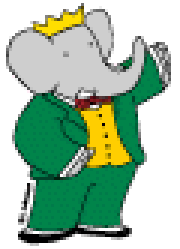
BaBar Computing
From Collisions to Physics Results



Peter Elmer – Princeton University
For the BaBar Computing Group

CHEP'04 – 27 September, 2004

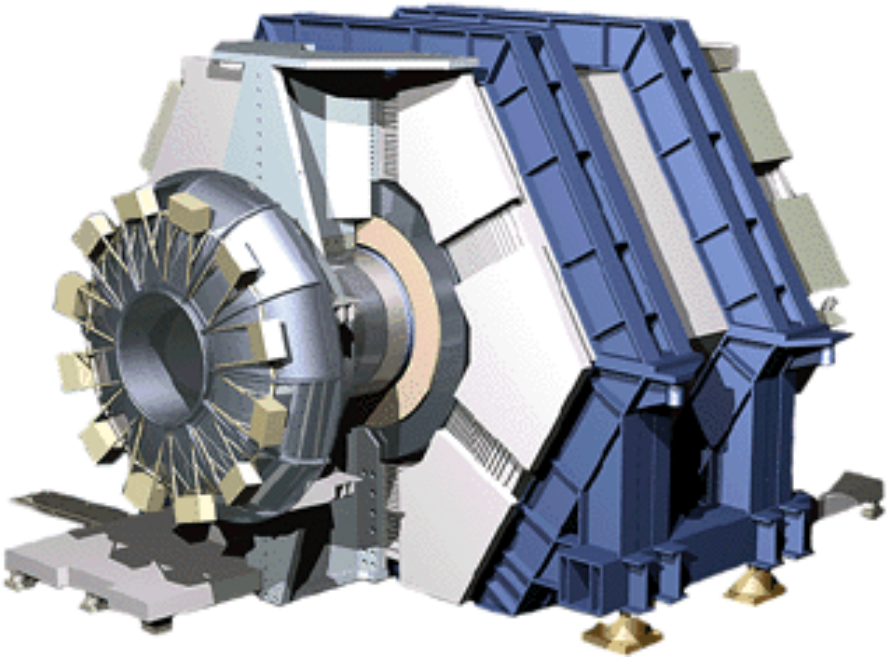
BaBar/PEP-II at SLAC

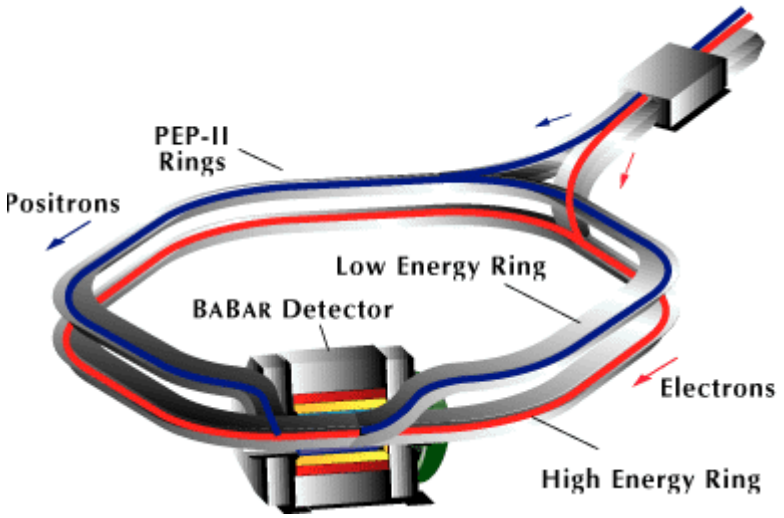
593 physicists and engineers

77 institutes/11 countries

BaBar studies primarily B physics

Taking data since May, 1999





PEP-II at the Stanford Linear Accelerator Center

Asymmetric e+e- collider

9.0 x 3.1 GeV beams

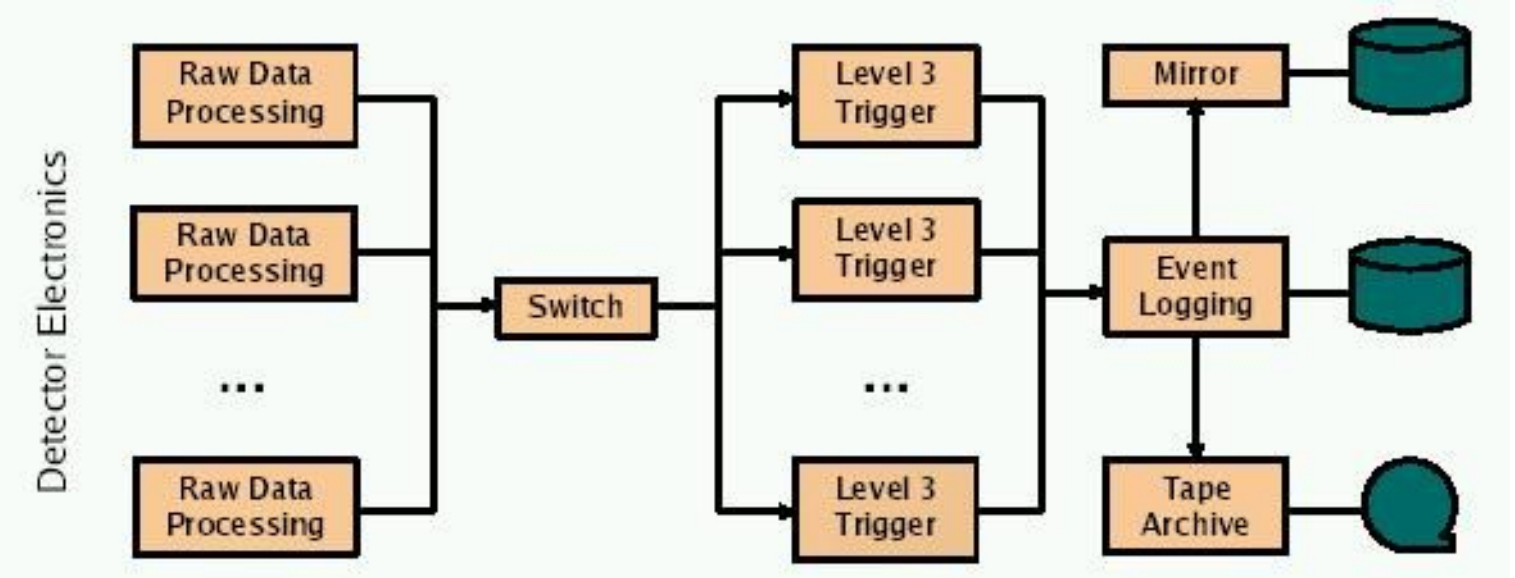
Upsilon(4s) resonance

Overview

- L3 trigger
- Prompt Reconstruction
- Simulation Production
- Analysis
- New Computing Model

L3 Trigger

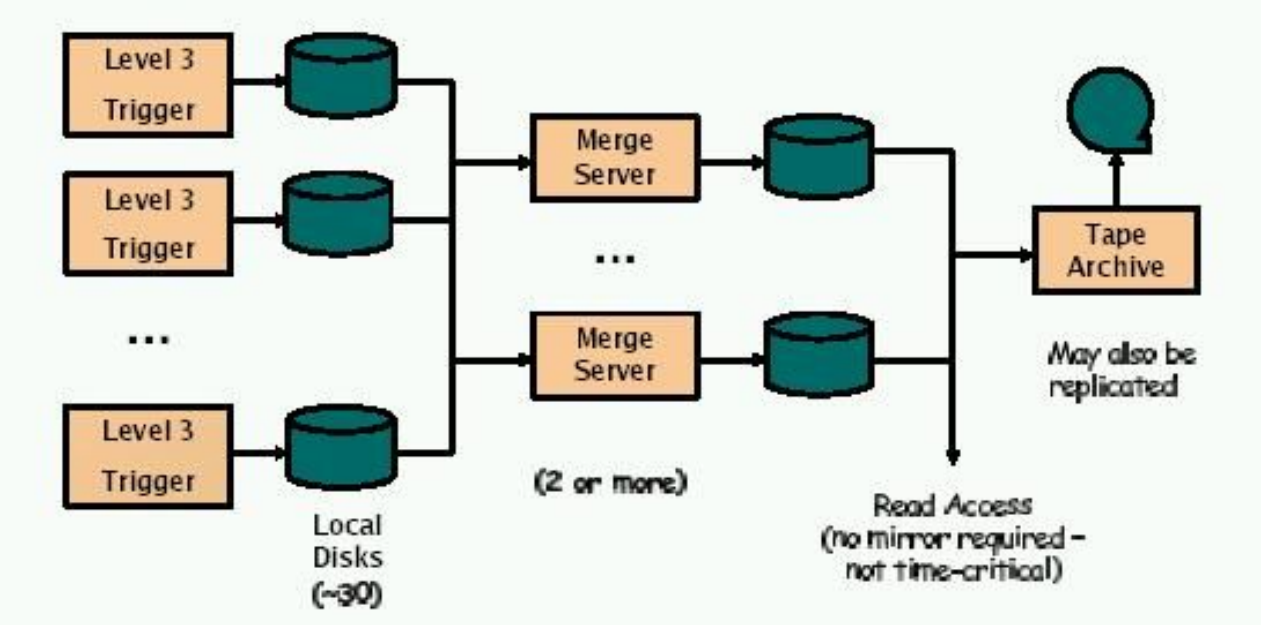
- Input rate from L1 trigger – 3kHz
- Output ~300Hz at peak lumi (32kB/evt) ~ 9MB/s
- Runs on 30 linux cpus



- Events output via TCP connection to a single “logging” server
- Adequate for first years, but does not scale forever.

