CMS

Compact Muon Solenoid

The Computing Project

Technical Design Report

LHCC Meeting June 29, 2005

Goals and Non-Goals

- Goals of CTDR
  - Extend / update the CMS computing model
  - Explain the architecture of the CMS computing system
  - Detail the project organization and technical planning
- Non-Goals
  - *Computing* TDR, so no details of ‘application’ software
  - It is not a ‘blueprint’ for the computing system
- Must be read alongside the LCG TDR

Computing Model

- CTDR updates the computing model
  - No major changes to requirements / specifications
  - LHC 2007 scenario has been clarified, is common between experiments
    - ~ 50 days @ $x.10^{32}$ $cm^{-2}s^{-1}$ in 2007
  - Additional detail on Tier-2, CAF operations, architecture
- Reminder of 'baseline principles' for 2008
  - Fast reconstruction code (reconstruct often)
  - Streamed primary datasets (allows prioritization)
  - Distribution of RAW and RECO data together
  - Compact data formats (multiple distributed copies)
  - Efficient workflow and bookkeeping systems
- Overall philosophy:
  - Be conservative; establish the 'minimal baseline' for physics

Data Tiers

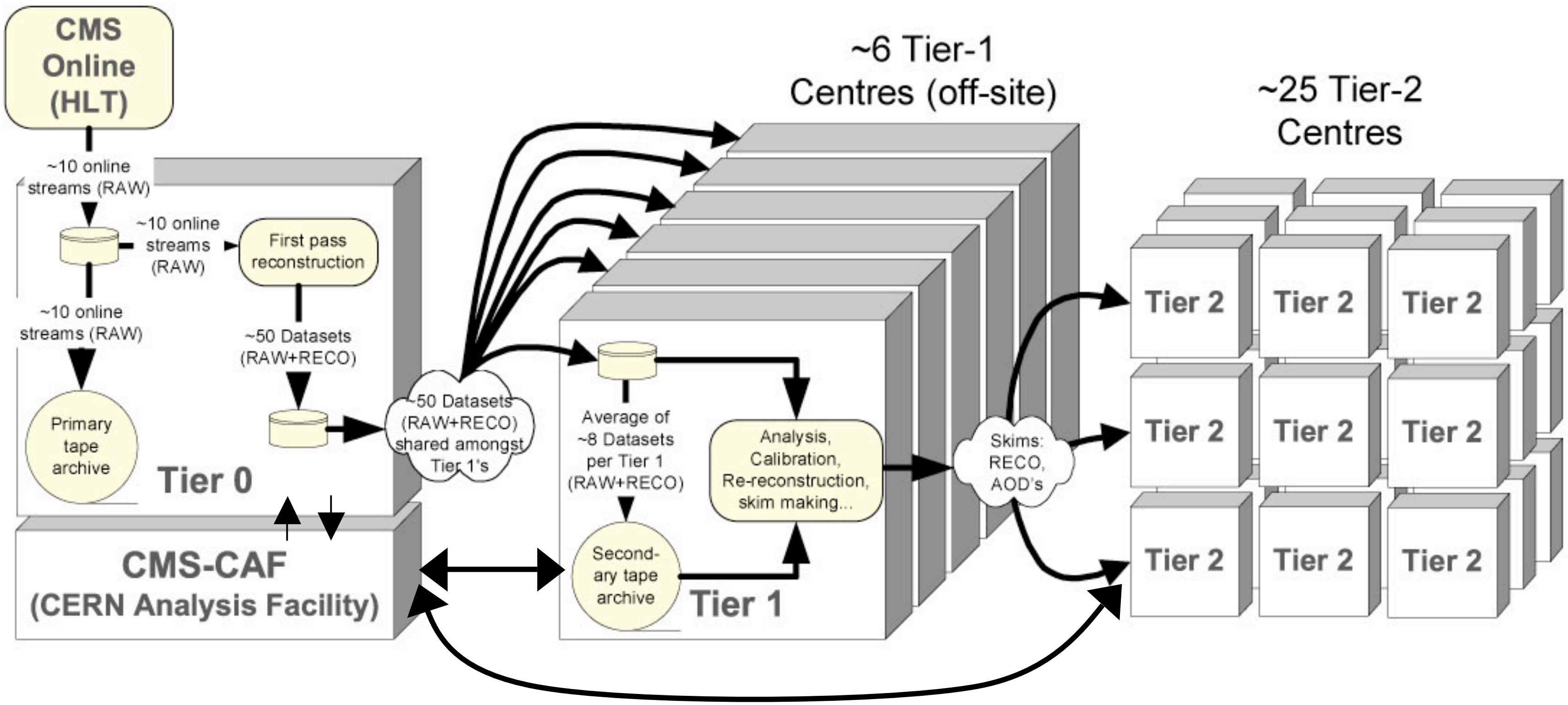
- RAW
  - Detector data + L1, HLT results after online formatting
  - Includes factors for poor understanding of detector, compression, etc
  - 1.5MB/evt @ <200Hz; ~ 5.0PB/year (two copies)
- RECO
  - Reconstructed objects with their associated hits
  - 250kB/evt; ~2.1PB/year (incl. 3 reproc versions)
- AOD
  - The main analysis format; objects + minimal hit info
  - 50kB/evt; ~2.6PB/year - whole copy at each Tier-1
- TAG
  - High level physics objects, run info (event directory); <10kB/evt

