

## ***BABAR*: Requirements & Plans for Data Storage & Access**

David R. Quarrie  
*Software Technologies and Applications Group  
Information and Computing Sciences Division  
Lawrence Berkeley National Laboratory*

*LBL - MS 50B-3238  
1 Cyclotron Road  
Berkeley, CA 94720  
(510) 486-4868  
DRQuarrie@LBL.Gov*

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## **Overview**

- Requirements
- Macro-organisation: Event Collections
- Micro-organisation: Event Structure
- Plans
- Status & Issues

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Requirements

Table 1: Requirements

Characteristic	Size
No. of Detector Subsystems	7
No. of Electronics Channels	~220,000
Raw Event Size	~25 kBytes
DAQ to Level 3	2000 Hz 50 MByte/sec
Level 3 to Spool Area prior to PASS1 Reconstruction	100 Hz 2.5 MByte/sec
PASS1 Reconstruction (pseudo real-time)	100Hz 10.0 MByte/sec (2.5 input, 7.5 output)
Storage Requirements (includes Monte Carlo)	10 <sup>9</sup> events/year ~85 TBytes/year
Timescales	January 1998: Detector Commissioning October 1998: Cosmic Ray Run April 1999: Detector on Beamline June 1999: Physics Run

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Processing Model

- Rather conventional
- Multiple phases
  - Level 3
    - Final stage in real-time trigger/dataflow
    - Source of “raw” data
  - PASS1 Reconstruction
    - Reconstruction of tracks, vertices, clusters, etc.
    - Categorising of events into physics signatures
    - Preliminary calibration constants
  - PASS2 Reconstruction
    - Same as PASS1 but with better algorithms and/or constants
  - DST Analysis
    - Separation of physics signatures
    - Physics-specific reconstruction (constrained vertices etc.)
  - Statistics Analysis
    - Multiple scans over event sample applying cuts, iterating etc.
  - etc.

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Level 3 & PASS1 Reconstruction

- Raw Data spooled to disk from Level 3 farm
  - Several Level 3 nodes (~10)
    - Even if technically feasible for a single node to do everything
  - Units of ~30mins worth of data
- PASS1 Reconstruction run on spooled data
  - Pseudo real-time
  - Several Reconstruction nodes (~10)
  - Added reconstruction data expected to be 2x raw data (50KB)
- Baseline model is to use flat files for spooling
- Should we be writing into OODBMS directly from Level 3?
  - Advantages:
    - ~20% decrease in total bandwidth through PASS1
    - Uniformity - OODBMS is only storage medium
  - Disadvantage of “real” real-time rather than pseudo real-time
    - Limited upstream buffering to smooth over problems
- Requires some R&D

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Macro-Organisation: Event Collections

- $10^9$  objects not manageable in a single collection
  - Not a space issue
  - Even if only an event header is kept for each event
- Need Hierarchy of Collections (collections of collections)
  - Ideally have transparent iterators, but not essential
- Natural Granularity
  - Runs
    - Periods of order of hours of “stable” conditions
    - Management unit - a convenience in referring to data dating interval
  - Run Periods
    - Periods of several weeks/months of “stable” accelerator/detector conditions
    - Laboratory scheduling unit
- Use natural granularity?
  - Look at implications on space

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Run-based collections

- Raw Data: 25KB per event at 100Hz
  - Second: 100ev; 2.5MB
  - Hour:  $3.6 \times 10^5$ ev; 8GB
  - Day:  $9 \times 10^6$ ev; 200GB
- Multiple data writers (~10)
  - Level 3 nodes
  - PASS1 Reconstruction Nodes
- Multiple data readers (~10)
  - PASS1 Reconstruction Nodes
- Natural parallelism
  - Decreases database size per writer
    - ~800MB/hour raw data
    - ~1.6GB/hour for reconstructed data
  - Decreases bandwidth per writer/reader
    - 250KB/sec for raw data
    - 500KB/sec for reconstructed data

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Event Collections

- Choose a database size
  - Of Order of 100's of MB
  - Following discussion based on 100MB
    - Probably a bit small, but convenient base for calculations
- 100MB file size
  - ~8 mins of raw data per writer
  - ~4 mins of reconstructed data per writer
- Hierarchy
  - Experiment (10's of Run Periods)
    - Timescale of years
  - Run Period (100's of Runs)
    - Timescale of months
  - Run (100's of Databases)
    - Timescale of hours
  - Databases (~ $5 \times 10^4$  events)
    - Timescale of minutes
- All components in hierarchy necessary? (e.g. Run Period)

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Micro-organisation: Event Structure

- Goals
  - Extensible event structure
    - Reconstruction is essentially process of decorating an event
  - Support different clustering strategies
  - Avoid schema evolution
    - It will be necessary but only if we change our minds or forget something
- Many items within an event are collections
  - List of tracks
  - List of vertices
  - List of drift chamber hits
  - etc.
- Treat Event as a collection of collections
  - Make queries on event to locate collections

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Event Structure

- Take advantage of phases
  - Raw data
  - Reconstructed data
  - DST data
  - etc.
- Design top-level Event object as containing a VArray, with each element being an *ooRef* to another object corresponding to a processing phase.
  - Naturally extensible to accommodate more phases
  - Use of *ooRef* allows all other data to be elsewhere (other databases)
- Easy to locate data for a particular phase
  - Fixed offset, linear search, small hash table
  - Choose something
- Top-level Event object is small
  - Could contain some summary information as well as VArray

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Event Structure (Cont.)

