Computing Model and Operational Status

Gustaaf Brooijmans

Run II Computing Review, September 14th, 2004

- Status in a nutshell
- Data Flow: processing steps
- Remote farms: MC and reprocessing
- (Important) loose ends: databases, disk and tape, operations
- Hardware tools: status and plans
- Software tools: status and plans
- Organization
- Conclusions

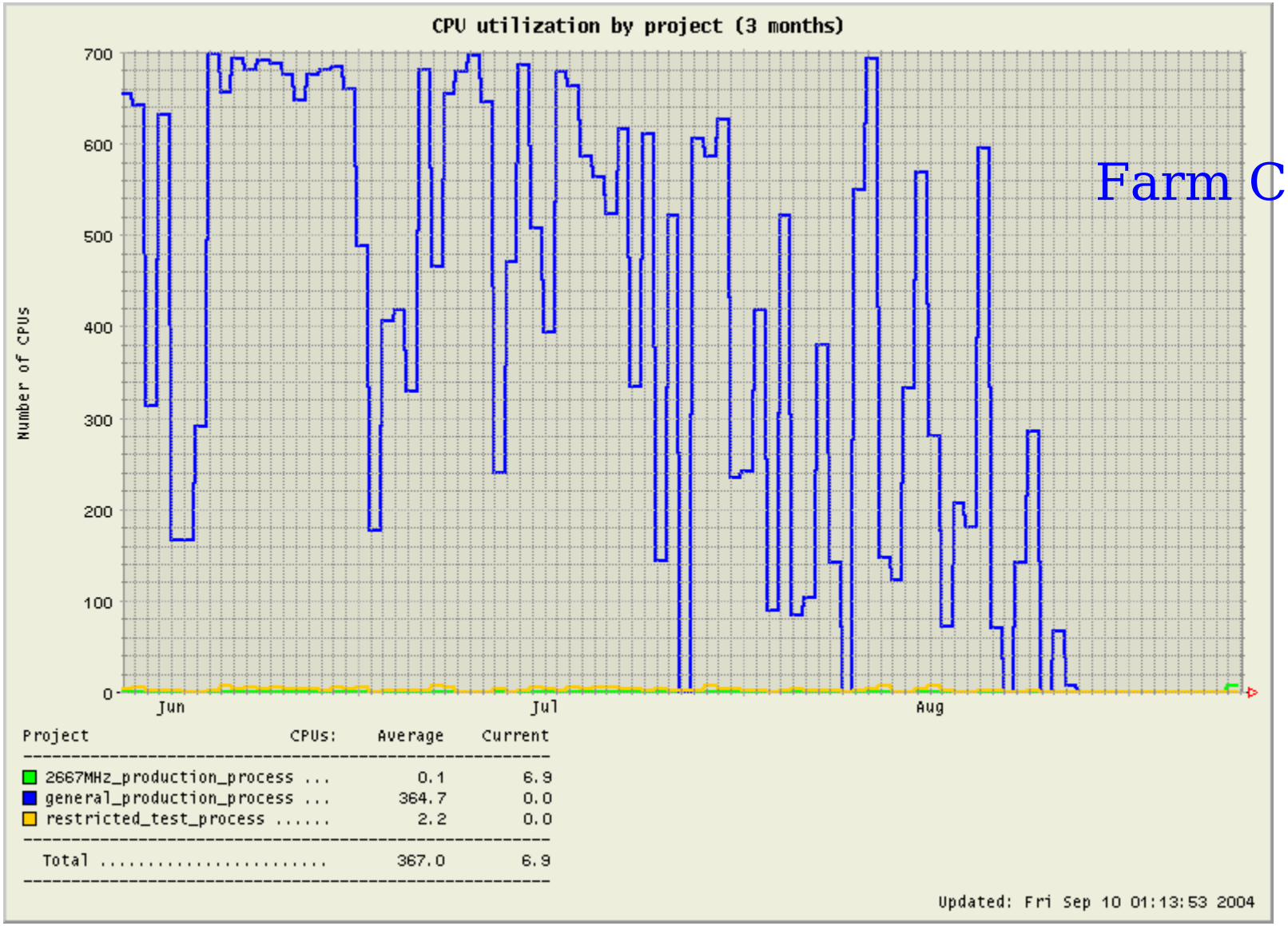
Current Status in a Nutshell

- Reconstruction is keeping up with datataking
- Datahandling is performing very well
- GEANT Monte-Carlo production has reached 1M events/week, all remote
- Reprocessing data remotely
- Analysis CPU power is adequate – expanding
