BaBar Computing

Gregory Dubois-Felsmann, SLAC
BaBar Computing Coordinator

SLAC DOE High Energy Physics
Program Review

7 June 2006

Overview

- The *BaBar* Computing Task
  - Data Volume
  - The Data Model
  - Distributed Computing Infrastructure
  - SLAC-Provided Resources
- Data Processing Pipeline
  - Online
  - Calibration, Reconstruction
  - Skimming
  - Simulation
  - Data Distribution
  - Data Access
  - Use of Grid Technologies
- Current Activities
- Future Plans

The *BaBar* Computing Task

- What must be accomplished:
  - Accumulate physics data from the detector
    - Acquire it, filter it, and record it while monitoring its quality
    - Calibrate and reconstruct it
  - Generate corresponding simulated data
    - Using the recorded history of the condition of the detector...
    - Generate and reconstruct simulated events (globally distributed)
  - Divide all the data into skims for convenient access
  - Distribute it to many sites
  - Provide an analysis environment
    - A software environment to be run at all sites
    - Substantial physical resources at major sites

Data Volume

- Rates:
  - Output from the detector:
    - Typically ~2000-2500 events/s at present $1*10^{34}$ lumi
    - Recently ~33-40 kB/event (*BaBar* custom binary format)
    - ~250-350 events/s selected by Level 3 trigger (~1/7)
  - Output of reconstruction:
    - ~1/3 of events selected as “AllEvents” data for analysis
    - ~3-3.5 kB “micro” & ~6.5-7.3 kB “mini” (separate files)
    - Bulk size averaged over history of experiment:
      - Micro: 44 GB/fb^{-1}; Mini: 81 GB/fb^{-1}
    - Simulation:
      - Larger: events include truth data; BBbar generated at x3 multiple
      - Micro: 89 GB/fb^{-1}; Mini: 126 GB/fb^{-1}

