



Grand Challenge Architecture and its Interface to STAR

Sasha Vaniachine

presenting for the **Grand Challenge** collaboration

(<http://www-rnc.lbl.gov/GC/>)

March 27, 2000

STAR MDC3 Analysis Workshop

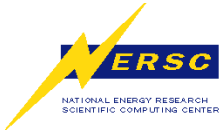


Outline

- **GCA Overview**
- **STAR Interface:**
 - *fileCatalog*
 - *tagDB*
 - *StGCAClient*
- **Current Status**
- **Conclusion**

GCA: Grand Challenge Architecture

- ***An order-optimized prefetch architecture for data retrieval from multilevel storage in a multiuser environment***
- **Queries select events and specific event components based upon tag attribute ranges**
 - **query estimates are provided prior to execution**
 - **collections as queries are also supported**
- **Because event components are distributed over several files, processing an event requires delivery of a “bundle” of files**
- **Events are delivered in an order that takes advantage of what is already on disk, and multiuser policy-based prefetching of further data from tertiary storage**
- **GCA intercomponent communication is CORBA-based, but physicists are shielded from this layer**



Participants

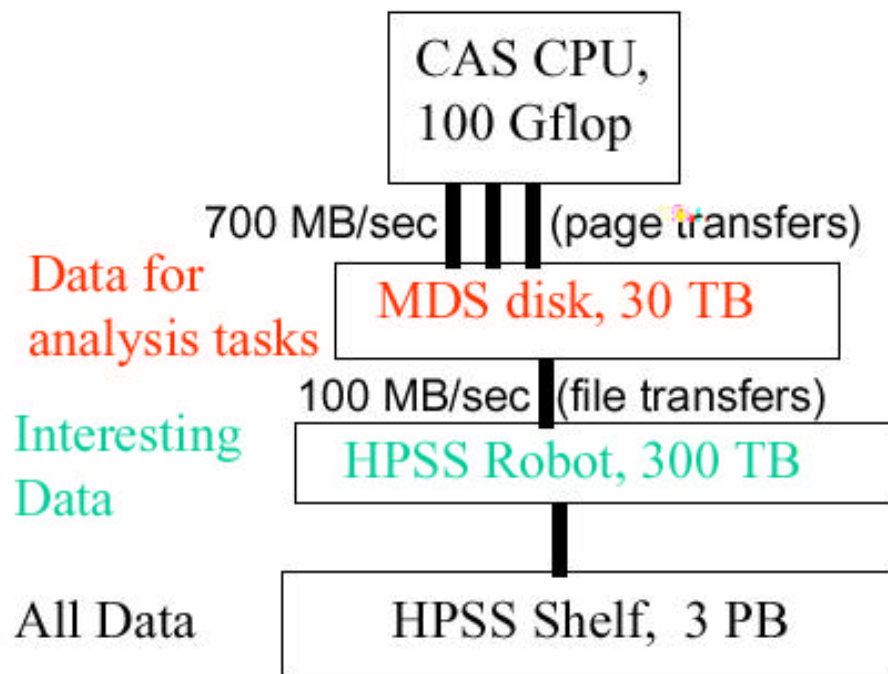
- **NERSC/Berkeley Lab**
 - L. Bernardo, A. Mueller, H. Nordberg, A. Shoshani, A. Sim, J. Wu
- **Argonne**
 - D. Malon, E. May, G. Pandola
- **Brookhaven Lab**
 - B. Gibbard, S. Johnson, J. Porter, T. Wenaus
- **Nuclear Science/Berkeley Lab**
 - D. Olson, A. Vaniachine, J. Yang, D. Zimmerman



Problem

- **There are several**
 - **Not all data fits on disk (\$\$)**
 - **Part of 1 year's DST's fit on disk**
 - What about last year, 2 year's ago?
 - What about hits, raw?
 - **Available disk bandwidth means data read into memory must be efficiently used (\$\$)**
 - **don't read unused portions of the event**
 - **Don't read events you don't need**
 - **Available tape bandwidth means files read from tape must be shared by many users, files should not contain unused bytes (\$\$\$\$)**
 - **Facility resources are sufficient only if used efficiently**
 - **Should operate steady-state (nearly) fully loaded**

Bottlenecks



Bulk bandwidth numbers meet estimated requirements assuming 100% efficiency.

How to achieve bulk bandwidth?

What fraction of data transferred is useful to programs?!!!

Keep recently accessed data on disk, but manage it so unused data does not waste space.