- Plus MC in ~ 1:1 ratio with data

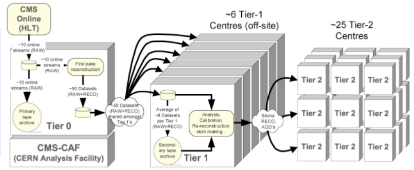
Data Flow

- Prioritization will be important
  - In 2007/8, computing system efficiency may not be 100%
  - Cope with potential reconstruction backlogs without delaying critical data
  - Reserve possibility of 'prompt calibration' using low-latency data
  - Also important after first reco, and throughout system
    - E.g. for data distribution, 'prompt' analysis
- Streaming
  - Classifying events early allows prioritization
  - Crudest example: 'express stream' of hot / calib. events
  - Propose O(50) 'primary datasets', O(10) 'online streams'
  - Primary datasets are immutable, but
    - Can have overlap (assume ~ 10%)
    - Analysis can draw upon subsets and supersets of primary datasets

Tiered Architecture

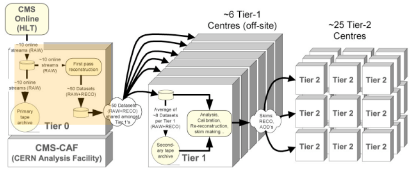


Computing System



- Where do the resources come from?
  - Many quasi-independent computing centres
  - Majority are 'volunteered' by 'CMS collaborators'
    - Exchange access to data & support for 'common resources'
    - …similar to our agreed contributions of effort to common construction tasks
  - A given facility is shared between 'common' and 'local use.
    - Note that accounting is essential
- Workflow prioritization
  - We will never have 'enough' resources!
  - The system will be heavily contended, most badly so in 2007/8
  - All sites implement and respect top-down priorities for common resources
- Grid interfaces
  - Assume / request that all Grid implementations offer agreed 'WLCG services'
  - Minimize work for CMS in making different Grid flavors work
    - And always hide the differences from the users

Tier-0 Center



Functionality

- Prompt first-pass reconstruction
  - NB: Not all HI reco can take place at Tier-0
- Secure storage of RAW&RECO, distribution of second copy to Tier-1

Responsibility

- CERN IT Division provides guaranteed service to CMS
  - Cast iron 24/7
- Covered by formal Service Level Agreement

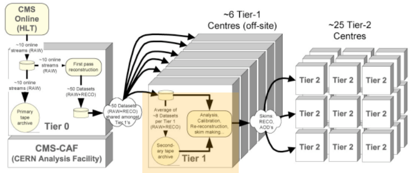
Use by CMS

- Purely scheduled reconstruction use; no ‘user’ access

Resources

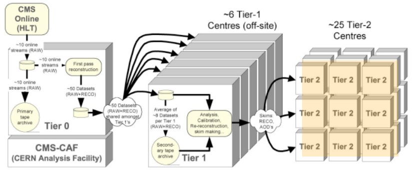
- CPU 4.6MSI2K; Disk 0.4PB; MSS 4.9PB; WAN 5Gb/s

Tier-1 Centers



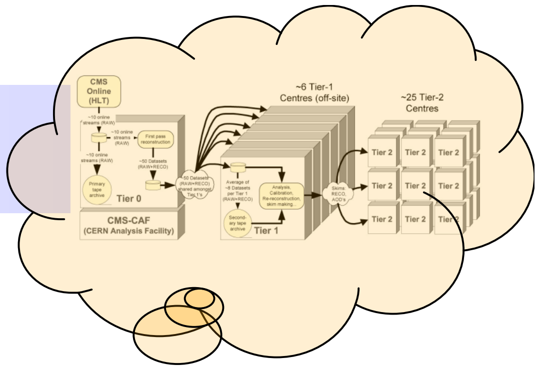
- Functionality
  - Secure storage of RAW&RECO, and subsequently produced data
  - Later-pass reconstruction, AOD extraction, skimming, analysis
    - Require rapid, scheduled, access to large data volumes or RAW
  - Support and data serving / storage for Tier-2
- Responsibility
  - Large CMS institutes / national labs
    - Firm sites: ASCC, CCIN2P3, FNAL, GridKA, INFN-CNAF, PIC, RAL
  - Tier-1 commitments covered by WLCG MoU
- Use by CMS
  - Access possible by all CMS users (via standard WLCG services)
    - Subject to policies, priorities, common sense, …
  - ‘Local’ use possible (co-located Tier-2), but no interference
- Resources
  - Require six ‘nominal’ Tier-1 centers; will likely have more physical sites
  - CPU 2.5MSI2K; Disk 1.2PB; MSS 2.8PB; WAN >10Gb/s