The Data Model

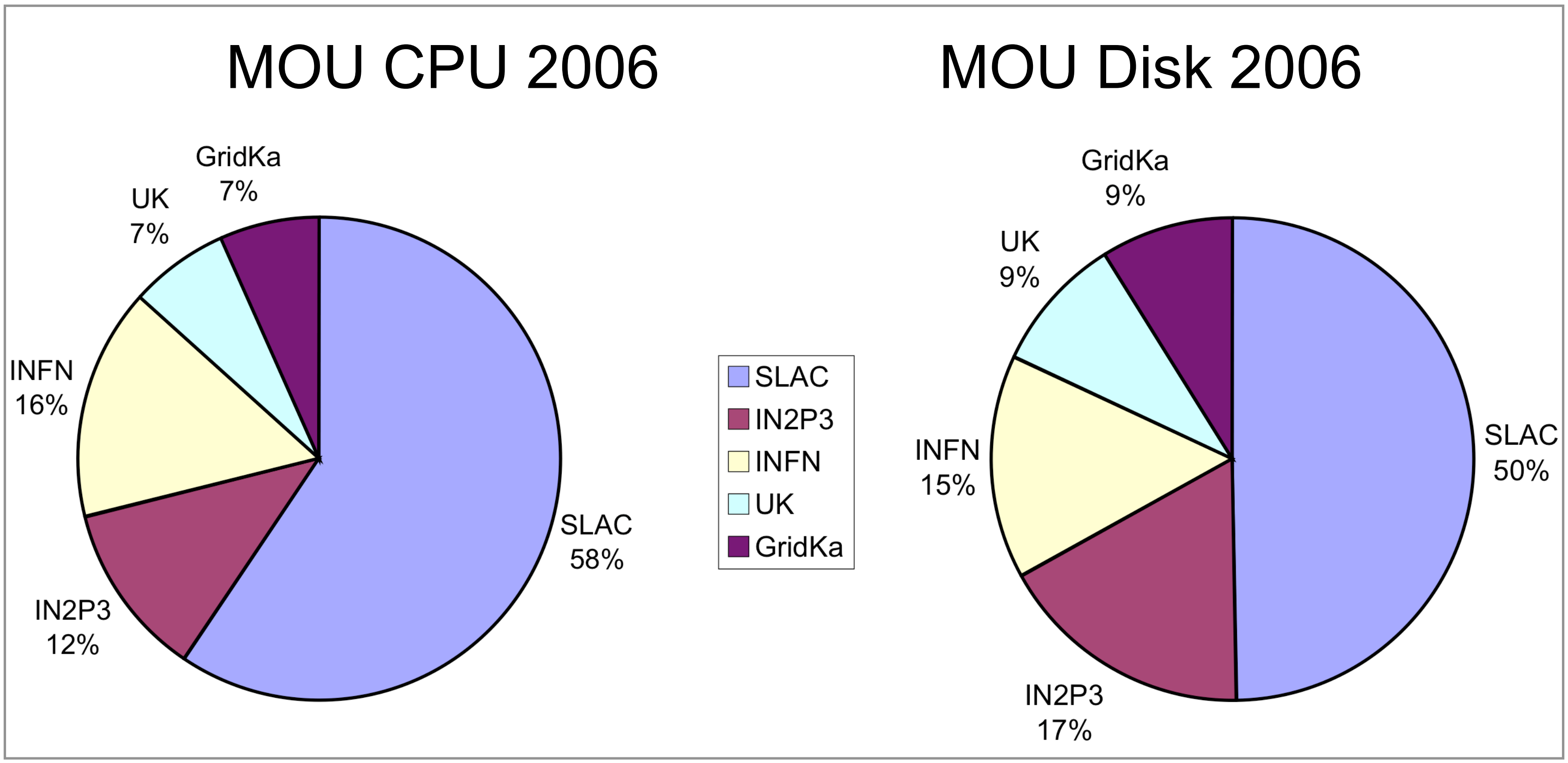
- After Computing Model 2 (CM2) re-engineering:
  - Reconstructed (beam & simulated) data:
    - ROOT trees, composed of *BaBar*-specific data objects
    - Optionally striped across multiple files by *component* (micro, mini, truth, ...)
    - Copies of events can be done by value or by reference (selectable by component)
    - ~140 TB by end of this year
  - Data skimmed for convenience of use & distribution
    - Currently ~215 skims (170 "deep copy", 35 pointer)
      - Pointer skims used for largest selection fractions
      - Deep-copy skims can be exported to small sites
    - Convenience comes at a cost:
      Total size of skims is about 5.1x larger than micro
    - ~300 TB by end of this year

Distributed Computing Infrastructure

- *BaBar* computing resources are distributed
  - SLAC
  - Four “Tier A” centers supporting both central production and user analysis
    - IN2P3 (Lyon, France)
    - RAL (Rutherford Lab, UK)
    - INFN (Padova and CNAF/Bologna, Italy)
    - GridKa (Karlsruhe, Germany)
  - A total of up to ~40 laboratory and university sites running the BaBar simulation
  - “Tier C” sites ranging from large departmental clusters to users’ laptops

