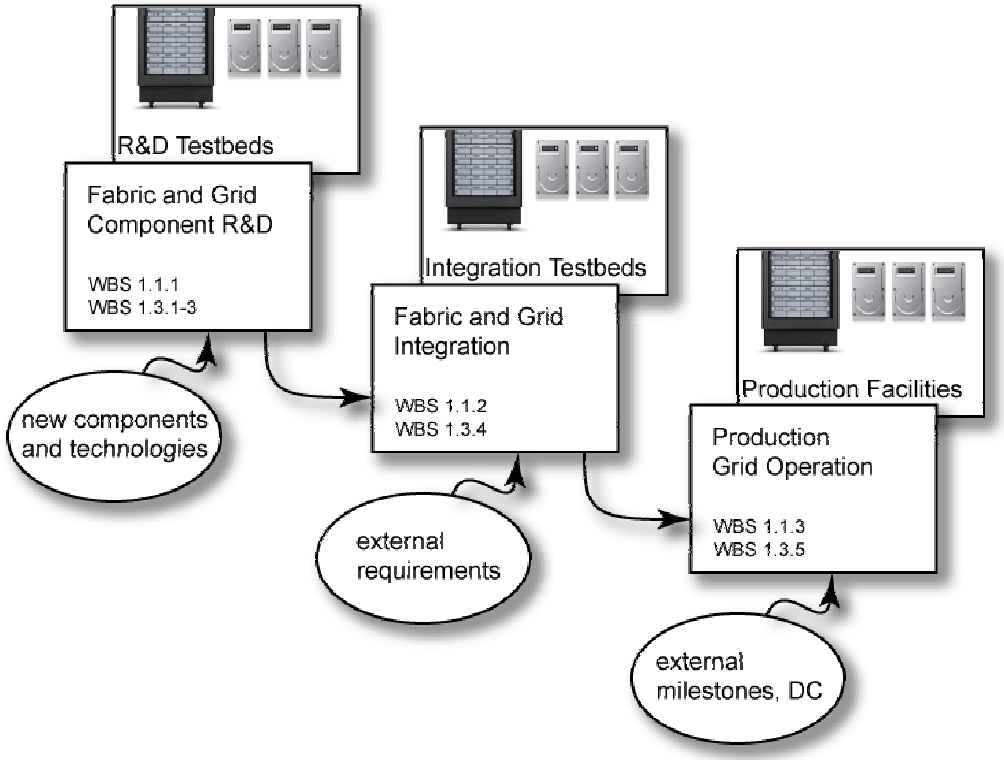
Introduction

- US CMS is positioning itself to be able to learn, prototype and develop while providing a production environment to cater to CMS, US CMS and LCG demands
  - R&D
  - Integration
  - Production



- 3 phased approach is not a new idea, but is mission critical!
  - Development Grid Testbed (DGT), Integration Grid Testbed (IGT), and Production Grid

The Integration Grid Testbed (IGT)

Resource Allocations (1 GHz equiv. CPU) in 2002/2003 for IGT and Production Grid. (R&D Grid not included.)

	2002(IGT)	2002(PG)	2003(New)	2003(IGT)	2003(PG)
FNAL	60	0	260	10	310
Florida	80	0	175	5	250
Caltech	120	0	88	5	203
UCSD	128	0	88	5	211
Total	388	0	611	25	974

New resources for Tier-2 are from iVDGL.

