ATLAS MC Use Cases

Zhongliang Ren

12 June 2006

WLCG Tier2 Workshop at CERN

Table of Contents

- ATLAS computing TDR and roles of Tier2s
- Current MC productions
- Basic requirements at Tier2s for ATLAS MC production
- ATLAS job requirement and site configurations
- Missing tools and grid middleware functionality
- Data services at Tier2s
- Near future plan & schedule

Operation parameters, types and sizes of event data and processing times

Raw event data rate from online DAQ	Hz	200
Operation time	seconds/day	50000
Operation time	days/year	200
Operation time (2007)	days/year	50
Event statistics	events/day	10^7
Event statistics (from 2008 onwards)	events/year	$2 \cdot 10^9$
Raw Data Size	MB	1.6
ESD Size	MB	0.5
AOD Size	KB	100
TAG Size	KB	1
Simulated Data Size	MB	2.0
Simulated ESD Size	MB	0.5
Time for Reconstruction	kSI2k-sec/event	15
Time for Simulation	kSI2k-sec/event	100
Time for Analysis	kSI2k-sec/event	0.5

Roles of Tier2

From the ATLAS computing TDR:

- Monte Carlo event simulation
  - Full chain MC production including event generation, simulation, digitization, event pileup and reconstruction
  - MC production rate equivalent to >20% of ATLAS raw event rate
  - Total Tier2 CPU resources should cover the MC production requirements
  - MC data to be stored at Tier1 to which the Tier2 is associated
- User physics analyses (See separate talk by Dietrich Liko)
  - Hosting (part of) AOD (200TB/year) & TAG (2TB/year) data
  - For physics working groups, sub-groups
  - AOD & TAG data based
  - Chaotic, competitive with 1,000+ users (resource sharing) !
- Special Tier2s dedicated for detector calibrations

MC Production: Requirements & Deliverables

- ATLAS C-TDR assumes:
  - 200 Hz and 320 MB/sec
  - 50,000 second effective data-taking, 10 million events per day
  - MC data production equivalent to 20% of raw data rate
  - 2 million MC events per day
- Assuming ATLAS data-taking at full efficiency:
  - 17.28 million events per day
  - ~3.5 million MC events per day
- MC production deliverables in 2007-2008:
  - World-wide distributed MC production at T2s
    - 50 events per job for physics events, 100 events/job for single particle events
    - Central MC productions ~100K jobs/day
  - Additional capacity for calibration and alignment (dedicated T2s)
  - Global operation with 100+ sites
  - Needs 7x24 stable services

Current MC Productions

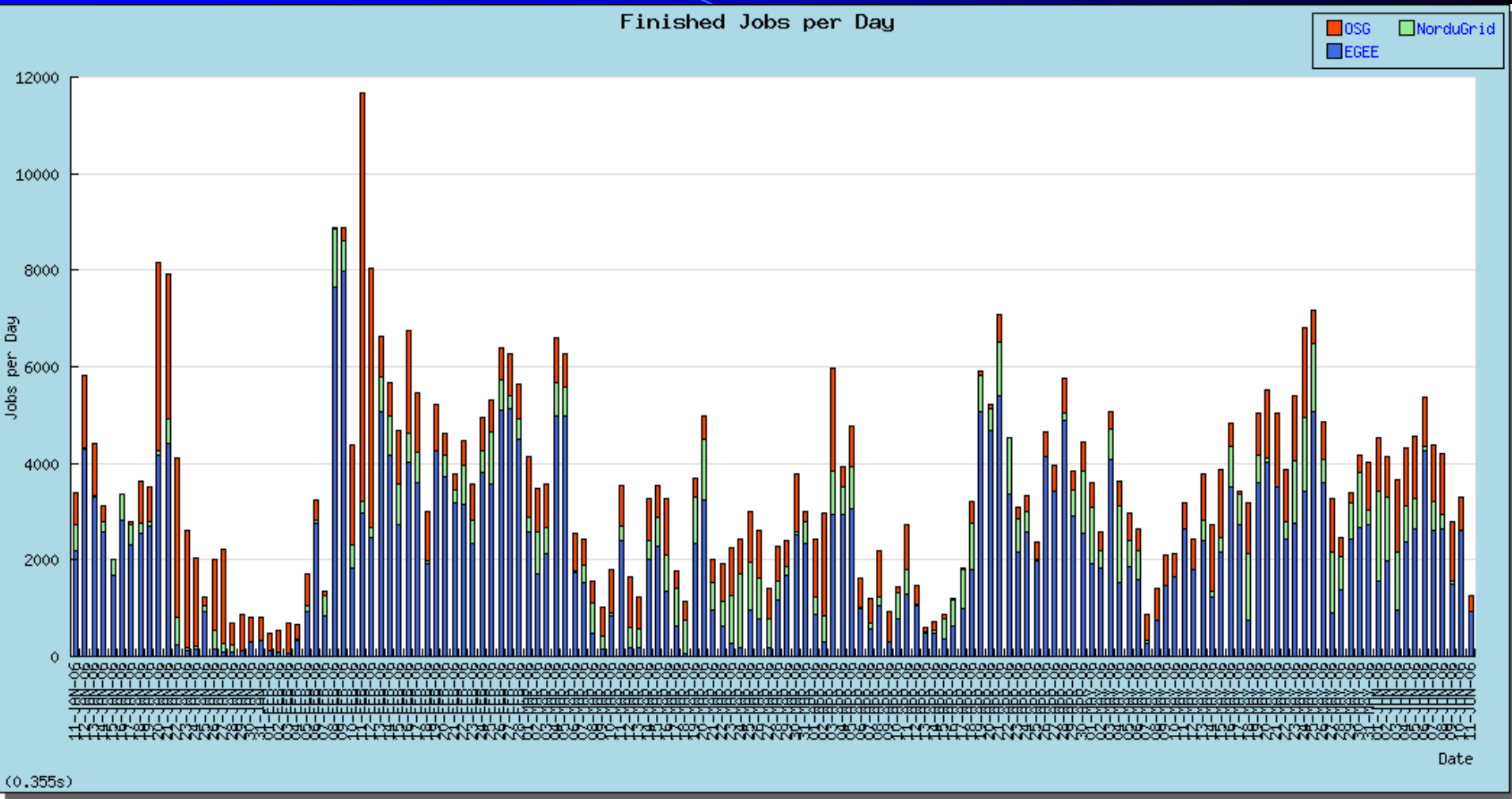
- Major objective has been the Computing System Commissioning (CSC) in 2006
- Several goals for on-going MC productions:
  - Started in Oct. 2005 as a running-in computing operations toward LHC startup in 2007
  - Testing & commissioning of the infrastructure and production systems
  - Offline software validations to catch the rare bugs at $1/10^3$ event level
  - Provide data for ATLAS detector commissioning and physics analysis
- With 150 different physics samples (of which 61 single particles)
- Used also Tier1 resources for MC production
- Being performed outside of LCG SC3/SC4 activities

