CineGrid GRID & Networking

Cees de Laat

University of Amsterdam

With grid slides thanks to David Groep (NIKHEF)

CineGrid Mission

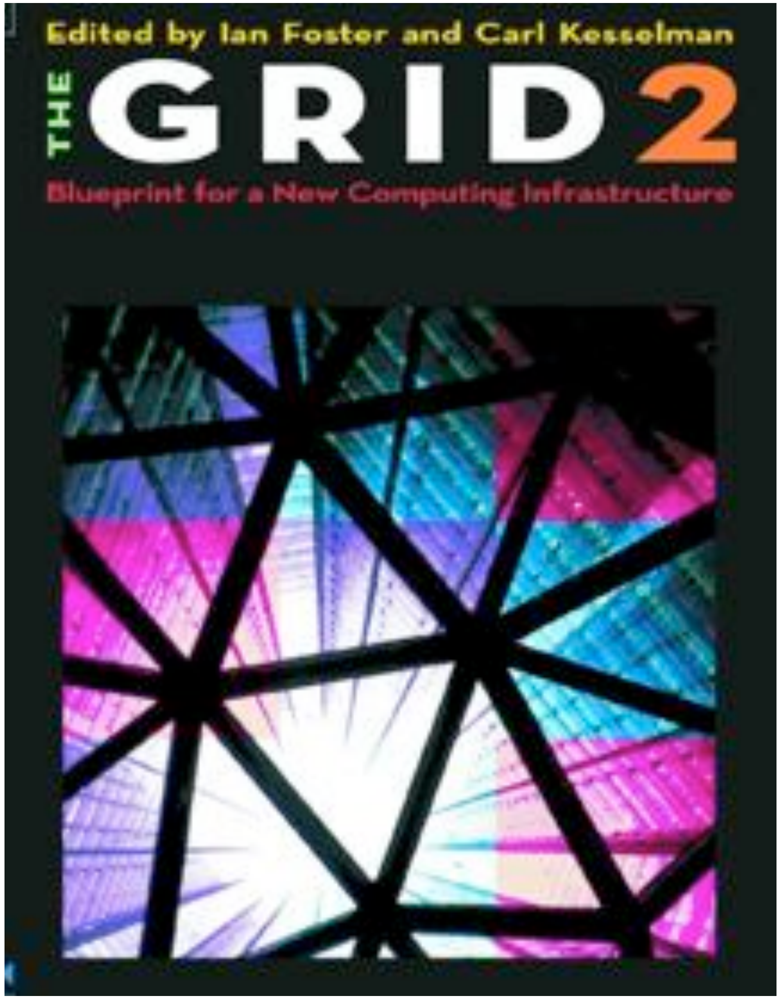
To build an interdisciplinary community that is focused on the research, development, and demonstration of networked collaborative tools to enable the production, use and exchange of very-high-quality digital media over photonic networks.

http://www.cinegrid.org/

The Grid

- Grid 'coined' in 1997 by
  - Carl Kesselman (ISI/USC) and
  - Ian Foster (ANL)

- builds on a tradition of distributed computing
  - 1969: Creaper & Reaper
  - 1978: RPC concept
  - 1985: Condor
  - 1991: CORBA
  - 1991: DCE/DFS


Edited by Ian Foster and Carl Kesselman
THE GRID 2
Blueprint for a New Computing Infrastructure

Exploding Data

Collected data in science (and industry) grows exponentially:

The Bible	**5**	**MByte**
X-ray image	5	MByte/day
Functional MRI	1	GByte/day
Bio-informatics databases	500	GByte each
Refereed journal papers	1	TByte/yr
Satellite world imagery	5	TByte/yr
US LoC contents	20	TByte
Internet Archive 1996-2002	100	TByte
Particle Physics today	1	PByte/yr
LHC era physics	**20**	**PByte/yr**

The Grid label

Many distributed computing middlewares are now called “grid”

- Oracle 10*g*
- BOINC (formerly SETI@home)
- Sun Grid Engine
- Unicore
- Globus Toolkit 4
- gLite
- …

And then there is middleware to build grids
that is not usually branded as such

- Condor
- …

But they are not all that ‘griddy’

- ~~Oracle 10*g*~~ = database on a cluster with node function changes
- BOINC (formerly SETI@home) = single application client/server
- Sun Grid Engine = cluster batch system
- Unicore
- Globus Toolkit 4
- gLite
- …

