

# **CMS Distributed Production**



**Greg Graham**  
**for the CMS Collaboration**  
**CHEP'01 4-031**  
**Beijing, 2001-9-2**



# Introduction

- ⌘ CMS Monte Carlo Software
- ⌘ Production Sites
- ⌘ File Transfers
- ⌘ Objectivity Usage in CMS Production
- ⌘ Production Results
- ⌘ Random Pitfalls
- ⌘ Conclusions and Future Goals

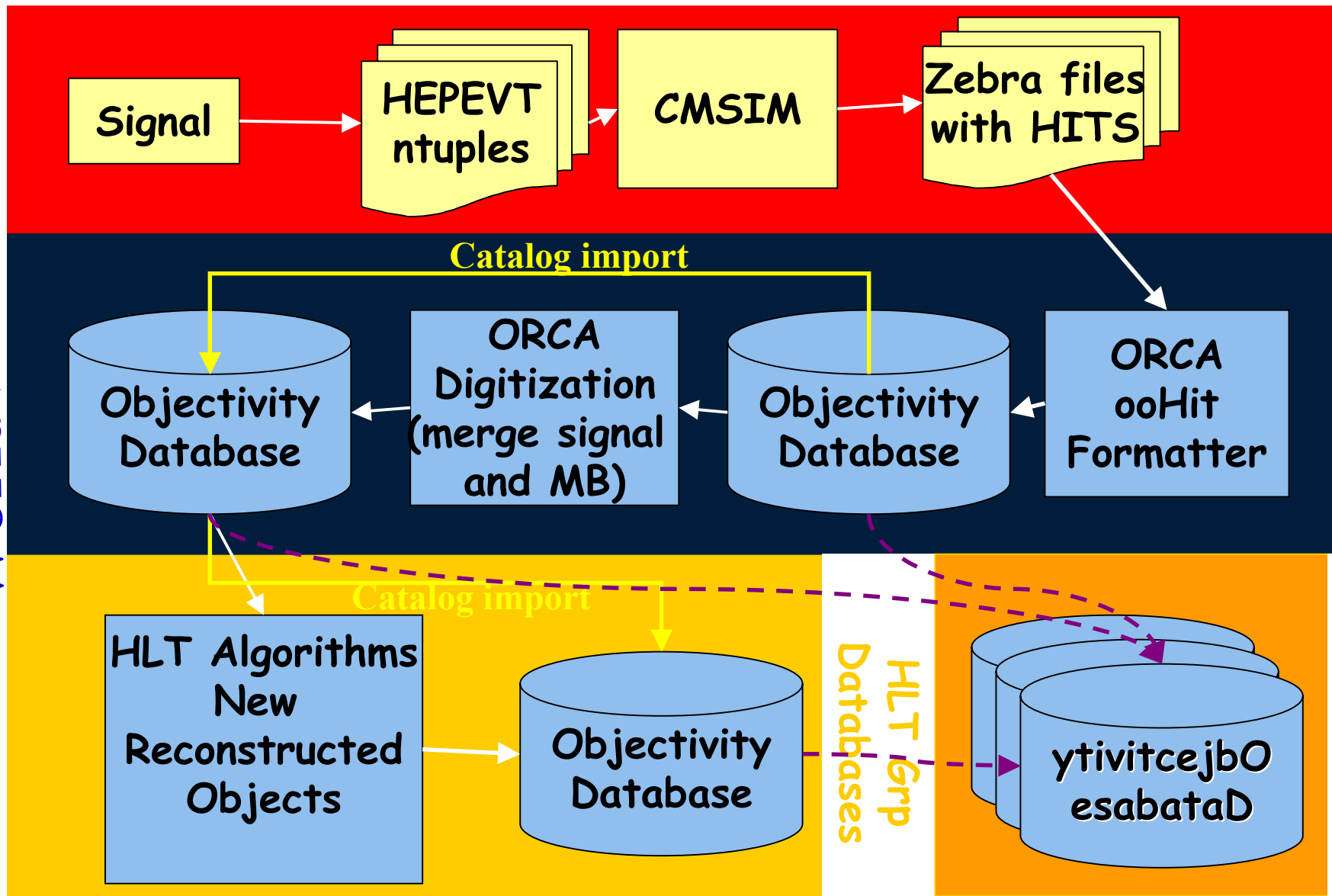
# CMS Monte Carlo 2001



Worldwide

Worldwide

Worldwide



MC Prod.