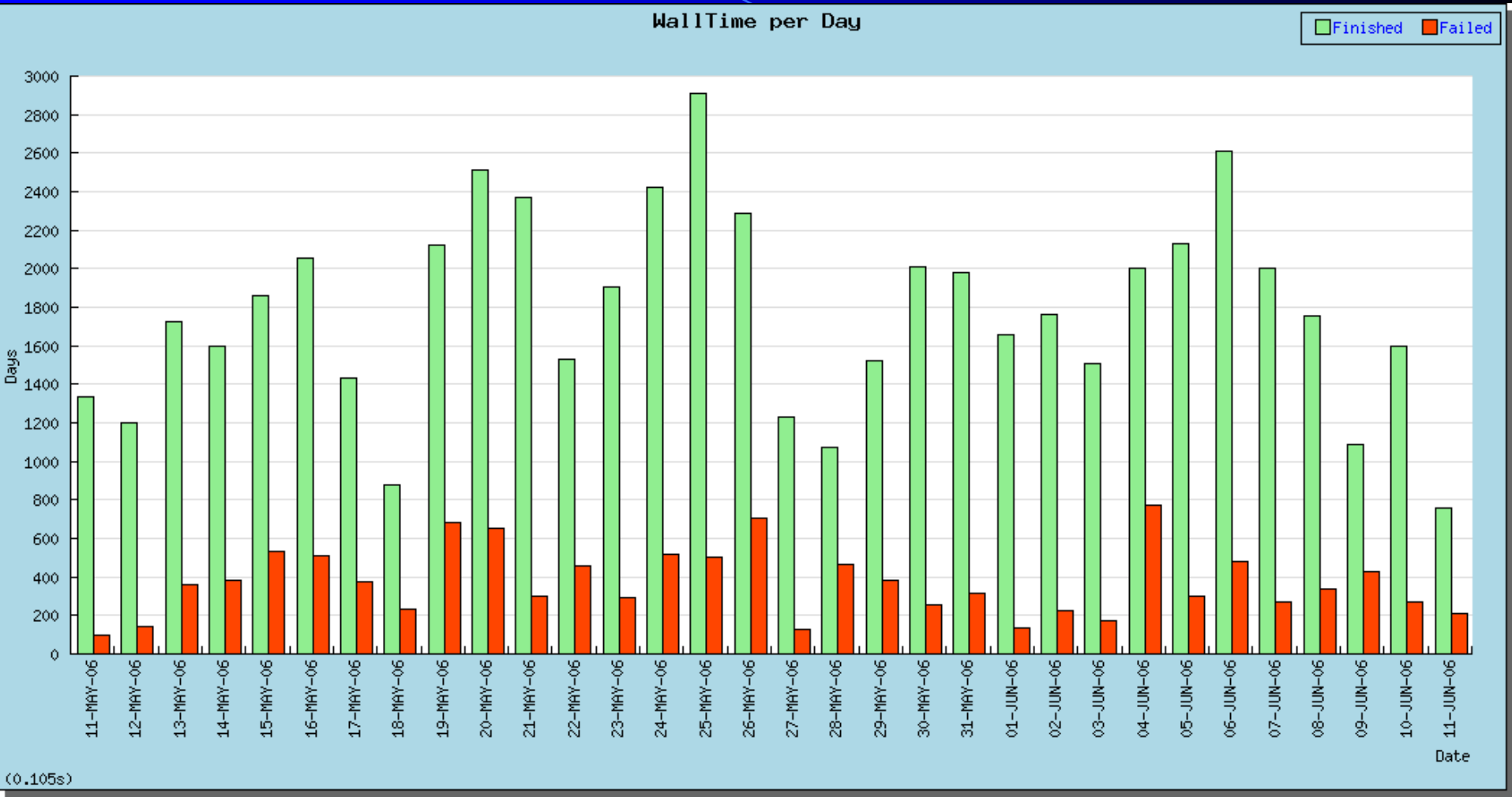
The Infrastructure and Production System

- Globablly distributed MC production using three Grids:
  - LCG/EGEE, OSG and NG
- 4 different production systems:
  - LCG-Lexor
  - LCG-CondorG
  - OSG-PANDA
  - NG – Dulcinea
- DQ2 Distributed Data Management (DDM) system
  - Integrated in OSG-PANDA, LCG-CondorG & LCG-Lexor
  - To be integrated in NG
- DDM operations
  - ATLAS VO-boxes and DQ2 servers ready at CERN and 10 T1s. FTS channel configurations done.
  - Need to configure FTS channels to the T2s!

