BABAR Computing & Software

DOE Program Review

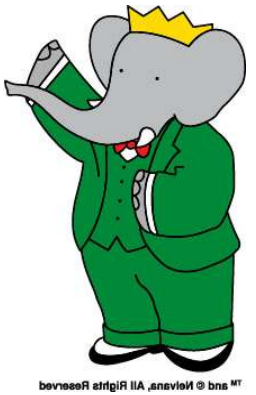
SLAC

Experimental Research Breakout Session

15 June 2005

Rainer Bartoldus

SLAC



Outline

- **Data Production Overview**
  - Getting the events from the detector into the plot: BaBar Computing in a nutshell
- **Software Development Projects**
  - Highlight 3 different, major BaBar Computing projects over the past year (which all happen to be driven by SLAC staff)
    - Breaking the limits on data logging: The new Logging Manager
    - Phasing out *Objectivity*: The database re-implementation project
    - Serving events faster: Scalable high performance data access with **xrootd**, a BaBar solution for the community

    (Disclaimer) Obviously, BaBar is a large international collaboration; Computing would not happen without the enormous contributions of many people all over North America and Europe!

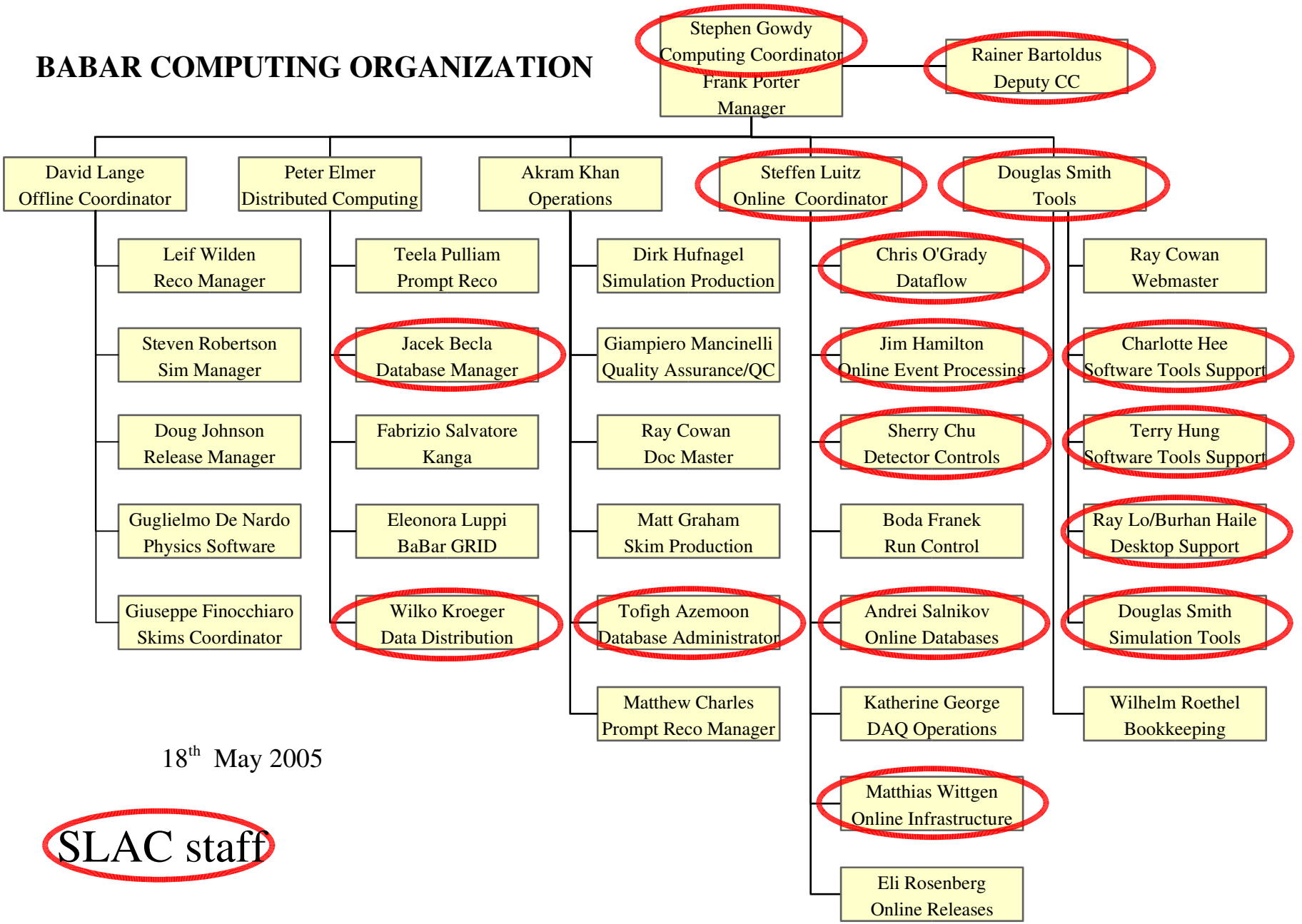
A Brief Look Back...