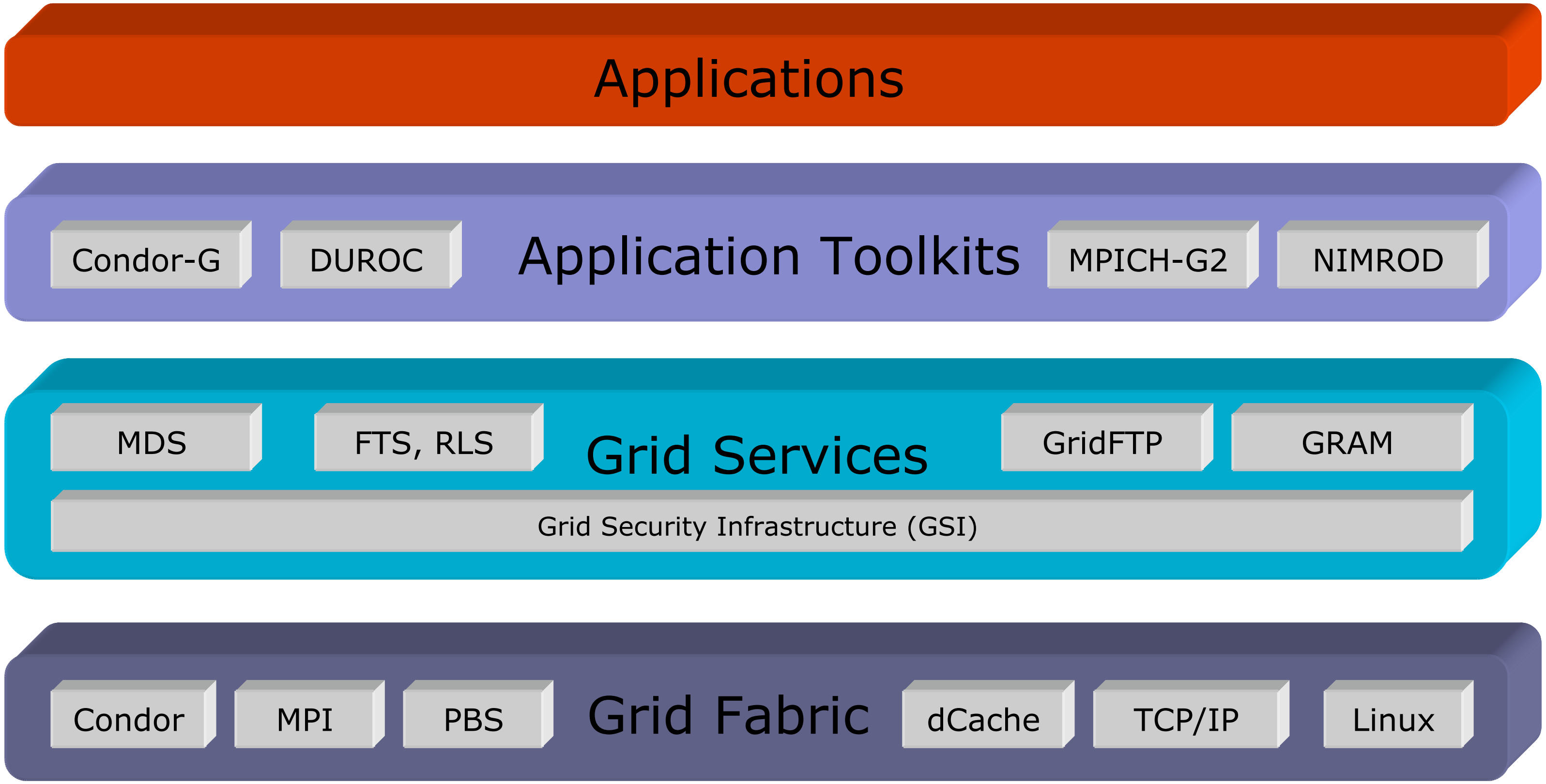
- Condor
- …

Grid Middleware

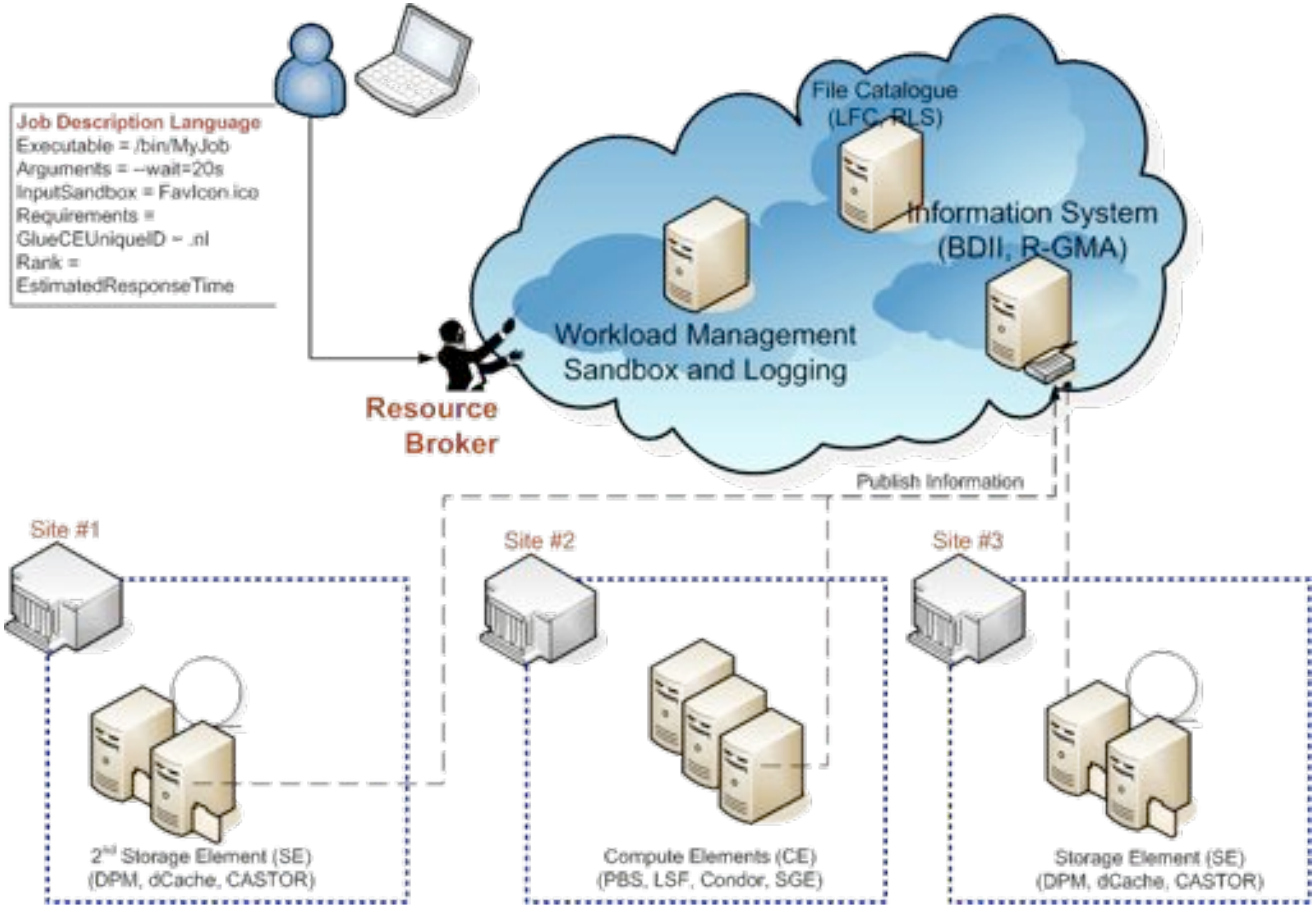
- Software that enables Grid
  - term deliberately vague, like the term Grid itself
  - but, from experience, one needs at least these services
    - resource discovery
    - resource scheduling
    - uniform compute access
    - uniform data access (to both files and structured data)
    - asynchronous information sources
    - authentication, delegation and secure communications
    - identity management
    - system management and system access
  - and these services should have a standard, common, interface

- In general, ‘middleware’ is used to describe the layer between network and application

Grid Middleware and their position



Typical Grid Topology



Job Description Language

This is JDL that the user might send to the Resource Broker

```
Executable         = "catfiles.sh";
StdOutput          = "catted.out";
StdError           = "std.err";
Arguments          = "EssentialJobData.txt
                         LogicalJobs.jdl /etc/motd";

InputSandbox       = {"/home/davidg/tmp/jobs/LogicalJobs.jdl",
                      "/home/davidg/tmp/jobs/catfiles.sh" };
OutputSandBox      = {"catted.out", "std.err"};

InputData          = "LF:EssentialJobData.txt";
ReplicaCatalog     =
              "ldap://rls.edg.org/lc=WPSIX,dc=cnrs,dc=fr";
DataAccessProtocol = “gsiftp";

RetryCount         = 2;
```

How to you see what’s in the Grid?

