



---

# Progress and Plans for CDF Databases

ODS/DBA “Taking Stock”  
Meeting

15 Oct 2002

Alan Sill, TTU



## What Is The CDF Database?

---

- **Consists of 7 basic applications (sets of tables with defined schema)**
  - Hardware: Contains configuration parameters for data taking hardware
  - Run Configurations: What the conditions were for a given data run
  - Trigger: The decision criteria and paths by which events are chosen
  - Calibrations: Measured responses of detectors & hardware under known or reproducible conditions to map out variations and instrumental drifts
  - Slow Controls: Long-term monitoring of voltages, temperatures, etc.
  - Data File Catalog: An offline index of the files containing data taken
  - SAM: A “super-DFC” with enhanced functionality for labeling and access
- **Each of the above applications has a person in charge of its operation, care & feeding (both schema and code)**
  - These people are called “Application Coordinators”
- **In addition, there are people in charge of matching calibrations, etc. to experimental conditions.**



## How CDF Databases Are Used

---

- Contains information critical to correct analysis of events:
  - Variation of detector response and calibration vs. conditions.
  - Changes of known settings and commands to hardware.
  - Information needed to be able to gather similar data together.
  - Database contains only <0.1% of information from experiment, but it is crucial to proper analysis.
- Need for access to information depends on the analysis stage:
  - Some constants and derived parameters (e.g. beam lines and alignment) are only known through extensive analysis
  - Interdependencies between tables exist and need to be kept consistent.
- Information can change through later analysis (alignments, calibrations, etc.), and a need exists to be able to apply such changes retroactively.
- Traceability and reproducibility of analyses are required and essential.



