On-line Monitoring of Grid Computing Applications

Henry Nebrensky,

School of Engineering and Design
Brunel University, UK

Topics

- What is the Grid?
- Why do we want one ?
- GridPP – and how Brunel fits in
- Applications monitoring:
  - BOSS
  - R-GMA
  - and how they fit together

What is Grid Computing?

(apart from an over-used buzz-phrase?)

Hyped as the solution: bold new breakthroughs in commerce, science, medicine; saves money saves lives, saves the planet!

http://www.climateprediction.net

http://gridcafe.web.cern.ch/gridcafe/animations/amalthea.rm ☺

Relates to making computing power easily available

Many solutions, and even more hype!

One authoritative view:

http://www-fp.mcs.anl.gov/~foster/Articles/WhatIsTheGrid.pdf

What is Grid Computing?

More than just "distributed computing":

- **General purpose**
- **"Virtual Organisation" VO – allows its resources to be shared among its members**
- **Wide-Area deployment – heterogeneous resources spread over globe**

- **Foster and Kesselman – the Globus Project**
  - Anatomy of the Grid
    http://www.globus.org/research/papers/anatomy.pdf
  - Physiology of the Grid
    http://www.globus.org/research/papers/ogsa.pdf
  - I. Foster and C. Kesselman: *"The Grid: Blueprint for a New Computing Infrastructure"*

What is Grid Computing?

- **The Web provides seamless access to data**
  Underlying protocols don't care about who you are

- **The Grid provides seamless access to computing**
  Users identified via X.509 certificates:
  must be in a suitable VO, then single sign-on
  move from individual to *group* working

- **Lots of separate Grid testbeds. I refer to the EDG/LCG model** http://lcg.web.cern.ch/LCG/

- **Constantly changing acronyms - terminology here is a personal mish-mash ☹**

Why do we want the Grid?

Increasing need to process lots of data:

- **Commerce – data mining**
- **Engineering – simulations**
- **Physics – HEP data processing (that's "us"!)**
- **Earth Sciences – climate and weather prediction**
- **Biology – protein folding and genomics**
- **Medicine – imaging and tomography**
- **and many more…**

What is Grid computing?

“It's significant that the UK is the first country to develop a national *e*-science grid, which intends to make access to computing power, scientific data repositories and experimental facilities as easy as the web makes access to information. One of the pilot *e*-science projects is to develop a digital mammographic archive, together with an intelligent medical decision support system for breast cancer diagnosis and treatment. An individual hospital will not have supercomputing facilities, but through the grid it could buy the time it needs.”

Tony Blair July 2002
http://politics.guardian.co.uk/speeches/story/0,11126,721029,00.html

From each according to his means…

A Grid is a collection of "resources" provided by participating sites:

- a Computing Element provides CPUs
- a Storage Element provides storage space

A CE consists of a Gatekeeper (GK) which receives the job, and a set of Worker Nodes (WN) that do the actual work – often similar to traditional batch farm

An SE could be a PC with a big disk – but it can also be a Petabyte-scale tape silo

…to each according to his needs.

- **At a “UI”, specify executable, data files and other requirements (using JDL) and submit**
- **The UI client passes this to a Resource Broker (RB) which identifies the best place to run the job:**
  - which CE has the most suitable CPUs?
  - which SE has the data files?
- **Eventually the job will have been processed, and the output can be retrieved by the UI…**
  the results are on your PC!

To each his own.

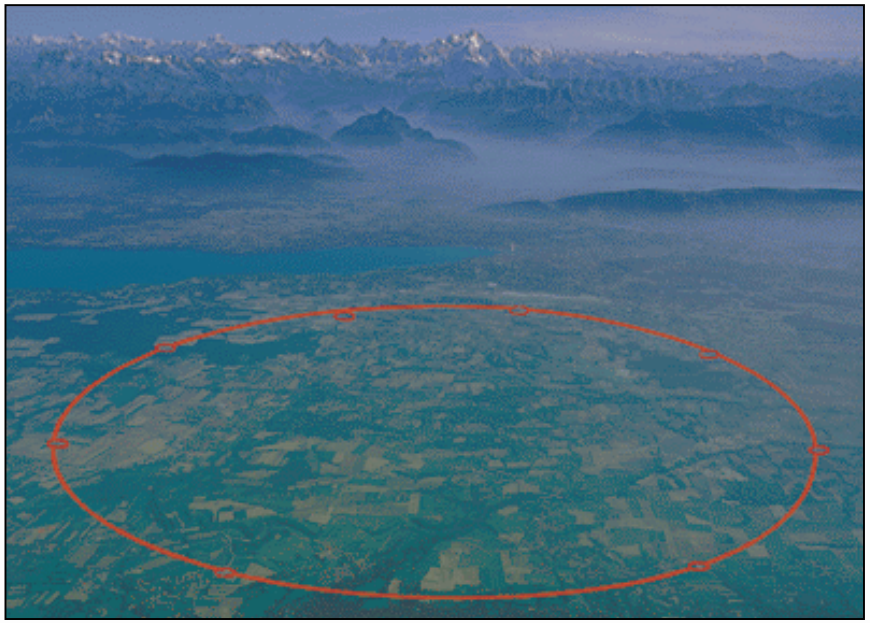
The user needn’t ever know – let alone care – where the job ran, just as few people worry about which power station supplies their wall socket.

Hence “The Grid”.

A better analogy than many appreciate.

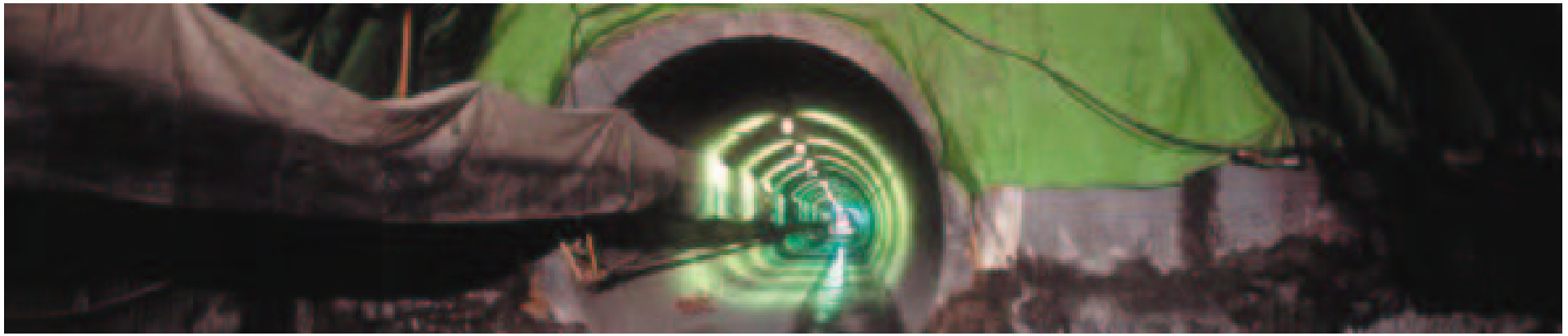
GridPP
UK Computing for Particle Physics

CERN Large Hadron Collider



- The world's most powerful particle accelerator
- Due to turn on in 2007
- Will collide protons at 7 Tera Electron Volts….