- At the time of last year's Program Review, we had just completed the **GreenCircle** dataset, starting on **BlueSquare**, with PEP-II and BaBar still running in full steam
  - On-peak data taking ended on Tuesday morning, July 13, 2004
  - The data was “promptly” calibrated, sent to Europe for reconstruction, transferred back to SLAC within 24 h, was QAed by Friday, skimmed, merged, and cataloged in the bookkeeping over the weekend
- Within less than a week, the following Monday, July 19, the DQG announced the **BlackDiamond** dataset
  - *“A great number of people worked very hard to bring you this data, please use it wisely and well.”*
- Eleven days later, the paper ***"Observation of Direct CP Violation in B0 -> K+ pi- Decays”,*** using these data, was submitted to PRL

Computing Model 2

- What followed was an outpour of results for ICHEP'04
- All this was made possible by the enthusiastic effort over the previous year to completely change the way BaBar does Computing, resulting in a **new Analysis Model**, and a **new Event Store**

BABAR COMPUTING ORGANIZATION



18th May 2005

SLAC staff

SLAC's Role

- **Staff (Person Power)**
  - Nearly half (16/38) of Computing management roles filled by SLAC staff
    - Dominant in Online and Tools areas
    - Key roles in Distributed Computing
    - Successfully filled Offline and Operations tasks with collaborators, including typical 6-12 month term roles:
      - Operations, Prompt Reco/Skims/Simulation Production, Data Quality etc.
- **Facilicites (CPU Power)**
  - SCS still the largest computing resource in the collaboration
  - Driving force for offloading to Tier-A, so can focus on special/urgent tasks

BaBar Tier A Sites

- **BaBar Grows into Distributed Computing**

- First, **IN2P3** in Lyon, France, originally as a replica of SLAC, taking a considerable share of analysis users

- Then **RAL**, UK, entered as an alternative analysis site for the “classic” Kanga data

- In 2002, **INFN**/Padova, Italy, joined as first Tier A to take on a production task (reprocessing of Run1-3 data); last year also Bologna (CNAF) for analysis

- Latest site is **GridKa**, Karlsruhe, Germany, to participate in skimming and now also analysis

Today all Tier As are involved in production tasks and support analysis on equal footing with SLAC; an enormous gain for BaBar!

Data Distribution

- **Tier-A Allocations**
  - Initiated Analysis Resources Task Force to evaluate "Skim" allocations across the computing centers
    - Goal is to optimize resource usage for BaBar. Doing this by concentrating AWGs at specific sites

      - **SLAC**: Tracking, PID, BReco, SemiLep, LepBC
      - **IN2P3**: Charmonium, TwoBody, RadPenguin, PartSpec
      - **RAL**: Q2Body, ThreeBody, TauQED
      - **INFN**: Charm
      - **GridKa**: S2BMix, IHBD

  - Eventually removed (to tape) *Objectivity* events at SLAC, Dec 31st
  - Removed "classic" Kanga from RAL, Jan 31st

