CMS Use Cases for SC4

Ian Fisk

February 12, 2006

The Computing TDR

CMS spent most of the last year working on data management and the computing model

- CMS RTAG on Data Management in the fall of 2004
- Release of the Computing Technical Design Report in the summer of 2005

Document describes baseline computing model for CMS

- Explanation of computing capacity
- Interconnectivity
- Baseline services and activity descriptions

Document describes a fairly traditional and largely statically partitioned distributed computing model

- Makes realistic requirements on grid services

More information in

- http://cmsdoc.cern.ch/cms/cpt/tdr/index.html

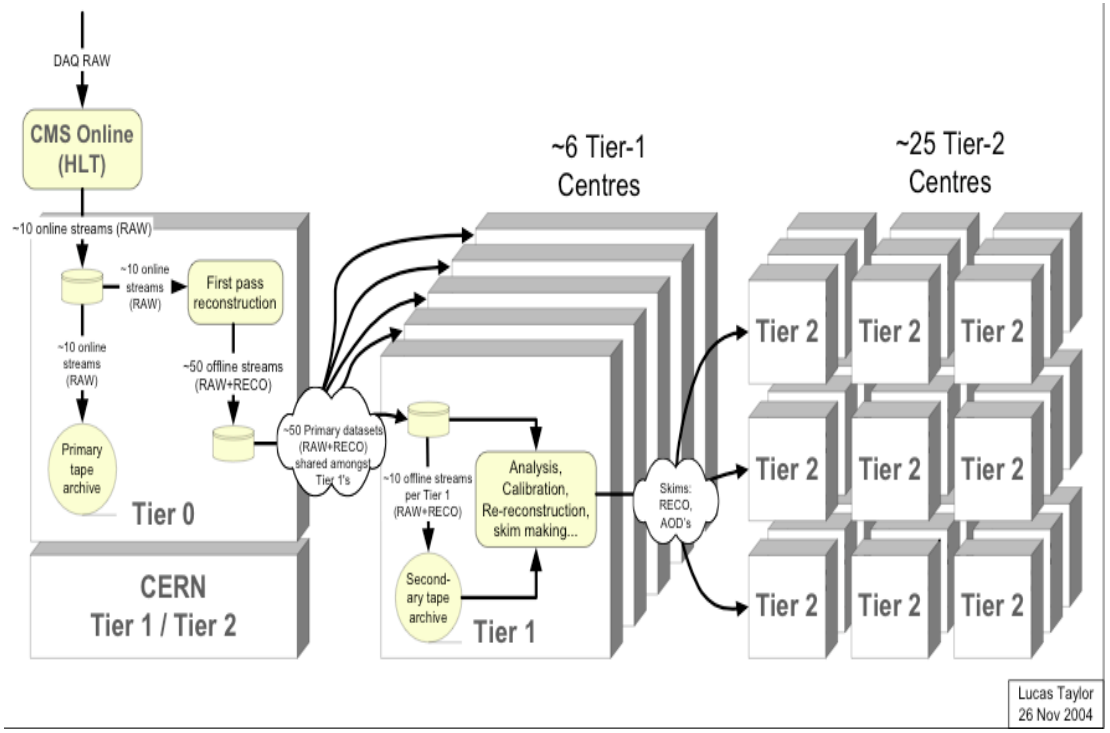
CMS Computing Model

CMS has proposed a computing model where the site activities and functionality is largely predictable

- Activities are driven by data location
- Opportunistic computing is largely restricted to limited activities

