BABAR Computing

DOE Program Review

SLAC

Breakout Session

3 June 2004

Rainer Bartoldus

BaBar Deputy Computing Coordinator

STANFORD LINEAR ACCELERATOR CENTER

Outline

- **The New Computing Model (CM2)**
  - New Kanga/ROOT event store, new Analysis Model, new Micro/Mini, new Bookkeeping
- **Overview of Production Activities**
  - Online, Calibration, Reconstruction, Skimming
- **Distributed Computing at Tier A Sites**
  - IN2P3, RAL, INFN, GridKa, SLAC
- **Production Summary**

Computing Model 2 (CM2)

- **Builds on the experience from the original Computing Model**
  - Have learned what works, what doesn't
  - Addressing issues that are becoming more important with increasing data volume
- **Result of a thorough 6 months study by CM Working Group 2**
  - Decision to implement in December 2002
  - Implementation launched (and completed!) in 2003

New Event Store

- **CM2 Kanga/ROOT**
  - Simple, ROOT-based file format
  - Minimal size overhead to keep disk space costs low
  - Used everywhere from Tier-A/C to workstation/laptop
  - Written directly from Production (ER and SP)
  - Support for multiple data components (tag, aod, esd, tru, cnd)
  - Supports requirements of new analysis model

New Analysis Model

- **Storage of composite candidate lists**
  - User-defined combinatorics
  - Any user-calculated quantities associated with event or candidate
- **Centralized Skim Production**
  - Opportunity for new ideas 4 times a year
- **Provision for easy/fast access to event store**
  - Option of interactive access from ROOT prompt

Addresses the main reasons for AWG "ntuple production"

Mini and Micro

- **Introduced Mini (DST) format**
  - Highly compact output of event reconstruction
  - Contains high-level objects as well as hit information
    - Allows to redo track fitting, vertexing, etc.
    - Very powerful source for analysis+diagnostics
- **Changed to New Micro (DST) format**
  - Subset of the Mini (“reduced” Mini)
  - Restricted to analysis-level objects, also composites
  - Behaves like Mini in cache mode

Overview of BaBar Production

- **Online**
- **Prompt Calibration**
- **Event Reconstruction**
- **Simulation Production**
- **Skimming**

Online Computing

- **Online Event Processing / Level 3 Farm**
  - Recording data with consistently high efficiency and minimal deadtime (1-2 % at peak luminosity)
  - Reached 300 Hz logging rates out of Level 3 Trigger (3 times design)
- **Logging Manager (LM)**
  - Single server connected to ~30 Level 3 farm nodes (serializing the data streams)
  - Capability of 500 Hz at 30 KB per event is the current limit (seen in L1 “passthrough” running)

Online Computing (cont.)

- **Logging Manager Upgrade**
  - Basic Idea
    - Log data in parallel streams to local farm node disks
    - Harvest data asynchronously through one or more servers, merging into single event stream
  - Goals
    - Eliminate dead time due to LM
    - Accommodate logging rates of 50 kB x 500 Hz sustained, and 50 kB x 5 kHz peak (scalable)
  - Ready to be deployed this month!

Prompt Calibration (PC)

- **Tracks changes of detector conditions in time**
  - Runs off 5 Hz of calibration events selected by Level 3
    - Bhabha, radiative Bhabha, mu pair, hadronic events
  - Extracts calibration constants for each run
    - *e.g.*, beam spot position, SVT timing, alignment, etc.
    - Writes to Conditions Database
  - "Rolling" calibrations:
    - Slowly varying cnds propagated from one run to next
    - Only stage that needs to process runs sequentially
    - Has to keep up with IR2 logging rate

PC Performance

- **PC Farms keeping up with data taking**
  - Current farms consist of 16 x dual 1.4 GHz PIII CPUs
  - Typically 1 farm for new data, 1-3 for reprocessing
  - Events read from xtc file, fanned out to nodes
    - Used to be capable of 600 pb^{-1} per day, recently added new server to achieve up to **1000 pb^{-1} per day**
  - Safe again, but still do not want PC farm to have to scale with luminosity
    - Developed fast Filter to generate small calib. xtc file
    - Will keep PC farms ahead of high luminosities

Event Reconstruction (ER)

- **Performs full reconstruction of raw event data**
  - Applies conditions data from PC pass
  - Only needs conditions from current run
    - Completely parallelizable at run granularity
    - Allows to scale ER farms to higher luminosities
  - Writes into (new) Event Store

