CMS Productions on the Grid in 2002

Claudio Grandi

(INFN – Bologna)

Outline

- CMS production tools
- CMS/EDG Stress Test
- USCMS IGT production
- Summary

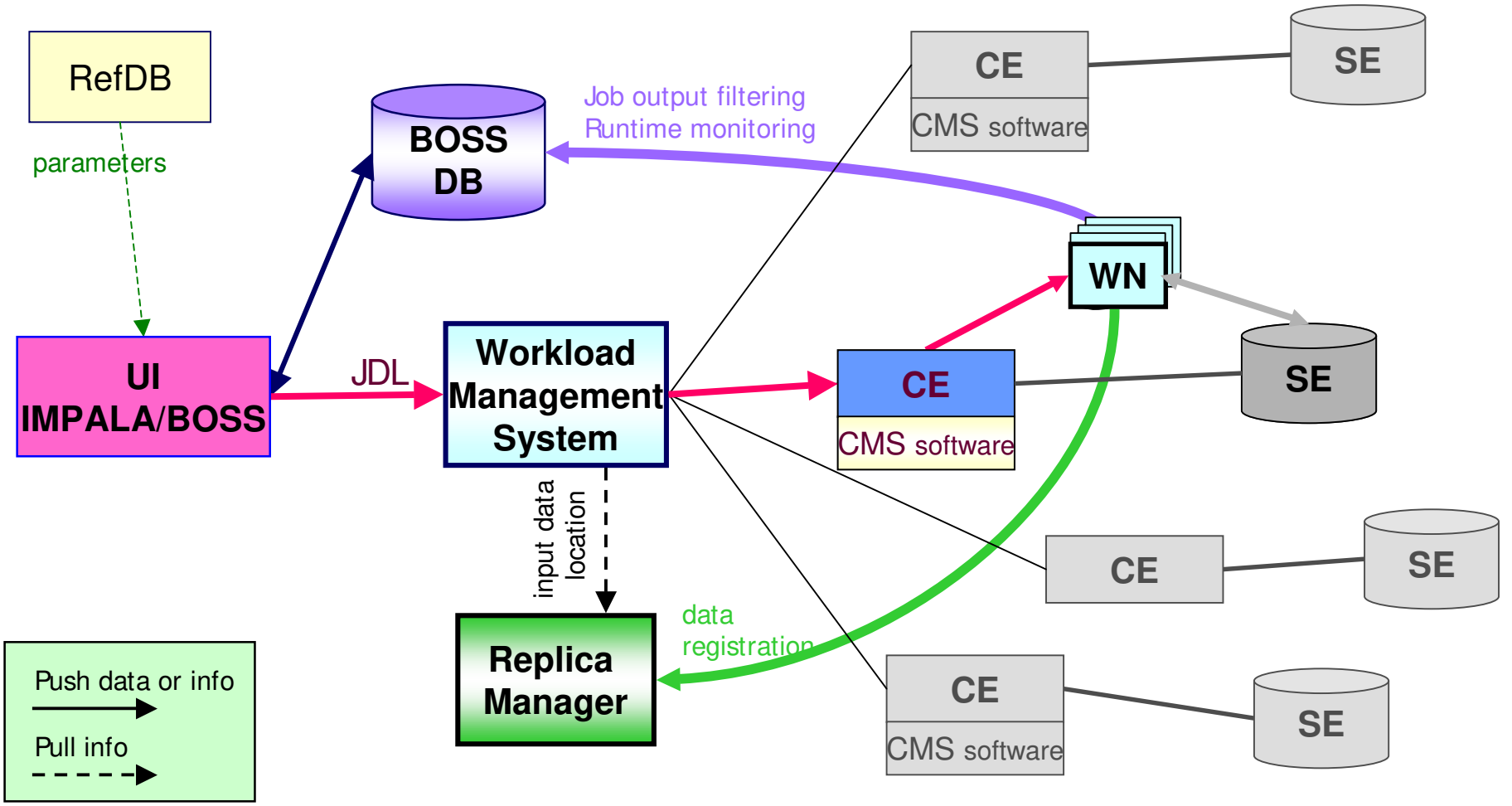
CMS Production Tools

- RefDB
  - Contains production requests with all needed parameters
- IMPALA
  - Accepts a production request
  - Produces the scripts for each single job that needs to be submitted (all steps sequentially)
  - Submits the jobs and tracks the status
- MCRunJob
  - Evolution of IMPALA: modular (plug-in approach)
  - Allows “chaining” of more steps in a single job
- BOSS
  - Real-time job-dependent parameter tracking