A system we think we can build

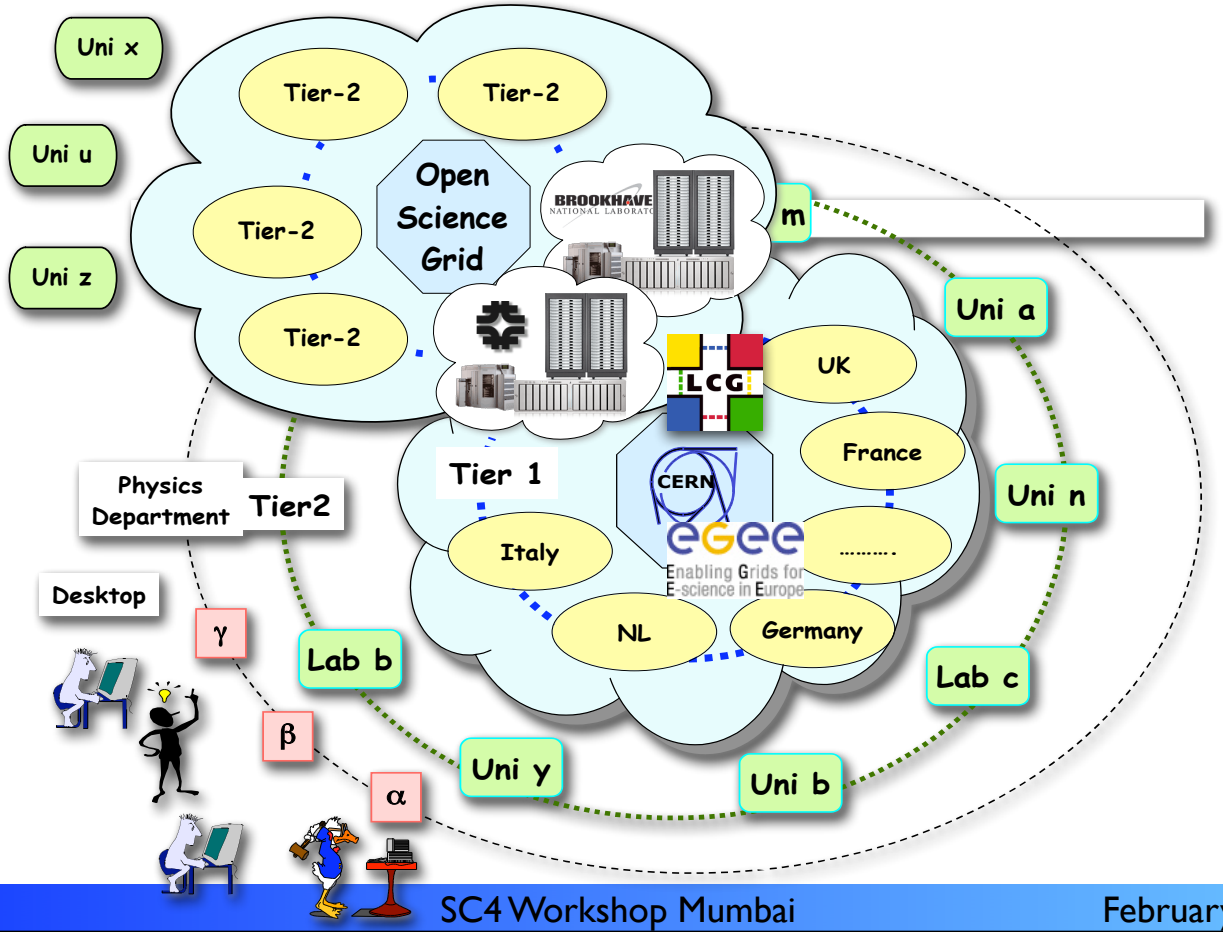
- Does not prevent more dynamic computing models if functionality and capacity are available.
  - Scheduling decisions are simplified



Cloud Model

The CMS model differs from the cloud model because it's easy to determine the function of a site by data placement

- Eventually data placement will be automated, but at the beginning it will be tuned



Input Parameters For Model

Event Sizes

- Current estimate of raw data event size is 1.5MB (1-2MB)
- Size of Reconstructed Event is 0.25MB
- Analysis Object data is 0.05MB per event

CMS best estimate is about 150Hz for the DAQ target Event rate

- ~ 200MB/s

During normal CMS running we expect to log about 2PB of data per year of raw data

- Raw data will be sent to Tier-1 centers for archiving and serving
  - CERN will have all raw, but will only serve some of the raw data
- During the first several years of the experiment the analysis will have to access more raw data
  - The understanding of the detector will be a long process
  - Leads to larger data sets for analysis and larger selected datasets transferred from the Tier-1 center to remote analysis resources.

