



Running CMS software on Grid Testbeds



Paolo Capiluppi
Dept. of Physics and INFN
Bologna



On behalf of the CMS Collaboration

Authors:

M. Afaq, A. Arbree, P. Avery, I. Augustin, S. Aziz, L. Bauerdick, M. Biasotto, J-J. Blaising, D. Bonacorsi, D. Bourilkov, J. Branson, J. Bunn, S. Burke, P. Capiluppi, R. Cavanaugh, C. Charlot, D. Colling, M. Corvo, P. Couvares, M. Ernst, B. MacEvoy, A. Fanfani, S. Fantinel, F. Fanzago, I. Fisk, G. Graham, C. Grandi, S. Iqbal, J. Kaiser, E. Laure, V. Lefebure, I. Legrand, E. Leonardi, J. Letts, M. Livny, C. Loomis, O. Maroney, H. Newman, F. Prelz, N. Ratnikova, M. Reale, J. Rodriguez, A. Roy, A. Sciaba', M. Schulz, I. Semeniuk, M. Sgaravatto, S. Singh, A. De Smet, C. Steenberg, H. Stockinger, Suchindra, T. Tannenbaum, H. Tallini, J. Templon, G. Tortone, M. Verlatto, H. Wenzel, Y. Wu

Acknowledgments:

Thanks to the EU and National funding agencies for their support of this work.



Outline



- ✓ **CMS Grid Computing**
- ✓ **CMS jobs for “Production”**
- ✓ **IGT and EDG CMS Testbeds**
- ✓ **Results obtained**
- ✓ **Conclusions**



CMS Grid Computing



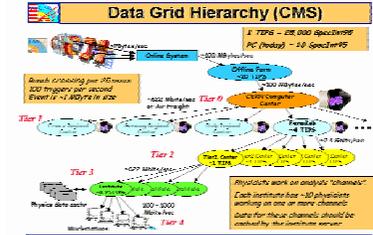
→ Large scale distributed Computing and Data Access

- Must handle PetaBytes per year
- Tens of thousands of CPUs
- Tens of thousands of jobs



→ Evolving heterogeneity of resources (hardware, software, architecture and Personnel)

- Must cope with Hierarchies and sharing among other Applications: a coordination is needed
- Must foster local capabilities
- Must allow for dynamical movement of responsibilities and target specific problems



→ Test now the functionalities to be adopted tomorrow (via CMS Data Challenges)

- Current Grid Testbeds and current CMS software
- Provide feedback to the Architecture and Implementation of Grid Middleware and CMS Software

→ Use the current implementations of many Projects: European DataGrid, GriPhyn, PPDG, DataTag, iVDGL, Trillium, National Grids, etc.(including GLUE and LCG)



CMS Jobs and Tools used for the Tests



CMS official jobs for “Production” of results used in Physics studies: Real-life testing

→ CMS Jobs:

- **CMKIN**: MC Generation of the proton-proton interaction for a physics channel (dataset)
- **CMSIM**: Detailed simulation of the CMS detector, processing the data produced during the CMKIN step
- **ORCA**: reproduction of detector signals (Digis) and reconstruction of physical information producing final analysis Ntuples
- **Ntuple-only**: The full chain in a single step (single composite job)

→ CMS Tools for “Production”

- **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
- **MCRunjobs**: Modular (plug-in approach) metadata-based workflow planner
 - Allows “chaining” of more steps in a single job
- **BOSS**: Real-time job-dependent parameter tracking

| Eye- BigJets Dataset | Size/event | Time*/event |
|----------------------------|------------------------|---------------|
| CMKIN | ~ 0.05 MB (Ntuple) | ~ 0.4-0.5 sec |
| CMSIM | ~ 1.8 MB (Fz file) | ~ 6 min |
| ORCA | ~ 1.5 MB (Objy DB) | ~ 18 sec |
| Ntuple | ~ 0.001 MB (Ntuple) | ~ 380 sec |

* PIII 1GHz

A “Complex” Process



IGT and EDG Testbeds



→ **IGT and EDG Testbeds are both part of CMS program to exploit Grid functionalities.**

