Networking and the Grid

Ahmed Abdelrahim
NeSC

PPARC e-Science Summer School
10th May 2005

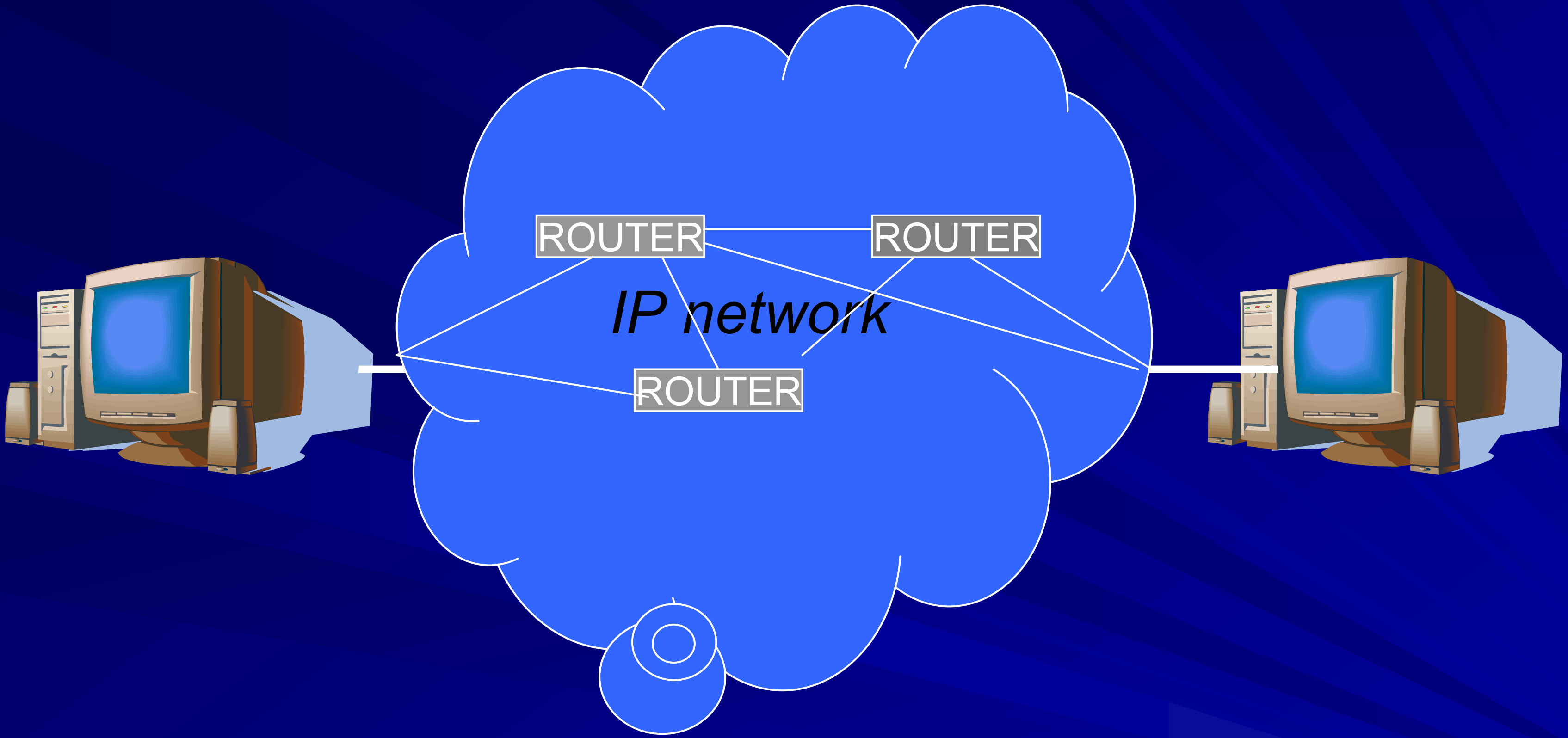
Overview

- Project Area
- Networking
- High Performance Transfer and Problems
- EGEE JRA4 involvement
- Future Directions

Project Area

- The Grid is interconnected through networks.
- The performance of a network affects the performance/operation of a Grid.
- How does it affect it ?
- Need to measure network performance metrics and correlate them with grid performance metrics.

Networking

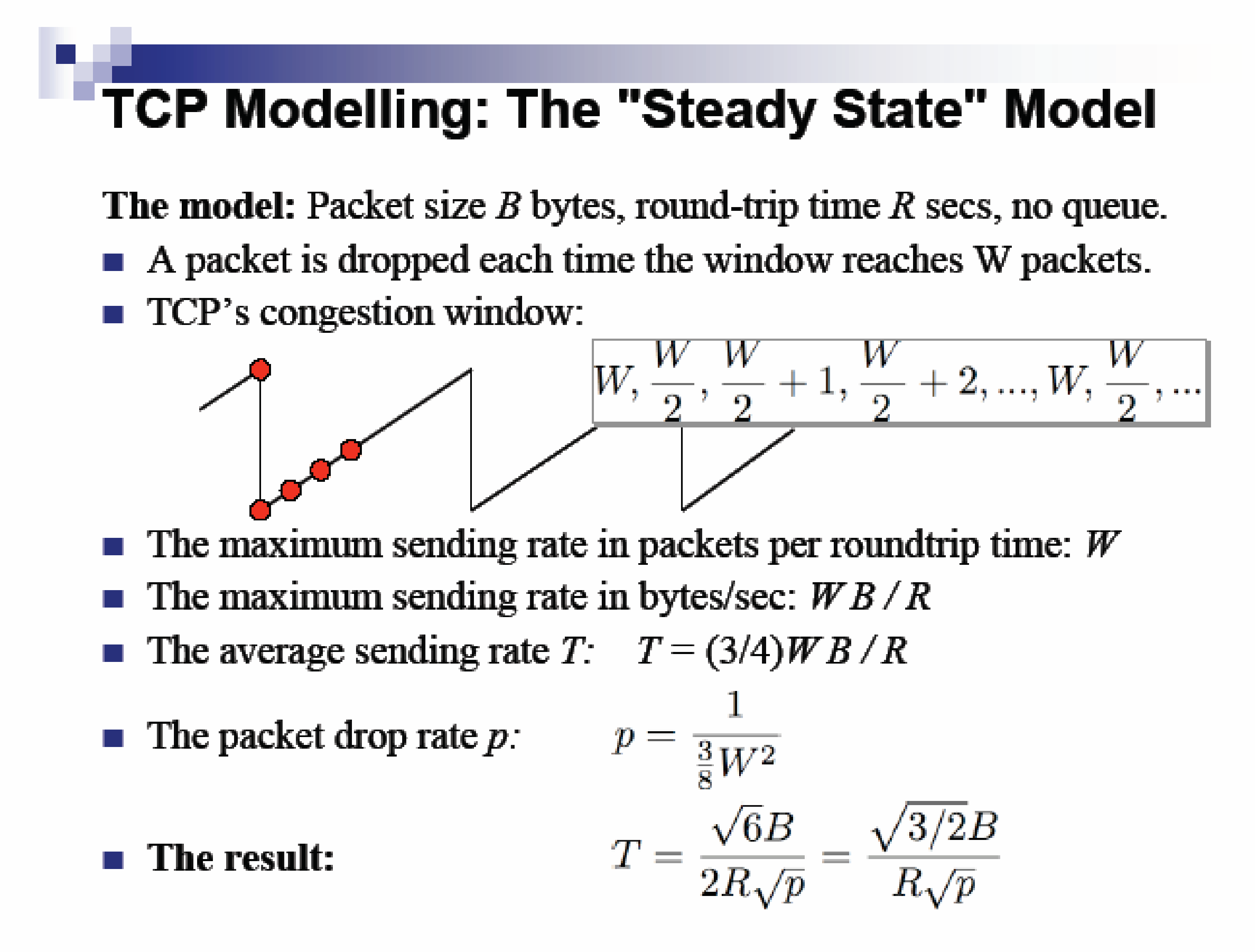


IP – Internet Protocol used to transfer packets from source to destination.

Problems :Packets can get lost , delayed at router queues , arrive out of order!

