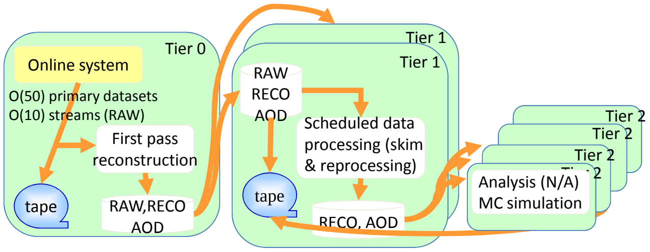
CMS production and proccesing system – design & experiences

CMS computing

- Distributed model
- Multiple sites – varying size / grid / storage technologies

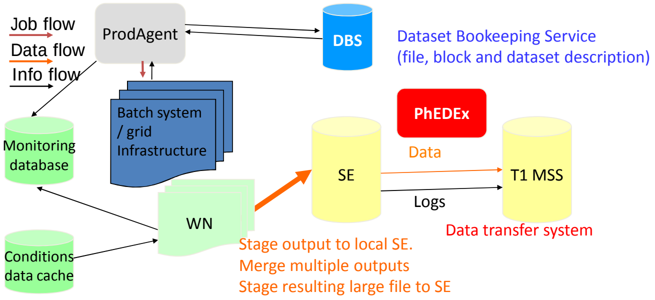

Tier 0
Online system
O(50) primary datasets
O(10) streams (RAW)
First pass reconstruction
tape
RAW,RECO AOD
Tier 1
Tier 1
RAW
RECO
AOD
Scheduled data processing (skim & reprocessing)
tape
RECO, AOD
Tier 2
Tier 2
Tier 2
Tier 2
Analysis (N/A)
MC simulation

- Low latency critical processing (Tier 0):
  - Prompt Reconstruction
  - Express Stream
  - Alignment and Calibration
  - Data Quality Monitoring
- Offline processing (7 Tier 1's and ~50 Tier 2's):
  - Reprocess data 3 times a year
  - Dataset skims of events with physics criteria post reconstruction
  - Monte Carlo production (1:1 ratio to collected data)

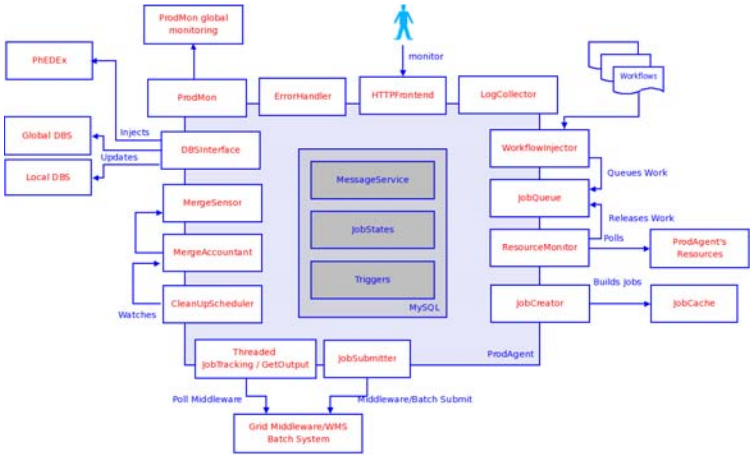
Design

- Workflow management used by all organized CMS processing
- Multiple independent instances (called ProdAgent's)
- In use for 2+ years
  - More recently adopted as base for Tier0 quasi-real time data processing.
- Automation
- Scalability
- Highly configurable/extensible:
  - Production and Processing, Real and MC, Online and Offline.
  - Grid and Non grid
  - Work with different site setups (storage, batch system)

