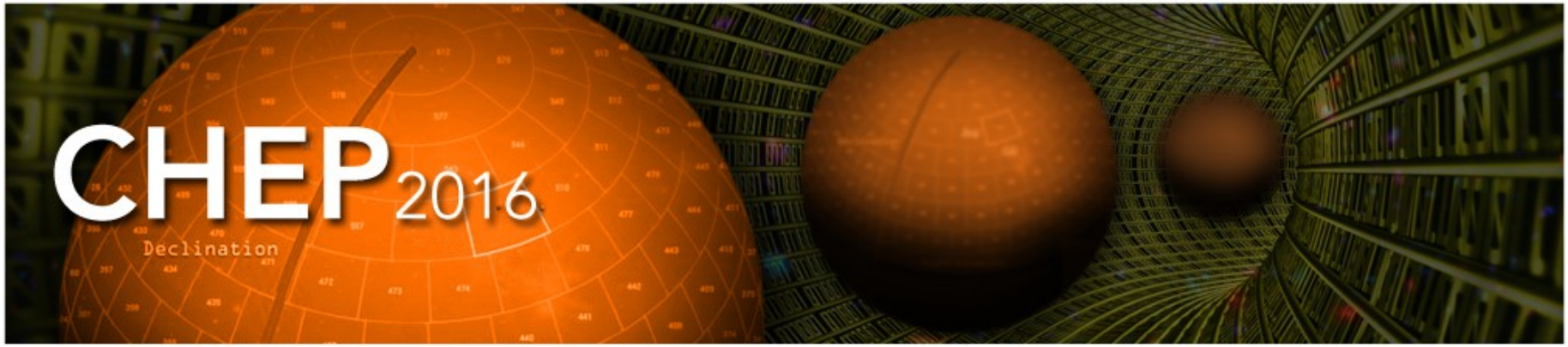
Software and Experience with Managing Workflows for the Computing Operation of the CMS Experiment



CHEP, October 10, 2016

Jean-Roch Vlimant for the CMS Collaboration

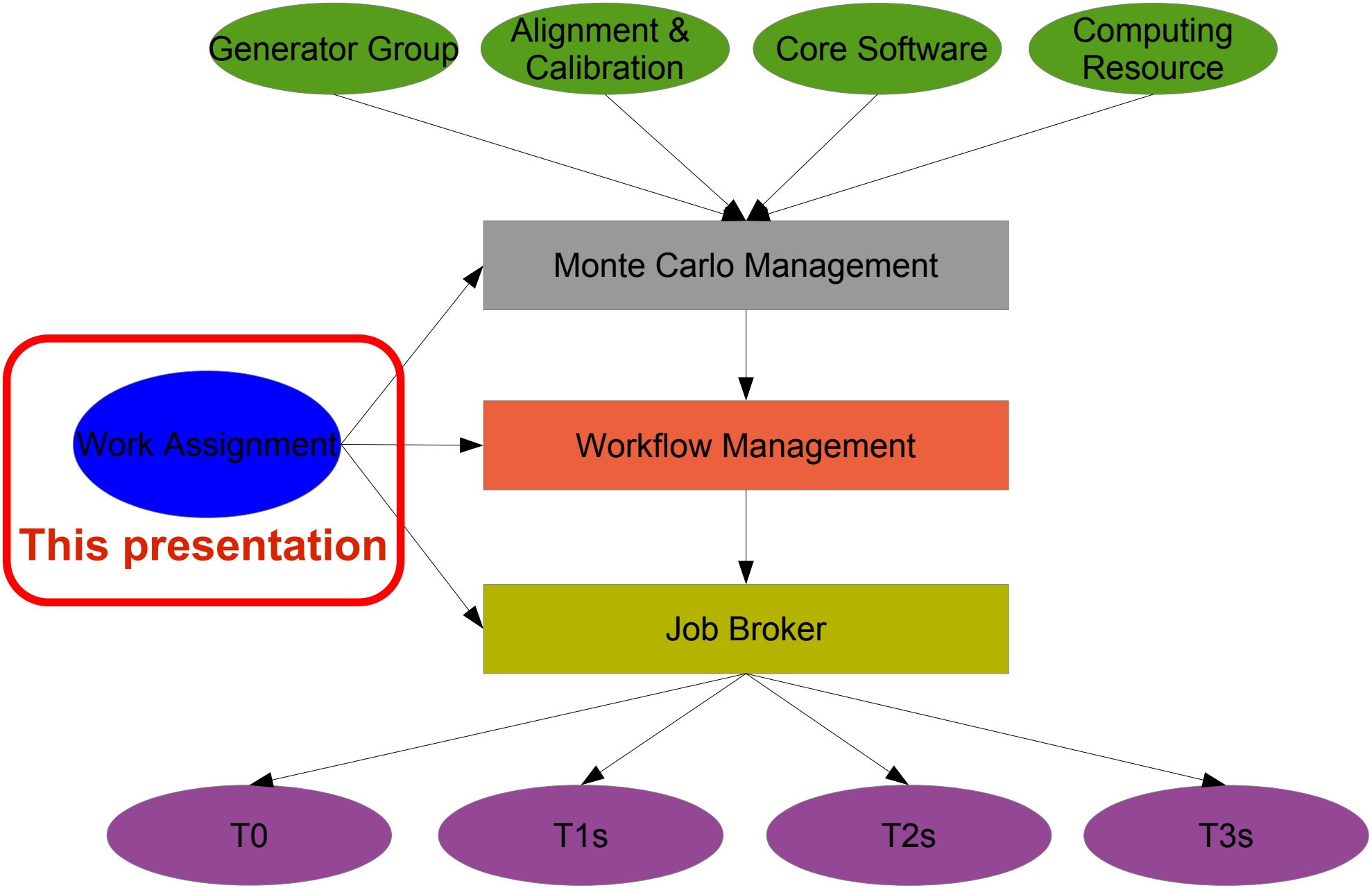
Production Overview

- Analyzing CMS data requires a **large volume of Monte-Carlo**
  - ✓ Billions of events in 10s of thousands of datasets
  - ➔ Need a **system that scales**
- Production is done in **successive steps** towards the production of the analysis datasets
  - ✓ Arrangement dictated by software requirements, flexibility, resource utilization, ...
  - ➔ Working towards **all-in-one workflow**
  - ➔ Requires a **flexible system**
- Production is performed over the LHC grid
  - ✓ 1 Tiern 0 (CERN), 6 Tier 1, ~60 Tier 2, hundreds of T3
  - ✓ **Heterogeneous clouds** summing up **200k cores** available
  - ➔ **Automation is the key** to using diverse resource
- Sites can develop features or become available very fast
  - ✓ Also opportunistic resource
  - ➔ Need a **dynamic system**

Outline

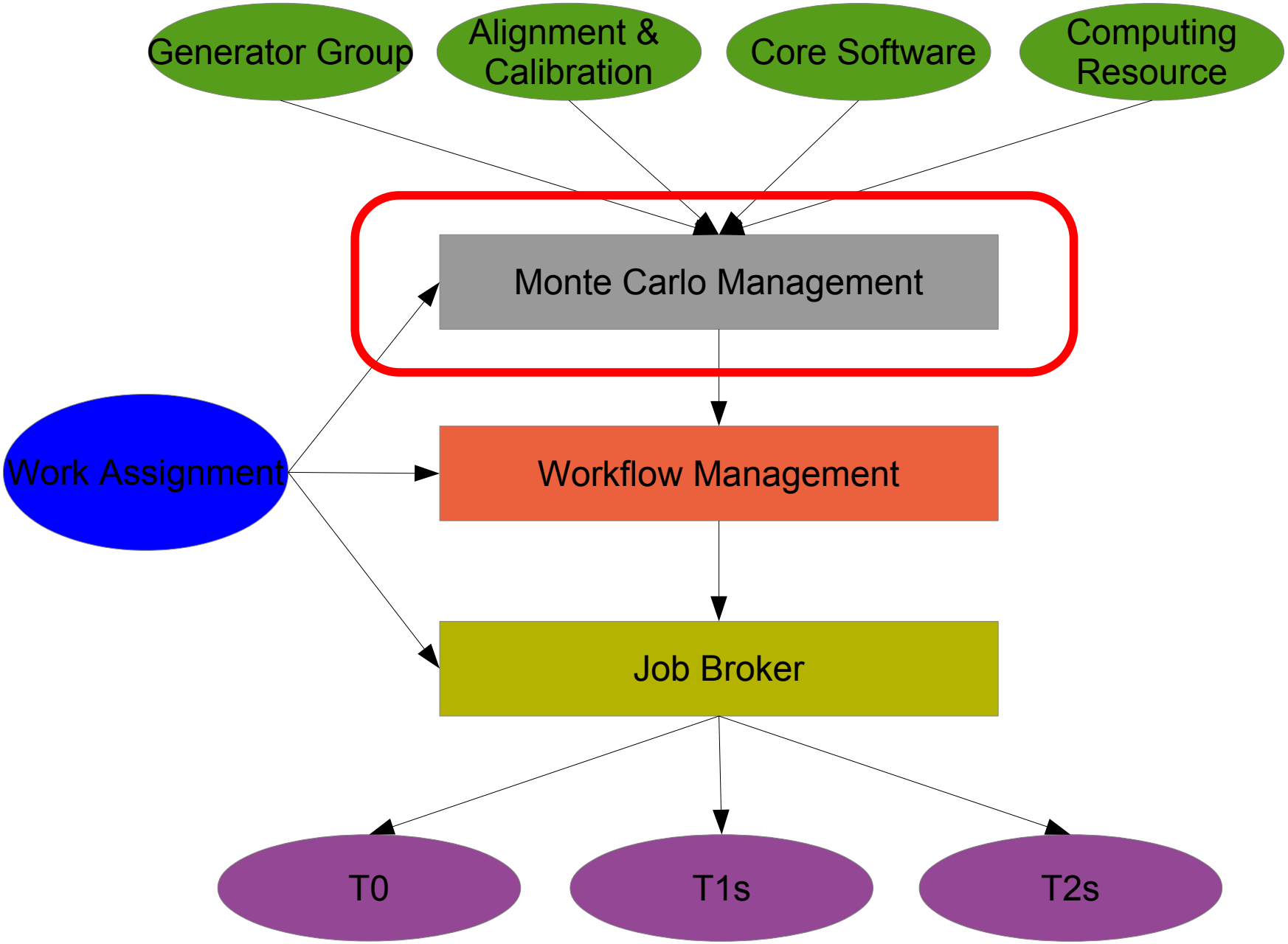
- Handling production
  - Quick Overview
  - Data Placement
  - Work Distribution
  - Work Routing
  - Monitoring
- Summary & Outlook