Tier-2 Centers



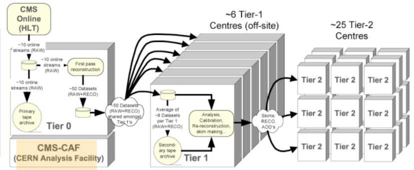
- Functionality
  - The 'visible face' of the system; most users do analysis here
  - Monte Carlo generation
  - 'Specialized CPU-intensive tasks, possibly requiring RAW data
- Responsibility
  - Typically, CMS institutes; Tier-2 can be run with moderate effort
  - We expect (and encourage) federated / distributed Tier-2's
- Use by CMS
  - 'Local community' use: some fraction free for private use
  - 'CMS controlled' use: e.g., host analysis group with 'common resources'
    - Agreed with 'owners', and with 'buy in' and interest from local community
  - 'Opportunistic' use: soaking up of spare capacity by any CMS user
- Resources
  - CMS requires ~25 'nominal' Tier-2; likely to be more physical sites
  - CPU 0.9MSI2K; Disk 200TB; No MSS; WAN > 1Gb/s
  - Some Tier-2 will have specialized functionality / greater network cap

Tier-3 Centers



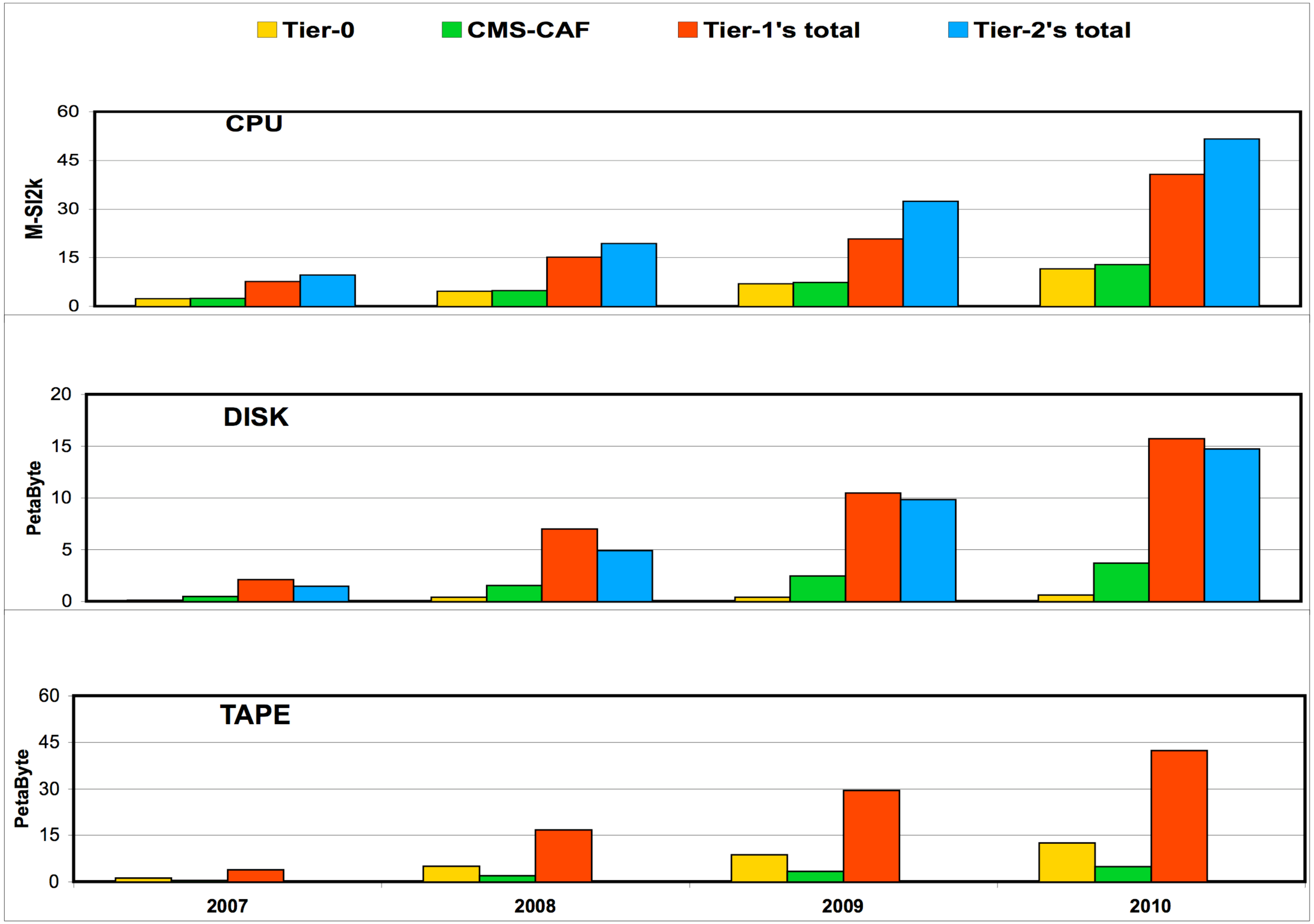
- Functionality
  - User interface to the computing system
  - Final-stage interactive analysis, code development, testing
  - Opportunistic Monte Carlo generation
- Responsibility
  - Most institutes; desktop machines up to group cluster
- Use by CMS
  - Not part of the baseline computing system
    - Uses distributed computing services, does not often provide them
  - Not subject to formal agreements
- Resources
  - Not specified; very wide range, though usually small
    - Desktop machines -> University-wide batch system
  - But: integrated worldwide, can provide significant resources to CMS on best-effort basis

