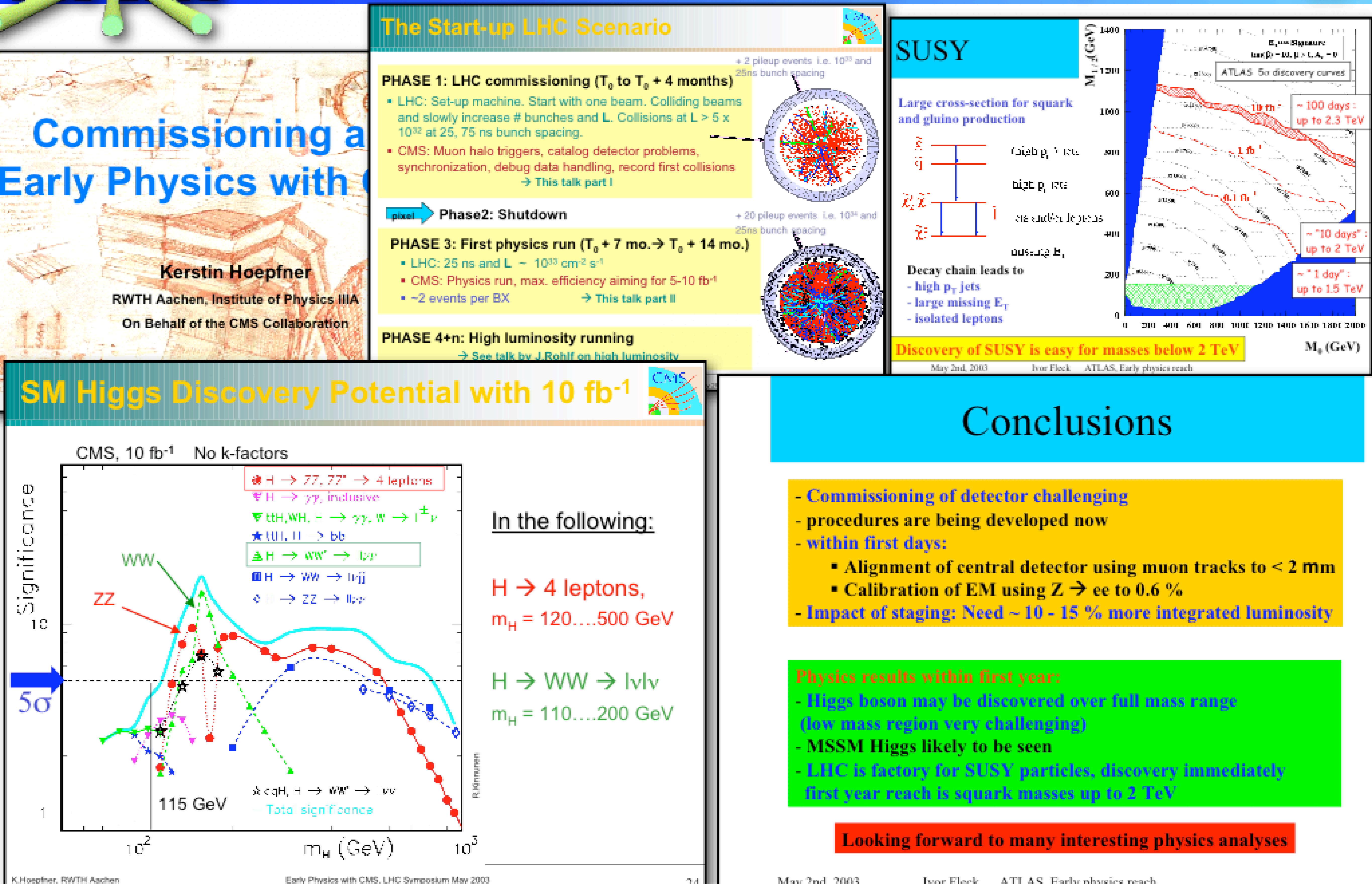
Talking Points on US CMS Analysis Plans

LATBauerdick, US CMS S&C Project Manager

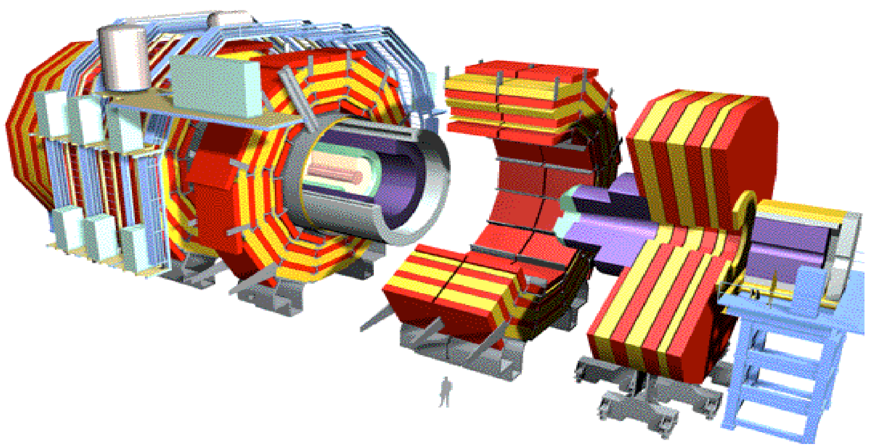
Caltech Workshop on Grid Enabled Analysis
June 23, 2003

We got to be ready on day 1!