ORCA Prod.

Mirrored Db's  
(US, Russia, Italy..)

# CMS Monte Carlo 2001

- ⌘ GEANT-3 based cmsim to be replaced by OO GEANT-4 based OSCAR within 1 year
- ⌘ I/O and CPU intensiveness :
  - ⊞ cmsim: CPU intensive, not very I/O intensive
    - ⊞ 1 min/evt on 700 MHz CPU
  - ⊞ OO Formatting : I/O intensive, not very CPU
    - ⊞ Up to 50 Mb/sec network load, network limited
  - ⊞ OO Digitization : CPU Intensive and I/O intensive when using full pileup

# Production Sites

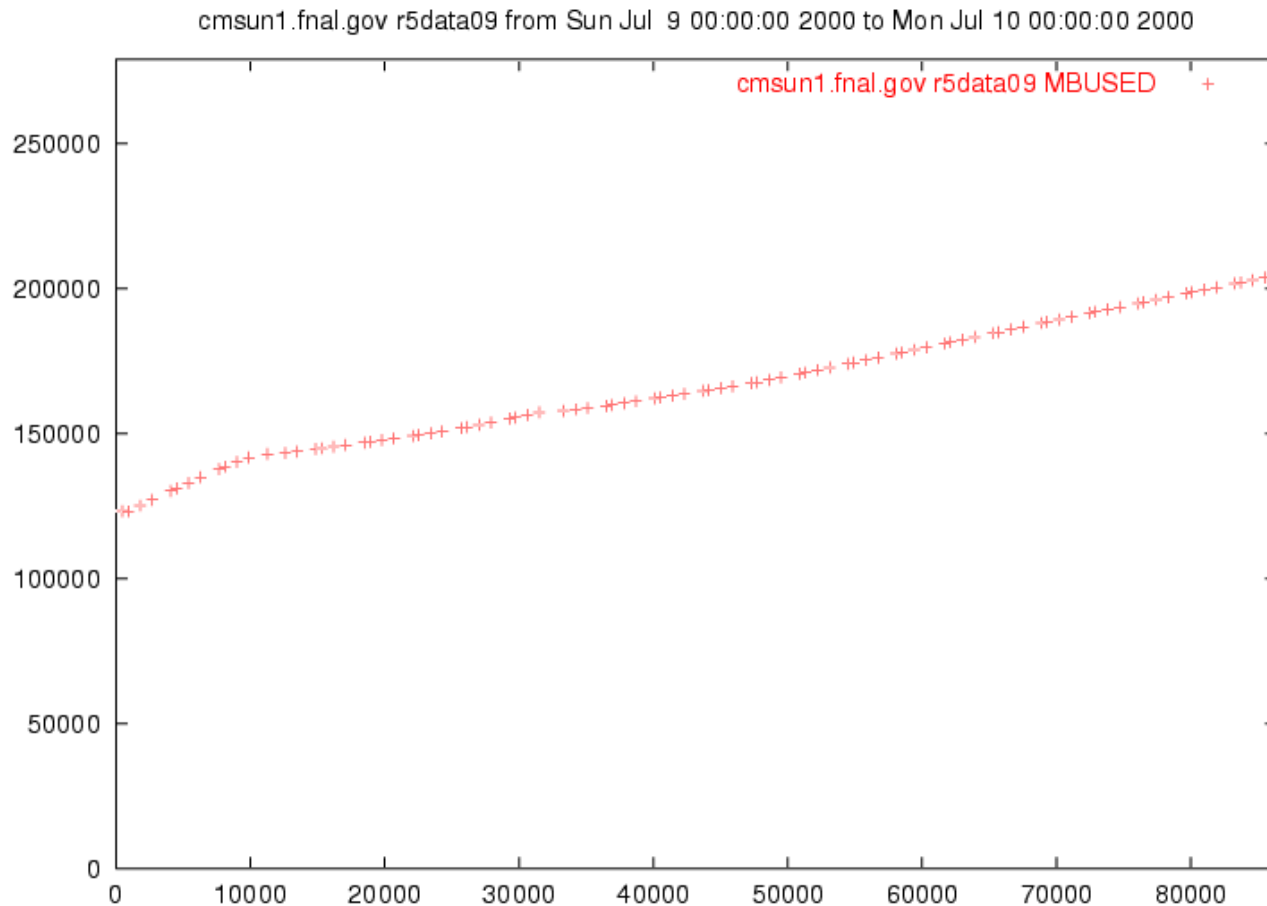
## ⌘ CMS Distributed Production Sites

- ☒ CERN : 200+ CPU, HPSS/CASTOR (Tier-0)
- ☒ INFN : Condor based system, BOSS (Tier-1)
- ☒ FNAL : 80+ CPU, Enstore (Tier-1)
- ☒ IN2P3 : (Tier-1)
- ☒ Helsinki (Tier-2)
- ☒ Moscow U (Tier-2)
- ☒ UCSD/Caltech (Tier-2)
- ☒ U. Florida (Tier-2)
- ☒ Bristol (Tier-2)

# File Transfers

- ⌘ BGE (Before Grid Era) : Use combination of parallel ftp and shell/perl scripts
  - ☑ Between FNAL and CERN, typically 5 GB/Hour
- ⌘ GE (Grid Era) : Use GDMP
  - ☑ Uses grid ftp to transfer Objectivity DB files and attach them to remote federations.
  - ☑ Now uses Globus replica catalog as a flat file catalog so FZ file transfer is possible.

# File Transfers (BGE)



# File Transfers

⌘ Other average transfer rates to CERN  
(measured in Spring 2000) :

- ⊗ FNAL: 4.7 GB/hour
- ⊗ HIP: 4.0 GB/hour
- ⊗ IN2P3: 3.0 GB/hour
- ⊗ Moscow: 1.0 GB/hour
- ⊗ Caltech: 5.0 GB/hour
- ⊗ Bristol: 6.1 GB/hour

# Objectivity Usage in CMS Production



## ⌘ Production Federations and User Federations

- ⊗ A “federation” is a collection of database files numbered from 1-64k (DBIDs)