CMS-CAF

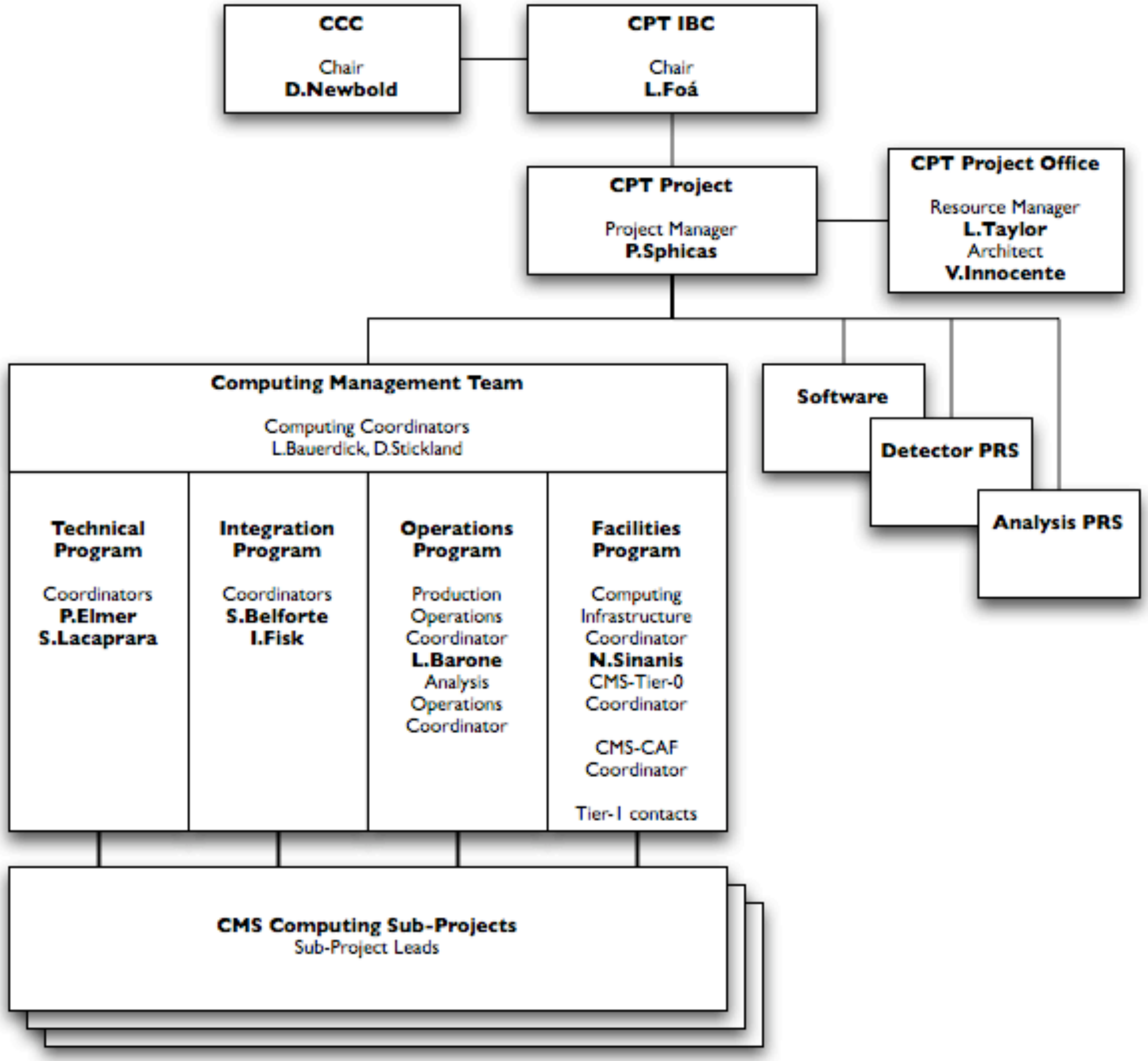


- Functionality
  - CERN Analysis Facility: development of the CERN Tier-1 / Tier-2
    - Integrates services associated with Tier-1/2 centers
  - Primary: provide latency-critical services not possible elsewhere
    - Detector studies required for efficient operation (e.g. trigger)
    - Prompt calibration ; ‘hot’ channels
  - Secondary: provide additional analysis capability at CERN
- Responsibility
  - CERN IT Division
- Use by CMS
  - The CMS-CAF is open to all CMS users (As are Tier-1 centers)
  - But: the use of the CAF is primarily for urgent (mission-critical) tasks
- Resources
  - Approx. 1 ‘nominal’ Tier-1 (less MSS due to Tier-0)+ 2 ‘nominal’ Tier-2
  - CPU 4.8MSI2K; Disk 1.5PB; MSS 1.9PB; WAN >10Gb/s
  - NB: CAF cannot arbitrarily access all RAW&RECO data during running
    - Though in principle can access ‘any single event’ rapidly.