Data Production Overview

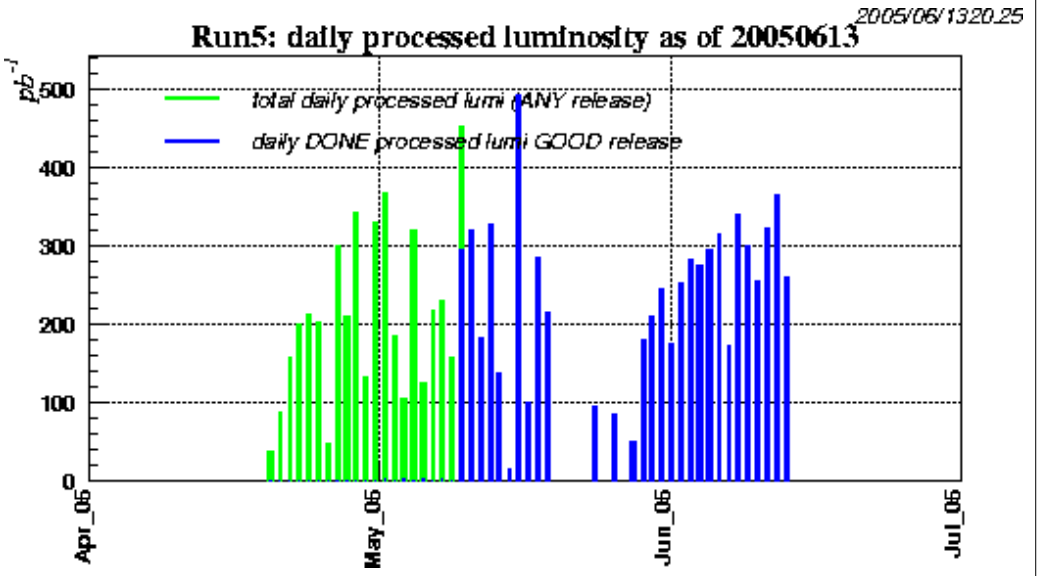
- **Online ("IR2")** — SLAC
  - We read out the detector at 2000-3000 Hz (L1), log data at 200-300 Hz (L3 farm) with minimal (1-2%) deadtime
- **Prompt Calibration ("PC")** — SLAC
  - We prefilter 7 Hz of events to perform "rolling" calibrations (PC farm, sequential, constant rate)
- **Event Reconstruction ("ER")** — INFN
  - We filter and fully reconstruct about 16 nb (ER farms, parallel, scales with luminosity)
- **Skimming** — All
  - We skim between a fraction of a percent and a few percent of reconstructed events into >100 different streams (batch farms)

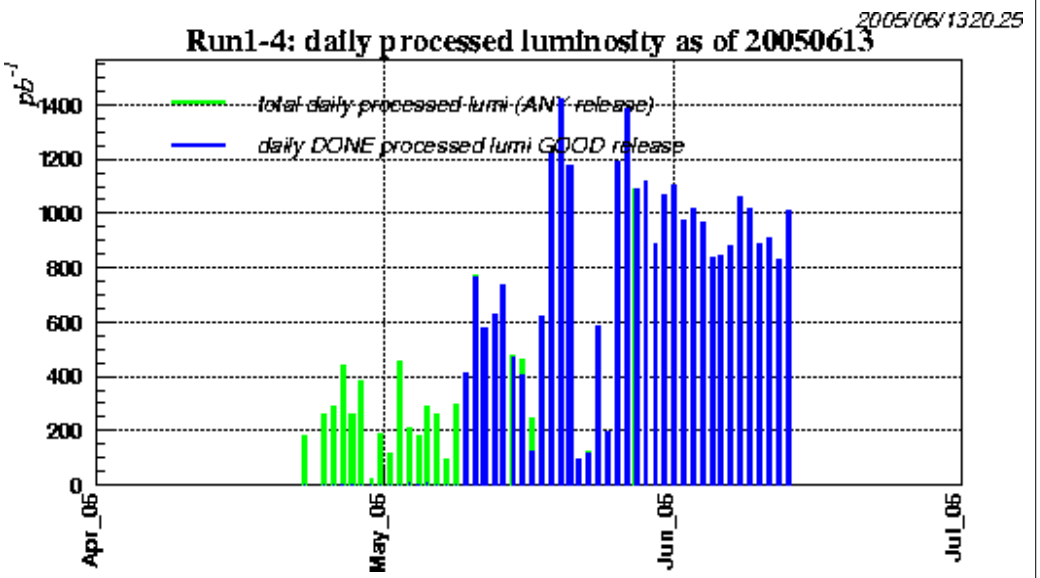
Notes on BaBar Production

- We are in a great advantage with our deeply pipelined DAQ, our Open Trigger and fast Level 3/OEP, which allows us to log more physics with higher efficiency
- The Prompt Calibration pass is the only stage that has to be serialized in order to support rolling calibrations; by adding an extra step of filtering out calibration events we made this independent of luminosity
- Skimming is essentially our answer to the “random access problem” imposed by sparse collections; a typical replication factor of 3-4 is being optimized by controlling the cut-off between deep-copy and pointer skims
  - Richard will have to say more about this in his talk this afternoon