- ⊗ Database files can contain data or metadata
- ⊗ Data is written into production federations and then the DB files (data+metadata) are attached to a user federation for analysis
- ⊗ DBID's must be unique !!!

# Objectivity Usage in CMS Production



## ⌘ Management of DBIDs

- ☑ Production sites are assigned non-overlapping ranges of DBIDs
- ☑ If they run out, they must request more.
  - ☒ In some cases, this has been ignored. The consequences are that datasets with overlapping DBIDs cannot be access from the same user federation later.

# Objectivity Usage in CMS Production



## ⌘ Objectivity Servers

### ☑ AMS (Advanced Multithreaded Server) :

- ☑ responsible for managing access to data in the Objectivity database files
- ☑ responsible for writing journal files for transaction support

### ☑ Lock server :

- ☑ responsible for granting locks for safe access to objects in database files

# Objectivity Usage in CMS Production



## ⌘ How does Objectivity get/put your data ?

- ☒ 1) Run an AMS server on the host where the database files exist; host and pathnames are reflected in the federation catalog
  - ☒ Can use OO\_DB\_HOST and OO\_DB\_PATH to control where new files are written
- ☒ 2) Use request redirection protocol (RRP)
  - ☒ Allows data requests to be redirected to different AMS servers based on information in a configuration file.
  - ☒ In use at several sites including CERN and UCSD
- ☒ A rich set of server configurations is possible

# Example Objy Server

## Depl

4 Production Federations at FNAL.

(Uses catalog only to locate database files.)