Production Overview



Handling Production

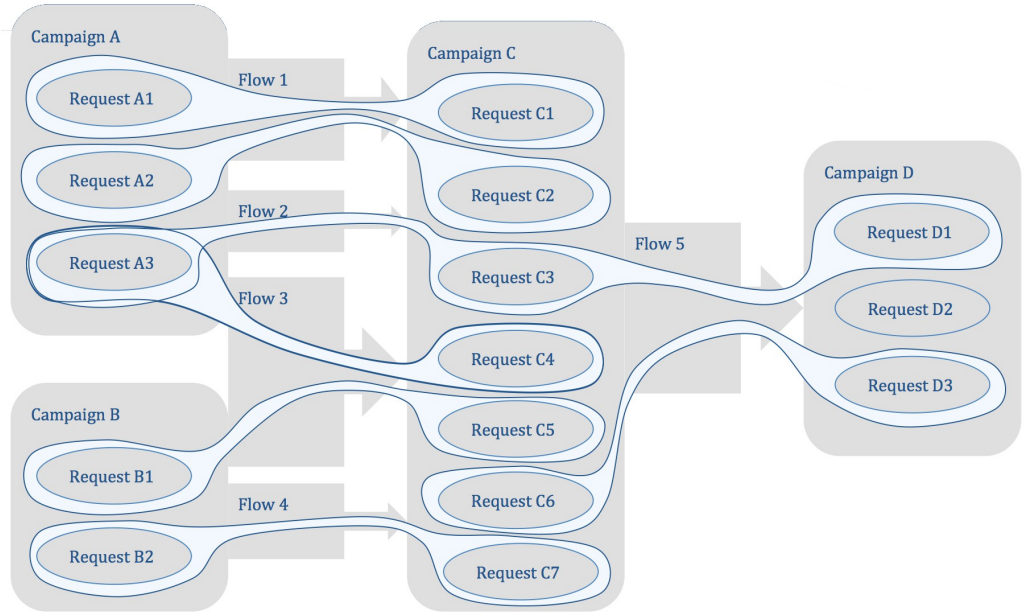
Production Overview



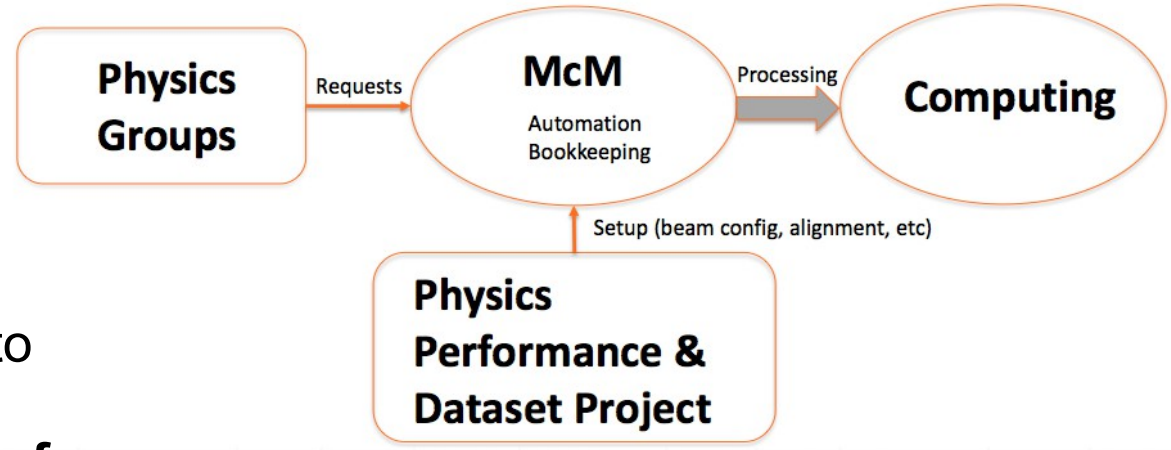
Configuration Assembling

Monte-**C**arlo **M**anagement (McM)

- CMS Software configuration and ingredients for production steps aggregated in **campaigns**
- Subsequent steps of production materialize in **chains of campaigns**
- Flow implement campaign modifiers
- Allow for complex chaining
- **Flexibility** for defining any specific request

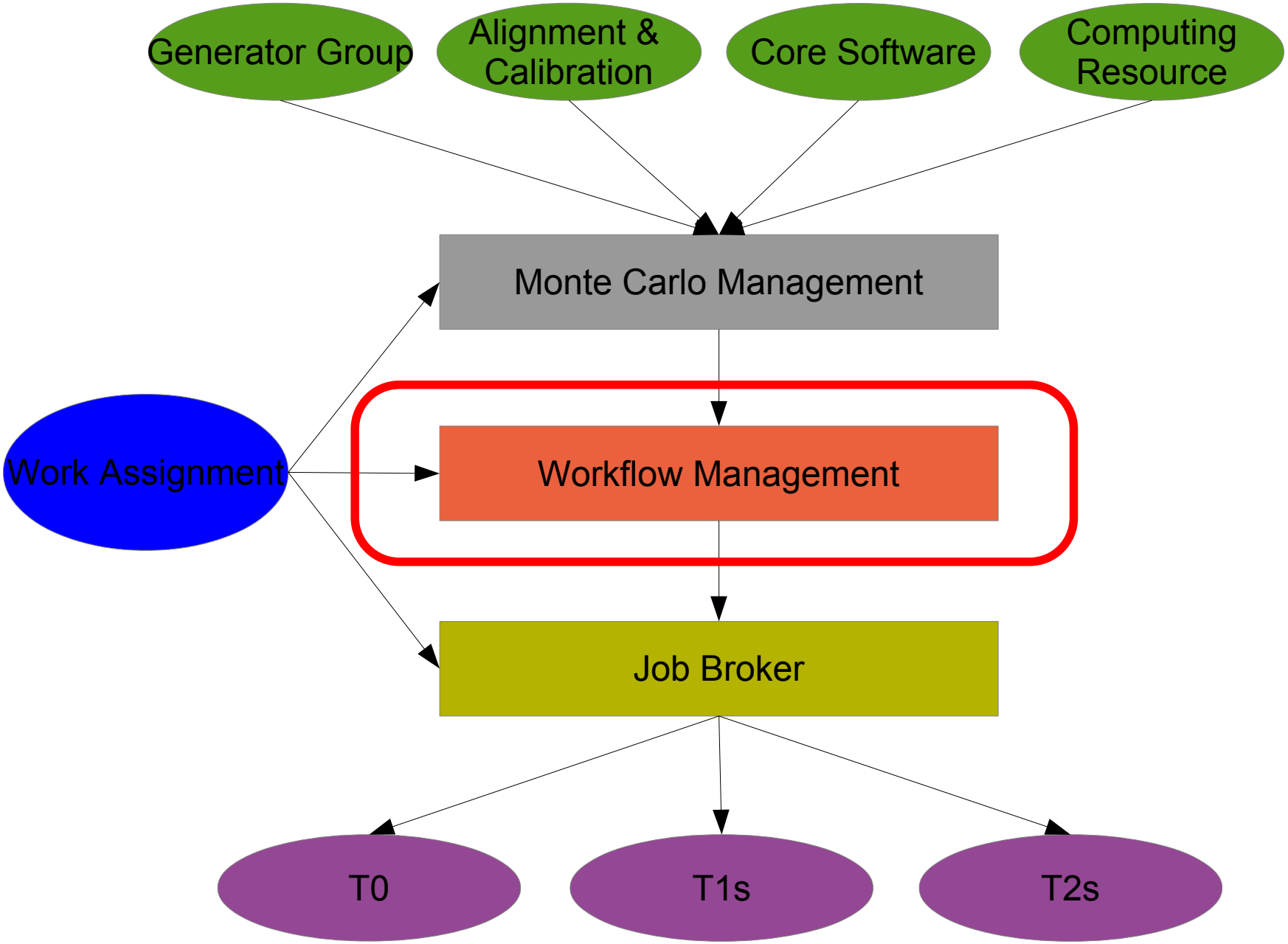


- ✓ Samples requests added by generator contact person
- ✓ Chaining operated by production managers
- ✓ **Automation** where relevant
- ✓ **Validation** histogram provided
- ✓ **Performance run-test** executed



