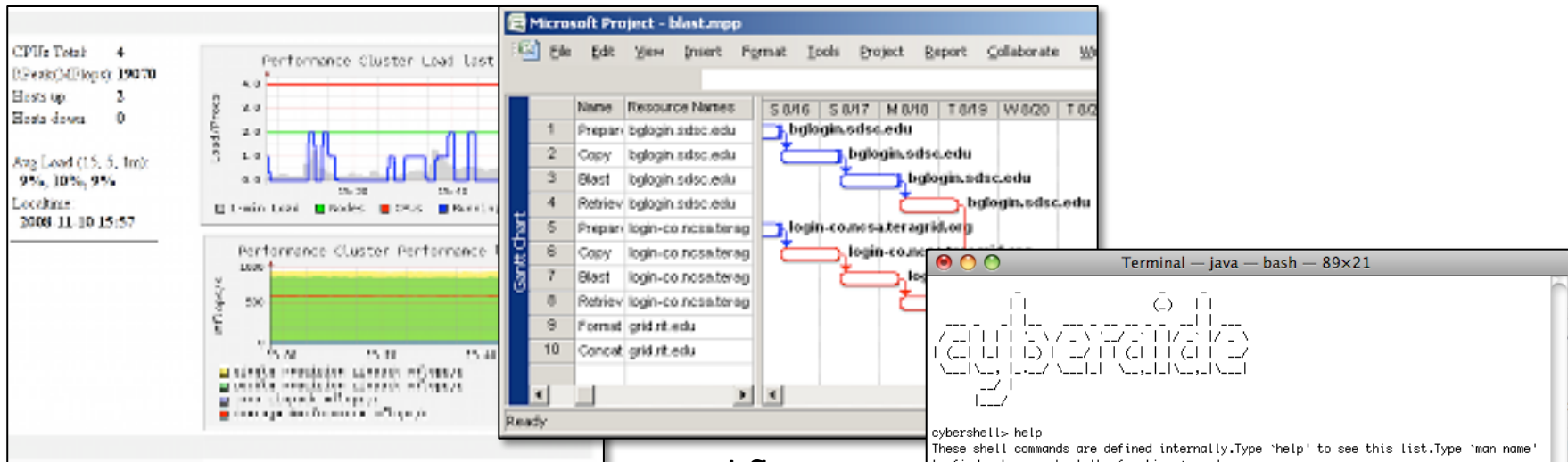
# 2008: Cyberaide Web 2.0

The screenshot displays the Grid Portal interface with several key components:

- Authentication:** A green box shows a successful login for 'myproxy server host' at 'myproxy.teragrid.org' with port '7512' and username 'myproxy'. The 'Authenticate' button is highlighted.
- Workflow Management:** A central window titled 'Historical Jobs Management' lists various workflow IDs and their results, such as '1625495365 [Results]: 1625495365' and '1953257326 [Results]: newdata'. A 'Submit workflow' button is visible.
- Resources:** A table lists available resources with columns for Type, SiteID, ResourceID, and Version. The table includes entries for 'gsi-openssh' at various sites like 'ncar.teragrid.org' and 'bigben.psc.teragrid.org'.
- Network Diagram:** A central diagram shows connections between sites: LA, Den, IU, TACC, IPGrid, Purdue, and SDSC. Data transfer rates are indicated on the links, such as '4.13 mb/s' between LA and Den, and '15.89 kb/s' between IPGrid and Purdue.
- Results:** A window shows the output for 'Workflow id: 819384657', listing queue names like 'workq', 'debug', and 'standard' along with their memory, CPU, and time usage.
- Teragrid News RSS:** A window displays a list of recent events and news, including 'Network Connectivity' and 'SDSC IA-64 cluster back in production'.

