Working towards the Computing Model for CMS

David Stickland

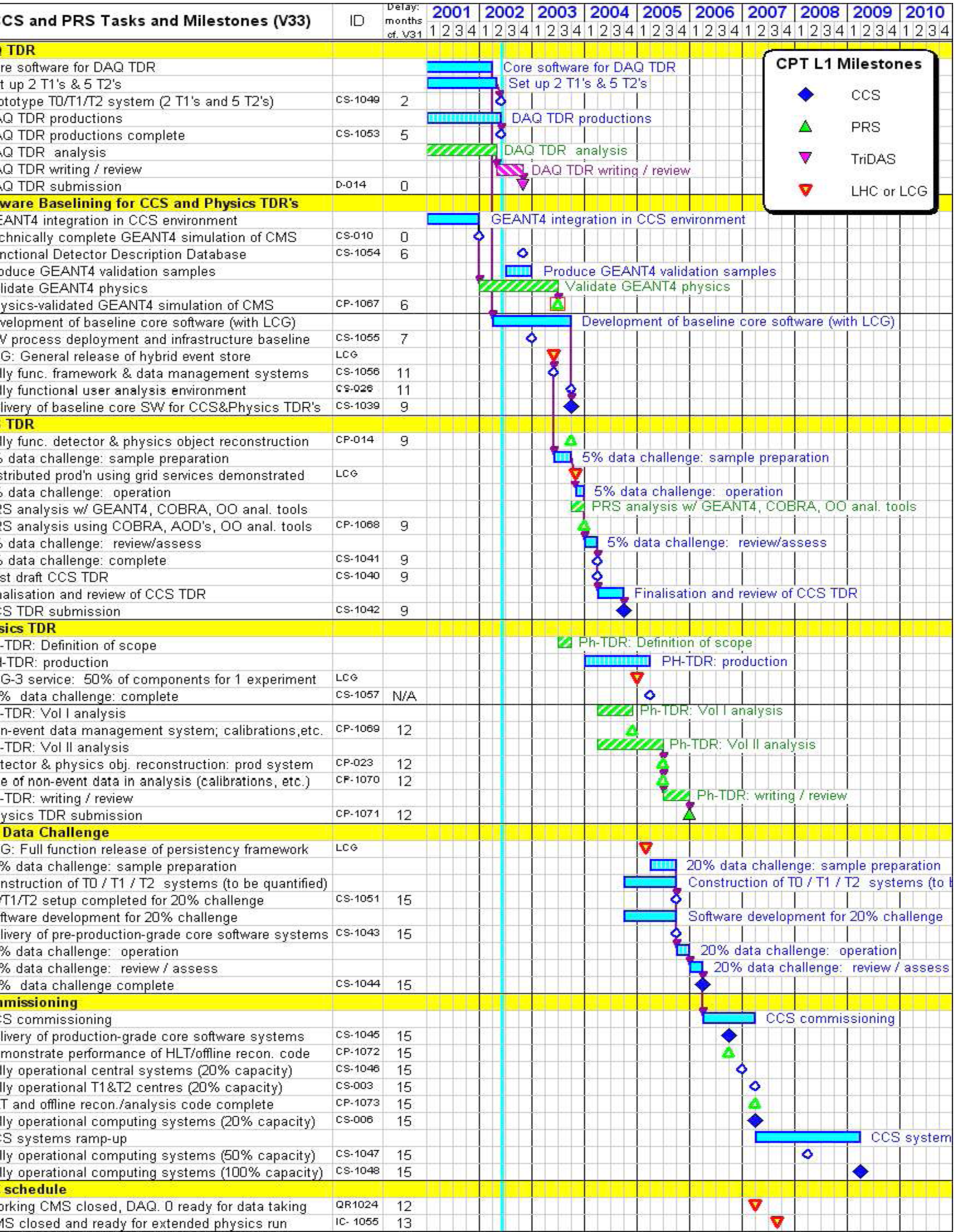
CMS Core Software and Computing

DPS/ LCG Review Nov 2003

Outline

- Computing TDR
- Computing Model
- Data Challenges
- CMS and LCG

Phases of Computing Activities



CS and PRS Tasks and Milestones (V33)	ID	Delay: months cf. V31
TDR		
re software for DAQ TDR		
t up 2 T1's & 5 T2's		
ototype T0/T1/T2 system (2 T1's and 5 T2's)	CS-1049	2
AQ TDR productions		
AQ TDR productions complete	CS-1053	5
AQ TDR analysis		
AQ TDR writing / review		
AQ TDR submission	D-014	0
ware Baselining for CCS and Physics TDR's		
EANT4 integration in CCS environment		
chnically complete GEANT4 simulation of CMS	CS-010	0
nctional Detector Description Database	CS-1054	6
oduce GEANT4 validation samples		
lidate GEANT4 physics		
ysics-validated GEANT4 simulation of CMS	CP-1067	6
velopment of baseline core software (with LCG)		
V process deployment and infrastructure baseline	CS-1055	7
G: General release of hybrid event store	LCG	
lly func. framework & data management systems	CS-1056	11
lly functional user analysis environment	CS-026	11
livery of baseline core SW for CCS&Physics TDR's	CS-1039	9
TDR		
lly func. detector & physics object reconstruction	CP-014	9
data challenge: sample preparation		
stributed prod'n using grid services demonstrated	LCG	
data challenge: operation		
S analysis w/ GEANT4, COBRA, OO anal. tools		
S analysis using COBRA, AOD's, OO anal. tools	CP-1068	9
data challenge: review/assess		
data challenge: complete	CS-1041	9
st draft CCS TDR	CS-1040	9
alisation and review of CCS TDR		
S TDR submission	CS-1042	9
sics TDR		
-TDR: Definition of scope		
-TDR: production		
G-3 service: 50% of components for 1 experiment	LCG	
% data challenge: complete	CS-1057	N/A
-TDR: Vol I analysis		
n-event data management system; calibrations,etc.	CP-1069	12
-TDR: Vol II analysis		
tector & physics obj. reconstruction: prod system	CP-023	12
e of non-event data in analysis (calibrations, etc.)	CP-1070	12
-TDR: writing / review		
ysics TDR submission	CP-1071	12
Data Challenge		
G: Full function release of persistency framework	LCG	
% data challenge: sample preparation		