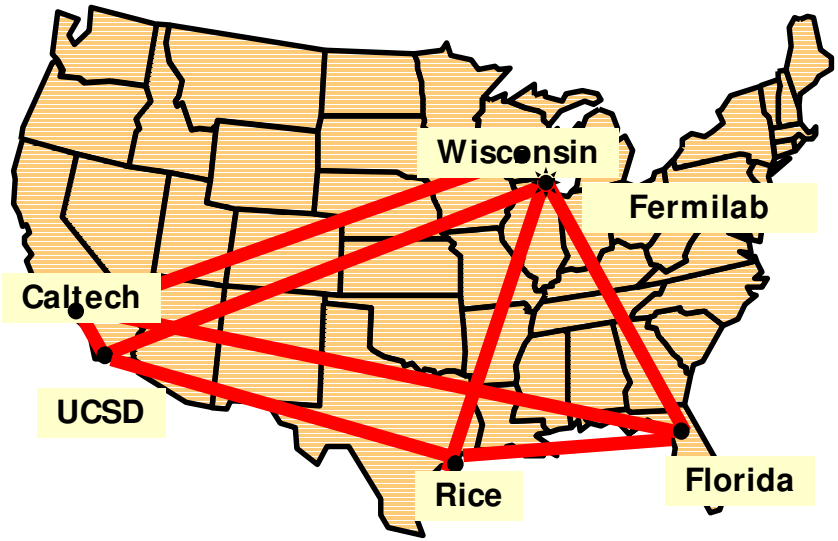
The Integration Grid Testbed (IGT)

- Controlled environment on which to test middleware and grid enabled software in preparation for release
  - Currently, the IGT uses USCMS Tier-1/Tier-2 designated resources.
  - Soon (by the end of March), most of the IGT resources will be turned over to a production grid.
    - IGT will retain a small number of resources deemed necessary to do integration testing
    - Virtual Organization management should be flexible enough to allow PG to "loan" resources to IGT when needed for scalability tests
- In the meantime, the IGT has been commissioned with real production assignments
  - Testing Grid operations, troubleshooting procedures, and scalability issues

US-CMS Development Grid Testbed

- Fermilab
  - 1+5 PIII dual 0.700 GHz processor machines
- Caltech
  - 1+3 AMD dual 1.6 GHz processor machines
- San Diego
  - 1+3 PIV single 1.7 GHz processor machines
- Florida
  - 1+5 PIII dual 1 GHz processor machines
- Rice
  - 1+? machines
- Wisconsin
  - 5 PIII single 1 GHz processor machines

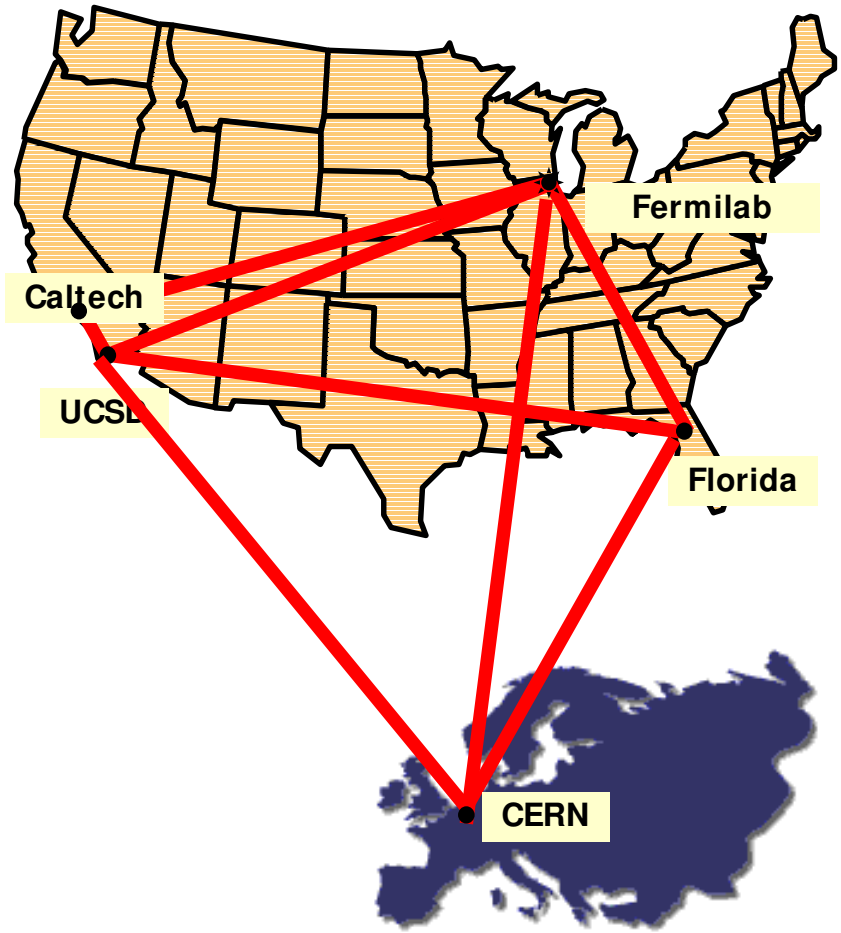
- Total:
  - ~40 1 GHz dedicated processors