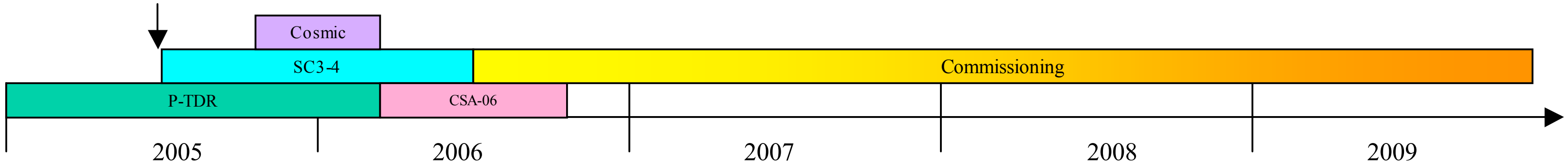
Resource Evolution



Project Organization



Project Phases

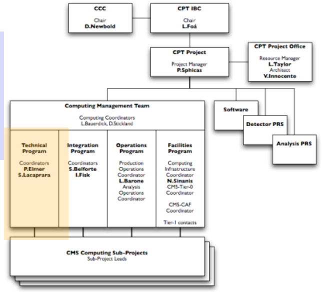


- Computing support for Physics TDR, -> Spring '06
  - Core software framework, large scale production & analysis
- Cosmic Challenge (Autumn '05 -> Spring '06)
  - First test of data-taking workflows
  - Data management, non-event data handling
- Service Challenges (2005 - 06)
  - Exercise computing services together with WLCG + centres
  - System scale: 50% of single experiment's needs in 2007
- Computing, Software, Analysis (CSA) Challenge (2006)
  - Ensure readiness of software + computing systems for data
  - 10M's of events through the entire system (incl. T2)
- Commissioning of computing system (2006 - 2009)
  - Steady ramp up of computing system to full-lumi running.

CPT L1 and Computing L2 Milestones V34.2

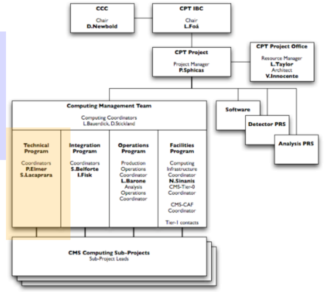
L1 Parent milestone	Date (version 34.2)	Milestone title	Level	ID
CPT-1	Aug-04	DC04 (5%) data challenge complete	2	CPT-101 / C
	Jan-05	Computing Model paper complete (1st draft Computing TDR)	2	CPT-102 / C
	Jun-05	Submission of Computing TDR	1	CPT-1
CPT-2	Jul-05	Initial integration of baseline computing components	2	CPT-202 / C
	Sep-05	Computing systems ready for Service Challenge SC3	2	CPT-204 / C
	Dec-05	Computing systems ready for Cosmic Challenge	2	CPT-212 / C
	Dec-05	Baseline Computing / Software Systems & Physics Procedures for Cosmic Challenge & Physics TDR	1	CPT-2
CPT-3	Apr-06	Submission of Physics TDR (Vols I and II)	1	CPT-3
CPT-4	Mar-06	Computing systems ready for Service Challenge SC4	2	CPT-402 / C
	Jun-06	Computing systems at Tier-0, 1, 2 centres ready for CSA-2006	2	CPT-404 / C
	Sep-06	Computing, Software, and Analysis Challenge (CSA-2006) complete	1	CPT-4
CPT-9	Dec-06	Submission of addenda to Physics TDR	1	CPT-9
CPT-5	Oct-06	Computing systems re-visited based on CSA-2006 lessons-learned	2	CPT-502 / C
	Dec-06	Integration of Computing Systems at Tier-0, 1 and 2 centres	2	CPT-504 / C
	Feb-07	Computing and Software Systems and Physics Procedures ready for data-taking	1	CPT-5
CPT-6	Feb-07	Tier-0 centre and CERN Analysis Facility ready for pilot run	2	CPT-601 / C
	Apr-07	Tier-1 and 2 centres ready for pilot run	2	CPT-602 / C
	Jun-07	Tier 0, 1, and 2 Computing Systems Operational (pilot run capacity)	1	CPT-6
CPT-7	Apr-08	Tier 0, 1, and 2 Computing Systems Operational (low luminosity capacity)	1	CPT-7
CPT-8	Apr-09	Tier 0, 1, and 2 Computing Systems Operational (high luminosity capacity)	1	CPT-8

Technical Program



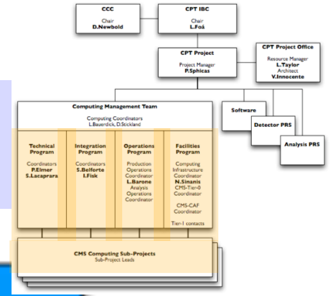
- Computing services:
  - Functionality and interfaces provided at the computing centres
  - Tools and mechanisms to allow use of the resources
    - Respecting CMS policy / priorities
  - Databases, bookkeeping and information services
- Strategy for the TDR
  - Cannot in 2004/5 specify a 'blueprint'
  - We specify 'baseline' targets and a development strategy
  - Aim to provide a continually 'up' system with incremental performance and functional improvements
    - Feed back results into next stages of development
- Use of the Grid
  - Most underlying functions provided by 'Grid services'
  - Grid - application interfaces need to be well-defined, but will evolve
  - Must accommodate a variety of Grid flavors