TCP

- TCP – reliable Transmission Control Protocol.
- Standardised in 1981
- Transfers majority of internet traffic – email, web ,ftp, ssh, telnet and GridFTP- all built upon TCP.
- Acknowledgement and sliding window based protocol
- Allows ordering of packets at receiver
- Retransmits lost packets.



TCP Modelling: The "Steady State" Model

The model: Packet size B bytes, round-trip time R secs, no queue.

- A packet is dropped each time the window reaches W packets.
- TCP's congestion window:

$$W, \frac{W}{2}, \frac{W}{2}+1, \frac{W}{2}+2, \ldots, W, \frac{W}{2}, \ldots$$

- The maximum sending rate in packets per roundtrip time: W
- The maximum sending rate in bytes/sec: $W\,B\,/\,R$
- The average sending rate T: $T = (3/4)W\,B\,/\,R$
- The packet drop rate p: $p = \dfrac{1}{\frac{3}{8}W^2}$
- **The result:** $T = \dfrac{\sqrt{6}B}{2R\sqrt{p}} = \dfrac{\sqrt{3/2}B}{R\sqrt{p}}$

Slide from NFNN Workshop University College London July 2004
from presentation by Professor Mark Handley.

TCP

Simple model for TCP “Steady State Model”

Average sending rate

- proportional to packet size

- Inversely proportional to packet loss

- Inversely proportional to Round Trip Time

TCP

- TCP was designed for LANs (and not WANs)
- Get performance problems when using standard TCP in WANs because have high Round Trip Time (RTT).

High Performance Transfers.

- Example transferring data from NeSC to CERN.
- Data is going to go across a number of networks. (in any single international path between two end points, at least five networks are involved:

Two university or campus networks

Two national backbones (the participating NRENs)

The European backbone (GÉANT or GÉANT2).

- Can have problems on any of these networks !

What will affect my Transfer ?

- End host specification

Network interface card. (could be bottleneck)

Motherboard, CPU, memory, Disk Drive

- So need a high specification PC.

What will affect my Transfer ?

- End host TCP tuning
- Need to calculate BDP (Bandwidth delay product) to determine adequate TCP buffer sizes at sender and receiver.
- EXAMPLE
- On a link where OC3 155Mbps and 100 BT Ethernet used throughout and RTT = 50 ms (obtained by ping)– require
- = 50 ms x (100Mbps / 8 bits) = 625 KB
- (Note most OS default TCP buffers size is 24KB or 32 KB and Linux is only 8KB) (this is ok for LANs but not WANs)
- Using untuned buffers you often get less than 5% utilisation !

(source: TCP tuning guide for distributed application on wide area networks , Brian L. Tierney*)*

What will affect my Transfer ?

- Version of TCP used
- Can get newer versions of TCP which are suitable for High performance transfer.
- Ongoing research – Fairness of these versions in terms of affect on other users.

How can I monitor my end-host TCP parameters?

- Web100

Provides information on TCP parameters

What else will affect my Transfer ?

- Performance of the networks my data traverses

- Delays

- Packet loss

- Jitter (IPDV)

Network transfer problems (In order of most probable cause)

- End host (OS, architecture, disk, TCP)
- End host application itself
- Local network limitation (switches)
- Firewall
- And only then the WAN

- Understanding each area will help you to reduce the Wizard Gap (difference between transfer rates a network expert and normal user achieve).

OK – how do I obtain information about the network?

- Need Network Monitoring

- Active monitoring e.g., iperf, udpmon, PingER

-(can put load on to the network though)

- Passive monitoring. Passive monitoring tools capture packets using either a standard NIC (network interface card) and libpcap library or using a specialized hardware monitoring adapters.
- (privacy issues)

Network monitoring other networks?

- Each network has its own monitoring infrastructure so would be ideal to share networking information.
- Need standard interface to share information.

Network performance Characteristics for Grids (METRICS) - GGF NMWG

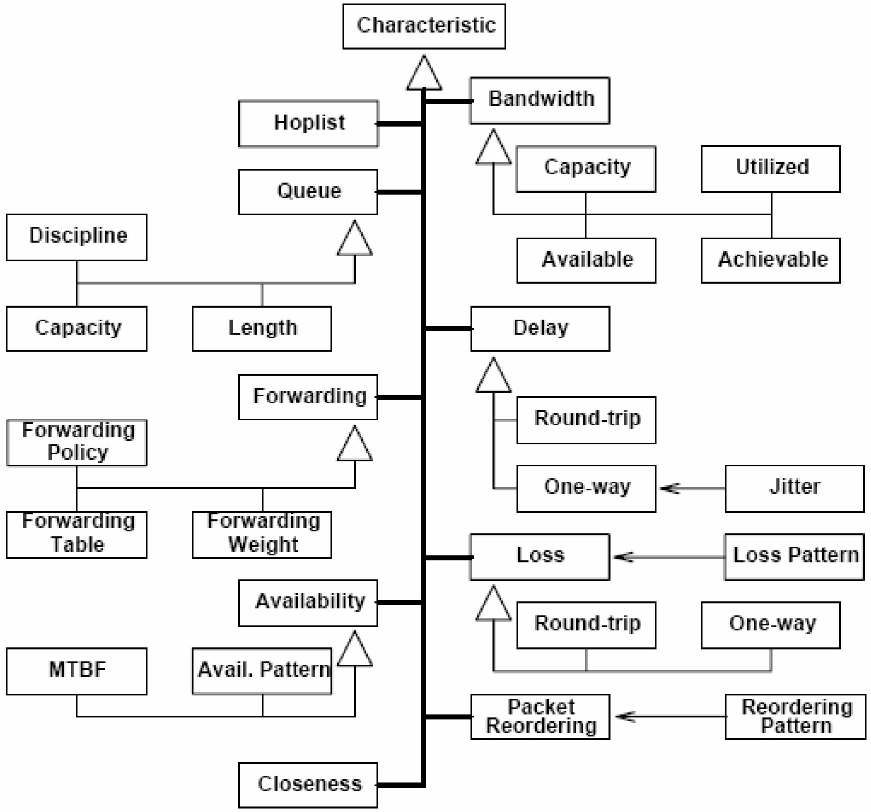
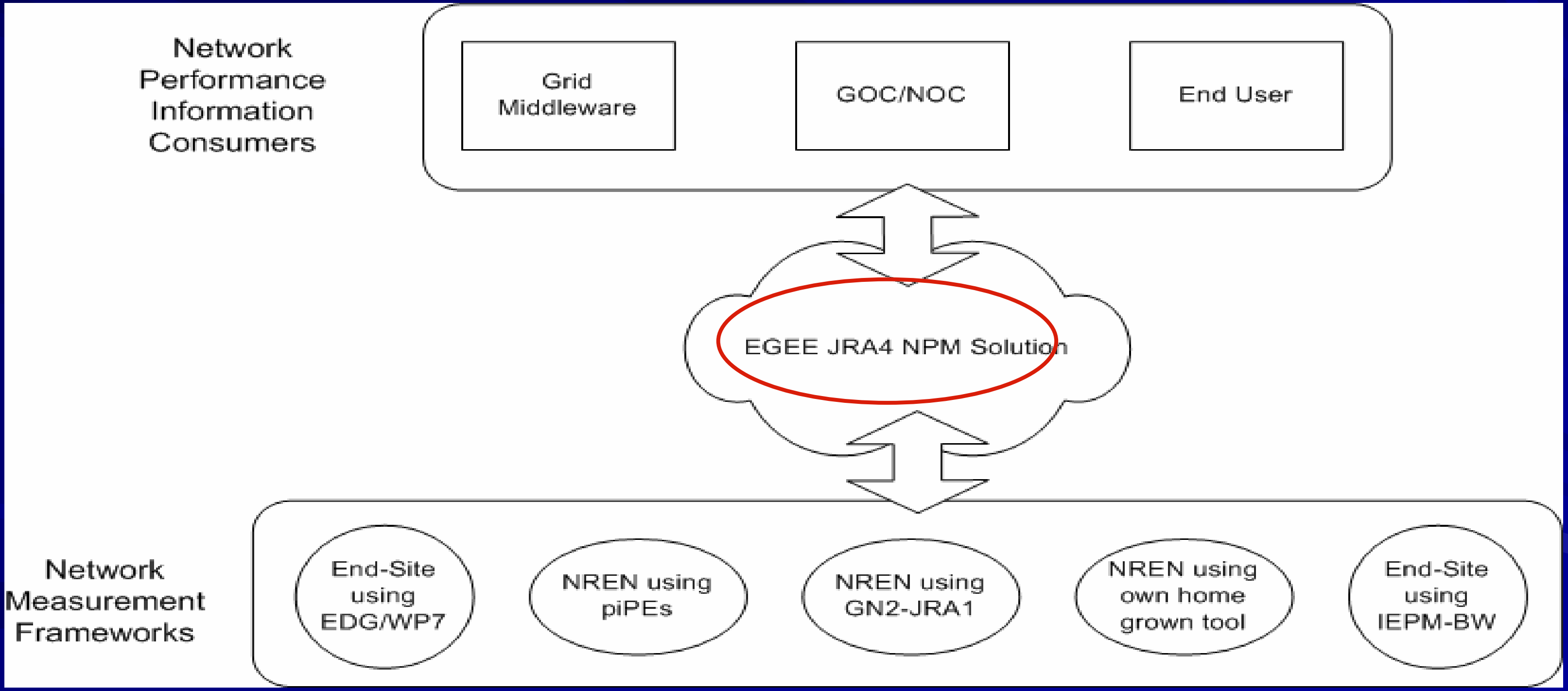