Roles and Responsibilities

Tier-0

- Primary reconstruction
- Partial Reprocessing
- First archive copy of the raw data

Tier-1s

- Share of raw data for custodial storage
- Data Reprocessing
- Data selection and skimming Tasks
- Data Serving to Tier-2 centers for analysis
- Archive Simulation From Tier-2

Tier-2s

- Monte Carlo Production
- Analysis

Computing Center Specifications

Tier-0 Center

		Running Year 2007	2008	2009	2010	
Conditions		Pilot	2E33+HI	2E33+HI	E34+HI	
Tier-0	CPU	2.3	4.6	6.9	11.5	MSi2k
	Disk	0.1	0.4	0.4	0.6	PB
	Tape	1.1	4.9	9	12	PB
	WAN	3	5	8	12	Gb/s

Tier-1 Centers

- 1/7
- Wide variation in size of Tier-1 centers

A Tier-1	CPU	1.3	2.5	3.5	6.8	MSi2k
	Disk	0.3	1.2	1.7	2.6	PB
	Tape	0.6	2.8	4.9	7.0	PB
	WAN	3.6	7.2	10.7	16.1	Gb/s

US-Tier-2 Centers

US Tier-2	2008	
CPU	~1	MSi2k
Disk	~.200	PB
WAN	1--10	Gb/s

Data Driven Baseline

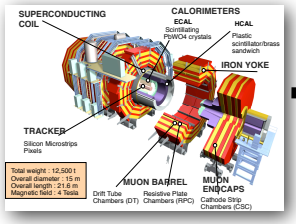
Data placement drives activity at the Tier-0 and Tier-1 centers in the CMS baseline model.

- Data is partitioned by the experiment as a whole
- Tier-0 and Tier-1 are resources for the whole experiment
- Leads to very structured usage of Tier-0 and Tier-1
  - Tier-0 and Tier-1 centers are CMS experiment resources and activities are nearly entirely specified
    - Primary reconstruction, Re-reconstruction, Data and Simulation Archiving, Data and Simulation Serving, and Data Skimming

Tier-2 Centers are the place where more flexible, user driven activities can occur

- Portion of resources are controlled by the local community
- More chaotic analysis activities
- Very significant computing resources in need of good access to data

Modified Hierarchical Model



Tier-2 centers may have relationships with Tier-1 centers for management, support, and operations

- Data access may come from a variety of Tier-1 centers

Operational Goals

CMS needs to be at production scale services in 2008

- Assuming we cannot easily more than double the scale each year, we should be able to demonstrate 25% of the 2008 scale now
  - Network transfers between T0-T1 centers
    - 2008 scale is roughly 600MB/s
  - Network transfers between T1-T2 centers
    - 2008 Peak rates from Tier-1 to Tier-2 of 50-500MB/s
  - Selection Submissions and Transfers to Tier-1 centers
    - 2008 submission rate 50k jobs per day to integrated Tier-1 centers
  - Analysis Submissions to Tier-2 Centers centers
    - 2008 Submission rate 150k jobs to integrated Tier-2 centers
  - MC Production jobs at Tier-2 centers
    - 2008 rate is 1.3 x10^9 Events per year