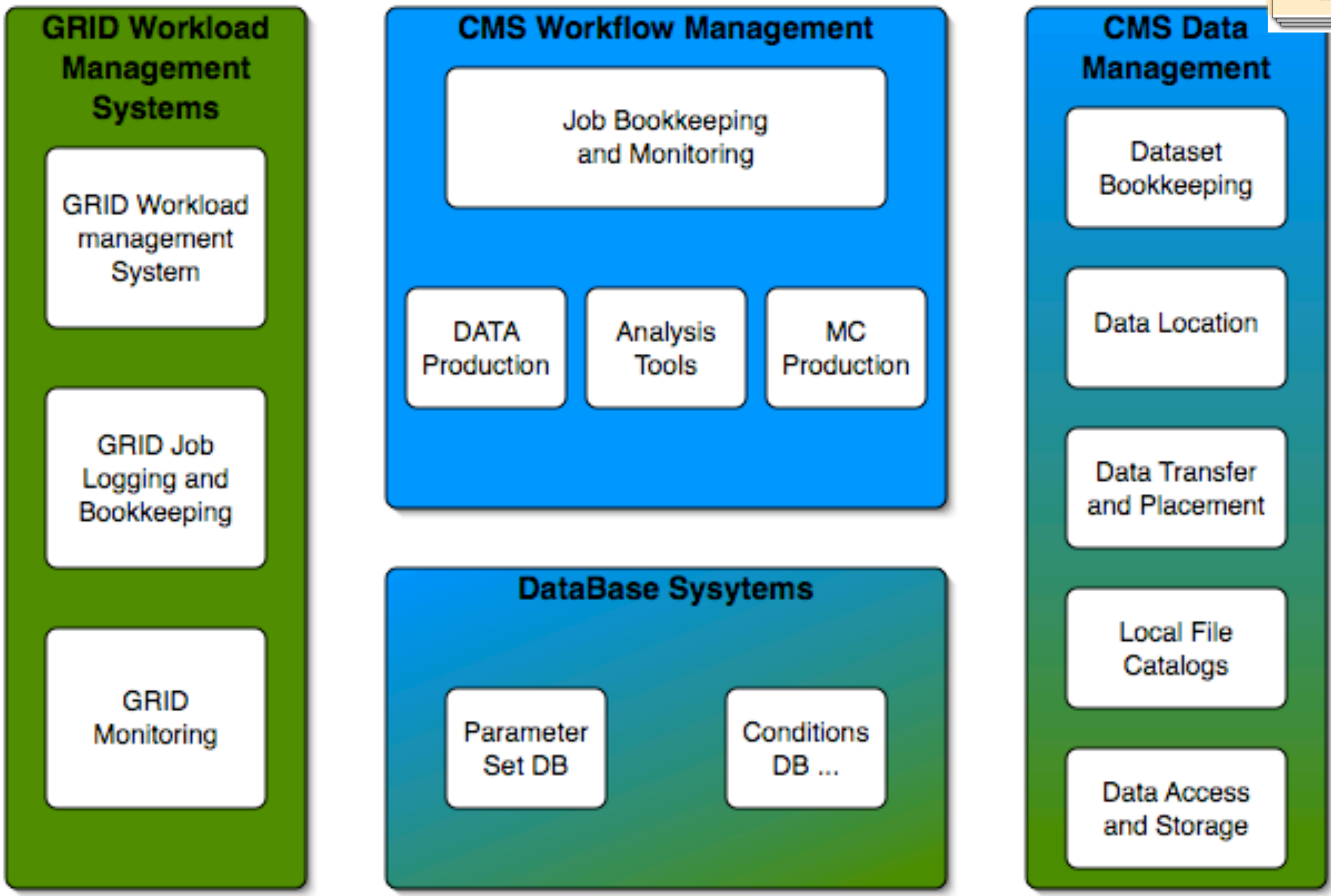
Design Philosophy



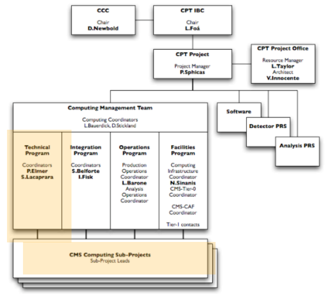
- Optimize for the common case:
  - Optimize for read access
    - Most data is write-once, read-many
  - Optimize for bulk processing, but without limiting single user
- Decouple parts of the system:
  - Minimize job dependencies
    - Allow parts of the system to change while jobs are running
  - Site-local information stays site-local
- Know what you did to get here:
  - ‘Provenance tracking’ is required to understand data origin
- Keep it simple!
- Also: Use explicit data placement
  - Data does not move around in response to job submission
  - All data is placed at a site through explicit CMS policy