nstruction of T0 / T1 / T2 systems (to be quantified)		
T1/T2 setup completed for 20% challenge	CS-1051	15
ftware development for 20% challenge		
livery of pre-production-grade core software systems	CS-1043	15
% data challenge: operation		
% data challenge: review / assess		
% data challenge complete	CS-1044	15
mmissioning		
S commissioning		
livery of production-grade core software systems	CS-1045	15
monstrate performance of HLT/offline recon. code	CP-1072	15
lly operational central systems (20% capacity)	CS-1046	15
lly operational T1&T2 centres (20% capacity)	CS-003	15
T and offline recon./analysis code complete	CP-1073	15
lly operational computing systems (20% capacity)	CS-006	15
S systems ramp-up		
lly operational computing systems (50% capacity)	CS-1047	15
lly operational computing systems (100% capacity)	CS-1048	15
schedule		
orking CMS closed, DAQ. 0 ready for data taking	QR1024	12
MS closed and ready for extended physics run	IC- 1055	13

2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
1 2 3 4
CPT L1 Milestones
CCS
PRS
TriDAS
LHC or LCG
Core software for DAQ TDR
Set up 2 T1's & 5 T2's
DAQ TDR productions
DAQ TDR analysis
DAQ TDR writing / review
GEANT4 integration in CCS environment
Produce GEANT4 validation samples
Validate GEANT4 physics
Development of baseline core software (with LCG)
5% data challenge: sample preparation
5% data challenge: operation
PRS analysis w/ GEANT4, COBRA, OO anal. tools
5% data challenge: review/assess
Finalisation and review of CCS TDR
Ph-TDR: Definition of scope
PH-TDR: production
Ph-TDR: Vol I analysis
Ph-TDR: Vol II analysis
Ph-TDR: writing / review
20% data challenge: sample preparation
Construction of T0 / T1 / T2 systems (to b
Software development for 20% challenge
20% data challenge: operation
20% data challenge: review / assess
CCS commissioning
CCS system

❖Support of HLT studies and the DAQ TDR

- ◆ (to end 2002)

❖Baseline Core Software for Computing and Physics TDRs

- ◆ (to end 2003)

❖"5%" Challenge DC04 and Computing TDR

- ◆ (to end 2004)
  (T0-30 months, LCG-6 months)
- ◆ Required for LCG TDR Mid 2005

❖"10%" Challenge DC05 and Physics TDR

- ◆ (to end 2005) (T0-18 months)