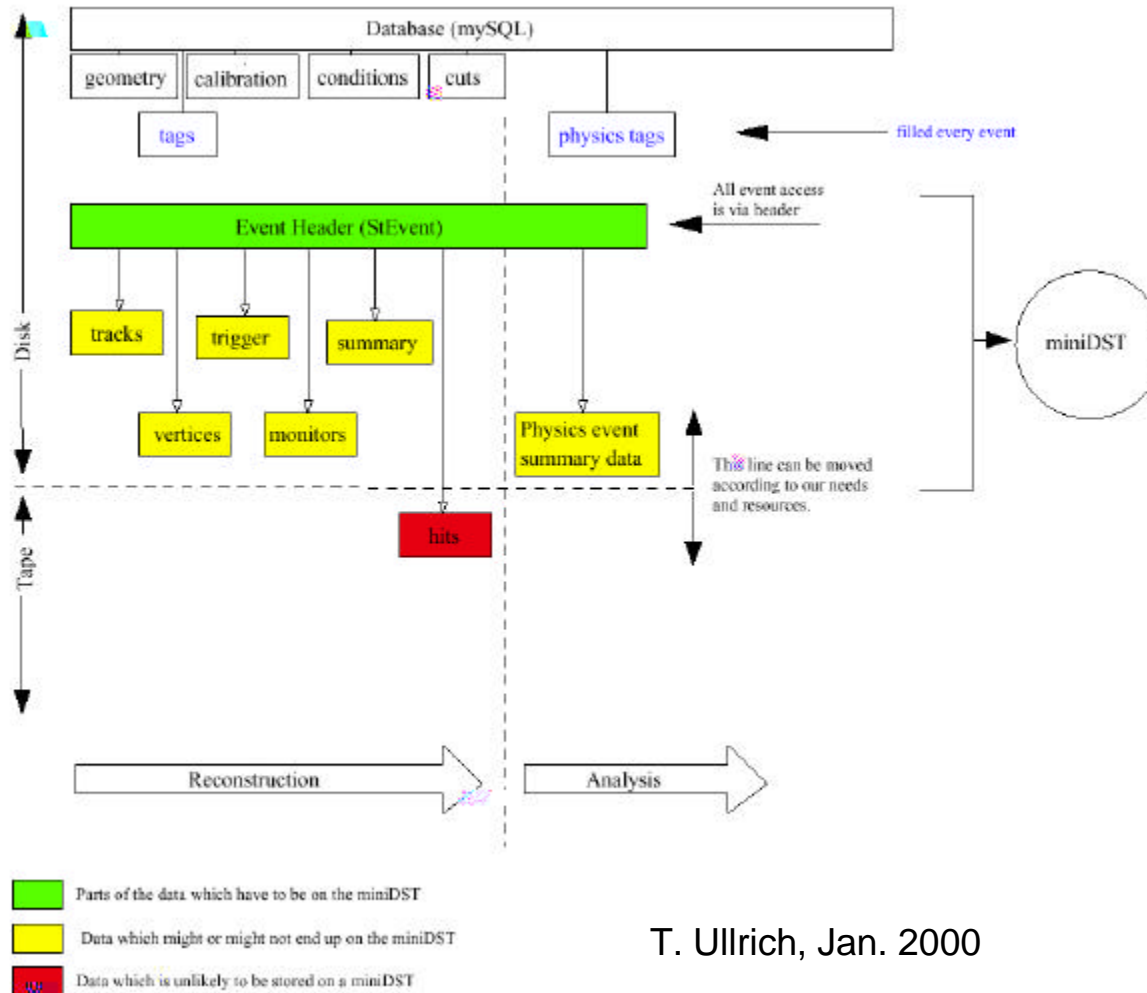
Try to arrange that 90% of file access is to disk and only 10% are retrieved from tape.



Solution Components

- **Split event into components across different files so that most bytes read are used**
 - Raw, tracks, hits, tags, summary, trigger, ...
- **Optimize file size so tape bandwidth is not wasted**
 - 1GB files, → means different # of events in each file
- **Coordinate file usage so tape access is shared**
 - Users select all files at once
 - System optimizes retrieval and order of processing
- **Use disk space & bandwidth efficiently**
 - Operate disk as cache in front of tape

STAR Event Model



T. Ullrich, Jan. 2000



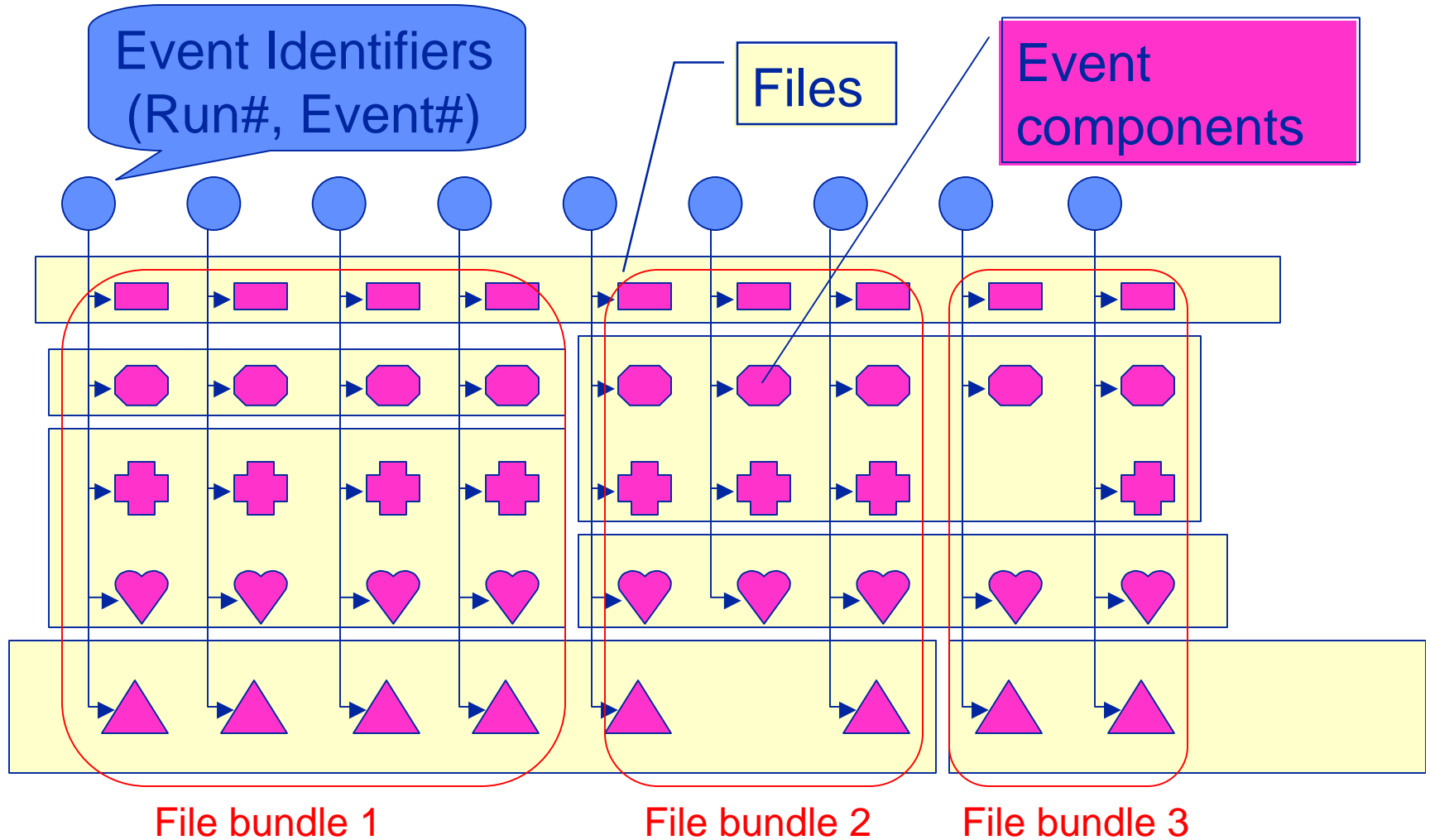
Analysis of Events

- **1M events = 100GB – 1TB**
 - **100 – 1000 files (or more if not optimized)**
- **Need to coordinate event associations across files**
- **Probably have filtered some % of events**
 - **Suppose 25% failed cuts after trigger selection**
 - **Increase speed by not reading these 25%**
- **Run several batch jobs for same analysis in parallel to increase throughput**
- **Start processing with files already on disk without waiting for staging from HPSS**