- Design Processing Phase objects as containing VArrays, with each element being an *ooRef* to another object containing the actual collections
  - Naturally extensible to accommodate more collections
  - e.g. Raw Data object
    - Each VArray element *ooRef*'s an object containing to data from a detector subsystem
  - e.g. Reconstructed Data object
    - Each VArray element *ooRef*'s an object containing the list of tracks, or the list of vertices or...
  - Processing Phase objects can themselves contain summary information
- Easy to locate a particular component
  - Fixed offset
  - Linear search
  - Small hash table
  - Choose something
- Use of *ooRef*'s allows different clustering strategies

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Summary Information

- Can be kept at several levels
  - Event Header
  - Processing Phase Objects
  - Track List object
  - Detector Subsystem Object
- Useful for making queries with minimal data access
  - e.g. query based on number of tracks doesn't have to loop over tracks in track list to determine the number
- Most useful when supported by predicate language
  - Access to both data and function members

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Clustering Strategies

- Event headers in hierarchical collections as discussed earlier
  - Small
  - Lend themselves to replication for particular physics signatures
  - Note: Iterators for these collections need to know where they are in the hierarchy so that the application just “sees” a sequence of events.
- Raw Data for each event clustered together
  - Normal mode of access is by PASS1 and most of the information per event is accessed
  - Optimal for fast database population
- Data clustering for other phases?
  - Either cluster by event or by type
  - Needs better understanding of access patterns
  - Generally migrate from by-event to by-type for later phases
  - Might need to replicate data in both strategies

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

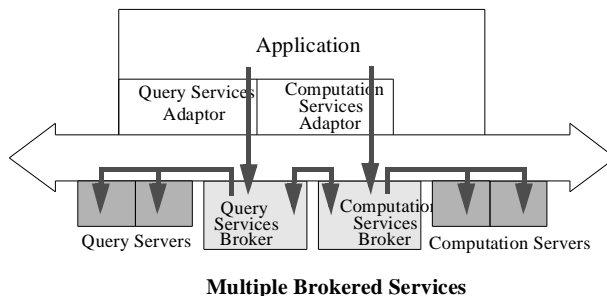
## Migration, Replication etc.

- Expect to replicate data
  - Event samples with particular physics signatures
  - Physics analyses taking place at different institutions
  - Based on access patterns
- Not database replication as supported by Objectivity
  - Goal is to reconcentrate a dilute signal
  - Cluster interesting data from interesting events close together
- PASS project raised many of the issues
  - Didn't necessarily address them

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Query Servers

- Definition
  - Given an input collection (of collections) and a predicate, return an output collection (of same type) of elements that satisfy the query
  - Essentially ODMG-93 query
- Wish to parallelise the query
- Plan is to use PASS Architectural Model (CORBA)
  - OID to output collection (of collections) in OODBMS is result



*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Plans

- Implement the BaBar Event Structure based on this extensible model
  - Details hidden from application programmers via abstract API
  - Requires some concrete implementation classes
  - Timescale April-May 1996
- Implement an Input Module within BaBar framework to access OODBMS
  - (c.f. CDF Analysis\_Control)
  - Application Modules independent of source of data
  - “Next event” just locates the next Event Header object in the collection
- Populate a small (GB’s) database to test Event Collection hierarchy strategy
  - Timescale June-July 1996

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Plans (Cont.)

- Investigate use of NERSC for large-scale prototypes
  - NERSC now located at LBNL
  - PDSF from SSC also now located at LBNL
  - Integration of Objectivity with HPSS
    - Aug-Sept 1996
- BaBar decision on adoption of this technology
  - Nov 1996
- Medium-scale prototype at NERSC (100's GB)
  - Incorporate Query Broker & Servers
  - Early 1997
- Prototype BaBar implementation available at SLAC
  - Nov 1997
- Production BaBar implementation available at SLAC
  - Dec 1998

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*

## Status & Issues

- Issues
  - Manpower
    - ~ 1FTE available in 1996 (not enough)
  - Licensing
    - Restricted to SLAC?
    - Regional centers?
    - Run-time vs application developers licenses
  - Objectivity support for ODMG-9x
    - Objy predicate language only allows access to data members
    - Support for ODMG queries
    - Iterators on VArrays
    - etc.
- Status
  - Programmer evaluating Objectivity for calibration constants database
    - Hope to use at least 50% on data storage project
  - Some PASS close-out funds still available
    - Need to discuss use in this context (~ 6-9 programmer-months)
  - Some of my time & Chris Day's time (both worked on PASS)
  - Preliminary design underway
    - This talk pretty much summarises it

*David Quarrie - BABAR: Requirements & Plans for Data Storage & Access*