Computing Status and Improvement Plans

Charles Young, SLAC

BaBar Technical Review
October 11-12, 2000
Outline

- Computing status.
  - Working at design luminosity.
  - Physics results reported at Osaka Conference.
  - Challenges along the way.

- Improvement plan.
  - Improvements at design luminosity.
  - Additional challenges from luminosity upgrade.
Flow of Data

- Raw data to tape in XTC format.
- Prompt Reconstruction (PR).
  - Reads XTC tape.
  - Filters out background events.
  - Reconstructs “isPhysics” events.
  - Performs “rolling calibration” to be used by subsequent run.
  - Writes output to an event store.
- Centralized skim production to select sub-samples.
- Data distributed to analysis sites for analyses.
PR Performance at Design Luminosity

- PEP-II performance numbers.
  - Best day ~150 pb\(^{-1}\) compared with “design day” of 135 pb\(^{-1}\).
  - Best week ~900 pb\(^{-1}\).
  - Best month ~3.3 fb\(^{-1}\) ~ “design month”.

- Kept up with best month so far.
  - Continual performance improvements from ongoing development.
  - Heroic effort by PR operations group.
  - But no headroom.
Physics Contributions to Osaka Conference

- After one year of operations.
- Plenary talk by David Hitlin.
- 14 contributions in all.
- Many different topics.
  - $\sin(2\beta)$.
  - $B^0$ branching fractions.
  - Measurement of $B_d$ oscillation.
  - Lifetime measurements.
  - Search for $B \to K l^+l^-$. 

*Bottom line -- it worked.*
Challenges in Last Year

- Problems in summer 1999.
- Prompt Reconstruction capability.
  - Unable to keep up with data taking.
  - Scaling limit.
  - Speed.
- Analysis access difficulties.
  - Contention among jobs.
  - Aggregate throughput decreases for more jobs.

- Problems overcome by early 2000.
Prompt Reconstruction in Summer 1999

- PR not able to keep up with data taking.
  - Trigger rate ~100 Hz.
  - PR limited to ~8 Hz.

- Problem with Objectivity DB event store.
  - Throughput initially scales with # nodes.
  - Saturation at ~40 nodes -- scaling issue.

- Plan called for 100 Hz with 190 nodes.
  - No background filter.
  - Program speed.
PR Bottlenecks in Summer 1999

![Graph showing events/sec vs. # nodes with labels: No Obey, Asymptotic limit, Baseline, Production.]

BaBar Database Group
6 internal and 3 external reviewers.

Recommendations:
- Understand and fix PR access problems.
- Continue Objectivity development for regional centers.
- Development test stand.

[/afs/slac.stanford.edu/g/babar/Computing/te
Objectivity Test Program

- Made measurements and tuned using dedicated test bed.
  - Real and *unchanging* data.
  - *Fixed* real-life PR code.
  - Replica of all hardware.
  - Minimal extrapolation from test to real life.

- Significant improvements in ~2 months.
  - Little or no saturation effect.
  - ~5x improvement in maximum rate from Objectivity tuning.

- Importance of dedicated test bed.
  - Sensitive to small changes.
  - Not influenced by beam condition, luminosity, code changes, etc.
Improvements from Objectivity Tuning

![Graph showing improvements from Objectivity tuning. The graph illustrates the performance of various configurations over the number of nodes, with an asymptotic limit indicated.]

Asymptotic limit

8/99

10/99
Additional Prompt Reconstruction Improvements

- Background filter to reduce load.
  - Full reconstruction on 10-nb equivalent according to Technical Design Report.
  - No filter in 8/99.
  - Now accepting ~15 nb.
- Improved program speed.
- Hardware upgrades.
- Additional Objectivity tuning.
- Achieved ~130 Hz peak processing rate in 12/99.
PR Status in 12/99

![Graph showing trigger rate](image)

- **8/99:**
- **10/99:**
- **12/99:**

Asymptotic limit

**Trigger rate**

- 12/99
- 10/99
- 8/99
Peak Rate vs Average Rate

- Job start-up and close-down overheads.
  - Set up, read constants.
  - Write histogram, compute and write constants.
- Initial efficiency 25-50% depending on length of run.
- Many improvements.
  - Objectivity tuning in reading/writing constants.
  - Reconstruction code improvements.
  - Remove excessive histograms.
  - Longer runs.
  - Some operations in parallel with previous and next runs.
- Efficiency typically ~80%.
Inefficient Processing

Event processing

Start up

Close down

outputLMM

Time [min]
More Efficient Running
PR Operational Efficiency

- Careful monitoring of down times.
  - Automated daily and weekly monitoring plots.
  - Log book entries by shifters.