L3 Trigger Upgrade

- Goal was output at 50kB@500Hz, 50kB@5kHz peak
- Local logging of data with subsequent harvest/merge
- Demonstrated that we can log the full L1 rate (32kB@~3kHz) w/o additional deadtime



To date: ~34k raw data files and 260 TB of data (for 244 ifb)

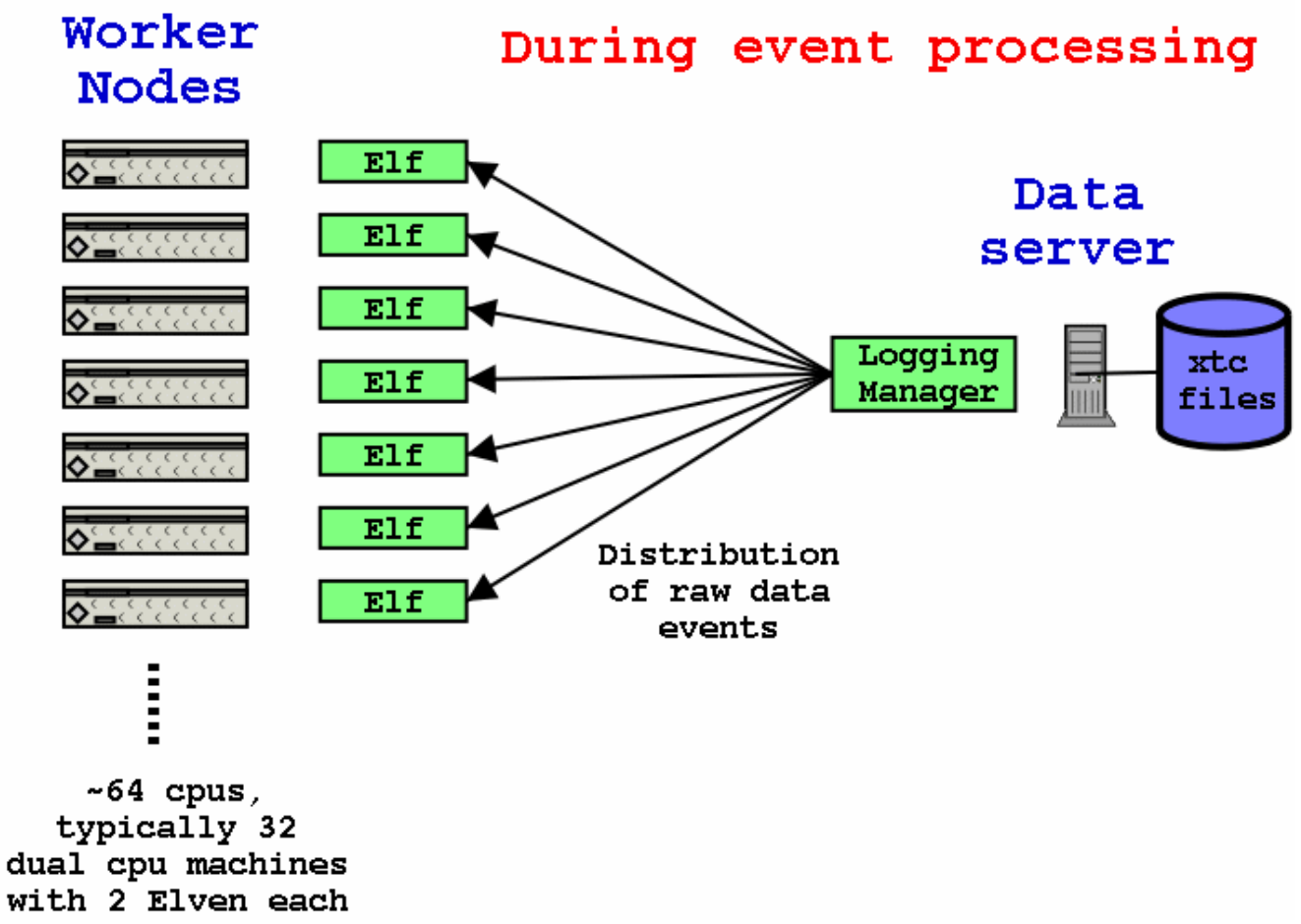
Prompt Reconstruction

Talk by Antonio Ceseracciu
Monday at 14:00 (ID 210)

- **Reconstruction of real data uses a dedicated (re)processing center setup by INFN-Padova/BaBar-Italia**
- **1TB/ifb raw out from SLAC, 150GB mini/micro back from Padova**
- **490GHz (PIII equivalent) to reco 1 ifb/day, including inefficiencies**

Prompt Reconstruction

Talk by Antonio Ceseracciu
Monday at 14:00 (ID 210)

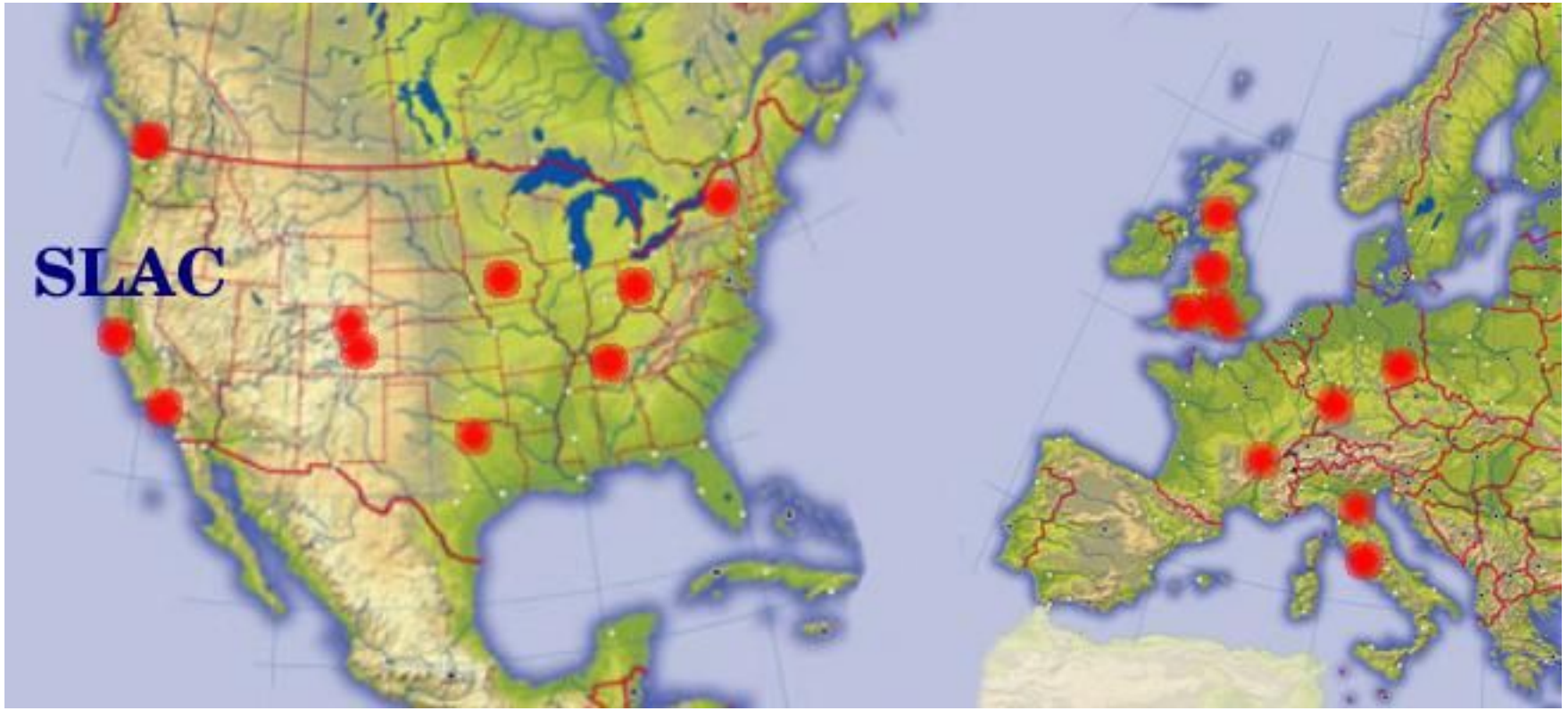


Single server reads the raw data file and serves events to reco processes ("Elven").