- ➔ Injection of **consolidated workflow** to production system
- ➔ Ability to inject a workflow with **trees of processing steps**

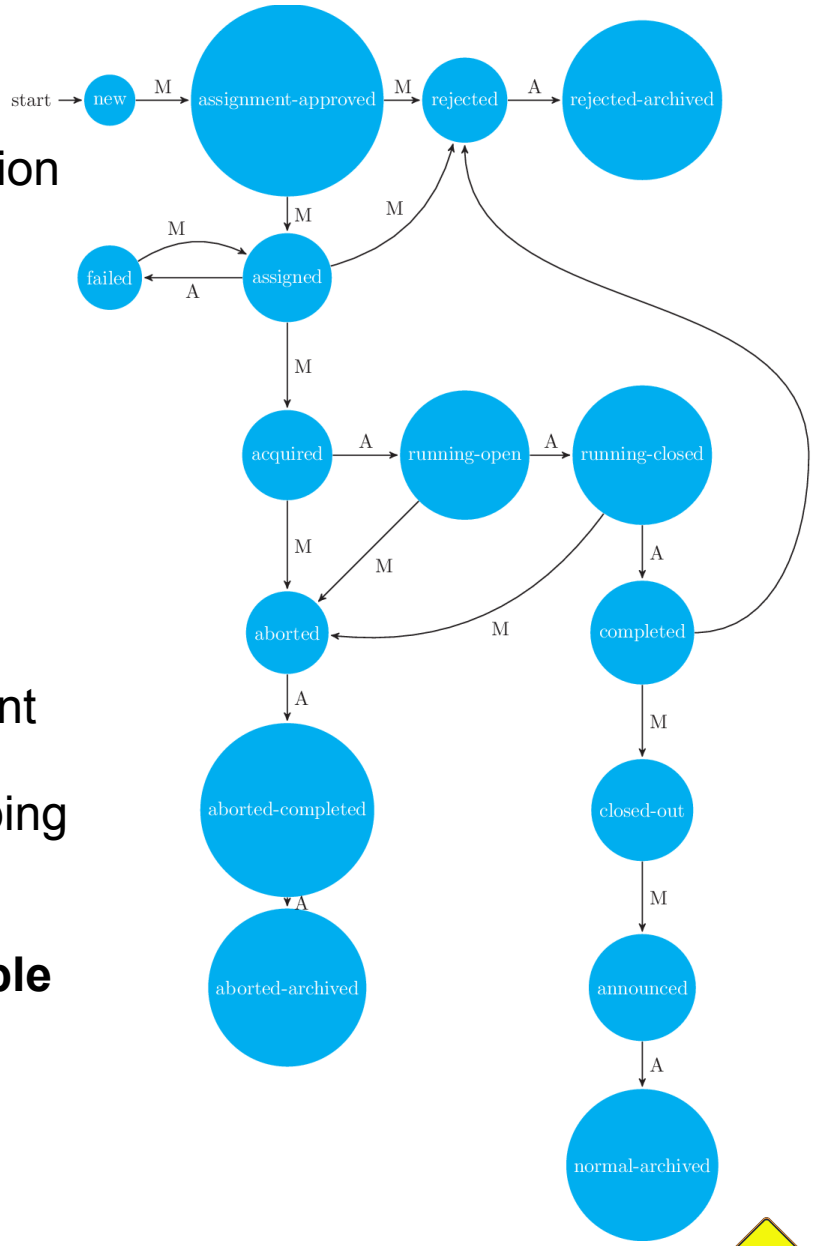
REFERENCE


Generator Group
Alignment & Calibration
Core Software
Computing Resource
Monte Carlo Management
Work Assignment
Workflow Management
Job Broker
T0
T1s
T2s

Workflow Management

- **Receive assembled configuration**
- Driven by work assignment agent

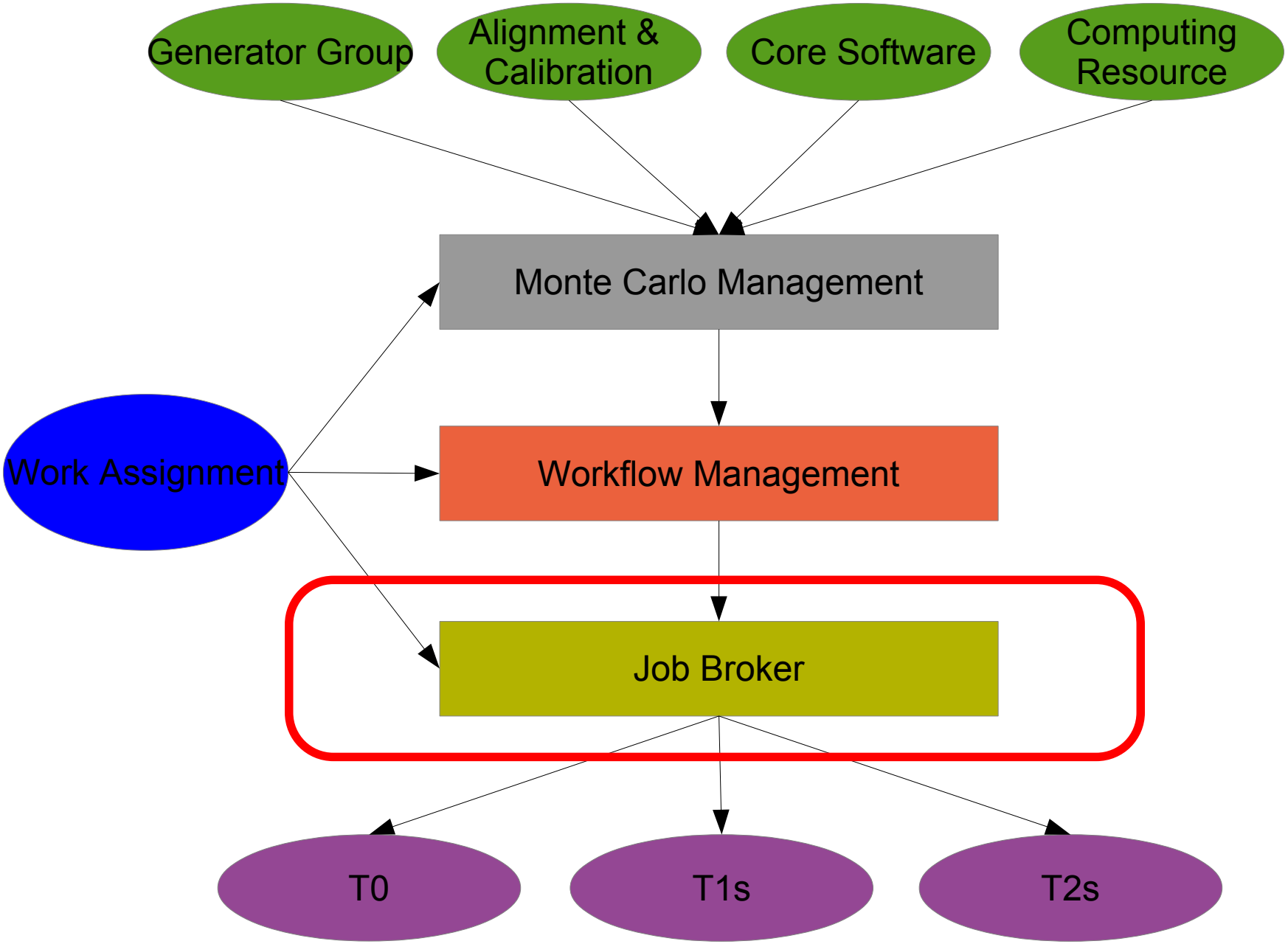
- Prepare the **full tree of processing** towards the production of the final output
  - ✓ Actual data processing and production
  - ✓ Additional steps: merging small output files for transfer efficiency, cleaning of outputs, collecting of running log files, ...
- **Split jobs** according to workload specifications and data content
- Submit jobs to broker (HTCondor)
- Resubmit certain types of failures
- Keep the books of production data location for subsequent processing
- **Inject the produced data** with parentage into book keeping system



➔ System composed **central request manager** and **multiple agents** supporting high load
  - ✓ 5k workflows
  - ✓ 200k jobs pending
  - ✓ 150k jobs running