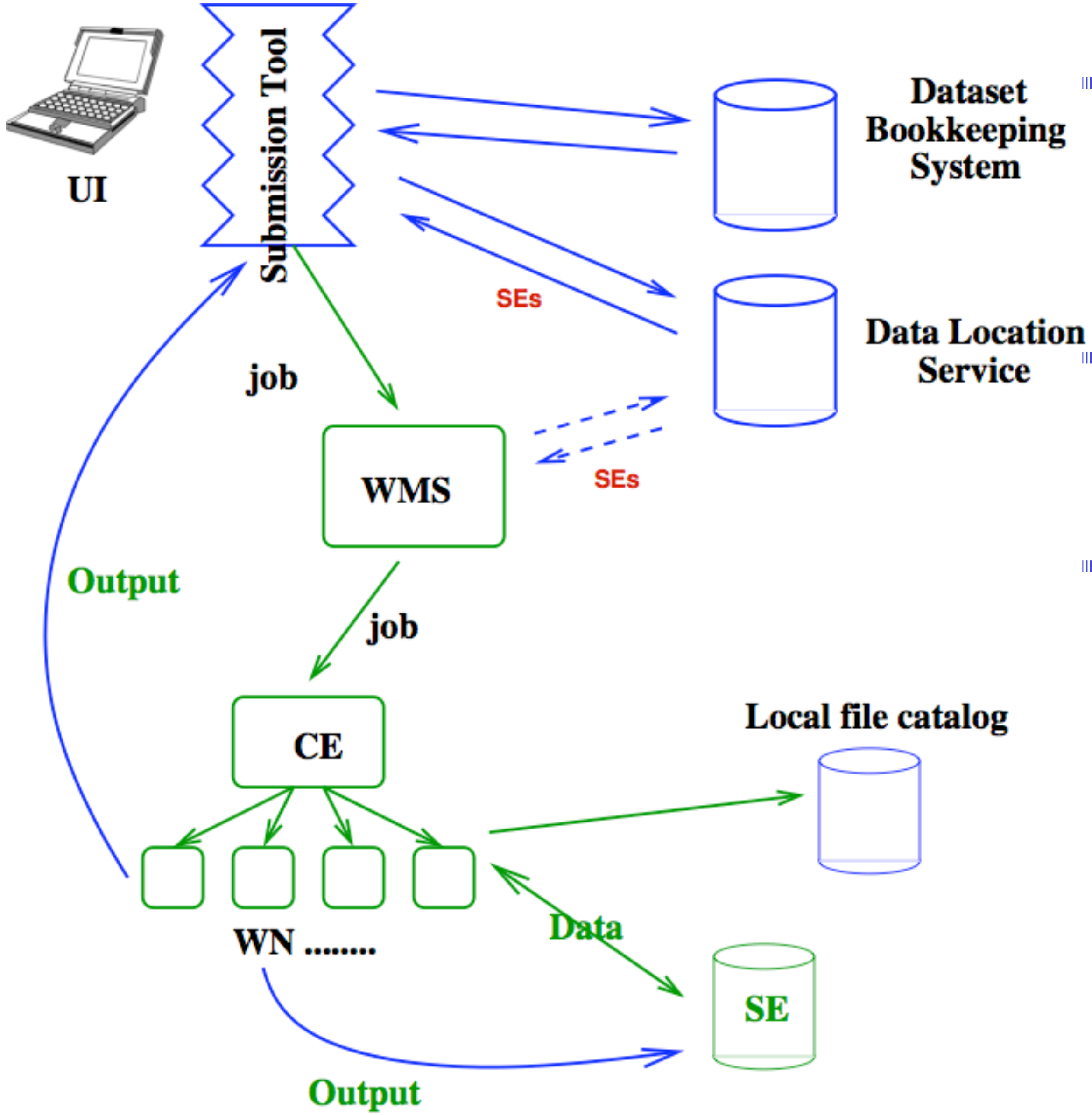
Services Overview





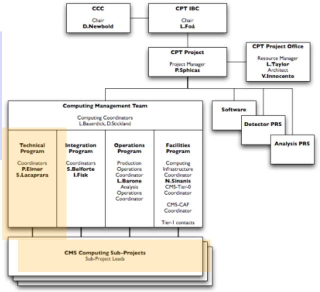
Basic Distributed Workflow



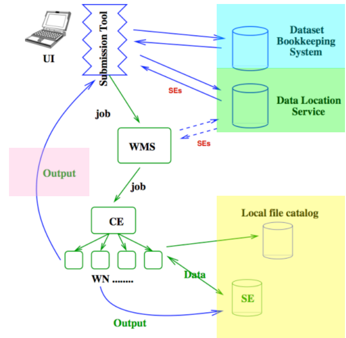


- The CTDR has served to converge on a basic architectural blueprint for a baseline system.
- We are now beginning the detailed technical design of the components
- It should be possible to bring up such a system over the next 6-9 months for the cosmic challenge and then CSA 2006

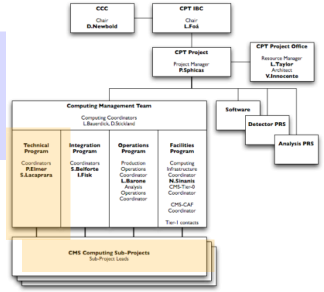
Data Management



- Data organization
  - ‘Event collection’: the smallest unit larger than one event
    - Events clearly reside in files, but CMS DM will track collections of files (aka blocks) (Though physicists can work with individual files)
  - ‘Dataset’: a group of event collections that ‘belong together’
    - Defined centrally or by users
- Data management services
  - Data book-keeping system (DBS) : “what data exist?”
    - NB: Can have global or local scope (e.g. on your laptop)
    - Contains references to parameter, lumi, data quality information.
  - Data location service (DLS) : “where are the data located?”
  - Data placement system (PhEDEx)
    - Making use of underlying Baseline Service transfer systems
  - Site local services:
    - Local file catalogues
    - Data storage systems