Prompt Reconstruction

- **Run 5 Processing**
  - Padova (easily) keeping up with IR2 data taking
  - Could do > 1 fb^{-1} a day
  - Some bootstrapping of new 18-series release
- **Run 1-4 Reprocessing**
  - "Grand" reprocessing for latest reconstruction an homogeneous data set
  - Averaging ~1 fb^{-1} per day
  - To finish in time for summer conferences 2006



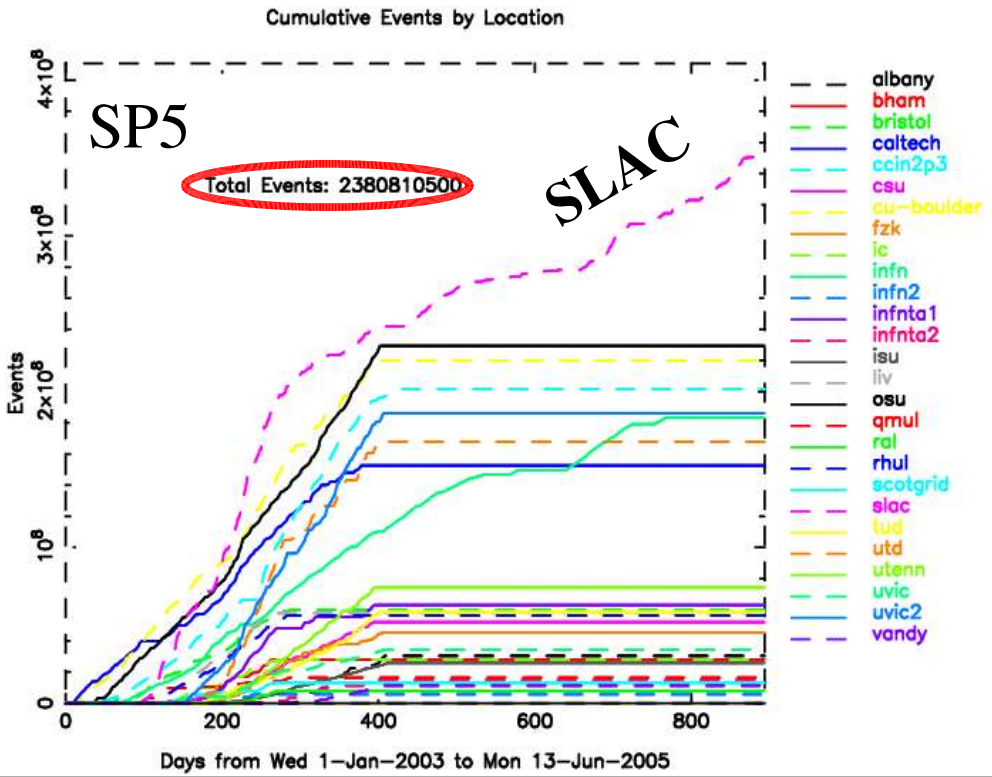


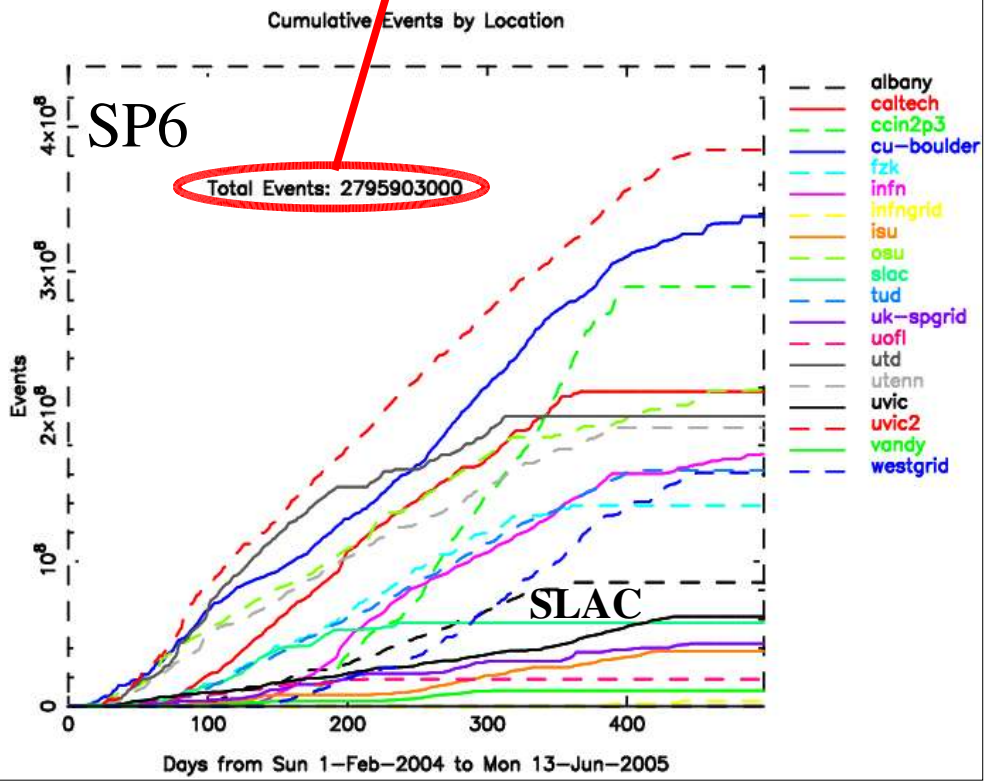
Simulation Production

- **Simulation Releases**
  - SP5 to match Run1-3 data reconstruction level
  - SP6 to match Run 4 data reconstruction
  - Prepared SP7 for Run 5, but skipped as a consequence of the 6 months down following the accident
  - Now about ready to start SP8 to match Run 1-5 (re-) processing
- **Allocations**
  - Highly parallelizable, performed at university sites (Tier C), or as background on Tier A farms to fill CPU cycles
  - Distributed over ~20 different sites

SP5 + SP6

- Target: generic BB 3x lumi, udsc + τ 1x, plus signal MC
  - SP5 (still *Objectivity*) 2.3 B events for Run 1-3
  - SP6 (Kanga) far exceeded Run 4 goal with 2.8 B events





Jim Hamilton, Steffen Luitz et al

New Logging Manager

- **The Challenge**
  - Data logging presents a serialization point
    - In the original design all O(30) Level 3 Trigger farm nodes were connected via TCP/IP to a single server that serialized the events into a single raw data file
    - Occasionally we observed this to relate to event-level deadtime (“backpressure”) under bad background conditions -> bottleneck