Figure 3: Network characteristics that can be used to describe the behaviour of network entities.

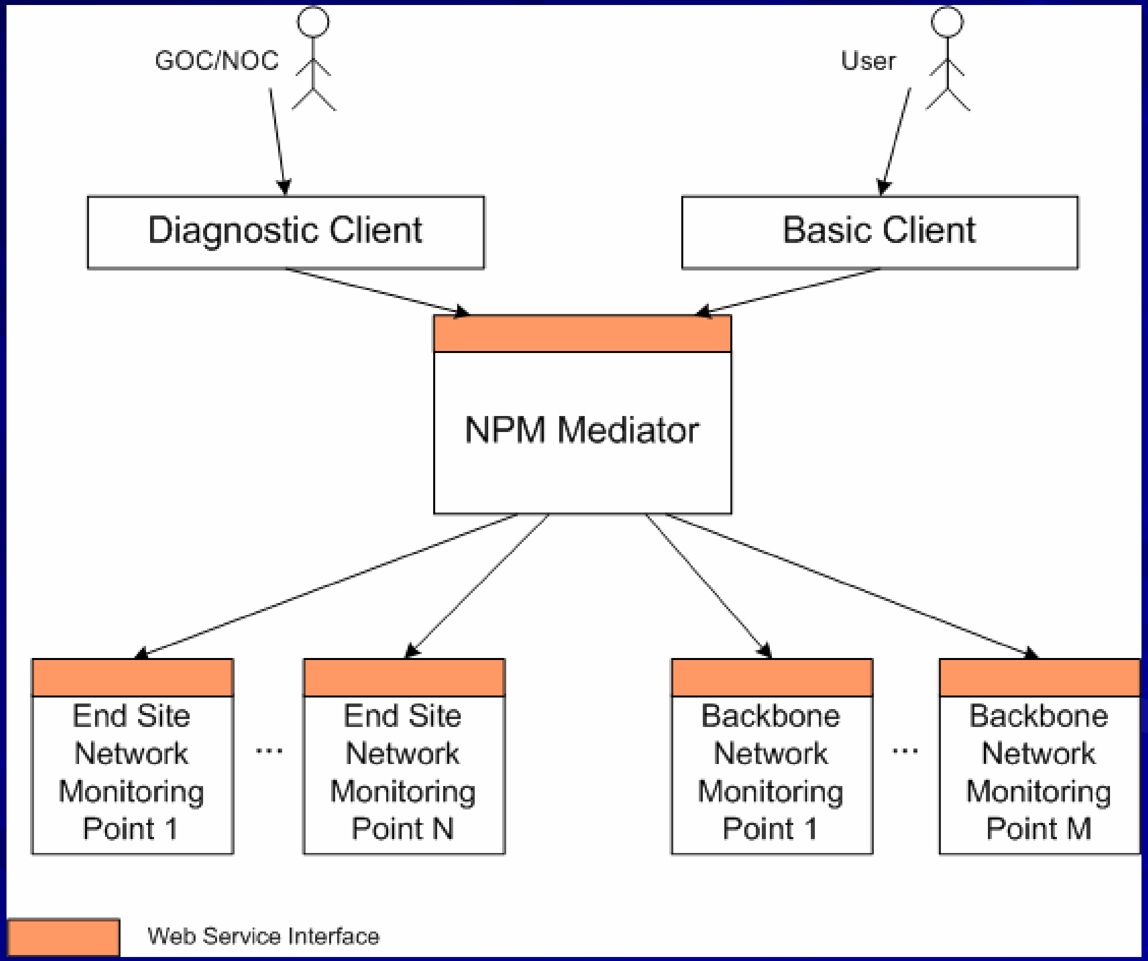
GGF have designed an NMWG Schema

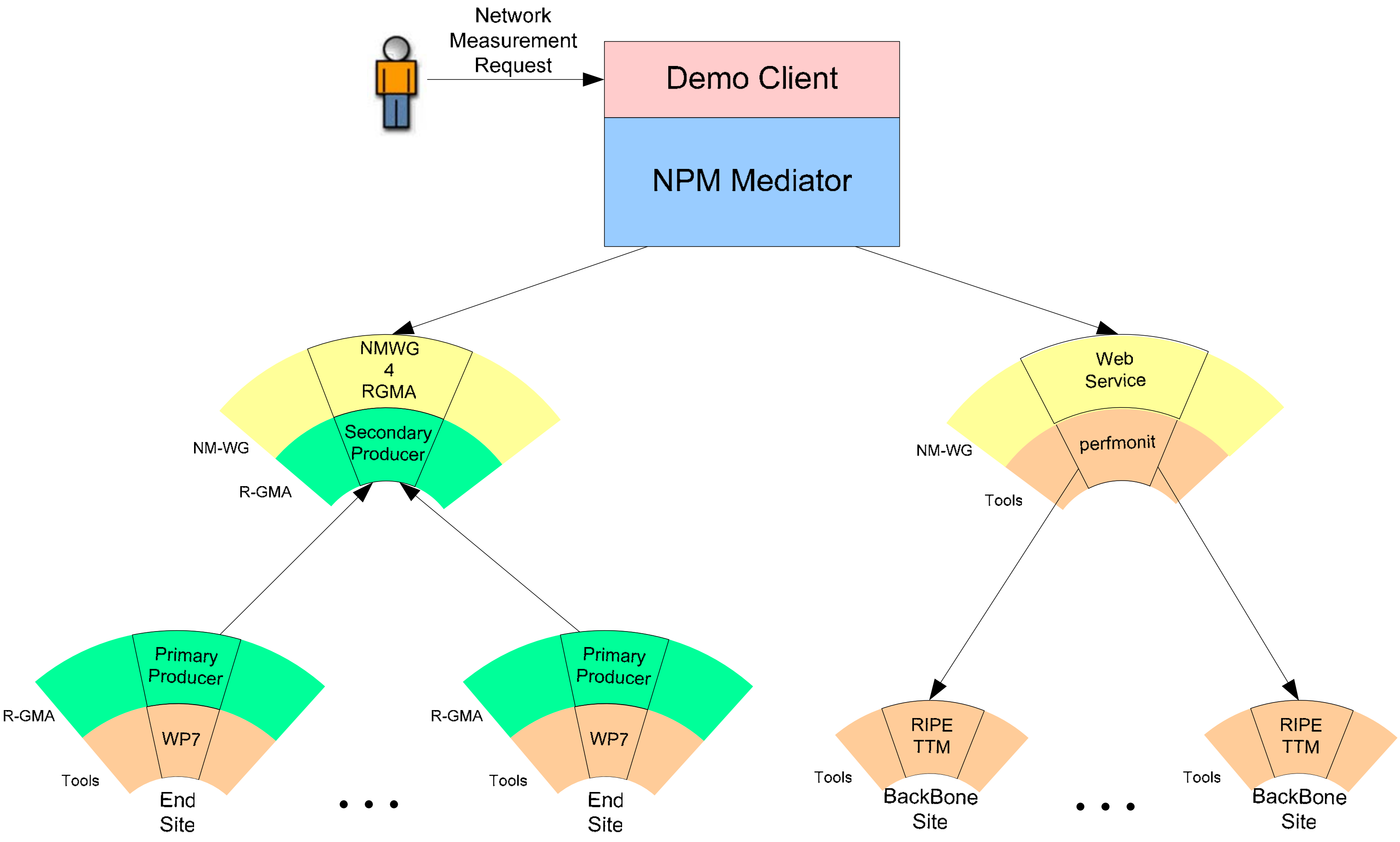
EGEE JRA4 - NPM Aims

- **JRA4/NPM provides uniform access to network performance information from a heterogeneous set of monitoring frameworks**