ER Performance

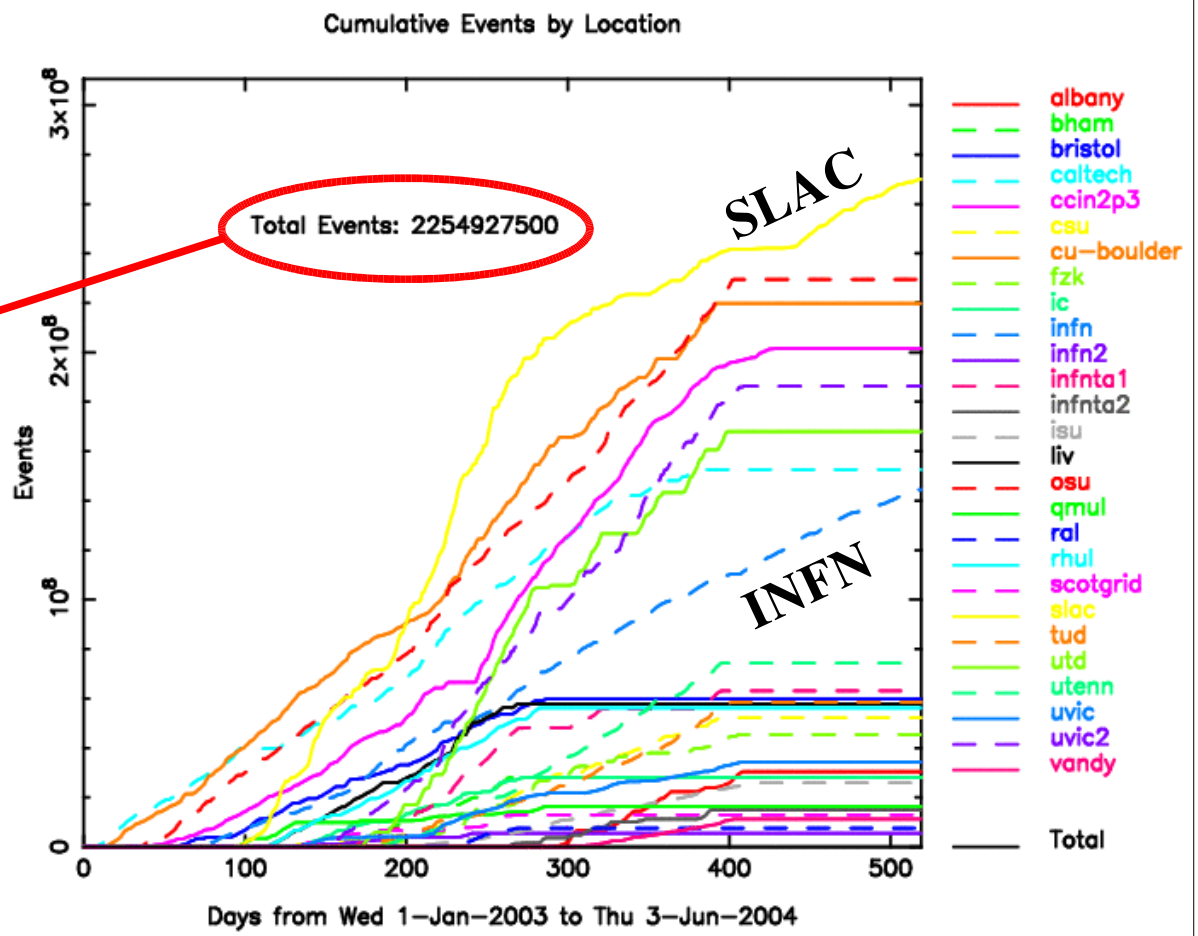
- **ER Farms doing processing and reprocessing**
  - Currently 4 farms of ~32 x dual 1.4 GHz PIII CPUs plus 2 farms of dual 2.7 GHz PIV
    - Each farm capable of **150 pb^{-1} (220 pb^{-1}) per day**
    - Currently 3 farms for new data, 2 to be used for skimming, 1 for reprocessing
  - Can keep up with new data
    - At the same time allows to reprocess early fraction of the Run with optimized calibrations

Simulation Production (SP)

- **Distributed over 27 production sites**
  - Mostly Tier-C (universities)
  - Little dependence on a single site
    - stable production rates
  - Between 10 and 600 CPUs per site
  - Total of ~2000 CPUs available for SP
- **Resources to produce 60 M events/week**
  - **SP5 for Run 1-3** data (12-series) ramping down
  - **SP6 for Run 4** (14-series) ramped up