Tier-A Computing Resource Commitments

- Proportions of commitments…



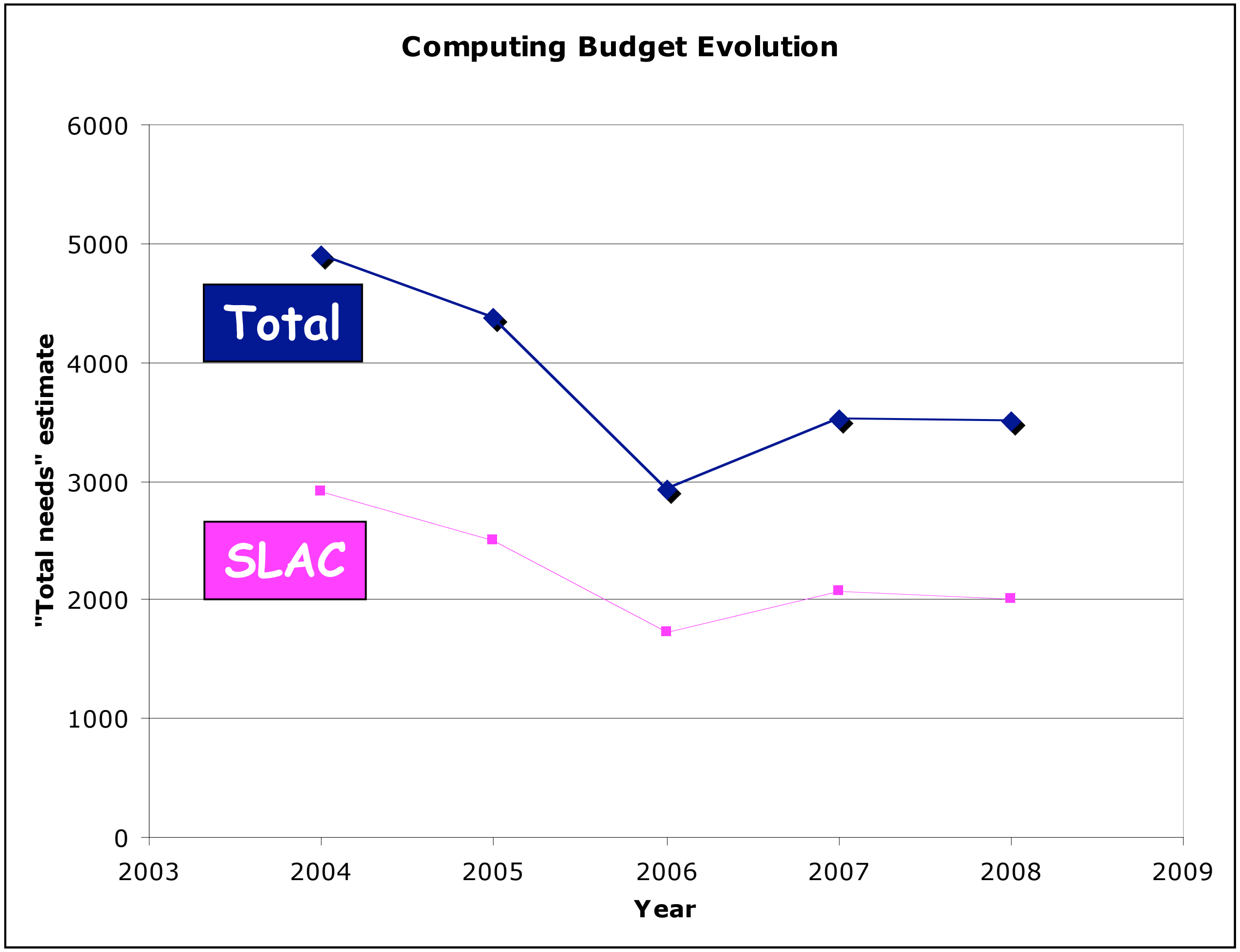
SLAC and Tier A Resources

- Some numbers:

2006 Commitments		
Site	**CPU (SLAC CPU-weeks)**	**Disk (TB = 2^{40} bytes)**
IN2P3	2580	205
INFN	3407	176
UK	1450	110
GridKa	1493	104
SLAC	13029	587
Total	21959	1182