CMS/EDG Stress Test

- Running from November 30th to Xmas
- Sites:
  - Official EDG Appl. testbed
  - RC + 4 dedicated RB's
  - CMS sites: Legnaro, Padova, Poly, IC, CNAF Tier-1
  - About 350 CPU available (RH 6.2) but not dedicated!
- Middleware:
  - EDG release 1.4.0, 1.4.1 and 1.4.3 (and added patches!)
- CMS task:
  - FZ files production with IMPALA/BOSS:
    - 2 steps: uses RC for input data location of CMSIM. No Objy (licenses)!
    - Deliverables are "big" FZ files

Stress test set-up



User Interface set-up

- IMPALA
  - Installed on the EDG user interface (UI) machine
  - “JDL” files are produced along with the job scripts
  - Computing resources are matched to the job requirements:
    - Installed CMS software, MaxCPUTime, etc.
    - CE near to the input data or to a given storage element
  - Output files are registered in the replica catalog
- BOSS
  - An “EDG scheduler” is registered
  - Additional info is stored in the BOSS DB:
    - Logical file names of input/output files
    - Name of the SE hosting the output files
    - Outcome of the copy and registration in the RC of files
    - Status of the replication of files

UI's and RB's

- Four UI's controlling the production:
  - Bologna / CNAF
  - Ecole Polytechnique
  - Imperial College
  - Padova
- Several resource brokers (each seeing all resources):
  - CERN (dedicated to CMS) (EP UI)
  - CERN (common to all applications) (backup!)
  - CNAF (common to all applications) (Padova UI)
  - CNAF (dedicated to CMS) (CNAF UI)
  - Imperial College (dedicated to CMS and BABAR) (IC UI)
- Replica catalog at CNAF
- Top MDS at CERN
- II at CERN and CNAF
- VO server at NIKHEF