- **The Solution**
  - Take serialization out of the deadtime path
    - New design implements single node logging to local disk, plus one or more asynchronous servers that perform merging of individual contributions

Logging Manager (cont)

- **Performance**
  - This was ready to be deployed at the time of the last Program Review
  - Since, we have been able to put this to the test in IR2
  - With only ~10 min added latency between the last event taken and the availability of the merged raw data file, the new design by far exceeds its performance goal
  - We can easily log 5 kHz of 50 kB events sustained, and could do >10 times that rate at peak!

Objectivity Phase-out

- **CM2 Event Store**
  - With the realization of CM2, the BaBar (Kanga) event store is implemented in ROOT I/O
  - Still, we cannot do much without *Objectivity*; every production farm, every analysis site, every laptop needs it. Why?
- **Non-Event Store Data**
  - Along with the main event store, *Objectivity* was also the choice for most non-event database needs in BaBar, *i.e.*, configuration, conditions data that are needed for DAQ calibration, reconstruction
  - External Computing Review recommended that BaBar consider the eventual complete phase-out of *Objectivity*

Non-Event Store Databases

- **Ambient**
  - Internal to **IR2**
- **Configuration**
  - Written in **IR2**, read everywhere
- **Conditions**
  - Written in **IR2 + PC**, read everywhere
- **Spatial+Temporal**
  - Internal to **PC**, intermediate