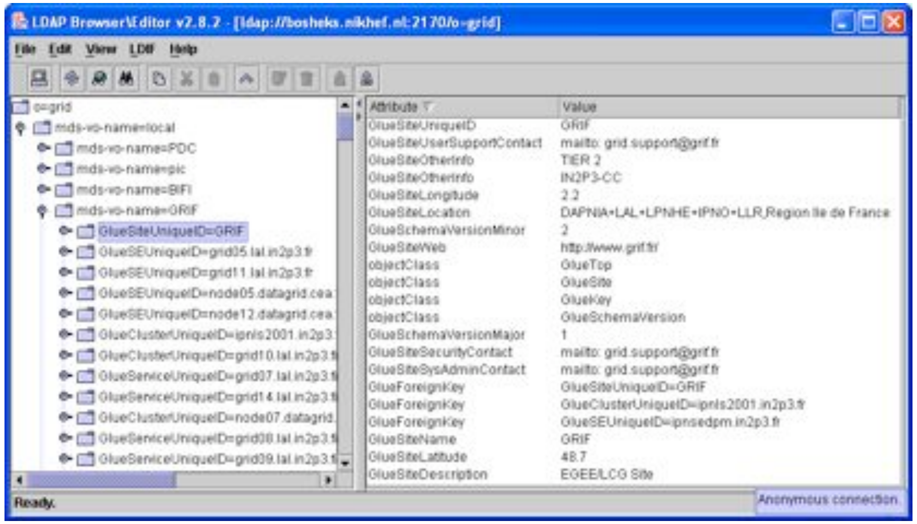
Broker matches the user’s request with the site

- ‘information supermarket’ matchmaking (using Condor Matchmaking)
- uses the information published by the site

Grid Information system

‘*the only information a user ever gets about a site’*

- So: should be reliable, consistent and complete
- Standard schema (GLUE) to describe sites, queues, storage *(complex schema semantics)*
- Currently presented as an LDAP directory


LDAP Browser\Editor v2.8.2 - [ldap://bosheks.nikhef.nl:2170/o=grid]
File Edit View LDIF Help
o=grid
mds-vo-name=local
mds-vo-name=PDC
mds-vo-name=pic
mds-vo-name=BIFI
mds-vo-name=GRIF
GlueSiteUniqueID=GRIF
GlueSEUniqueID=grid05.lal.in2p3.fr
GlueSEUniqueID=grid11.lal.in2p3.fr
GlueSEUniqueID=node05.datagrid.cea
GlueSEUniqueID=node12.datagrid.cea
GlueClusterUniqueID=ipnls2001.in2p3
GlueClusterUniqueID=grid10.lal.in2p3.f
GlueServiceUniqueID=grid07.lal.in2p3.f
GlueServiceUniqueID=grid14.lal.in2p3.f
GlueClusterUniqueID=node07.datagrid
GlueServiceUniqueID=grid08.lal.in2p3.f
GlueServiceUniqueID=grid09.lal.in2p3.f
Ready.
Anonymous connection.

Attribute	Value
GlueSiteUniqueID	GRIF
GlueSiteUserSupportContact	mailto: grid.support@grif.fr
GlueSiteOtherInfo	TIER 2
GlueSiteOtherInfo	IN2P3-CC
GlueSiteLongitude	2.2
GlueSiteLocation	DAPNIA+LAL+LPNHE+IPNO+LLR,Region Ile de France
GlueSchemaVersionMinor	2
GlueSiteWeb	http://www.grif.fr/
objectClass	GlueTop
objectClass	GlueSite
objectClass	GlueKey
objectClass	GlueSchemaVersion
GlueSchemaVersionMajor	1
GlueSiteSecurityContact	mailto: grid.support@grif.fr
GlueSiteSysAdminContact	mailto: grid.support@grif.fr
GlueForeignKey	GlueSiteUniqueID=GRIF
GlueForeignKey	GlueClusterUniqueID=ipnls2001.in2p3.fr
GlueForeignKey	GlueSEUniqueID=ipnsedpm.in2p3.fr
GlueSiteName	GRIF
GlueSiteLatitude	48.7
GlueSiteDescription	EGEE/LCG Site

LDAP Browser Jarek Gawor: www.mcs.anl.gov/~gawor/ldap

Attributes set per Site

- Site information
  - SiteSysAdminContact: mailto: grid-admin@example.org
  - SiteSecurityContact: mailto: security@example.org
- Cluster info

GlueSubClusterUniqueID=gridgate.cs.tcd.ie

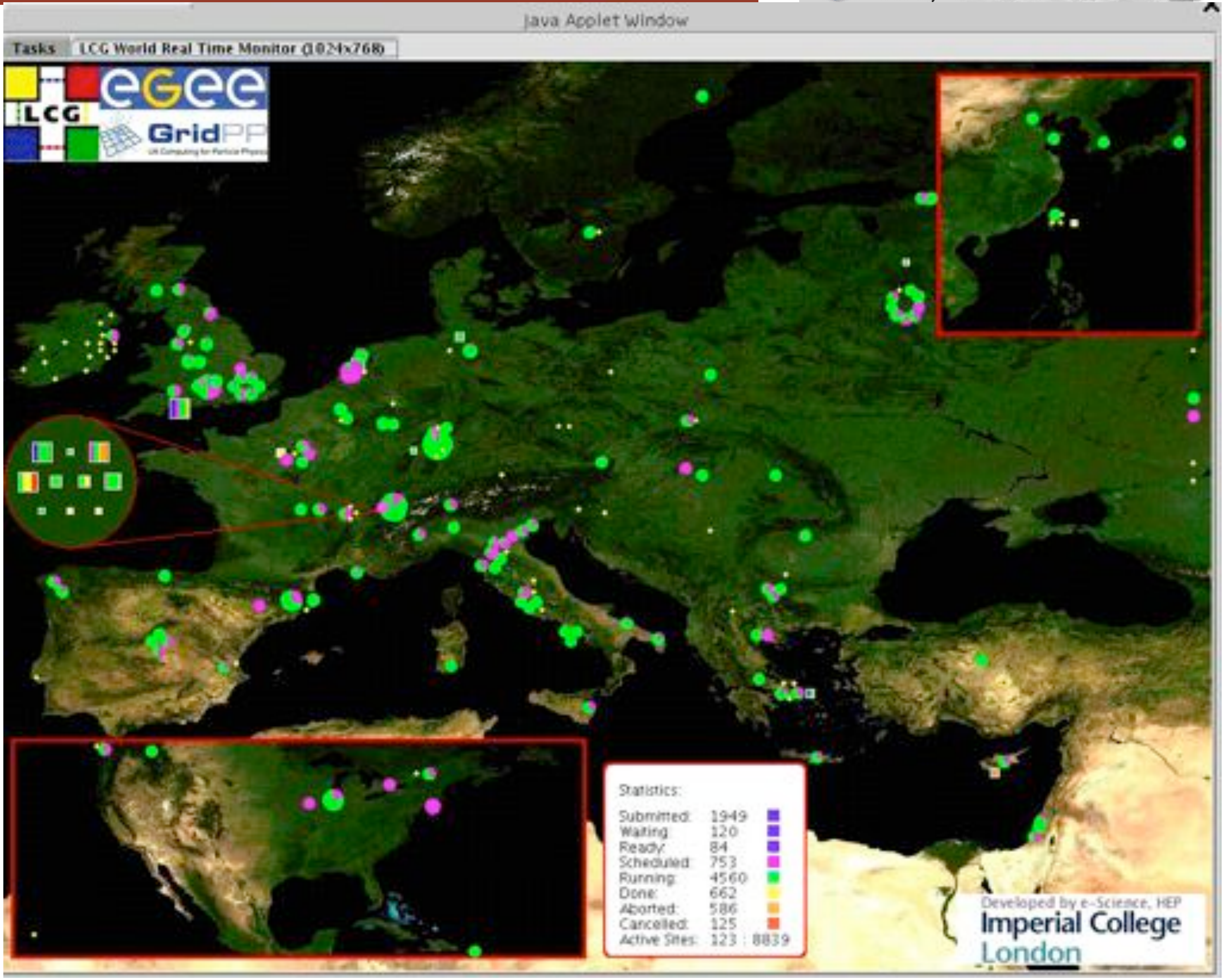
```
HostApplicationSoftwareRunTimeEnvironment: LCG-2_6_0
HostApplicationSoftwareRunTimeEnvironment: VO-atlas-release-10.0.4
HostBenchmarkSI00: 1300
GlueHostNetworkAdapterInboundIP: FALSE
GlueHostNetworkAdapterOutboundIP: TRUE
GlueHostOperatingSystemName: RHEL
GlueHostOperatingSystemRelease: 3.5
GlueHostOperatingSystemVersion: 3

GlueCEStateEstimatedResponseTime: 519
GlueCEStateRunningJobs: 175
GlueCEStateTotalJobs: 248
```