- **IGT Testbed is: Integration Grid Testbed in US (a US CMS initiative)**
- **EDG Testbed is: European DataGrid in EU (a EU Science shared initiative)**
- **Similar dimensions and available resources**
- **Complementary tests and information (for CMS Experiment and for Grid Projects)**

→ **CMS IGT**

- **Running from October 25th to Xmas 2002**
- **Both Ntuple-only and FZ files productions with MCRunjob/MOP (Single step)**

→ **CMS EDG**

- **Running from November 30th to Xmas 2002**
- **FZ files production with IMPALA/BOSS (Two steps)**



CMS/EDG Strategy



→ EDG Stress Test Goals were:

- **Verification of the portability of the CMS Production environment into a grid environment;**
- **Verification of the robustness of the European DataGrid middleware in a production environment;**
- **Production of data for the Physics studies of CMS, with an ambitious goal of ~ 1 million simulated events in a 5 weeks time.**

→ Use as much as possible the High-level Grid functionalities provided by EDG:

- **Workload Management System (Resource Broker),**
- **Data Management (Replica Manager and Replica Catalog),**
- **MDS (Information Indexes),**
- **Virtual Organization Management, etc.**

→ A “Top-down” Grid approach.

→ Interface (modify) the CMS Production Tools to the Grid provided access methods



CMS/IGT Strategy



→ IGT main goals were:

- Provide a large Testbed of CMS-US Tier1 and Tier2, stable and robust;
- Produce a large number of CMS usable events;
- Demonstrate the reduction of Personnel in comparison to “traditional” CMS Production;
- Test the scalability of underlying Condor/Globus middleware

→ Use as much as possible the low-level Grid functionalities provided by basic components:

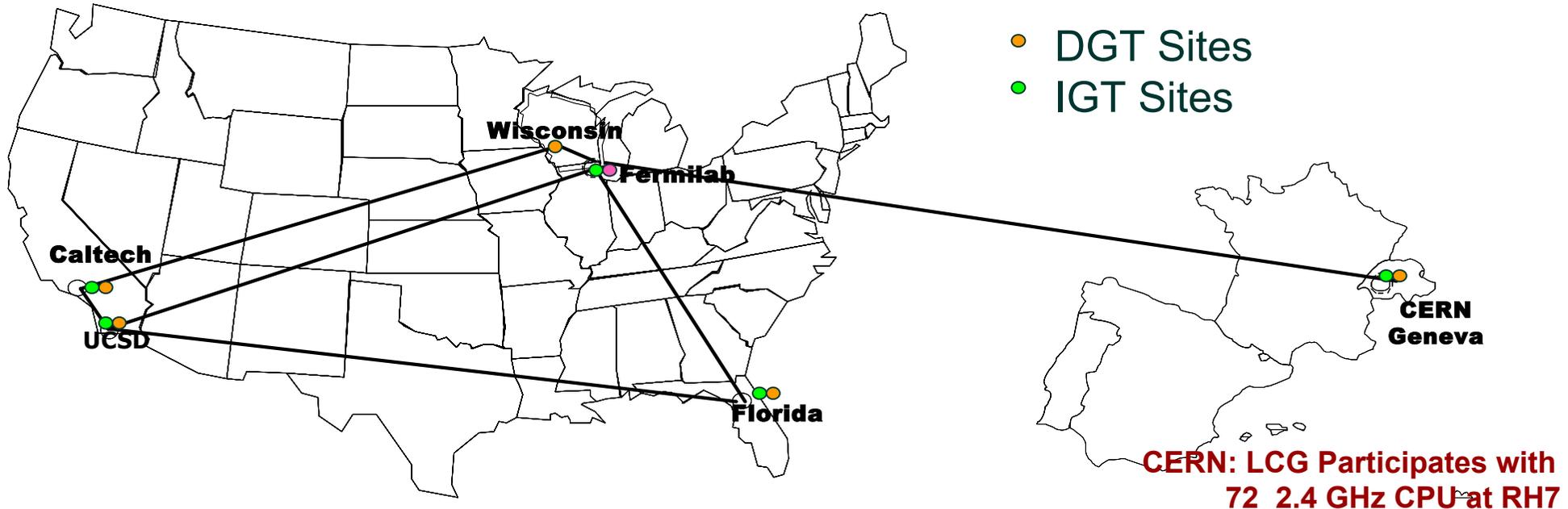
- Globus,
- Condor,
- DAGMan,
- Basic VO, etc.