CERN



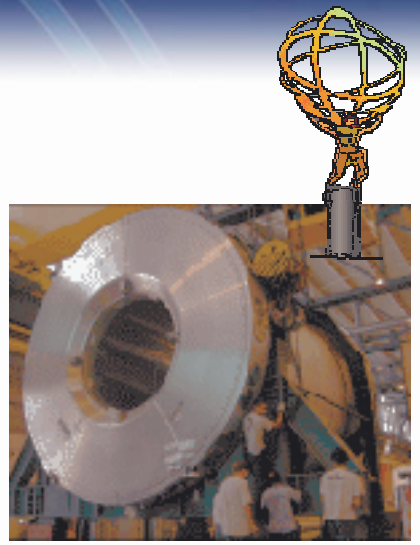
GridPP
UK Computing for Particle Physics

4 LHC Experiments

ATLAS

- general purpose: origin of mass, supersymmetry, micro-black holes

- 2,000 scientists from 34 countries



CMS

- general purpose detector

- muon tracking, electromagnetic calorimeter, central tracking and hadron calorimeter

LHCb

- to study the differences between matter and antimatter

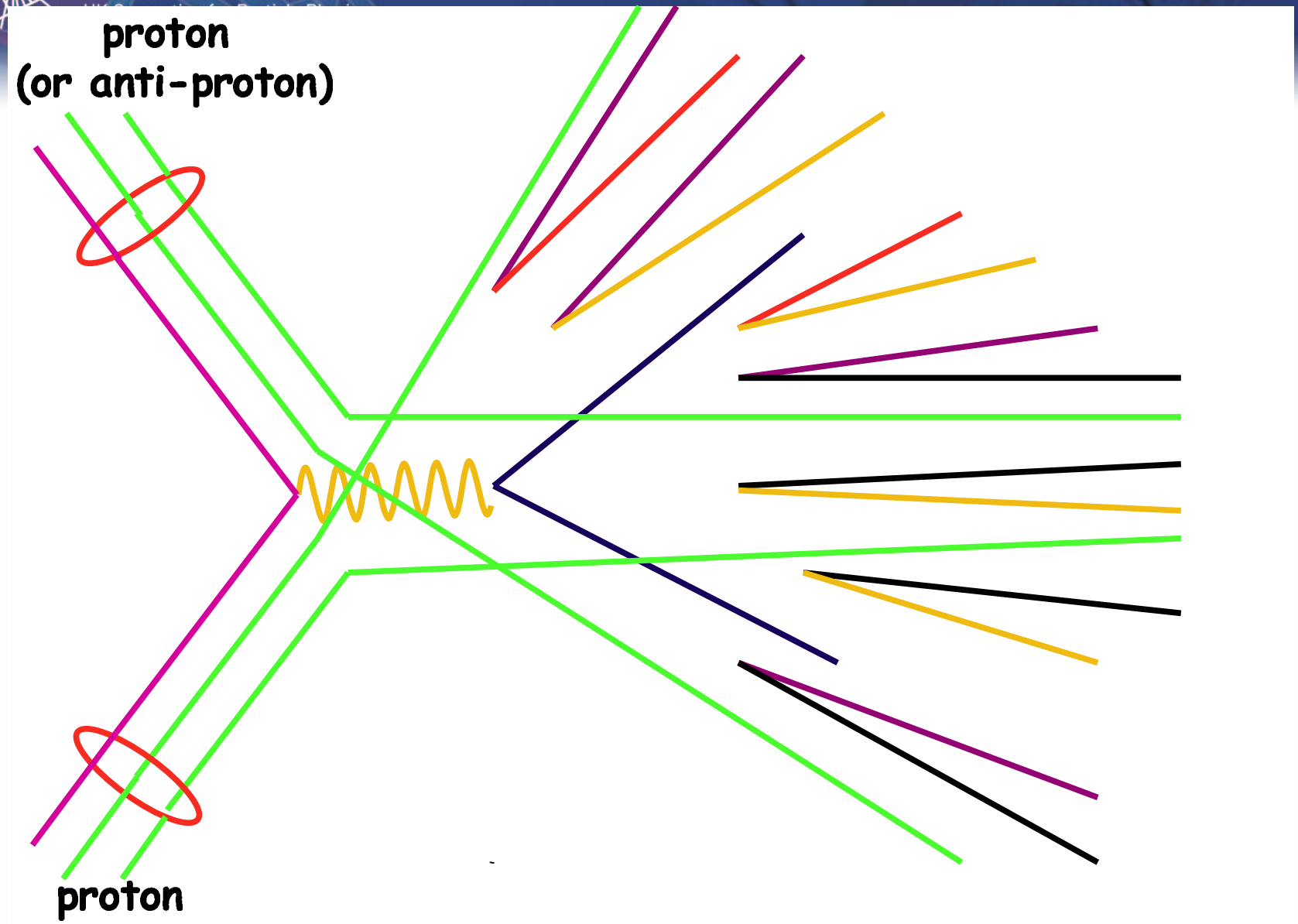
- producing over 100 million b and b-bar mesons each year

ALICE

- heavy ion collisions, to create quark-gluon plasmas

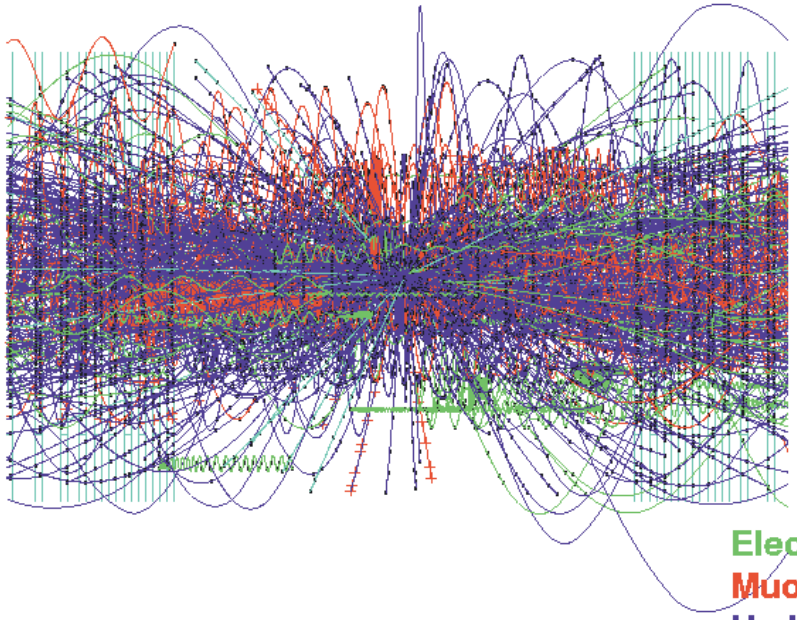
- 50,000 particles in each collision

Proton collisions produce a lot of debris ....