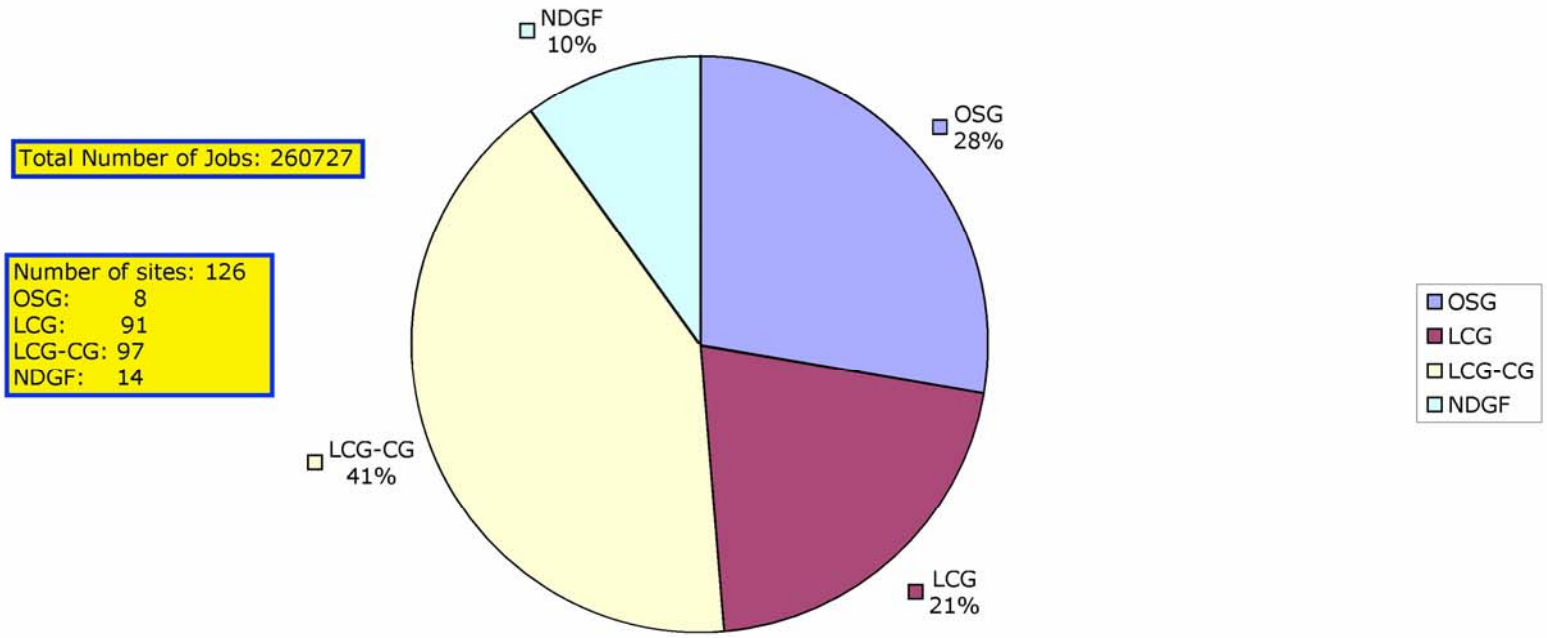
The MC Production Workflow

- The typical workflow for distributed MC production:
  - ATLAS offline software release and pacman distribution kit
    - This has been driving the production schedule
  - Validation of the release: A multiple step process
    - Pre-release nightly build and RTT (Run Time Tester) testing
    - Physics validation team signs off
    - Site and software installation validation group (central computing operations)
  - Installation of the release
  - Physics samples from ATLAS physics coordination
    - Sample A (~250k events) validation on all grids
    - Sample B (~1M events) production
    - Sample C (~10M events)
  - Job definitions
  - Validation production with Sample A and B
  - Large scale production for CSC, physics, CTB, cosmic commissioning, etc.


Finished Jobs per Day
OSG
NorduGrid
EGEE
12000
10000
8000
6000
4000
2000
0
Jobs per Day
Date
(0.355s)