Global Schedule

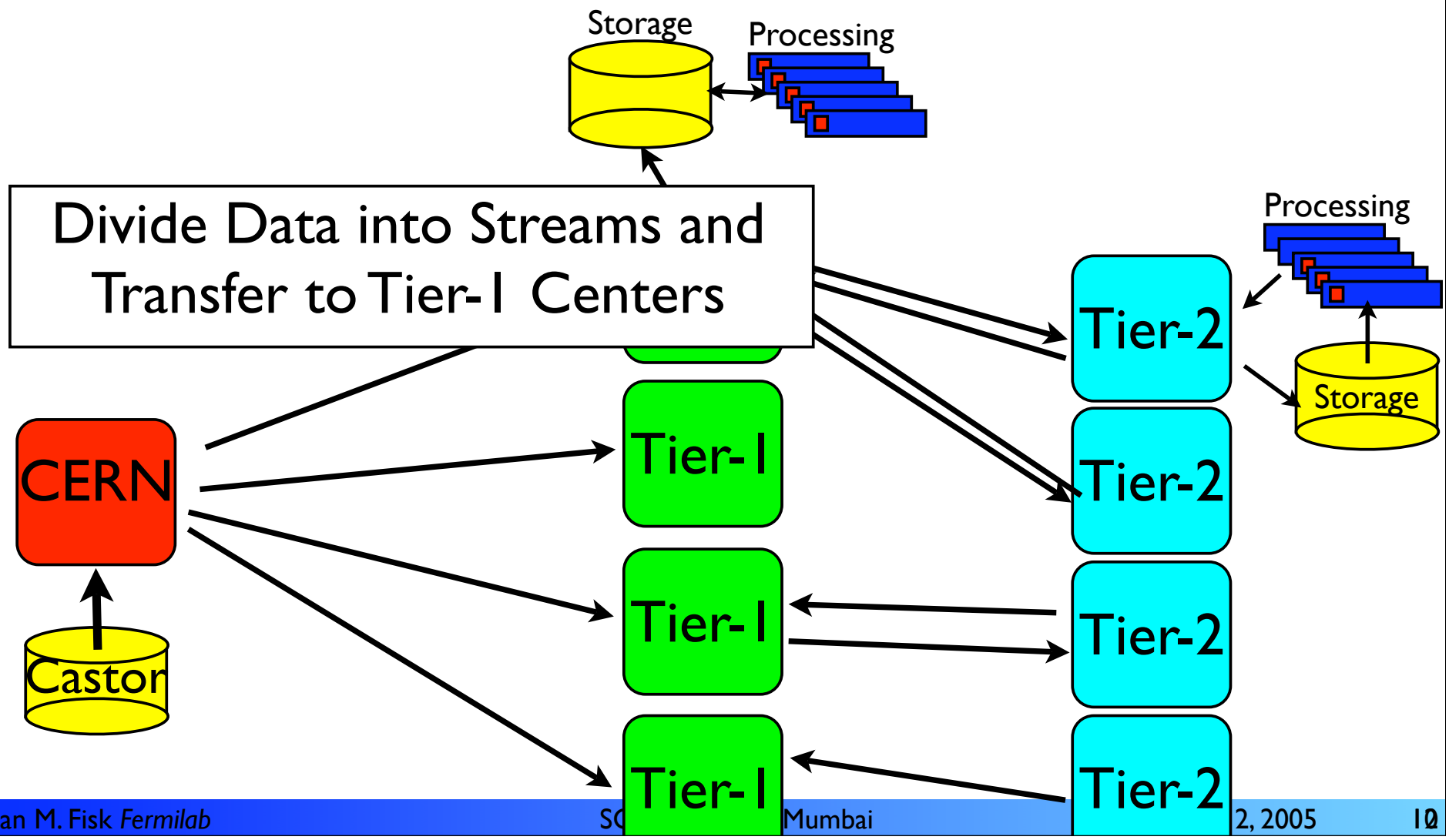
External Items

- gLite 3.0 is released for pre-production testing March 1st
- gLite 3.0 is rolled onto production in the middle of April

CMS Items

- CMS is rolling out a new framework, deploying new data management system and new simulation production infrastructure
  - CMS will have 10TB of new event data at the beginning of April
  - CMS will be able to produce 10M with the new production infrastructure in May
  - Beginning of June CMS would like a two week functionality rerun of of goals of SC3
    - Demonstration and preparation for the 2006 Data Challenge CSA06
  - July and August CMS will produce 25M events per month (roughly 1TB per day of data for CSA06
  - September - October is CSA06

Tier-0 to Tier-1 Basic Flows



Tier-1 Transfers

In 2008 CMS expects ~300MB/s being transferred from CERN to Tier-1 centers on average

- Provision roughly twice that
- Assume peaks to recover from downtime and problems with a factor of two