And We Want to be Ready in the US

Physics Analysis requires Information Technology and Computing Infrastructure

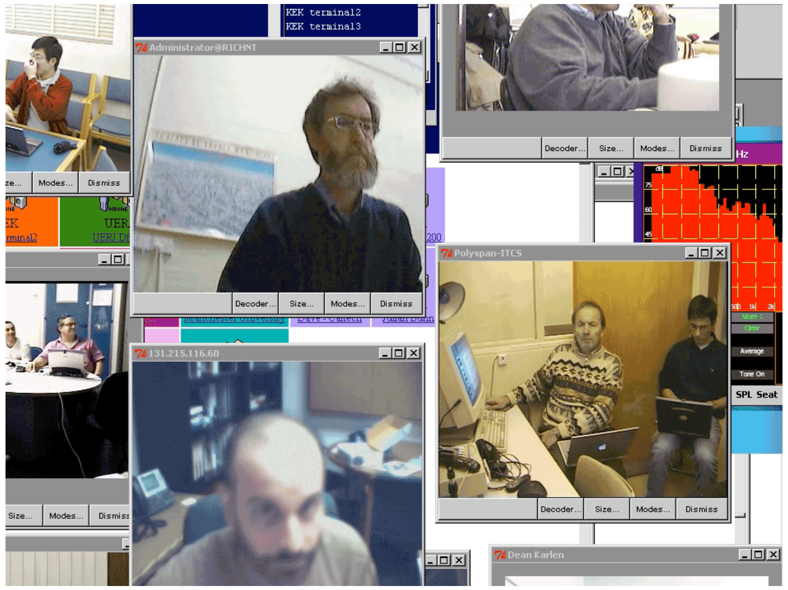


\+

—> Need an Advanced Coherent Global "**Info**rmation-Infra**structure**"

International and Interdisciplinary Partnerships

US LHC: Empower the LHC Scientists at Universities and Labs to do Research on LHC Physics Data



This is why we are pushing Grids and other Enabling Technology

e.g. Gigabit/sec access through WAN may well completely change the way we will do Analysis