→ A “Bottom-up” Grid approach

→ Adapt (integrate) the CMS Production tools to access the Grid basic components



The IGT Hardware Resources



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 prototype Tier-2 center, support

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



IGT Middleware and Software



→ Middleware was: Virtual Data Toolkit (VDT) 1.1.3

- **Virtual Data Client**
 - Globus Toolkit 2.0 (with improved GASS cache)
 - DAGMAN: A package that models production jobs as Directed Acyclic Graphs
 - Condor-G 6.4.3: A backend that allows DAGMAN to manage jobs on Globus Job Managers
- **Virtual Data Server**
 - (the above, plus:)
 - mkgridmap: A tool to help manage the gridmap authorization files
 - GDMP 3.0.7: The EDG WP2 replica manager

→ Software distribution (mostly) via PACMAN

- PACMAN keeps track of what is installed at each site

→ Virtual Organization Management

- GroupMan (from EDG, PPDG)
- Uses DOE Science Grid CA

→ Monitoring via MonaLisa:

- Dynamic discovery of monitoring targets and schema
- Interfaces to/from MDS implemented at FNAL and Florida
- Interfaces with local monitoring systems like Ganglia at Fermilab



CMS/IGT MOP Tool



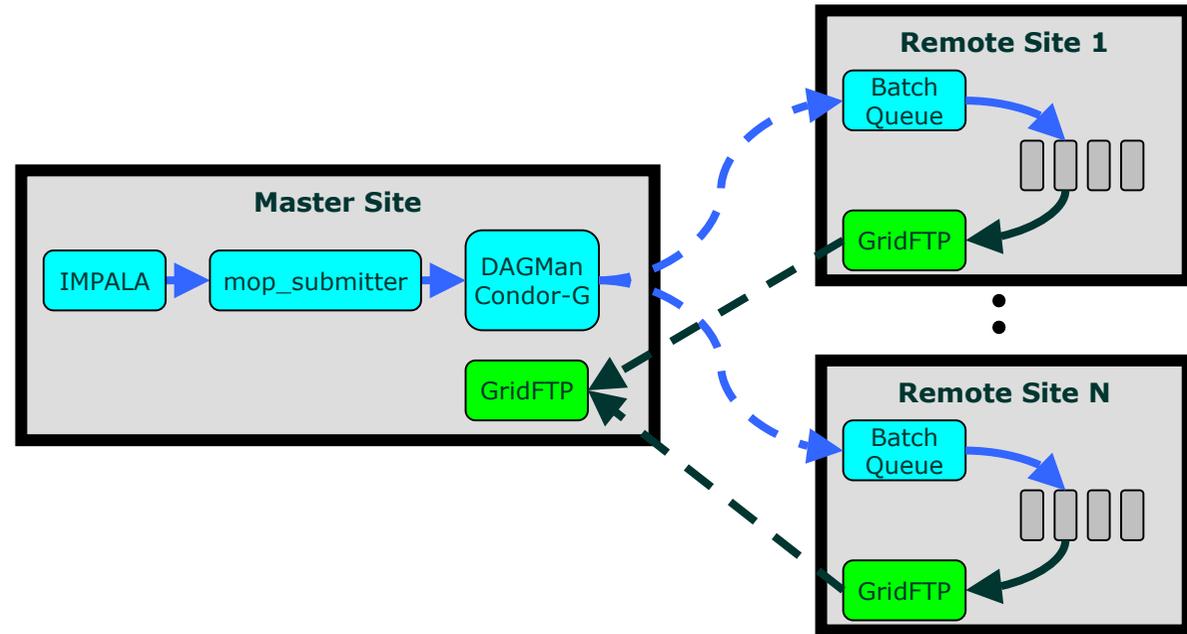
MOP is a system for packaging production processing jobs into DAGMAN format