❖"20%" Challenge DC06. Readiness Review

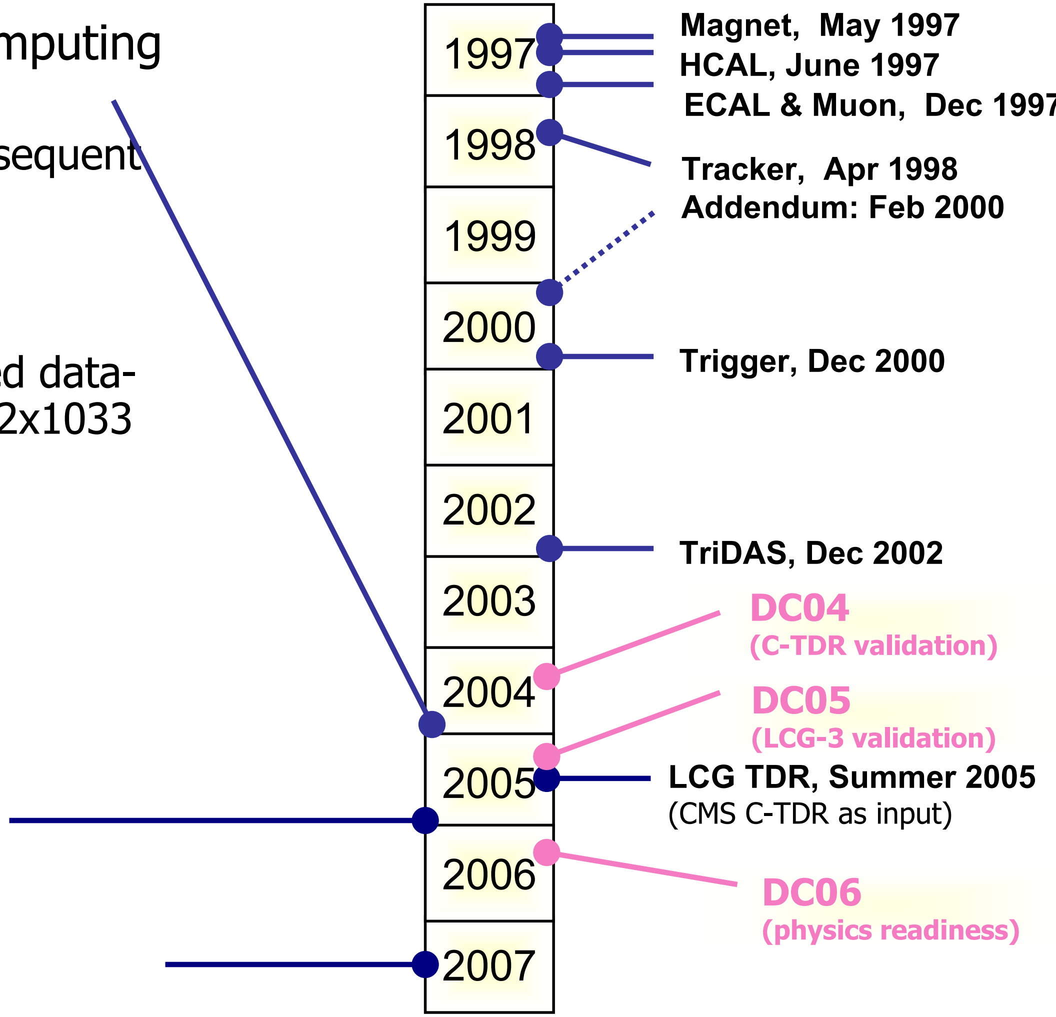
- ◆ (to mid-2006) (T0-1year)

❖Staged Commissioning of Software and Computing Systems

- ◆ (T0 -1 to +2 years)

Computing TDR

- Computing TDR, End 2004
  - Technical specifications of the computing and core software systems
    - for DC06 Data Challenge and subsequent real data taking
  - Includes results from DC04 Data Challenge
    - successfully copes with a sustained data-taking rate equivalent to 25Hz at 2x1033 for a period of 1 month
- Physics TDR, Dec 2005
- CMS Physics, Summer 2007