CMS Strategy: Computing and Physics Model

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

Technologies
Evaluation and evolution

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

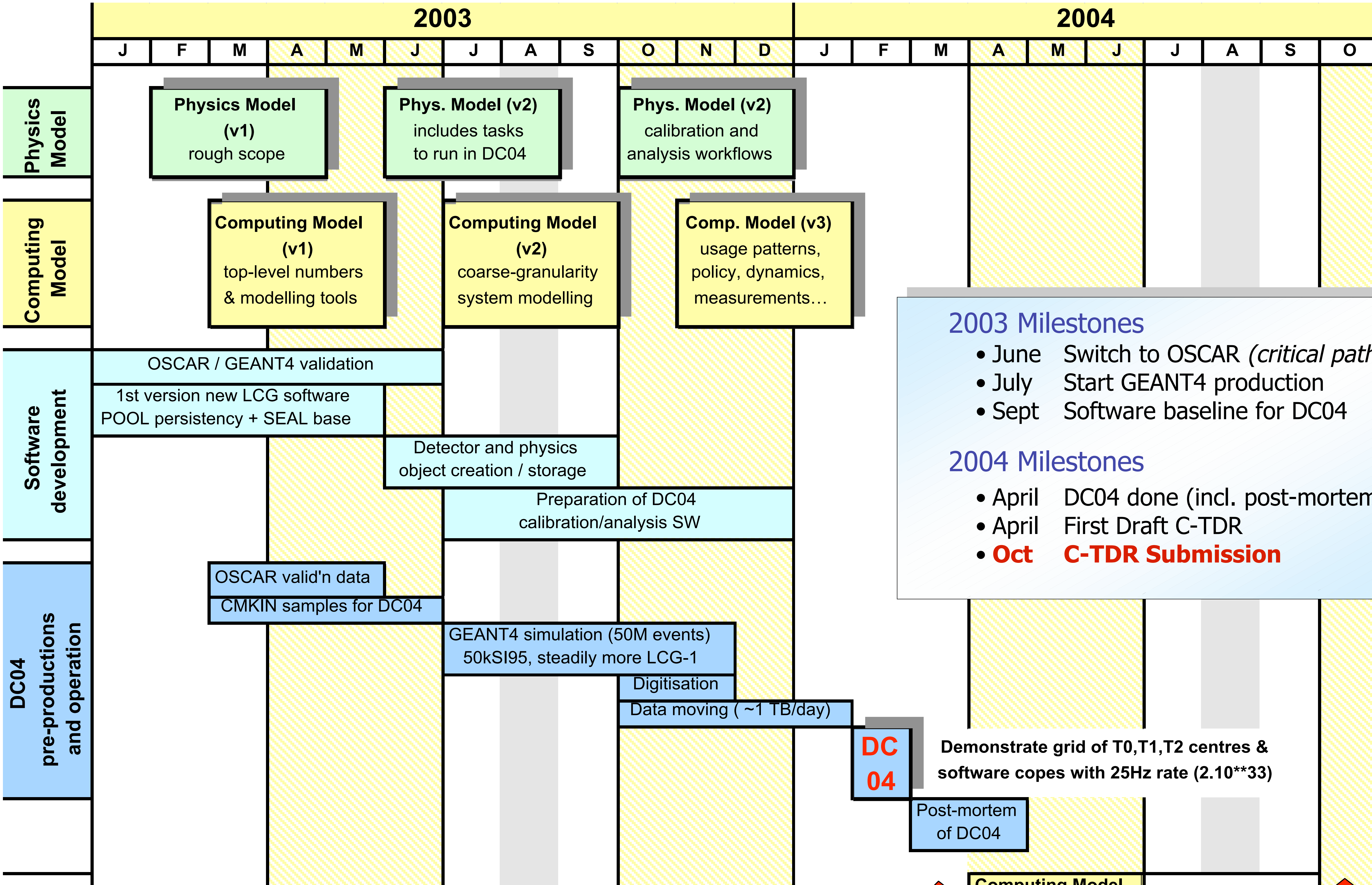
Validation of Model

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

Simulations
Model systems & usage patterns

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

CMS Data Challenges and TDR Schedule



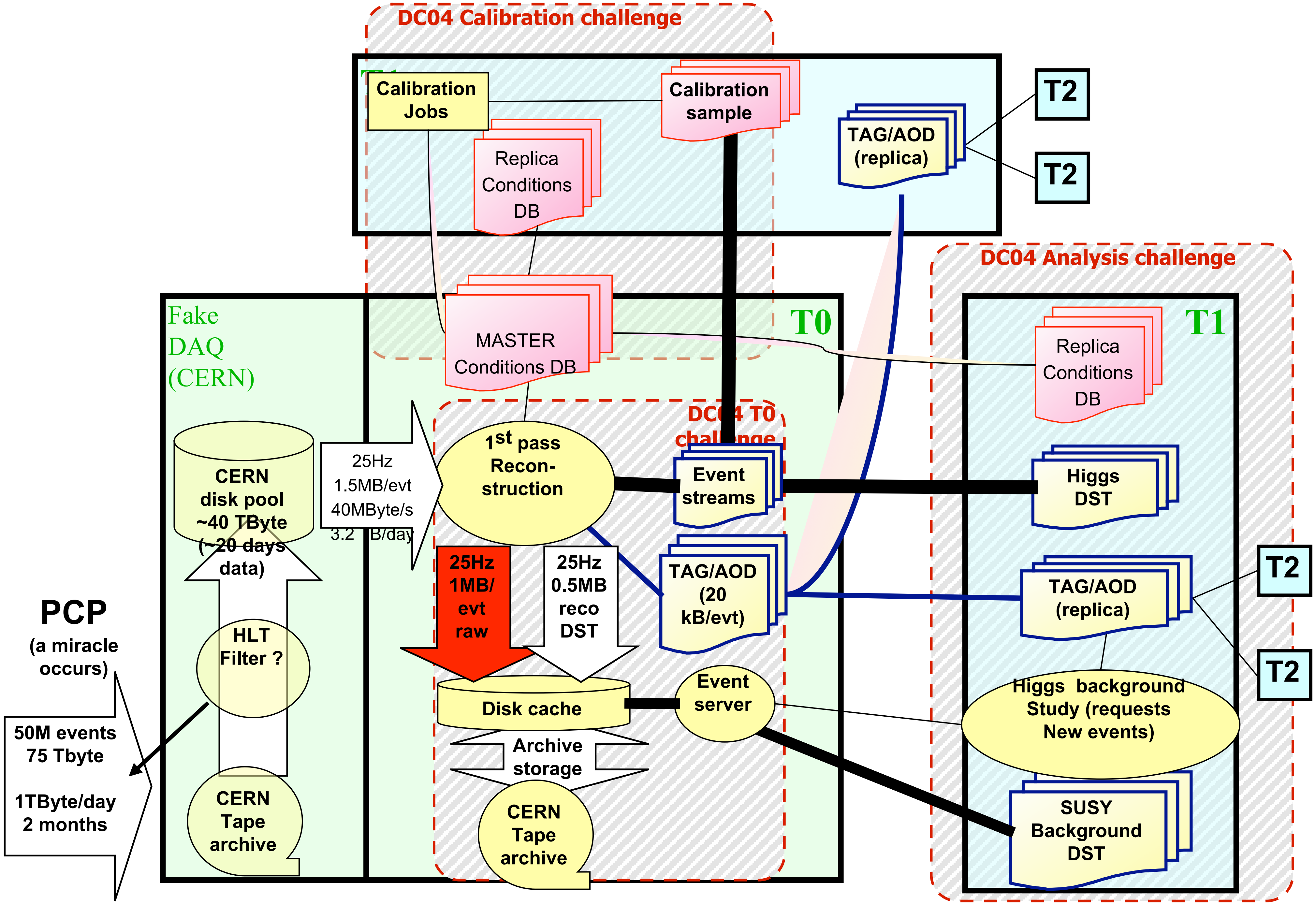
2003 Milestones

- June Switch to OSCAR *(critical path)*
- July Start GEANT4 production
- Sept Software baseline for DC04

2004 Milestones

- April DC04 done (incl. post-mortem)
- April First Draft C-TDR
- **Oct C-TDR Submission**

Plan for CMS DC04



Pre-Challenge Production

US CMS working full-speed with CMS to prepare the Data Challenge 04 Many pieces involved

- Preparing the Software Framework and the Software Support Components, including the new Object persistency and file handling
- Preparing the Tools for Distributing Software and Running Jobs
- Setting up the Grid Environment, Packaging and Deployment
- Developing the "Storage Element" at it's interfaces, including Data and "Replica" Management
- Commissioning Robust and Efficient File Transfers and Data Movement
- Providing the basic "Authentication" services for a (~static) VO
- Simulation of the distributed computing model

- So, how about Analysis Services?

Distributed Analysis

Unclear in the LHC community how we should approach that new focus