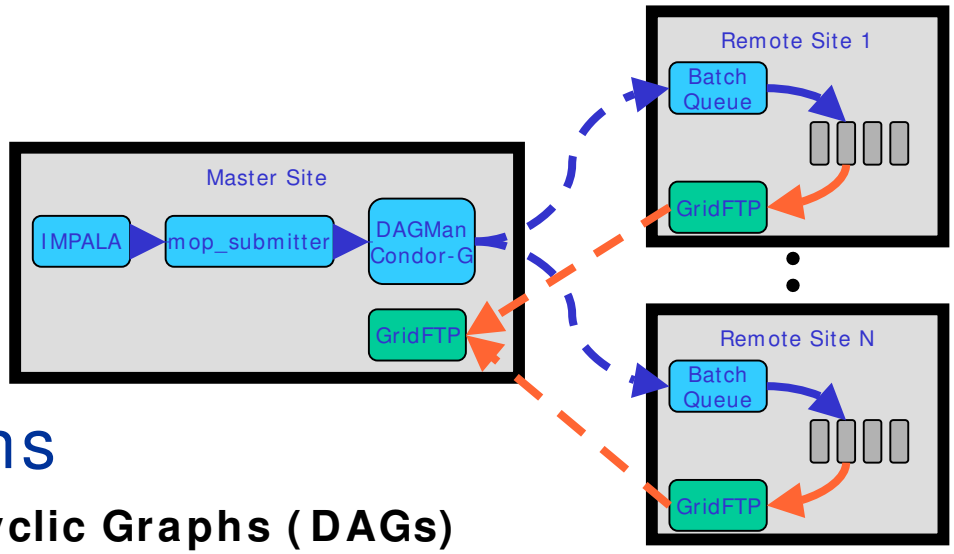
US-CMS Integration Grid Testbed

- Fermilab (Tier1)
  - 40 dual 0.750 GHz processor machines
- Caltech (Tier2)
  - 20 dual 0.800 GHz processor machines
  - 20 dual 2.4 GHz processor machines
- San Diego (Tier2)
  - 20 dual 0.800 GHz processor machines
  - 20 dual 2.4 GHz processor machines
- Florida (Tier2)
  - 40 dual 1 GHz processor machines
- CERN (LCG Tier0 site)
  - 36 dual 2.4 GHz processor machines

- Total:
  - 240 0.85 GHz processors: Red Hat 6
  - 152 2.4 GHz processors: Red Hat 7