Demonstrate aggregate transfer rate to 300MB/s sustained on experiment data by the end of year to tape at Tier-1 by the end of the challenge

- 150MB/s in the spring

At the start of the experiment, individual streams will be sent to each Tier-1 center

- During the transfer tests, data to each tier-1 will likely have a lot of overlap

Pieces Needed for CMS

Phedex integration with FTS

- Expected middle of March

CMS Data Management Prototype

- Version 0 is released. Version 1 expected in the middle of Match
- Ability to define and track datasets and locations

New Event Data Model Data

- Expect our first 10TB sample roughly on April 1
- The ability to transfer data on files is a component, but the experiment needs to transfer defined groups of files, validate integrity and make them accessible.

SC3 Rerun clearly demonstrated capabilities of transfer at this rate to disk

- CMS would like to replicate multiple copies of 10TB sample from Tier-0 to Tier-1 tape at a rate of 150MB/s
  - This should coincide with tape throughput tests planned in SC4
  - We would also like to exercise getting them back for applications

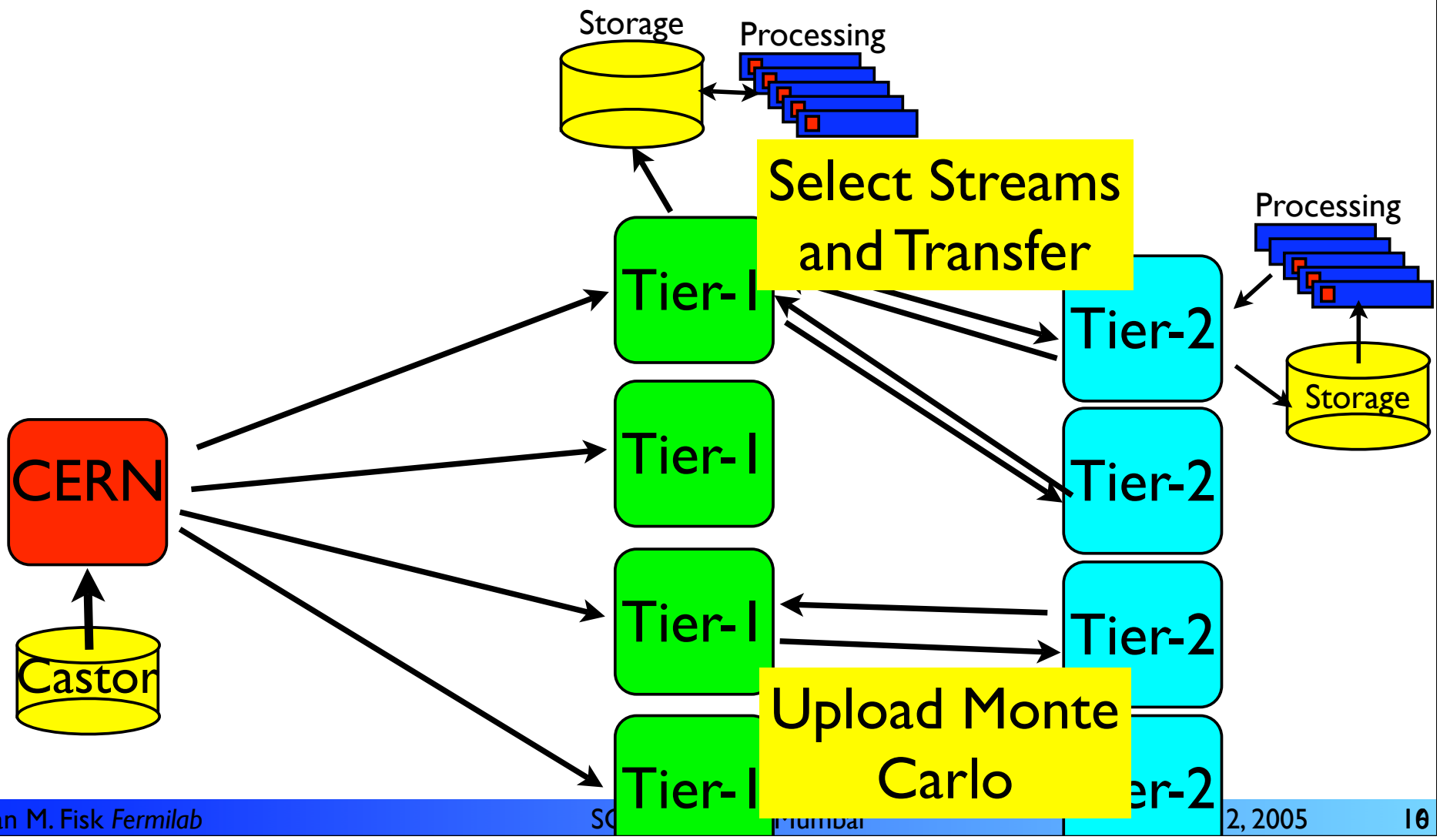
Rate Goals Based on Tier-1 Pledges

Tape Rates during first half of the year with experiment datasets

- ASGC: 10MB/s to tape
- CNAF: 25MB/s to tape
- FNAL: 50MB/s to tape
- GridKa: 15MB/s to tape
- IN2P3: 15MB/s to tape
- PIC:30MB/s to tape
- RAL: 10MB/s to tape

By the end of the year, we should double the estimates

Tier-1 to Tier-2 Basic Flows



CMS Pieces

CMS Tier-1 to Tier-2 transfers in the computing model are likely to be very bursty and driven by analysis demands

- Network to Tier-2 are expected to be between 1Gb/s to 10Gb/s assuming 50% provisioning and a 25% scale this spring
  - Desire is to reach ~10MB/s for worst connected Tier-2s to 100MB/s to best connected Tier-2s in the Spring of 2006.

The Tier-2 to Tier-1 transfers are almost entirely fairly continuous simulation transfers

- The aggregate input rate into Tier-1 centers is comparable to the rate from the Tier-0.
  - Goal should be to demonstrate 10MB/s from Tier-2s to Tier-1 centers
    - 1TB per day
    - The production infrastructure will be incapable of driving the system at this scale, so some pre-created data will need to be used.