- Distributed Analysis effort not yet projectized in the US CMS "WBS"
- Need to understand what should be on CMS, in LCG AA, in R&D projects
  - perception of (too many) independent (duplicating) efforts (?)

What can we test/use in DC04?

- Some prototypes can be tested soon and for DC04
  - What are the assumptions they make on the underlying GRID
  - On Physicists work patterns?
  - How are their architectures similar/different?
  - Are their similarities that can sensibly be abstracted to common layers?
    - Or is it premature for that
  - Diversity is probably good at this time!

LCG RTAG on "An Architectural Roadmap towards Distributed Analysis"

- review existing, confront with HEPCAL use cases, consider interfaces between Grid, LCG and Application services,
- To develop a roadmap specifying wherever possible the architecture, the components and potential sources of deliverables to guide the medium term (2 year) work of the LCG and the DA planning in the experiments.

Distributed Analysis: Shifting the Focus

going forward to analysis means a significant paradigm shift

- from well-defined production jobs to interactive user analysis
- from DAGs of process to "Sessions" and state-full environments
- from producing sets of files to accessing massive amounts of data
- from files to data sets and collection of objects
- from using essentially "raw data" to complex layers of event representation
- from "assignments" from the RefDB to Grid-wide Queries
- from "user registration" to enabling sharing and building communities