- Operational meeting.
  - Initially daily, now biweekly.
  - Resolve problems.
  - Coordinate changes.
  - Schedule tests.
  - Careful testing before deployment.
56% efficiency in a bad week in 3/00
75% efficiency in a better week in 3/00
Performance of PR

- Latency = time from data taking to successful reconstruction.

- Goal to keep up with data taking.
  - Starting 3/15/00.
  - Latency typically no more than 8 hours.

- Started “factory mode” operation on 3/1/00.
  - Latency ~3 days on 3/1/00.
  - Rapidly brought down to ~12 hours.
  - Keeping up with data taking since about 3/7/00.
  - More than 8-hour latency goal.
  - Acceptable.
PR Latency in March 2000

Prompt Reconstruction

Latency (hr)

Run Number

3/1  3/15  4/1
PR Latency in April 2000
PR Latency in July 2000

Waiting for constants
Prompt Reconstruction Summary

- Continual improvements to match daily luminosity.
  - Factor of two since March.
  - Performance and operational improvements.

- No headroom.
  - Cannot keep up if PEP-II delivered 30 “design days” a month.

- Need additional development.
  - To track PEP-II efficiency.
  - To keep up with luminosity upgrade.
BaBar Daily Luminosity

BaBar Daily Recorded Luminosity

- ~60 pb\(^{-1}\) / day
- >120 pb\(^{-1}\) / day
Data Access for Analysis in Summer 1999

- Contention between jobs.
  - No prioritization or resource allocation scheme.
  - Uncontrolled access to event store.
- Aggregate throughput very poor.
- Inefficient use of CPU resources.
  - ~20% utilization.
Poor Analysis Access in August 1999

- No systematic monitoring
- Very low CPU utilization
Additional recommendations:

- Understand and fix analysis access problems.
- Prioritize resource intensive tasks.
- Non-Objectivity micro-DST level analysis.
  - Now known as Kanga.
- Kanga as near-term solution for smaller sites.

Prioritization Scheme for Analysis Access

- Batch system with hierarchical resource allocation.
  - Experiment-wide allocation: detector studies, physics, etc.
  - Control number of jobs in parallel.
  - Operate near optimal point for aggregate throughput.
- Increase the job limit as problem became understood.
  - Based on Prompt Reconstruction (PR) studies.
  - More chaotic access pattern compared with PR.
    - Many users.
    - Different job characteristics.
  - Harder to tune and control.
- Routinely operate at ~90% CPU utilization.
Improved Analysis Access to Data

Typical CPU utilization ~80%

need continued tuning
Current Batch Queue Definitions

- Skim production queue.
  - Used the same server resources as Objectivity users.

- Queue limits controlled two ways.
  - Queue definition.
  - Allocation of batch workers to queues.

<table>
<thead>
<tr>
<th>Usage</th>
<th>Queue Name</th>
<th>Queue Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skim production</td>
<td>bfprod</td>
<td>150</td>
</tr>
<tr>
<td>Objectivity users</td>
<td>bfobjy</td>
<td>100</td>
</tr>
<tr>
<td>Kanga users</td>
<td>kanga</td>
<td>50</td>
</tr>
</tbody>
</table>
bfobjy Queue in 6/00

Up to ~200 jobs
bfojy CPU Utilization in 6/00

~75% CPU utilization for ~200 jobs
bfobjy Queue in 7/00

Raised skim production queue limit
bfobjy CPU Utilization in 7/00
Objectivity Outage

■ Why?
  ● Transfer data from PR to analysis area.
    • Already transferred file may be updated.
  ● General maintenance such as load balancing.

■ How long?
  ● Was 2 x 8 hours *scheduled* each week, i.e. 10%.
  ● One 8-hour outage a week since August.
  ● Reduced to 2-3 hours a week since September.

■ Impact -- user jobs stalled during outage.