They write data (in parallel) to their local disks. It is then subsequently harvested & merged.

Simulation Production

Talk by Douglas Smith
Wed. at 14:00 (ID 339)



- **25 centers (typically small university sites), ~1800 cpus**
- **Try to keep manpower ~ 0.1FTE/site after initial setup**
- **Total data output of all simulation sites currently 200-500GB/day**

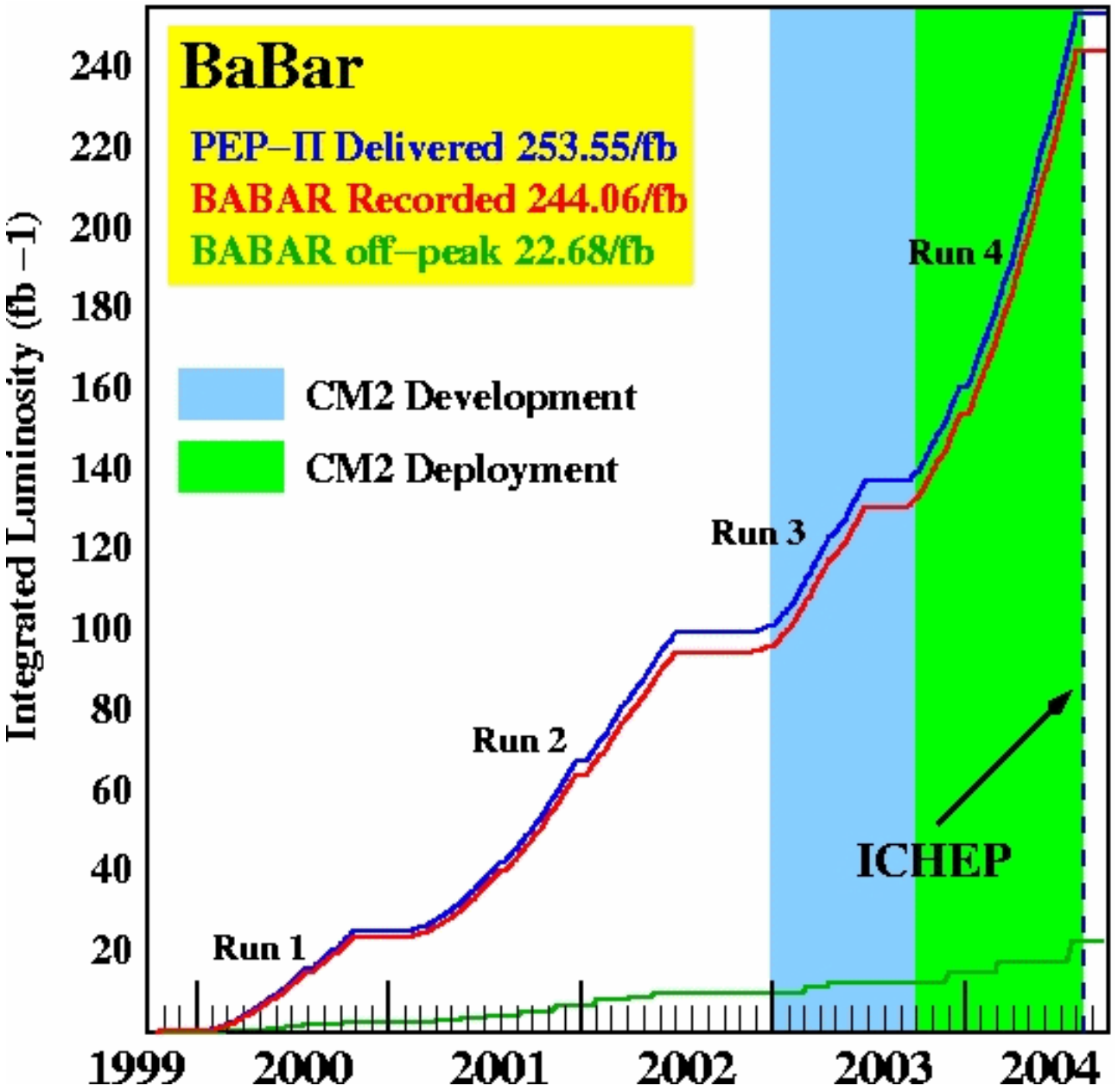
Analysis Computing

- **User analysis at 5 large computing centers plus many universities.**
- **Datasets are assigned by BaBar to each site such that users can (mostly) work at one site and so we use all of the available resources.**
- **Users log on to a front-end machine at each site to work.**

New Computing Model – CM2

- **Large changes in BaBar offline computing**
- **New eventstore**
- **New data content**
- **New analysis model**
- **New bookkeeping/data distribution/etc.**

New Computing Model – CM2



- **Prototypes and collaboration decision during 2002**
- **Real development during 2003**
- **Deployment fall 2003, early 2004**
- **Convert last processing of all data taken before CM2 deployment**

Processing model

- Mental model: **parallel data processing and data reduction**

Processing model

- Mental model: **parallel data processing and data reduction**
- The "core mission" of a job running on a worker node is to:
  - **read data, do calculations, write data**
- For stability/scalability each job should be independent of all other jobs:
  - limit or remove read and update access to global data structures and resources

Processing model

- Mental model: **parallel data processing and data reduction**
- The "core mission" of a job running on a worker node is to:
  - **read data, do calculations, write data**
- For stability/scalability each job should be independent of all other jobs:
  - limit or remove read and update access to global data structures and resources
- The largest problems with the Objectivity-based eventstore were the file catalog and the collection layer.

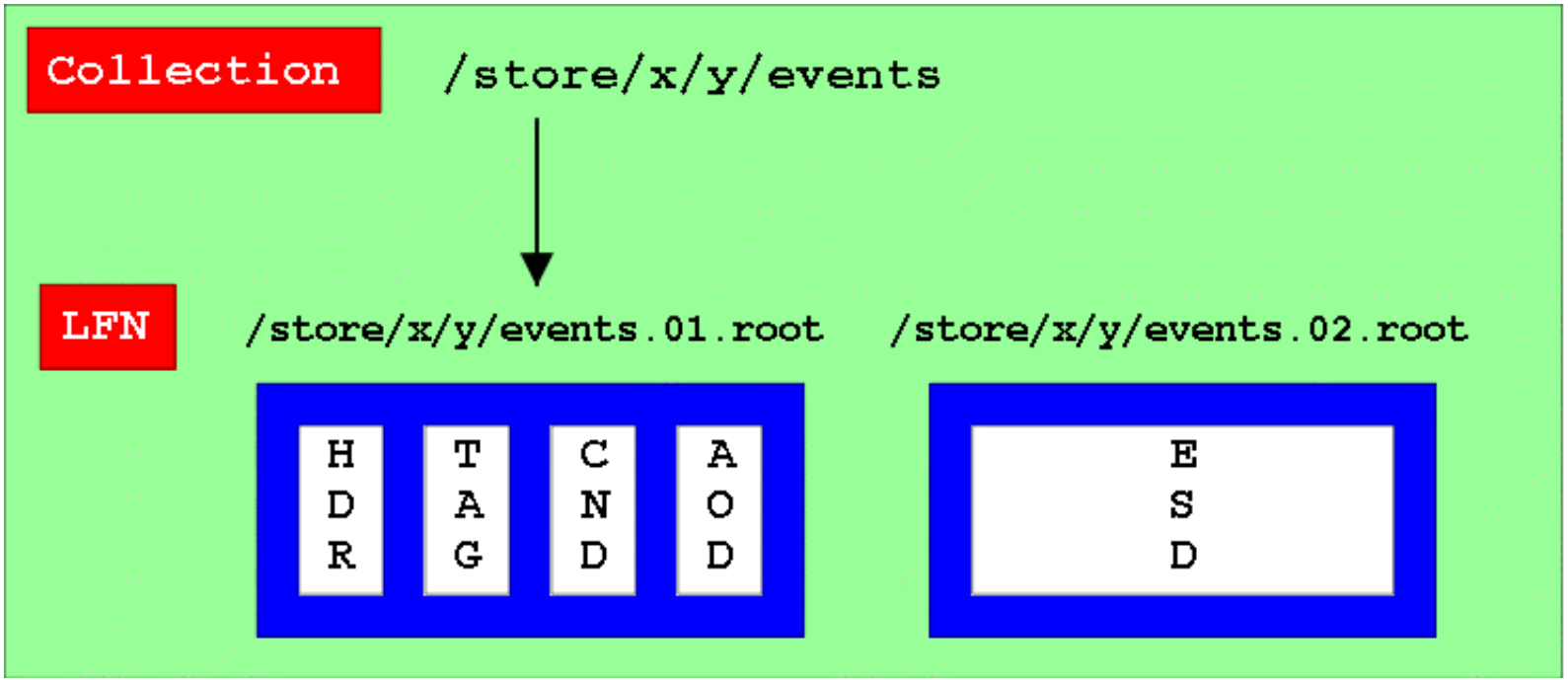
The real question for us was:

How close can we be to "Embarrassingly Parallel"?