- Storage: similar info (paths, max number of files, quota, retention, …)

Grid in operation

LCG Real Time Monitor, Gidon Moont, RAL GOC


Java Applet Window
Tasks
LCG World Real Time Monitor (1024x768)
eGee
LCG
GridPP
UK Computing for Particle Physics
Developed by e-Science, HEP
Imperial College
London

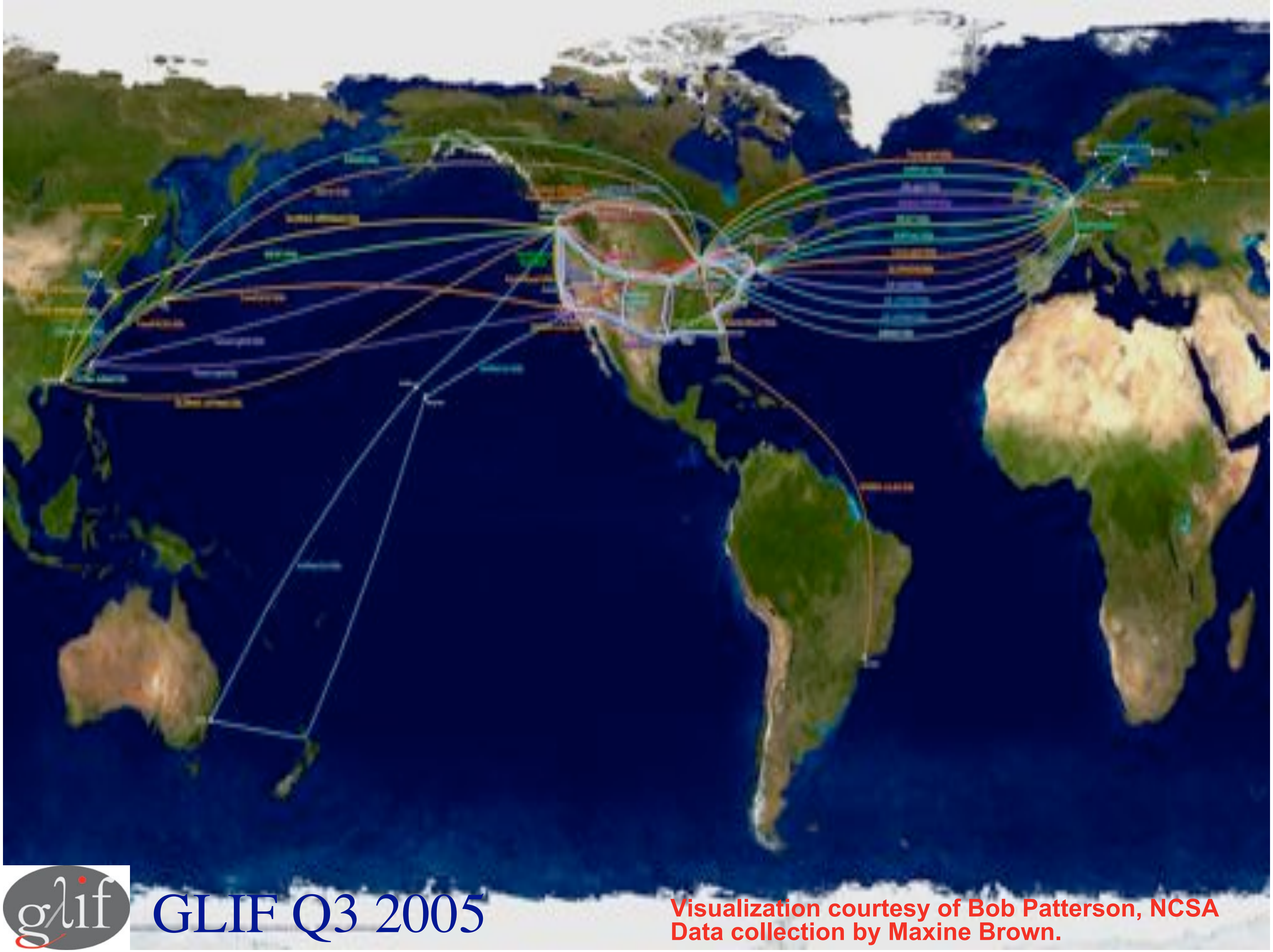
Statistics:

Submitted:	1949	🟦
Waiting:	120	🟦
Ready:	84	🟦
Scheduled:	753	🟪
Running:	4560	🟩
Done:	662	🟨
Aborted:	586	🟧
Cancelled:	125	🟥
Active Sites:	123 : 8839	

Format - Numbers - Bits (examples!)

Format	X	Y	Rate	Color bits/pix	Frame pix	Frame MByte	Flow MByt/s	Stream Gbit/s
720p HD	1280	720	60	24	921600	2.8	170	1.3
1080p HD	1920	1080	30	24	2073600	6.2	190	1.5
2k	2048	1080	24 48	36	2211840	10	240 480	1.2 2.4
SHD	3840	2160	30	24	8294400	25	750	6.0
4k	4096	2160	24	36	8847360	40	960	7.6

Note: this is excluding sound!
Note: these are raw uncompressed data rates!