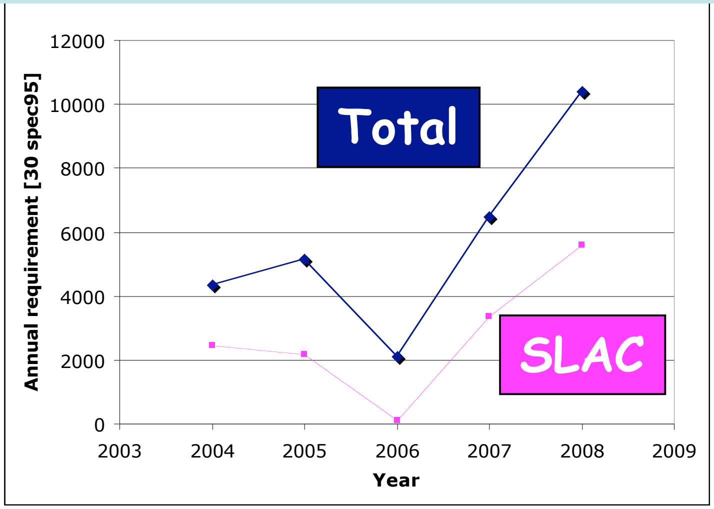
- Funding for SLAC *BaBar* computing hardware is shared between DOE and the collaborating national agencies
  - International Finance Committee mechanism supervises this
  - Offsetting credit is given for national contributions to Tier A centers