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



EDG hardware resources



| Site | Number of CPUs | Disk Space GB | Availability of MSS |
|----------------------------------|------------------|-----------------------|---------------------|
| CERN (CH) | 122 | 1000* (+100) | yes |
| CNAF (IT) | 40* | 1000* | |
| RAL (UK) | 16 | 360 | |
| Lyon (FR) | 120 (400) | 200 | yes |
| NIKEF (NL) | 22 | 35 | |
| Legnaro (IT)* | 50 | 1000* | |
| Ecole Polytechnique (FR)* | 4 | 220 | |
| Imperial College (UK)* | 16 | 450 | |
| Padova (IT)* | 12 | 680 | |
| Totals | 402 (400) | 3000* + (2245) | |

*Dedicated to CMS Stress Test



CMS/EDG Middleware and Software



→ Middleware was: EDG from version 1.3.4 to version 1.4.3

- Resource Broker server
- Replica Manager and Replica Catalog Servers
- MDS and Information Indexes Servers
- Computing Elements (CEs) and Storage Elements (SEs)
- User Interfaces (UIs)
- Virtual Organization Management Servers (VO) and Clients
- EDG Monitoring
- Etc.

→ Software distribution was via RPMs within LCFG

→ Monitoring was done through:

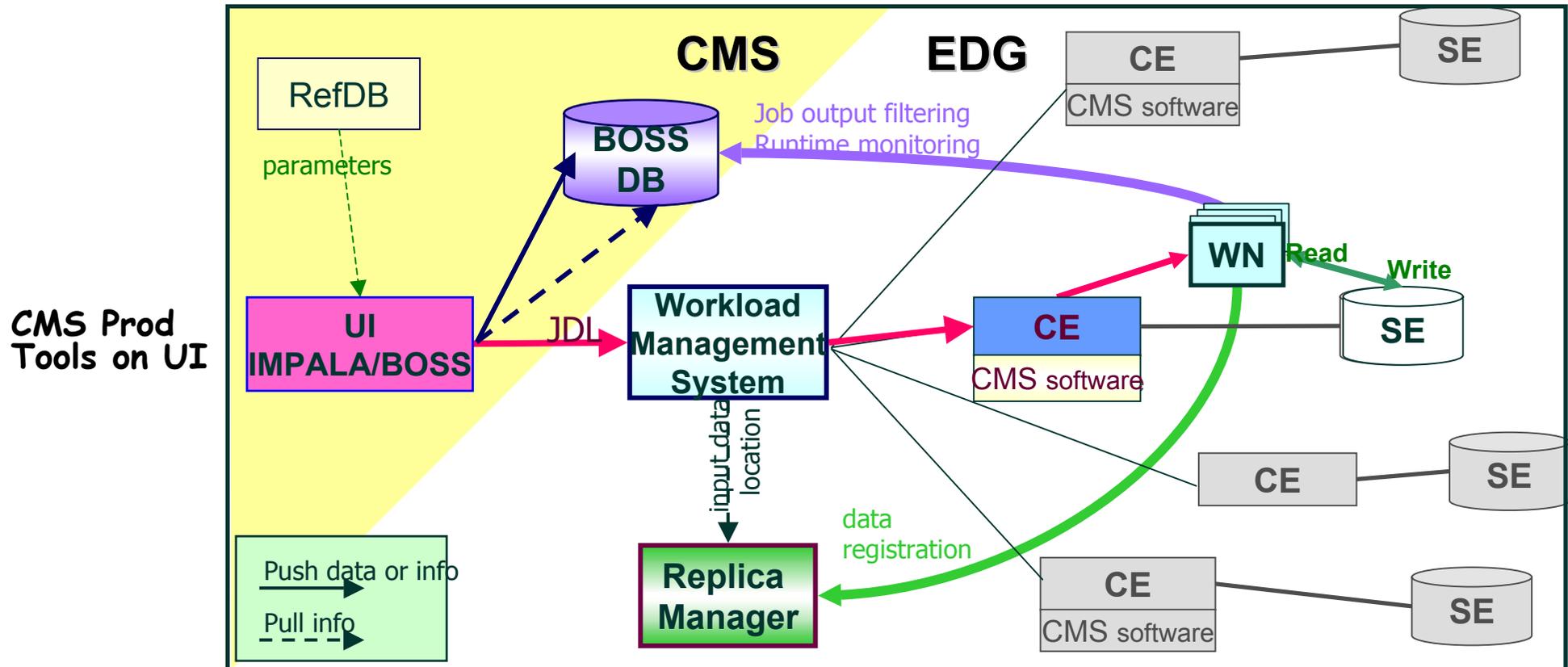
- EDG monitoring system (MDS based)
 - collected regularly by scripts running as cron jobs and stored for offline analysis
- BOSS database
 - permanently stored in the MySQL database
- Both sources are processed by `boss2root` and the information is put in a Root tree
- Online monitoring with Nagios