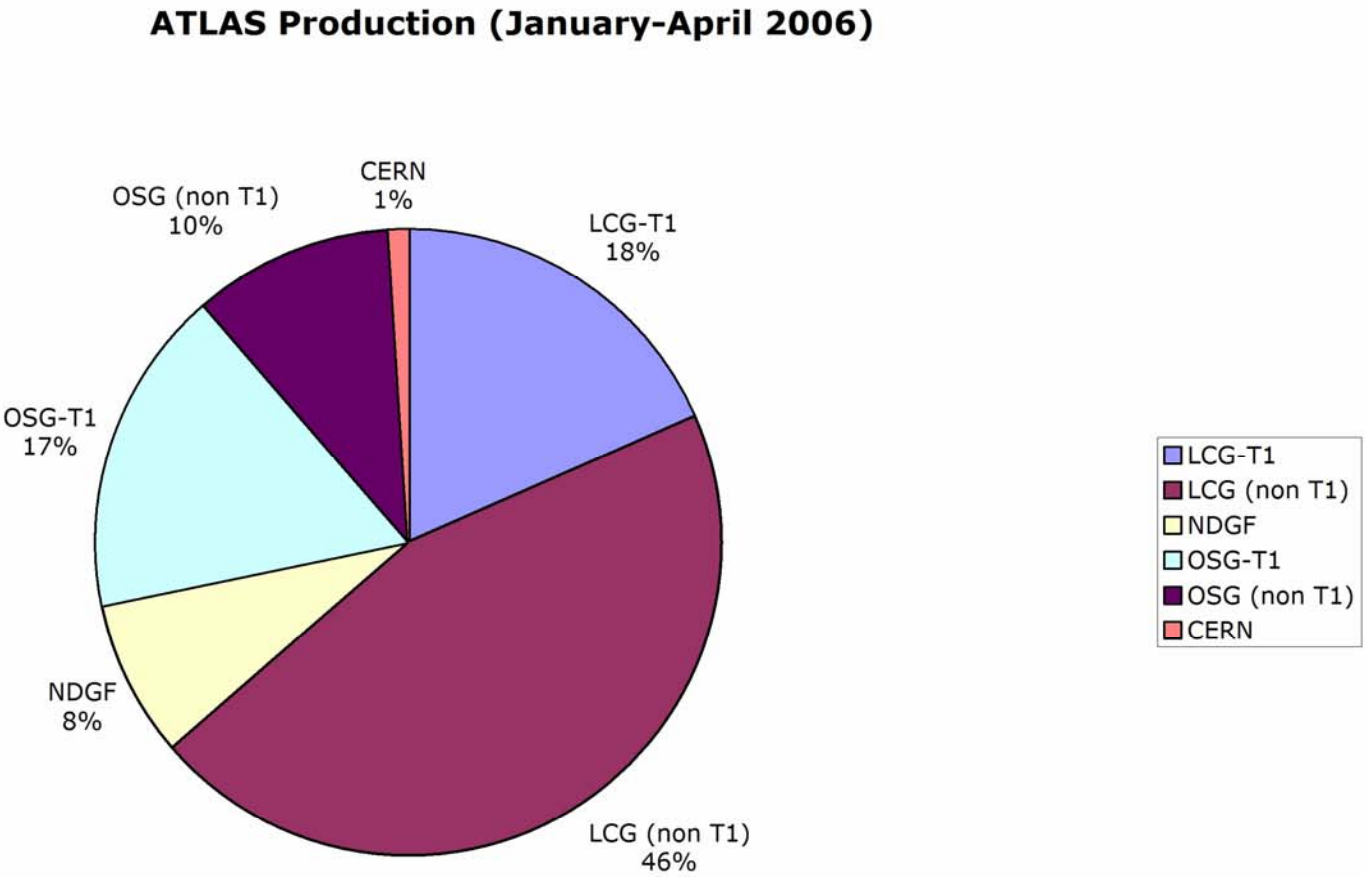

WallTime per Day
Finished
Failed
Days
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
11-MAY-06
12-MAY-06
13-MAY-06
14-MAY-06
15-MAY-06
16-MAY-06
17-MAY-06
18-MAY-06
19-MAY-06
20-MAY-06
21-MAY-06
22-MAY-06
23-MAY-06
24-MAY-06
25-MAY-06
26-MAY-06
27-MAY-06
28-MAY-06
29-MAY-06
30-MAY-06
31-MAY-06
01-JUN-06
02-JUN-06
03-JUN-06
04-JUN-06
05-JUN-06
06-JUN-06
07-JUN-06
08-JUN-06
09-JUN-06
10-JUN-06
11-JUN-06
Date
(0.105s)

ATLAS Production (January 1st - March 15th) - Number of jobs



2

Worldwide Distributed Productions



- Productions not using ProdSys
- Information on non-Grid use

Not taken into account

Summary of Current MC Productions

- Finished many cycles of validation runs:
  - ~100 bug reports filed so far
  - Found many exotic software bugs at $1/10^3$ event level in ATLAS offline software as well as in Geant4
- Provided MC data for ATLAS physics analysis and detector commissioning
- Reached an average MC production rate of ~2 million events per week, still to ramp up by a factor of 10 according to ATLAS computing model
- Thanks to all Grids (LCG/EGEE, OSG and NG) and sites!
  - Direct contacts between ATLAS and sites have been set up
  - Need synchronized information channels with LCG!
- Requires increasing resources at T2s and a more stable, production quality infrastructure towards the end of 2006 ?!