- Databases are satisfying most of our needs
- Globally, we're doing OK

Data Flow

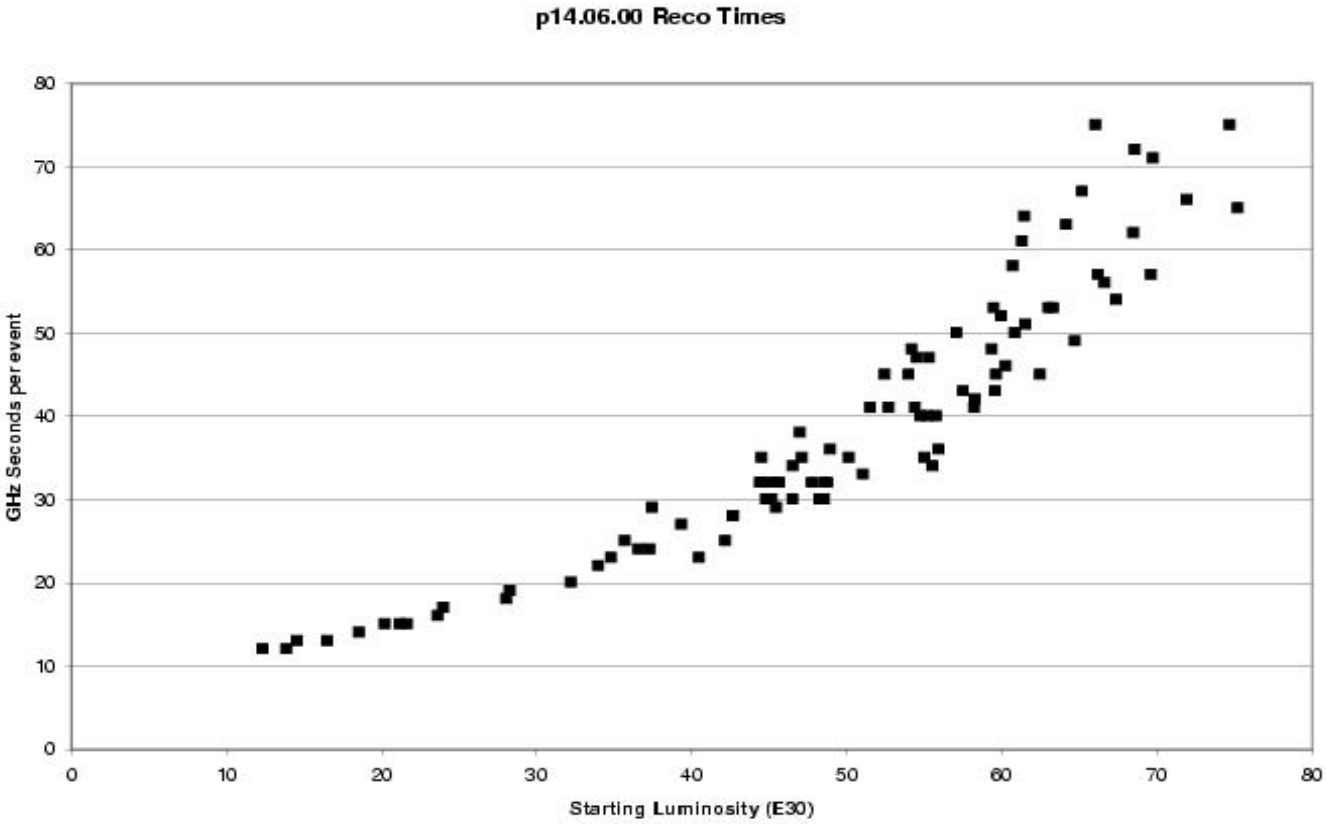
- Data acquisition (raw data in evpack format)
  - Currently limited to ~50 Hz L3 accept rate, but hardware limitation will be increased to ~100 Hz this Fall (with further increase possible at minimal cost: ~$5k/100 Hz)
- Reconstruction (DST/tmb in evpack format)
- Fixing (tmb in evpack format)
  - Correct for problems discovered after reconstruction, re-run high-level algorithms
- Skimming (tmb in evpack format)
  - Event streaming based on reconstructed physics objects
- Analysis (in: tmb or root; out: root histograms)