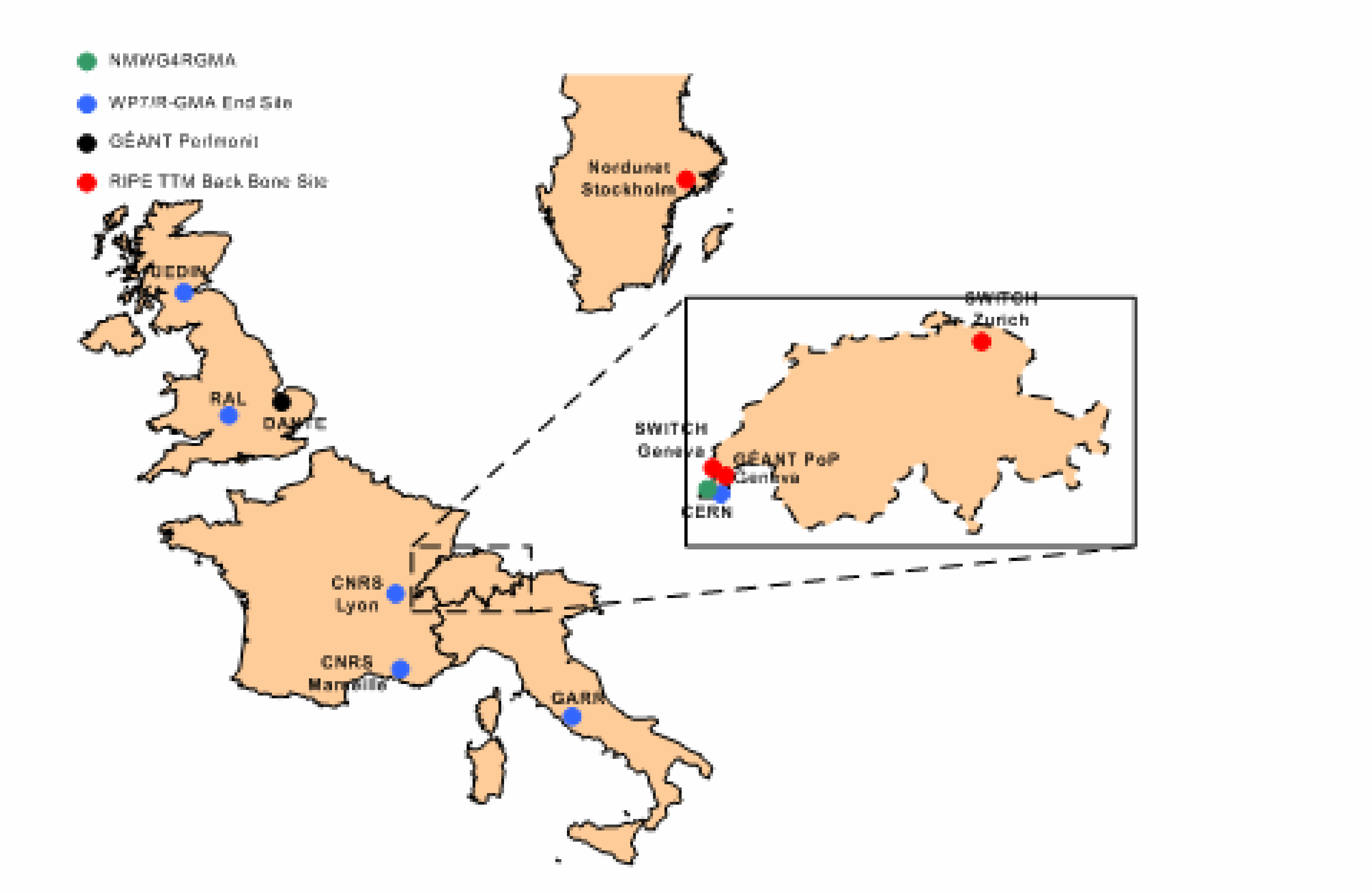


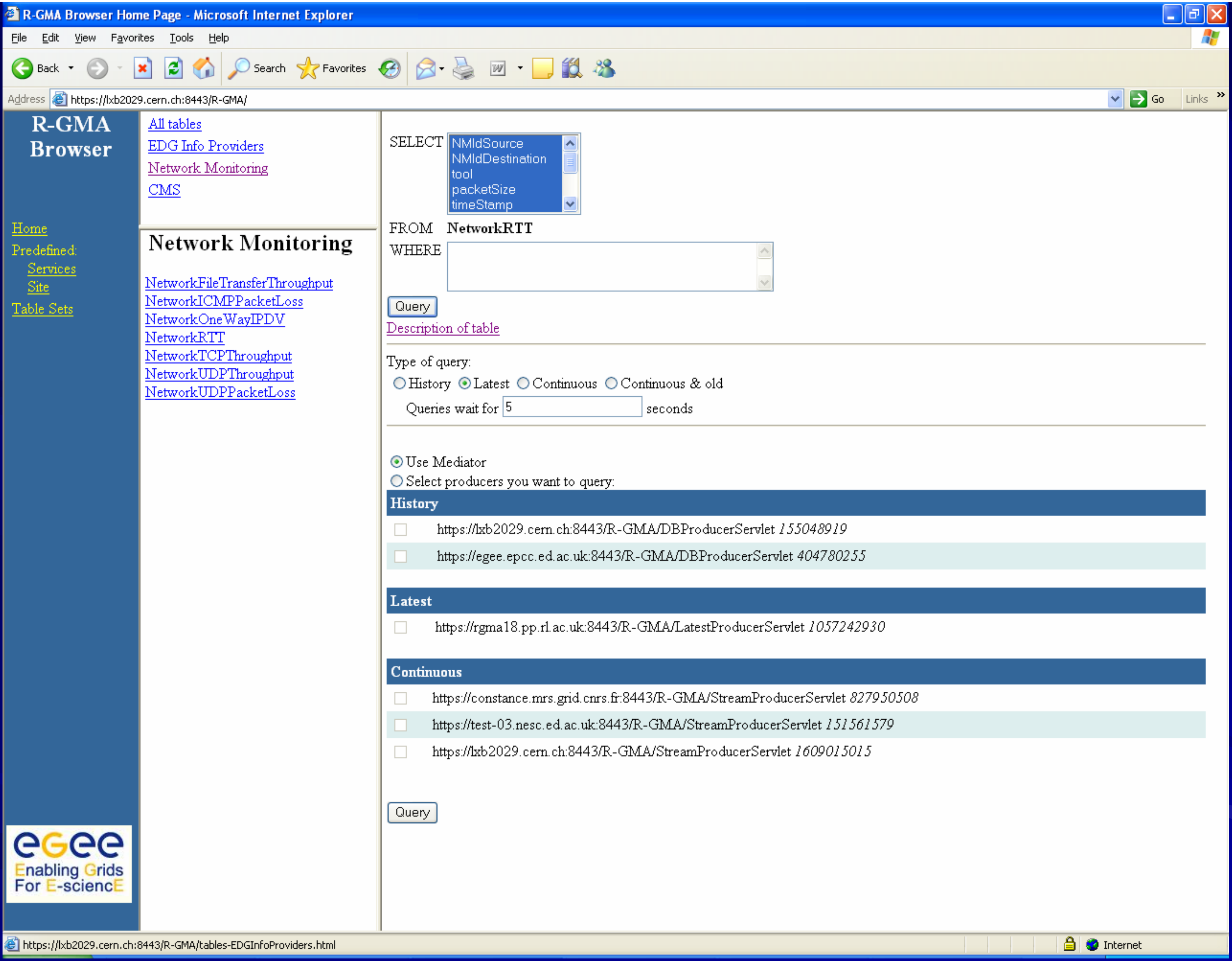
NPM Architecture


GOC/NOC
User
Diagnostic Client
Basic Client
NPM Mediator
End Site Network Monitoring Point 1
...
End Site Network Monitoring Point N
Backbone Network Monitoring Point 1
...
Backbone Network Monitoring Point M
Web Service Interface


Network Measurement Request
Demo Client
NPM Mediator
NMWG 4 RGMA
Secondary Producer
NM-WG
R-GMA
Web Service
perfmonit
NM-WG
Tools
Primary Producer
WP7
R-GMA
Tools
End Site
Primary Producer
WP7
R-GMA
Tools
End Site
RIPE TTM
Tools
BackBone Site
RIPE TTM
Tools
BackBone Site

The current state of Network Monitoring Point deployment is shown below, with deployments at JRA4 partners (University of Edinburgh, CNRS, GARR and DANTE) and on the JRA1 testbed (at CERN and RAL).