Magnet, May 1997

HCAL, June 1997

ECAL & Muon, Dec 1997

Tracker, Apr 1998

Addendum: Feb 2000

Trigger, Dec 2000

TriDAS, Dec 2002

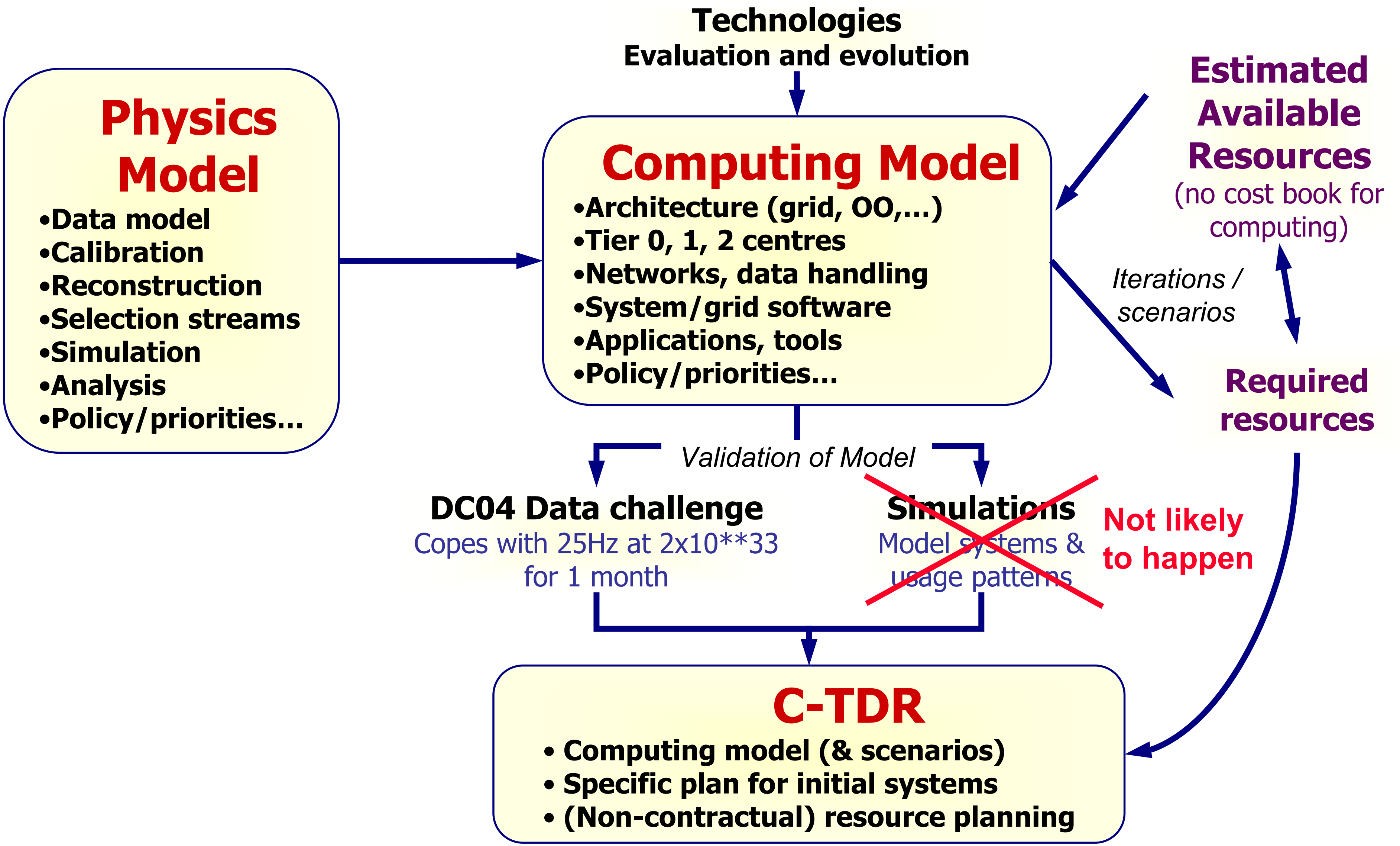
DC04
(C-TDR validation)

DC05
(LCG-3 validation)

LCG TDR, Summer 2005
(CMS C-TDR as input)

DC06
(physics readiness)

Computing TDR Strategy



Technologies
Evaluation and evolution

Physics Model

- Data model
- Calibration
- Reconstruction
- Selection streams
- Simulation
- Analysis
- Policy/priorities…

Computing Model

- Architecture (grid, OO,…)
- Tier 0, 1, 2 centres
- Networks, data handling
- System/grid software
- Applications, tools
- Policy/priorities…

Estimated Available Resources
(no cost book for computing)