Reconstruction



Farm CPU

Concern



Laurent will talk about this …

Fixing and Skimming

- Both performed centrally by the Common Samples Group (CSG)
  - Some problems in data quality are only (can only be) discovered after a significant sample has been processed
  - Some can be fixed offline, only require re-running of high-level algorithms (i.e. fast) -> “ Fixing”
- Skimming is event selection based on physics object
  - Basic samples for all analyses

Analysis Format

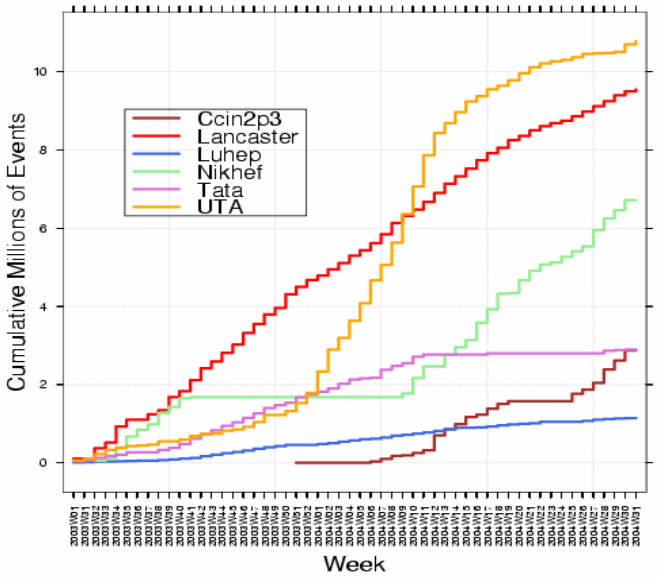
- Output format from reco/skims is “ tmb” (evpack)
- Physics plots are made using root
- Different groups have been using different approaches to this
  - No intermediate format (i.e. work directly on tmb)
    - Painful due to long linking times, recently introduced dynamic linking to improve this significantly
  - Intermediate root tree produced by physics groups or individuals which gets distilled further by end-user
    - Various incarnations of this -> strain on resources (cpu - convert all data, disk – store it, people – run the jobs)