➔ Constant improvement for scalability

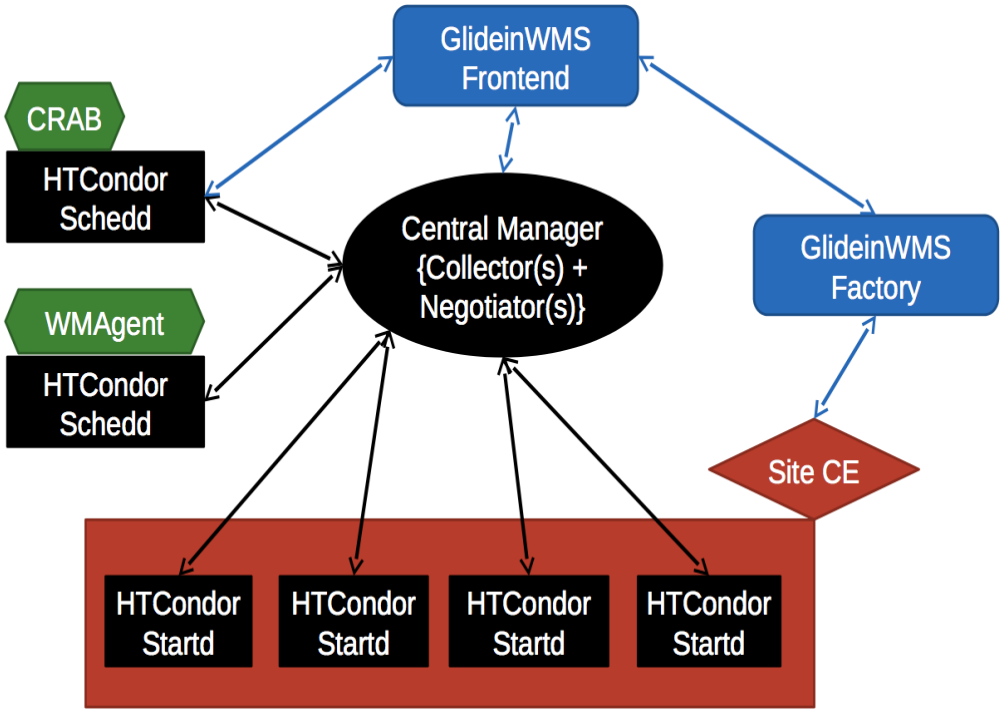
REFERENCE


Generator Group
Alignment & Calibration
Core Software
Computing Resource
Monte Carlo Management
Work Assignment
Workflow Management
Job Broker
T0
T1s
T2s

Job Brokering

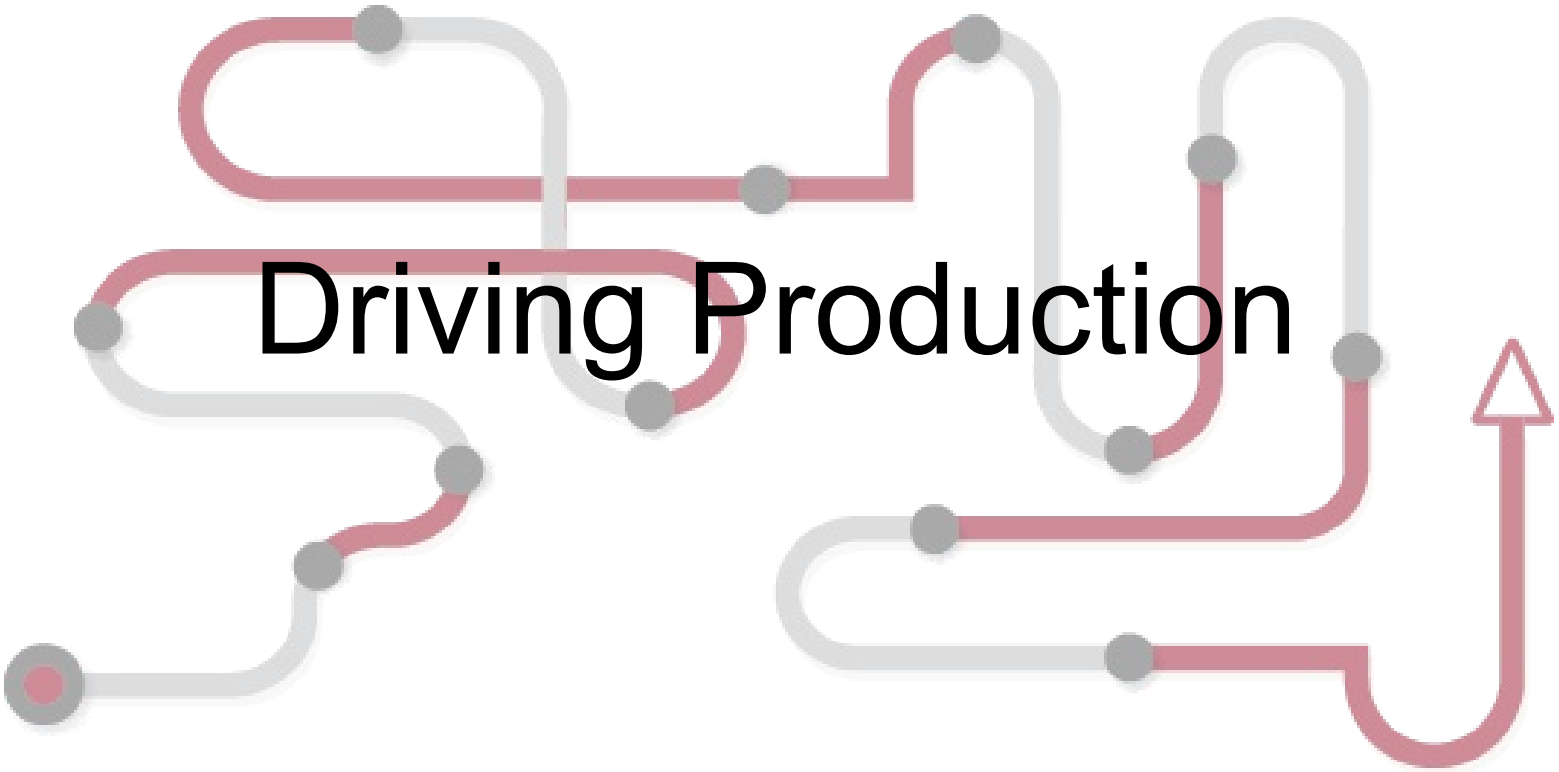
- **Shared resource** between analyzer and central production in a global pool
  - ✓ T0 production on a specific pool
- Use of HTCondor + glidein mechanism
  - ✓ Wrapper job : pilot running on site
  - ✓ Receive and execute trusted jobs
- Double stage of matchmaking
  - ✓ Jobs to resource (start pilots)
  - ✓ Jobs to pilots (claim pilots)
- Migrated for a large fraction to **multi-core partitionable pilots**
  - ✓ Allows multi-thread application
  - ➔ Moving most workflows to 4+ threads
- High Throughput computing solution

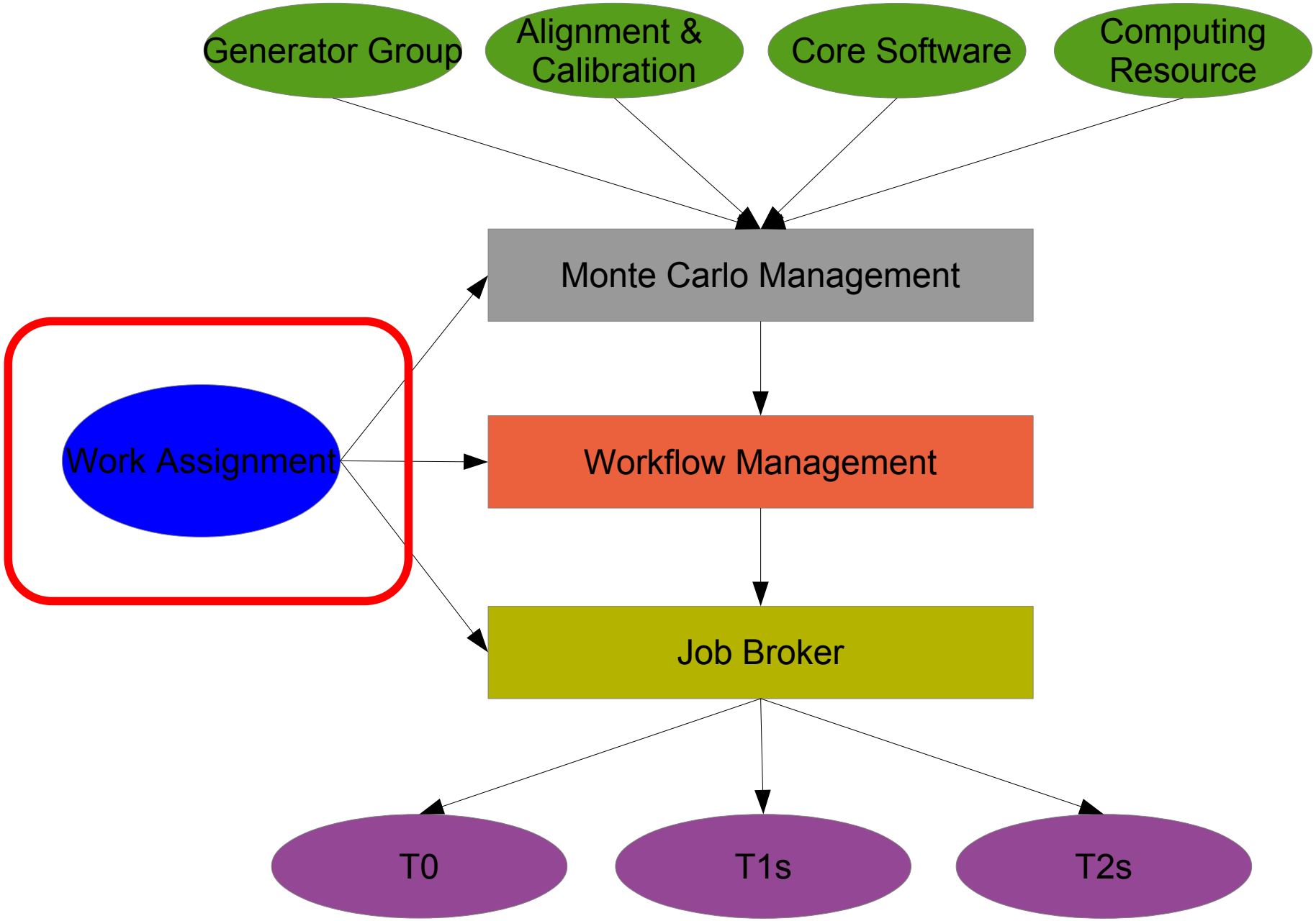
- ~30 schedds for production and analysis with redundancy
  - ✓ Record **200k concurrent jobs**
  - ✓ **Steady >150k job**
- ➔ Constantly working towards scaling up