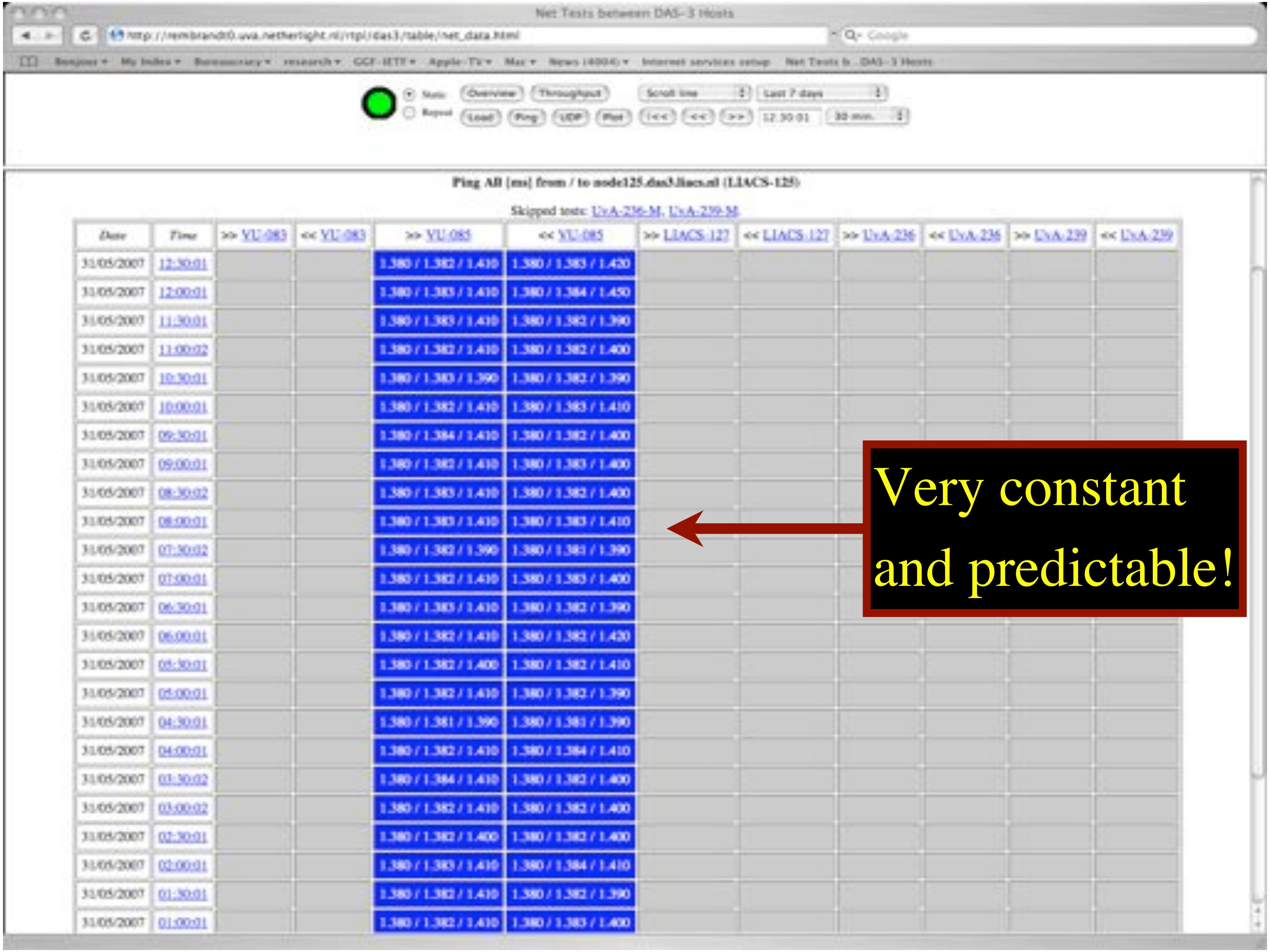


GLIF Q3 2005

Visualization courtesy of Bob Patterson, NCSA
Data collection by Maxine Brown.

What is a LightPath

- A LightPath is a circuit like connection that connects end systems to each other. This uses usually the same infrastructure as the Internet, but a LightPath gets dedicated resources next to Internet.
- A LightPath can be a combination of:
  - A color in a fiber (Lambda)
  - Sonet/sdh circuit in a sonet infrastructure
  - Vlans and dedicated ports in an ethernet switch
  - Etc.
- Aim is to get predictable and knowable connection characteristics

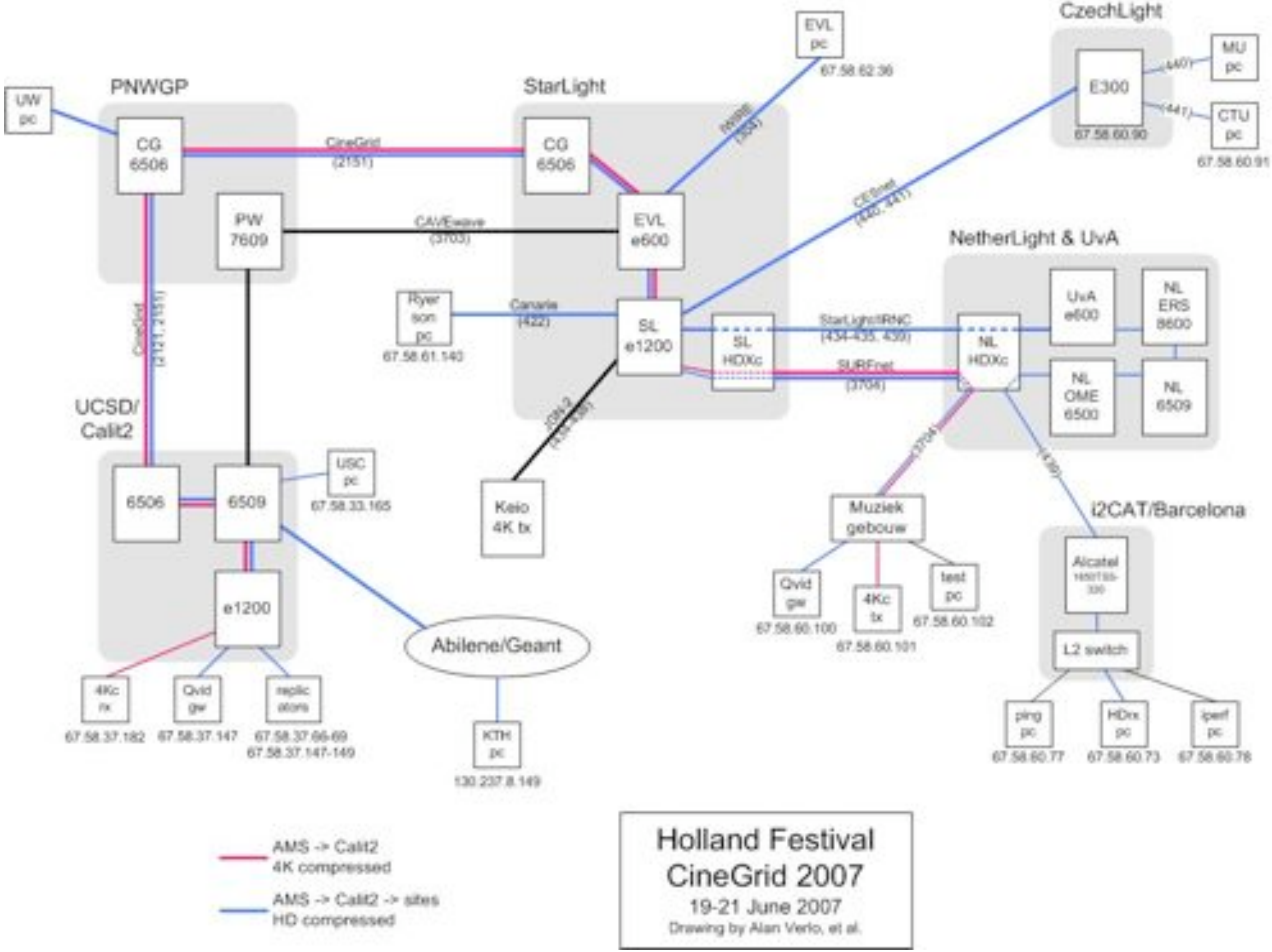

Net Tests between DAS-3 Hosts
http://rembrandt0.uva.netherlight.nl/rtpi/das3/table/net_data.html
Google
Bonjour ▾ My Index ▾ Bureaucracy ▾ research ▾ GGF-IETF ▾ Apple-TV ▾ Mac ▾ News (4004) ▾ Internet services setup Net Tests b...DAS-3 Hosts
◉ Static ○ Repeat
Overview
Throughput
Load
Ping
UDP
Plot
Scroll time
Last 7 days
|<<
<<
>>
12:30:01
30 min.