R-GMA Browser Home Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address https://lxb2029.cern.ch:8443/R-GMA/
Go Links

R-GMA Browser

All tables
EDG Info Providers
Network Monitoring
CMS

Home
Predefined:
Services
Site
Table Sets

Network Monitoring

NetworkFileTransferThroughput
NetworkICMPPacketLoss
NetworkOneWayIPDV
NetworkRTT
NetworkTCPThroughput
NetworkUDPThroughput
NetworkUDPPacketLoss

SELECT NMIdSource NMIdDestination tool packetSize timeStamp

FROM NetworkRTT

WHERE

Query

Description of table

Type of query:
○ History ● Latest ○ Continuous ○ Continuous & old

Queries wait for 5 seconds

● Use Mediator
○ Select producers you want to query:

History

□ https://lxb2029.cern.ch:8443/R-GMA/DBProducerServlet *155048919*
□ https://egee.epcc.ed.ac.uk:8443/R-GMA/DBProducerServlet *404780255*

Latest

□ https://rgma18.pp.rl.ac.uk:8443/R-GMA/LatestProducerServlet *1057242930*

Continuous

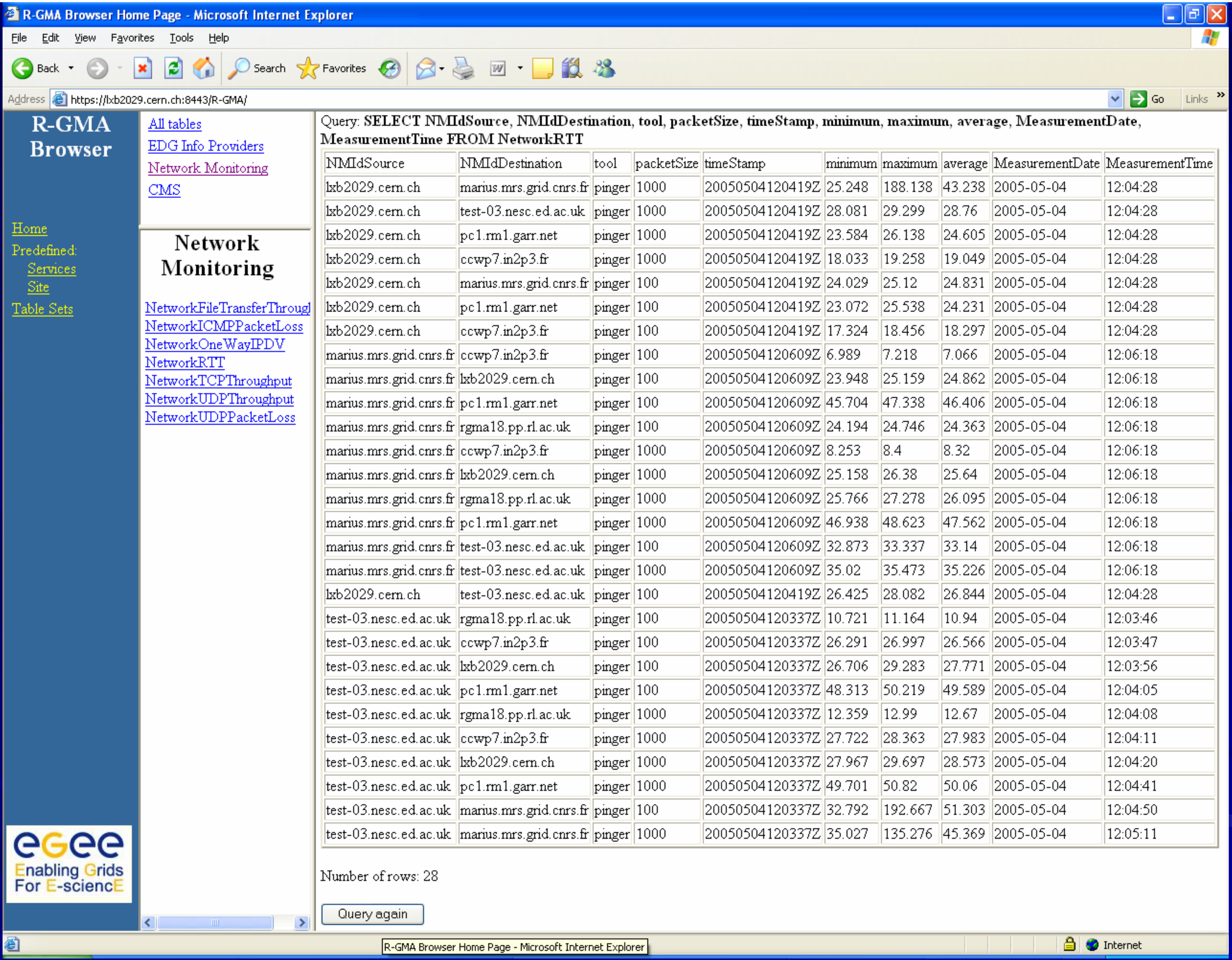
□ https://constance.mrs.grid.cnrs.fr:8443/R-GMA/StreamProducerServlet *827950508*
□ https://test-03.nesc.ed.ac.uk:8443/R-GMA/StreamProducerServlet *151561579*
□ https://lxb2029.cern.ch:8443/R-GMA/StreamProducerServlet *1609015015*

Query

egee
Enabling Grids For E-sciencE
https://lxb2029.cern.ch:8443/R-GMA/tables-EDGInfoProviders.html
Internet