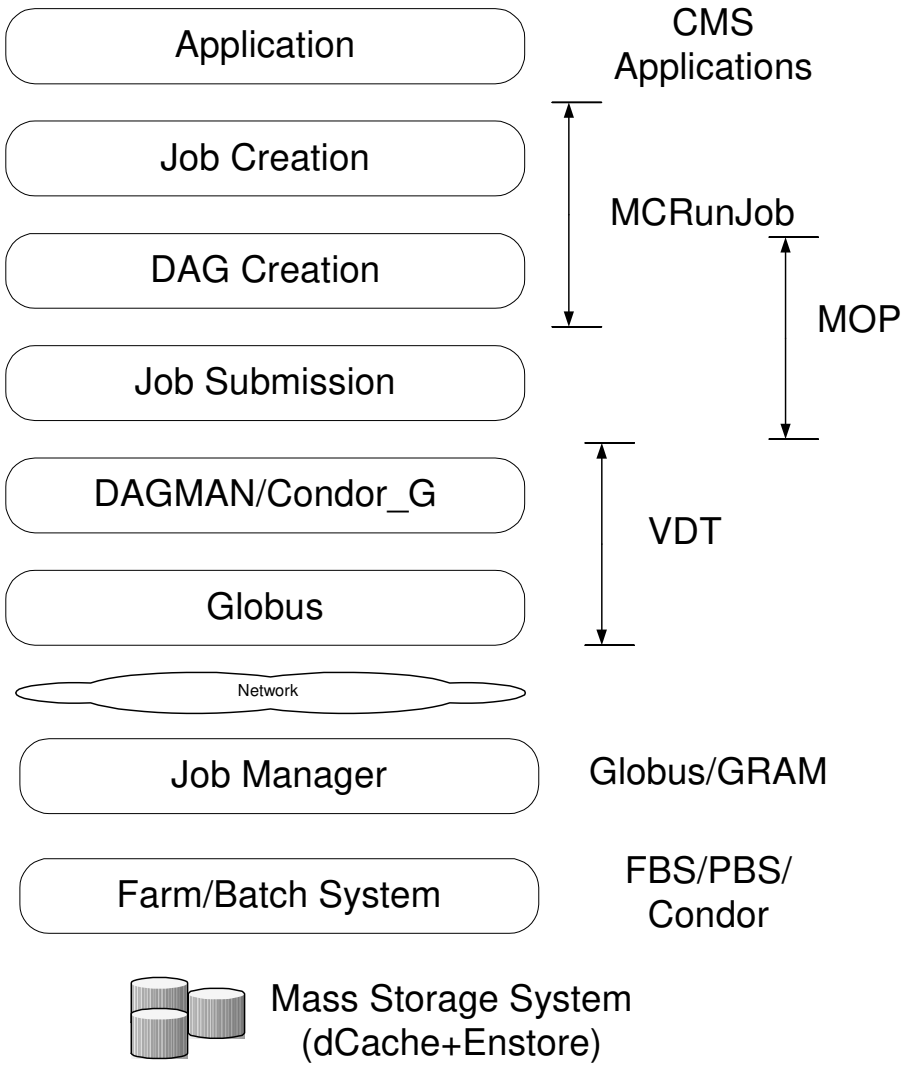


Special IGT Software: MOP



- MOP is a system for packaging production processing jobs into DAGMan job descriptions
  - **DAGMan jobs are Directed Acyclic Graphs (DAGs)**
  - **MOP uses the following DAG Nodes for each job:**
    - Stage-in: Stages in needed application files, scripts, data from the submit host
    - Run: The application(s) run on the remote host
    - Stage-out: The produced data is staged out from the execution site back to the submit host
    - Clean-up: Temporary areas on the remote site are cleansed
    - Publish: Data is published to a GDMP replica catalogue after it is returned

The CMS IGT "Stack"



The CMS IGT "stack" comprises nine layers. The Application layer contains only CMS executables. The Job Creation layer comprises CMS provided tools MCRunJob and Impala. Neither MCRunJob nor Impala are specifically "grid aware." Then there is a DAG Creation layer and a Job Submission layer. Both functionalities are provided by MOP. Jobs are submitted to DAGMAN which, through Condor-G, manages jobs run on remote Globus Job Managers. Finally, there is a local Farm or Batch System used by Globus GRAM to manage jobs. In the case of the IGT, the local Batch manager was always FBSNG or Condor. Scheduling and Integrated monitoring are not present.