REFERENCE

Driving Production

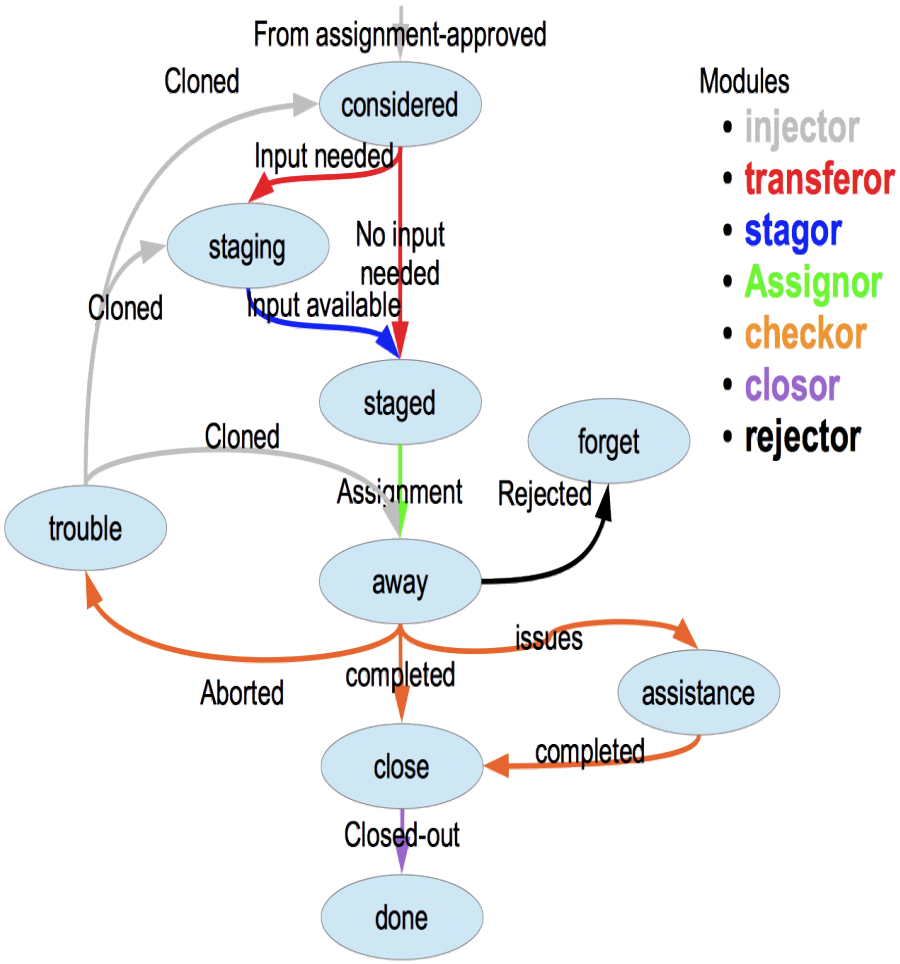



Generator Group
Alignment & Calibration
Core Software
Computing Resource
Monte Carlo Management
Work Assignment
Workflow Management
Job Broker
T0
T1s
T2s

Production State Transition

- **Considered** : received from submission tool
- **Staging** : primary of secondary input is being placed
- **Staged** : all inputs are in place
- **Away** : submitted to htcondor
- **Close** : is fully ready for delivery
- **Done** : delivered
  - Most workflow go through **untouched automatically**

- **Trouble** : the workflow had to be removed and replaced
- **Forget** : the workflow is too much trouble and is just removed
- **Assistance** : goes
  - That's where trouble begins (see in later slide)



Unified® state diagram

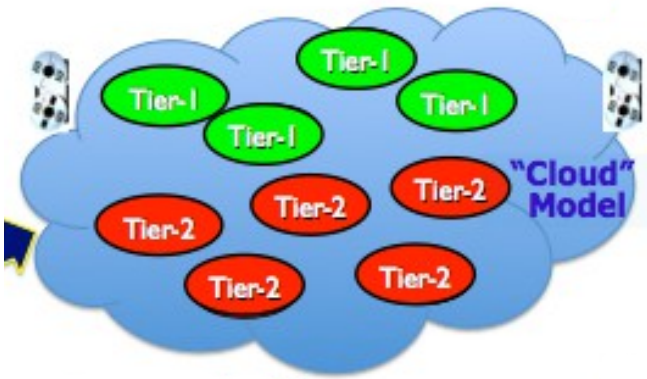
Input Data Placement

- Heterogeneous size/event in primary input and time/event
  - Usually anti-correlated : constant size/time.
  - **Automated transfers** for 1-3 copies over the sites
- Data might be held by someone else
  - **Re-use existing copies** if possible
- Disk space is handled by DDM/Dynamo (see Yutaro's talk)
  - **80% of the allowed quota** is used as operation quota for placing input, leaving enough room for growing output datasets
- Not all workflows can go run everywhere
  - **Pre-matching job/resource** to decide destination according to pledge CPU resource and within quota
- The more sites the better for load sharing
  - Input are **split in 4T chunk** and distributed
- Simulation of LHC event overlay requires event mixing
  - **Secondary inputs are positioned automatically** according to adopted strategy
- Transfers are subject to storage and network issues
  - Stuck-ness of transfers are monitored, solved or by-passed with starting with less than optimal copies

Monitoring of used space, quota and output buffer

Work Assignment

- Not all workflows can go run everywhere
  - **Pre-matching job/resource** to decide
- Simulation of LHC event overlay requires event mixing
  - Standard mixing with **high-read restricts the list of sites**
  - Pre-mixing with **lower-read read over the network (xrootd)**
- The more sites the better for load sharing
  - Use **all sites that hold part of the input** are candidates
- Input dataset can be too large, and still need to be processed in many places
  - Setup **reading the primary over xrootd** to sites holding full copies and their **WAN neighbors**
- Diversity of workflow and complexity of submission
  - Some **job splitting tweaks** are performed upon rules
- Some resource might get available temporarily
  - Flexibility to add **specific assignment rules**
- DDM/Dynamo is managing the disk space for production
  - All input and outputs are **locked from deletion**

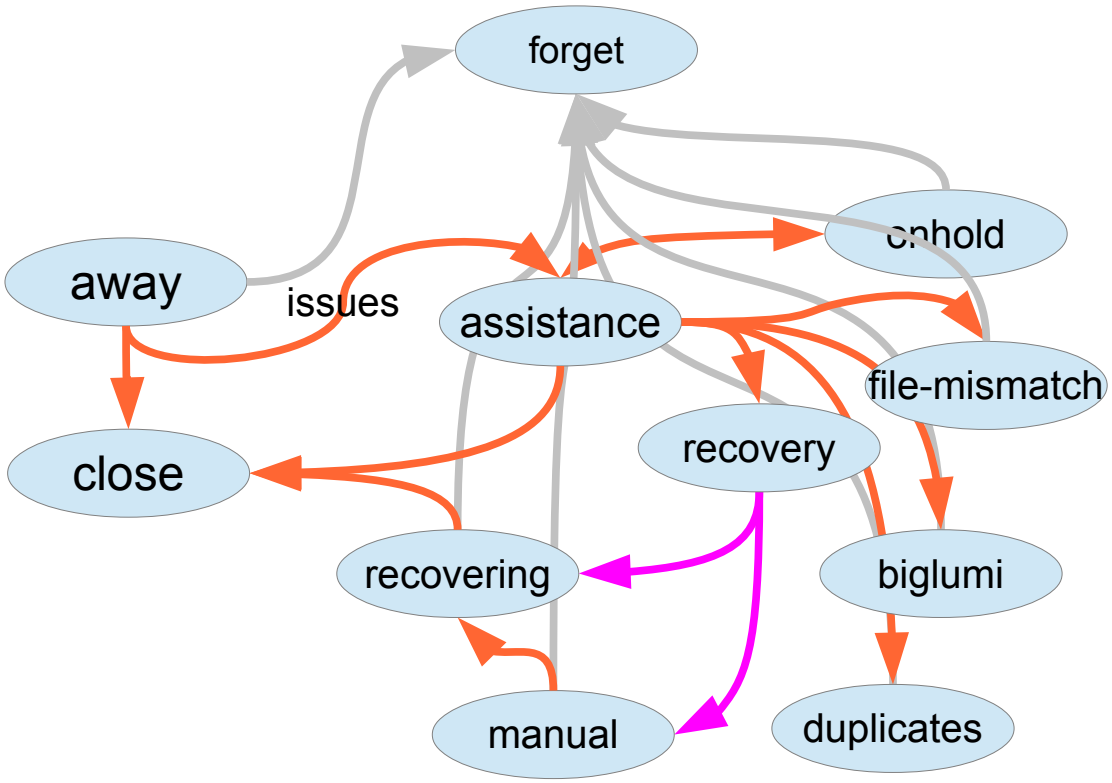


Not much difference between tiers within a cloud model

Where Trouble Begins

- **Assistance** : some level of scrutiny from operator is required
- **onhold** : decision is taken to hold on until further notice (issue to be understood)
- **recovery** : inline for automatic recovery of missing statistics
- **biglumi** : big lumi-section size (production artifact in simulation)
- **duplicates** : a lumi-section are appearing in multiple files
- **File mismatch** : a file mismatch appeared in the book keeping system
- **manual** : requires visual inspection from operator