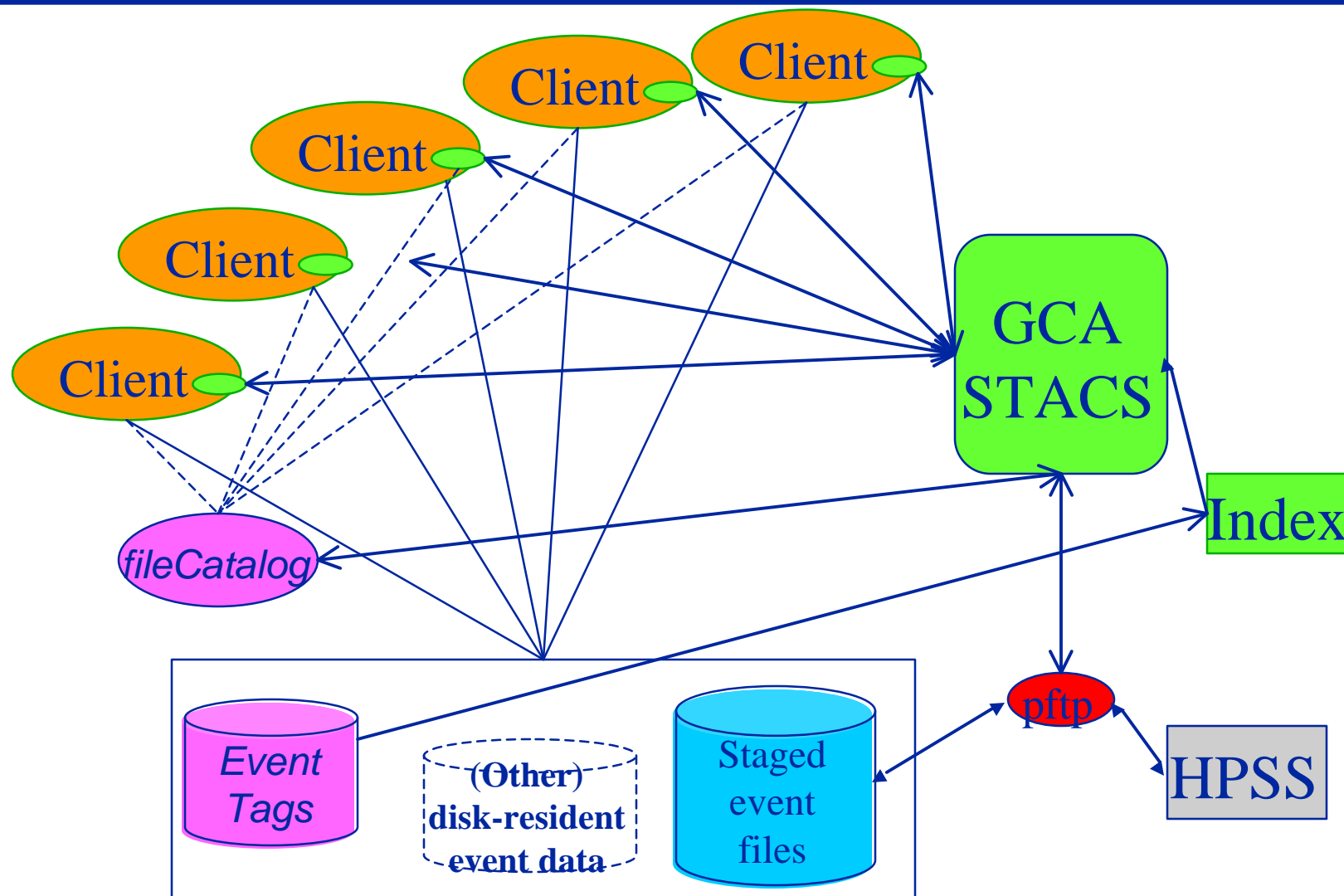
In the Details

- **Range-query language, or query by event list**
 - “NL>700 && run=101007”,
 - {e1,r101007;e3,r101007;e7;r101007 ...}
 - Select components: dst, geant, ...
- **Query estimation**
 - # events, # files, # files on disk, how long, ...
 - Avoid executing incorrect queries
- **Order optimization**
 - Order of events you get maximizes file sharing and minimizes reads from HPSS
- **Policies**
 - # of pre-fetch, # queries/user, # active pftp connections, ...
 - Tune behavior & performance
- **Parallel processing**
 - Submitting same query token in several jobs will cause each job to process part of that query