Sites and resources

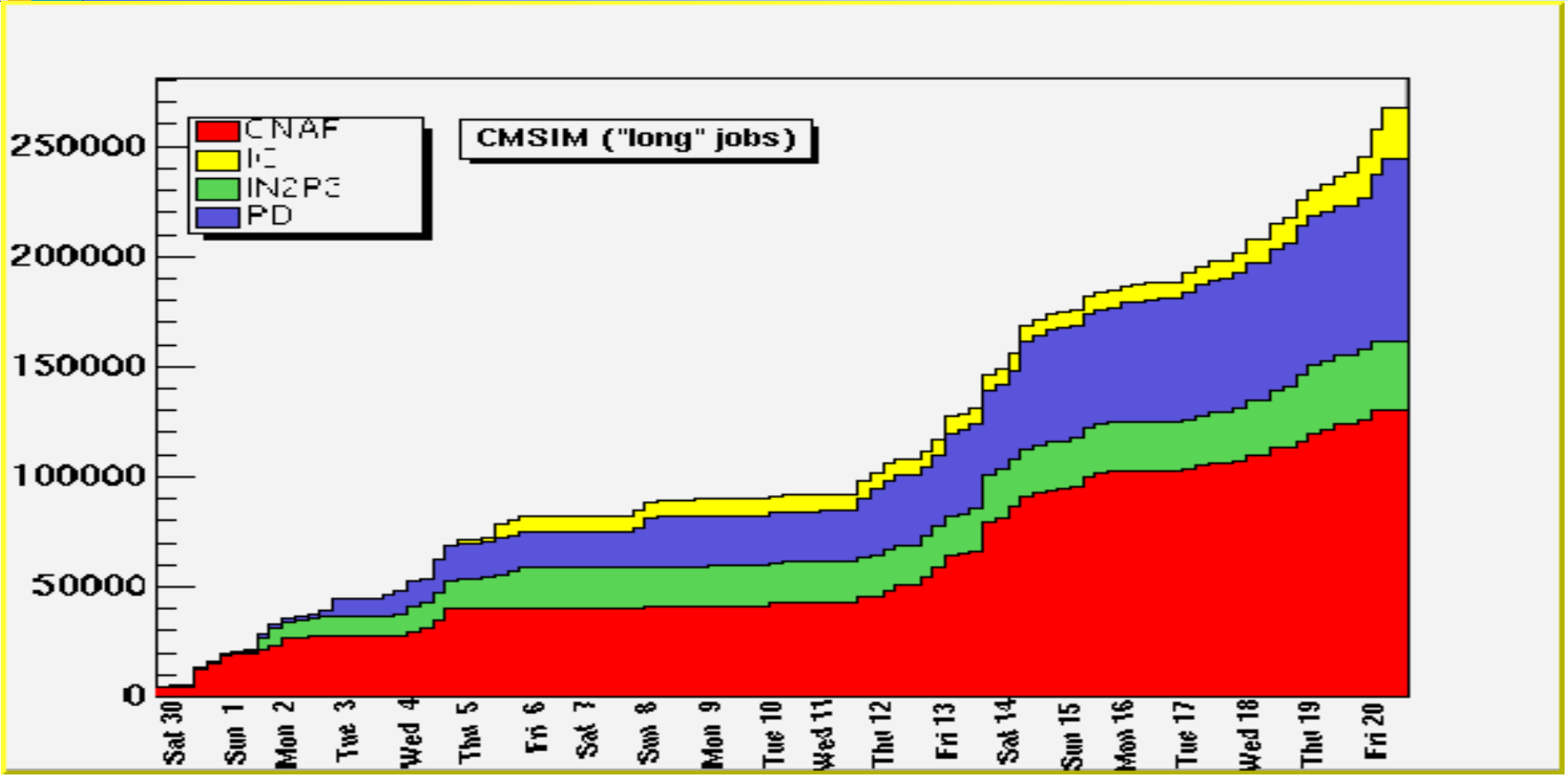
- Standard CE and SE set-up
- CMS software distributed as rpm files

Site	CE	No. of CPU	SE	Disk space (GB)
CERN	lxshare0227	122	lxshare0393 lxshare0384	100 1000(=100×10)
CNAF	testbed008	40+	grid007g	1000
RAL	gppce05	16	gppse05	360
NIKHEF	tbn09	22	tbn03	35
Lyon	ccgridli03 ccgridli08	120 mainly non-grid 400 mainly non-grid	ccgridli07	200
Legnaro	cmsgrid001	50	cmsgrid002	513(+513)
Padova	grid001	12	grid005	680
Ecole Poly.	polgrid1	4	polgrid2	220
Imperial C.	gw39	16	fb00	450
TOTAL		802		5271

Monitoring

- Two main sources of information:
  - EDG monitoring system (MDS based)
    - collected regularly by scripts running as cron jobs and stored for offline analysis
  - BOSS database
    - permanently stored in the MySQL database
- Both sources are processed by **`boss2root`** and the information is put in a Root tree
- Online monitoring with Nagios

Stress Test results



~260K events produced

~7 sec/event average

~2.5 sec/event peak (12-14 dec)