Secondary producer – at CERN

Access requires a user certificate



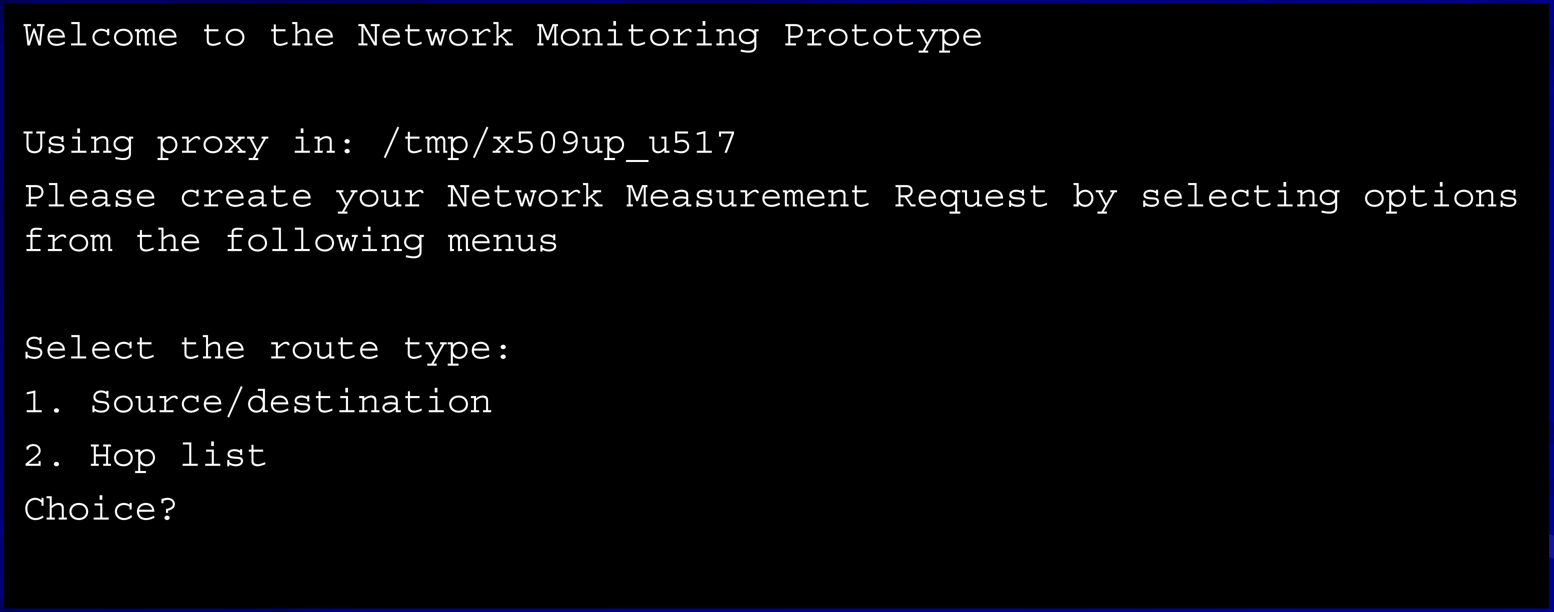
Query: **SELECT NMIdSource, NMIdDestination, tool, packetSize, timeStamp, minimum, maximum, average, MeasurementDate, MeasurementTime FROM NetworkRTT**

NMIdSource	NMIdDestination	tool	packetSize	timeStamp	minimum	maximum	average	MeasurementDate	MeasurementTime
lxb2029.cern.ch	marius.mrs.grid.cnrs.fr	pinger	1000	20050504120419Z	25.248	188.138	43.238	2005-05-04	12:04:28
lxb2029.cern.ch	test-03.nesc.ed.ac.uk	pinger	1000	20050504120419Z	28.081	29.299	28.76	2005-05-04	12:04:28
lxb2029.cern.ch	pc1.rm1.garr.net	pinger	1000	20050504120419Z	23.584	26.138	24.605	2005-05-04	12:04:28
lxb2029.cern.ch	ccwp7.in2p3.fr	pinger	1000	20050504120419Z	18.033	19.258	19.049	2005-05-04	12:04:28
lxb2029.cern.ch	marius.mrs.grid.cnrs.fr	pinger	100	20050504120419Z	24.029	25.12	24.831	2005-05-04	12:04:28
lxb2029.cern.ch	pc1.rm1.garr.net	pinger	100	20050504120419Z	23.072	25.538	24.231	2005-05-04	12:04:28
lxb2029.cern.ch	ccwp7.in2p3.fr	pinger	100	20050504120419Z	17.324	18.456	18.297	2005-05-04	12:04:28
marius.mrs.grid.cnrs.fr	ccwp7.in2p3.fr	pinger	100	20050504120609Z	6.989	7.218	7.066	2005-05-04	12:06:18
marius.mrs.grid.cnrs.fr	lxb2029.cern.ch	pinger	100	20050504120609Z	23.948	25.159	24.862	2005-05-04	12:06:18
marius.mrs.grid.cnrs.fr	pc1.rm1.garr.net	pinger	100	20050504120609Z	45.704	47.338	46.406	2005-05-04	12:06:18
marius.mrs.grid.cnrs.fr	rgma18.pp.rl.ac.uk	pinger	100	20050504120609Z	24.194	24.746	24.363	2005-05-04	12:06:18
marius.mrs.grid.cnrs.fr	ccwp7.in2p3.fr	pinger	1000	20050504120609Z	8.253	8.4	8.32	2005-05-04	12:06:18
marius.mrs.grid.cnrs.fr	lxb2029.cern.ch	pinger	1000	20050504120609Z	25.158	26.38	25.64	2005-05-04	12:06:18
marius.mrs.grid.cnrs.fr	rgma18.pp.rl.ac.uk	pinger	1000	20050504120609Z	25.766	27.278	26.095	2005-05-04	12:06:18
marius.mrs.grid.cnrs.fr	pc1.rm1.garr.net	pinger	1000	20050504120609Z	46.938	48.623	47.562	2005-05-04	12:06:18
marius.mrs.grid.cnrs.fr	test-03.nesc.ed.ac.uk	pinger	100	20050504120609Z	32.873	33.337	33.14	2005-05-04	12:06:18
marius.mrs.grid.cnrs.fr	test-03.nesc.ed.ac.uk	pinger	1000	20050504120609Z	35.02	35.473	35.226	2005-05-04	12:06:18
lxb2029.cern.ch	test-03.nesc.ed.ac.uk	pinger	100	20050504120419Z	26.425	28.082	26.844	2005-05-04	12:04:28
test-03.nesc.ed.ac.uk	rgma18.pp.rl.ac.uk	pinger	100	20050504120337Z	10.721	11.164	10.94	2005-05-04	12:03:46
test-03.nesc.ed.ac.uk	ccwp7.in2p3.fr	pinger	100	20050504120337Z	26.291	26.997	26.566	2005-05-04	12:03:47
test-03.nesc.ed.ac.uk	lxb2029.cern.ch	pinger	100	20050504120337Z	26.706	29.283	27.771	2005-05-04	12:03:56
test-03.nesc.ed.ac.uk	pc1.rm1.garr.net	pinger	100	20050504120337Z	48.313	50.219	49.589	2005-05-04	12:04:05
test-03.nesc.ed.ac.uk	rgma18.pp.rl.ac.uk	pinger	1000	20050504120337Z	12.359	12.99	12.67	2005-05-04	12:04:08
test-03.nesc.ed.ac.uk	ccwp7.in2p3.fr	pinger	1000	20050504120337Z	27.722	28.363	27.983	2005-05-04	12:04:11
test-03.nesc.ed.ac.uk	lxb2029.cern.ch	pinger	1000	20050504120337Z	27.967	29.697	28.573	2005-05-04	12:04:20
test-03.nesc.ed.ac.uk	pc1.rm1.garr.net	pinger	1000	20050504120337Z	49.701	50.82	50.06	2005-05-04	12:04:41
test-03.nesc.ed.ac.uk	marius.mrs.grid.cnrs.fr	pinger	100	20050504120337Z	32.792	192.667	51.303	2005-05-04	12:04:50
test-03.nesc.ed.ac.uk	marius.mrs.grid.cnrs.fr	pinger	1000	20050504120337Z	35.027	135.276	45.369	2005-05-04	12:05:11