Iterations / scenarios

Required resources

Validation of Model

DC04 Data challenge
Copes with 25Hz at 2x10**33 for 1 month

~~**Simulations**~~
~~Model systems & usage patterns~~

Not likely to happen

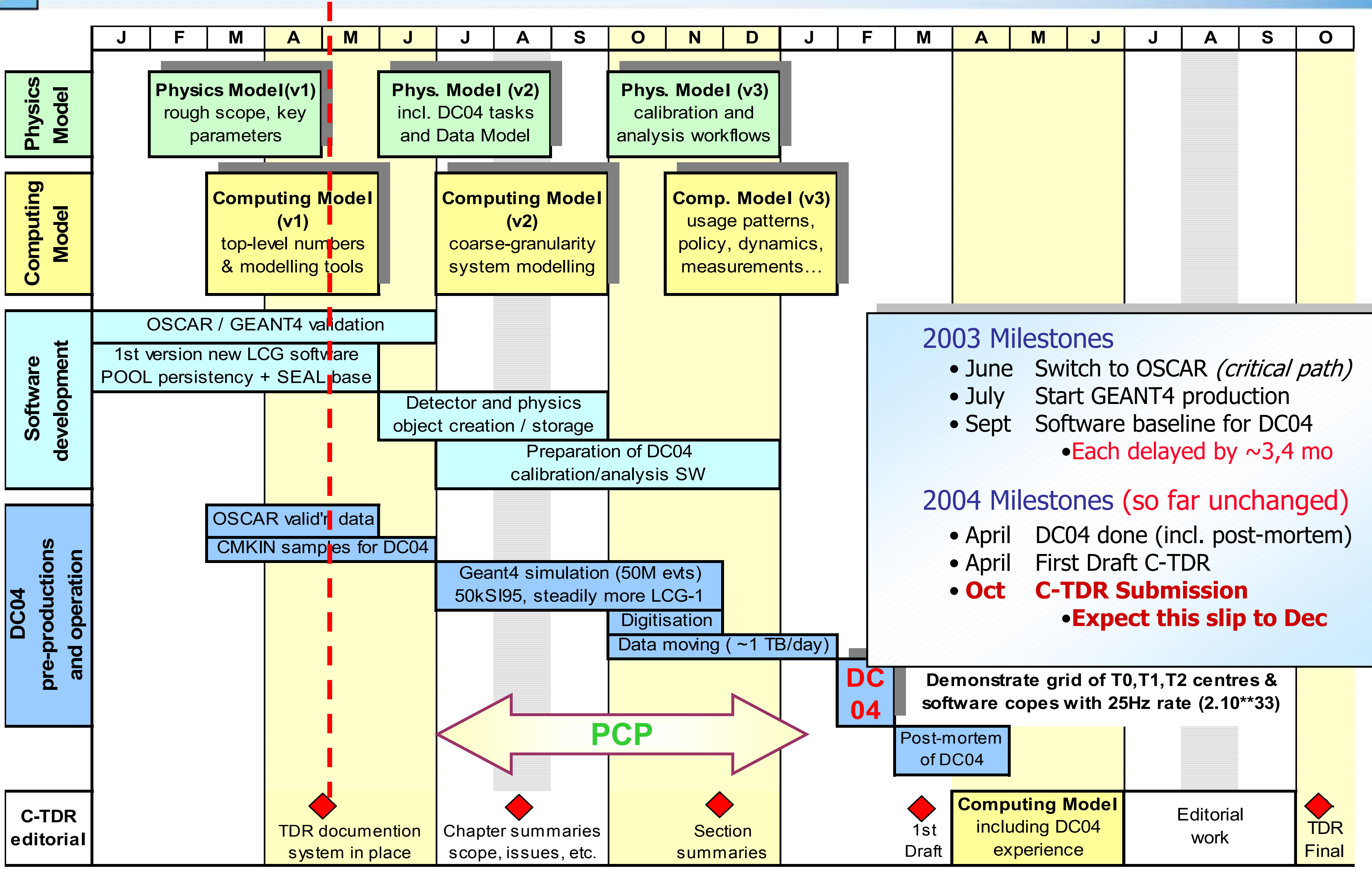
C-TDR

- Computing model (& scenarios)
- Specific plan for initial systems
- (Non-contractual) resource planning

CTDR Status

- First studies of “Physics model” completed
- Key people in experiments very heavily overworked
  - Schedule will be very hard to meet.
- Starting weekly CTDR meeting (Actually this week)
  - CMS focused, but open to anyone interested
  - Develop two or three Strawmen Computing Models
- Expect the CTDR to describe two models
  - One which can be realistically achieved in the remaining time to LHC with the planned scope and manpower of CMS and projects such as the LCG
  - One which represents a much reduced scope and resources (say 50%)
  - In both cases focus on the initial two years of LHC operation

Schedule: PCP, DC04, C-TDR …



J F M A M J J A S O N D J F M A M J J A S O

Physics Model

Physics Model(v1) rough scope, key parameters

Phys. Model (v2) incl. DC04 tasks and Data Model

Phys. Model (v3) calibration and analysis workflows

Computing Model

Computing Model (v1) top-level numbers & modelling tools

Computing Model (v2) coarse-granularity system modelling

Comp. Model (v3) usage patterns, policy, dynamics, measurements…

Software development

OSCAR / GEANT4 validation

1st version new LCG software POOL persistency + SEAL base

Detector and physics object creation / storage

Preparation of DC04 calibration/analysis SW

DC04 pre-production and operation

OSCAR valid'n data

CMKIN samples for DC04

Geant4 simulation (50M evts) 50kSI95, steadily more LCG-1

Digitisation

Data moving (~1 TB/day)

DC 04

Demonstrate grid of T0,T1,T2 centres & software copes with 25Hz rate (2.10**33)

Post-mortem of DC04

PCP

C-TDR editorial

TDR documentation system in place

Chapter summaries scope, issues, etc.

Section summaries

1st Draft

Computing Model including DC04 experience

Editorial work

TDR Final

2003 Milestones

- June Switch to OSCAR *(critical path)*
- July Start GEANT4 production
- Sept Software baseline for DC04
  - Each delayed by ~3,4 mo

2004 Milestones (so far unchanged)

- April DC04 done (incl. post-mortem)
- April First Draft C-TDR
- **Oct C-TDR Submission**
  - **Expect this slip to Dec**

Strawman Model I. Guiding Principles

- Data access is a much harder problem than CPU access.
  - Avoid bottlenecks, distribute data widely, quickly.
- Require a balance between the common a-priori goals of the experiment and the individual goals of its collaborators
  - The experiment must be able to partition resources according to policy
- No dead-time can be introduced to the data acquisition by the offline system
  - All potential points of blockage must have “relief valves” in place.
  - The Tier-0 must keep up in real time with the DAQ. Latencies must be no more than of order 6-8 hours.
- Tier-1 centers are largely resources of the experiment as a whole
  - Data intensive tasks need to run at Tier-1 centers
- Tier-2 centers are focused more at geographic and/or physics groupings
  - It must be possible to replicate modest sized data sets to the Tier-2’s in a timely way.

