



HENP Grand Challenge Meeting 30 June - 1 July 1997

BABAR Event Store

David R. Quarrie
Lawrence Berkeley National Laboratory
DRQuarrie@LBL.Gov



HENP Grand Challenge Meeting 30 June - 1 July 1997

2

Overview

- Goals
- Overall Design
- Data Model
- Event API
- Data Distribution Strategy
- Data Logging Strategy
- Management Tools
- Monitoring Tools
- Summary

6/29/97

David R. Quarrie: BaBar Event Store



Goals

- Provide complete coverage of all stages of event processing
 - Simulated Data ➔ Raw Data ➔ Reconstructed Data ➔ Event Summary Data ➔ Analysis Object Data ➔ Tag Data
 - ◆ Terminology taken from Atlas
 - Access to event data should be transparent apart from performance
- Deal with >150TB per year
- Deal with very large event samples
 - $\gg 10^9$
- Provide optimized data distribution across collaboration
 - 500 Collaborators, 80 Institutions, 10 countries

6/29/97

David R. Quarrie: BaBar Event Store



Overall Design

- BABAR DBMS covers several *domains*
 - Event Store
 - Conditions Database
 - Online Databases
- Objectivity is underlying technology
- Domains are logically independent
 - Actually tied together by underlying Objectivity Federated Database
 - ◆ Also transaction boundaries
- Allows independent development with late binding

6/29/97

David R. Quarrie: BaBar Event Store



Fundamental Concepts

- Physics event samples in named event collections
 - Traditional view
- Clustering of data at multiple levels
 - If you have to access a tape, try to access multiple databases
 - If you have to open a database, try to access multiple pages
 - If you have to access a page, try to access multiple objects
 - Predictive clustering & heuristic monitoring
- Clustering cannot solve problem of doubly-dilute data
 - Cluster for small fraction of events in sample
 - Cluster for small fraction of data per event
 - Only solution to both is to duplicate relevant data/events

6/29/97

David R. Quarrie: BaBar Event Store



Authorization Levels

- Several authorization levels to protect database contents
 - Domain specific
 - ◆ System, Group, User
 - System level grants update access to complete database
 - Group level grants update access to particular group's databases
 - ◆ e.g. Subdetector (Conditions DB) or physics group (Event Store)
 - User level grants update access only to user's data
 - ◆ Not meaningful for Conditions DB
- Anyone has read access to complete database
 - Including all subdetectors, physics groups & other users data
- Authorization levels are additional level of protection
 - File access permissions also give protection

6/29/97

David R. Quarrie: BaBar Event Store



Event Collections

- Event collections
 - Small
 - ◆ Limited to $\sim 10^6$ events
 - Large
 - ◆ Essentially unlimited (current implementation limited to $\sim 10^{11}$ events)
 - Run-based
 - ◆ Designed for possible data logging
 - Other implementations planned
 - ◆ Container-based, database-based etc.
 - All share common abstract interface (iteration etc.)
- Can be named and given aliases
 - Current implementation doesn't yet support aliases

6/29/97

David R. Quarrie: BaBar Event Store



Event Dictionaries & Registry

- One dictionary per user, group and system
- Contain references to named event collections
 - Names must be unique within a dictionary
 - Names may have aliases (not yet implemented)
- Mechanism by which a collection may be located and used as a source of events
 - Application locates the desired dictionary & then the required event collection
- Event Registry
 - Set of all dictionaries & other management information