  - **Partly automated**
  - Issues **efficiently reported**
  - Error collecting and analysis **towards automation** of decision



Unified® state sub-diagram

While a Workflow is “Away” ...

- Site might come out of production status because of schedule intervention, emergency shutdown, intermittent failures, ... (see sites monitoring)
- Workload backlog might develop on local site queue
  - ✓ **Mechanism to overflow** to neighboring site
    - ➔ Quicken delivery with reliable remote read
  - ✓ **Reposition blocks** of data accordingly
    - ➔ Can be used to divert work to resource becoming available

- Jobs requirement are just estimation from limited test-run
  - ✓ Job memory requirement is edited when possible to values observed in running over the grid
  - ✓ Job runtime requirement can be edited
  - ➔ **Better partitioning of resource** into job slots

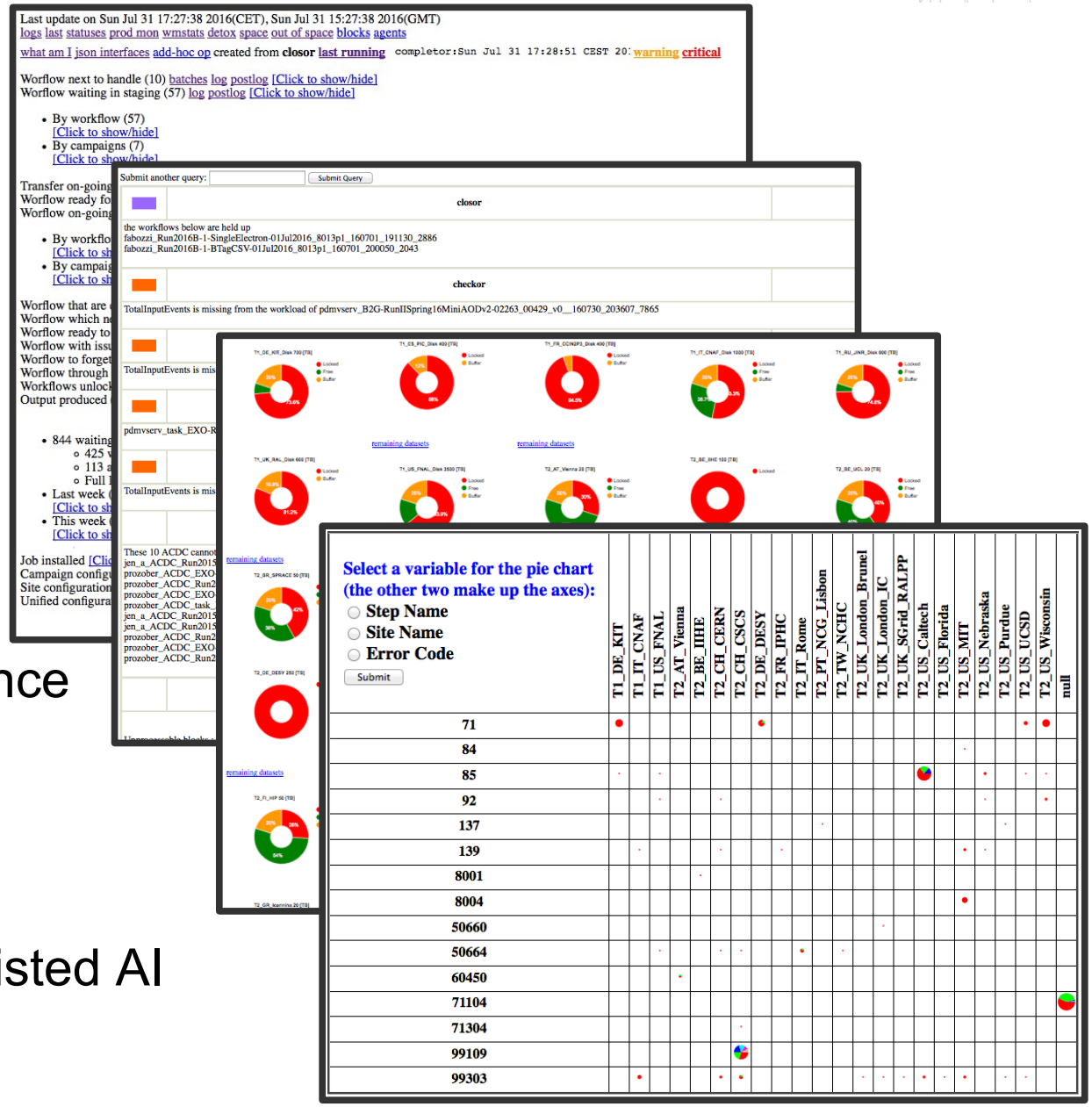
- Shorten workflow processing above agreed completion fraction and running time

➔ Working towards much **more flexibility**, using a more granular data-driven processing strategy

Operation Monitoring

- Overview of work at each level
  - ✓ Provide links to all services
- Logging **heart beats**
  - ✓ Dashboard of **critical items**
  - ✓ Single **workflow history**

- Expose information relevant to other services in json
- Production disk space at a glance
- **Notification** to requesters
  - ✓ Log redundancy
- Display all relevant errors
  - ✓ **Guidance** to operators
  - ➔ Working towards human-assisted AI operation



Outlooks

- Towards even more dynamic job/data placements for load balancing
- Incorporate more opportunistic resources
- Towards network-aware workload balancing
- More automation in dealing with errors, over the sites, over jobs, ...
- Dreaming of AI-assisted computing operation

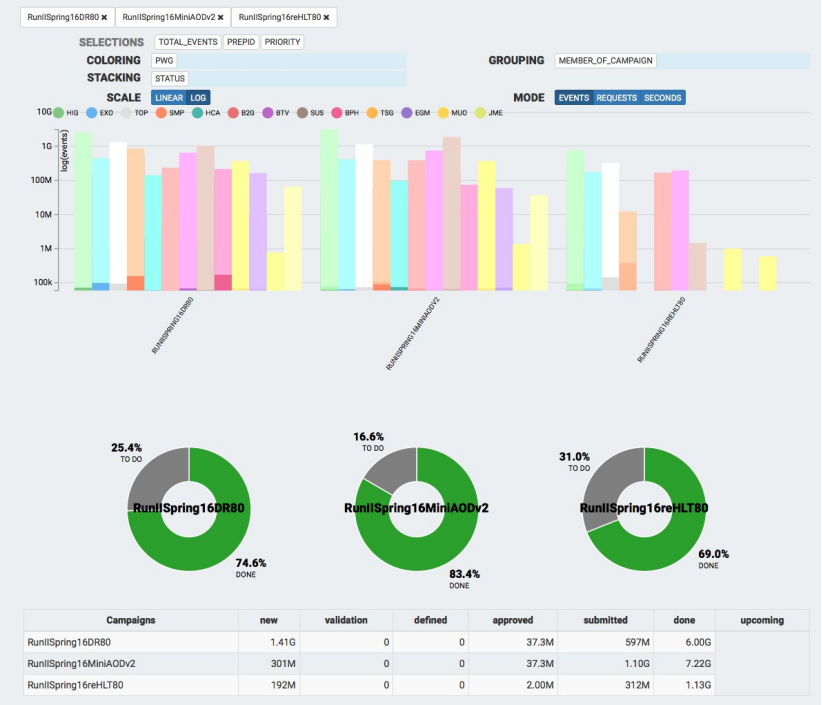
SUMMARY

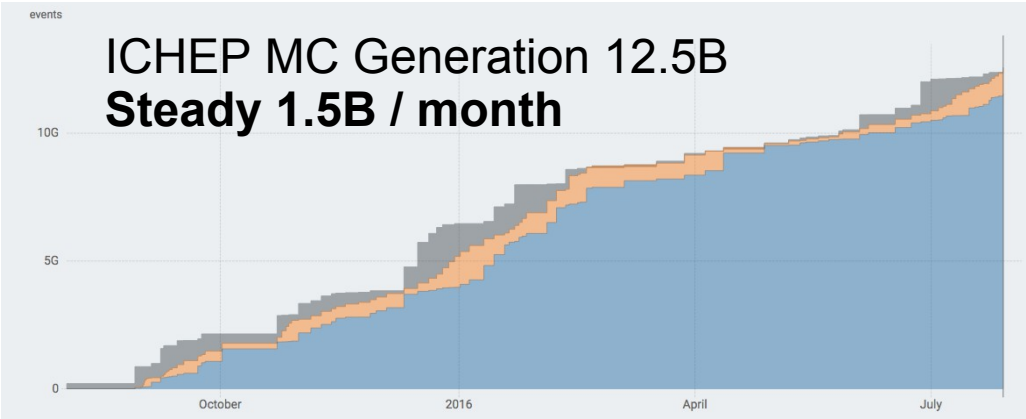
- Large scale production and reprocessing for RunII
- Automated operation helps improving through-put
- Complex work assignment helps reaching more resource
- Dynamic work reallocation helps reducing backlog
- Constantly working on improvements