CMS production components interfaced to EDG

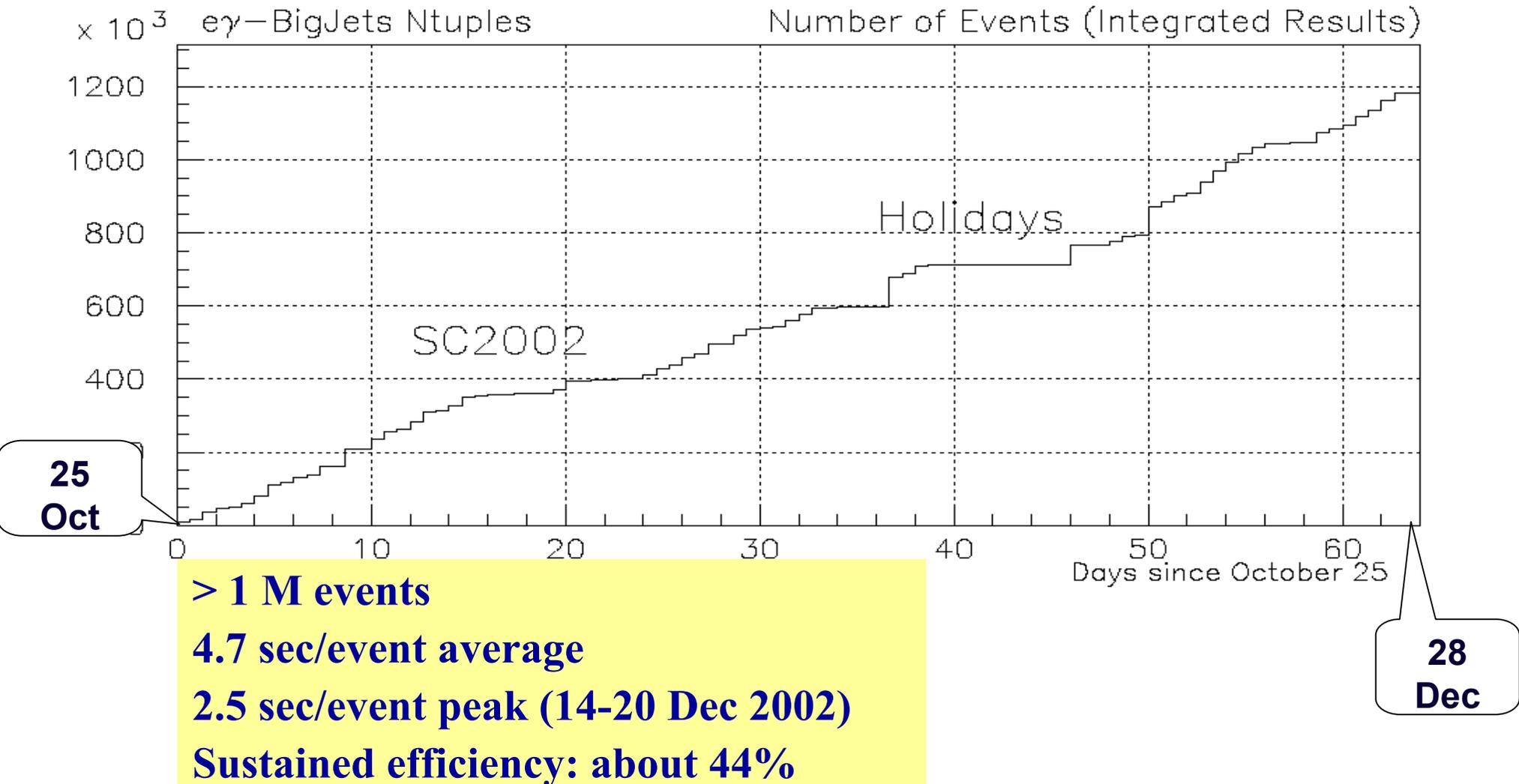


- Four submitting UIs: Bologna/CNAF (IT), Ecole Polytechnique (FR), Imperial College (UK), Padova/INFN (IT)
- Several Resource Brokers (WMS), CMS-dedicated and shared with other Applications: one RB for each CMS UI + "backup"
- Replica Catalog at CNAF, MDS (and II) at CERN and CNAF, VO server at NIKHEF