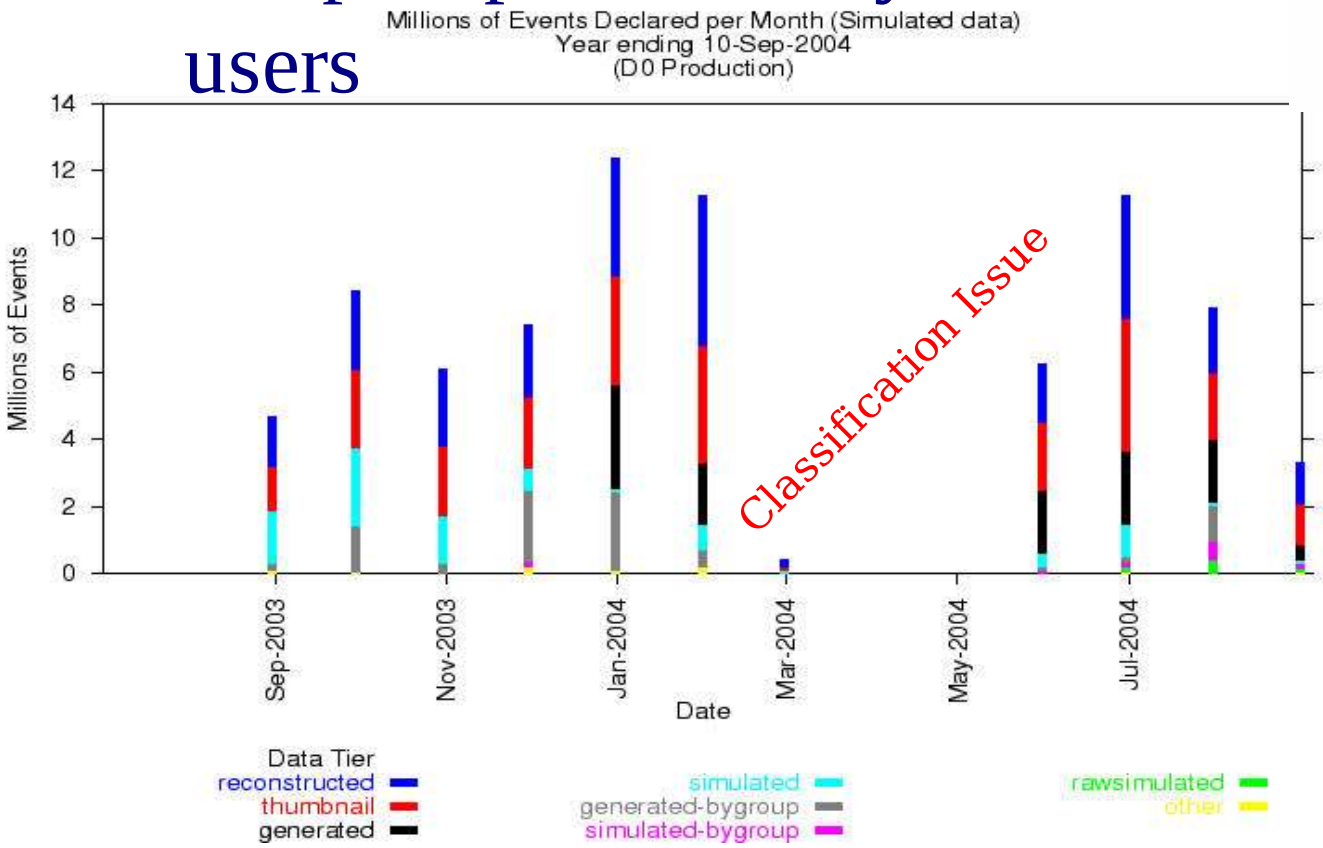
Common Root Format

- Committee formed to evaluate options
  - They recommend creating a general root tree containing the full thumbnail info
    - Would be centrally produced
- Working group now formed to implement this (start from existing “ TMBTree”)
  - Expect them to finish by the end of the year
- Expect 2 types of analysis then: tmb- and common root tree-based

Monte-Carlo

- All MC is produced offsite, with the exception of small samples produced by individual users





Have reached 1 Mevts/week

(conversion to JIM in March - now, only small perturbations)

37 Mevts in FY04

Remote Data Processing

- When make significant improvements to the reconstruction, want to reprocess old data
  - For p14 reprocessing (Fall '03), 100 M events were processed from DST offsite (out of ~500 M)
  - For p17 reprocessing (anticipated in Winter '04), expect to process bulk of data remotely (total volume ~1 B events), this time from raw (!)
  - See Mike's talk
- These represent important steps towards our future computing model