Ping All [ms] from / to node125.das3.liacs.nl (LIACS-125)

Skipped tests: UvA-236-M, UvA-239-M.

Date	*Time*	>> VU-083	<< VU-083	>> VU-085	<< VU-085	>> LIACS-127	<< LIACS-127	>> UvA-236	<< UvA-236	>> UvA-239	<< UvA-239
31/05/2007	12:30:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.420						
31/05/2007	12:00:01			1.380 / 1.383 / 1.410	1.380 / 1.384 / 1.450						
31/05/2007	11:30:01			1.380 / 1.383 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	11:00:02			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	10:30:01			1.380 / 1.383 / 1.390	1.380 / 1.382 / 1.390						
31/05/2007	10:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.410						
31/05/2007	09:30:01			1.380 / 1.384 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	09:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.400						
31/05/2007	08:30:02			1.380 / 1.383 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	08:00:01			1.380 / 1.383 / 1.410	1.380 / 1.383 / 1.410						
31/05/2007	07:30:02			1.380 / 1.382 / 1.390	1.380 / 1.381 / 1.390						
31/05/2007	07:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.400						
31/05/2007	06:30:01			1.380 / 1.383 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	06:00:01			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.420						
31/05/2007	05:30:01			1.380 / 1.382 / 1.400	1.380 / 1.382 / 1.410						
31/05/2007	05:00:01			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	04:30:01			1.380 / 1.381 / 1.390	1.380 / 1.381 / 1.390						
31/05/2007	04:00:01			1.380 / 1.382 / 1.410	1.380 / 1.384 / 1.410						
31/05/2007	03:30:02			1.380 / 1.384 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	03:00:02			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	02:30:01			1.380 / 1.382 / 1.400	1.380 / 1.382 / 1.400						
31/05/2007	02:00:01			1.380 / 1.383 / 1.410	1.380 / 1.384 / 1.410						
31/05/2007	01:30:01			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	01:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.400						

Very constant and predictable!



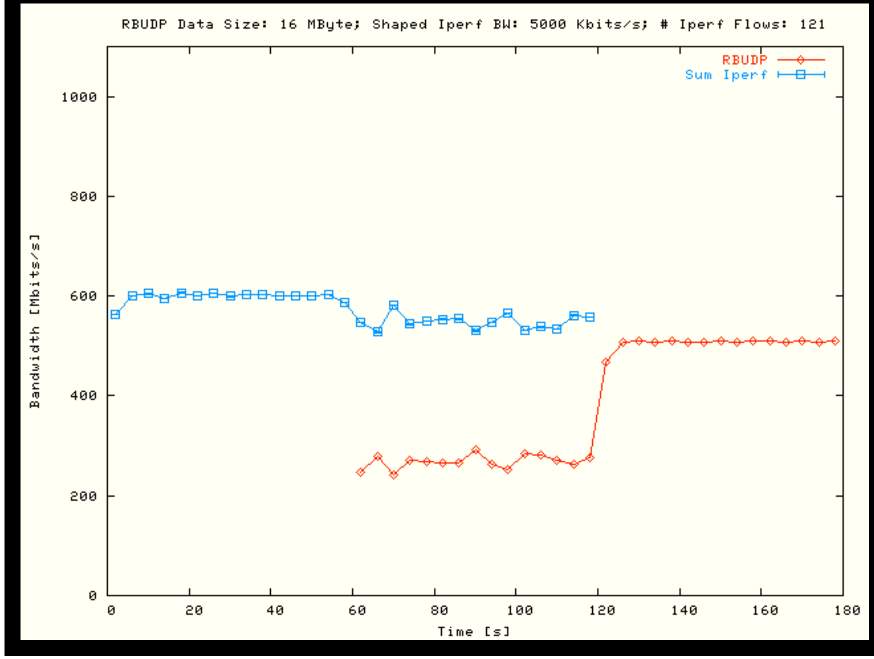
Holland Festival
CineGrid 2007
19-21 June 2007
Drawing by Alan Verlo, et al.

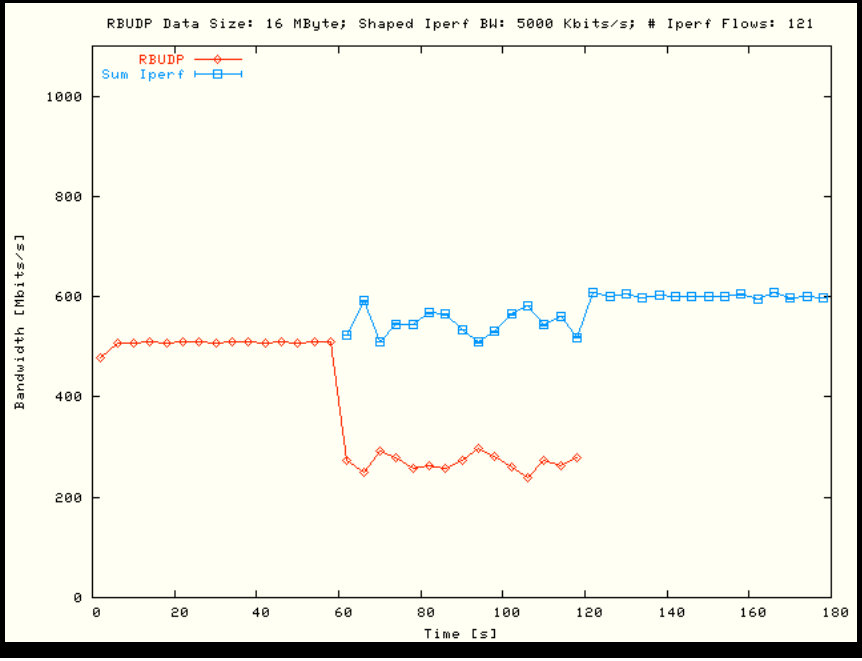
Internet Transport Protocols

- IP = Internet Protocol
  - Connectionless packet transport service
  - Datagrams of max 64 kByte that can be fragmented down the way
  - Packets can get lost, duplicated or out of order!
- TCP/IP = Transmission Control Protocol
  - Reliable byte-stream over potentially unreliable packet service
  - Connection oriented, exactly once and in order, end to end duplex
- UDP = User Datagram Protocol
  - Packet service up to 64 kByte
  - Connectionless, unidirectional, L2 switches may start flooding
  - Unreliable delivery, can get out of order, duplicated, lost