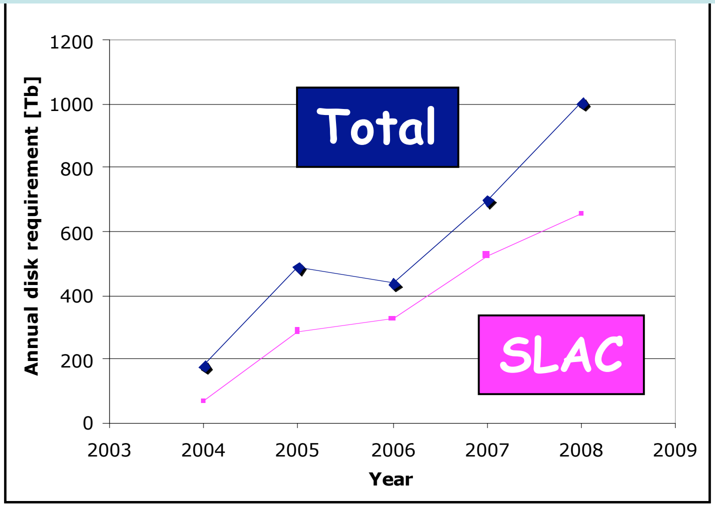
Predicted evolution of computing budget

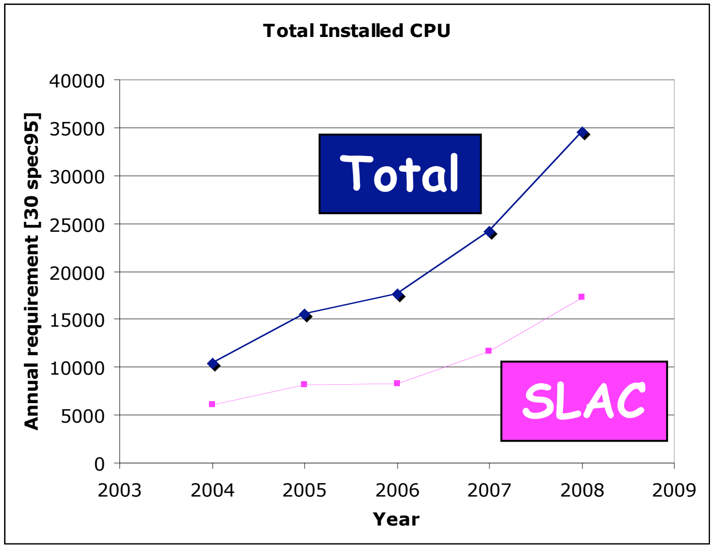


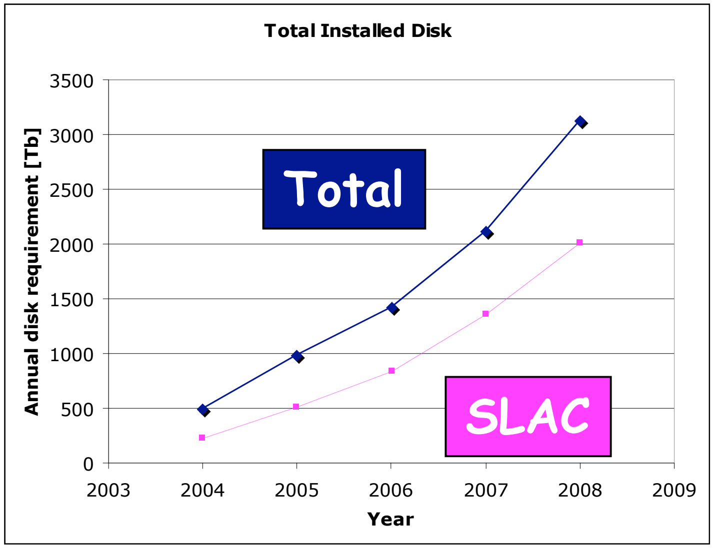
Dip in budget is caused by difference between anticipated and delivered luminosity in 2005

Installed computing capacity









SLAC-Provided Staffing

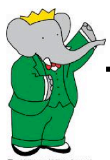
SLAC staffing for B Factory

SLAC B-Factory Staffing Level (FTE's)

		Phys	Comp Prof	Eng	Techs	Admin	Students [1]	Others	TOTAL
FY05	BaBar	39	50	18	22	21	1	14	165
	PEP-II	35	23	118	124	20	3	53	376
	FY05 Total SL	226	137	201	204	151	35	137	1,091
FY06	BaBar	34	48	16	19	18	1	14	150
	PEP-II [2]	36	21	101	127	21	2	49	358
	FY06 Total SL	217	140	177	196	140	41	123	1,035
FY07	BaBar								144
	PEP-II [2]								325
	FY07 Total SLAC								1,020
FY08	BaBar								142
	PEP-II [2]								320
	FY08 Total SLAC								985

[1]Graduate Students are Stanford University Assistants who typically have half-time appointments, i.e. 0.5 FTE each.
[2] FY06 is the first year of a multiple-year transition of linac operations to BES. PEP-II FTE's include those supported from BES linac operations funding.

(From D. MacFarlane's presentation at 2006 Operations Review)



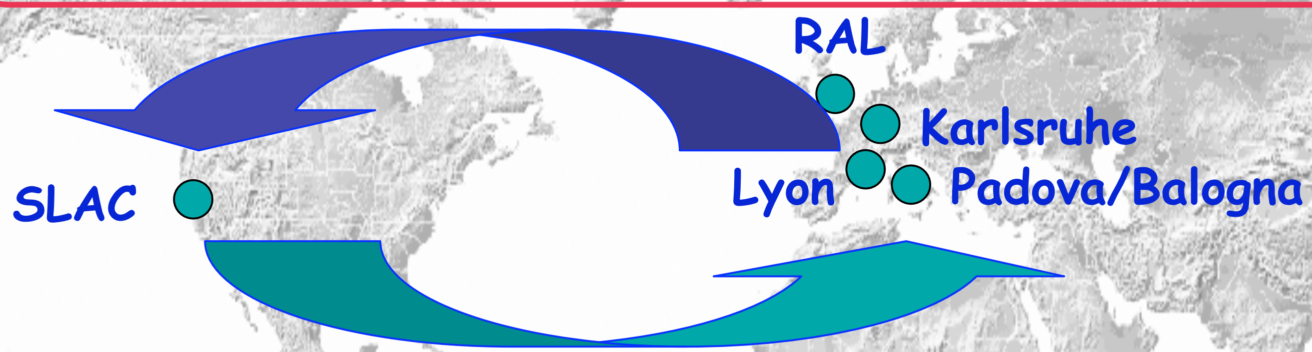
Data Processing Pipeline

- Online
- Calibration, Reconstruction
- Skimming
- Simulation
- Data Distribution

Online

- Online computing group now mostly from SLAC
  - Management as well as most development and operational staffing
- Online system provides
  - Data acquisition: front-end readout, feature extraction, event building, software triggering (Level 3), data logging
    - Rates and volumes as mentioned above
  - Data quality monitoring
  - Detector operation (slow control and run control)
  - Configuration and operational status and conditions databases
- Infrastructure
  - O(100) computers (compute, file, database, console servers, workstations...)
  - Gigabit Ethernet networks for event building, file service backbone, link to SCCS; switched 100 Mb for rest of systems
  - Release management, user environment for development
- Data logged finally transferred to HPSS at SCCS