The LHC Data Challenge

Each event is complicated:

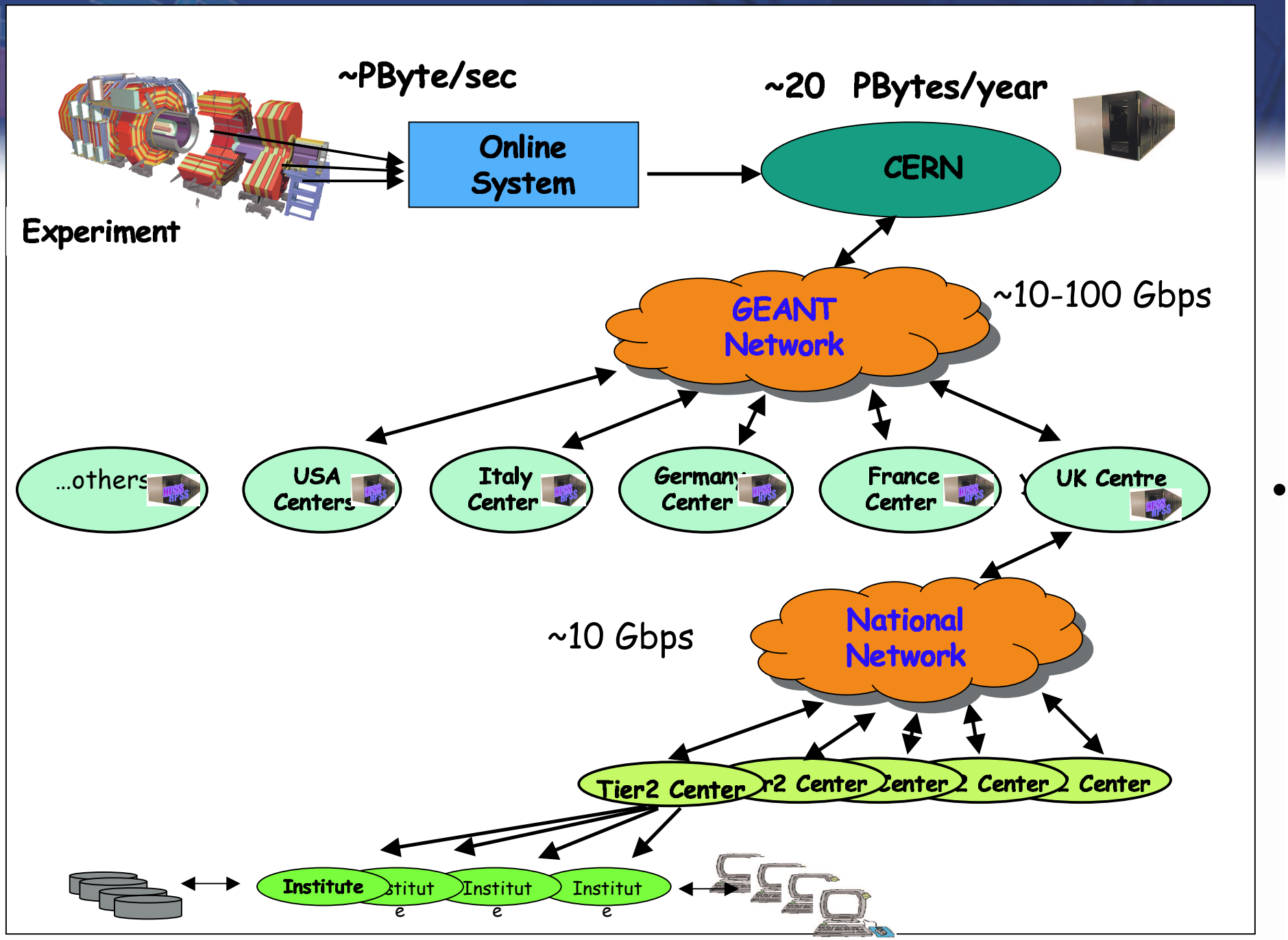


Typical Selectivity:
1 in 10^{11}

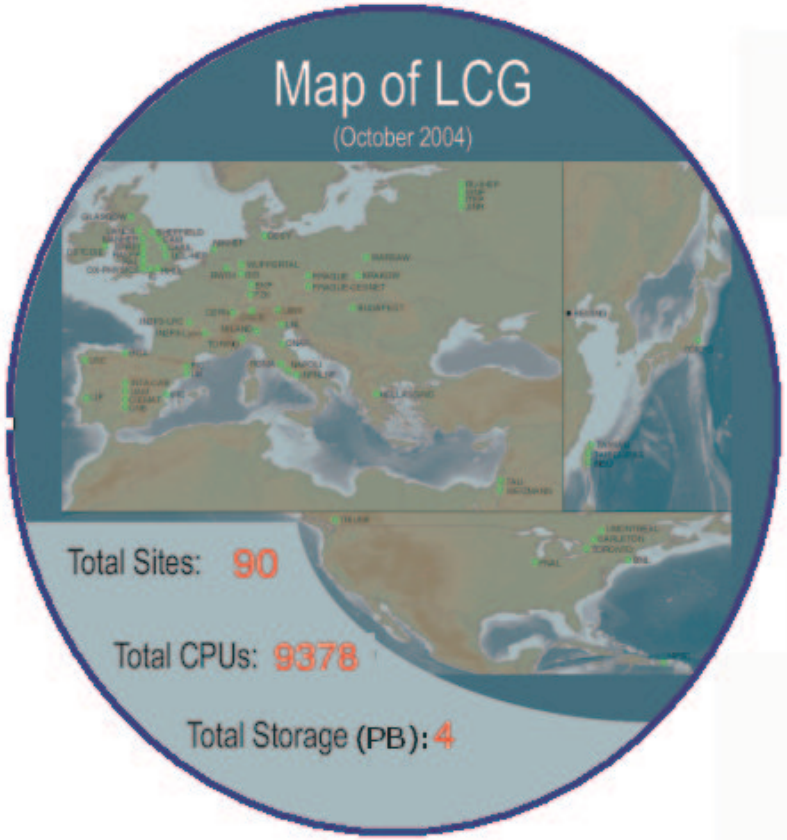
Like looking for 1 person in a ten world populations

Or for a needle in 20,000 haystacks!

CERN


~PByte/sec
~20 PBytes/year
Online System
CERN
Experiment
GEANT Network
~10-100 Gbps
…others
USA Centers
Italy Center
Germany Center
France Center
UK Centre
~10 Gbps
National Network
Tier2 Center
Institute
Institut e
Institut e
Institut e

LHC Computing Grid (LCG)



By 2007:
- 100,000 CPU
- 100 institutions worldwide

- building on software being developed in advanced grid technology projects, both in Europe and in the USA

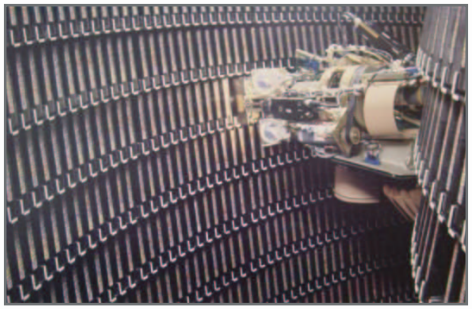
- prototype went live in September 2003 in 12 countries

GridPP
UK Computing for Particle Physics

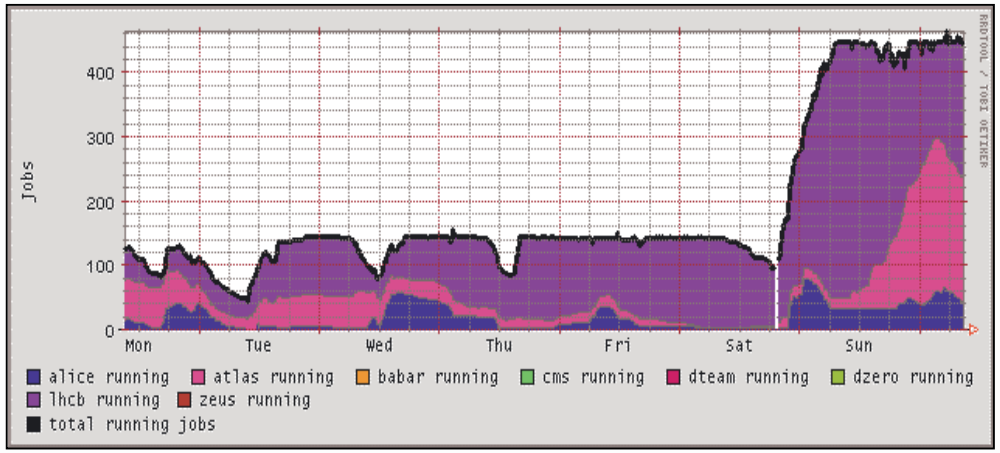
Deployment : UK Tier-1@ RAL

- High quality data services
- National and International Role
- UK focus for International Grid development