# Cyberaide

- Access advanced cyberinfrastructure easily



Application Monitoring

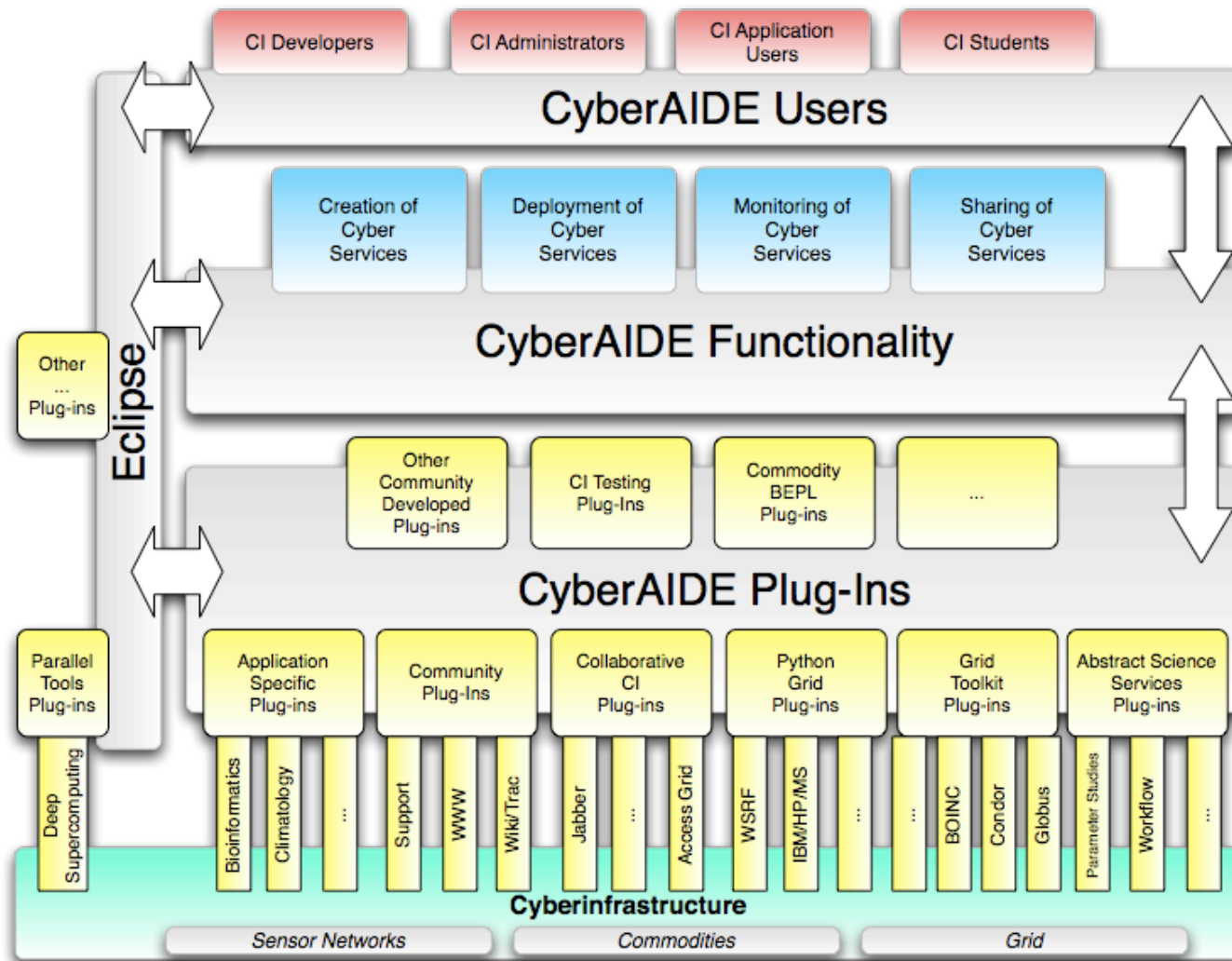
Workflow

Object Shell

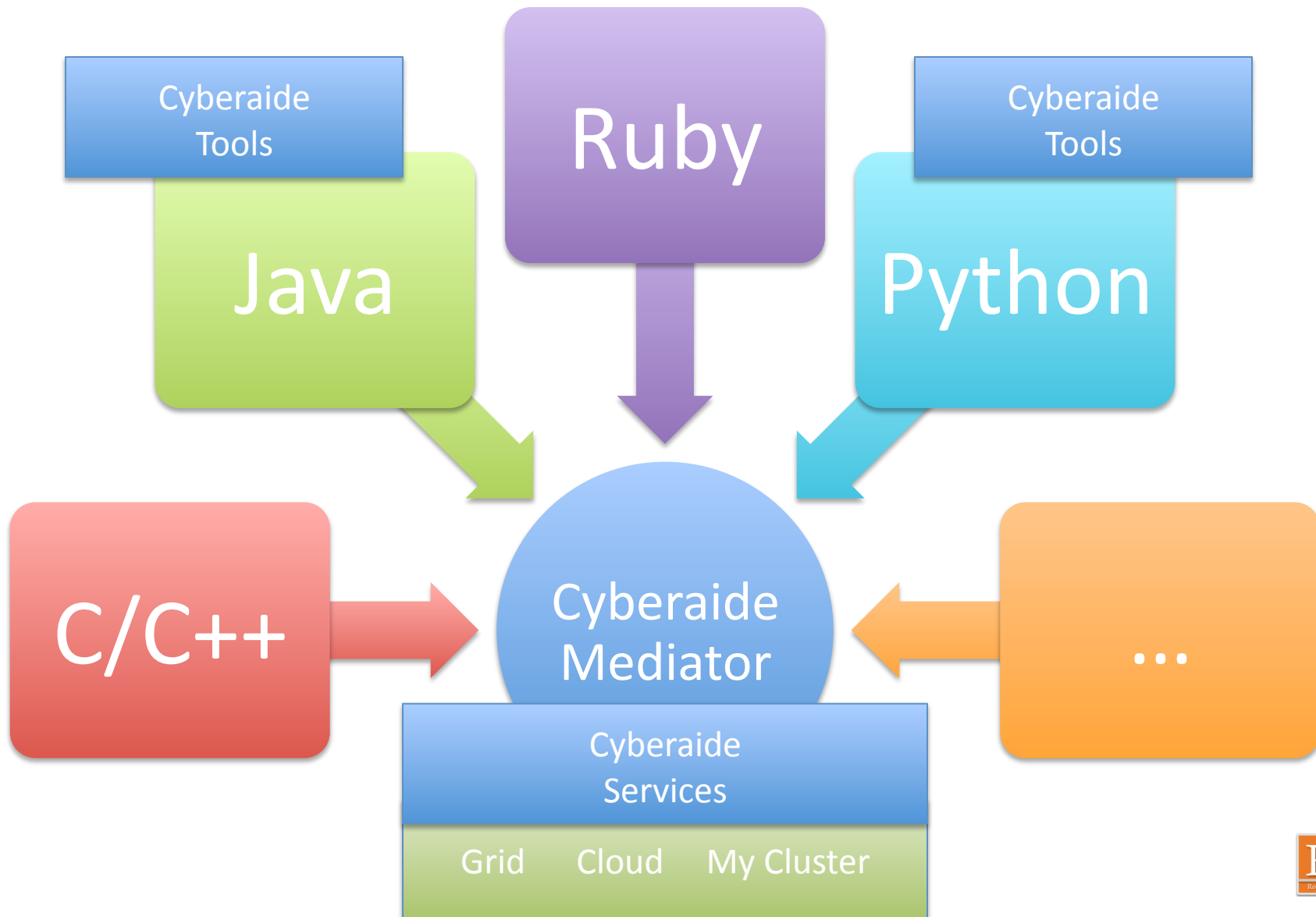


Shared Resource Calendar

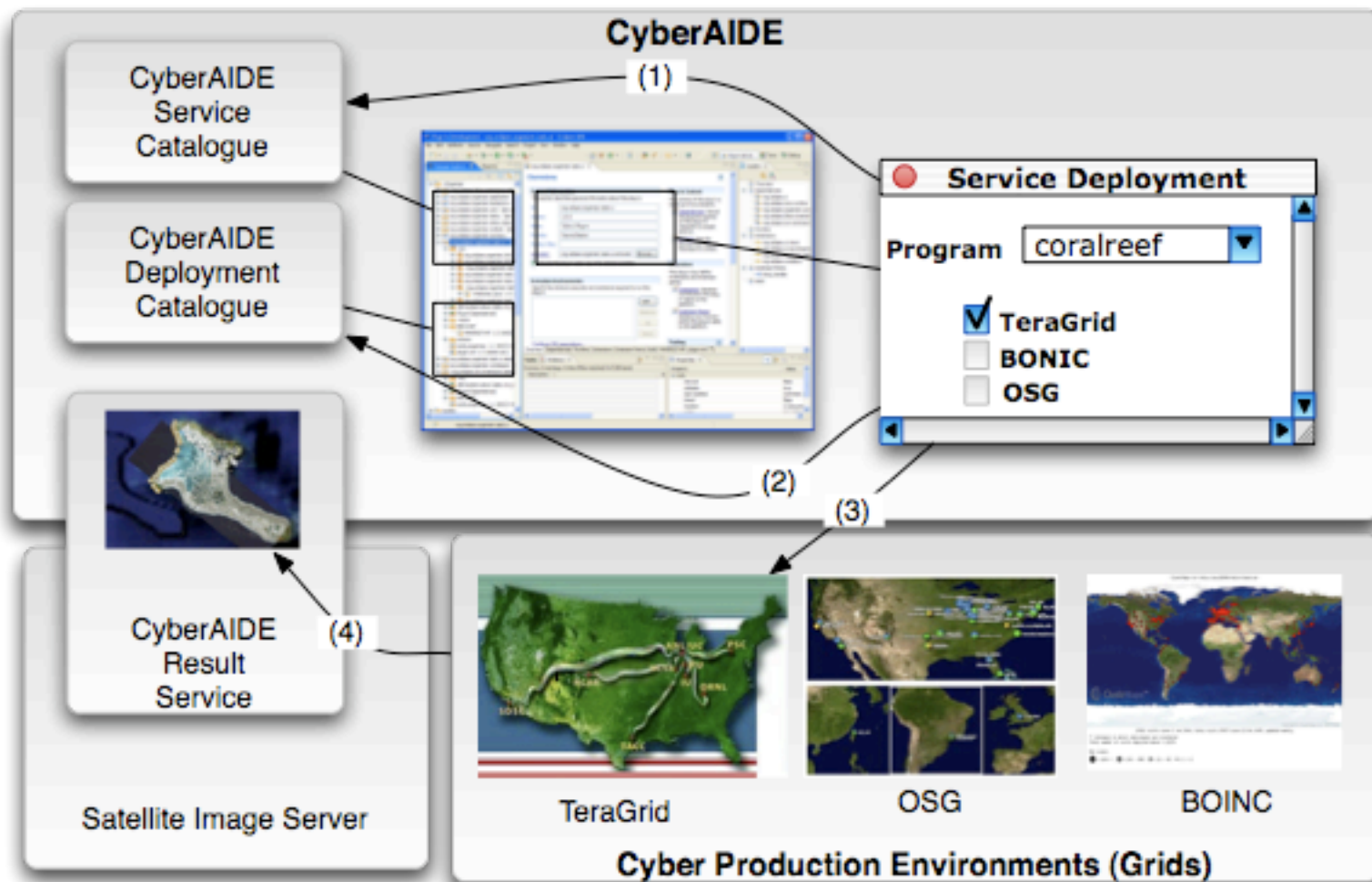
# Cyberaide Framework



# Mediating to the Future



# Real Time Science Infrastructure



# Politics, Administration, and other unfortunate hurdles

# Administration

- Separate domains
  - Don't touch my stuff
- Network
  - Network between buildings
  - Network to sectors
  - It was faster to send data from one building to the other via Internet2 ;-)
- Accounting
  - Account administration