Issues & protocols

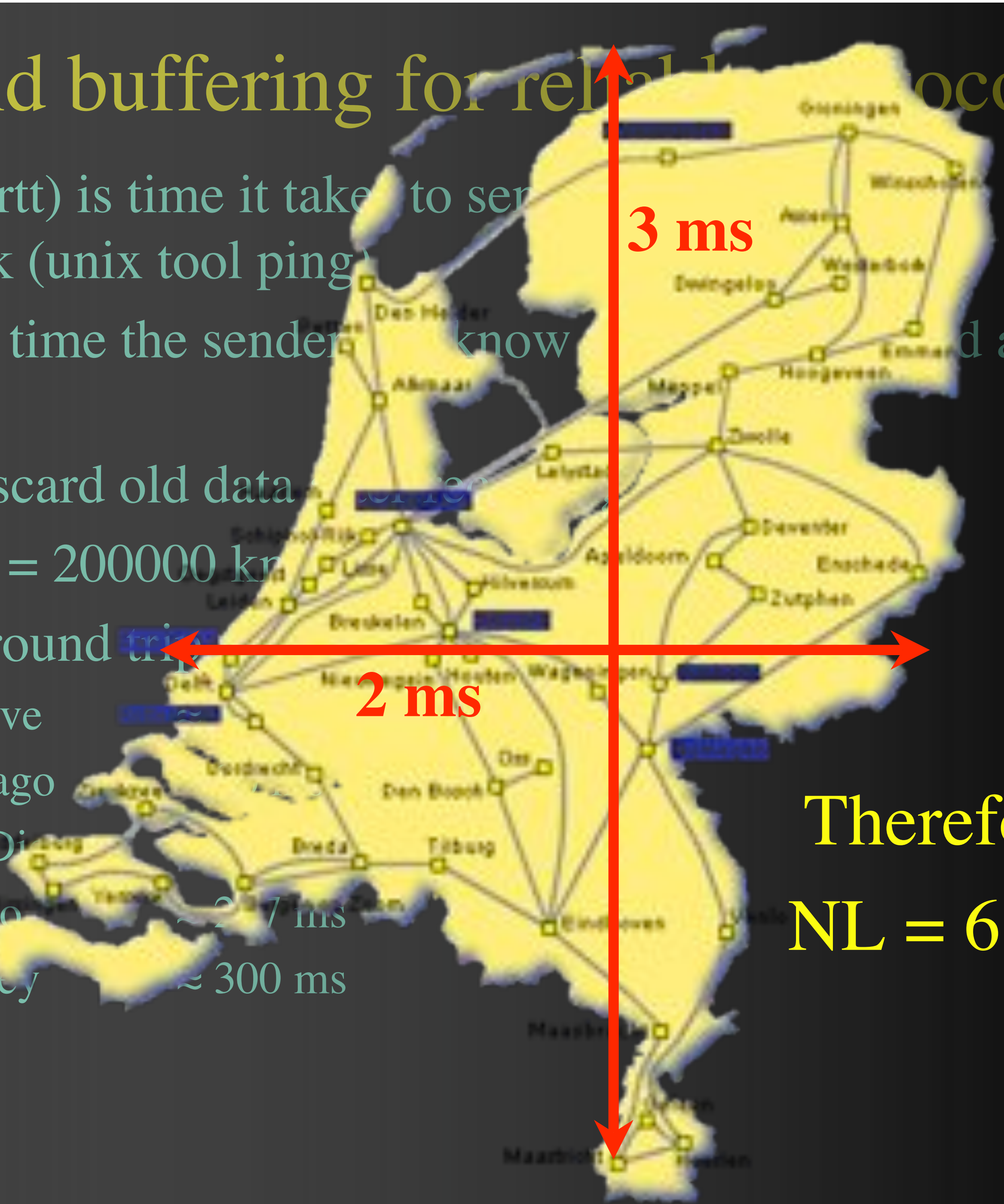
- When using UDP watch for bottleneck!
- About 10 other non standard protocols
- FAST TCP
  - Modified receiver algorithms
- RBUDP
  - Runs on top of UDP, simple back-off and retransmission scheme





Windows and buffering for rel ocols

- Round Trip Time (rtt) is time it take to ser and get the answer back (unix tool ping
- That is the shortest time the sender know d at the other end
- Sender can only discard old data
- Lightspeed in fiber = 200000 kn
- 100 km = 200 km round trip
  - Amsterdam - Geneve
  - Amsterdam - Chicago
  - Amsterdam - San Di
  - Amsterdam - Tokyo ≈ 2 7 ms
  - Amsterdam - Sydney ≈ 300 ms



Therefore:

NL = 6 ms^2

Buffer space

Window = RTT * BW

RTT	100 Mbit/s	1 Gbit/s	10 Gbit/s
1	12.5 kB	125 kB	1.25 MB
2	25 kB	250 kB	2.5 MB
5	62.5 kB	615 kB	6.15 MB
10	125 kB	1.25 MB	12.5 MB
20	250 kB	2.5 MB	25 MB
50	625 kB	6.25 MB	62.5 MB
100	1.25 MB	12.5 MB	125 MB
200	2.5 MB	25 MB	250 MB
500	6.25 MB	62.5 MB	625 MB
1000	12.5 MB	125 MB	1250 MB

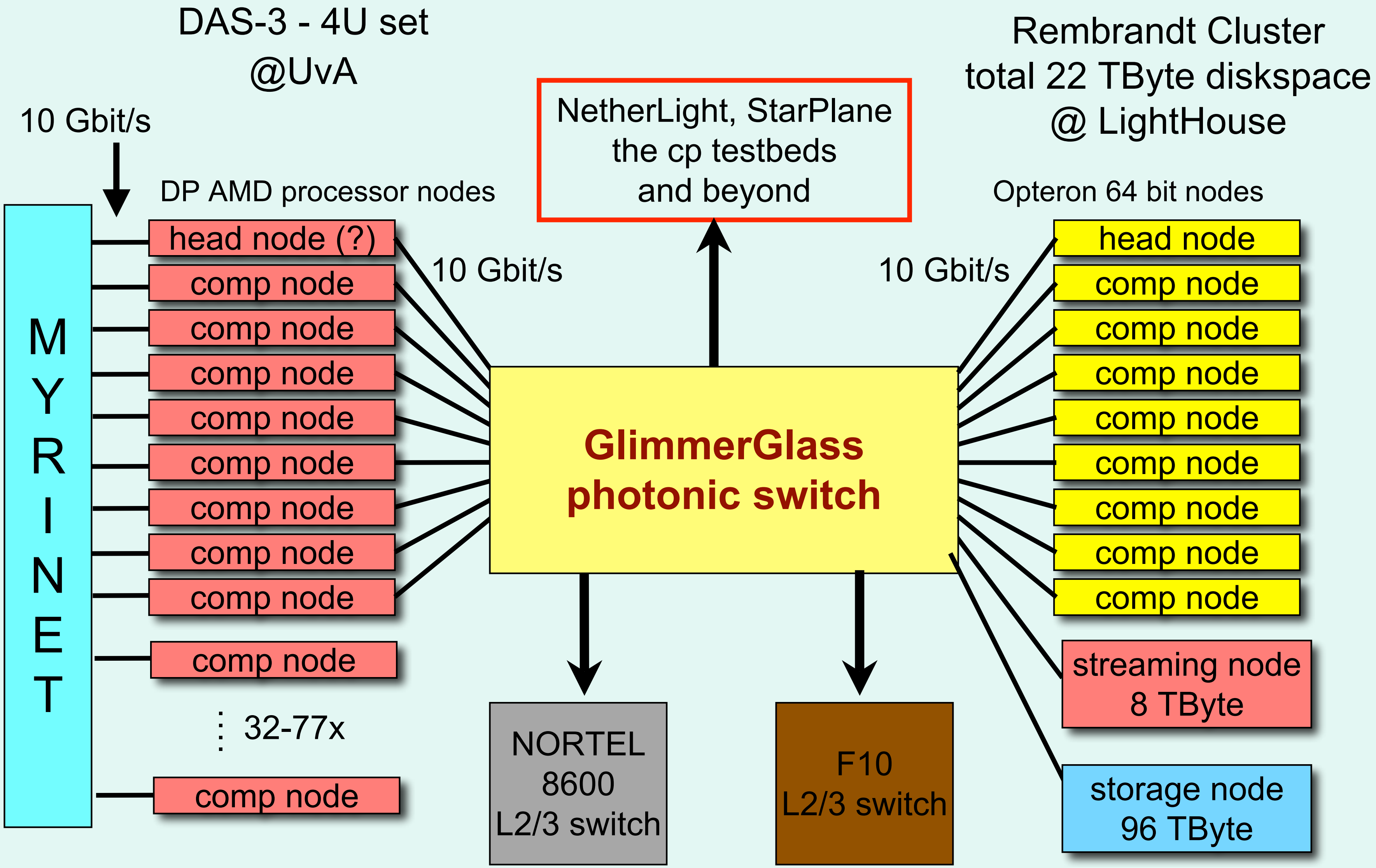
TCP Tuning (if not auto-tuning)