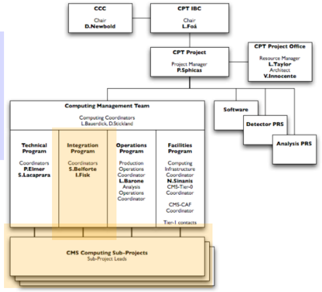


Workload Management



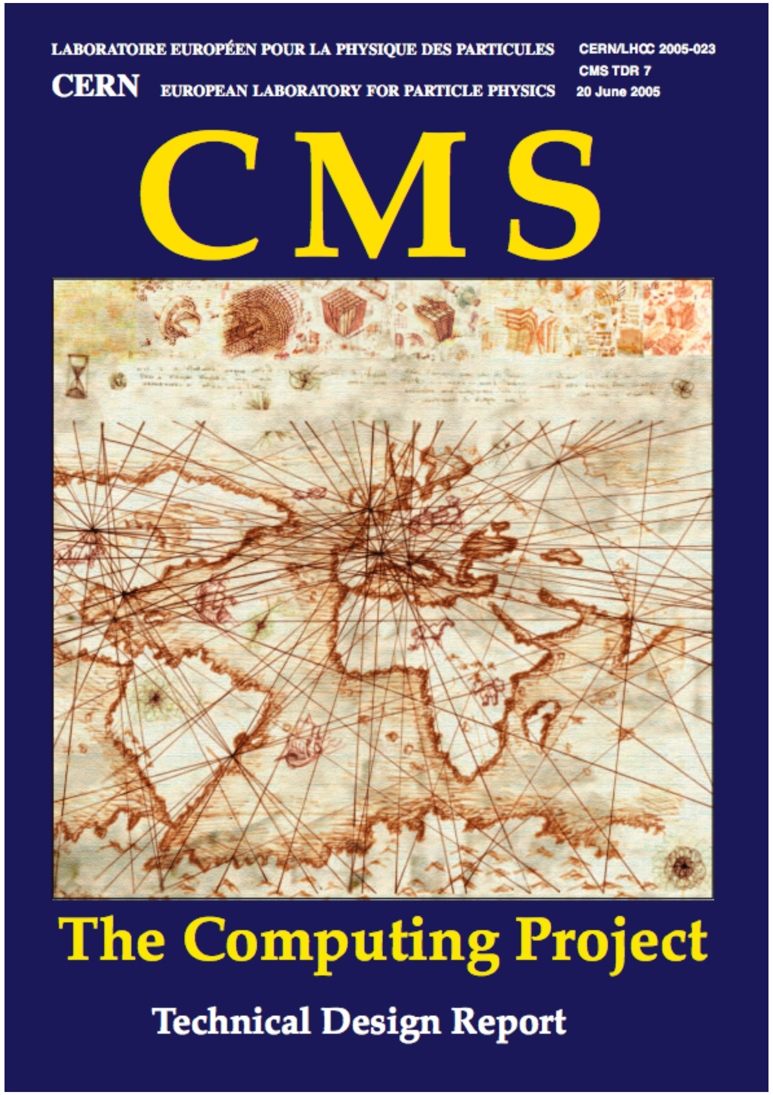
- Running jobs on CPUs…
- Rely on Grid workload management, which must
  - Allow submission at a reasonable rate: O(1000) jobs in a few sec
  - Be reliable: 24/7, > 95% job success rate
  - Understand job inter-dependencies (DAG handling)
  - Respect priorities between CMS sub-groups
    - Priority changes implemented within a day
  - Allow monitoring of job submission, progress
  - Provide properly configured environment for CMS jobs
- Beyond the baseline
  - Introduce 'hierarchical task queue' concept
  - CMS 'agent' job occupies a resource, then determines its task
    - I.e. the work is 'pulled', rather than 'pushed'.
  - Allows rapid implementation of priorities, diagnosis of problems

Integration Program



- This Activity is a recognition that the program of work for Testing, Deploying, and Integrating components has different priorities than either the development of components or the operations of computing systems.
  - The Technical Program is responsible for implementing new functionality, design choices, technology choices, etc.
  - Operations is responsible for running a stable system that meets the needs of the experiment
    - Production is the most visible operations task, but analysis and data serving is growing.
    - Event reconstruction will follow
  - Integration Program is responsible for installing components in evaluation environments, integrating individual components to function as a system, performing evaluations at scale and documenting results.
    - The Integration Activity is not a new set of people nor is it independent of either the Technical Program or the Operations Program
    - Integration will rely on a lot of existing effort

Conclusions



- CMS gratefully acknowledges the contributions of many many people to the data challenges that have led to this TDR
- CMS believes that with this TDR we have achieved our milestone goal to describe a viable computing architecture and the project plan to deploy it in collaboration with the LCG project and the Worldwide LCG Collaboration of computing centers

Let the games begin