Databases

- Calibration:
  - With p17, all systems (except muon) get calibration from DB during reconstruction
    - Cache for Fermilab farm (600 d0reco processes)
    - Proxy servers for remote reconstruction
- Luminosity
  - Going from ASCII files to DB, migration complete by end of CY04
- Trigger
  - Complexity reflects complexity of trigger system and its continuous development, help very much needed and on the way (from CD)

Disk and Tape

- Tape access has been a bottleneck a few times in the past year, mainly due to inadequate SAM disk cache
  - Current system's robustness "saved" us so far
  - Help on the way:
    - Adding 10 LTOII drives, will probably activate 2nd arm in ADIC robot
    - Very significant expansion of disk resources, both SAM cache and project space for final samples
  - Starting tests of dCache

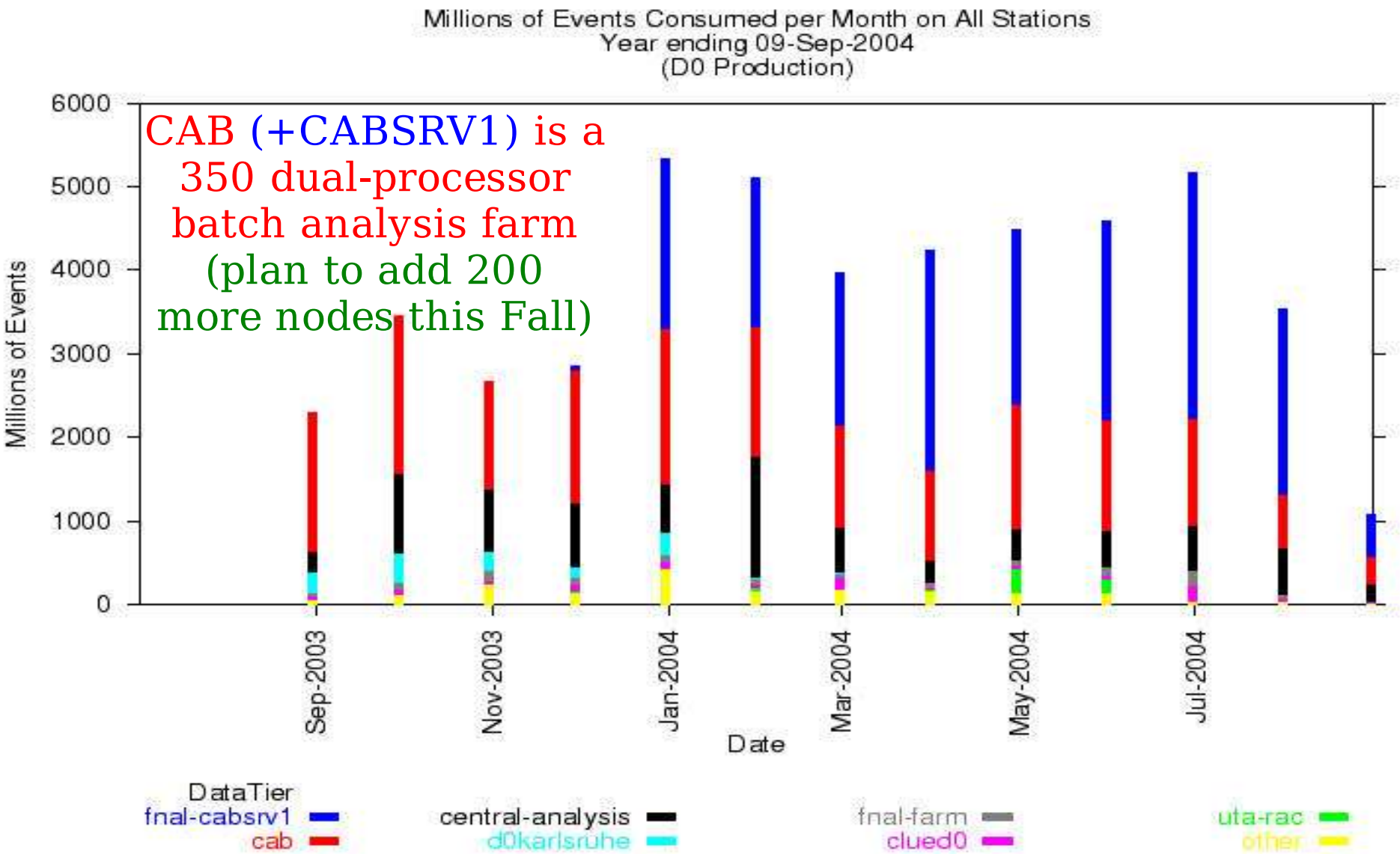
Operations

- Very successful "SAM shifts" model
  - Remote collaborators take these instead of control room shifts (like physics monitoring shifts)
  - Form first line of defense
    - Can answer questions, restart some services
    - Work underway to give them better diagnostics (currently much is done by "grep" on logfiles)
- Maintenance tasks that require downtime done on "first Tuesday of the month"
  - People are prepared
  - Downtime rarely exceeds expectation

Current Cluster Usage

- Reco farm for reconstruction (+ fixing)
- Remote farms for MC production and data reprocessing
- CAB (batch analysis farm) for processing of farm output
  - Fixing and skimming
  - Direct analysis
  - Group-organized translation to root format
  - CPU-intensive analysis jobs