- 700 Dual CPU
- 80 TB Disk
- 60 TB Tape (Capacity 1PB)



Grid Operations Centre



GridPP
UK Computing for Particle Physics

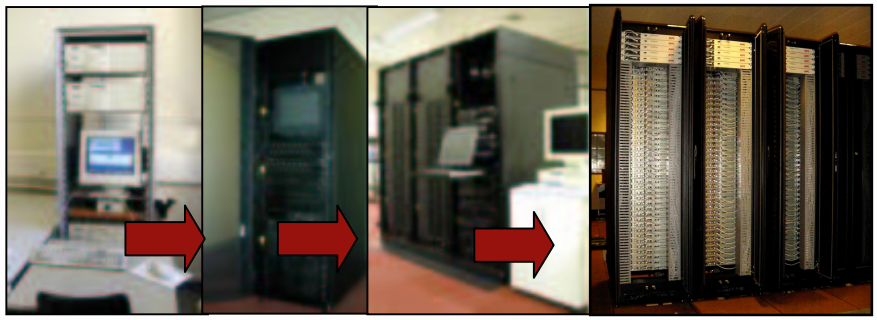
Deployment: UK Tier-2 Centres

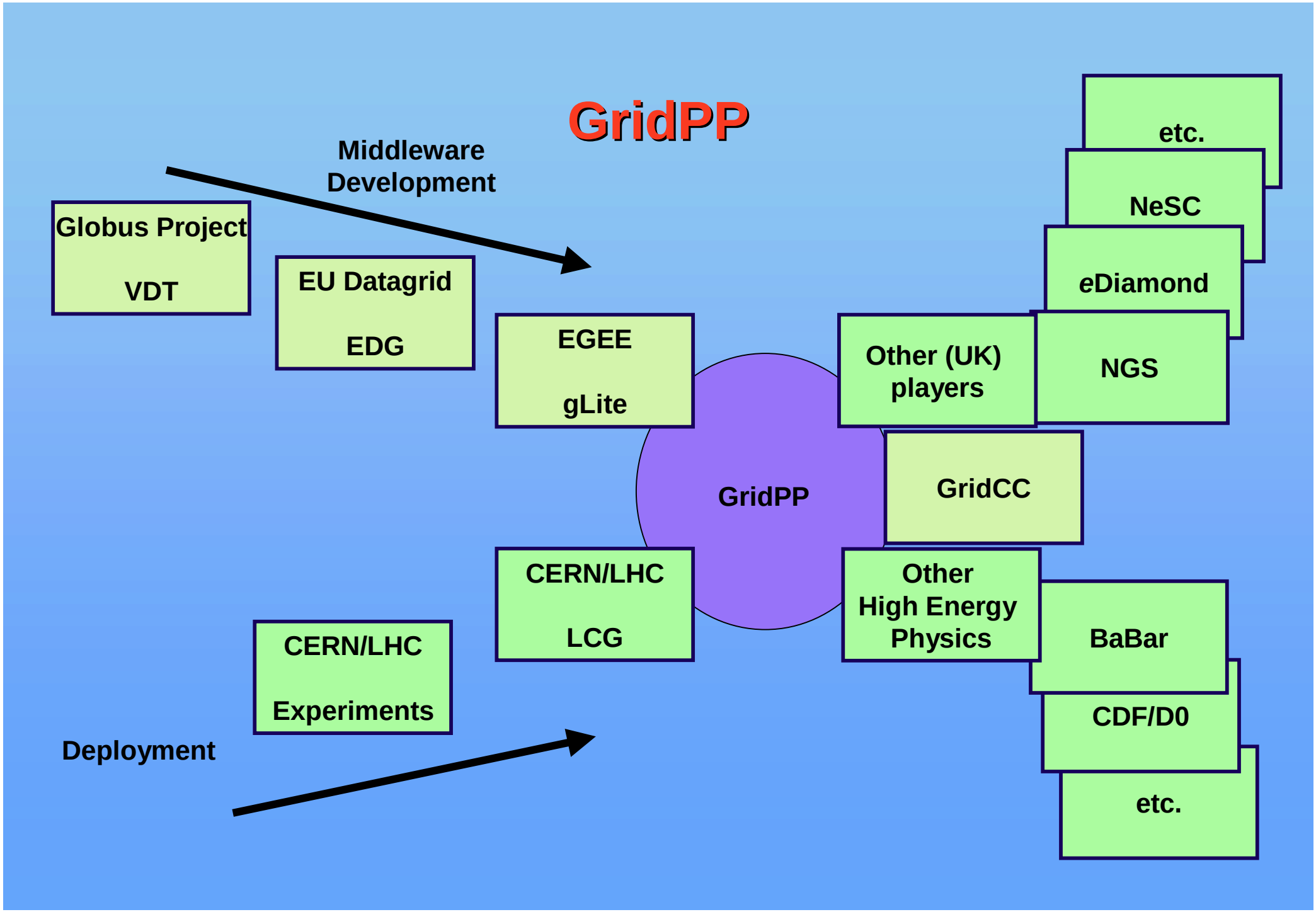
ScotGrid
Durham, Edinburgh, Glasgow

NorthGrid
Daresbury, Lancaster, Liverpool, Manchester, Sheffield

SouthGrid
Birmingham, Bristol, Cambridge, Oxford, RAL PPD, Warwick

London Tier2
Brunel, Imperial, QMUL, RHUL, UCL



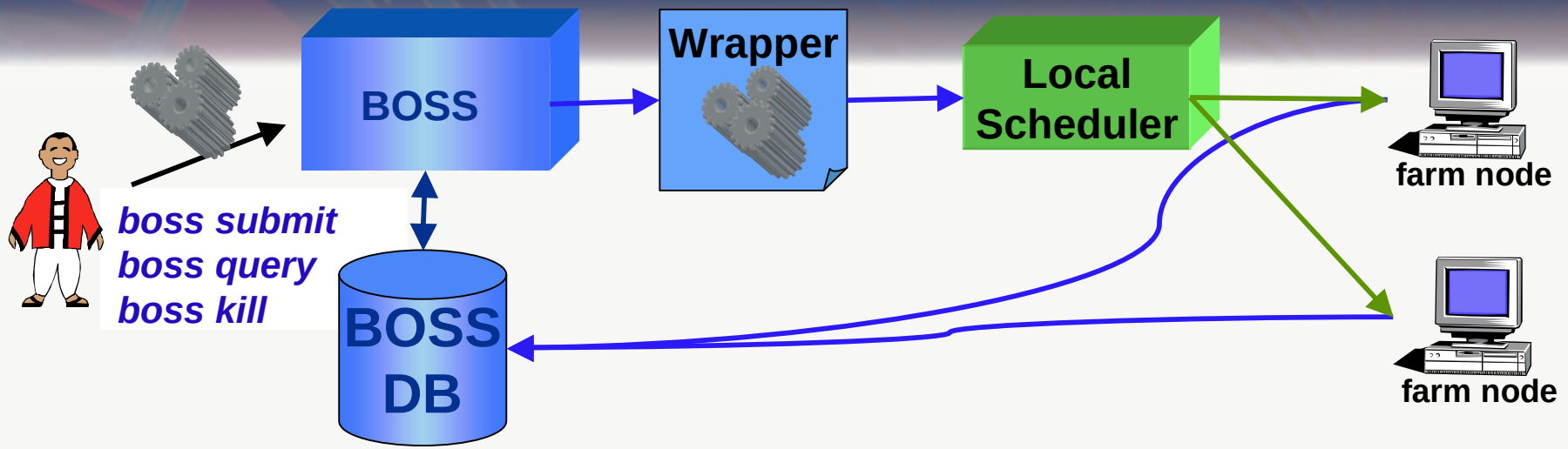

GridPP
Middleware Development
Globus Project
VDT
EU Datagrid
EDG
EGEE
gLite
CERN/LHC
LCG
CERN/LHC
Experiments
Deployment
GridPP
Other (UK) players
etc.
NeSC
eDiamond
NGS
GridCC
Other High Energy Physics
BaBar
CDF/D0
etc.

Job Monitoring