New Eventstore

Talk by Matthias Steinke
Wed. at 17:10 (ID 172)

- Use ROOT I/O for object persistence. We support:
  - Event "components" placed in one or more files
  - Object references (within the event), etc.

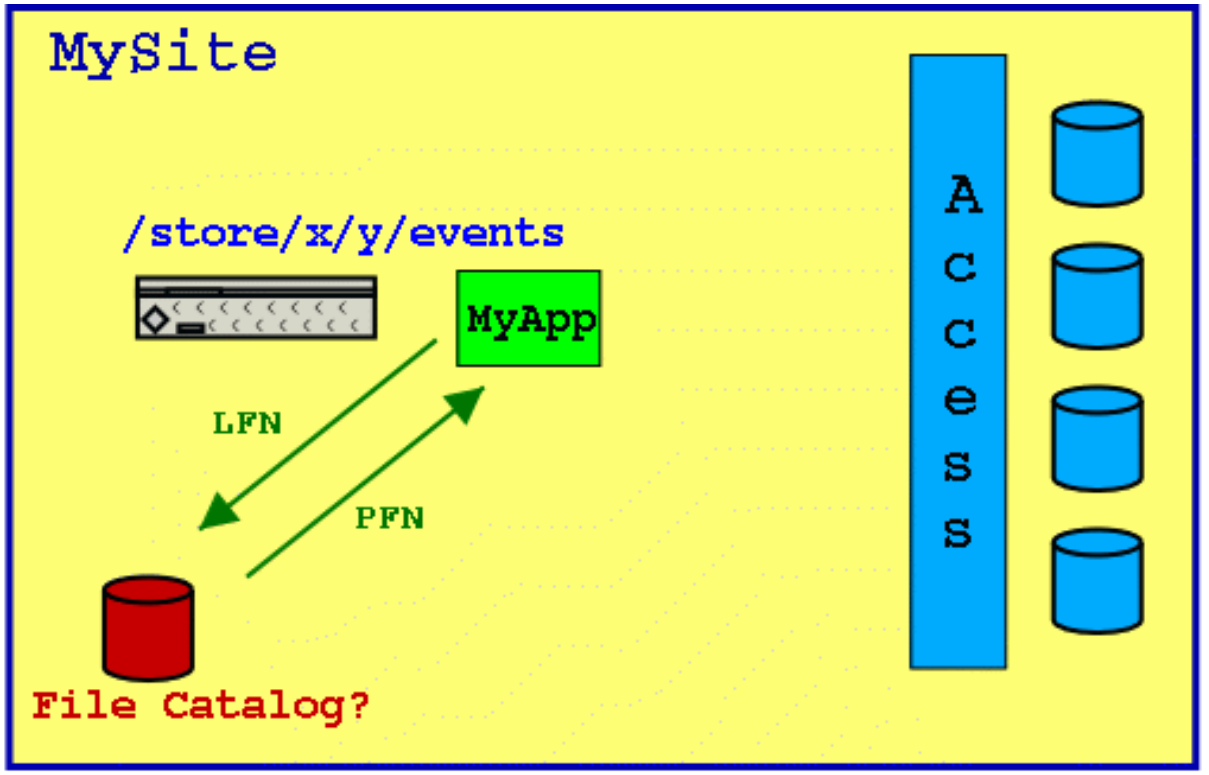


- The event header tells us which LFN contains the event data for each component (and where to find it in that LFN).

Data Access

Each job is configured with a ***site-independent*** collection name

Only when a job is actually running on a worker node is this turned into ***site-specific*** data access PFN's:

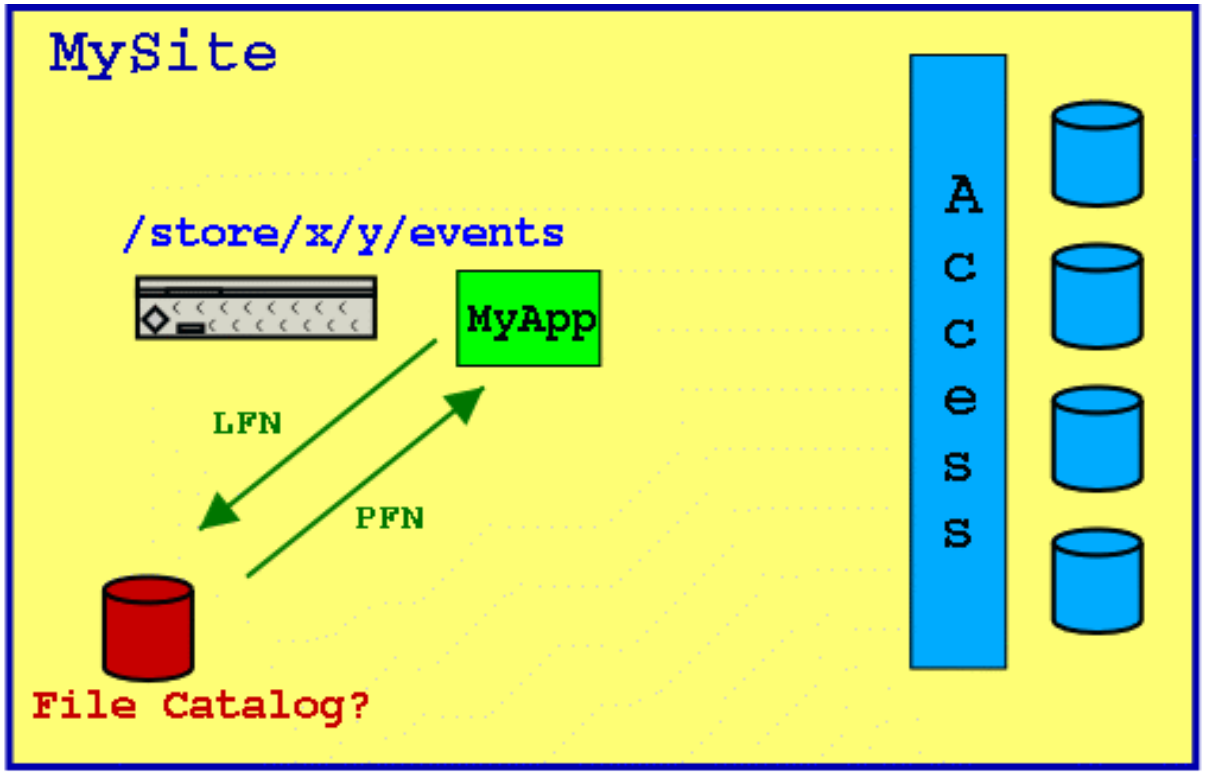



```
LFN = /store/x/y/events.01.root
PFN = root://mysiteaccess//store/x/y/events.01.root
```

Use file catalog to map from LFN to PFN?

Data Access

For us the PFN could always be constructed by placing the LFN namespace in some other namespace (i.e. by prepending something)



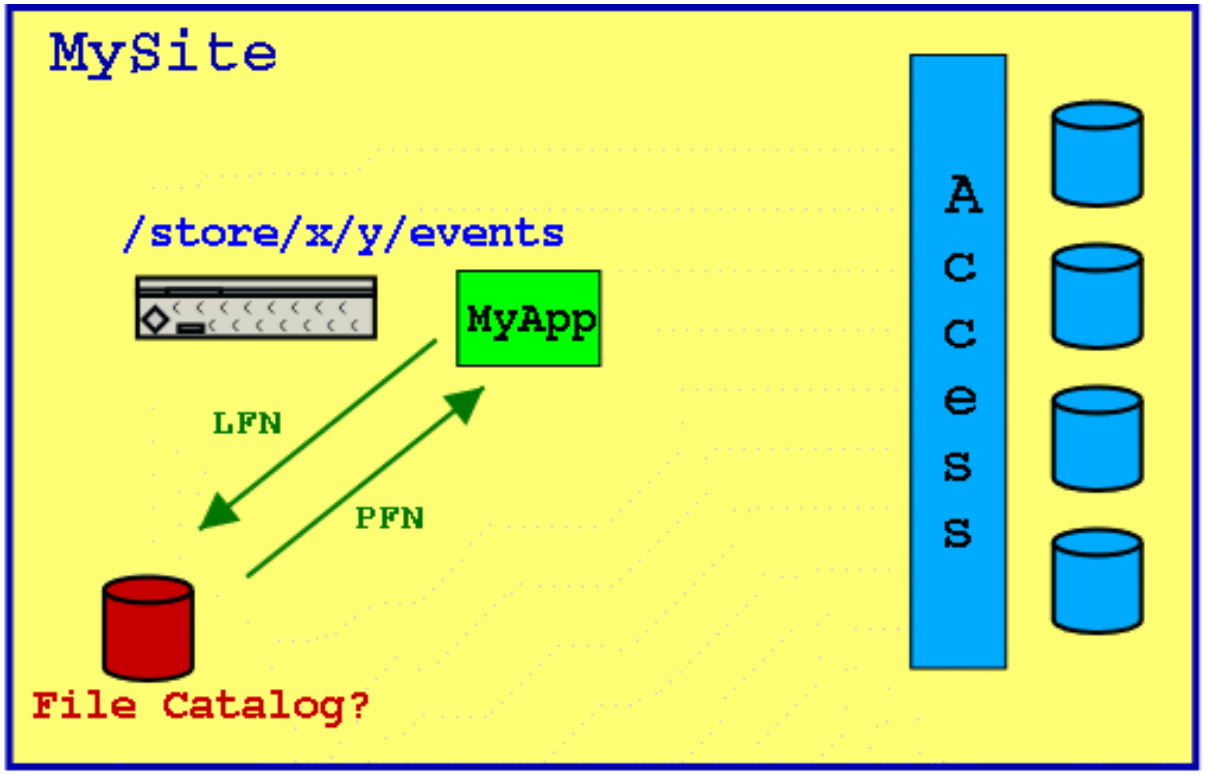
For example:

```
LFN = /store/x/y/events.01.root
PFN = root://mysiteaccess//store/x/y/events.01.root
PFN = rfio://castor/zzz/store/x/y/events.01.root
PFN = /mnt/bigdiskarea/store/x/y/events.01.root
```