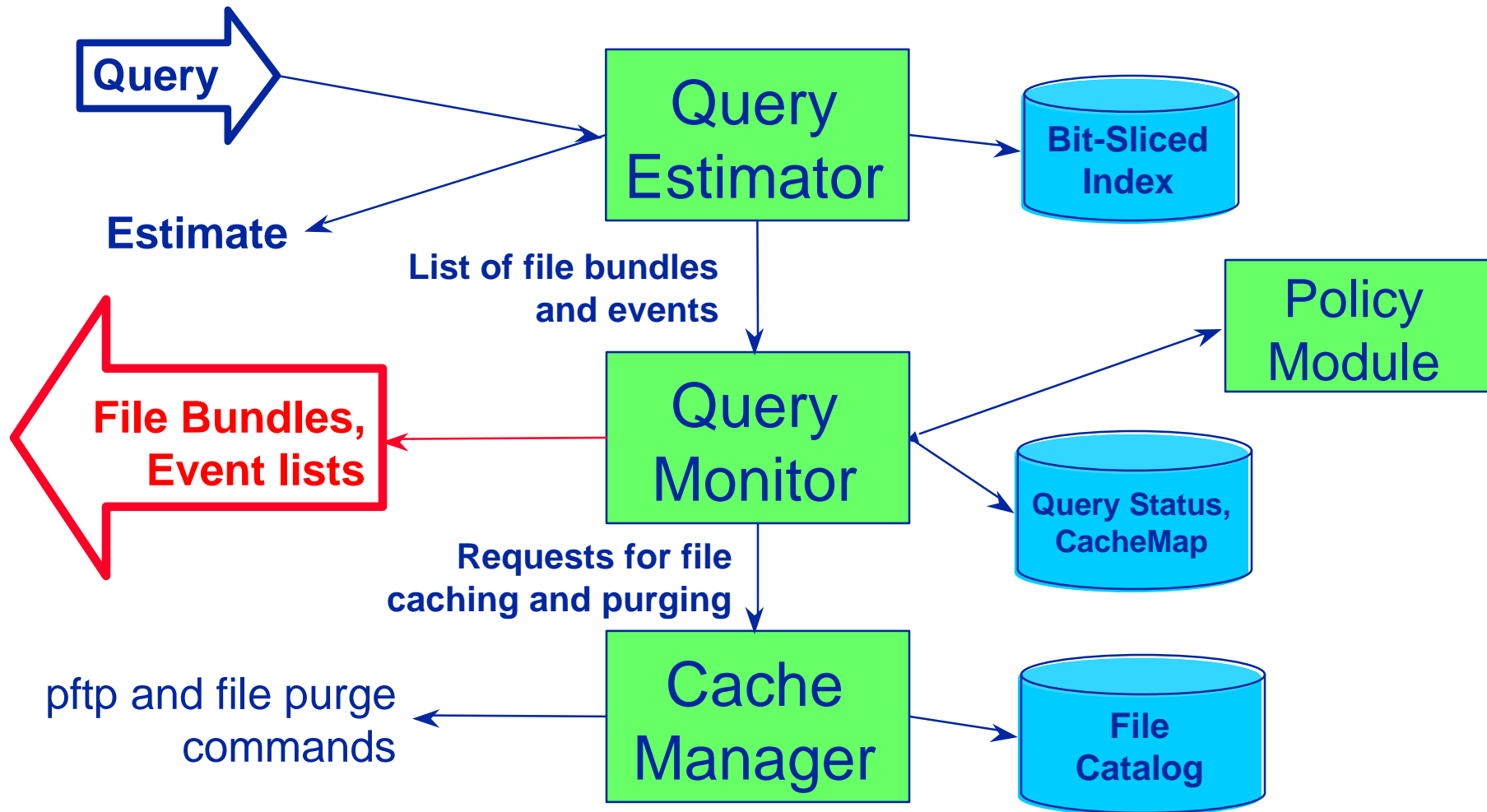
Organization of Events in Files



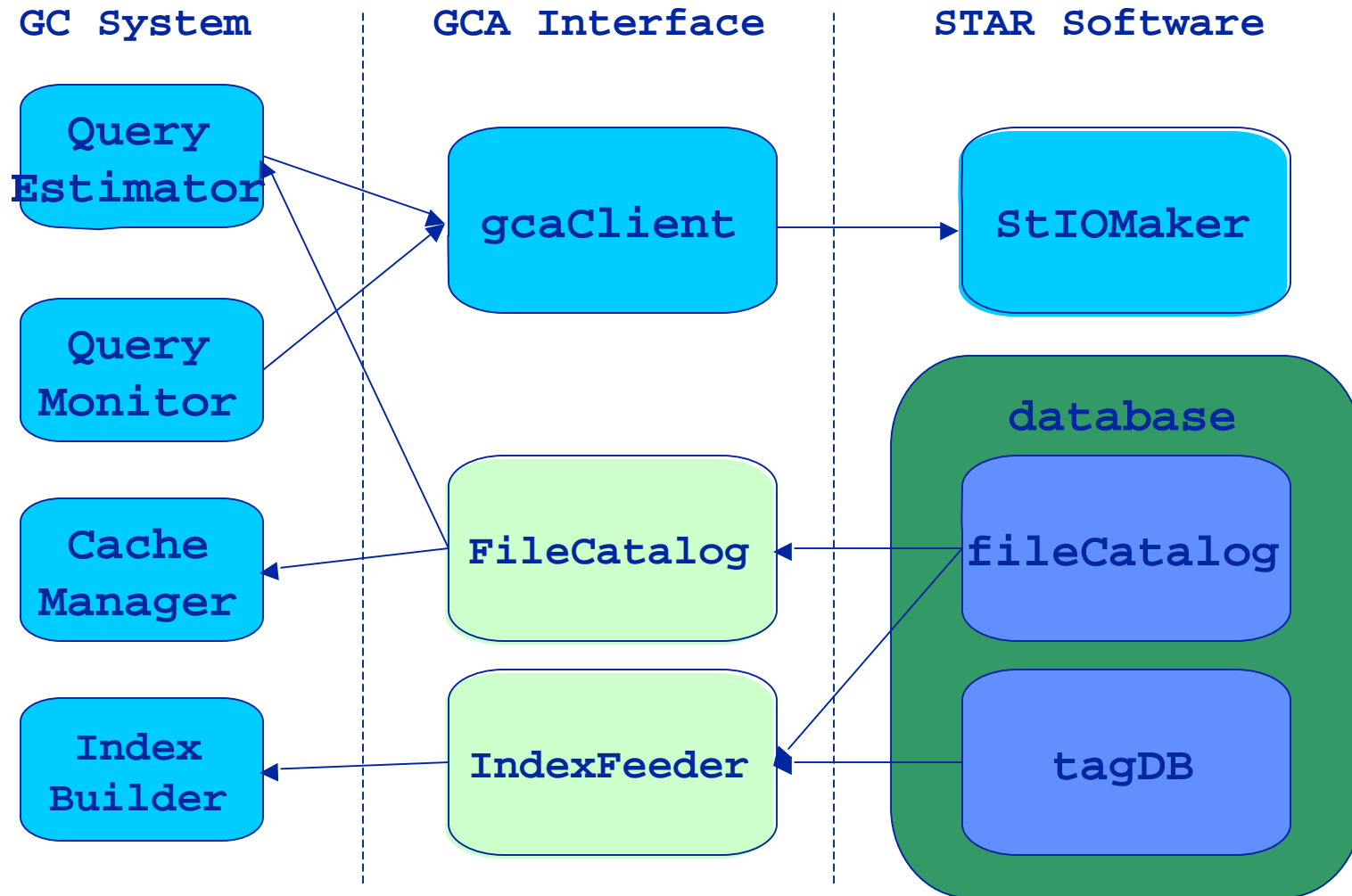
GCA System Overview



STACS: S**T**orage Access Coordination System



Interfacing GCA to STAR



Limiting Dependencies

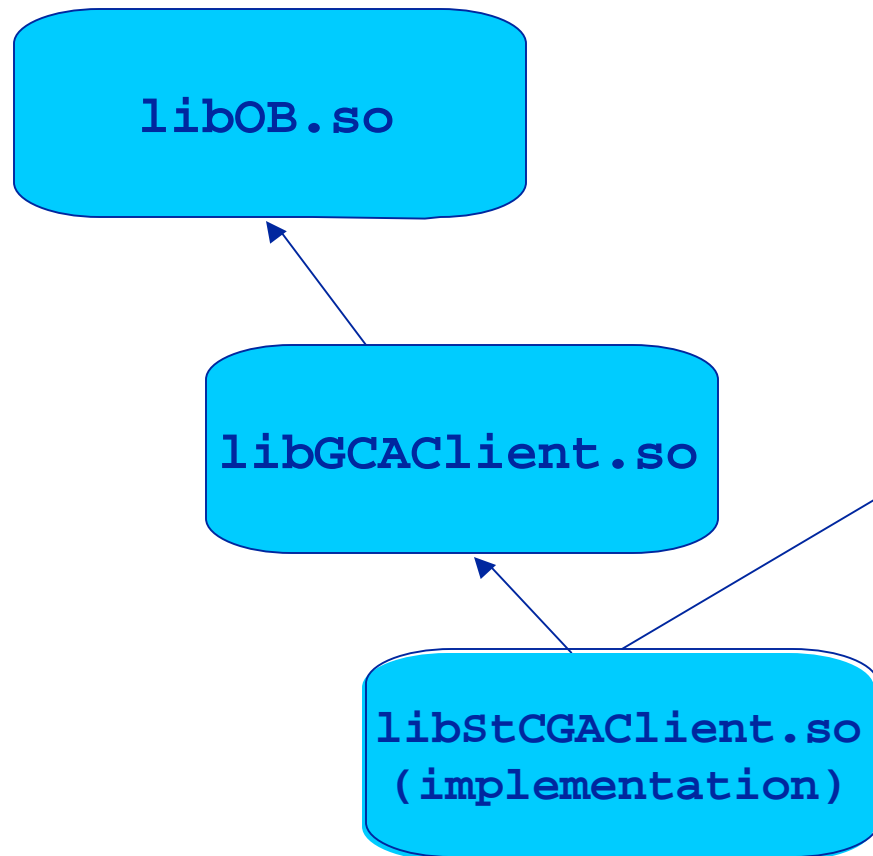
STAR-specific & GCA-dependent

- **IndexFeeder server**
 - IndexFeeder read the “tag database” so that GCA “index builder” can create index
- **FileCatalog server**
 - FileCatalog queries the “file catalog” database of the experiment to translate fileID to HPSS & disk path