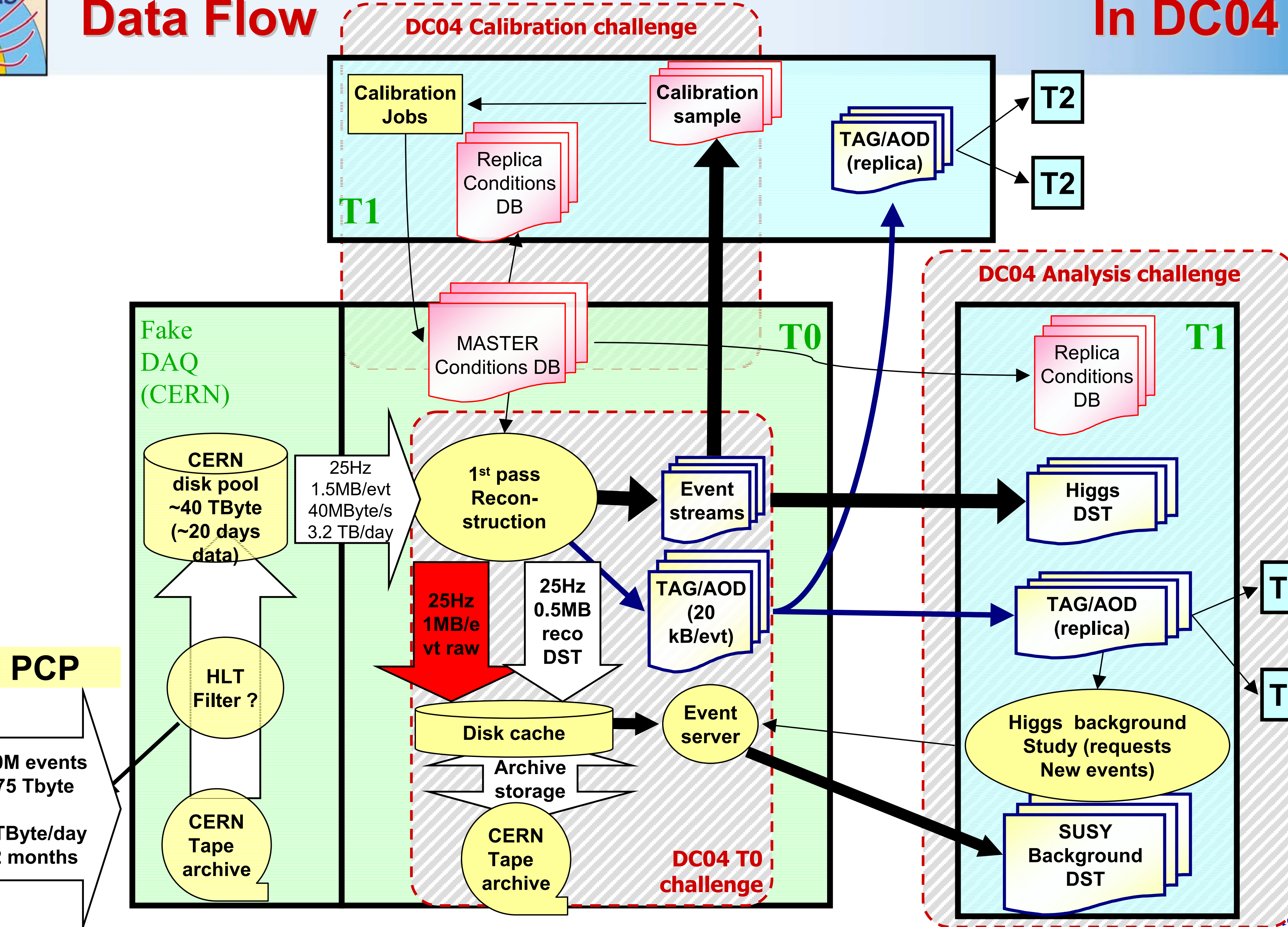
Building a more detailed Strawman Model

1. Raw data is “streamed” from the online system at 100 MB/s.
    1. The streams may not be the final ones
2. Raw data is sent to MSS at the Tier-0
    1. A second copy of the raw data is sent offsite
3. First pass reco of (some of) the raw data at the Tier-0
    1. Some first pass reconstruction may be carried out away from CERN
4. The Tier-0/1 first pass keeps up with the DAQ rate
    1. Tier-0 is available outside the LHC running period for rerunning etc.
5. The DST may be further/differently streamed w.r.t. DAQ Streams.
    1. Some (10%?) event duplication is allowed
6. Calibration “DST’s” are sent to the Tier-1/2 responsible for the processing
7. The full DST is kept at the Tier-0 and at each Tier-1
8. The full TAG (selection data) is stored at the Tier-0-1 and-2
9. Scheduled Analysis passes on DST/TAG data are run at the Tier-1’s
10. Tier-2 centers are the point of access for most user analysis/ physics preparation.
11. …

Extracting Information

- Use a more detailed model to extract more detailed numbers, for example:
  - Summing T0 <>T1 traffic
    - 3Gb/s output traffic from CERN
    - @ 50% efficiency would mean CMS needs 5-6Gb/s at startup at CERN
    - Approximately 1/5 of this at each T1 input, double that to support T2’s
  - 2GB files match well CPU times of jobs
    - 1 file every 20s from CMS
    - 4 hour CPU in reconstruction
    - 600 // jobs running to keep up
    - 100TB 20 day input buffer to T0
  - …