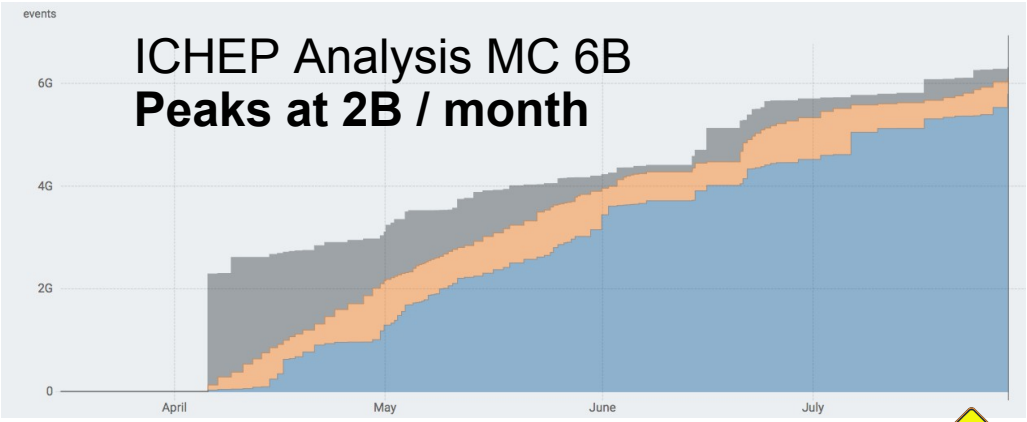
Backup Slides

Sample Monitoring

- **P**roduction **M**onitoring **P**latform (pMp)
- Display current statuses of campaigns
- Track **evolution of single requests** and aggregates several ways
- Help guide the user waiting for samples
- Allows for **production planning**



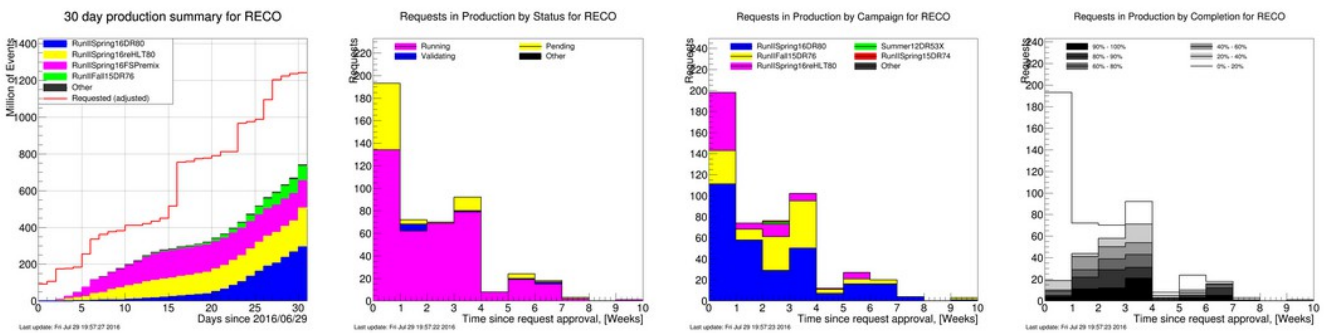


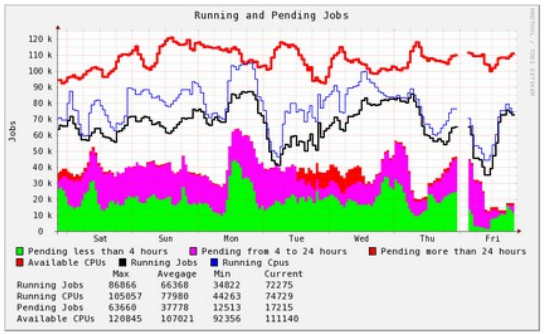


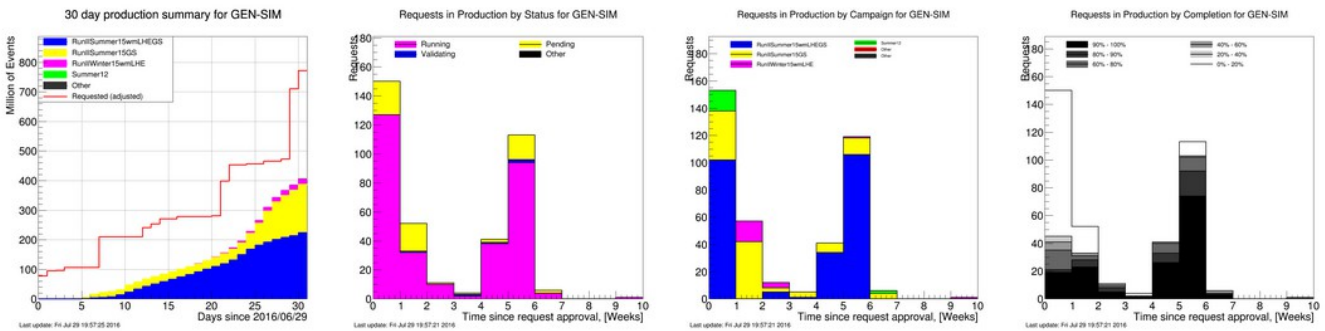
REFERENCE

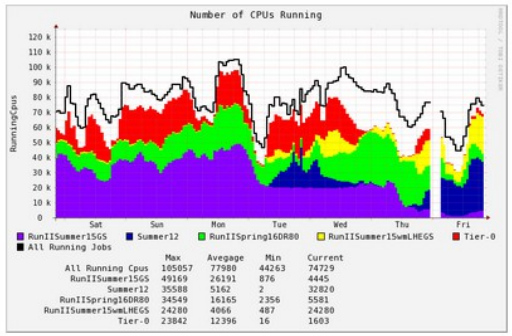
Production Monitoring

- Amount of work left for production at a glance
- Monitors overall resource utilization
- Identifies tails in production
- Aggregate information from several services
- ✓ Average 2000 datasets released per week
- ✓ **Peak 5000 analysis datasets** per week









REFERENCE

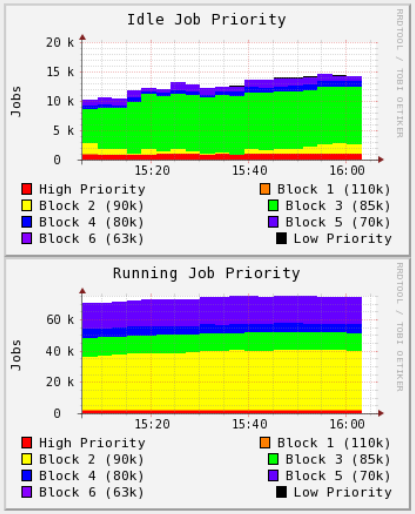
Job Monitoring

- Aggregate and present information from **HTCondor and Glideinwms**
  - ✓ Number of **CPU and jobs per task**, per workload, per site, ...
  - ✓ **Status of sites** with respect to HTCondor
  - ✓ Show the **load on the schedd**
  - ✓ Job **production/analysis share** at sites
- **Feedback loop** on how sites, tasks, and jobs are performing

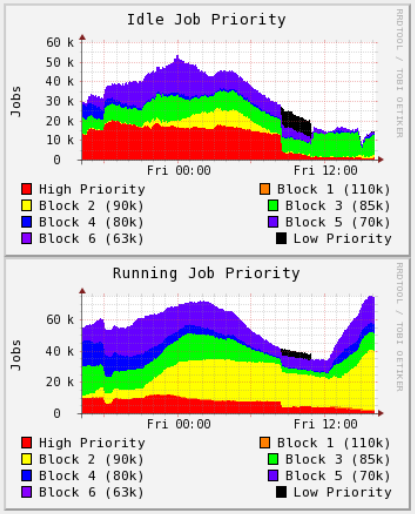
➔ Working on using more of the feedback loop for **processing optimization**

Priority Type	Running	Idle
High Priority	1832	1022
Block 1 (110k)	663	0
Block 2 (90k)	37748	1674
Block 3 (85k)	11187	9762
Block 4 (80k)	5452	934
Block 5 (70k)	17412	788
Low Priority	0	6