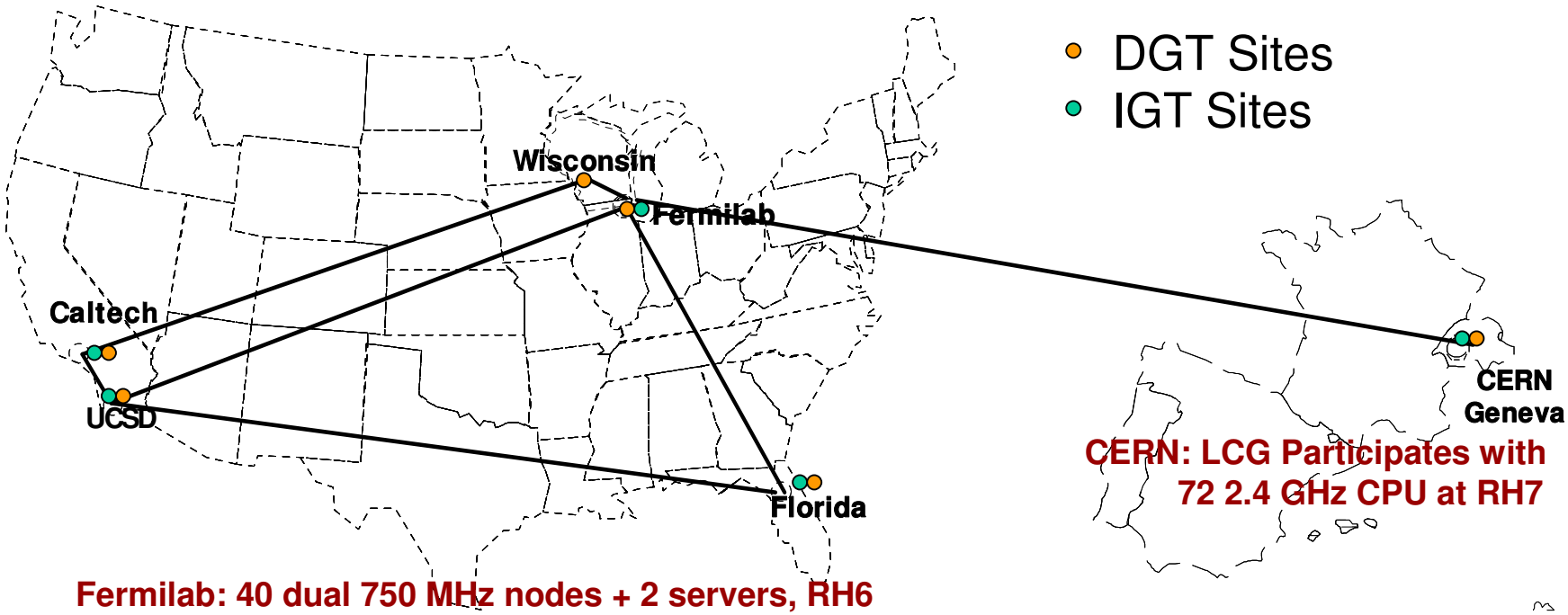
Observations

- Were able to quickly add new sites to provide extra resources
- Fast turnaround in bug fixing and installing new software
- Test was labour intensive (since software was developing and the overall system was fragile)
  - WP1 At the start there were serious problems with long jobs- recently improved
  - WP2 Replication Tools were difficult to use and not reliable, and the performance of the Replica Catalogue was unsatisfactory
  - WP3 The Information System based on MDS performed poorly with increasing query rate

USCMS IGT Production

- Running from October 25th to Xmas
- Sites:
  - Caltech, FNAL, UCSD, Uflorida
  - About 230 CPU (RH 6.1) (+80 CPU at 2.4 GHz, RH 7.X)
- Middleware: VDT 1.1.3
  - Globus, DAGMAN, Condor-G, EDG VO management scripts, Ganglia, MDS, Mona Lisa
- CMS task:
  - Both Ntuple only and FZ files productions with MCRunJob/MOP:
    - All steps in one job. No Objy data distribution
    - Deliverables are “small” ntuples (ntuple-only) or “big” FZ files

The Current IGT - Hardware



CERN: LCG Participates with 72 2.4 GHz CPU at RH7

Fermilab: 40 dual 750 MHz nodes + 2 servers, RH6
Florida: 40 dual 1 GHz nodes + 1 server, RH6
UCSD: 20 dual 800 MHz nodes + 1 server, RH6
New: 20 dual 2.4 GHz nodes + 1 server, RH7
Caltech: 20 dual 800 MHz nodes + 1 server, RH6
New: 20 dual 2.4 GHz nodes + 1 server, RH7
UW Madison: Not a protoype Tier-2 center, support