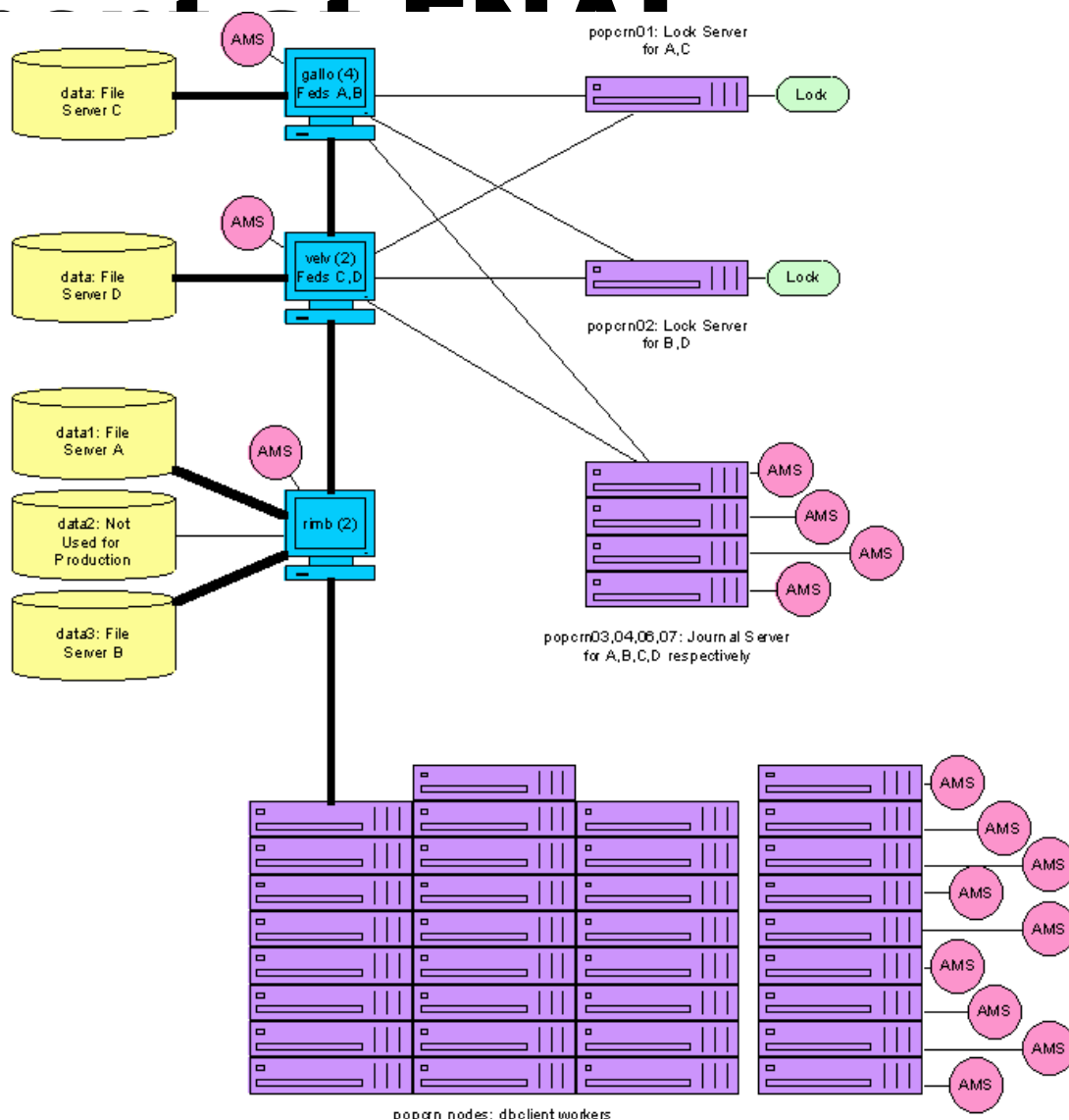
3 FNAL servers plus several worker nodes used in this configuration.