Any given site would in general chose one data access type, so...

Data Access

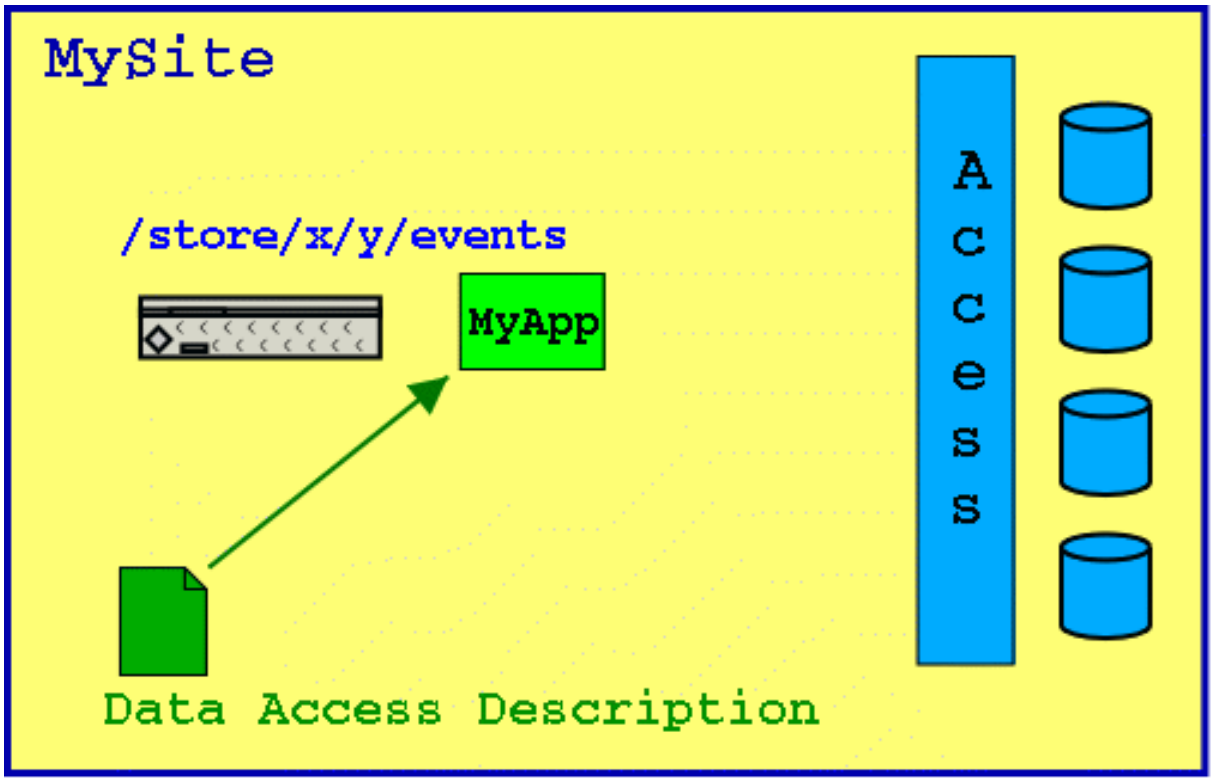
If one data access source is being used, then for any given site (e.g. MySite) the catalog would look like:




```
LFN = /store/x/y/events.01.root
PFN = root://mysiteaccess//store/x/y/events.01.root

LFN = /store/x/y/events.02.root
PFN = root://mysiteaccess//store/x/y/events.02.root

LFN = /store/a/b/events.01.root
PFN = root://mysiteaccess//store/a/b/events.01.root

etc. etc.
```

Data Access

- Use a site-specific **data access description** instead of a full file catalog and defer to the data access system
- Collapse “file catalog” into “rules” for mapping LFN to PFN at a given site.
- This file doesn't change as more data files are added at the site.



Example rules (not actual syntax):

to read /store/* use root://mysiteaccess/

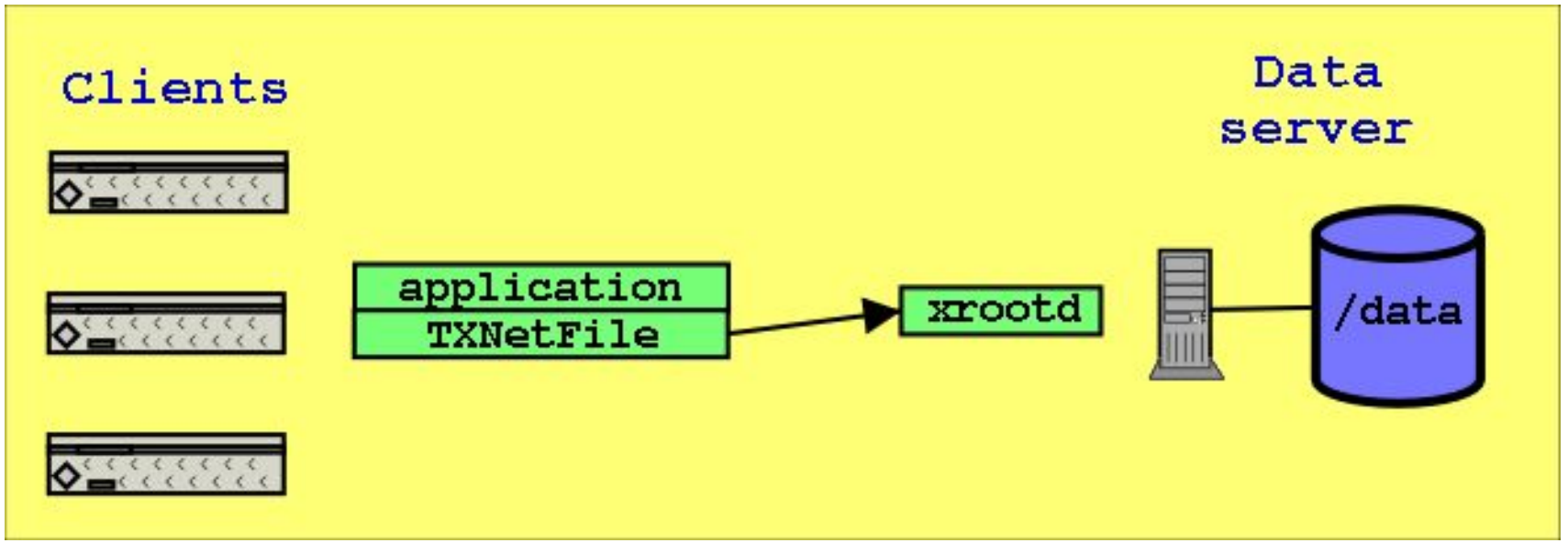
or

to read /store/* use /mnt/bigdiskarea/

xrootd

Talk by Andy Hanushevsky today at 16:30 (ID 328)