- **gcaClient interface**
 - Experiment sends queries and get back filenames through the gcaClient library calls

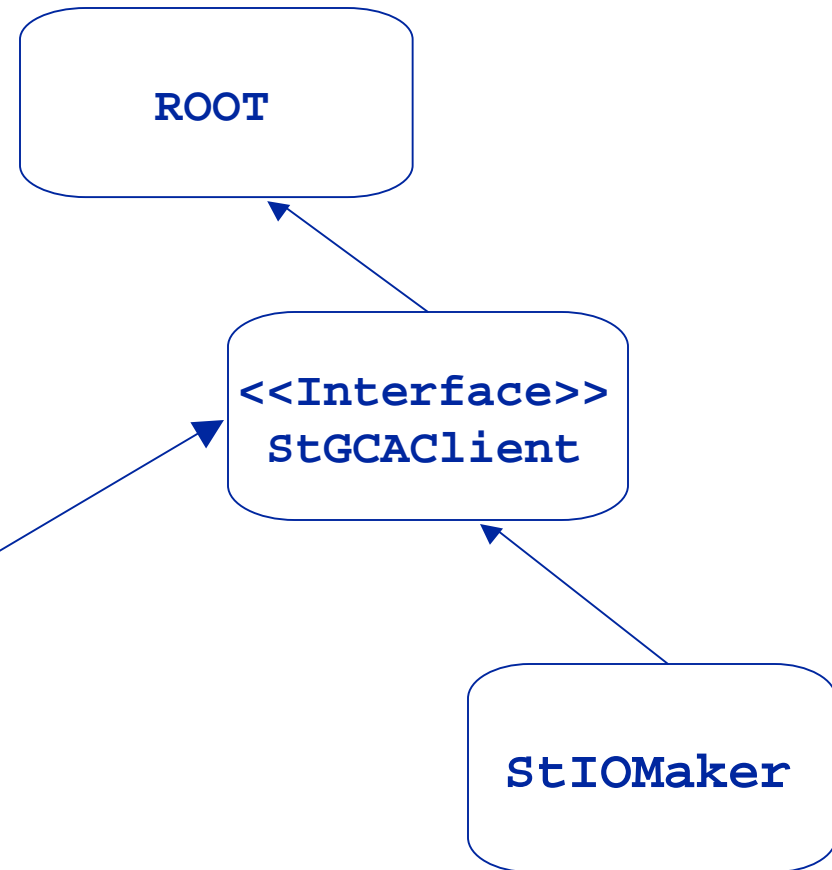
Eliminating Dependencies

CORBA + GCA software

`/opt/star/lib`



ROOT + STAR Software





STAR *fileCatalog*

- **Database of information for files in experiment.
File information is added to DB as files are created.**
- **Source of File information**
 - **for the experiment**
 - **for the GCA components (Index, gcaClient,...)**