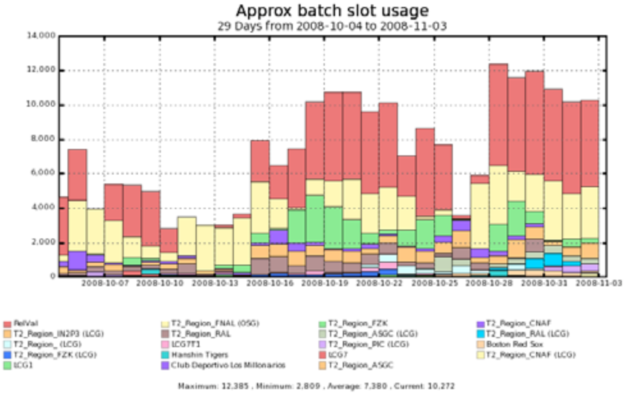
Production system workflow

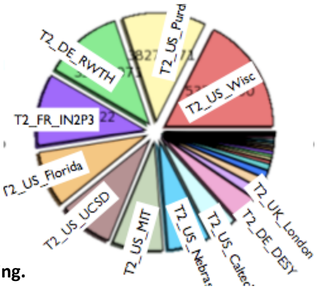

Job flow
Data flow
Info flow
ProdAgent
DBS
Dataset Bookeeping Service
(file, block and dataset description)
Batch system / grid Infrastructure
Monitoring database
Conditions data cache
WN
SE
PhEDEx
Data
T1 MSS
Logs
Data transfer system
Stage output to local SE.
Merge multiple outputs
Stage resulting large file to SE

- Independent asynchronous components
- Written in python for flexibility and ease of development
- MySQL database.
  - Persistency
  - Communication between components
- Plugins used for specialized behavior

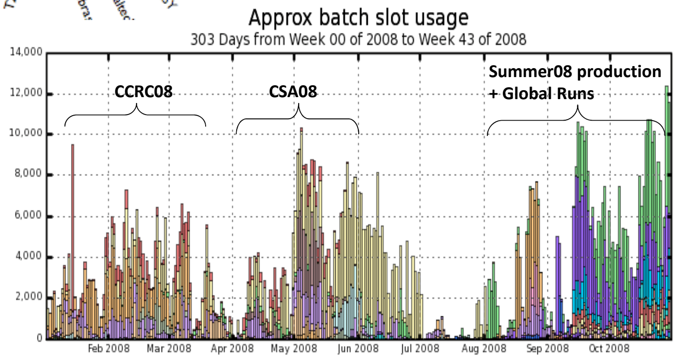

ProdMon global monitoring
PhEDEx
ProdMon
ErrorHandler
HTTPFrontend
monitor
LogCollector
Workflows
Global DBS
Local DBS
Injects
Updates
DBSInterface
WorkflowInjector
Queues Work
MessageService
JobStates
Triggers
MySQL
JobQueue
Releases Work
MergeSensor
MergeAccountant
CleanUpScheduler
Watches
ResourceMonitor
Polls
ProdAgent's Resources
Builds Jobs
JobCreator
JobCache
ProdAgent
Threaded JobTracking / GetOutput
JobSubmitter
Poll Middleware
Middleware/Batch Submit
Grid Middleware/WMS Batch System

Experience – incl. beam and cosmic runs



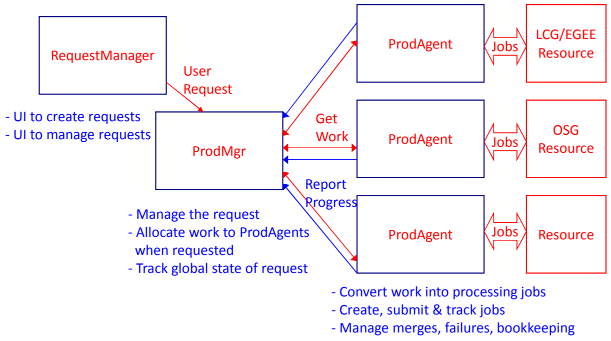


- Used within computing challenges (CCRC08 & CSA08), production activities and global runs
- Recently reached nominal startup goals
  - 100M events a month
- Utilized multiple prodAgents to reach goal
  - Only 1 submission technology per instance.
  - Reaching limits with grid middleware interaction
    - Move to Bulk/threaded operations



Development plans

- Service to manage physics requests (Request Manager):
  - Enforce testing of workflows
  - Priorities
  - Improve physicist feedback
  - Distribution of work to ProdAgents
- Distribute work to instances (ProdMgr):
  - Improve scalability
  - Spread work over different grids
- Create new common layer for distributed projects:
  - Production & analysis systems – also Tier 0
  - Avoid separate implementations for each project
  - Take best of breed from each project
  - Refactor - improve testing, style etc.


RequestManager
User Request
- UI to create requests
- UI to manage requests
ProdMgr
Get Work
Report Progress
- Manage the request
- Allocate work to ProdAgents when requested
- Track global state of request
ProdAgent
ProdAgent
ProdAgent
Jobs
Jobs
Jobs
LCG/EGEE Resource
OSG Resource
Resource
- Convert work into processing jobs
- Create, submit & track jobs
- Manage merges, failures, bookkeeping