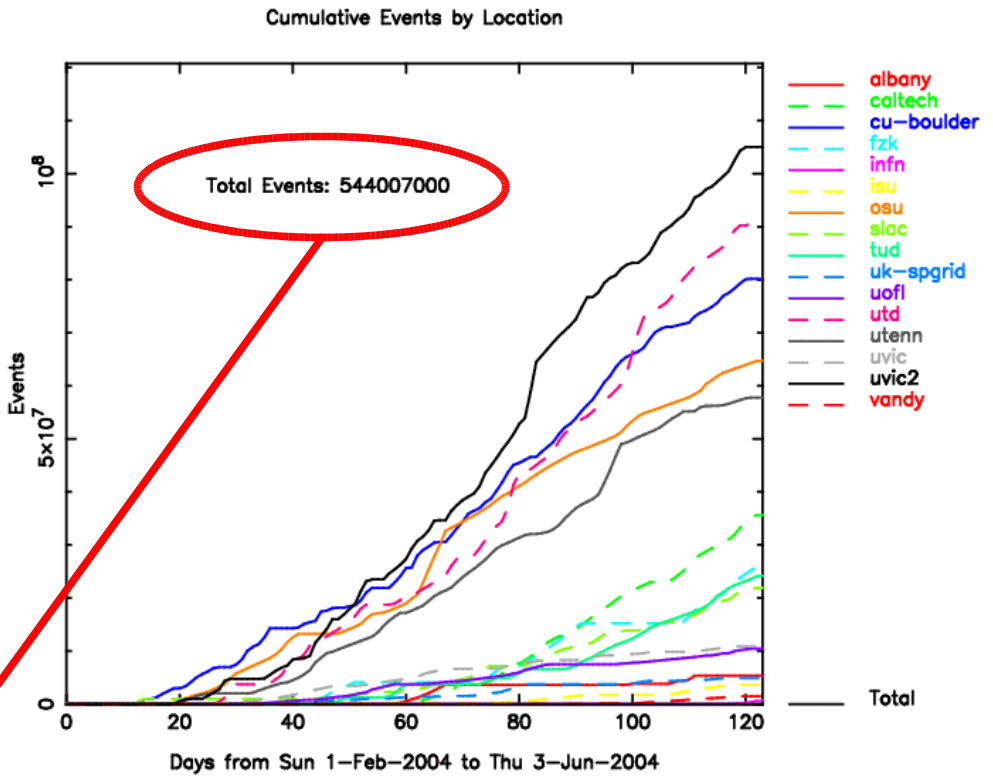
SP5

- **Generic BB production complete**
  - **2.2 billion events** available
  - SLAC and INFN continue to produce signal MC
    - 3M events/wk



SP6

- **Target and Status**
  - Generic BB: 3x lumi
    - 3.2 M / fb^{-1}
  - udsc + tau: 1x lumi
    - 4.3 M / fb^{-1}
  - Signal (replicated)
    - 3.3 M / fb^{-1}
  - 1 B events for Run 4
  - **544 M events done**
    - All signal + generics up to May 1 done by June 30



BaBar Tier A Sites

- **The "pillars" of the distributed model**


– Originally, **IN2P3** in Lyon, France as a replica of SLAC, taking a considerable share of analysis users


– Then **RAL**, UK entered as an alternative analysis site for the (initial) Kanga data


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


– Latest addition is **GridKa**, Karlsruhe, Germany to participate in skimming and soon also analysis

IN2P3

- **Analysis**
  - Seeing continuously high load of BaBar analysis jobs
    - About 200 jobs running in parallel on average during last month
    - The maximum is > 400 jobs
- **Simulation Production**
  - MC production has now switched from SP5 (Run 1-3) to SP6 (Run 4)
  - Objectivity Mini + Truth were imported for SP5 conversion into CM2 format

IN2P3 (cont.)