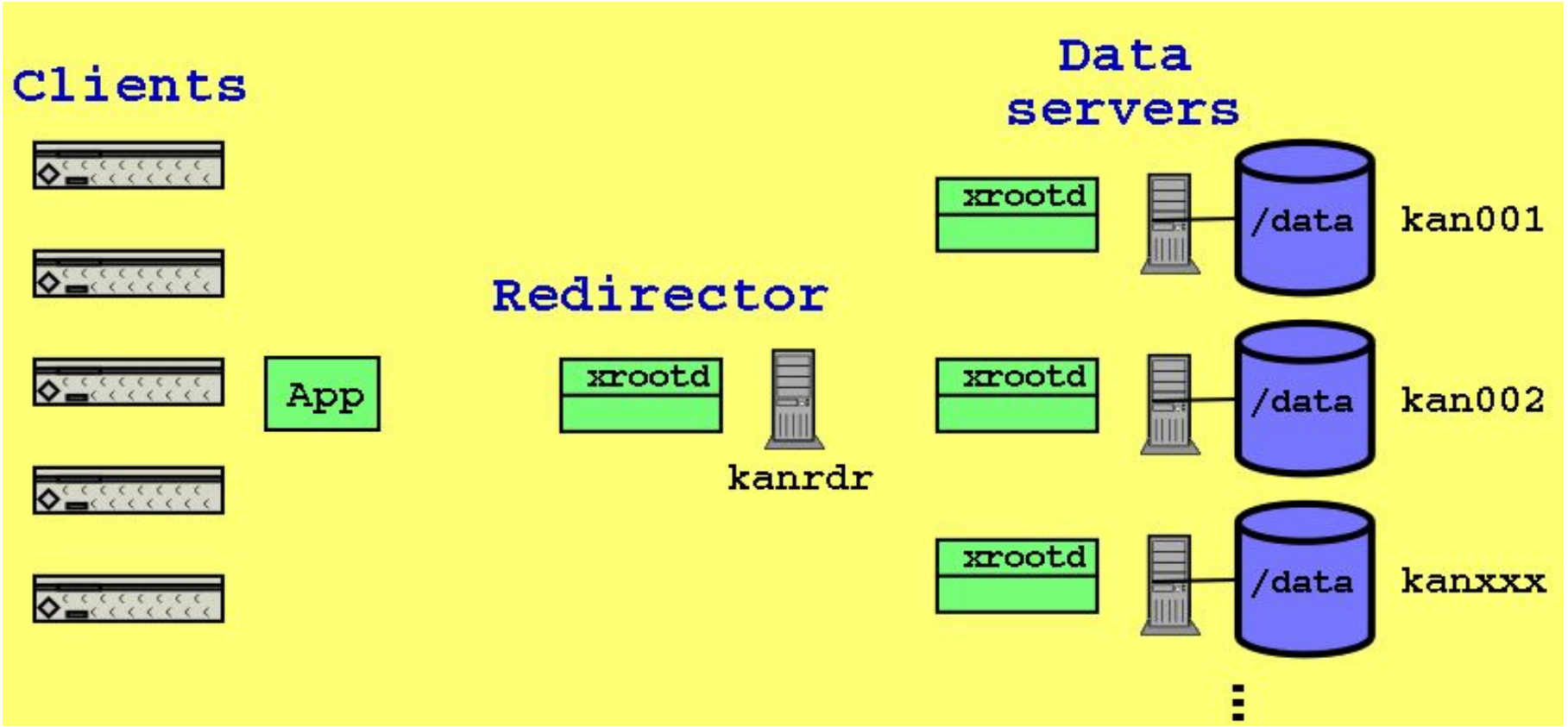
- High performance file server – improves on rootd/AMS/NFS
- TXNetFile - ROOT client plugin which speaks the xrootd protocol (A.Dorigo and F.Furano, INFN-Padova).



- Goals: scalability, performance and fault tolerance
- Rich protocol supports interesting system architectures....

xrootd

Talk by Andy Hanushevsky today at 16:30 (ID 328)

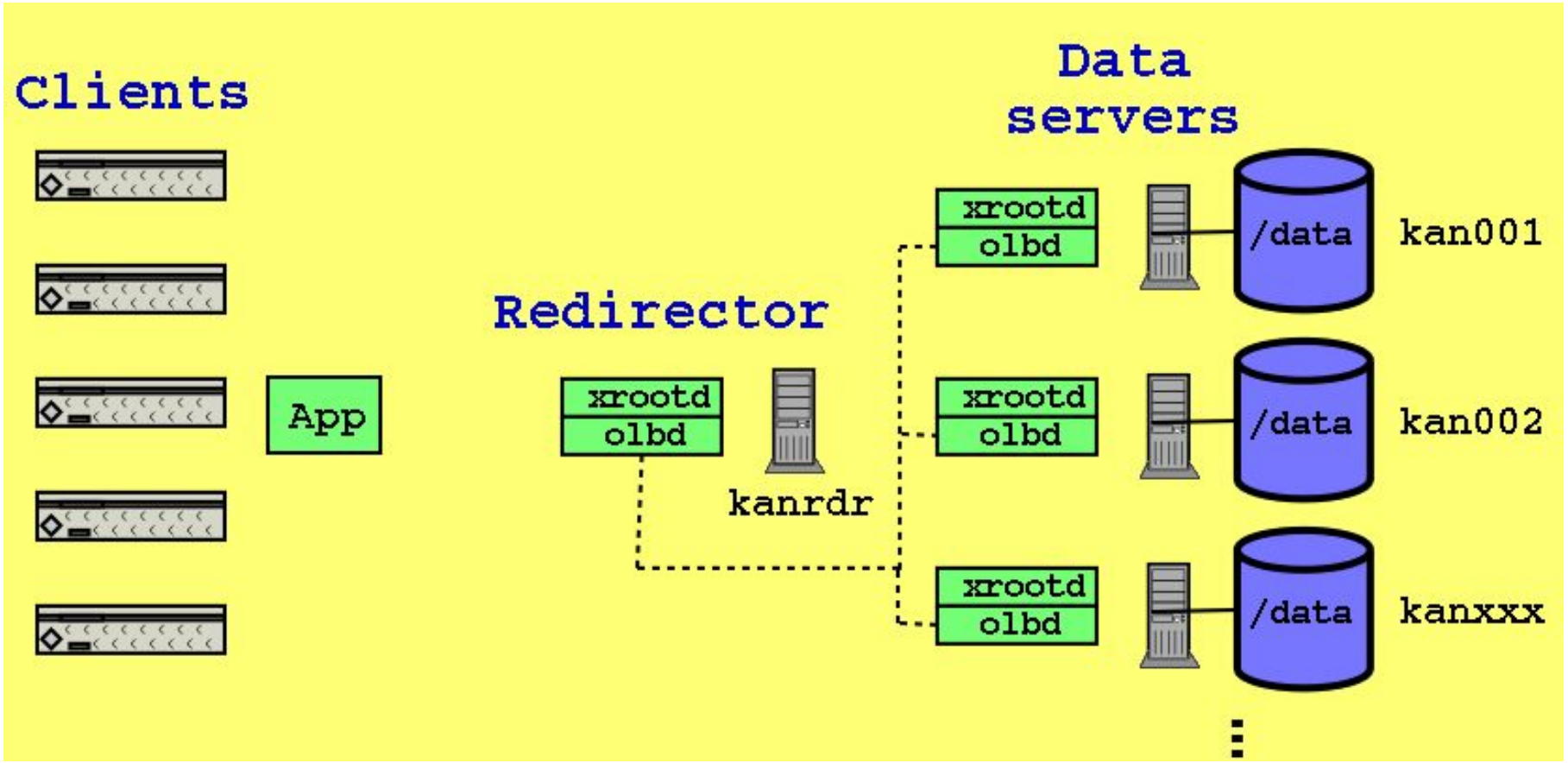


Redirection

Fault Tolerance

xrootd system

Talk by Andy Hanushevsky today at 16:30 (ID 328)