Basic requirements at Tier2s for ATLAS MC Production

- General requirements for ATLAS offline software installation:
  - A working WLCG MW installation
  - Scientific Linux 3.0.x or fully compatible OS
  - A shared file system among all WNs
  - ~20-30GB free disk space (one release needs ~10GB)
  - The glibc-headers and libX11.a (should be dropped!) installed in the WNs
  - A supportive site administrator
- Installation is done via a Pacman kit
- Managed by a central operation team. Different people take care of installations on different grids
- NG has mixed installations of the centrally-built pacman kit and recompiled RPMs for clusters with different OS

ATLAS Job Requirement and Site Configurations

- Different jobs require different minimum sizes of main memory to avoid frequent paging:
  - > 600MB for a simulation job
  - > 1.0 GB for a reconstruction job
  - > 1.3 GB for a pileup job
- Current LCG MW does not support passing individual job parameters to the local batch system! Only information published is GlueHostMainMemoryRAMSize via the Glue schema
- Problem with LCG sites having different main memory PCs! Complicated cluster and batch system configurations (double CPU PC or with Hyper-Threading)

Missing Tools and Grid MW Functionality

- Need tools for resource allocation and job priority setting among the VOs and the individual users
  - VOMS and GPBox (Grid Policy Box)?
- Need better monitoring and accounting tools for resource usage
- Require new middleware functionality to be able to pass individual job parameters down to the local batch system

Data Services at Tier2s

- MC data produced at Tier2s are transferred to and stored at the T1s they are associated to
- T2s Provide SRM SE for both raw and MC AOD data (disk resident)
- LCG Tier2 sites:
  - MC data transfer is done with the DQ2 services at the T1
  - DQ2 installations at Tier2s are not needed for now
- OSG Tier2 sites:
  - DQ2 installation is needed at every T2 in the PANDA production system
- NG Tier2 sites:
  - DQ2 not yet used!
- Data services at T2s are complicated with DB replications which still need to be worked out in the context of LCG 3D project

Event Data at Tier2s

- Not only from MC production
- T2 keeps AOD/TAG data for user analysis
- T2 with enough storage can hold complete AOD data
  - AOD of real data: 200 TB/year
  - AOD of MC data: 40 TB/year
  - ATLAS C-TDR now says 1/3 of AOD for each T2
- Complete TAG data
  - 2 TB/year
  - Can be relational TAG database and/or TAG data files
  - Relational database is presumed to be MySQL at Tier2s

Non-Event Data at Tier2s

- Geometry DB
  - Master in OracleDB@CERN, replicas distributed as SQLite files (file-based “database”)
- Calibrations and Conditions DB
  - Mixed file- and database-resident data (MySQL at Tier2s)
- Some conditions will be needed for analysis, some calibrations for simulation: no good volume estimates yet
  - Small compared to TAG database

Data File Distribution Using DQ2

- Event data and calibration “datasets” arrive by subscription in distributed data management (DDM) system (DQ2)
- Subscribing to a dataset implicitly adds data transfer requests to DQ2 transfer queue
- DDM decides the site(s) from which files in dataset will be transferred
- Tier2s will in general get their data from the associated Tier1

Database-Resident Data Distribution

- Assumption is that all database-resident data has a single master copy in OracleDB@CERN
- Tier0 -> Tier1 transfer using Oracle-specific “Streams” replication
- Tier1 -> Tier2 crosses technologies: Oracle->MySQL
- Tools for TAG and conditions database replication from Oracle to MySQL have been demonstrated, but have not been used in production to date
- Production model not well understood yet
- LCG 3D project is prototyping mechanisms for distributed database deployment, replication, and update operations

Web Caching of DB Data

- Possibility of using web caching technologies (Squid) for access to database-resident data is under investigation
- FNAL-developed FroNtier project (Squid-cache-based) is being evaluated in LCG 3D project
- Before the DB replication issue is settled, it is not very clear what will be needed at Tier2s for data services!
- The jobs will mostly remain with the ATLAS central computing operation team, and be kept minimum for the Tier2 services!