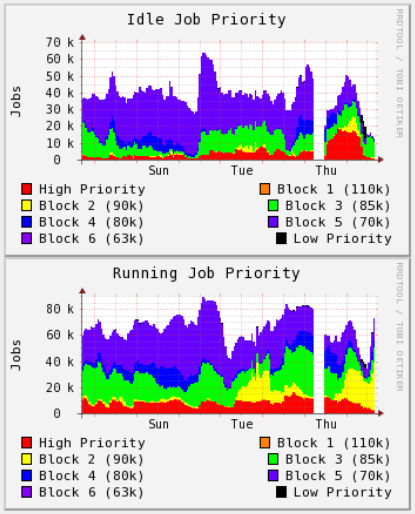
Last Hour



Last Day



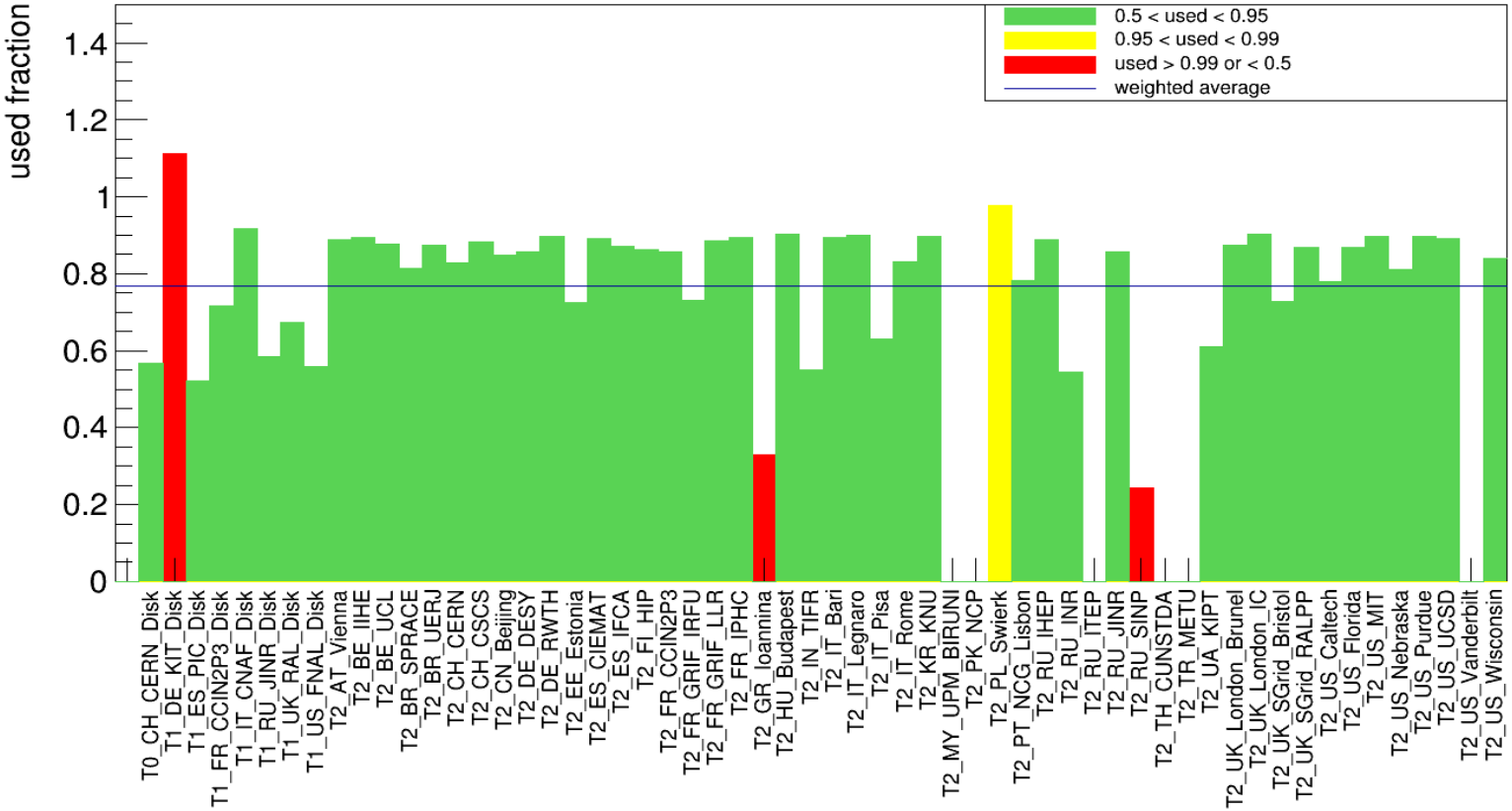
Last Week



REFERENCE

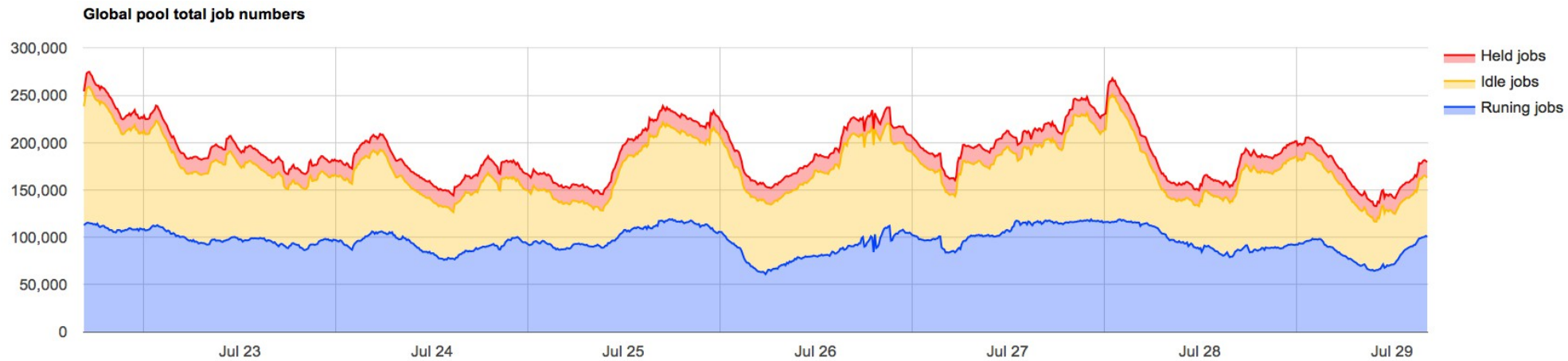
Storage Control

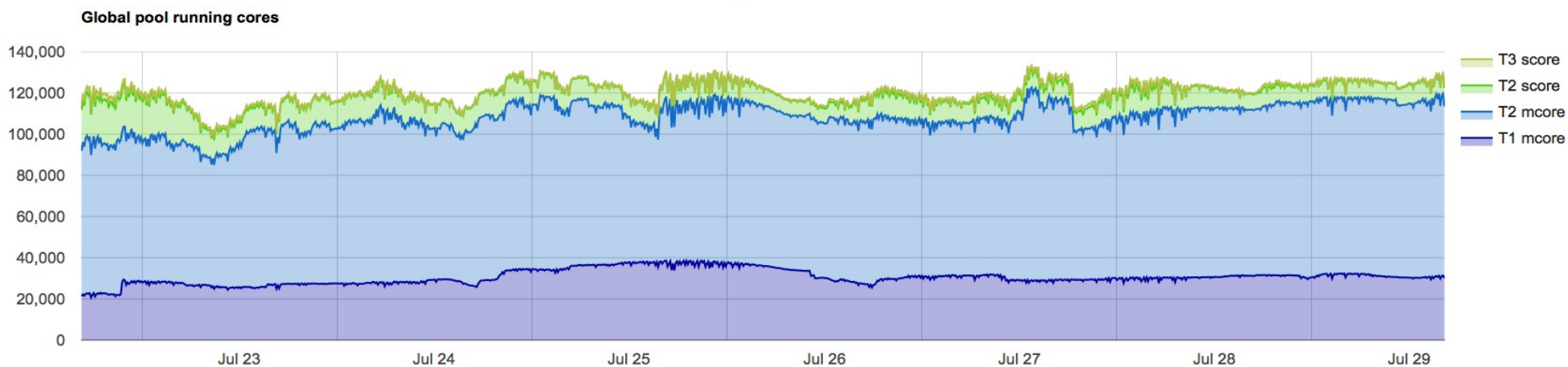
- Available tape space monitored
  - ✓ **Fair-share distribution** to long term storage
- Disk space managed with virtual quota for production and analysis
  - ✓ Automatic transfer and deletion
- ➔ Developing production strategy with a **smaller disk footprint**



REFERENCE

Resource Monitoring





- Steady 100k jobs running for CMS (production and analysis)
- Large contributions from T2
- Large fraction of multi-core pilots
- Spot trend in resource utilization

Sites Availability

- Aggregate **live information about sites**
  - ✓ Site Availability Monitor (SAM) compute and storage services
  - ✓ Hammer Cloud (HC) ability to run jobs
  - ✓ Data Transfer (PhEDEx) transfer links
- Determine the site ready-ness

➔ Working towards **more dynamic and specific** site status evaluation