Tier A centers in calibration and reconstruction



- Fast calibration pass at SLAC
- Data sent via network to Padova for reconstruction
- Data reconstructed and returned via network to SLAC

Calibration, Reconstruction

- Calibration
  - Calibrations for a run needed before full reconstruction
    - Some timing and beam position parameter change rates require this
  - Based on a small subset of simple, clean events
    - Bhabhas, dimuons, ...
    - Selected at an approximately constant luminosity-independent rate (~7Hz) to provide sufficient events (currently 2-3% of L3)
    - Selected in the online system and written to a separate file
  - Processed by the full reconstruction executable
    - Single-threaded as constants are fed forward (“rolling”)
  - Possible to execute on a subset of the SLAC CPU farm
    - Typically 10-15 nodes
      - Additional subfarms maintained for reprocessing and testing
    - Normally completed within 2.5-3 hours after completion of DAQ

Calibration, Reconstruction

- Reconstruction
  - Raw data are exported continuously to Padova site
    - Permanently archived there as a redundant copy
  - Constants resulting from calibration pass are exported in batches
  - Multiple runs can be processed in parallel
    - No feed-forward dependencies - full event independence
    - Allows use of multiple, smaller farms - alleviates scaling problems
  - Plenty of capacity for reprocessing
    - Current maximum is ~2 fb^{-1}/day (c.f. 0.737 fb^{-1} 24-hour record to date)
    - Overcapacity planned to be shifted to skimming shortly
  - Full reconstruction executable applies filtering
    - Reduces processing power required as well as output sample size
    - Retains ~1/3 of input events from Level 3
    - Output in ROOT files in CM2 data format
  - Typically completed in ~1.5 days, but very tolerant of outages
  - Outputs exported continuously back to SLAC

Skimming

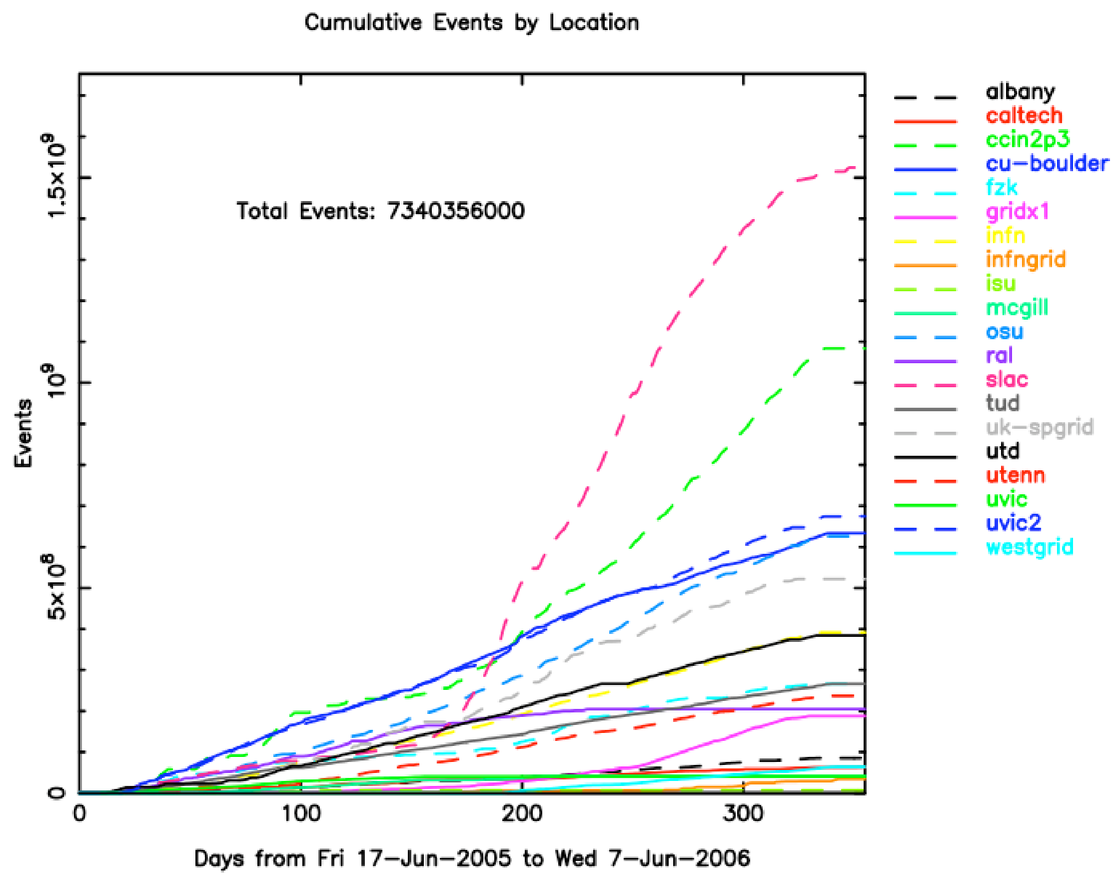
- Data skimmed for convenience of use & distribution
  - Currently ~215 skims (170 “deep copy”, 35 pointer)
    - Pointer skims used for largest selection fractions
    - Deep-copy skims can be exported to small sites
  - Convenience comes at a cost:
    Total size of skims is about 5.1x larger than micro
  - ~300 TB by end of this year
- Skim cycles 3-4 times/year
  - Want to keep latency low to enable new analyses to start quickly