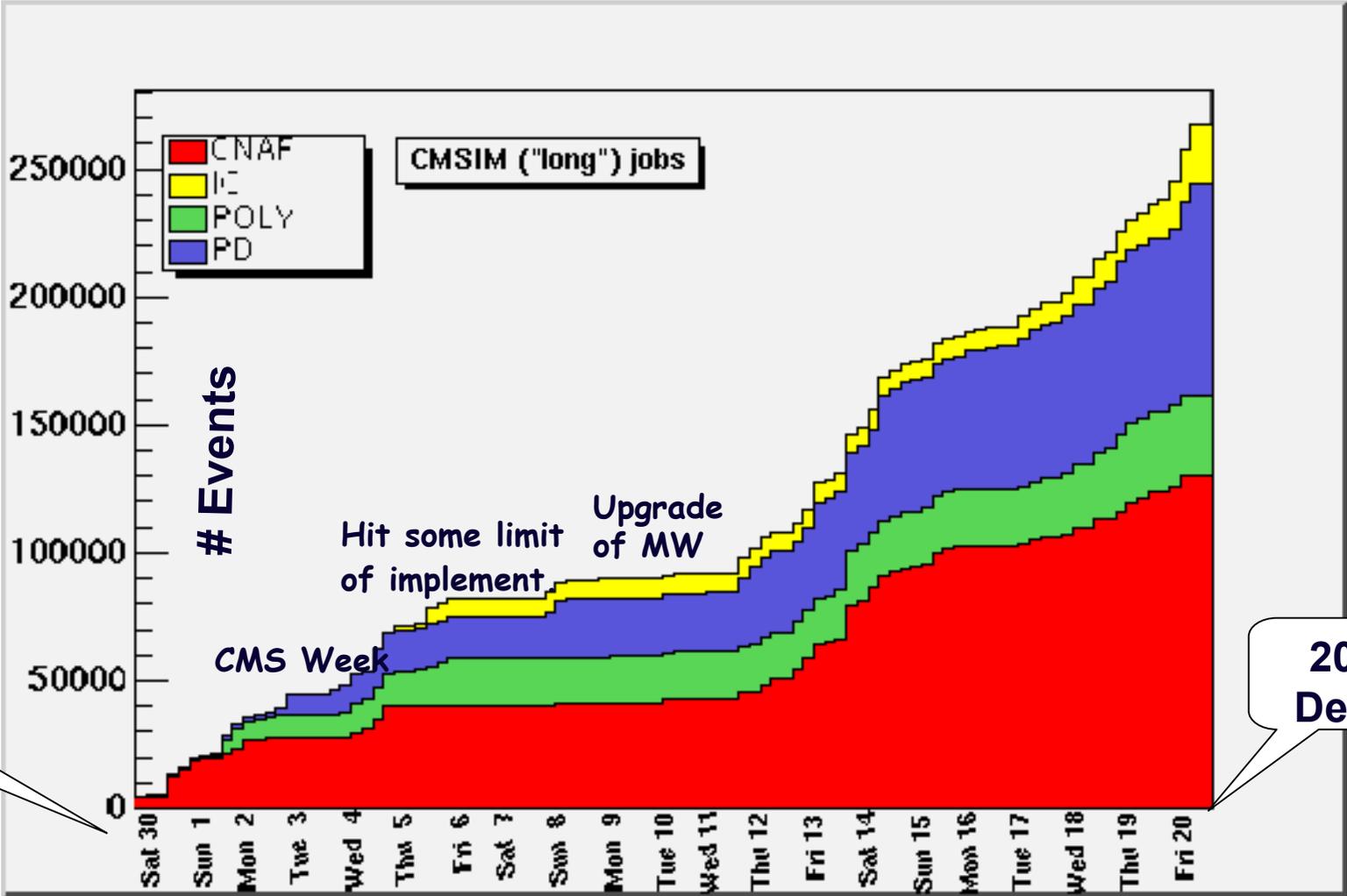
US-CMS IGT Production





CMS/EDG Production

~260K events produced
 ~7 sec/event average
 ~2.5 sec/event peak (12-14 Dec)





CMS/EDG Summary of Stress Test



Short jobs

After Stress
Test – Jan 03

| CMKIN jobs | | | |
|-----------------------------|---------------------------------------|----------------------|---|
| Status | EDG Stress Test evaluation | EDG ver 1.4.3 | "CMS" Stress Test evaluation |
| Finished Correctly | 5518 | 1014 | 4742 |
| Crashed or bad status | 818 | 57 | 958 |
| Total number of jobs | 6336 | 1071 | 5700 |
| Efficiency | 0.87 | 0.95 | 0.83 |

Long jobs

After Stress
Test – Jan 03

| CMSIM jobs | | | |
|-----------------------------|---------------------------------------|----------------------|---|
| Status | EDG Stress Test evaluation | EDG ver 1.4.3 | "CMS" Stress Test evaluation |
| Finished Correctly | 1678 | 653 | 2147 |
| Crashed or bad status | 2662 | 264 | 935 |
| Total number of jobs | 4340 | 917 | 3082 |
| Efficiency | 0.39 | 0.71 | 0.70 |

Total EDG Stress Test jobs = 10676 , successful = 7196 , failed = 3480



EDG reasons of failure (categories)

Short jobs

| CMKIN jobs | |
|---|---------------|
| Status | Totals |
| Crashed or bad status | 818 |
| Reasons of Failure for Crashed jobs | |
| No matching resource found | 509 |
| Generic Failure: MyProxyServer not found in JDL expr. | 102 |