Total: 240 0.8 equiv. RH6 CPU
152 2.4 GHz RH7 CPU

Software Distribution

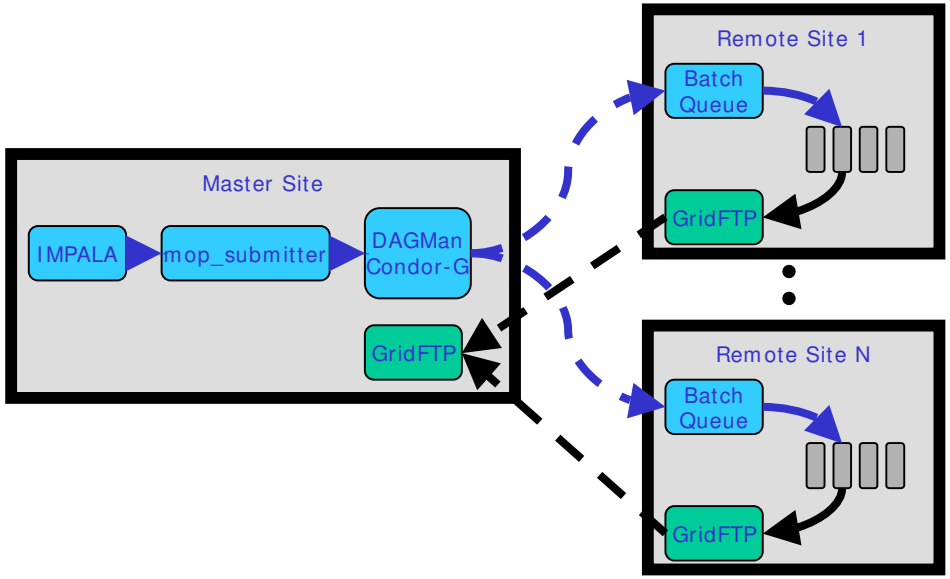
- IGT Software is (mostly) distributed using PACMAN
  - PACMAN caches set up for the USCMS R&D Grid Testbed and the IGT
    - The VDT is distributed this way
    - CMS Executables can in principle be distributed this way also
  - PACMAN keeps track of what is installed at each site
    - HTTP based package retrieval
    - MDS based monitoring system to come
  - Software not yet fully distributed by PACMAN:
    - CMS Executables, GroupMan, Clarens

Virtual Organization

- Virtual Organization Management
  - LDAP server maintained at FNAL
    - Keeps all DNs of USCMS Grid Users
  - GroupMan (from EDG, PPDG)
    - Periodically downloads LDAP information into grid mapfiles
  - Uses DOE Science Grid CA
    - Also accepts Globus CA for EDG compatibility and for certain file transfers
  - Currently using a “group certificate”, but will update to using personal certificates soon

MOP

- MOP is a system for packaging production processing jobs into DAGMAN format
- Mop_submitter wraps Impala jobs in DAG format at the “MOP master” site
- DAGMAN runs DAG jobs through remote sites’ Globus JobManagers through Condor-G
- Results are returned using GridFTP. Though the results are also returned to the MOP master site in the current IGT running, this does not have to be the case.