Management of large Monte Carlo productions (~3000 jobs) or data analyses and the quality assurance of the results requires careful monitoring and bookkeeping.

BOSS (Batch Object Submission System) has been developed within CMS to provide bookkeeping and real-time monitoring of jobs submitted to a compute farm system.

BOSS Basic flow



Accepts job submission from users

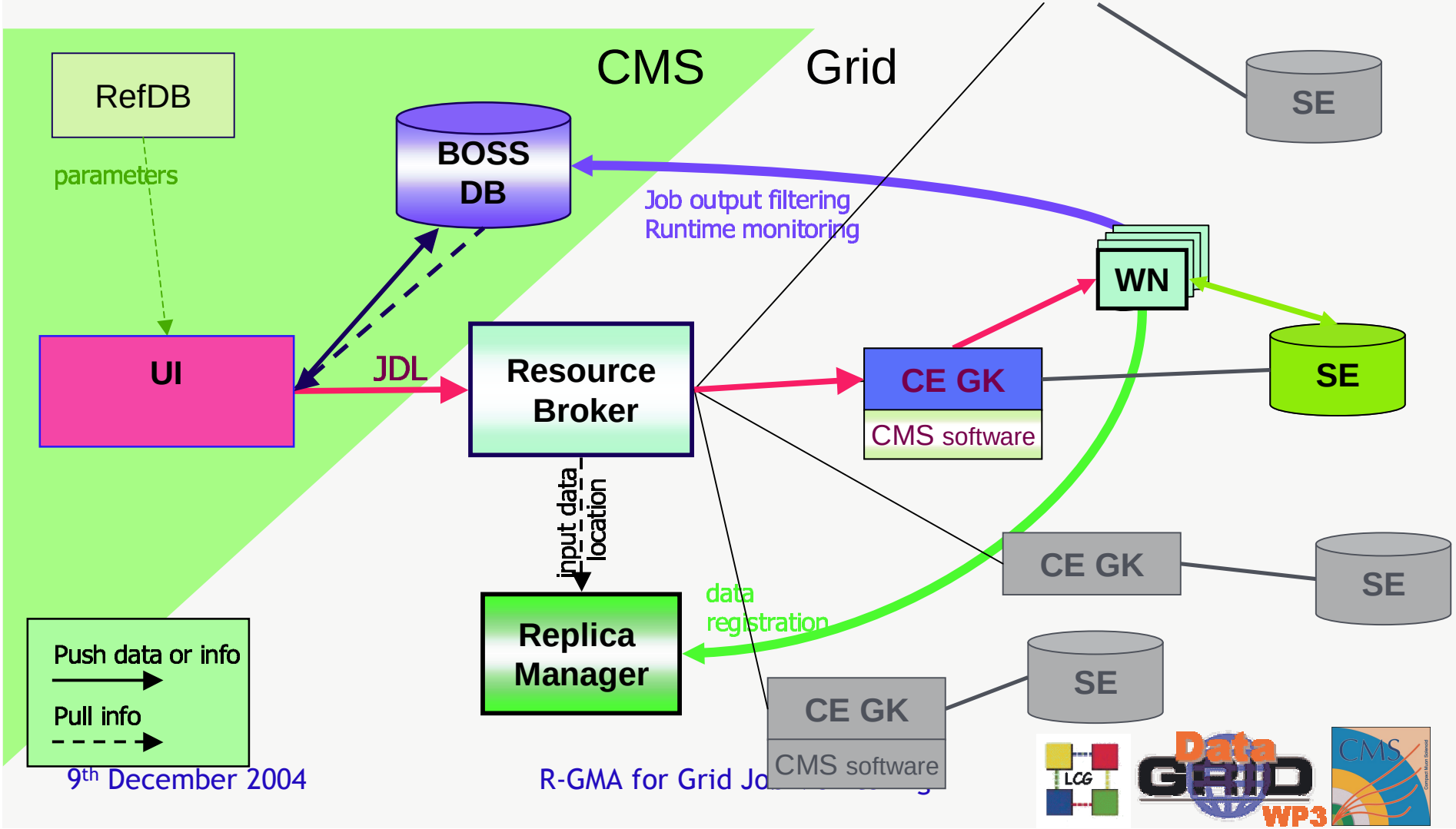
Stores info about job in a DB

Builds a wrapper around the job (*jobExecutor*)

Sends the wrapper to the local scheduler

The wrapper sends info about the job to the DB

Grid and BOSS



BOSS and the Grid

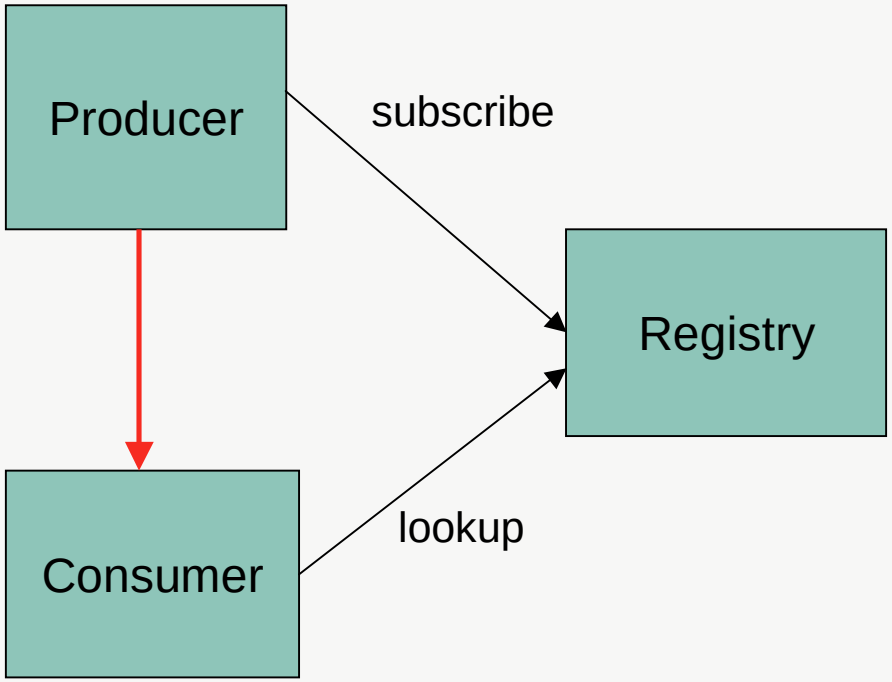
Problems in Grid context:

- Large number of simultaneous connections into DBMS
- DBMS must be open to world
- Unease over WN connectivity requirements

Can avoid by using R-GMA to move on-line monitoring data.