■ Recent improvements.
  ● Database files not extended after transfer.
    • Except for Conditions.
Non-Objectivity Format Data

- Known as Kanga.
  - Built on Objectivity infrastructure.
- Routine use since late 1999.
- Identical content between Objectivity and Kanga.
- Identical user level code.
  - Link in either (or both) persistent-to-transient code.
  - Select at run time if both linked.
Kanga and Objectivity

- Kanga simpler to install and maintain than Objectivity.
- Data distributions simpler for Kanga at this time.

- No direct access to information beyond micro in Kanga.
  - Crucial for more detailed studies.
- Longer term data management issues in Kanga.

- Kanga currently more suitable for small sites.
- Objectivity needed for data centers.
Kanga and Objectivity

- **Objectivity usage.**
  - At SLAC, LBNL and IN2P3 for analysis.
  - At all Monte Carlo production sites.
  - Data distribution labor intensive.

- **Kanga usage.**
  - At all sites for (Micro level) analysis.
Impact of Kanga

- Empowered remote sites.
- Relieved load at SLAC.
- Allowed time for Objectivity improvements.

*Tremendous success!*

- Duplication of resources.
- Much harder to go beyond Micro level.
- Meant to be short to medium term.

*How to proceed?*
Summary of First Year’s Run

- Major challenges along the way.
  - Largely resolved at SLAC.
  - Interim solution for remote sites.

- Physics results presented at Osaka Conference.
- Functional system.

- Improvements needed.
  - Based on success to date.
  - Evolution rather than revolution.
Review Our Strategy

- Luminosity upgrade proposal.
  - Faster than Moore’s Law can keep costs down.
- Computing Model Working Group (CMWG).
  - Requirements.
  - Technical strategy.
- Internal review.
- General consensus.
Improvement Plans

■ Do not disrupt working system.
  ● Evolutionary rather than revolutionary changes.

■ Objectivity as the long-term goal.
  ● Better matched to BaBar’s needs.
  ● Analysis at Micro level insufficient.

■ Continue use of Kanga for the next 1-2 years.
  ● Micro-DST level as before.

■ Details in parallel session.
Beyond Design Luminosity

- System is scalable.
- Parallelism to achieve greater capacity.
- Decisions needed on where to parallelize.

Example from Prompt Reconstruction:
  - Double the resources of one farm.
  - Two farms in parallel.
- Needs careful evaluation before making decision.
Objectivity Related Issues

- Very Large Data Base (VLDB).
  - More than 64K files.
  - Accommodates FY01 luminosity and beyond.
  - File sizes driven by usage.

- Performance.
  - Reduced lock traffic.
  - Read-only databases for analysis.
Data Streams and Distribution

- Organize data into ~20 self-contained streams.
- Agreed-upon unit for data distribution.
- Hierarchical distribution via network.
- Upgrade connectivity.
  - At least one OC12 (622 Mbps) at SLAC.
Implementation Plans

- Just started on implementation plan.
- Based on CMWG’s recommendations.
- Bottoms-up approach.
- First round plans for next year by 10/20.
  - Are any goals inconsistent?
  - What tradeoffs are needed?
  - No glaring issues so far.
- Schedules and milestones.
Costs

- Disk cost is a major item.
- Staging of micro level data.
  - Physics driven prioritization.
- Use of cheaper disks.
  - Two sets ordered for testing.
  - May use different disks in different applications.
- Possibility of IN2P3 to share SLAC’s load.
  - Multiple Tier-A sites.
  - Partial data sample at each one.
  - Perhaps more attractive to funding agencies.
Non/Less Technical Issues

- Propagate improvements from central site to remote sites.
  - Focus on SLAC the first year.
  - Prompt Reconstruction and data access.
  - More attention on remote sites.

- Greater attention to costs.
  - Focus on technical problems the first year.
  - Greater pressure from increasing luminosity.

- Continued vigilance on priorities and schedules.
  - Computing provides resources and facilities.
  - Turning them into physics output requires experiment-wide priorities and schedules.
Summary

- Working at design luminosity.
- Luminosity upgrade presents challenges.
- No fundamental problems.
- Computing Model Working Group and Internal Review agree on direction.
- Developing implementation plan.
- Success of computing is not purely technical problem.