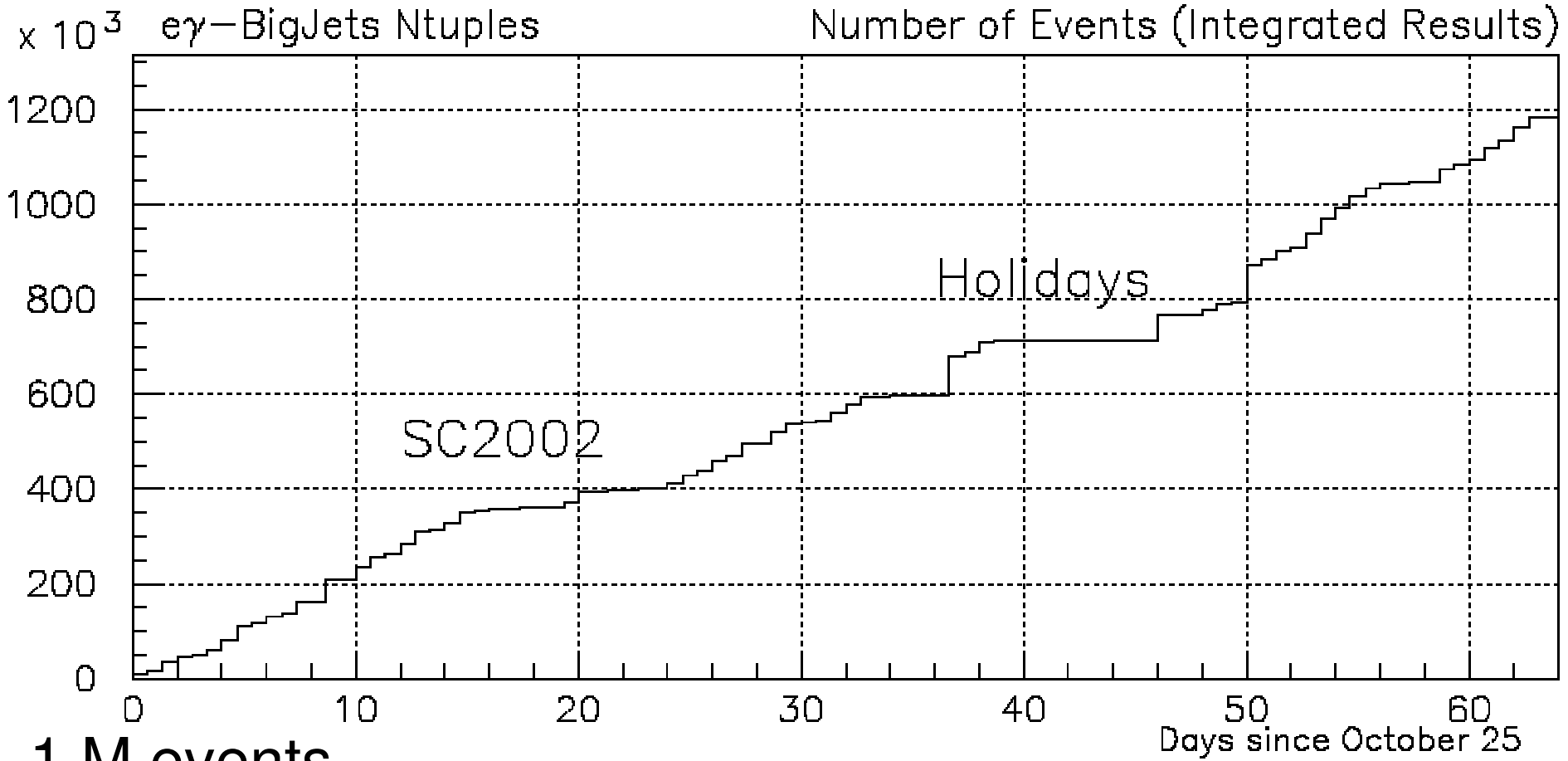
UW Madison is the MOP master for the USCMS Grid Testbed

FNAL is the MOP master for the IGT and the Production Grid

Monitoring

- MonaLisa is used as the primary Grid-wide monitor.
  - Dynamic discovery of monitoring targets and schema
  - Implemented with Java/Jini with SNMP local monitors
  - Interfaces to/from MDS implemented at FNAL and Florida
    - Important since most schedulers will talk to MDS
  - Interfaces with local monitoring systems
    - Ganglia interface used at Fermilab
- IGT Monitoring information was not used for automatic scheduling
  - For human benefit only so far

IGT Production Results



1 M events

4.7 sec/event average

2.5 sec/event peak (14-20 dec 2002)

Observations

- Failure semantics currently lean heavily towards automatic resubmission of failed jobs
  - Sometimes failures are not recognized right away
  - Need better system for spotting chronic problems
- The “wrong” people were often the ones to start troubleshooting problems
  - At one point, an application problem was misdiagnosed early as a GASS cache issue
  - Once the application expert looked into it, the problem was solved in 90 minutes
    - But middleware experts had already spent days :-(
- Little manpower needed (2.65 FTE during initial phase, 1.1 during smooth running, expected to be less than 1 with regular shift procedures)

Summary

- Two “official” CMS productions on the grid in 2002
  - CMS-EDG Stress Test on EDG testbed + CMS sites
    - ~260K events CMKIN and CMSIM steps
    - Top-down approach: more functionality but less robust, large manpower needed
  - USCMS IGT Production in the US
    - 1M events Ntuple-only (full chain in single job)
    - 500K up to CMSIM (two steps in single job)
    - Bottom-up approach: less functionality but more stable, little manpower needed