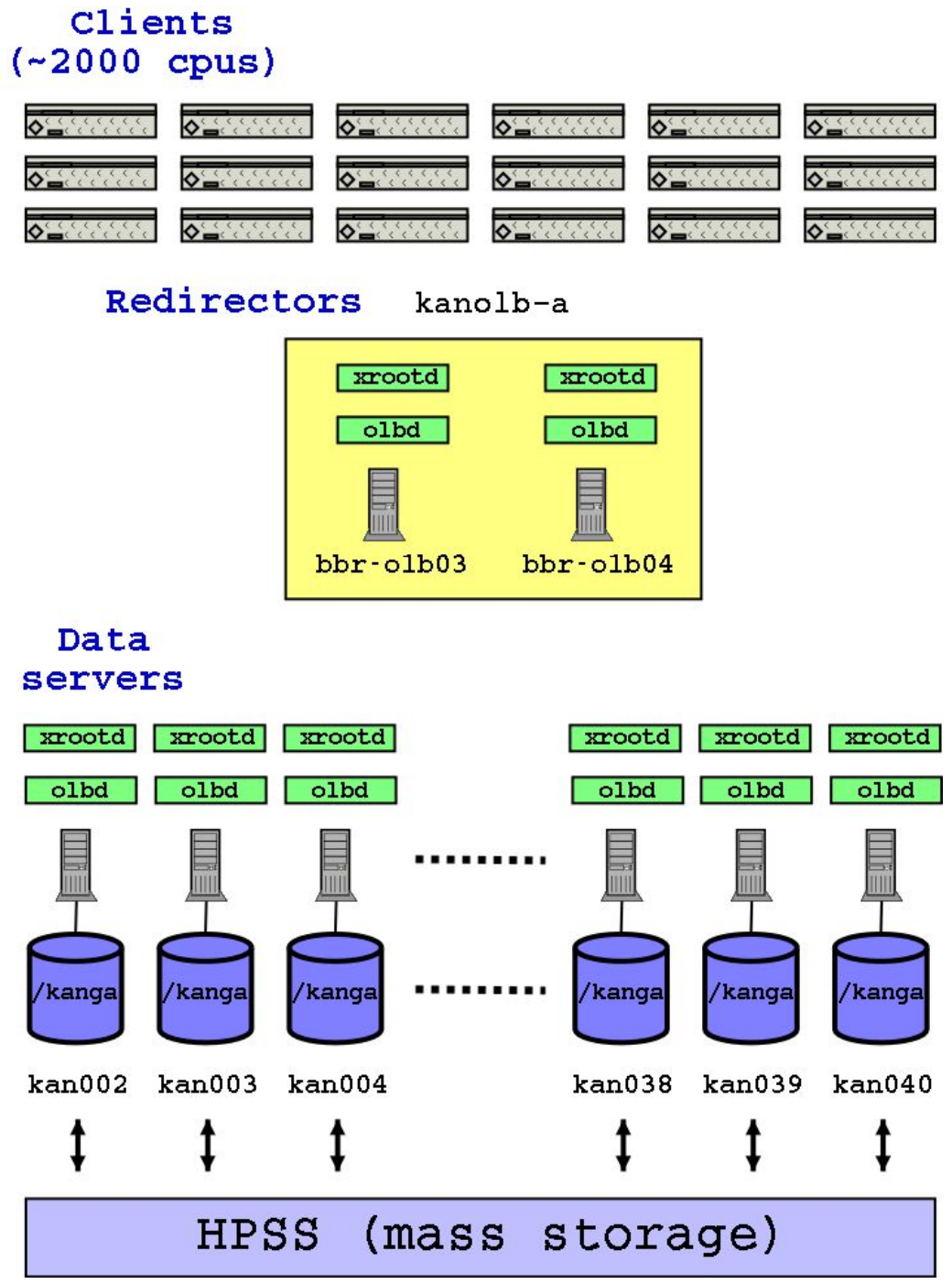


Dynamic cache of file locations

Load Balancing **Disk cache management/MSS**

SLAC

- Largest single data access system in BaBar
- Mix of analysis and production skim jobs reading data
- Production skimming uses dedicated output buffers
- Other sites (RAL, In2p3, Padova, CNAF) have similar systems



Bookkeeping

Talk by Douglas Smith
Thur. at 17:10 (ID 338)

File catalog – used by data distribution, not by batch jobs

Data distribution is a situation where a full view of all files may make sense.

Contains no site-specific "PFN" information other than an "is_local" flag, we defer the actual location to the data access system.

Collection/dataset catalog and metadata

Primary interface for users.

Users can perform complex queries (release, Run, skim type, etc.), but mostly they use predefined "datasets".

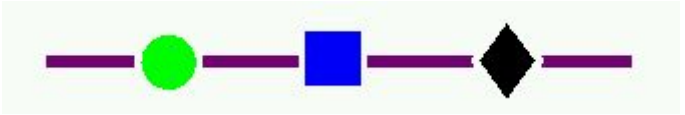
Datasets

A “dataset” consists of a list of collections:

```
prompt> BbkUser –dataset SP-uds-AllEventsSkim-Run4-R14 collection

/store/SPskims/R14/14.4.3d/AllEvents/00/000998/200310/AllEvents_000998_1539
/store/SPskims/R14/14.4.3d/AllEvents/00/000998/200309/AllEvents_000998_1540
```

Datasets are managed in a way analogous to CVS:

AllEventsSkim-Run4-OnPeak-R14-GreenCircle -- 44 collections (a tag)
AllEventsSkim-Run4-OnPeak-R14-BlueSquare -- 66 collections (a tag)
AllEventsSkim-Run4-OnPeak-R14-BlackDiamond -- 76 collections (a tag)
AllEventsSkim-Run4-OnPeak-R14 -- 80 collections (like HEAD)

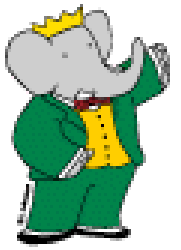


Analysis model

Talk by David Brown
Thur. at 15:40 (ID 347)

- CM2 analysis model builds on new eventstore to add end-user functionality previously missing
- Single data content for both analysis and reconstruction
  - Replace overly complex “rec” and overly simple CM1 micro with a “mini”, derived from reconstruction (and experience)
  - Analysis “micro” is now a proper subset of “mini”
- Support skimming and user-customized output content
  - Composite candidates and “User data”
- Give the users tools such that they don't need to do large ntuple productions to get the data into a usable format
- Provide more options for people doing analysis

Summary



- BaBar tried the “full frontal assault” approach to computing.
- It didn't work particularly well, but we have 5 years of experience to show for it. We now try to be smarter.
- In the past 1.5 years we have made many changes such that our computing *enables* our collaboration to produce physics results.
- The enthusiasm with which our collaborators have moved to the new model indicates that we are indeed succeeding.

Technology

"La perfection est atteinte non quand il ne reste rien à ajouter, mais quand il ne reste rien à enlever." - Antoine de Saint-Exupery

"The paper version of the agenda is more up-to-date than the electronic version"

- CHEP03

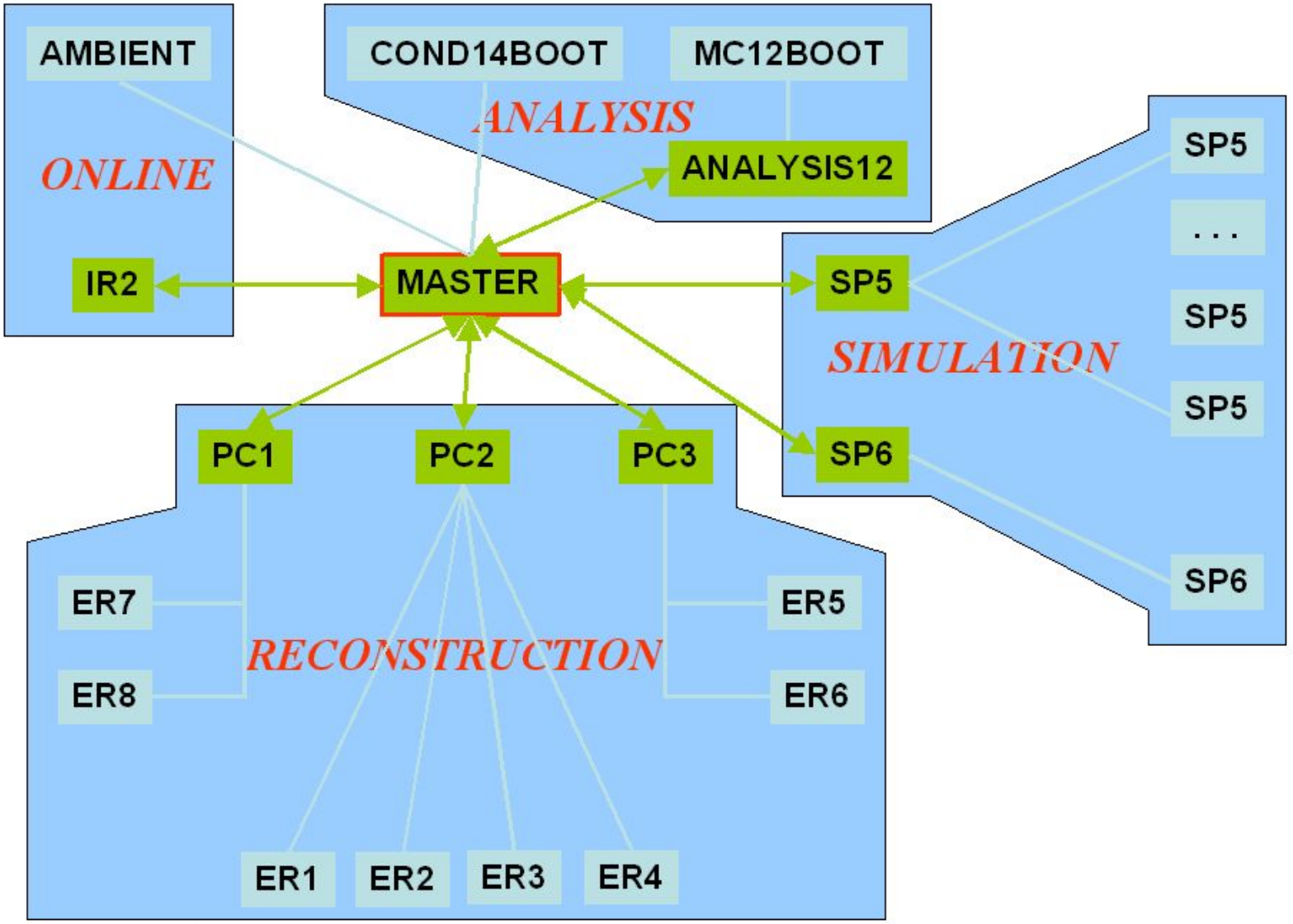
Backup slides

Breakdown of data

- Xtc - 34k files and 260 TB of data
- Objy – 900k files and 931 TB of data
- Classic Kanga ~ 40TB
- CM2 Kanga – 290k files and 162TB
- Conditions ~ 32GB
- Configurations ~100MB
- Ambient ~ 400GB