- Clued0 (desktop cluster) for interactive work, user analysis (has batch, SAM access, releases, ...)
- D0mino (Origin 2k): trying to find out what remains...

Statistics


Millions of Events Consumed per Month on All Stations
Year ending 09-Sep-2004
(D0 Production)
CAB (+CABSRV1) is a 350 dual-processor batch analysis farm (plan to add 200 more nodes this Fall)
Millions of Events
6000
5000
4000
3000
2000
1000
0
Sep-2003
Nov-2003
Jan-2004
Mar-2004
May-2004
Jul-2004
Date
DataTier
fnal-cabsrv1
cab
central-analysis
d0karlsruhe
fnal-farm
clued0
uta-rac
other

Cluster Evolution

- Continuously upgrading reco farm and CAB to keep up with increasing dataset size
- Clued0 has presumably reached steady size, machines get replaced by institutions
  - Number of sysadmins is permanently low, requires some twisting of arms
- D0mino retirement planned for end of CY04, will make it “ read-only” at end of October
  - Interactive linux cluster at FCC ready for users

Software Evolution

- Experiment (and computing) is in operations phase
  - Unstable situation however
    - Expect significant increase in instantaneous luminosity (x3)
    - Simultaneous with migration of resources to the LHC (both equipment and personpower)
  - We are therefore evolving our model
    - Make all our computing tasks GRID-compatible
    - Reduce maintenance load on tools

We're Not Alone ...

- CD is pursuing similar path
  - Run II Department
  - Planned consolidation of farms
- LHC (and other) experiments are looking at our experience, evaluating some of our tools, with development now common projects:
  - SAMGrid
  - Runjob

← Risk: now our colleagues have to follow through!