Number of rows: 28

RTT measured by PingER

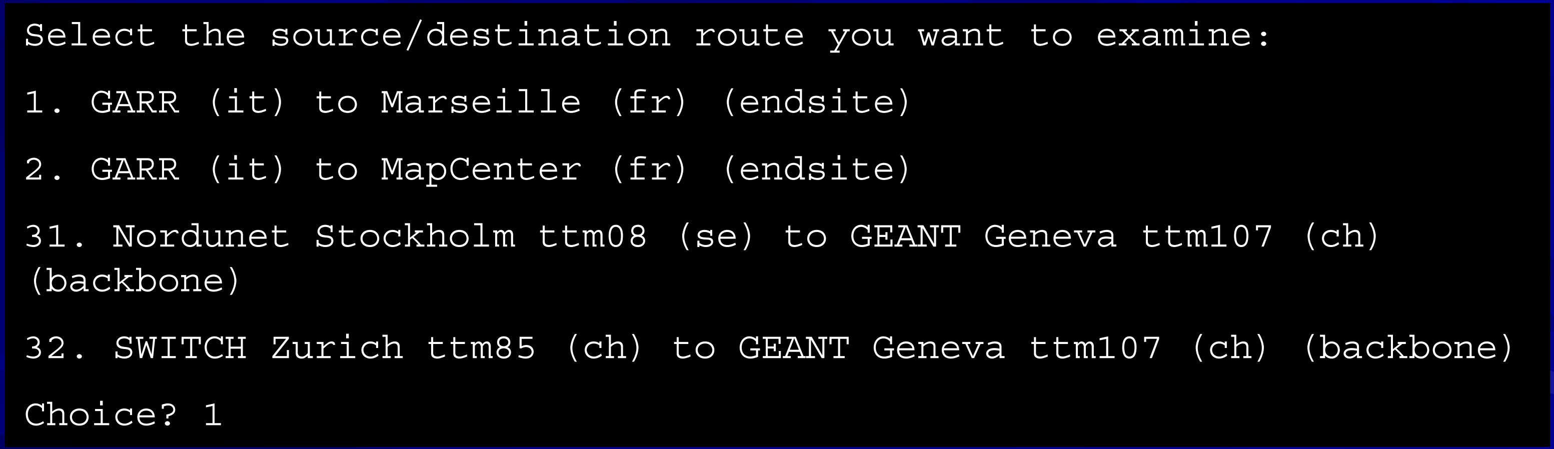
Demo of NPM prototype



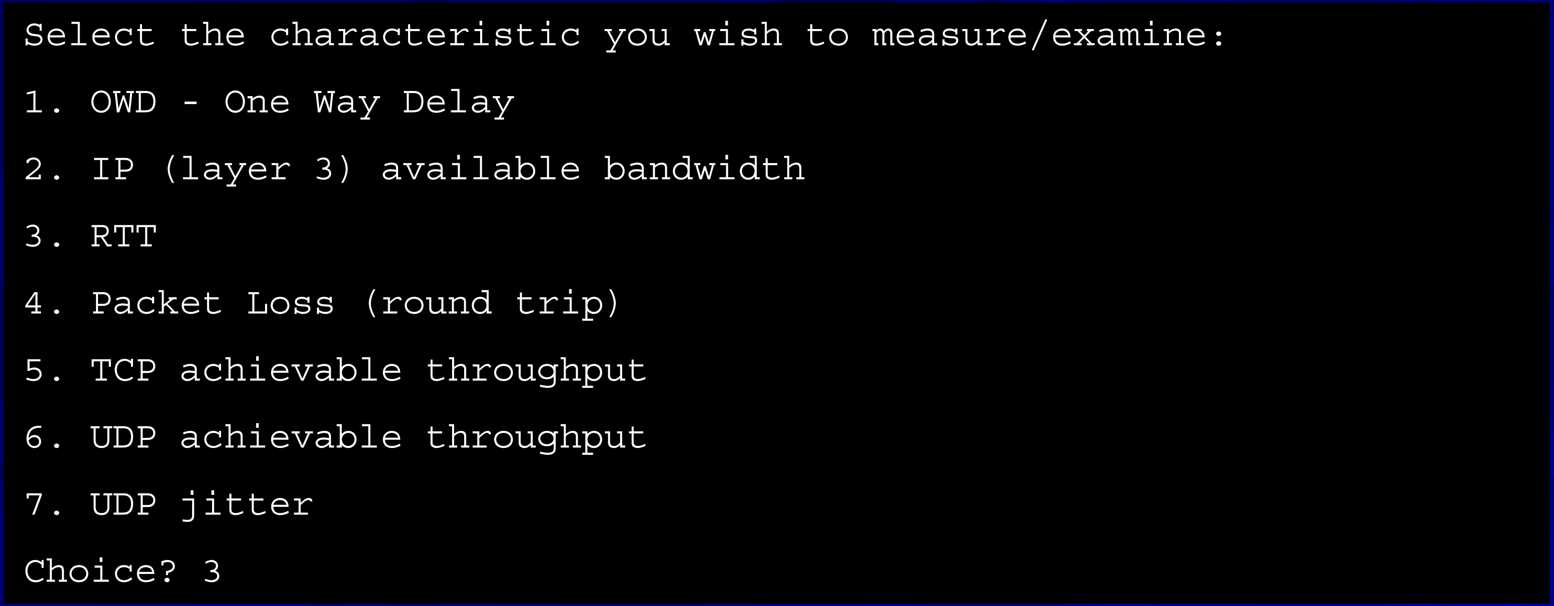
```
Welcome to the Network Monitoring Prototype

Using proxy in: /tmp/x509up_u517
Please create your Network Measurement Request by selecting options
from the following menus

Select the route type:
1. Source/destination
2. Hop list
Choice?
```

- The demo client states what type of route it is, but the user need not be aware of the different network monitoring points contacted

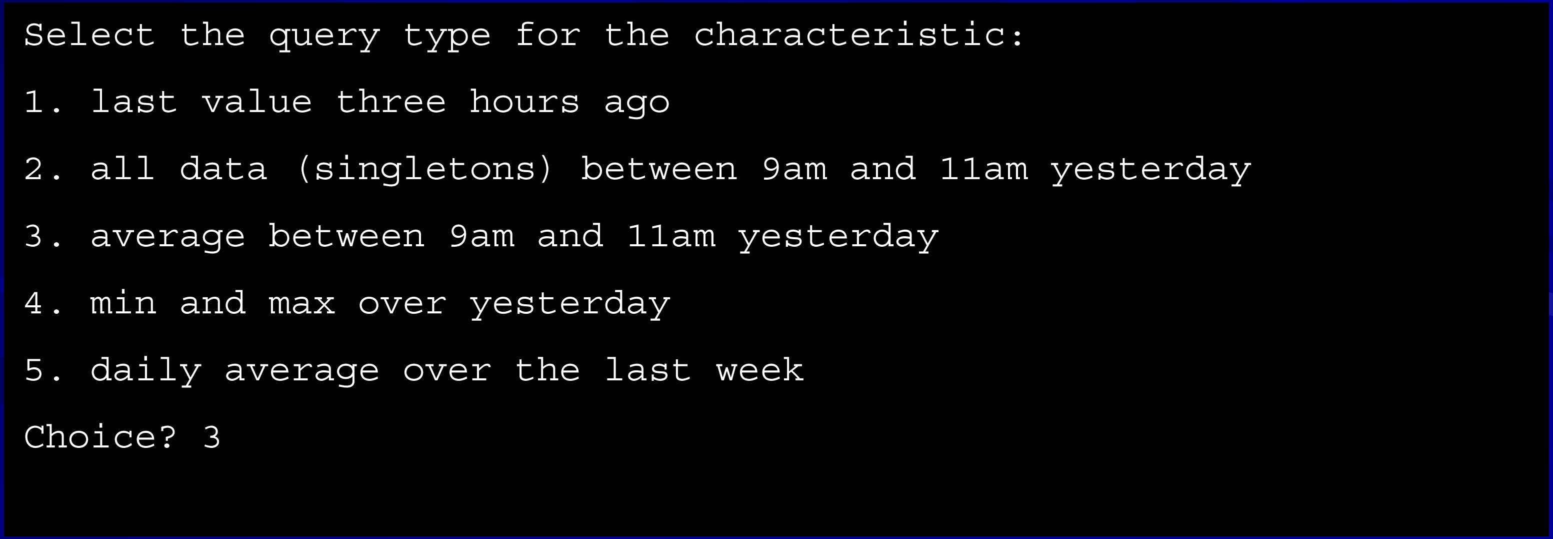
selected route 1: GARR to Marseille



```
Select the source/destination route you want to examine:

1. GARR (it) to Marseille (fr) (endsite)

2. GARR (it) to MapCenter (fr) (endsite)

31. Nordunet Stockholm ttm08 (se) to GEANT Geneva ttm107 (ch)
(backbone)

32. SWITCH Zurich ttm85 (ch) to GEANT Geneva ttm107 (ch) (backbone)

Choice? 1
```



```
Select the characteristic you wish to measure/examine:
1. OWD - One Way Delay
2. IP (layer 3) available bandwidth
3. RTT
4. Packet Loss (round trip)
5. TCP achievable throughput
6. UDP achievable throughput
7. UDP jitter
Choice? 3
```