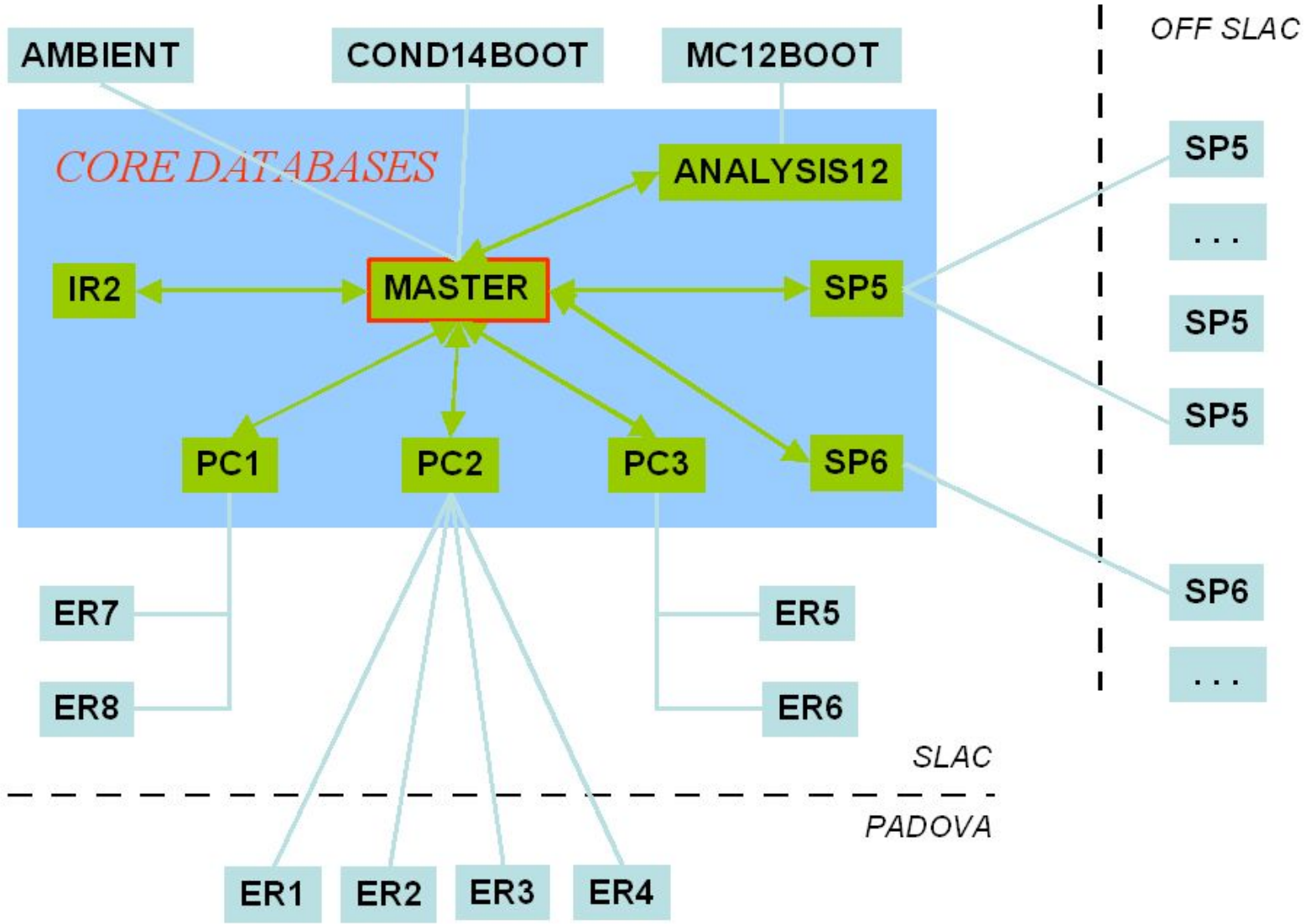
Conditions/Calibrations

Talk by Igor Gaponenko
Wed. at 17:50 (ID 316)



Conditions/Calibrations

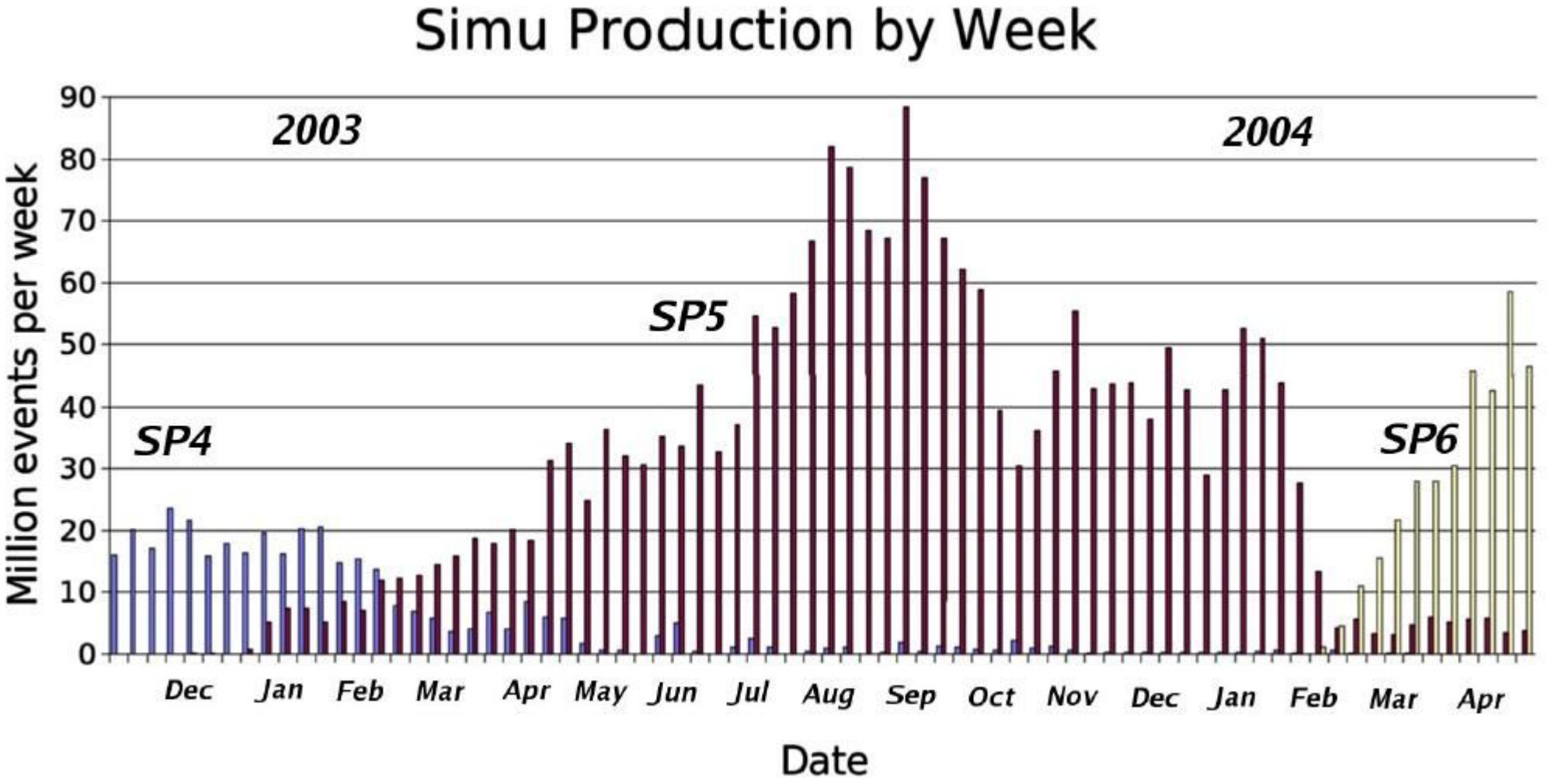
Talk by Igor Gaponenko
Wed. at 17:50 (ID 316)



Skimming

- Key element of CM2 analysis model
- Centralized production of skimmed data
- Runs over AllEvents, outputs 120 analysis-specific skims
- Part “data train”, part “calculation train” and part “do I really want to run 50k jobs myself?”
- Each output skim:
  - selects a subset of events
  - may deep-copy the micro, deep-copy the full mini or be a “pointer skim”
  - may write *skim-specific* data for selected events (results of combinatorics, “user data”)

Life Cycles


Simu Production by Week
Million events per week
90
80
70
60
50
40
30
20
10
0
2003
2004
SP4
SP5
SP6
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
Date