6/29/97

David R. Quarrie: BaBar Event Store



Events

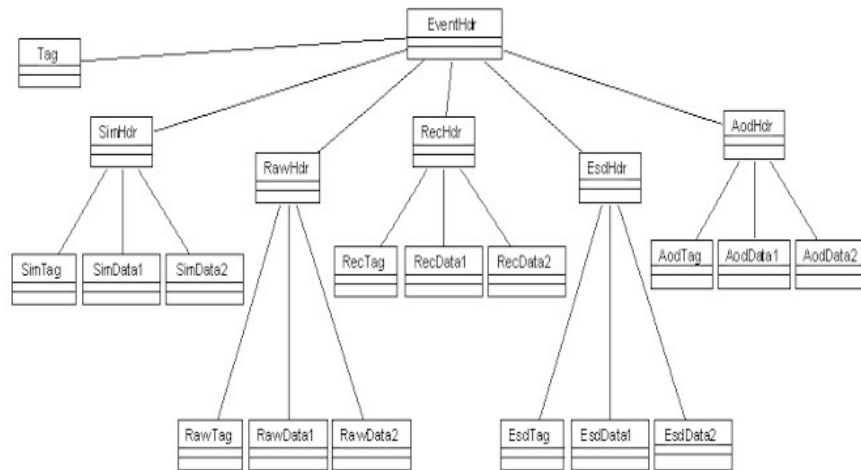
- Hierarchy of objects - *event header, stage headers & tags*
 - Stages
 - ◆ SIM - Simulated Truth where applicable
 - ◆ RAW - Raw Data (~32KB)
 - ◆ REC - Reconstructed Data (~100KB)
 - ◆ ESD - Event Summary Data (~10KB)
 - ◆ AOD - Analysis Object Data (~1KB)
 - ◆ TAG - Event Tag (100-200 bytes)
 - Terminology taken from Atlas
 - Stage tags ~1-10% of size of corresponding stage information
 - Event Header (<64 bytes)
- Goal is to keep navigation information small

6/29/97

David R. Quarrie: BaBar Event Store



Event Hierarchy



6/29/97

David R. Quarrie: BaBar Event Store



Event Tag

- Small amount of selection information
 - Attempt to make globally useful for all physics signatures
- Has a bidirectional association with event header
 - Provides the same navigational interface as the Event Header
 - ◆ Can traverse to locate any information from the event
- Event Collections can contain references either to Event Headers or Event Tags
 - Latter is most compact representation
 - ◆ Slightly more overhead for navigation
 - Another database to be opened & another navigation node to traverse

6/29/97

David R. Quarrie: BaBar Event Store



Event Collections & Access

- Different Event collections by default reference the same event headers for the same event
- Have flexibility
 - Copy the Event Header for dilute samples
 - ◆ Enhanced performance
 - Copy the Event Tag
 - Copy a portion of the event hierarchy
- Garbage collection is an issue
 - Delete those portions of the event that have been copied, but *not* the original portions

6/29/97

David R. Quarrie: BaBar Event Store



Clustering Strategy

- Strategy for creation of objects
 - Objects likely to be accessed together should be located close together within the same database
 - ◆ Access of one object pre-fetches adjacent objects (page-based server)
- Strategy for creation of database files
 - Final database will have >>>50,000 files
 - ◆ Cannot create them all in a single Unix directory!
 - Spread them throughout a directory hierarchy
 - ◆ Basis for file protection based on authorization levels
 - ◆ Basis for caching/migration with HPSS

6/29/97

David R. Quarrie: BaBar Event Store



Database Sizes etc.

- Fixed size database files
 - File size varies with processing stage & authorization level
 - 2GB seems a reasonable match to tape capacity
 - ◆ Tape seek performance 10^2 faster than read performance
 - Smaller for Group & User Authorization Levels
- Clustering Hint classes have to have a persistent component
 - Remember where to continue from
- Transient between stage based & class based clustering
 - Early stages use stage based strategy
 - ◆ Stage information for the event kept in same database set
 - Later stages use class based strategy
 - ◆ Objects of same class kept in same database set

6/29/97

David R. Quarrie: BaBar Event Store



Persistent Container Classes

- Objectivity intrinsic
 - ooVArray - extensible vector
 - ◆ Not persistent capable - must be embedded
 - ooMap - Hash Table
- Objectivity persistent versions of Rogue Wave Tools.h++
 - Obsoleted (based on Tools.h++ V6)
- STL-based transient & persistent classes
 - “ODMG-compliant”
 - Available on required platforms by spring 1998
- Bad timing for *BABAR*