Completed CMS Production on the IGT

- Assignment to produce 1 Million “eGamma Bigjets” events by Christmas 2002, all steps.
  - About 500 sec per event on a 750 MHz processor
    - Dominated by the cmsim step
  - Can run only on RH6 machines because of Objectivity licensing issues
- Assignment to produce 500K additional events, cmsim step only
  - This runs on the USCMS Tier-2 hardware that is currently RH7

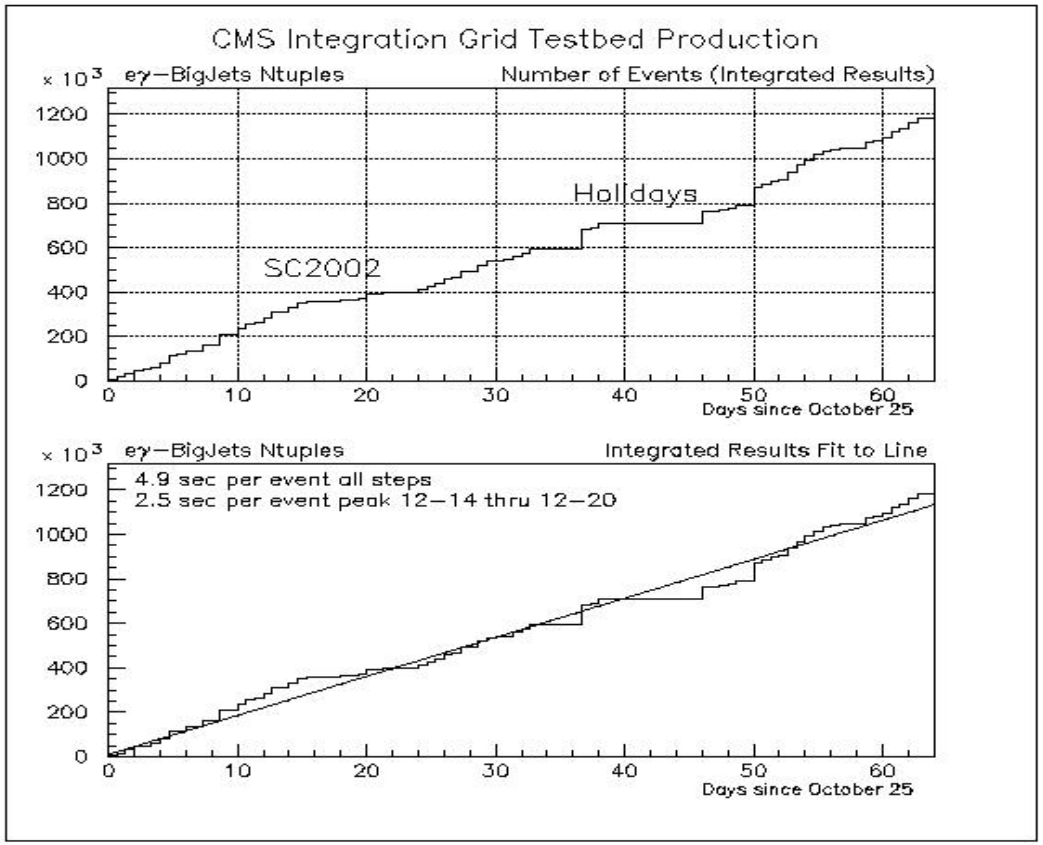
IGT Results

Time to process 1 event: 500 sec @ 750 MHz

Speedup: Avg factor of 100 speedup during current run

Resources: Approximately 230 CPU @750 MHz equiv.

Sustained efficiency: about 43.5%



1M fully simulated and reconstructed events!