3 federation hosts with attached RAID partitions

2 lock servers

4 journal servers

9 pileup servers



# Objectivity Usage in CMS Production



- ⌘ AMS provides an interface to use MSS backend for serving files
  - ☑ At CERN, this has been done for access to HPSS and CASTOR
  - ☑ At FNAL, this has been done for access to MSS using Enstore
    - ☒ 10 MB/s per tape drive X 4 drives for CMS
    - ☒ We use this both for serving data to user federations as well as in the production federations.

# Production Results

A thick, horizontal yellow brushstroke is positioned below the title, extending across most of the width of the slide.

- ⌘ CMS Spring 2000 Production

- ⌘ CMS Fall 2000 Production

  - ☒ It's been a long autumn ...

    - ☒ (ie- it's still winding down.)

# CMS Spring 2000 Production (CERN)



- ⌘ 140 Digitization jobs reading asynchronously from 30 AMS server's, writing to a high speed SUN server.
- ⌘ Data staged automatically from tape (AMS/HPSS interface) when needed.
- ⌘ Data archived to HPSS automatically by CDR.
- ⌘ 70 jetmet jobs at  $\sim 60$  seconds/event and 35MB IO/event
- ⌘ 70 muon jobs at  $\sim 90$  seconds/event and 20MB IO/event
- ⌘ Best Reading rate out of Objectivity 5  $\sim 70$ MB/sec
- ⌘ Continuous 50MB/sec reading rate out of Objectivity 5  
2 million events, digitized in two weeks

# CMS Fall 2000 Production (CERN, FNAL, INFN, ...)



- ⌘ Approx. 6 million events processed so far, with various pileup conditions (none,  $10^{33}$ ,  $10^{34}$ )
  - ☒ At  $10^{34}$  luminosity, an average of 200 pileup events are needed per signal event !
- ⌘ Most FZ files are successfully stored at CERN
- ⌘ Production is divvied up among Tier-1 sites roughly by physics group
  - ☒ CERN/Bristol : egamma production
  - ☒ INFN : Muon production
  - ☒ FNAL : Jets and Missing Et production

# CMS Fall 2000 Production (FNAL)



## ⌘ FNAL Hardware

- ⊞ 40 dual node 750 MHz Intel based worker nodes
- ⊞ 3 quad node 650 MHz Intel based server nodes
- ⊞ 1 250 GB RAID5 partitions (Dell Powervault) per server
- ⊞ (1 soon to be canned dual CPU server with 1.5 TB RAID; RAID will be salvaged ...)
- ⊞ 100 Mb/s Ethernet (soon to be upgraded to Gb ethernet)
- ⊞ 1 8 CPU 400 MHz Sun Server with 1 TB RAID for User federation

## ⌘ FNAL Experience :

- ⊞ Limited by AMS server to 15 concurrent formatting jobs, but overcame this by going to multiple federations.
  - ⊞ (We did not play with the OO\_RPC\_TIMEOUT, but we did find that this was insensitive to the thread count.)
- ⊞ >3 Hit formatting jobs will starve digitization per federation
- ⊞ File descriptor limit for AMS server was raised to 4096.
  - ⊞ This limit was reached at CERN, but not yet at FNAL.