Cataloguing Analysis Workflow



Job Catalog: Standard job catalog

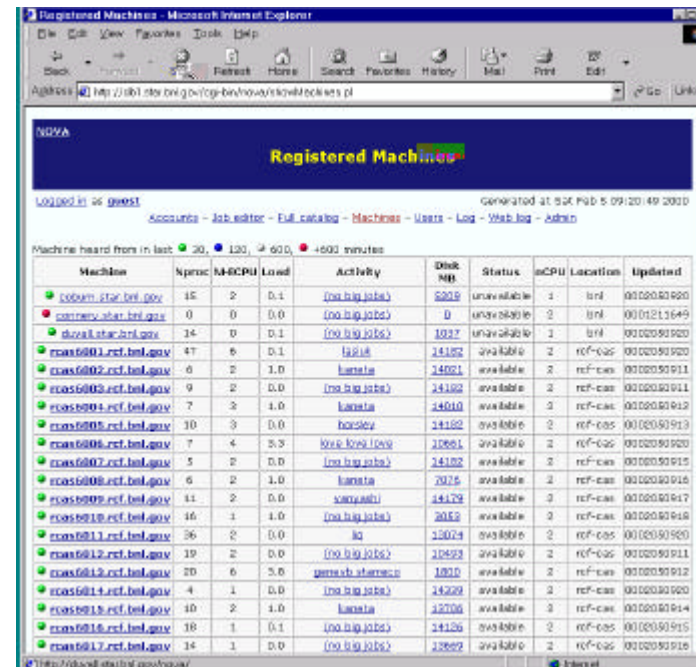
Logged in as guest Generated at Sat Feb 5 09:27:48 2010

accounts - job_editor - full_catalog - Machines - Users - Log - Web.log - Admin

standard Job 'test'

ExpGrid	sec0512_01_3Server.dat.coord	db@vanta.C	geo
ExpGrid	3D		
ExpGrid	3D		
ExpGrid	/data/xx00/ATAC/data/seqseq/		
Backend-01	StdEventBased		
BackendName-01	eventa		
Backend-02	StdAnalysisBased		
BackendName-02	analysia		

Job configuration manager



Registered Machines

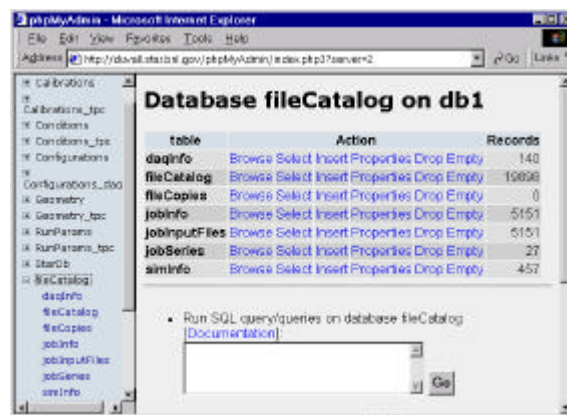
Logged in as guest Generated at Sat Feb 5 09:20:49 2010

accounts - job_editor - full_catalog - Machines - Users - Log - Web.log - Admin

Machine heard from in last 30, 120, 600, +600 minutes

Machine	Nproc	M-CPU	Load	Activity	Disk MB	Status	nCPU	Location	Updated
codevm_star.bnl.gov	45	2	0.1	(no job jobs)	2209	unavailable	1	bnl	00:00:00000
camerix_star.bnl.gov	0	0	0.0	(no job jobs)	0	unavailable	2	bnl	00:01:213649
duval_star.bnl.gov	14	0	0.1	(no job jobs)	1037	unavailable	1	bnl	00:00:00000
exas6001.rcf.bnl.gov	47	6	0.1	lsjls	14182	available	2	rcf-cas	00:00:00000
exas6002.rcf.bnl.gov	6	2	1.0	lanata	14021	available	2	rcf-cas	00:00:00011
exas6003.rcf.bnl.gov	9	2	0.0	(no job jobs)	14182	available	2	rcf-cas	00:00:00011
exas6004.rcf.bnl.gov	7	2	1.0	lanata	14010	available	2	rcf-cas	00:00:00012
exas6005.rcf.bnl.gov	10	3	0.0	hcsley	14182	available	2	rcf-cas	00:00:00013
exas6006.rcf.bnl.gov	7	4	3.3	lvsr_kvsr_jvsr	13881	available	2	rcf-cas	00:00:00020
exas6007.rcf.bnl.gov	5	2	0.0	(no job jobs)	14182	available	2	rcf-cas	00:00:00015
exas6008.rcf.bnl.gov	6	2	1.0	lanata	14075	available	2	rcf-cas	00:00:00016
exas6009.rcf.bnl.gov	11	2	0.0	scmashl	14179	available	2	rcf-cas	00:00:00017
exas6010.rcf.bnl.gov	16	1	1.0	(no job jobs)	2012	available	2	rcf-cas	00:00:00018
exas6011.rcf.bnl.gov	36	2	0.0	ls	13074	available	2	rcf-cas	00:00:00000
exas6012.rcf.bnl.gov	19	2	0.0	(no job jobs)	10493	available	2	rcf-cas	00:00:00011
exas6013.rcf.bnl.gov	20	6	3.6	gmsub_starters	10010	available	2	rcf-cas	00:00:00012
exas6014.rcf.bnl.gov	4	1	0.0	(no job jobs)	14209	available	2	rcf-cas	00:00:00000
exas6015.rcf.bnl.gov	10	2	1.0	lanata	13706	available	2	rcf-cas	00:00:00014
exas6016.rcf.bnl.gov	18	1	0.1	(no job jobs)	14126	available	2	rcf-cas	00:00:00015
exas6017.rcf.bnl.gov	14	1	0.0	(no job jobs)	13882	available	2	rcf-cas	00:00:00016