6/29/97

David R. Quarrie: BaBar Event Store



Event API Testbeds

- Understand code changes
- Understand migration issues
 - e.g. Whether things
- Decide on what persistent classes we can use
 - Container classes
- Decide on set of access macros & clustering hint classes
 - Initially intended to use RD45
 - ◆ Proven to be unstable
 - Decided to use our own set of macros & classes to isolate ourselves

6/29/97

David R. Quarrie: BaBar Event Store



Transient API Testbed

- About 20 classes given transient/persistent siblings so far
 - ~20% of data model
- A set of smart pointer classes defers references between objects
 - Otherwise would have to create transitive closure of transient objects from their persistent siblings on access.
- Hash table of transient objects maintains one-to-one transient/persistent relationship
- No changes necessary to existing reconstruction code so far
- No performance studies done yet
- Feasibility demonstrated

6/29/97

David R. Quarrie: BaBar Event Store



Schema Evolution Strategy

- Minimize its use whilst realizing that it's necessary
 - Objectivity allows several mechanisms to update existing data
- Also problem is tight schema coupling
 - Event Registry *etc.* dependent on leaf nodes of event
- Decouple at Stage Headers using Objectivities RTTI
 - Type safe access through templates
 - Break dependencies
 - Concept already used for transient *BABAR* event
- Problem with Objectivity query predicate language
 - Probably can't traverse these Stage headers
 - ◆ Provide access to Stage Tags

6/29/97

David R. Quarrie: BaBar Event Store



Federated Database Strategy

- Protect master federated database
 - Contains schema and database catalog
- *Reference* federated database
 - Contains schema only
 - Used for builds of new *BABAR* software releases
 - Schema evolution disabled by default - managed evolution.
- *Production* federated database
 - Database catalog
- *Developer* federated database
 - Copy per developer, copy of *Reference* FDDB

6/29/97

David R. Quarrie: BaBar Event Store



Data Distribution Strategy

- Goals
 - Efficient access to data from anywhere in collaboration
 - ◆ From bulk processing engines
 - ◆ From desktops in any Institution
 - Flexible placement of data
 - ◆ Strategy should allow replication of data to optimize access
- Components
 - Regional Centers
 - ◆ Secondary data repositories
 - ◆ Better bandwidth between Institutions & their Regional Center
 - Database *Partitioning*
 - ◆ Replicate portions of database & act as local cache

6/29/97

David R. Quarrie: BaBar Event Store



Regional Centers

- SLAC
 - Main Collaboration Center (Regional Center for USA & Canada)
 - Primary repository for all data
- IN2P3
 - Primary European Regional Center
 - Replicates the complete data repository (present plan)
- RAL
 - Restricted subset of data
- INFN
 - Restricted subset of data

6/29/97

David R. Quarrie: BaBar Event Store



Database Partitions

- Replicate part of the *Federated Database*
 - Main partition
 - ◆ Master copy of the schema (C++ class definitions) & data
 - Secondary partitions
 - ◆ Copy of the schema & a subset of the data
- Secondary partitions act as a cache
 - When client accesses data, retrieve it from the “closest” location
 - ◆ Client & data in secondary ⇒ from secondary partition
 - ◆ Client in secondary & data not in secondary ⇒ from main partition
 - Transparent to client application
 - Database maintains synchronization between partitions