Data Flow In DC04



DC04 Calibration challenge

T1

Calibration Jobs

Calibration sample

Replica Conditions DB

TAG/AOD (replica)

T2

T2

DC04 Analysis challenge

Fake DAQ (CERN)

T0

MASTER Conditions DB

T1

Replica Conditions DB

CERN disk pool ~40 TByte (~20 days data)

25Hz
1.5MB/evt
40MByte/s
3.2 TB/day

1st pass Recon- struction

Event streams

Higgs DST

TAG/AOD (20 kB/evt)

TAG/AOD (replica)

T

T

PCP

HLT Filter ?

M events
75 Tbyte

TByte/day
months

CERN Tape archive

25Hz 1MB/evt raw

25Hz 0.5MB reco DST

Disk cache

Event server

Higgs background Study (requests New events)

Archive storage

CERN Tape archive

DC04 T0 challenge

SUSY Background DST

DC04 Status Today

- Pre-Challenge Phase (MC Gen, Simu, Digitization)
  - Generation/Simulation steps going very well

	Requested	Completed
CMSIM G3	52M	48M
OSCAR G4	16M	0 (But started now)
Not Yet assigned	7M (Probably OSCAR)	

  - N.B. Of this
    - 1.5M with LCG0 (~40kSI2Kmonths)
    - 2.3M with USCMS/MOP (~50kSI2k months)
- Digitization Step getting ready
  - Complicated. May only have about ~30M Digitized by Feb 1
- Final schedule for DC04 could slip by ~1 month (March/April)

DC04 Scales at T0,1,2

Tier-0

- Reconstruction and DST production at CERN
  - 75TB Input Data (25TB Input buffer?)
  - 180kSI2k.month =400 CPU @24 hour operation (@500SI2k/CPU)
  - 25TB Output data
  - 1-2TB/Day Data Distribution from CERN to sum of T1 centers

Tier-1

- Assume all (except CERN) “CMS” Tier-1’s participate
  - CNAF, FNAL, Lyon, Karlsruhe, RAL
- Share the T0 output DST between them (~5-10TB each?)
  - 200GB/day transfer from CERN (per T1)
  - (Possibly stream ~1TB Raw-Data to Lyon/RAL to host full EGamma dataset?)
- Perform scheduled analysis group “production”.
  - ~100kSI2k.month total = ~50 CPU per T1 (24 hrs/30 days)

Tier-2

- Assume about 5-8 T2:
  - 2 US, 1UK, 2-3 Italian, 1 Spanish, + ?
  - Store some of TAG data at each T2 (500GB? 1TB?)
  - Estimate 20CPU at each center for 1 month

DC04 Tasks At T1 and T2 (under discussion)

- “Most” T1s participate to Analysis Group Scheduled Productions
- One T1 and Two T2 do pseudo-calibration
  - Analyzing calibration DST’s, exercising round-trip for calibration back to T0
- Two T1 and Two T2 exercise LCG RB/RLS tools to prepare and submit jobs, accumulate results running over DST at one or both T1 centers.
- One T1 and Two T2 centers exercise LCG tools (GFAL/POOL/RLS) for job preparation and execution. (runtime file access from WAN/MSS)
- 1-2 T2 centers exercise Tag processing, defining new collections, constructing deep-copies at T1 and exporting new collections back to T2
- ….

- Filling in detals of Milestone plan
  - Not realy milestone yet, but work areas. Still needs quantitative specification

Pre-Challenge Milestones

- ✔ PCP-1.Generation of approximately 50 million Monte-Carlo events.
- PCP-2.Simulation of the events with either CMSIM or OSCAR.
  - ✔ PCP-2a. At least a fraction *x* of the 50M events simulated with CMSI These events must be Hit-Formatted by ORCA and stored in the POOL forma
  - PCP-2b. At least a fraction *(1-x)* of the 50M events simulated with OSCA and directly stored in the same POOL format as in (a) with the sa cataloguing and reference information available.
- PCP-3.The Digitization of the 50M events at an effective luminosity 2 x 10^{33} cm^{-2} s^{-1}
- PCP-4. At least the Digitized data (not necessarily with the MC tru information) transferred to the CERN Mass Storage, together with t adequate catalog information for its later processing.
- PCP-5.To be able to run 75% of the PCP Simulation production in an LC Grid Environment.
  - This is not an integrated 75% over the PCP period, but a demonstration that a period of, say, a week we can reach this instantaneous level.

Software milestones for PCP

✔ ◆ SWBASE-1. CMSIM version complete for PRS requirements

✔ ◆ SWBASE-2. OSCAR version complete for PRS requirements

✔ ◆ SWBASE-3. Storage of MC truth and hits in POOL. Readabil guaranteed for xxx months/years.

✔ ◆ SWBASE-4. Digitization code in ORCA/COBRA complete for PR requirements.