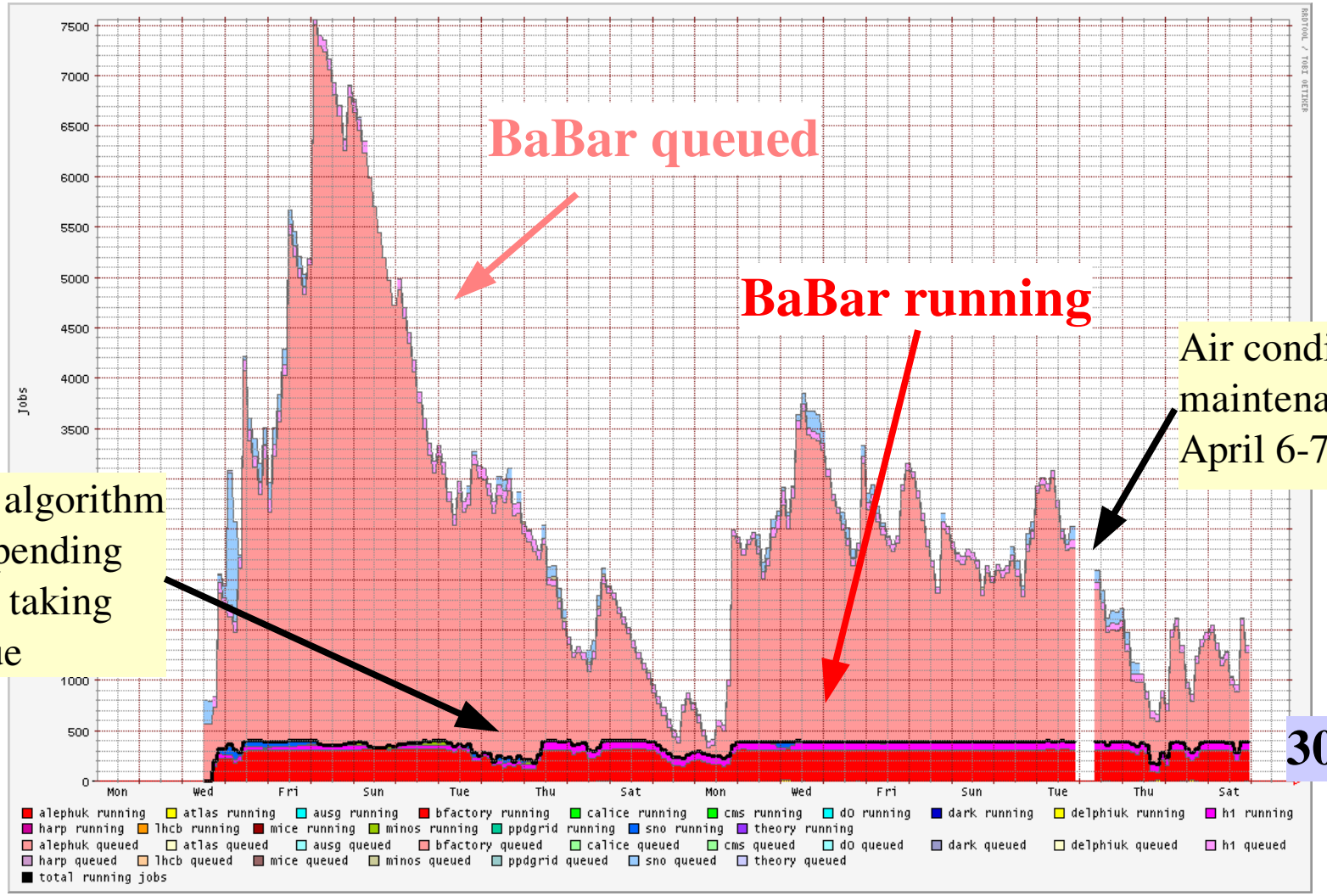
- **CM2 Data**
  - Import completed through SRB (= Storage Request Broker, a GRID tool!)
  - ***AllEvents*** plus all 41 allocated skims are now available at CCIN2P3 for Run 1, 2, 3
- **Future**
  - More disks expected
  - Preparing to switch from RH 7.2 to RHEL 3 (but BaBar is not alone at CCIN2P3)

RAL

- **Analysis Resources**
  - 304 x dual 1.4 GHz PIII CPUs
    - Queues full + 1000's of jobs waiting most of the time
    - 60-70 new dual 2.8 GHz PIVs to come in June/July
  - 49 TB disk space available
    - All 10- and 12-Series “classic” Kanga data
    - In process of replacing *Streams* with *Pointer Skims*
      - This is freeing up considerable space for CM2 data
    - 30-40 TB new disk coming online in June/July