- Stores slow control data (EPICS), detector voltages, temperatures etc.
- Stores Front-end and Trigger configurations, detector geometries
- Stores machine/detector conditions and calibration constants
- Accumulates data across farm nodes and “rolls” calibrations in time

Database Re-Implementation

- **Non-Event Store Databases**
  - Ambient, Configuration, Conditions, Spatial+Temporal DBs are left in *Objectivity (TM)* based implementations
- **Motivation**
  - While none of these implementations is fundamentally broken, they do impose constraints we have since eliminated from the Event Store; Replacing them would:
    - Free all read-only access in ER, SP, Skims + Analysis from having to install Objy servers at Tier-As, SP sites, laptops
    - Free us in our choice and schedule of compilers and operating systems (platforms)
    - Save BaBar licensing costs of O(100 k$)

Migration Strategy

- **Planning**
  - A study of possible migration options was done last year => BaBar Note 576
  - Key decision was to go for ROOT/CINT as replacement of *Objectivity*/DDL for user defined schema (“payload”)
  - For storage of payload objects and related metadata (CDB “intervals”, “revisions”, “views”) provide two solutions:
    - Read-only version implemented using ROOT I/O files
    - Read/write version implemented using a relational database (MySQL), which provides transactions/recovery. Payload objects are serialized, compressed and stored as SQL BLOBs

Migration Plan

- **Priorities**
  - **Phase I:** Implement ROOT I/O version first, deploy in Skims, Analysis, and eventually in SP, while keeping *Objectivity* for write-access in IR2, PR(PC)
    - At this point Objectivity will be confined to production environments and not be needed outside SLAC
  - **Phase II:** Implement RDBMS version to completely replace *Objectivity* in BaBar
  - This is a big job
  - After completing the design phase and prototyping last year, the implementation phase is now well underway
    - Parts of this are already complete to be deployed

Migration Status

- **Configuration Database** **Andy Salnikov et al**
  - All interfaces (APIs) restructured to become technology-neutral (18-series releases), already deployed still with Objy
  - ROOT/MySQL implementation practically done!
  - Now working out data distribution mechanism (using xrootd)
- **Conditions Database** **Igor Gaponenko (LBNL), Jane Tinslay et al**
  - Two parts: “Core” functionality (framework) exist in first implementation, expect revised version in a few weeks, “payload” migration (bulk of user-defined classes) being underway, including calibrations, with help from developers
- **Other (Ambient, PR Databases)**
  - APIs already technology neutral, implementation work to be done, but note that these are not used outside SLAC

Andy Hanushevsky et al

xrootd

- **Motivation**
  - Development stimulated by the huge amount of BaBar data being accessed by a high number of analysis users
  - Thought of as an extension of the *rootd*, after BaBar converted their event store from *Objectivy*/DB to ROOT I/O
    - Takes over the role of the *Objectivity* AMS (“Advanced Multithreaded Server”) (the AMS that wasn't...)
      - The same SLAC developer who was granted a source license with *Objectivity* to provide us with a (barely) workable AMS put his expertise (and that of others) to a real solution
  - Goals are