# CMS Fall Production 2000 (FNAL)

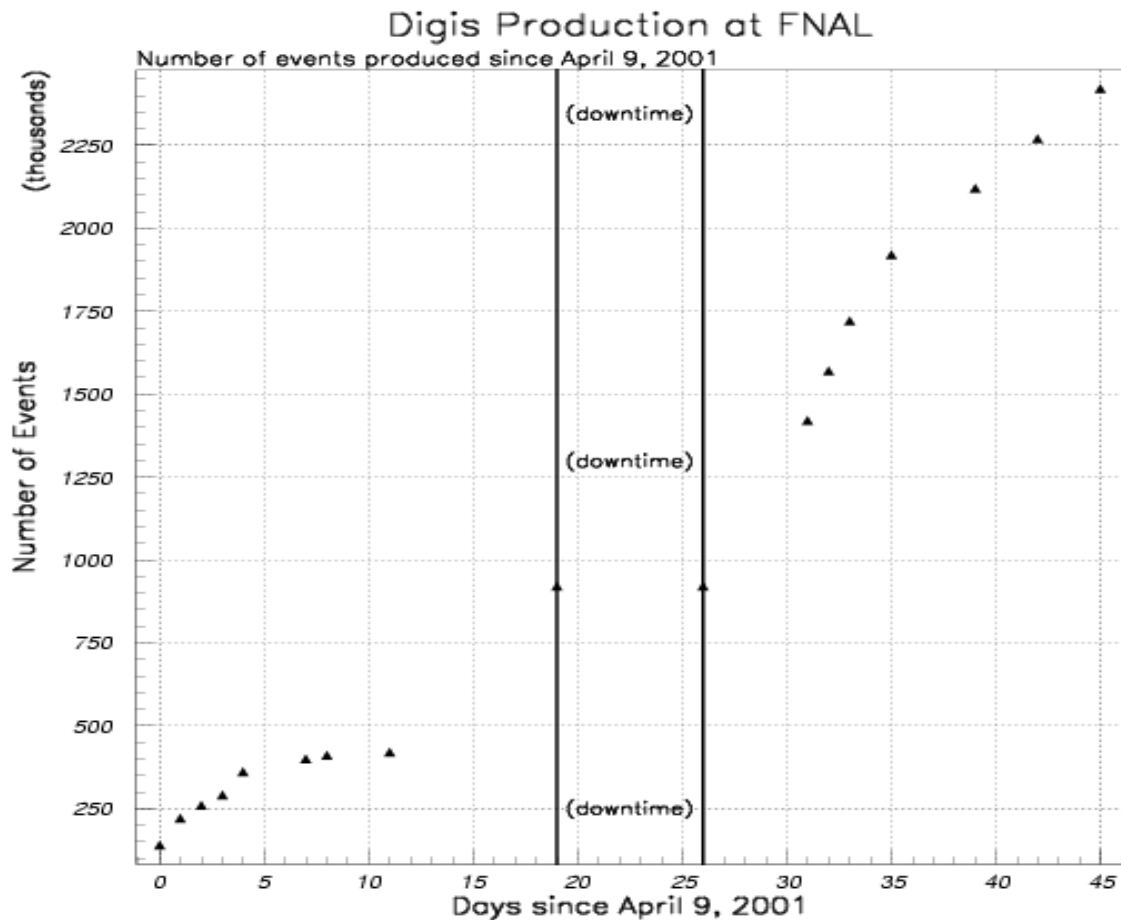


## ⌘ Pileup, pileup, and more pileup ...

### ☑ FNAL farm:

- ☑ 60 CPU processing digitization jobs - requires about 833 Mb/s of pileup data on average.
- ☑ Use 9 pileup servers on 100 Mb/s network for full pileup.
  - But we didn't reach the network limit
- ☑ FBSNG batch manager used to configure the farm
  - pileup intensive jobs required to NOT run on pileup serving worker nodes.