Near Future Plan and Schedule

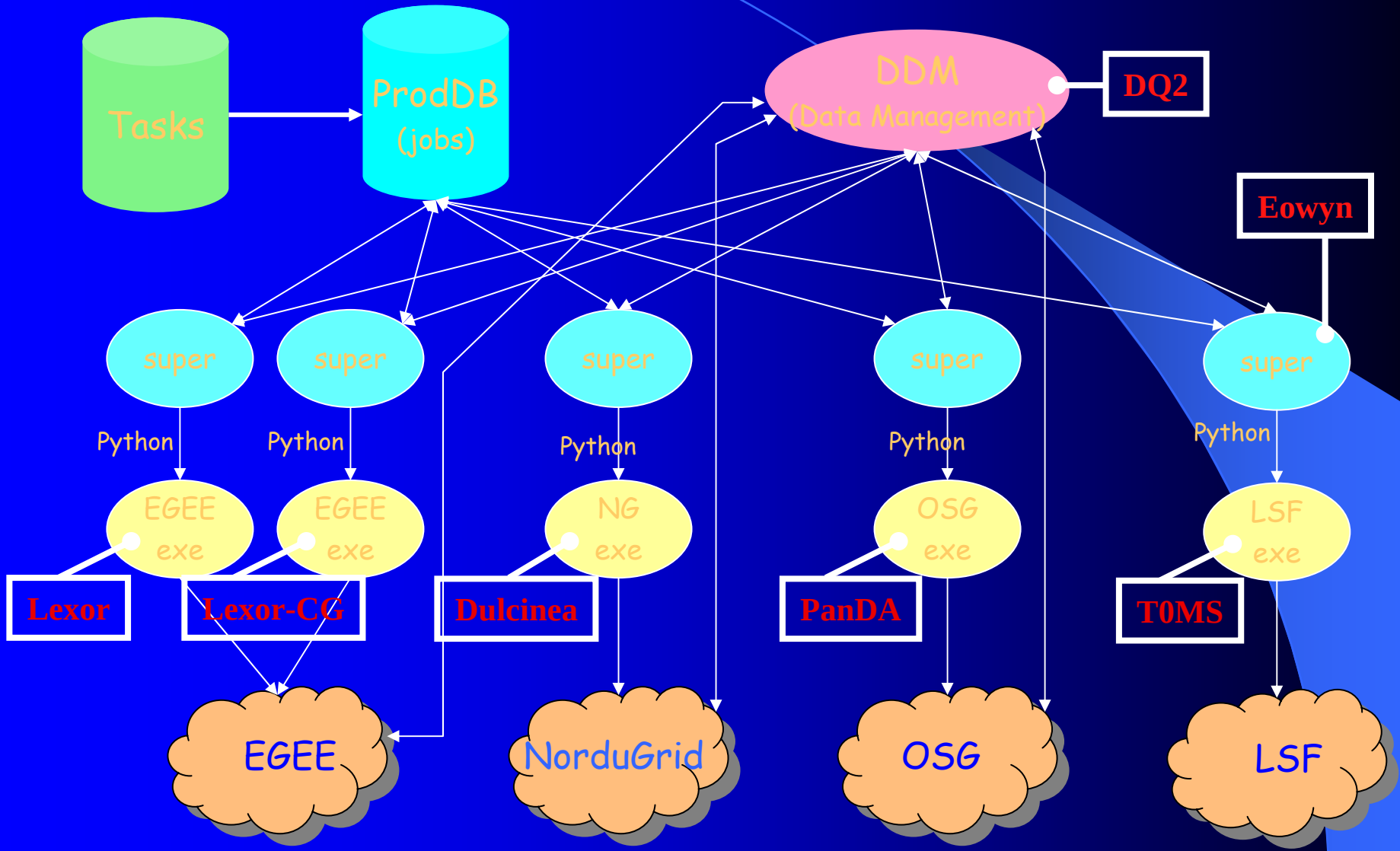
- Things urgently needed in the production system
  - DQ2/DDM fully integrated in the NG production system
- Require increased CPU resources with enough main memory for ATLAS MC production jobs
  - Need VOMS & GP-Box for resource allocation, job priority setting, etc.
- Continue ramping-up the production rate:
  - ~50K jobs/day by summer 2006
  - ~100K jobs/day by end of 2006
  - up to ~1M jobs/day (including distributed analysis jobs) by LHC startup
- Continue validation production, Perform CSC, Calibration Data Challenge production, etc. in 2nd half 2006
- Performed combined general dress rehearsal in Spring 2007 using MC produced data (ref. Next slide)

The Dress Rehearsal in Spring 2007

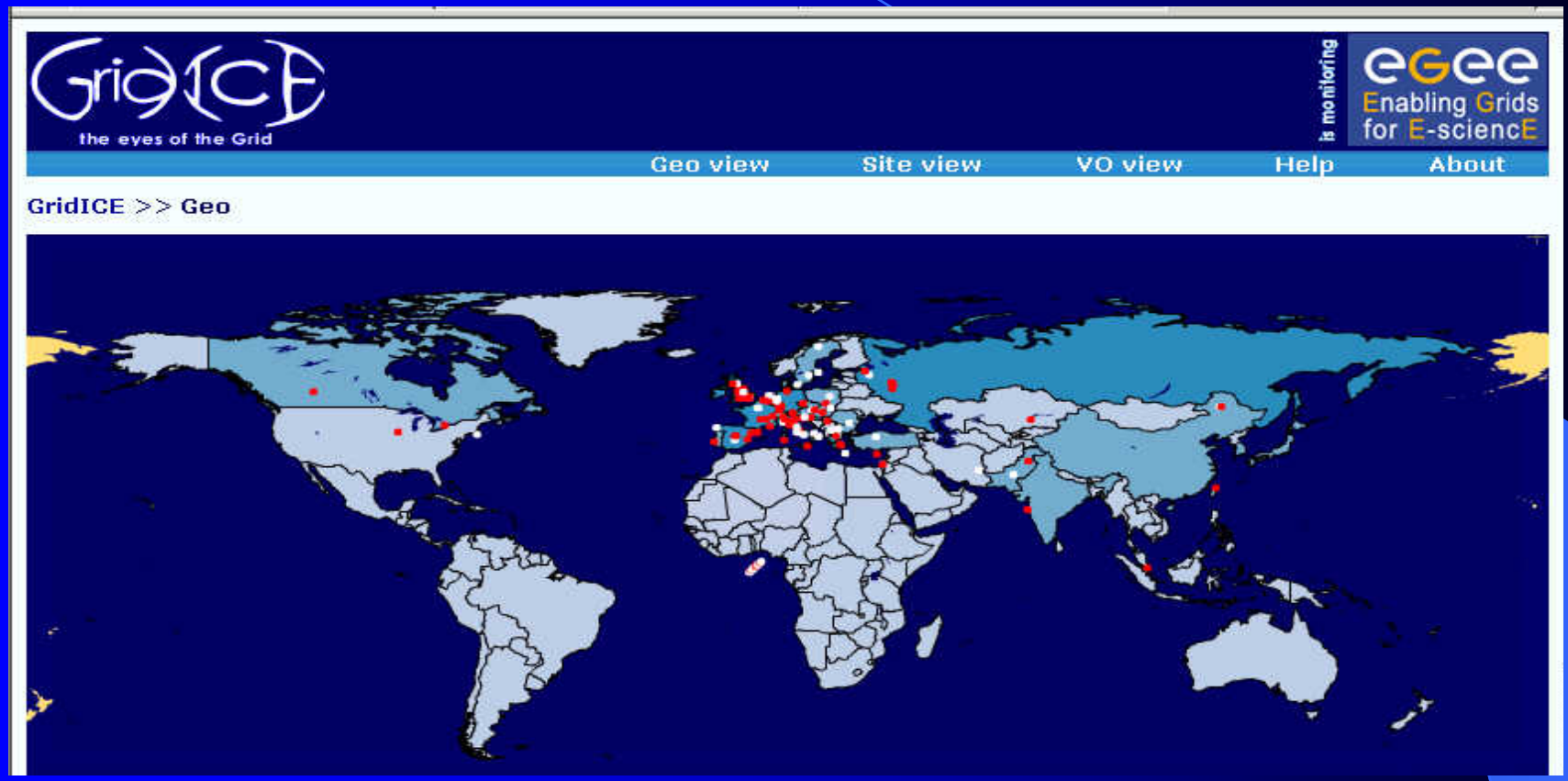
- Generate an amount of MC data close to what is expected in 2007.
- MC production scale of $N*10^7$ events ?
- Mix and filter events at MC generator level to get correct physics mixture as expected at HLT output
- Run G4 and trigger simulation and produce ByteStream data to mimic the raw data
- Send raw data to Point 1, pass through HLT nodes and SFO, write out events into streams, closing files at boundary of luminosity blocks
- Send events from Point 1 to Tier0, Perform calibration & alignment at Tier0, Run reconstruction at Tier0
- Distribute ESD, AOD, TAG data to Tier1s and Tier2s
- Perform distributed analysis at Tier2s