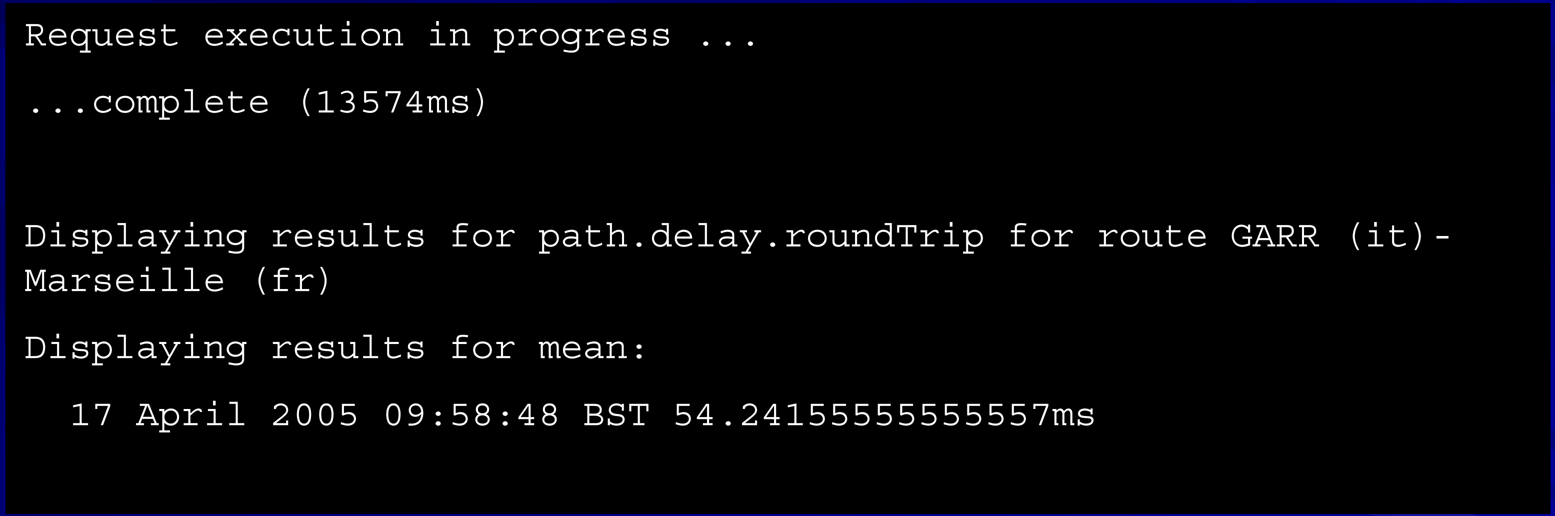


```
Select the query type for the characteristic:
1. last value three hours ago
2. all data (singletons) between 9am and 11am yesterday
3. average between 9am and 11am yesterday
4. min and max over yesterday
5. daily average over the last week
Choice? 3
```

- The client generates an appropriate NM-WG Request and outputs a summary to the screen
- It then sends the request to the appropriate NM-WG compliant network monitoring point (web service)

```
You selected to obtain RTT measurements (average between 9am and
11am yesterday) on Source/destination: GARR (it) to Marseille (fr)
(endsite)

Your selections resulted in a query for 1 result (mean over the
whole period) for path.delay.roundTrip for the route GARR (it)-
Marseille (fr) for measurements between Sun Apr 17 09:00:00 BST
2005 and Sun Apr 17 11:00:00 BST 2005
```

- Once the result is received, it is output to the user
- Note that the execution time is long – NPM team will be resolving this issue within the next development cycle



```
Request execution in progress ...

...complete (13574ms)


Displaying results for path.delay.roundTrip for route GARR (it)-
Marseille (fr)

Displaying results for mean:

  17 April 2005 09:58:48 BST 54.24155555555557ms
```

END OF NPM DEMO

Future Project/Research Directions

- Work on Network diagnostics
- Correlation of Grid and Network performance metrics.

- Do practical tests e.g. submit jobs. Vs network performance metrics.

- Important to understand grid traffic requirements. (for planning, SLAs etc.)

- Continued involvement with NPM

Future Project/Research Directions

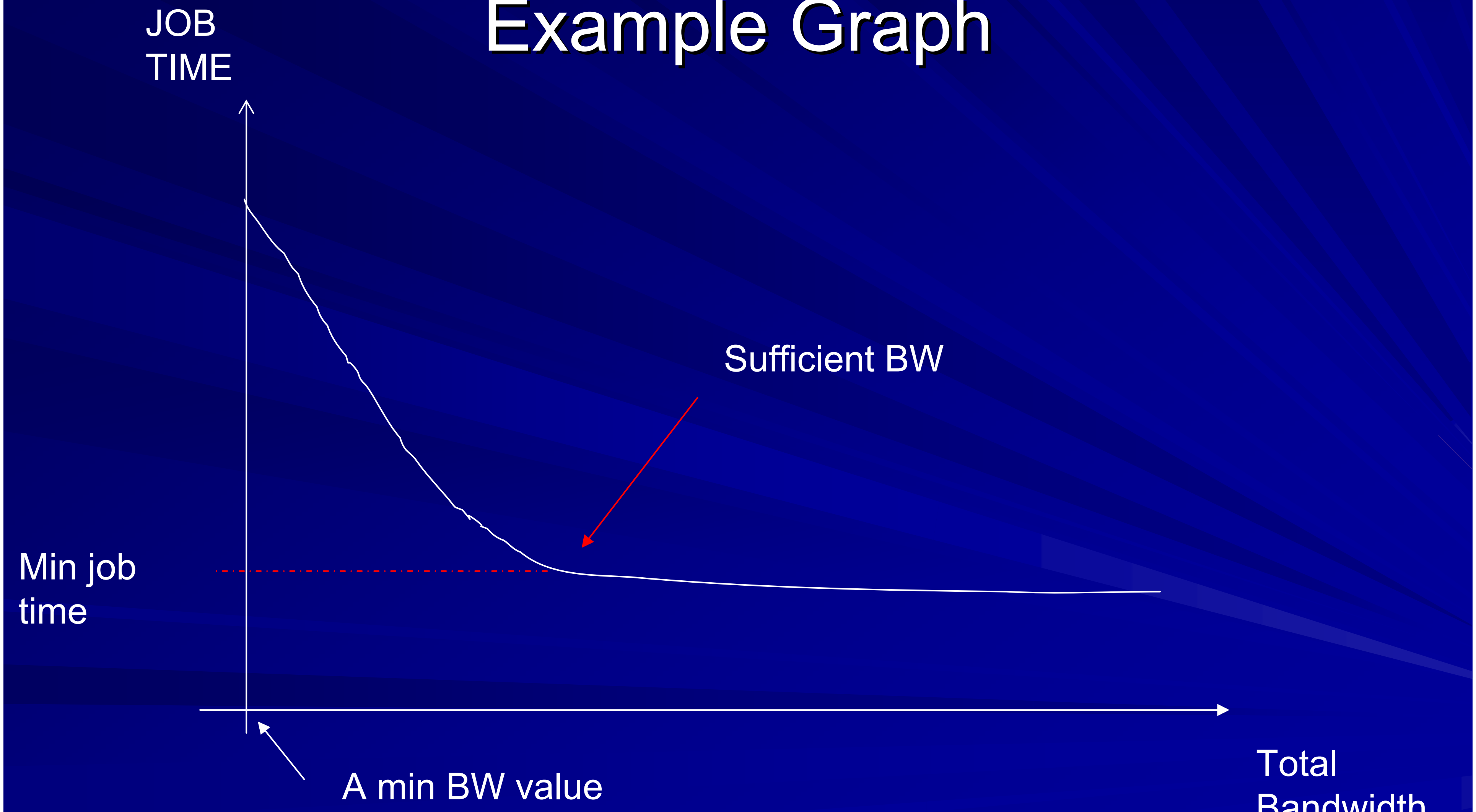
- Use of networking information

Network cost functions

e.g. Replica Placement

WMS – where to submit jobs

Grid and Network performance metrics.
Example Graph



Contribution to NPM

1st Prototype

– assisted with development

2nd prototype

–deployed monitoring node at NeSC.

(Installed R-GMA server, client and WP7 tools)

Next Prototype (due September)

– Will Work on Diagnostic tool functional specification.

- Maintain deployment node.

Resources

- Networks for non-Networkers workshop UCL July 2004

-http://grid.ucl.ac.uk/NFNN_Programme.html

- TCP tuning guide for distributed application on wide area networks, Brian L. Tierney

-http://dsd.lbl.gov/TCP-tuning/tcp-wan-perf.pdf

THE END
and
Question Time!