R-GMA is an EDG WP3 Middleware product.

R-GMA

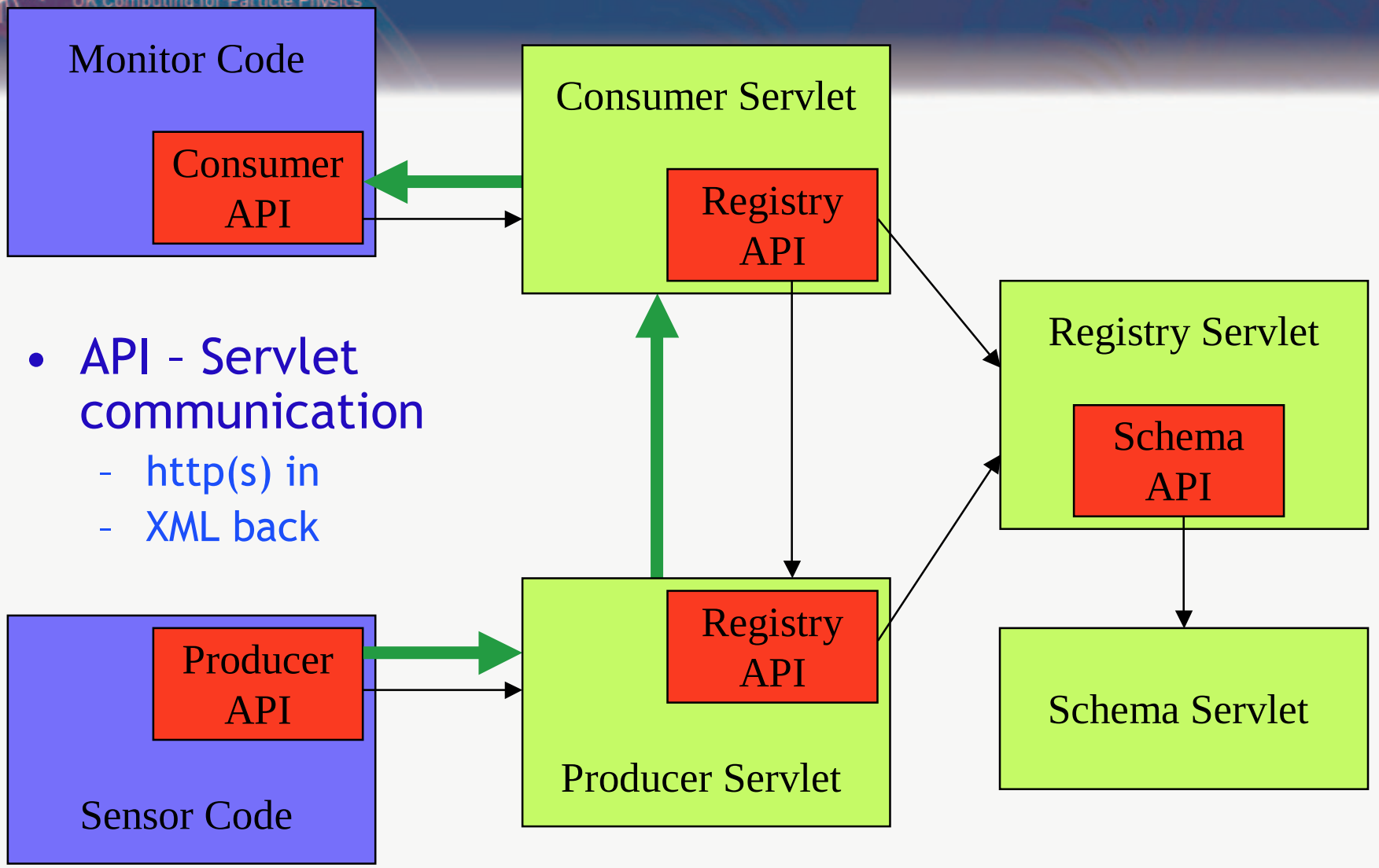


- Uses the GMA from GGF
- A relational implementation
- Applied to both information and monitoring
- Creates impression that you have one RDBMS per VO

R-GMA

- Not a general distributed RDBMS system, but a way to use the relational model in a distributed environment
- Producers
  - announce: SQL "**CREATE TABLE**"
  - publish: SQL "**INSERT**"
- Consumers collect: SQL "**SELECT**"

R-GMA



- API – Servlet communication
  - http(s) in
  - XML back

Schema & Contributions

CPULoad (Global Schema)				
Country	Site	Facility	Load	Timestamp
UK	RAL	CDF	0.3	19055711022002
UK	RAL	ATLAS	1.6	19055611022002
UK	GLA	CDF	0.4	19055811022002
UK	GLA	ALICE	0.5	19055611022002
CH	CERN	ALICE	0.9	19055611022002
CH	CERN	CDF	0.6	19055511022002

CPULoad (Producer 1)				
UK	RAL	CDF	0.3	19055711022002
UK	RAL	ATLAS	1.6	19055611022002

CPULoad (Producer 2)				
UK	GLA	CDF	0.4	19055811022002
UK	GLA	ALICE	0.5	19055611022002

CPULoad (Producer3)				
CH	CERN	ATLAS	1.6	19055611022002
CH	CERN	CDF	0.6	19055511022002

Contributions are Views

CPULoad (Producer 1)				
UK	RAL	CDF	0.3	19055711022002
UK	RAL	ATLAS	1.6	19055611022002

```
SELECT * FROM  cpuLoad
WHERE country = 'UK' AND site = 'RAL'
```

CPULoad (Producer 2)				
UK	GLA	CDF	0.4	19055811022002
UK	GLA	ALICE	0.5	19055611022002

```
SELECT * FROM  cpuLoad
WHERE country = 'UK' AND site = 'GLA'
```

R-GMA Summary

R-GMA is an information transport infrastructure for the Grid:

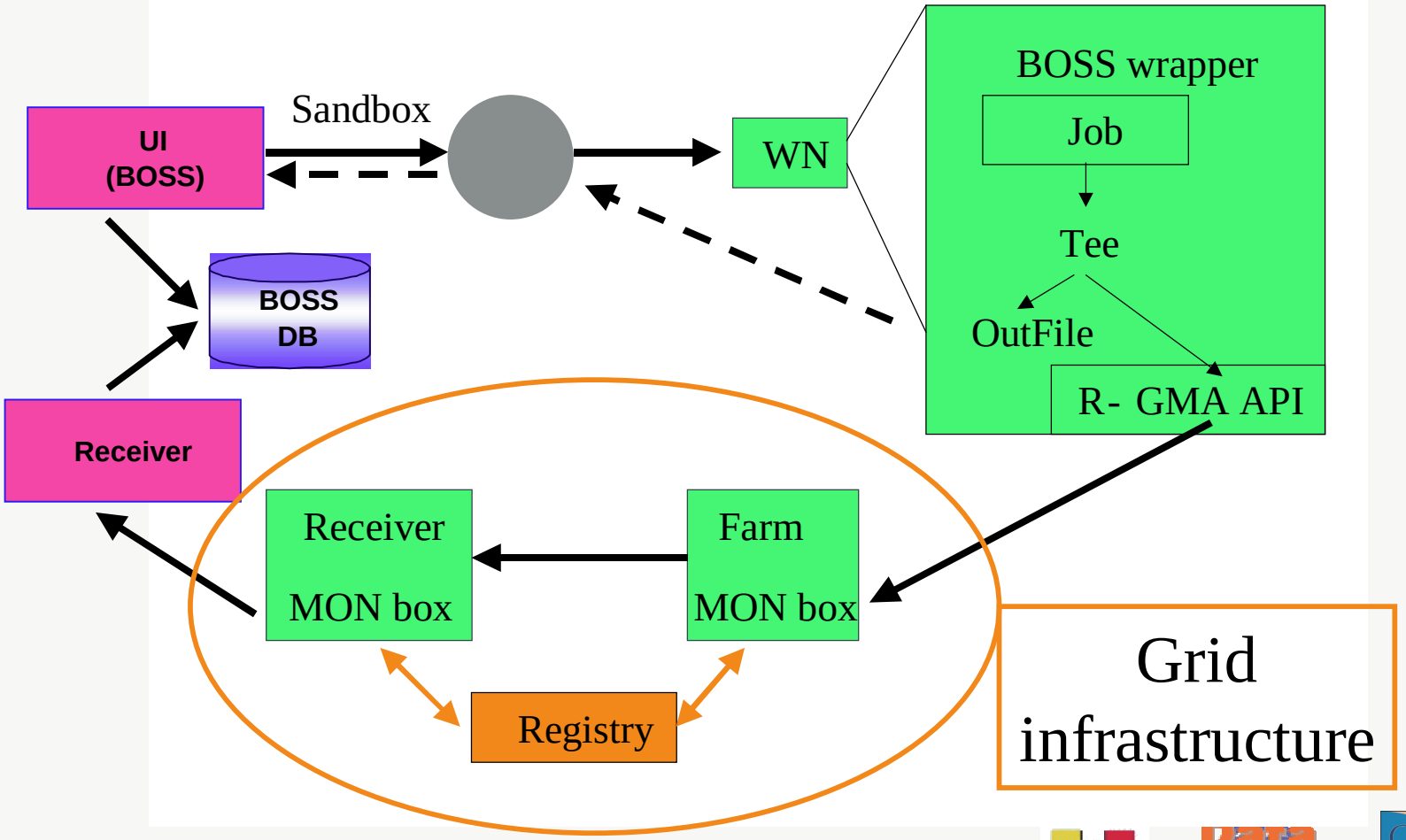
- It deals with structured data within a user-defined schema
- Routing is done inside the R-GMA layer: consumers don't have to know where the producers will be, nor do the producers care where or who the consumers are… information is simply published and consumed.

Use of R-GMA in BOSS

- Each running job has a Producer that announces host and name of "home" BOSS DB, and BOSS' jobId; this identifies it uniquely.
- Publish each update into R-GMA as a separate message – separate "row".
- Receiver **SELECT**s for all rows relating to its DB; uses jobId and jobType to do MySQL **UPDATE**.
- The Registry is a matchmaker between producers and consumers.

Use of R-GMA within BOSS

Use of R - GMA in BOSS



BOSS/R-GMA Demo

The screenshot below shows the streamed output messages from a Brunel job (ID 112) being sent through R-GMA and displayed using the Pulse tool. As Pulse can monitor multiple producers, it also shows the output from a longer job already running at Imperial (ID 72).

The receivers that update the BOSS databases use the bossDatabaseHost and bossDatabaseName fields to select only the relevant messages, so that the database at each institute is updated with only the information about its own jobs.

SELECT * FROM bossJobExOutMessage

bossDatabaseHost[]	bossDatabaseName[]	bossJobId[]	bossJobtype[]	bossVarName[]	bossVarValue[]	timeStamp[]
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	comment	I_am_fully_operational_and_all_my_circuits_are_functioning_perfectly.	1043425943
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	majorcount	204	1043425943
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	tick	15	1043425943
young.brunel.ac.uk:0	boss_v3_3_young	112	JOB	E_HOST	young	1043426585
young.brunel.ac.uk:0	boss_v3_3_young	112	JOB	E_PATH	/home/boss/boss-v3_3_pre5/CounterDemo	1043426585
young.brunel.ac.uk:0	boss_v3_3_young	112	JOB	E_USR	eesrjjn	1043426585
young.brunel.ac.uk:0	boss_v3_3_young	112	JOB	T_START	1043426579	1043426585
young.brunel.ac.uk:0	boss_v3_3_young	112	counterdemo	comment	START...	1043426585
young.brunel.ac.uk:0	boss_v3_3_young	112	counterdemo	majorcount	0	1043426585
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	comment	Message_7:_This_is_message_number_7._Message_7_ends.	1043425948
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	majorcount	207	1043425948
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	tick	6	1043425949
young.brunel.ac.uk:0	boss_v3_3_young	112	counterdemo	majorcount	0	1043426590
young.brunel.ac.uk:0	boss_v3_3_young	112	counterdemo	tick	1	1043426590
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	comment	Brain_the_size_of_a_planet_and_he_has_me_count_to_twenty!_Bah.	1043425954
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	majorcount	209	1043425954
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	tick	17	1043425954
young.brunel.ac.uk:0	boss_v3_3_young	112	counterdemo	comment	I''m_sorry_Dave,_I''m_afraid_I_can''t_do_that.	1043426595
young.brunel.ac.uk:0	boss_v3_3_young	112	counterdemo	majorcount	2	1043426595
young.brunel.ac.uk:0	boss_v3_3_young	112	counterdemo	tick	13	1043426595
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	comment	There''s_a_pain_in_the_diodes_all_the_way_up_my_left_side.	1043425959
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	majorcount	212	1043425959
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	tick	8	1043425959
young.brunel.ac.uk:0	boss_v3_3_young	112	JOB	RET_CODE	0	1043426600
young.brunel.ac.uk:0	boss_v3_3_young	112	JOB	T_STAT	0.07s user 0.01s sys	1043426600
young.brunel.ac.uk:0	boss_v3_3_young	112	JOB	T_STOP	1043426600	1043426600
young.brunel.ac.uk:0	boss_v3_3_young	112	counterdemo	comment	That''s_all,_folks!	1043426600
young.brunel.ac.uk:0	boss_v3_3_young	112	counterdemo	majorcount	5	1043426600
young.brunel.ac.uk:0	boss_v3_3_young	112	counterdemo	tick	20	1043426600
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	majorcount	214	1043425964
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	tick	19	1043425964
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	majorcount	217	1043425969
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	tick	9	1043425969
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	comment	I''m_sorry_Dave,_I''m_afraid_I_can''t_do_that.	1043425974
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	majorcount	220	1043425974
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	tick	20	1043425974
gw30.hep.ph.ic.ac.uk:0	boss_v3_3	72	counterdemo	majorcount	222	1043425979

R-GMA and BOSS

- R-GMA smoothes firewall issues
- Consumer can watch many producers; producers can feed multiple consumers
- Provides uniform access to range of monitoring data (network, accounting...)
- BOSS job wrapper uses an R-GMA StreamProducer and C++ API
- Can define minimum retention period, but ultimately no guarantees
- BOSS receiver implemented in Java

Assessing BOSS/R-GMA

Need to ensure R-GMA can cope with volume of expected traffic and is scalable.