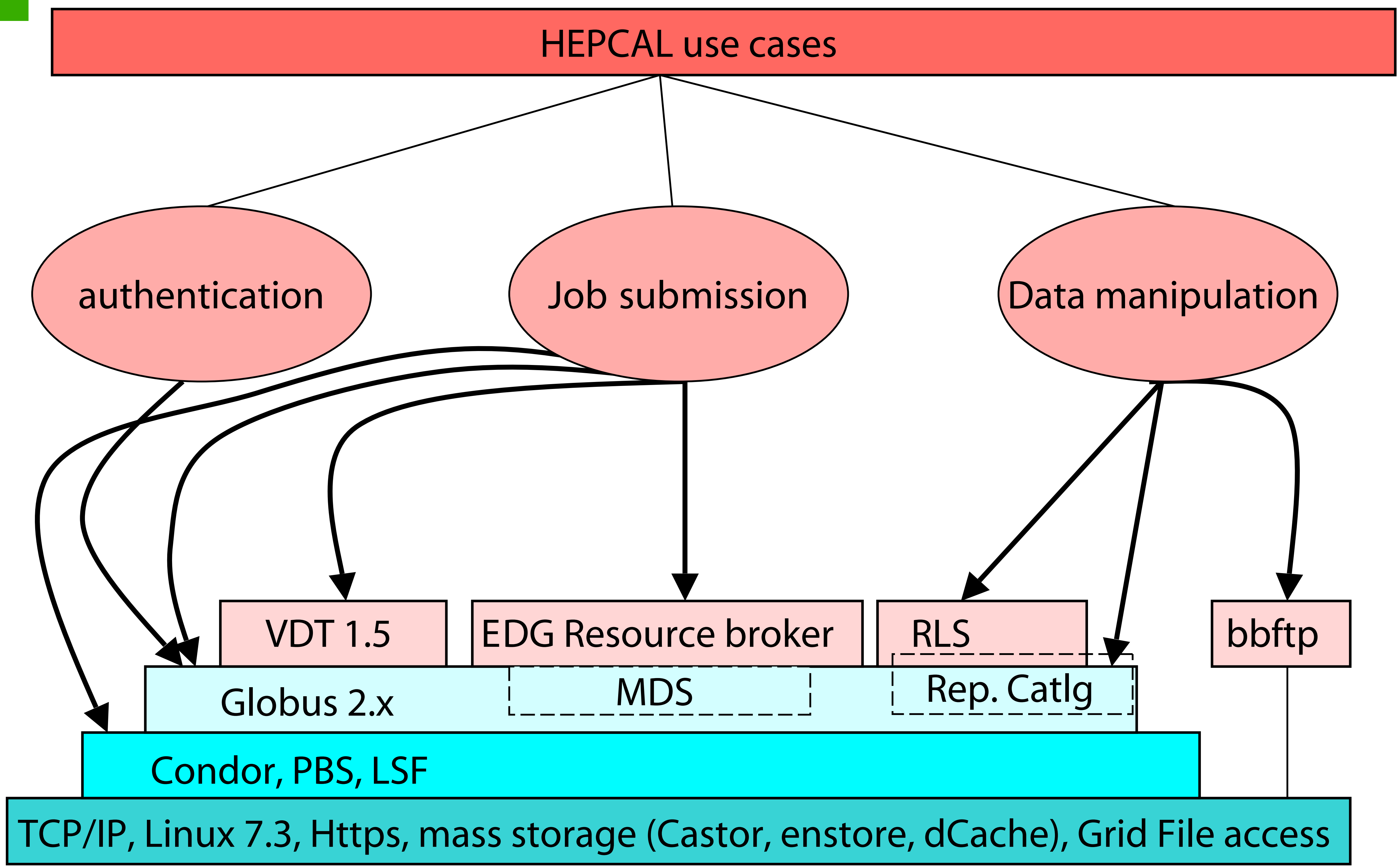
are the (Grid) technologies ready for this?

- there will be a tight inter-play between prototyping the analysis services and developing the "lower level" services and interfaces
- how can we approach a "roadmap towards an Architecture"?

what are going to be the "new paradigms" that will be exposed to the user?

- user analysis session transparently extended to a distributed system
  - but requires a more prescriptive and declarative approach to analysis
- set of services for "collaborative" work
  - new paradigms beyond "analysis"

LCG Prototype: LCG-1 Architecture of Middleware Layers



Harvey et al: HEP Layers, End-to-end Services

HEP Grid Architecture: (H. Newman)

Layers Above the Collective Layer

Physicist's Application Codes

- Reconstruction, Calibration, Analysis

Experiments' Software Framework Layer

- Modular and Grid-aware:
  Architecture able to interact effectively with the lower layers (above)

Grid Applications Layer
(Parameters and algorithms that govern system operations)

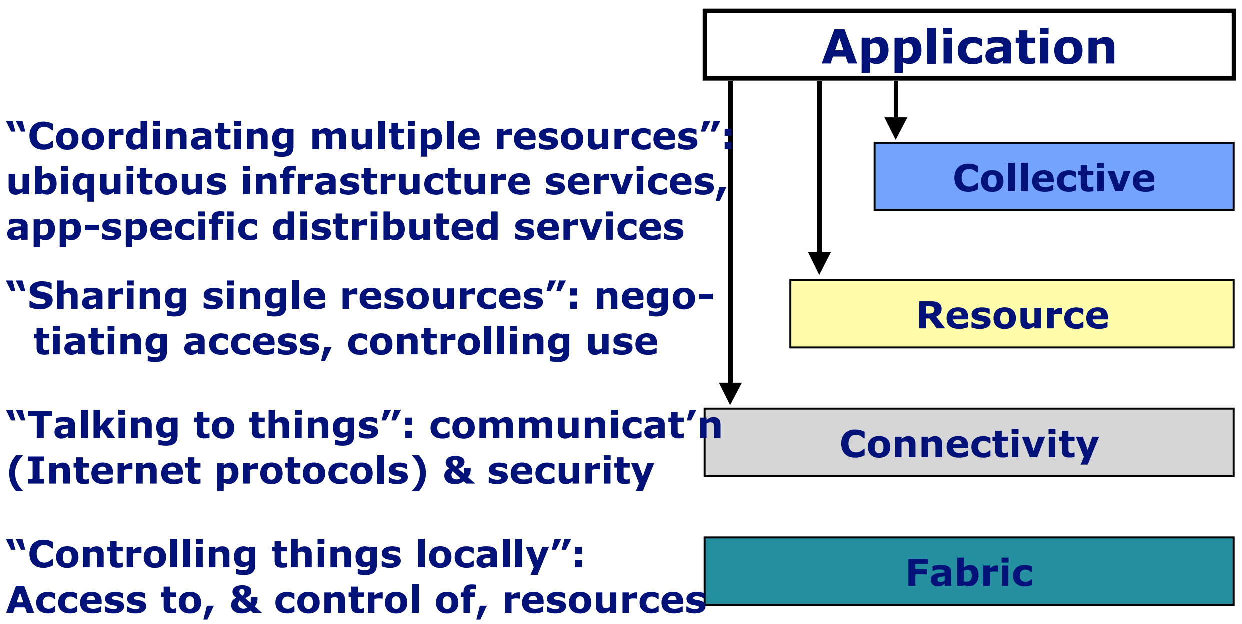
- Policy and priority metrics
- Workflow evaluation metrics
- Task-Site Coupling proximity metrics