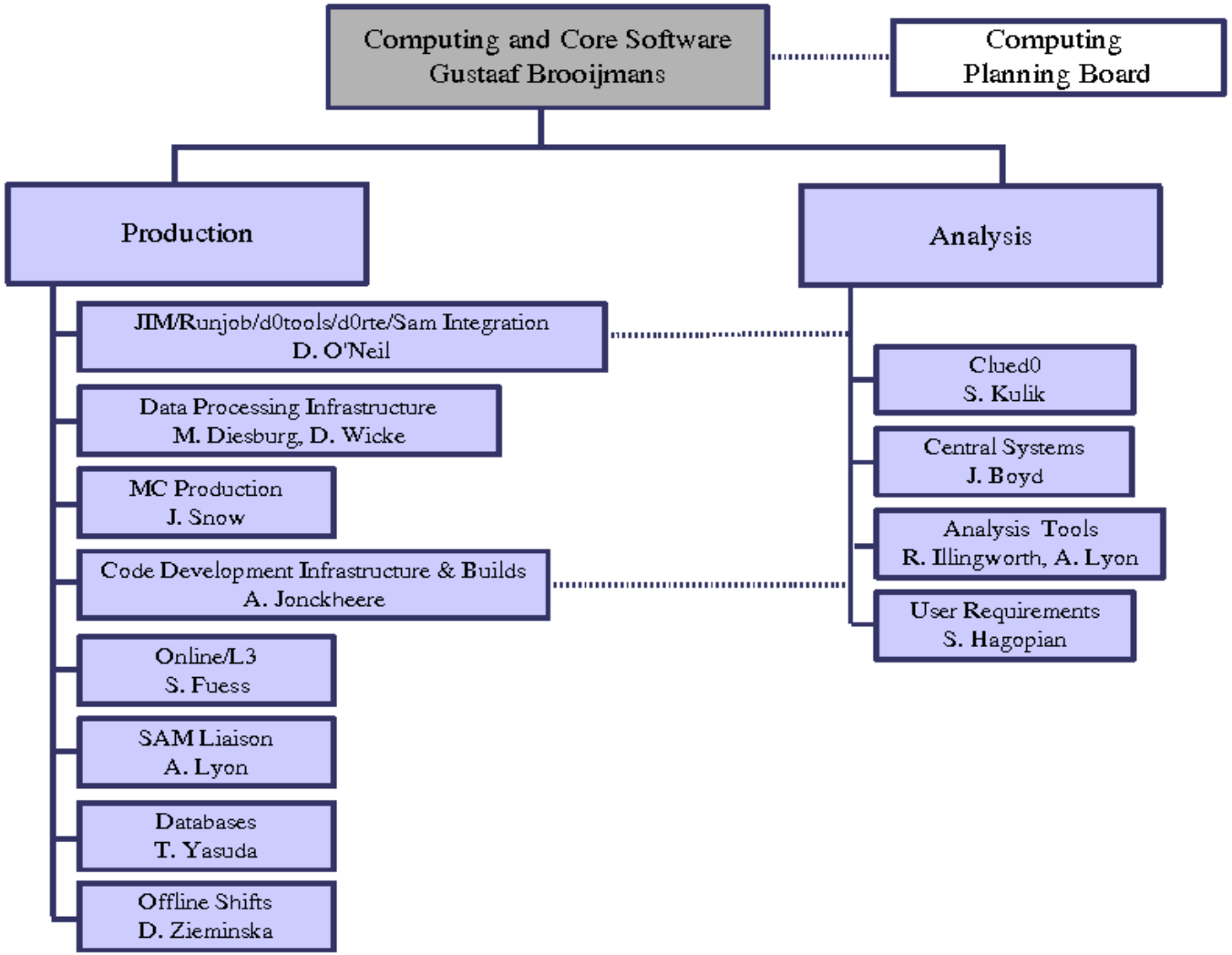
Getting There

- First step was p14 remote processing
  - Successful, a lot learned
- Next step: p17
  - Process from raw data, requires DB access
  - Increased use of GRID tools, reduce manual intervention
- After that, “ export” other common tasks (like fixing), then finally user jobs

Tool Interfaces

- Currently using different tools for different purposes:
  - (mc_)runjob for remote production and user MC production
  - d0tools: (popular) user interface for job submission
  - SAM(Grid) for data handling, accounting
  - d0rte for specification of run-time needs
  - OS compatibility to run on different linux flavors
- Integration of these important step towards “simple” GRID job submission

Organization



Conclusion

- Computing is *not* a bottleneck in data analysis, want to keep it that way
- Potential issues being addressed
  - Short term:
    - Disk space expansion, increase in analysis cpu, data format consolidation, reconstruction cpu consumption
  - Long term:
    - Common set of tools for production and analysis, pursue overall GRID-compatibility, develop tools in collaboration with others
- Globally doing rather well