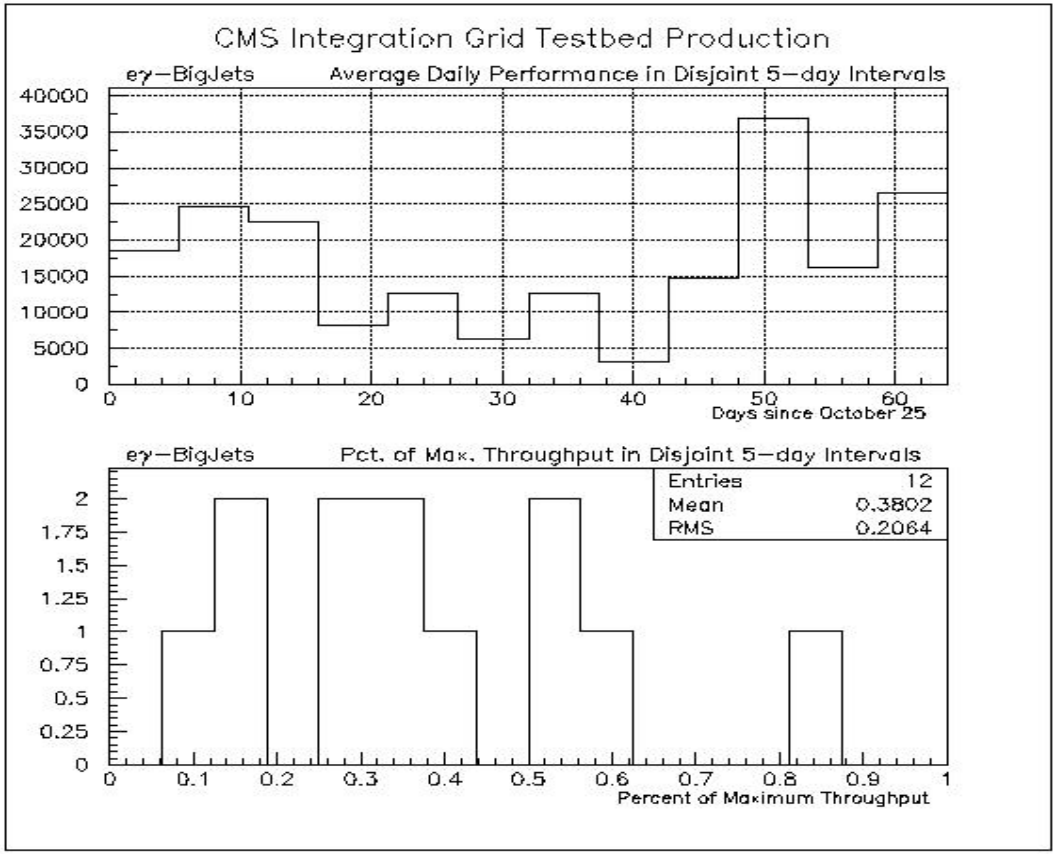
IGT Results

Time to process 1 event: 500 sec @ 750 MHz

Speedup: Avg factor of 100 speedup during current run

Resources: Approximately 230 CPU @750 MHz equiv.

Sustained efficiency: about 43.5%



1M fully simulated and reconstructed events!

Analysis of Inefficiencies

- 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 :-(

IGT Results

- Manpower Estimates
  - 2.65 FTE equivalent during initial phase and debugging
    - Reported voluntarily in response to a general query
  - 1.1 FTE equivalent during smooth running periods
    - The MOP person plus periodic small file transfers
  - Expected to be less than 1 FTE when regular shift procedures are adopted
- File Transfers are a small job so far
  - Only ntuples are kept in much of the production
  - All input files were staged from Fermilab and output staged to Fermilab using globus-url-copy
    - ie- No replica catalog was used DURING production

Comparison to Spring 02 and EDG Stresstest Efforts

- CMS Spring 2002 "Manual" Production
  - CPU utilization was 10-40% during Spring02
    - Comparison not fair because: I don't know how many CPU should be counted in the denominator and the IGT currently doesn't have any significant file transfers
  - Manpower was a lot higher
- Comparison to EDG Stresstest
  - Also completed CMS Production in Fall 2002
  - More manpower than IGT run, but also more advanced functionality was tested.
    - We have gained much experience from both "bottom up" IGT and "top down" EDG approaches.

US-CMS and LCG-1

- US-CMS has significant experience with integrating and deploying the components that will be chosen for LCG-1 Pilot in a large scale production environment
  - Capitalize on previous development efforts
- US Facilities will be early participants in LCG-1
  - First will be Fermilab
  - Other US sites may follow as the most efficient interface is determined

Conclusions

- The 3 phased approach (DGT, IGT, and PG) provides a lot of flexibility to the USCMS Grid operations.
  - Mission critical software, such as LCG-1, can be introduced quickly on the IGT and supported from there for Production Grid operations.
  - Speculative development and future oriented ideas can be developed and tried out on the DGT.