Backup Slides

ATLAS Production System (2006)

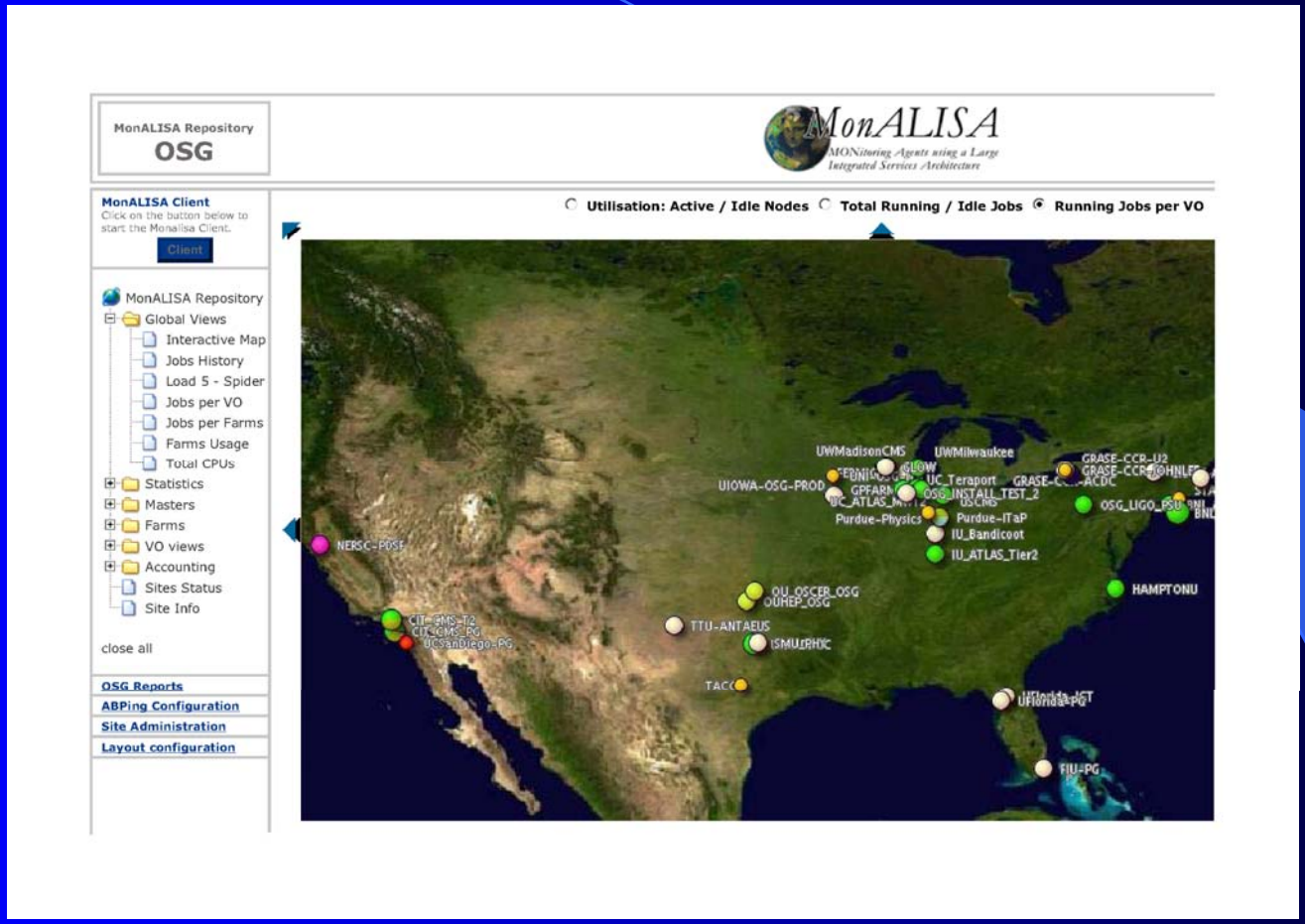


LCG World View



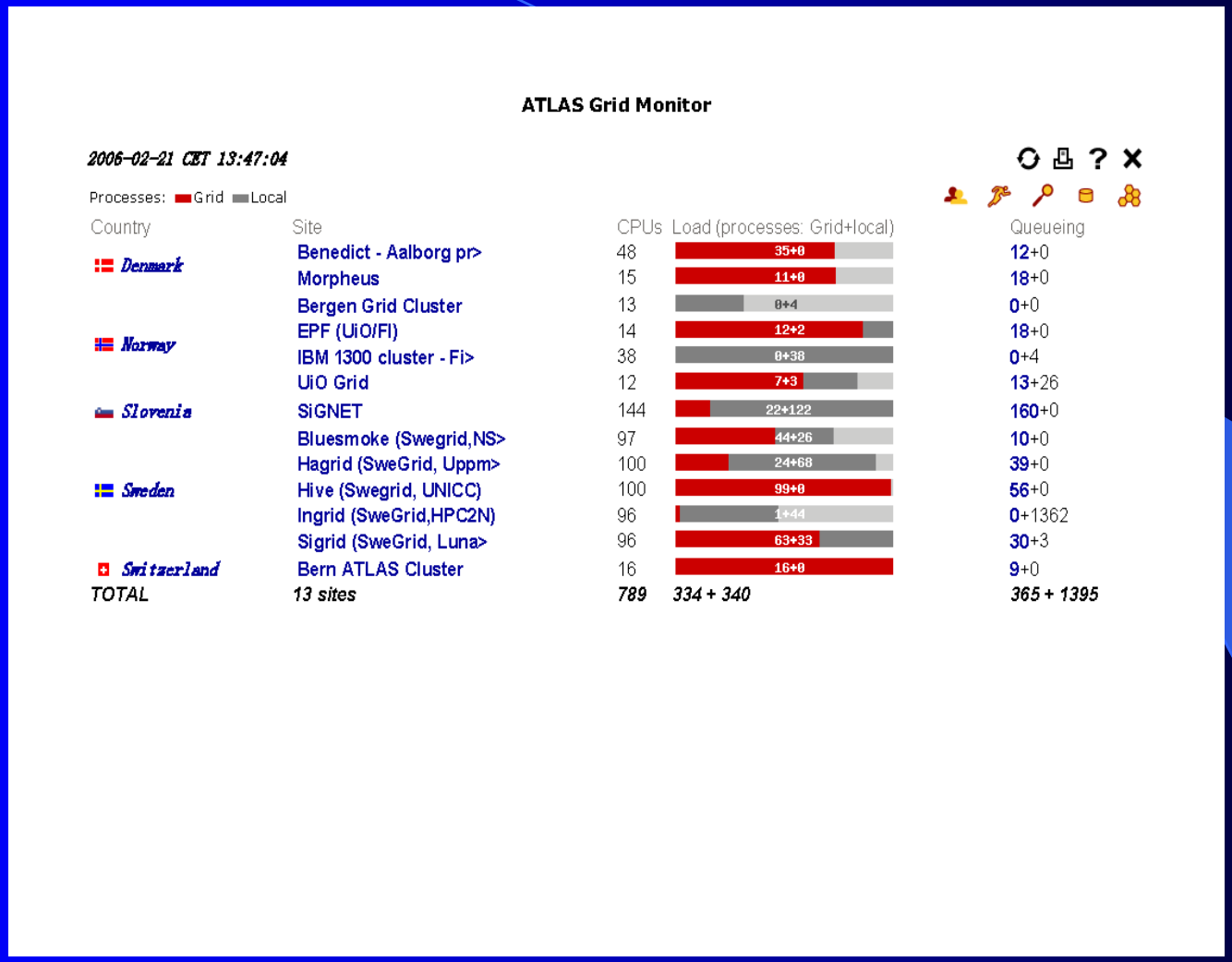
Currently 51 sites with ATLAS SW installed
+9300 CPUs (shared with other VOs)

OSG World View



Currently ~50 sites, ~5000 CPU's
ATLAS dedicated ~1000 CPU's

NorduGrid World View



ATLAS Grid Monitor

2006-02-21 CET 13:47:04

Processes: Grid Local

Country	Site	CPUs	Load (processes: Grid+local)	Queueing
Denmark	Benedict - Aalborg pr>	48	35+0	12+0
	Morpheus	15	11+0	18+0
	Bergen Grid Cluster	13	0+4	0+0
Norway	EPF (UiO/FI)	14	12+2	18+0
	IBM 1300 cluster - Fi>	38	0+38	0+4
	UiO Grid	12	7+3	13+26
Slovenia	SiGNET	144	22+122	160+0
Sweden	Bluesmoke (Swegrid,NS>	97	44+26	10+0
	Hagrid (SweGrid, Uppm>	100	24+68	39+0
	Hive (Swegrid, UNICC)	100	99+0	56+0
	Ingrid (SweGrid,HPC2N)	96	1+44	0+1362
	Sigrid (SweGrid, Luna>	96	63+33	30+3
Switzerland	Bern ATLAS Cluster	16	16+0	9+0
TOTAL	*13 sites*	*789*	*334 + 340*	*365 + 1395*

13 sites, 789 CPUs available for ATLAS now