- 1 Gbit/s on 160 ms RTT (= Amsterdam - San Diego) :
  - sysctl -w kern.ipc.maxsockbuf=50000000
  - sysctl -w net.inet.tcp.sendspace=21000000
  - sysctl -w net.inet.tcp.recvspace=21000000
  - sysctl -w net.inet.udp.maxdgram=57344
  - sysctl -w net.inet.udp.recvspace=74848
  - sysctl -w net.local.stream.sendspace=32768
  - sysctl -w net.local.stream.recvspace=32768
  - sysctl -w kern.ipc.somaxconn=512
  - sysctl -w net.inet.tcp.mssdflt=1460
  - sysctl -w net.inet.tcp.delayed_ack=2
  - sysctl -w net.inet.tcp.rfc1323=1
  - sysctl -w net.inet.tcp.rfc1644=1
  - sysctl -w net.inet.tcp.newreno=1

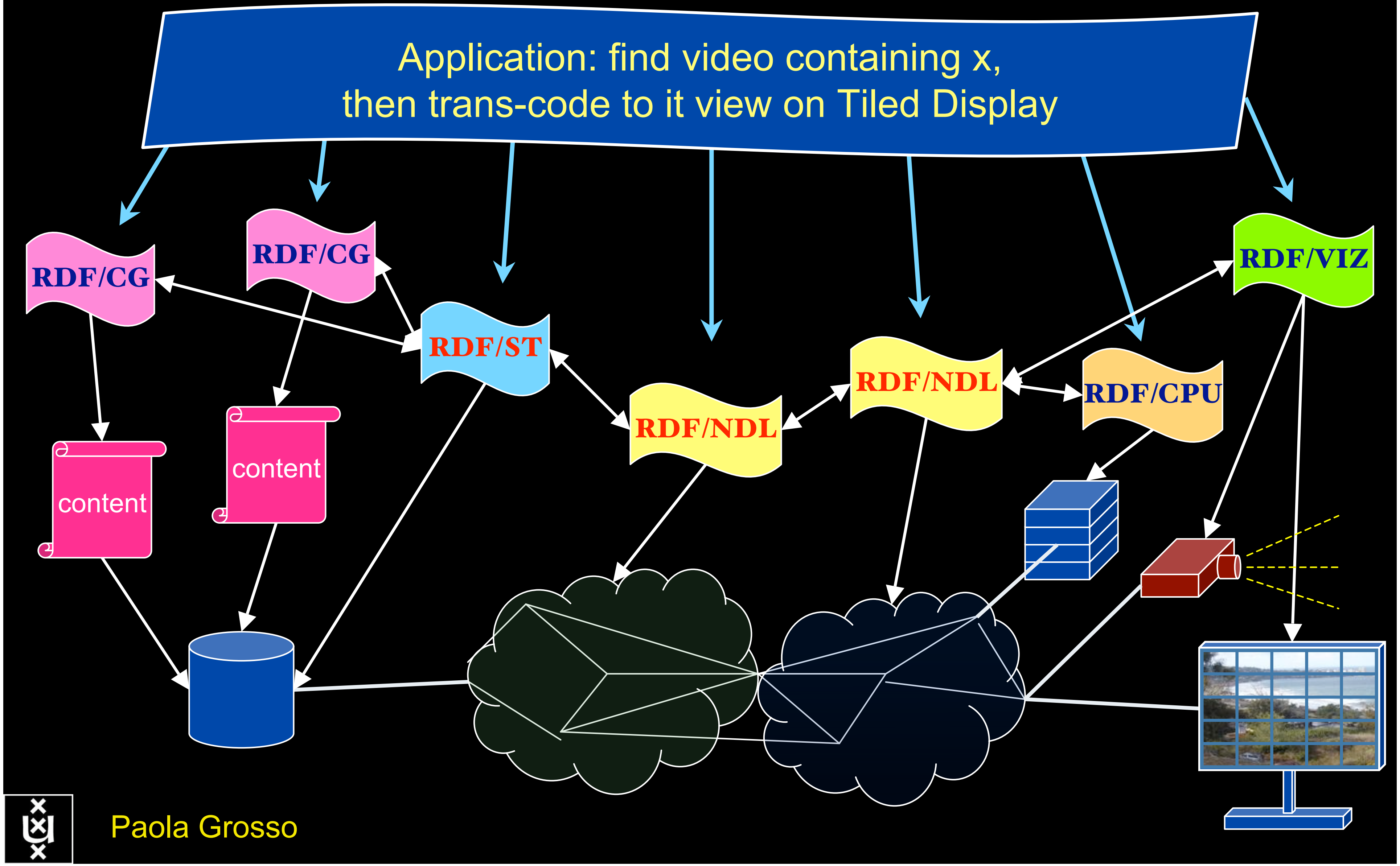
End System Issues

- Ethernet card interface to computer bus system
  - PCI-X
    - 32/64 bit 66/133/266 MHZ -> about 8 Gbit/s max in 133 MHZ mode
  - PCI-Express
    - 2.5 Gbit/s per lane, 4, 8, 16 lanes
- Memory organization
- CPU cache
  - Effect when things go out of cache (small windows, etc.)
- CPU core
  - Takes 1 core to handle network (affinity may help)
- Disk raid subsystem
  - raid0 twice as fast as raid5
  - One disk does typically 40 MB/s write, 60 MB/s read

Amsterdam CineGrid S/F node



RDF describing Infrastructure


Application: find video containing x,
then trans-code to it view on Tiled Display
RDF/CG
RDF/CG
RDF/ST
RDF/NDL
RDF/NDL
RDF/CPU
RDF/VIZ
content
content

Paola Grosso

Questions ?

www.cinegrid.org

www.cinegrid.nl

www.supertube.org

www.science.uva.nl/~delaat