xrootd: Data Delivery Goals

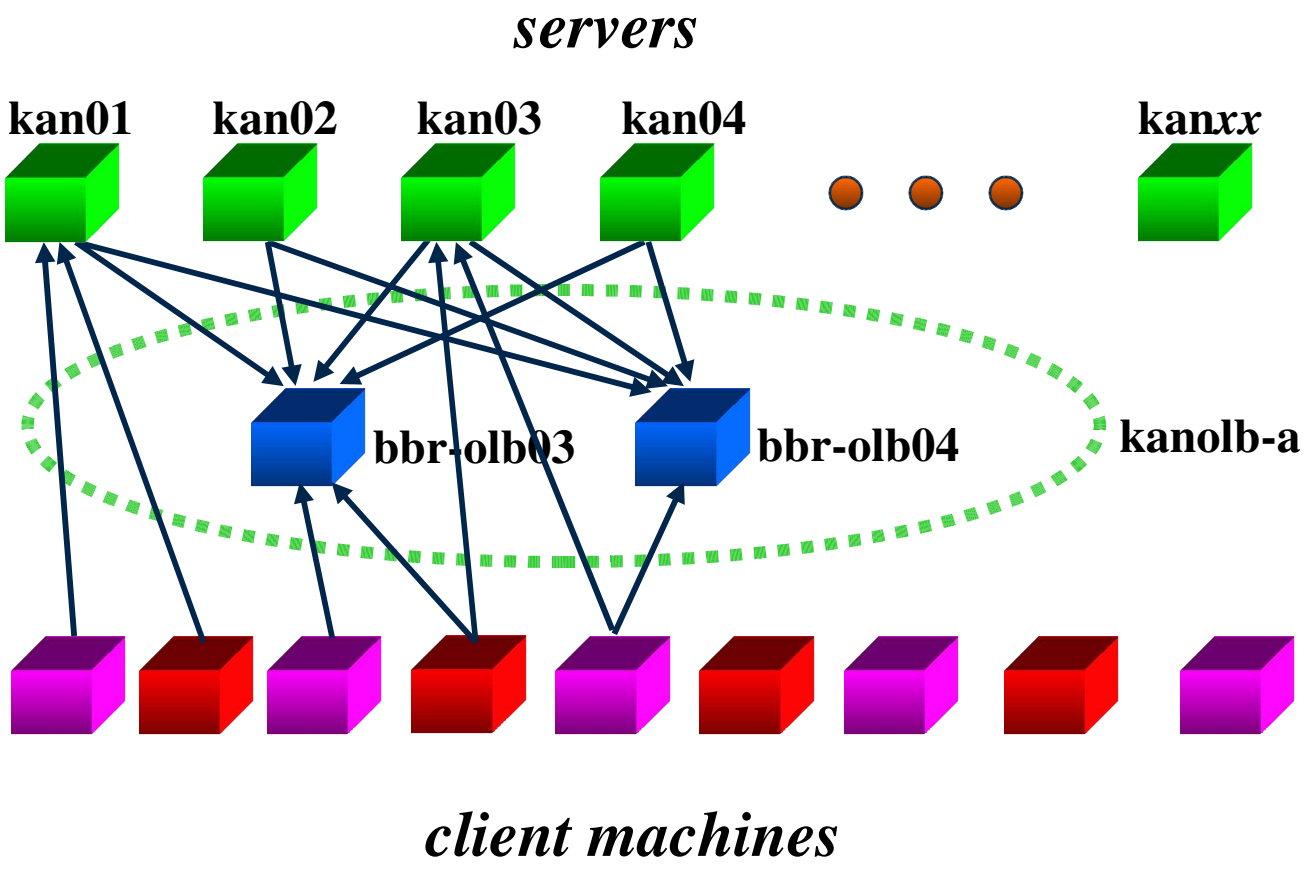
- High Performance File-Based Access
  - Scalable, extensible, usable
- Fault Tolerance
  - Servers may be dynamically added and removed
- Flexible Security
  - Allowing use of almost any protocol
- Simplicity
  - Can run out of the box (no config for small installations)
- Generality
  - Can be configured for ultimate performance (intermediate to large sites)
- Rootd Compatibility

xrootd: Features

- Rich and efficient protocol combines file serving with peer-to-peer elements (file sharing)
- Heavily multi-threaded
- Can deliver data at disk speed (streaming mode)
- Low CPU overhead
  - 75 % less CPU than NFS for same data load
- Supports massive clusters
  - 280 nodes self-cluster in about 7 s
- Monitoring: Provides selectable event traces
  - Ganglia Monitoring implemented for all BaBar Tier As
  - http://www-gmon.slac.stanford.edu:8080/ganglia/babar