CMS production load estimated at around 3000 jobs, each lasting about 10 hours.

- A Java MC simulation represents a typical CMS job: emulates "CMSIM" message publishing pattern, but with 10 hour run time compressed
- … so actually have fewer simultaneous simjobs than real case, but overall a much higher rate of message production.

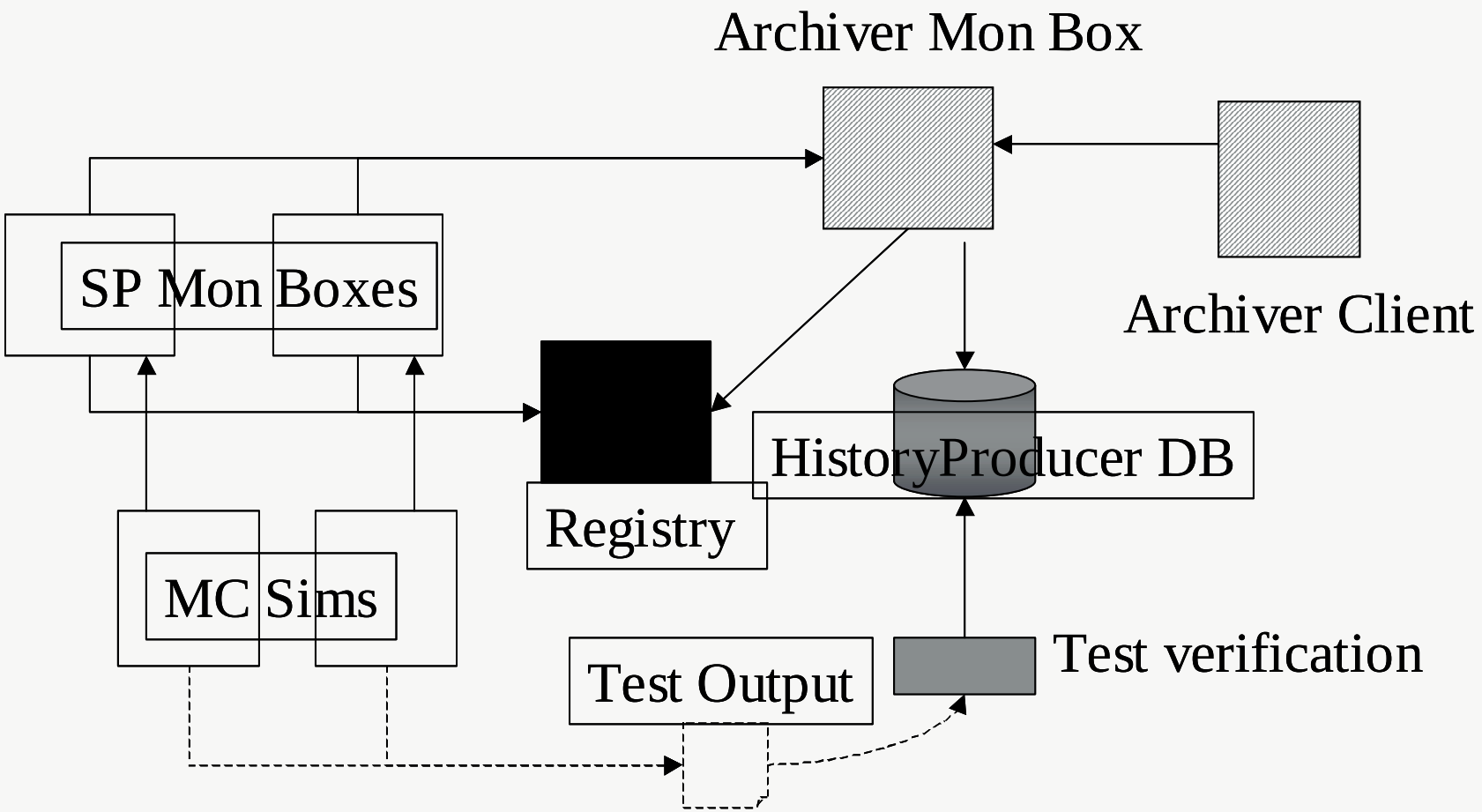
Assessing BOSS/R-GMA

- Submit simjobs to the Grid; see
  - if messages get back
  - how many come back
- Don't need to do the number crunching in between so have multiple threads producing – small number of Grid jobs can put large load on R-GMA

Assessing BOSS/R-GMA

- An R-GMA Archiver/HistoryProducer scoops up published tuples
- The HistoryProducer DB used is a representation of the BOSS DB, but stores history of received messages, rather than just a cumulative update – can compare with published tuples to verify the test outcome

Use of R-GMA within BOSS



Scalability tests of BOSS/R-GMA

For scalability tests the compressed simjob ran for just 60 s.

Initial tests with R-GMA v3.3.28 only managed about 400 simjobs (NSS '03 / *IEEE T. Nuc. Sci.*).

- Limitations in R-GMA implementation, e.g. too few streaming sockets – 1024; insufficient memory allocated to the JVM
- Various configuration problems at both Imperial and Brunel sites

Scalability tests of BOSS/R-GMA

Modified version of these CMS tests now forms part of R-GMA performance test suite – feedback into R-GMA development.

- 1 MC producer creating 2000 simjobs and publishing 7600 tuples proven to work without glitch (R-GMA v3.4.13)
- Demonstrated 2 MC producers each running 4000 simjobs (with 15200 published tuples)
  (EDG WP3 testbed)

BOSS/R-GMA on LCG 2

- R-GMA is part of the **LCG** 2.2.0 release
- We need to confirm that it can handle load of applications monitoring under "real-world" *deployment* and *operation* conditions

Deployment of BOSS/R-GMA on LCG 2

Significant problems getting jobs to run at all
- large proportion of sites misconfigured: e.g. on 13th October only 13 of 24 matching resources were able to run a test job and transfer data back to Brunel.

Deployment of BOSS/R-GMA on LCG 2

- At 3 sites messages were published to their MON box but didn't reach the consumer (all have since confirmed firewall issue)
- At 3 sites the MON box refused connections
- 3 sites "false advertising": R-GMA environment advertised even though not installed or configured
- 2 sites aborted job – due to other Grid problems

Deployment on LCG 2: Current Status

Successful deployment of a complex infrastructure spanning the globe is difficult: most sites are run not by Grid developers but by sysadmins with major non-Grid responsibilities. Confusing, missing or incorrect documentation causes major headaches.

R-GMA deployment has since improved drastically – but is still not 100%.

Operation on LCG 2: Current Status

Initially use the CMSIM "emulator" from previous tests, but now with a runtime of ~30 min.

Simultaneous submission of 50-simjob producers to 4 manually specified sites – all 14800 messages transferred successfully.

Risks: unlikely to break Registry

test MON boxes off-Grid or through aggregation

Operation on LCG 2: Current Status

- 4 400-simjob producers has worked … once.
- General problems with Registry – single point-of-failure. Not yet tested since work-around put in place.

LCG in action

http://www.hep.ph.ic.ac.uk/e-science/projects/demo/index.html

http://map.gridpp.ac.uk/

http://goc.grid-support.ac.uk/gppmonWorld/gppmon_maps/lcg2.html

http://goc.grid-support.ac.uk/gridsite/gocmain/monitoring/

The End