  - It took 3 month and 12 administrators to get me an account, which included rewriting my job description for doing the same thing ;-)
  - Dans group needed after me only 3 days ;-)
  - 1 admin
- Security
  - Firewalls
    - Firewalls are expensive
    - Limit bandwidth below what is possible
  - HW/SW Approval process
    - Software development process is faster than approval process -> always old version is installed
    - HW approval process was slow
    - No money for new HW
      - Here, let me give you my Windows 3.1 OS box that I have not used anymore for that last x years but its still here
      - That is no good



# User Demand Contradictions

- Easy access and no security
  - vs. don't give my data out
- I need lots of data vs.
  - I can not analyze the data due to lack of compute power
- Privacy
  - vs. I don't give my data but you sure should give me yours
- Do things quickly for me
  - vs. do things right for many users
- Let me focus on science
  - vs. but I can program this all myself

# Conclusion

- Globus is a Middleware Toolkit
  - We need more
- We need upperware
  - Cyberaide is just one example
- Administrative hurdles are significant
- Security is underestimated
- Networks are too slow
  - If you can string one wire, why not just put in 5 ;-)

# Collaborate

- We are interested to collaborate with you.
- Send e-mail to [laszewski@gmail.com](mailto:laszewski@gmail.com)
- <http://www.cyberaide.org>