Jacek Becla, Yemi Adesanya, Tofigh Azemoon et al

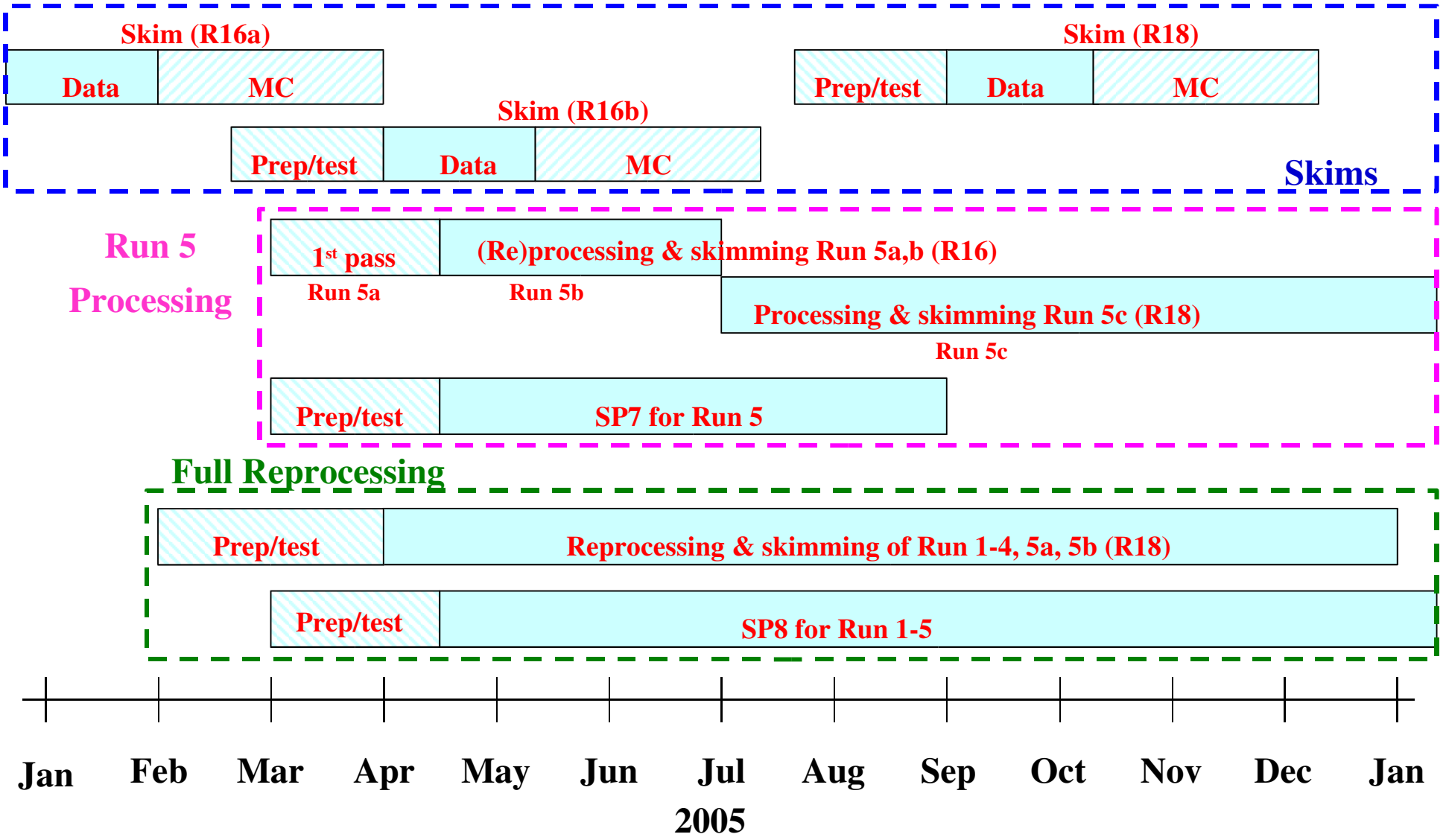
Example: SLAC Configuration



xrootd: Prospects

- One server can easily support 600 clients
  - A challenge to maintain scaling while interfacing with external systems
- Has been installed at all BaBar Tier As
  - Can not only serve event data but also Conditions etc.
    - Remember, these will soon be in ROOT, too
- Can provide data access at the LHC scale
- Ships with CERN's recent releases of ROOT
- People outside BaBar have started looking into this
- More than a spin-off: In my opinion a great example of how the lab can help the community share solutions to common challenges http://xrootd.slac.stanford.edu

2005 Computing Schedule



Conclusions

- By a year ago, BaBar had reinvented the way it does Computing, CM2 led us to an outpour of new results presented at ICHEP'04; Now CM2 has turned into routine and we are looking for ways to make it even better
- BaBar Computing has grown into a truly distributed endeavor, SLAC + the 4 Tier A Centers in Europe are turning around data in record time
- SLAC staff and resources play a key role in BaBar Computing, both in operating the experiment and in breaking ground for new technologies for the years to come
- We have to keep our technology at the cutting edge if we want to master a doubling of the data set twice in the next 3 years