Global End-to-End System Services Layer

- Workflow monitoring and evaluation mechanisms
- Error recovery and long-term redirection mechanisms
- System self-monitoring, steering, evaluation and optimization mechanisms
- Monitoring and Tracking Component performance

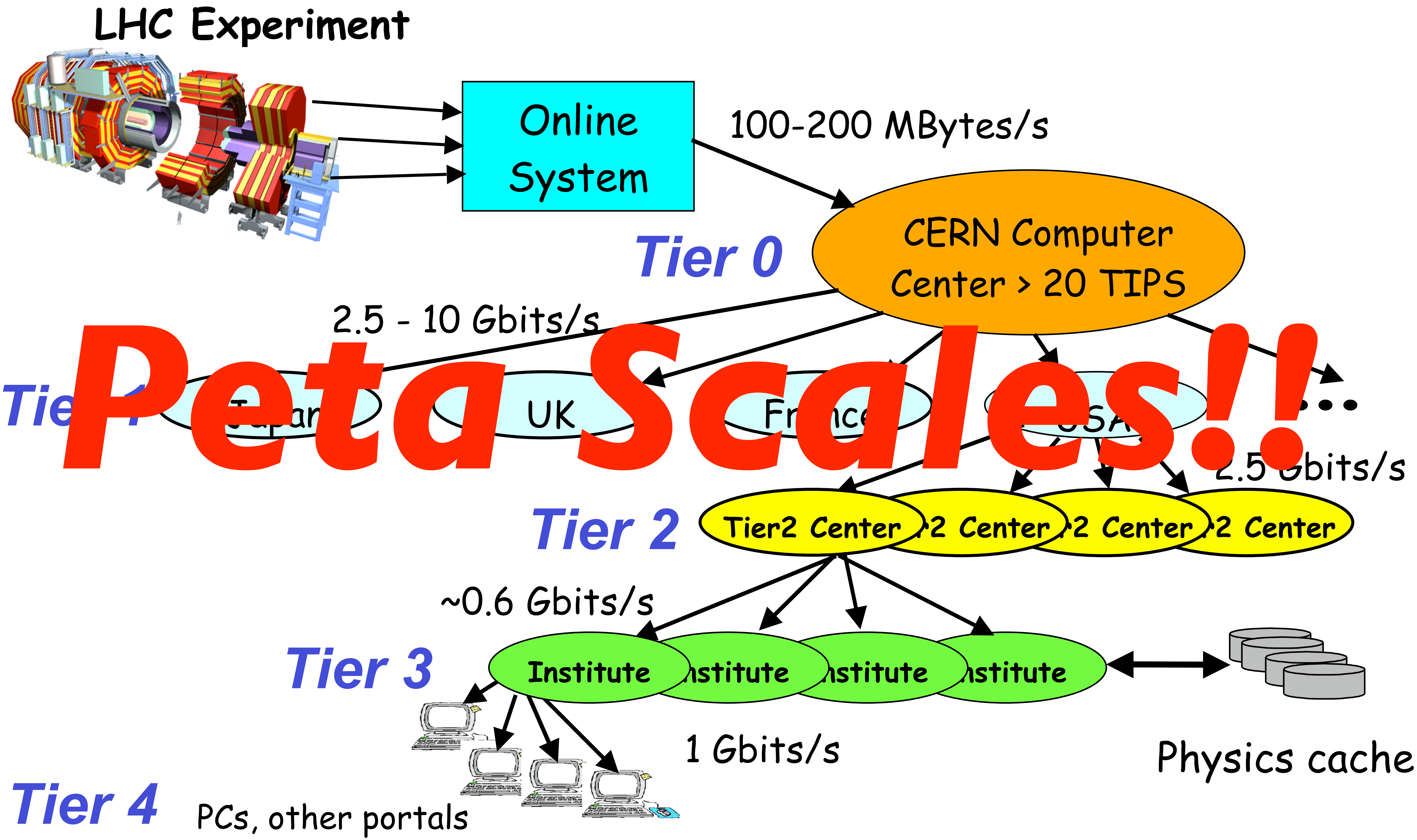
Layered Grid Architecture

(I.Foster et al.)