Job monitoring system



Database fileCatalog on db1

table	Action	Records
daqinfo	Browse Select Insert Properties Drop Empty	140
fileCatalog	Browse Select Insert Properties Drop Empty	10698
fileCopies	Browse Select Insert Properties Drop Empty	0
jobInfo	Browse Select Insert Properties Drop Empty	5151
jobInputFiles	Browse Select Insert Properties Drop Empty	6151
jobSeries	Browse Select Insert Properties Drop Empty	27
simInfo	Browse Select Insert Properties Drop Empty	457

Run SQL queries/queries on database fileCatalog
[Documentation]

Go

fileCatalog

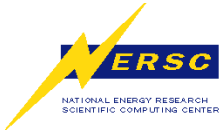
GCA MDC3 Integration Work

<http://www-rnc.lbl.gov/GC/meetings/14mar00/default.htm>

Goals:

14-15 March 2000

status	goal	description / summary (as of 16Mar2000)
done	1	Build index on new STAR files The index was build (several times) on the new STAR MDC3 data. This consisted of about 5,000 events. By the end of next week (start of MDC3) STAR expects about 140K events to put in the GC index. Sasha is continuing to accumulate additional event tag files as they are available.
done	2	Check that GCAClient and MinimalQuery work Modifications to GCAClient and the MinimalQuery (& MinimalQuery1) test programs were completed for the updated version of STACS, including the new file bundle flag on the iterator.
done	3	Run MinimalQuery on linux GCAClient & test program was compiled, run successfully on linux as well as Solaris. This included modifications to the Makefile to build both on linux & solaris.
done	4	Run multiple MinimalQuery simultaneously Run on linux. Not verified yet on solaris.
done	5	Test index update The feature of being able to update (add new events) to an existing index was just added. This feature was first tested during this period. A number of bug fixes were made and the basic procedure is working. John is continuing to investigate one or two bugs before the procedure is declared reliable.
done	6	Test index update while queries are running This is a functionality test and was successful. Any remaining work on the update functionality is not related to interlocks with running queries.
done	6.1	update between queries This check is to run a query before the update and then after and verify that the results are accurate. This was successful.
done	6.2	update while new queries are being submitted This tests the interlock mechanism so that queries do not run during the update process. This was successful.
in progress	7	Integrate GCAClient into root4star This is the final work to connect the GCA to STAR data analysis. There were various discussions among Victor, Sasha, Jeff, Frank, Dave, Doug. The basic idea of how to incorporate the GCAClient into StIOMaker has been worked out by Victor, Sasha & Jeff. Sasha & Victor will work on it.



Status Today

- **MDC3 Index**
 - **6 event components:**
 - fzd
 - geant
 - dst
 - tags
 - runco
 - hist
 - **179 physics tags:**
 - StrangeTag
 - FlowTag
 - ScaTag
 - **120K events**
 - **8K files**
- **Updated daily...**

User Query

ROOT Session:

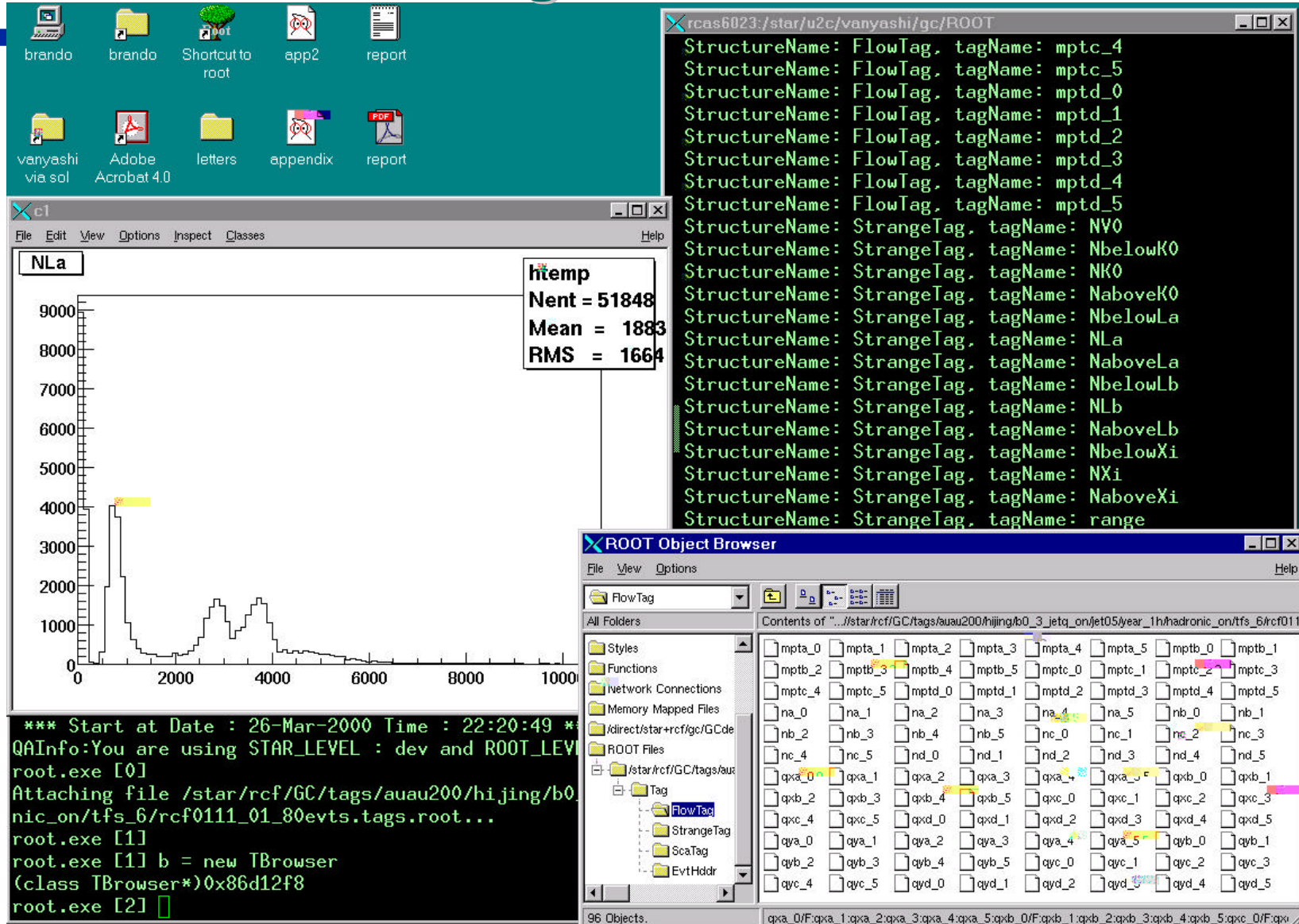
```
rcas6023:/star/u2c/vanyashi/gc/StGCAClient
1 mBeamPolarizationWest_0
1 mBeamPolarizationWest_1
1 mBeamPolarizationWest_2
1 mBImpact
1 mPhImpact
0 mGenerType
0 mBunchCrossingNumber
0 mEventNumber
0 mEventTime
0 mEventDate
0 mProdTime
0 mProdDate
qM: 0x86cb588
qE: 0x86cb158
fC: 0x86cba40
qF: 0x86cbd48
a->Init()
*** OldSource is not set.

Submitting query: SELECT dst
                WHERE -5<=qxa_3<0.3 && 22>qxc

qoF:: query created
qoF:: query added to list
query 0x86d0ce4
Full estimate is 205 events in 161 files ( unknown MBs).
```

```
rcas6023:/star/u2c/vanyashi/gc/StGCAClient
root.exe [0]
Processing test.C...
StGCAdapter::LoadGCAServer: libStGCAClient.so loaded
StGCAdapter::LoadGCAServer: new StGCAServer created.
StGCAServer::Init messages:
    I will not attempt to follow refs returned via the iterator.
gcaResources: Attempting to read configFile /star/rcf/GC/MDC3/stacs.rc
Using configuration file "/star/rcf/GC/MDC3/stacs.rc".
Narrowing QE reference found in /star/rcf/GC/MDC3/logs/SM_QE.ref
Converting (string_to_object) IOR:000000000000001549444c3a736d457374696d
00000005c000100000000000137273756e30302e7263662e626e6c2e676f76000006be00
2e676f763a5175657279457374696d61746f723a303a3a49523a736d457374696d61746f
Converted string_to_object
returning from findObjViaStringFile...
A Query Estimator has been contacted.
Converting (string_to_object) IOR:000000000000001849444c3a716d4576656e74
00000005c000100000000000137273756e30302e7263662e626e6c2e676f76000006c100
2e676f763a51756572794d6f6e69746f723a313a3a49523a716d4576656e744974657261
Converted string_to_object
returning from findObjViaStringFile...
A Query Monitor is available to your OrderOptIterator.
Narrowing FileCatalog reference found in /star/rcf/gc/GCdev/FC/FileCatal
Converting (string_to_object) IOR:000000000000001449444c3a46696c65436174
4000100000000000137273756e30302e7263662e626e6c2e676f760000883b00000000001
Converted string_to_object
returning from findObjViaStringFile...
A File Catalog has been found.
You are connected to a Query Factory.
Index Information
Name = simulated data for MDC3
Description =51749 events, 179 attributes, 6 components (all NULL FIDs a
00
number of components = 6
dst
fzd
geant
hist
runco
tags
0 tags
```

STAR Tag Database Access



The screenshot displays a Windows desktop environment with several open windows:

- Desktop Icons:** Includes 'brando', 'brando', 'Shortcut to root', 'app2', 'report', 'vanyashi via sol', 'Adobe Acrobat 4.0', 'letters', 'appendix', and 'report'.
- Terminal Window:** Shows the execution of 'root.exe' and the output of a STAR query. The output lists various tags and their corresponding structure names, such as 'FlowTag' and 'StrangeTag'.
- Plot Window:** Displays a histogram of 'NLa' with a callout box showing statistics: **htemp**, **Nent = 51848**, **Mean = 1883**, and **RMS = 1664**.
- Object Browser Window:** Shows a tree view of the STAR database structure, including 'FlowTag', 'StrangeTag', and 'ScaTag'.



Problem: SELECT NLa>700

3

STAR Tag Structure Definition

./pams/global/idl/FlowTag.idl

Version: [.DEV] [DEV00] [SL00b_2] [SL99f]

```

1 //
2 // $Id: FlowTag.idl,v 1.3 2000/01/13 23:18:06 snelling Exp $
3 //
4 // Event by event flow tag
5 //
6 // $Log: FlowTag.idl,v $
7 // Revision 1.3 2000/01/13 23:18:06 snelling
8 // Changed sum pt to mean pt
9 //
10 // Revision 1.2 1999/11/16 20:59:40 snelling
11 // Removed unused tags and added 6th harmonic
12 //
13 // Revision 1.1 1999/02/09 21:42:21 wenaus
14 // Final (?) versions of MDC2 PWG tags
15 //
16 // The tags are defined for 4 subevents (a,b,c,d) and 6 harmonics
17
18 struct FlowTag {
19     float qxa[6], qxb[6], qxc[6], qxd[6]; /* x component Q vector */
20     float qya[6], qyb[6], qyc[6], qyd[6]; /* y component Q vector */
21     long na[6], nb[6], nc[6], nd[6]; /* multiplicity */
22     float mpta[6], mptb[6], mptc[6], mptd[6]; /* mean pt */
23 };

```

**Selections like
 $\sum q_x a^2 + q_x b^2 > 0.5$
 can not use index**

Conclusion

- **GCA developed a system for optimized access to multi-component event data files stored in HPSS.**
- **General CORBA interfaces are defined for interfacing with the experiment.**
- **A client component encapsulates interaction with the servers and provides an ODMG-style iterator.**
- **Has been tested up to 10M events, 7 event components, 250 concurrent queries.**
- **Is currently being integrated with the STAR experiment ROOT-based I/O analysis system.**