| Running forever | 74 |
| Failure while executing job wrapper | 37 |
| Other failures | 96 |

Long jobs

| CMSIM jobs | |
|---|---------------|
| Status | Totals |
| Crashed or bad status | 2662 |
| Reasons of Failure for Crashed jobs | |
| Failure while executing job wrapper | 1476 |
| No matching resource found | 722 |
| Globus Failure: Globus down/Submit to globus failed | 144 |
| Running forever | 116 |
| Globus Failure | 90 |
| Other failures | 114 |



Conclusions



→ Two different, complementary approaches

- **CMS-EDG Stress Test on EDG testbed + CMS sites**
 - ~260K events CMKIN and CMSIM steps (~10.000 jobs in 3 weeks)
 - Identification of Bottlenecks and fast fixes implementation (High dynamicity)
 - Measures of (in)efficiencies
 - Able to quickly add new sites to provide extra resources
 - **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)
 - Identification of areas of more work (e.g. automatic resubmission, error reporting, ...)
 - **Bottom-up approach: less functionality but more stable, little manpower needed**
- **Comparison to CMS Spring 2002 “manual” Production**
 - Quite different processes simulated, with different environment (Pile-up, resources)
 - However the CPU occupancy (10-40%) and the “sec/event” (1.2-1.4) are not too far

→ Evolution of Testbeds:

- **EDG -> EDG 2 (2Q03) -> LCG-1 (3Q03)**
- **IGT -> Production Grid Testbed (1Q03) -> LCG-1 (3Q03)**