Schedule and Rates

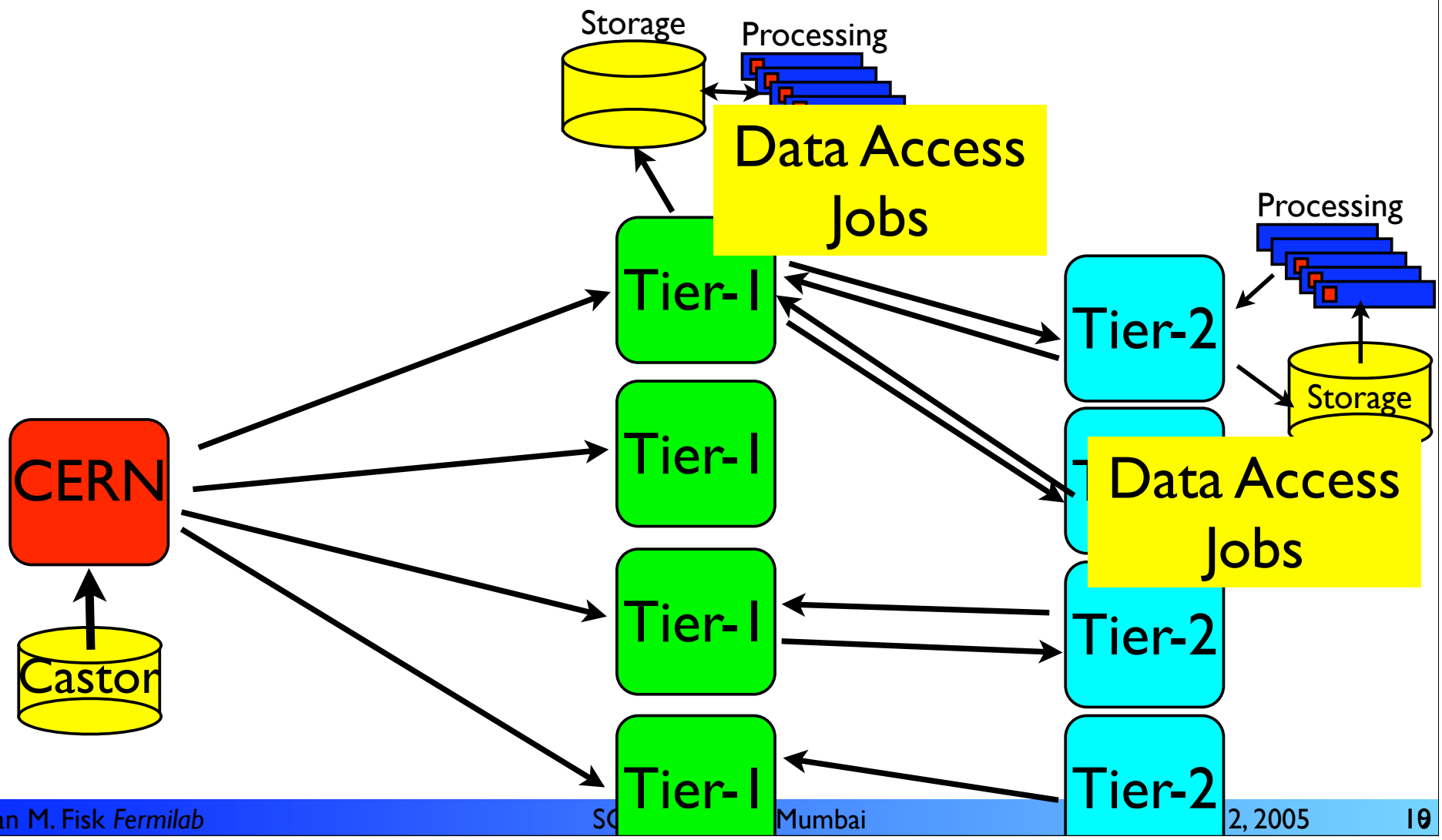
Schedule Items

- March 15 FTS driven transfers from PhEDEx
- Starting in April CMS would like to drive continuous low level transfers between sites that support CMS
  - On average 20MB/s (2TB per day)
  - There is a PhEDEx heartbeat system which is almost complete
  - Additionally CMS has identified 3 groups in opposing time zones to monitor transfers
  - Use new EDM data sample
    - In April we only expect to have 5 days worth of unique new EDM data

In addition to low level transfers, CMS would like to demonstrate the bursting nature of Tier-1 to Tier-2 transfers

- Demonstrate Tier-2 centers at 50% of their available networking for hour long bursts

Accessing the Data



Job Submission

CMS calculates roughly 200k job submissions per day in 2008

- Calculation makes a lot of assumptions about the number of active users and kind of data access.

Aim for 50k jobs per day during 2006.

- CMS will begin transitioning to gLite 3.0

A larger number of application failures come from data publishing and data access problems than from problems with grid submission

- Need new event data model and data management infrastructure to have reasonable test

Currently CMS has only basic data access applications.

- Data skimming and selecting applications will be available by June for new EDM

Job Submission Schedule

By the beginning of March we expect a analysis grid job submitter capable of submitting to new event data model jobs

By April we will enable user submissions to the limited sample of new EDM data

- At this point we can start scale tests of gLite 3.0

During the first two weeks of June during the CMS focused section we would like to hit 25k jobs per day.

July and August we will be performing simulation at a high rate

Exercising Local Mass Storage

The CMS data model involves access to data that it stored in large disk pools and in disk caches in front of tape

- Demonstrating reasonable rate from mass storage is important to success of analysis
- Goal in 2008 is 800MB/s at a Tier-1 and 200MB/s at a Tier-2
  - 2006 Goal is 200MB/s at Tier-1 or 1MB/s per batch slot for analysis

During scale testing with new EDM in April CMS will aim to measure local mass storage serving rate to analysis applications.