6/29/97

David R. Quarrie: BaBar Event Store

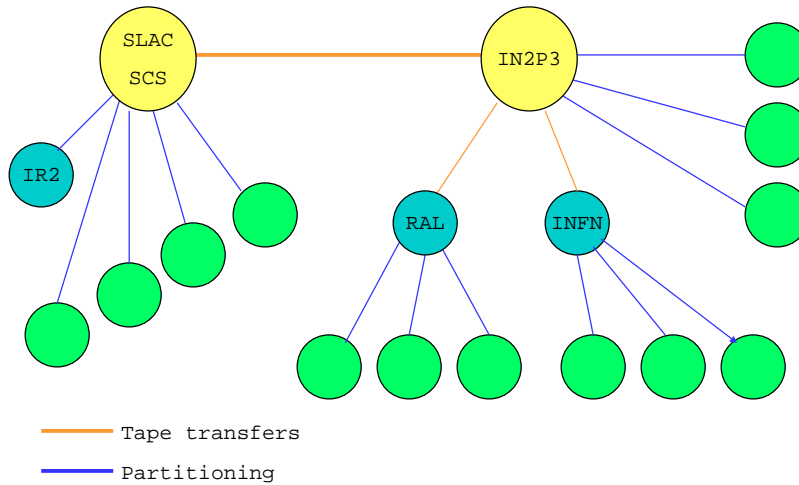


Database Partitions (2)

- Single Tier
 - Main partition can have multiple secondary partitions, but these themselves cannot have further children
- Database transactions are synchronized
 - Perform at the speed of the slowest secondary partition
 - Not a good idea to hold SLAC processing hostage to poor networking
- Ideal for IR2/SCS Synchronization of Conditions Database
 - IR2 as secondary partition with SCS as main partition



Data Distribution Strategy





Data Logging

- Current strategy (changing...)
 - OEP (Level 3 Trigger) outputs streamed event data
 - ◆ *Not* objects
 - Spooled to conventional disk files
 - ◆ Location of spool disk? Probably experiment
 - ◆ Stage to DLT tape if disk becomes full
 - Prompt Reconstruction inputs from spool disk files & outputs to database
 - ◆ First contact with HPSS
 - Each PR node writes to a different output database
 - ◆ Lock conflicts only when a database becomes full in updating event collection (single OID).

6/29/97

David R. Quarrie: BaBar Event Store



Data Logging - Conditions DB

- Both online & offline require access to conditions database
- Proposal is to make online a secondary Objectivity partition
 - SCS is primary partition
- Updates from online automatically propagated to offline
- Updates from offline automatically propagated to online
- Semi-autonomous in case of network failure

6/29/97

David R. Quarrie: BaBar Event Store



Management Tools

- Database management tools
 - Monitor usage & access efficiency
 - Cleanup database
 - Perform replication or re-clustering etc.
 - User Interface rather than programmatic interface
- Large Hole at the moment
 - Long term schedule details tasks but nothing available until next year
 - Hope can use SLAC Database/Web expertise to help us develop user interfaces

6/29/97

David R. Quarrie: BaBar Event Store



Performance Monitoring Tools

- Objectivity gives you some statistics gathering tools
- Not much thought gone into this yet
- I think it's crucial
- Part of the motivation for using BdbRef(T) rather than ooRef(T) *etc.* was to allow for some monitoring to be added
- Intend to spend more time on this in the next 6 months

6/29/97

David R. Quarrie: BaBar Event Store



Documentation

- Accessible from *BABAR* Computing Home Page
 - <http://www.slac.stanford.edu/BFROOT/doc/www/Computing.html>
 - ◆ Follow *Databases* link
- All documents in HTML format
 - Conceptual & detailed design documents
 - Reference Manual (class library)
 - User's Guide
 - Event API

6/29/97

David R. Quarrie: BaBar Event Store



Summary

- Good progress on overall design
 - Integration of database domains
 - Testbed prototypes
- Event API has been a really tough decision
 - Recent progress gives hope for light at end of the tunnel
- Much still to be understood about data distribution
- Scaling & HPSS integration issues
 - RD45/Objy/HPSS Meeting at Objectivity
- Staffing is an ongoing problem area
- Schedule looks tight but doable

6/29/97

David R. Quarrie: BaBar Event Store