RAL: Batch Usage

- One Month -

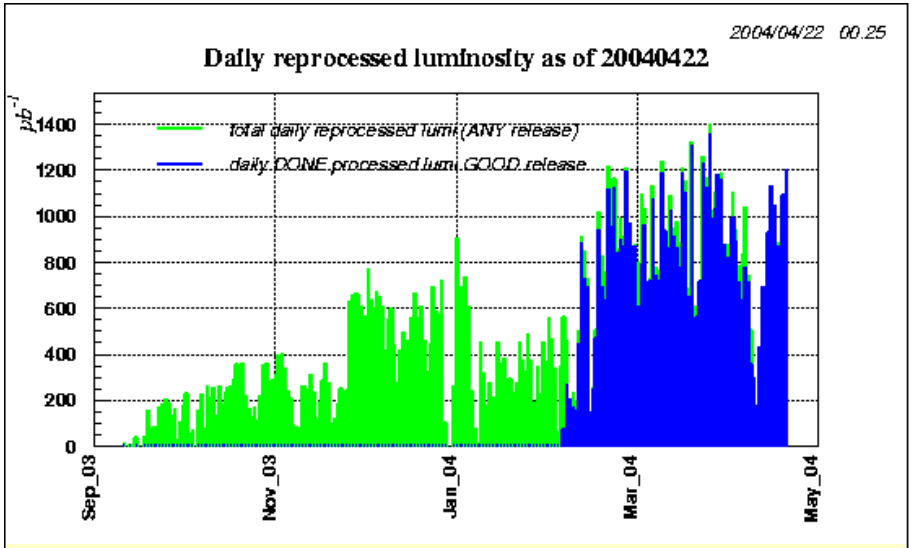


RAL: CM2 Data

- **CM2 Data Availability**
  - ARTF (Analysis Resources Task Force) specified:
    - ***Tau11***, ***Tau1N*** and ***Tau33***
  - Quick unofficial survey of users added:
    - ***BCCC03a3body***, ***BCCC3body***, ***BCCKs3body***, ***BCPi0Ks3body***, and ***BFourBody***
    - Will feed back to ARTF for final decision
  - First CM2 data made available last month
    - More to be added as space is being freed up
  - **xrootd** service for CM2 access starting soon

INFN

- **Event Reconstruction**
  - Imported **1.4 TB on best day**
  - Often **1.2 TB/day**
  - 60 fb^{-1} done in ER (same as IR2) but "110 fb^{-1}" processed
  - Always need to redo early fraction of a new Run



- **Run4 (re-)processing**
  - Needed to converge on reconstruction release, conditions (*e.g.* SVT local alignment)
  - Officially started on Feb 9th

INFN: Reconstruction

- **Control-System Development**
  - Developed while in production (now very stable and feature-rich)
    - New CM2 tools for merging, checking collections
    - Lots of new states/checks, and new post processing state machine
- **PR (= PC+ER) Managers**
  - System where managers look after farms on both sides of the Atlantic is working well for BaBar
    - Proved very helpful in raising efficiency

INFN: Analysis

- **CNAF (Bologna)**
  - Former Tier-B of Roma moved to CNAF end of 2003
    - CNAF also Tier-1 for LHC experiments
- **Hardware Resources**
    - 2 front-end servers
    - 30 dual PIII clients and 21 dual PIV clients
    - 6 TB of disk for scratch, AWG, conditions
    - 29 TB for event store
  - Major upgrade this summer
    - 88 new clients plus ~70 TB disk

GridKa

- **Activities**
  - Two main tasks in the past:
    - Skimming
    - Simulation Production (SP5)
  - In the future, in addition:
    - Analysis
  - Ramped down SP during skimming and ramped it up again now

GridKa: CM2 Skimming

- **Production Output**
  - January to early April
    - GridKa skimmed 1439 collections of Run3
    - Produced 9 merges of 110 collections each
    - Skimming/merging used 150 CPUs (sometimes 200 CPUs) corresponding to BaBar share at GridKa
    - Used > 5 TB of disk space for skimming/merging
  - At present
    - Skimming Run 4 data until Padova is ramped up

SLAC: CM2 Production

- **Run 1-3 Data:**
  - 1.731 x 10^9 events converted to CM2, skimmed, merged and put in bookkeeping
    - "This is all the available data."
- **SP5 Monte Carlo:**
  - 1.38 x 10^9 events converted and merged (target)
  - 147 x 10^6 BBbar events skimmed
    - Skimming is ramping up now that SP5 conversion is winding down (can do 25-30 M a day, signal 5x faster)
  - 94.1 x 10^6 events merged

CM2 Production (cont.)

- **Run 4 Data:**
  - ~83 fb^{-1} recorded (as of today)
  - 62.9 fb^{-1} in “Green Circle” dataset (up to May 1) processed, skimmed and merged and put in bookkeeping
    - This completes the Green Circle dataset in CM2!
- **SP6 Monte Carlo:**
  - 544 x 10^6 events generated
  - 131 x 10^6 events skimmed
  - 85 x 10^6 events merged

Summary

- Within a year's time, BaBar has reinvented the way it does Computing
- Computing Model 2 is in place and has successfully finished its first production cycle for ICHEP04
- SLAC + the 4 Tier A Centers in Europe are turning around data in record time
- Over the past year, BaBar has become a more distributed experiment than it ever was before
- We have to keep our resources at the cutting edge if we want to master the hoped-for explosion in data volume that PEP-II promises