# CMS Fall 2000 Production (FNAL)



# CMS Production Fall 2000 (CERN)



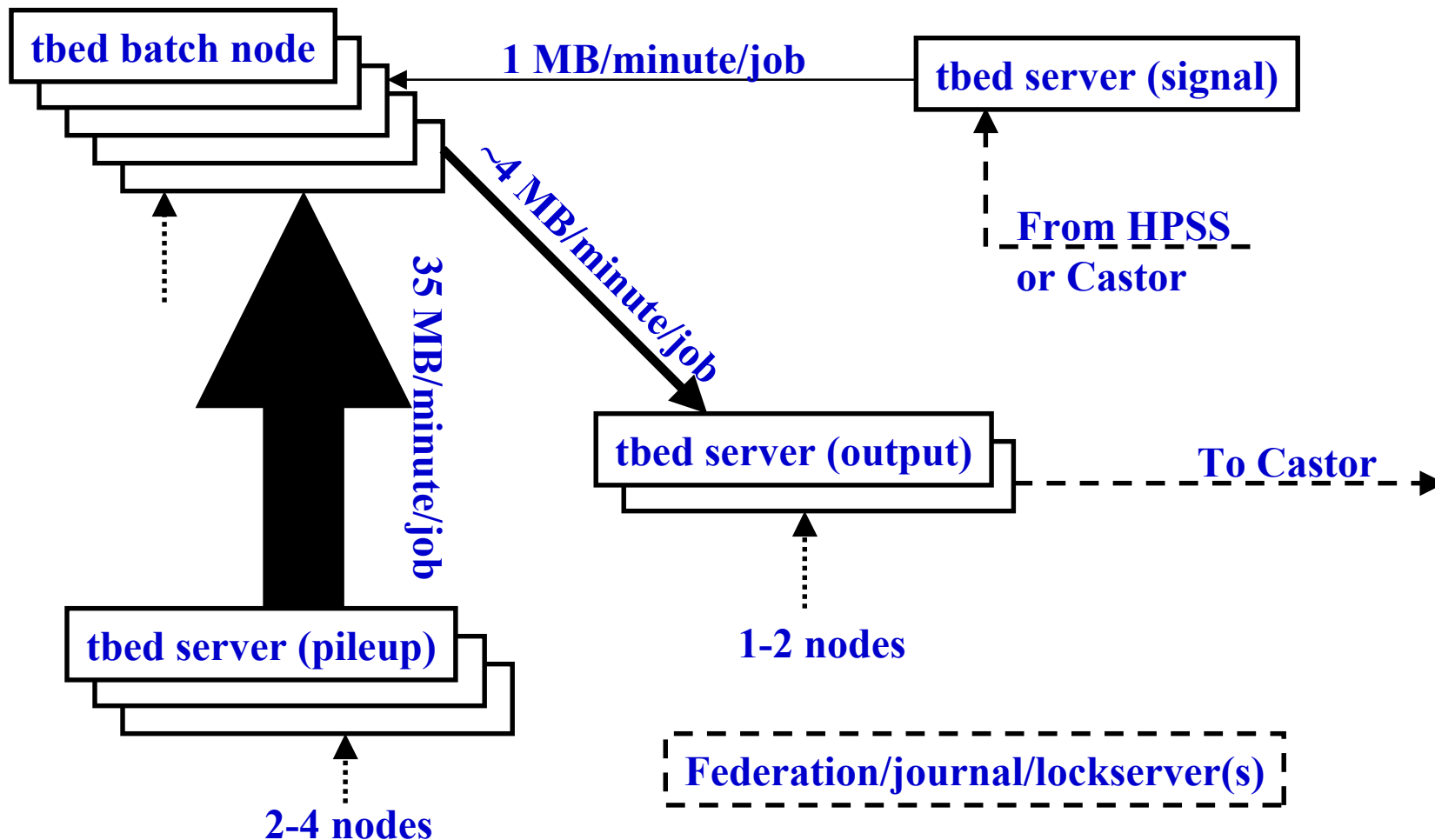
- ⌘ Farm is >150 Dual CPU 850 MHz Intel
- ⌘ Ran 300 full pileup jobs into the same federation
  - But throughput leveled off at 120 jobs
- ⌘ Serve pileup at 20 MB/sec sustained per server
  - ☒ We need 0.5 MB/sec/job, one server supports 40 jobs
  - ☒ For 300 jobs, would need 8 servers
  - ☒ CPU usage seen to be ~25% USR, ~140% SYS (dual CPU)
    - ☒ Not limited by the AMS CPU consumption itself, either the kernel or the I/O is the limiting step
    - ☒ 300 filehandles open on the AMS at this time