Already investigate a set of prototypical services and architectures

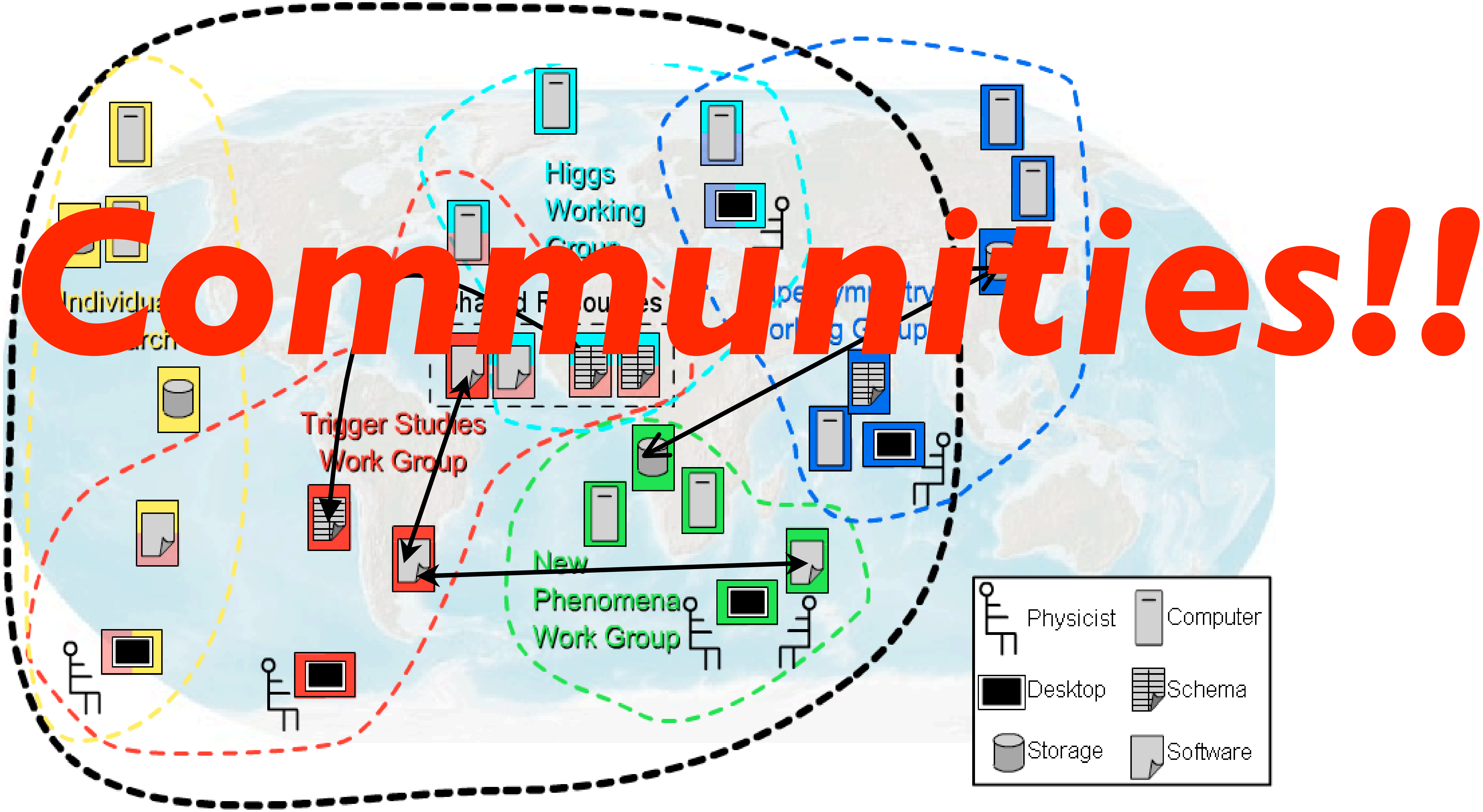
LHC Multi-Tier Structured Computing Resources



Scientists within Dynamic Workspaces!