Simulation

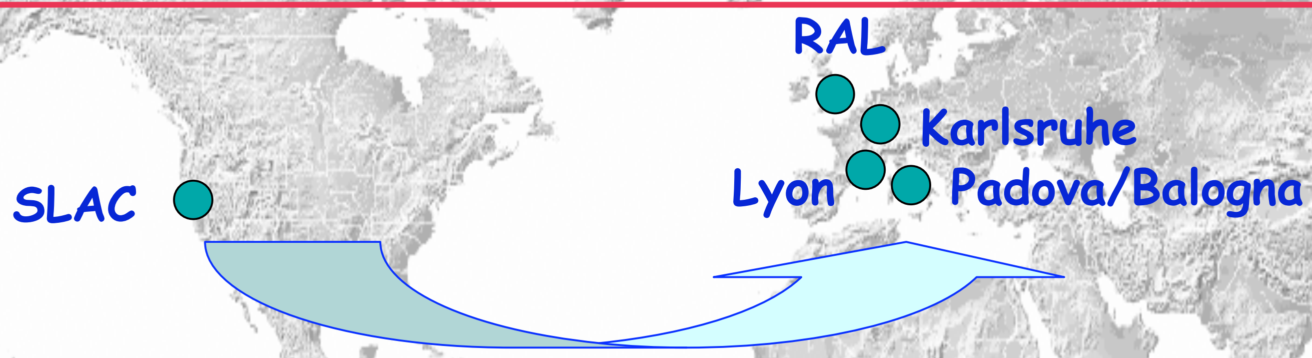
- Full resimulation pass under way
  - GEANT 4 v6-based core

- Distributed over network of 20-40 universities and labs

- All data returned to SLAC for skimming
  - GridKa for *uds* continuum
- Distributed to sites by AWG assignment



Data distribution to Tier A computing centers



Maximum transfer rates:

IN2P3	**3 Tb/day**
RAL	**2 Tb/day**
GRIDKA	**1 Tb/day**
CNAF	**1 Tb/day**

- **Data skims are uniquely assigned to Tier A Centers**
- **Disk space for corresponding Analysis Working Group located at same Tier A**

Tier A assignments

Each AWG is hosted by a Tier-A:

Tier-A	AWG
SLAC	AllEventsSkim, Breco, LeptBC, PID, SemiLep, Tracking, TauQED
Bologna	AllEventsSkim, Charm
IN2P3	AllEventsSkim, Charmonium, PartSpec, RadPenguin, ChlsTwoBody
GridKa	TDBC
RAL	ChlsQ2Body, ChlsThreeBody

Use of Grid Technologies

- Grid technologies beginning to be used to support BaBar
- Established: simulation
  - ~25-30 Mevents/week of total ~200 Mevents/week capacity provided by Grid systems in UK, Italy, Canada
  - Planning to increase utilization in future, especially when Objectivity phase-out is complete
- Nearing operational status: skimming
  - Developing capability to do skimming of simulated data on Grid
    - UK “Tier B” resources: Manchester, 2000 nodes; possibly others later
- Not expecting to provide general user analysis support

Current Activities

- Reprocessing
  - In final stages of completing full reprocessing of all data
    - First all-CM2 processing cycle
    - Reprocessing essentially complete
    - Generated corresponding ~7.3 billion simulated events
    - Finishing last ~1% of skimming
- Data-taking
  - In the midst of delivering June 1 data to analysts by next week to meet internal ICHEP deadlines
  - Additional data-taking through August ~21
- Skimming
  - Starting two new skim cycles, one now, one in ~2-3 months with a test cycle starting now

Future Plans

Through end of data

- Online farm upgrade
  - Current online farm (worker nodes and event-build switch):
    - Supports Level 3 trigger and online data quality monitoring
    - Reaching end of its hardware lifetime cycle
    - Likely to reach processing capacity limit around BaBar's highest luminosities
  - IFC (1/2006) approved online farm upgrade ($400K)
    - New high-capacity network switch, all gigabit-Ethernet
    - O(50) current-model AMD Opteron dual-CPU dual-core 1U workers
    - Miscellaneous server and disk capacity improvements
  - Will install during upcoming shutdown
    - Switch acquisition in progress
    - Farm node acquisition linked with next round of CPU acquisitions for SLAC computing center

Future Plans

Through end of data

- Completion of phase-out of Objectivity
  - Non-event-store applications
  - Configuration, Ambient, Calibration (Spatial, Temporal) databases all migrated to ROOT data format and ROOT (and mySQL) database framework
    - Will switch over to the non-Objectivity versions in 2006 shutdown
  - Conditions database partially migrated
    - All conditions needed for data analysis available
      - Preparing to put this into production
    - Remaining conditions still require significant engineering effort
      - Required for simulation, reconstruction, online
      - On tight timetable to still be able to put in production for Run 6
      - Additional engineering effort (physicist or comp. pro.) would be helpful

Future Plans

Through end of data

- Replacement of mass storage system at SLAC
  - Current STK multi-silo system has ~15-year history
    - Regular upgrades to drives, control software
    - Finally reaching end of its support lifetime
  - Replacement needed
    - For *BaBar* needs alone as well as for SLAC's further projects
  - Planning to move to STK 10000-series system
    - Acquisition of first silo late this year or early next
    - Still thinking about whether to recopy all data to latest high-capacity tapes
      - Would probably take O(1 year) to interleave with other demands on the mass storage system

Future Plans

After end of data

- After end of data
  - Expect ~two years of full-scale analysis effort with extensive central computing activities
    - Planning to maintain 3-4 skim cycles/year to keep latency low for new analyses
    - Need to accommodate well-motivated extensions of filter to look at classes of events not originally envisioned
    - Simulation capability must be maintained
    - Planning for possibility of one post-shutdown full reprocessing cycle if sufficiently compelling improvements in reconstruction are developed
      - Full reprocessing would require full resimulation pass
  - Should all be doable with infrastructure in place at end of data
    - Assumes retention of resources at Tier-A sites, network of simulation hosts
    - Continued funding will be needed to
      - Purchase tapes to store output of skimming and user-generated datasets
      - Replace hardware on its ordinary ~4-year lifetime cycle
  - Expect to support a substantially lesser effort for several more years
    - Will need to think about long-term archiving of datasets