# CMS Fall Production 2000 (CERN)



- ⌘ Serving signal is no effort, need 1 MB/min/job
- ⌘ Output server writing at up to 24 GB/hour (6-7 MB/sec)
  - ⏏ Then mostly busy in I/O, probably saturated
- ⌘ Tests with ~120 jobs did not seem to reach these hardware limits, but not clear what limit they did reach
  - ⏏ Ran out of time for investigating further with full farm
    - ⊗ Most of the time spent running 'real' production, not tests
  - ⏏ Smaller scale tests still going on

# CMS Fall 2000 Production (CERN)



# Random Pitfalls

## ⌘ FNAL:

- ☑ Dell Powervault RAID arrays failed often with complete data loss in RAID level 5.
  - ☒ Dell Solution: Drives in dual channel configuration to boost performance NOT a good idea ...
- ☑ 3ware IDE RAID servers often went down with SCSI controller errors.
  - ☒ Un-Solution: Might be server ... still investigating

# Random Pitfalls

## ⌘ CERN

- ⊞ Severe problems with 3Ware-based IDE disk servers
- ⊞ Data-loss at  $10^{**4}$  times higher than 3Ware quotes
  - ⊗ The same configuration ALICE used, same machines, same disks, directly before CMS. ALICE saw no such problems!
- ⊞ Cause unclear. Suspect bad disks, but also unclear if CMS data-access patterns are triggering specific problems
- ⊞ No tool to reproduce data-access patterns reliably.
  - ⊗ Need to run the whole farm to reproduce effects!

# Random Pitfalls

---

## ⌘ CMS intends to continue testing these servers

- ☒ Performance is excellent, clearly on a par with SCSI
- ☒ Performance per buck is unbeatable
- ☒ 3Ware provide frequent firmware upgrades
  - ☒ Both good news and bad!
    - They care, they are active
    - They do not have a fully mature product
  - ☒ Latest version of firmware has not yet been seen to fail
    - Tested artificially, but for several days. Cautiously optimistic ...

# Conclusion and Future Goals



- ⌘ Continue pushing Objectivity to the limit
  - ☑ Most centers have upgraded to Objectivity 6.0
  - ☑ Farms are getting larger ...
- ⌘ Make more use of grid tools
  - ☑ Early effort is going into integrating GDMP
    - ☒ FNAL runs a GDMP "heartbeat" file transfer several times daily from US Tier-2 sites

# Conclusion and Future Goals



⌘ Taking a harder look at the manpower intensive tasks with an eye towards automation.

- ⊗ Monte Carlo request formalism and automatic processing
- ⊗ Parameter Tracking
- ⊗ Automatic notification and request tracking
- ⊗ Systems are excellent at “farm” level, but still not “integrated” without lots of manpower

# Acknowledgements

---

⌘ I am heavily indebted to Tony Wildish at CERN for the CERN material.

⌘ Other related talks of interest at CHEP'01:

⌘ 10-051: CMS Grid Activities in Europe *C. Grandi, et al.*

⌘ **I didn't talk about BOSS ... an excellent Condor based system used at INFN.**

⌘ 10-052: CMS Grid Activities in the US *I. Fisk, et al.*

⌘ 10-053: CMS Requirements for the Grid K. *Holtmann, et al.*

⌘ 4-033: Tools and Infrastructure for Distributed Production of CMS Monte Carlo Samples *G. Graham, et al.*

⌘ 1-001: FBSNG - Batch System for Farm Architecture *J. Fromm, et al.*