Communities of Scientists Working Locally within a Global Context

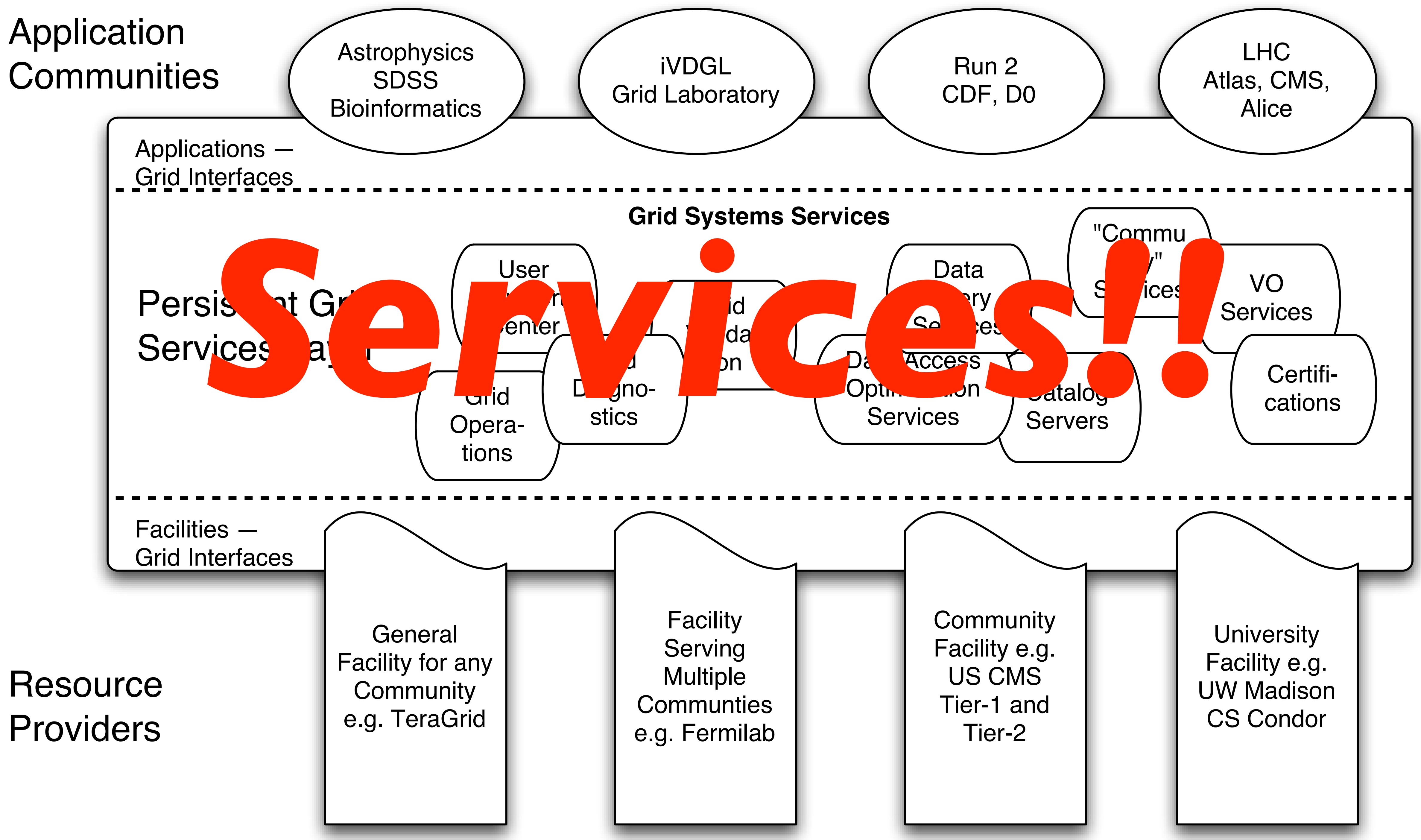
Infrastructure for sharing, consistency of physics and calibration data, software



Grid Services Infrastructure for Analysis

Grid Layer “Abstraction” of Facilities — Rich with Services!

Open Science Grid Services



Steps towards Grid Service **Infrastructure**

Initial Testbeds in US Atlas and US CMS, consolidation of middleware to VDT

- VDT agreed as basis of emerging LCG service, basis of the EDG 2.0 distribution

Build a functional Grid between Atlas and CMS in the US: **Grid03**

- based on VDT, with a set of common services:
  VO management, information services, monitoring, operations, etc
- demonstrate this infrastructure using well-defined metrics for LHC applications
  - November CMS demonstration of reliant massive production (job throughput), robust data movements (TB/day), consistent data management (#files, #sites)
  - to scale of the 5% data challenge DC04, planned for Feb. 2003

Get LHC Grid stake holders together in the US and form the **Open Science Consortium**

- LHC labs, Grid PIs, Tera Grid, Networking

develop plan for implementing and deploying **Open Science Grid** peering with the EGEE in Europe, Asia to provide LHC infrastructure

A Project to Build the Open Science Grid

Scope out services and interface layers between Applications and Facilities

- LHC already has identified funding for the fabric and it's operation

Work packages to acquire and/or develop enabling technologies as needed

- goal to enable "persistent organizations" like the national labs to provide those infrastructures to the application communities (CMS, Atlas, etc)
  - develop the "enabling technologies" and systems concepts that allow the fabric providers to function in a Grid environment, and the applications and users to seamlessly use it for their science
    - develop well defined interfaces and a services architecture
      - issues like distributed databases, object collections, global queries
    - work on the technologies enabling end-to-end managed resilient and fault tolerant systems: networks, site facilities, cost-estimates
- devise strategies for resource use, and dependable "service contracts"

Put up the initial operation infrastructure

Start here:

Caltech
Genus Loci