✔ ◆ SWBASE-5. At least a single-site complete catalog of all produced fil relevant to later processing. Meta-data adequate to perm dataset/collection processing

TIER-0 Milestones

- TIER0-1. Data serving pool to serve Digitized events at 25Hz to t computing farm with 20/24 hour operation.
  - Adequate buffer space (Digitized data set expected to be of order 100TB, aim keep 1/4 of this in the disk buffer).
  - Pre-staging software. File locking while in use, buffer cleaning and restocking files have been processed.
- TIER0-2. Computing Farm operating for 30 days . Approximately 3 running jobs 20/24 hours. Files in buffer locked till successful j completion. (500 events per job implies 3 hour batch jobs)
- TIER0-3. Output products of Tier-0 production stored to CERN MSS.
- TIER0-4. Data transfer performance defined.
  - How many streams at 100MB/s?
- TIER0-5. Secure and complete catalog of all data input/produc maintained at CERN.
- TIER0-6. Data catalog is accessible and/or replicable to the oth computing centers.

RPROM Software for T0

- RPROM-1. Tier-0 reconstruction software defined
- RPROM-2. DST persistent classes defined
- RPROM-3. Reconstruction and persistency code complete
- RPROM-4. TAG/NTUPLE defined.
  - (Information to characterize the events and allow efficient selection in la analysis)
- RPROM-5. TAG/NTUPLE production coded,
  - including cataloging information to allow at least ROOT and ORCA process of TAG/NTUPLE collections
- RPROM-6. TAG/NTUPLE analysis code ready from physics groups.
  - Critical point here. Is this NTUPLE or Analysis Object Data (AOD)?

Data Distribution

- DATA-1. Replication of the DST at one or more Tier-1 centers.
  - possibly using the LCG replication tools.
- DATA-2. Replication of at least those parts of the catalog that have be imported to each Tier-1.
- DATA-3. Transparent access of jobs at the Tier-1 sites to the local da whether in MSS or on disk buffer.
- DATA-4. Defined linkage between Tier-1 and Tier-2 sites. Tier-2 sites access the data only via the peer Tier-1 site.
  - (This is a linkage just for the duration of DC04 and subsequent analysis, no commitment for all time)
- DATA-5. Replication of the full Tier-0 TAG/NTUPLE at each Tier1 a further replication from the Tier-1 sites to requestingTier-2 sites
- DATA-6. Replication of any TAG/NTUPLEs produced at the Tier-1 sites the other Tier-1 sites and interested Tier-2 sites
- DATA-7. Monitoring of Data Transfer activites with for example Mona Lisa

Tier-1 Analysis Milestones

- TIER1-0. Participating Tier-1 centers define with approximate scales
- TIER1-1. All data distributed from Tier-0 safely inserted to local storage
- TIER1-2. Management and publication of a local catalog indicating status locally resident data
- TIER1-3. Operation of the PRS TAG/NTUPLE productions on the imported data.
- TIER1-4. Local computing facilities made available to Tier-2 users,
  - Possibly via the LCG job submission system.
- TIER1-5. Export of the PRS TAG/NTUPLE to requesting sites (Tier-0, -1 o 2)
- TIER1-6. Operation of a scheduled Analysis service, for example publicati of plots associated with the PRS TAG/NTUPLES for each dataset processe
- TIER1-7. Tier-1 data catalogue (either produced locally or replicated from the Tier-0) accessible remotely and made available to the “associated” Tier-2 centers.
- TIER1-8. Register the data produced locally to the Tier-0 catalog and mal them available to at least selected sites via the LCG replication tools.

Tier-2 Analysis Milestones

- TIER2-1. Pulling of data from peered Tier-1 sites as defined by the lo
  Tier-2 activities

- TIER2-2. Analysis on the local TAG/NTUPLE produces plots and
  summary tables.

- TIER2-3. Analysis on distributed TAG/NTUPLE or DST available at least
  the reference Tier-1 and “associated” Tier-2 centers.
  - Results are made available to selected remote users possibly via the LCG d
    replication tools.

- TIER2-4. Private analysis on distributed TAG/NTUPLE or DST is outsi
  DC04 scope but will be kept as a low-priority milestone.

CMS and LCG

Applications Area

- POOl work vital
  - SEAL as base for POOL and dictionary service vital
- GEANT4 collaboration much better. Now very good.
- ROOT collaboration effective. (Particularly with POOL/CMS (CERN & FNAL))
- SPI. Savannah excellent.
  - Misaligned expectations on SCRAM.
- CMS has reassigned some CMS/LCG manpower back to CMS in light of project status's and dire manpower situation in CMS

GRID Deployment and Technology

- CMS active tester and ready to use whatever is there in DC04
  - Looking forward to testing GFAL
- Excellent collaboration between CMS and LCG Grid Deployers

Fabrics

- Good collaboration on CERN T0/T1 specification
  - Role in worldwide computing less clear
- Need work on Disk Data Management
  - Important issue for all Tiers, only just beginning to be addressed