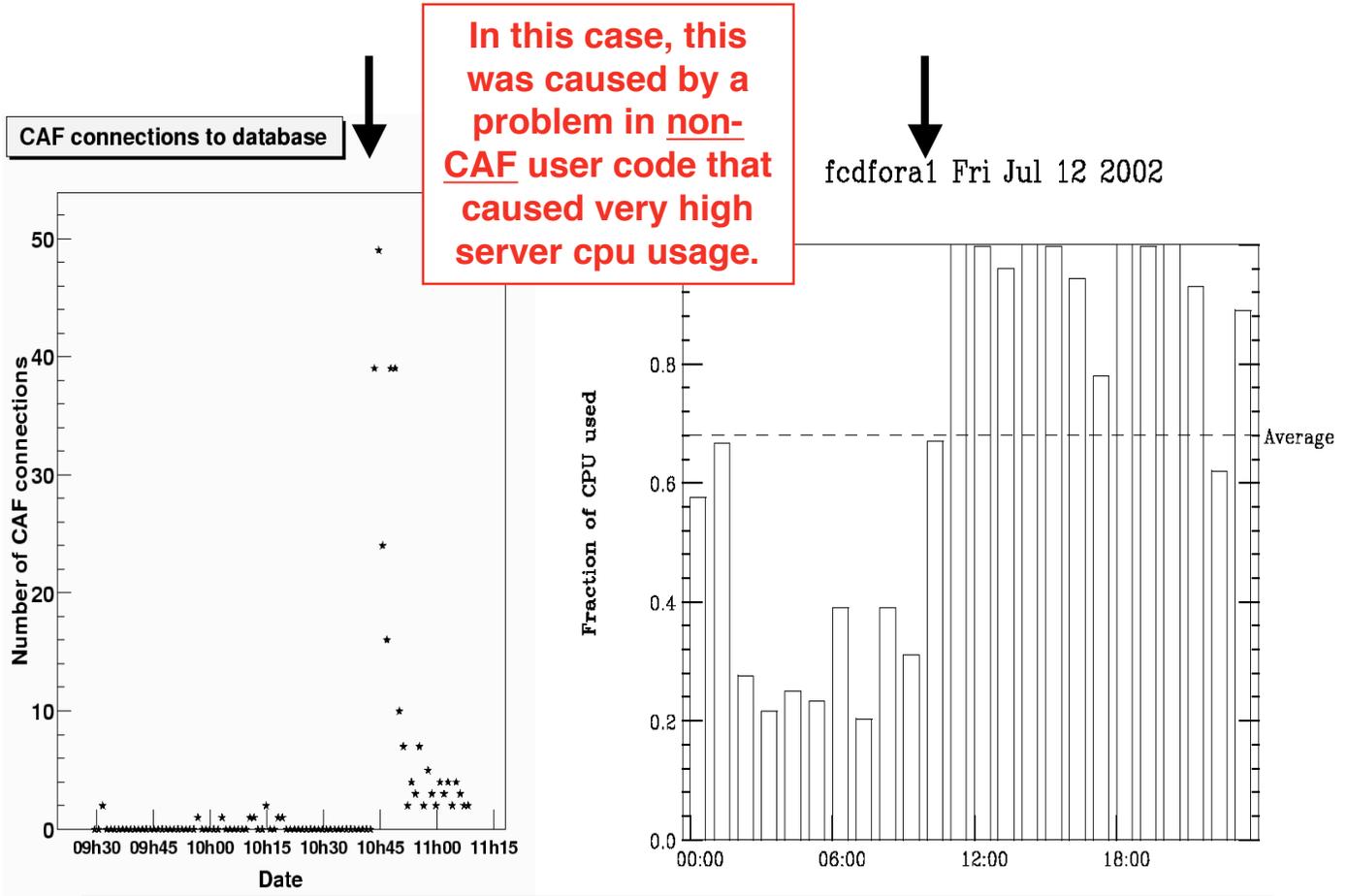
## Current status

---

- Online server cdfonprd for DAQ support, plus 2 offline servers (cdfofprd = basic replica of cdfonprd w/ writeable master of file catalog info.; cdfrep01 = 2nd basic replica of cdfonprd + basic replica of DFC) for all other uses.
  - License usage concerns.
  - Offline servers subject to overloads.
  - Spikes can be caused in usage due to bugs in code and increasing number of deployed cpus.
  - By tuning things carefully, we presently are able to keep up with the load.
- Large numbers of future analysis cpus will soon be deployed:
  - CAF plans to grow; DCAF, SAM, etc. being implemented
  - Off-site institutions plan to (and in some cases already have) implement farms of tens to hundreds of additional computers to do analysis and Monte Carlo production.
- Present usage patterns do not scale to fit within existing resources for serving CDF database contents to the world.
  - We will exceed available resources again in the near-term future unless we continue to improve server capabilities.



## Example of DB overload. Initiated by CAF? Not exclusively...



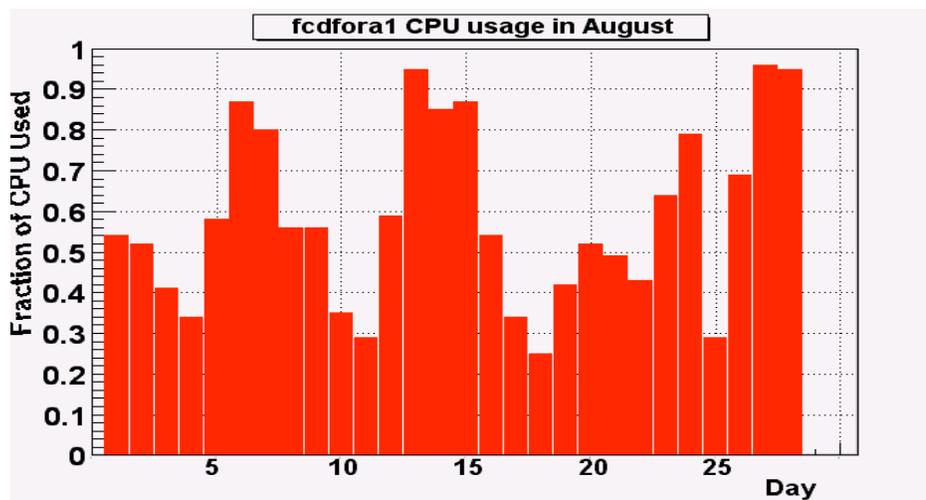
**Spikes of 100% usage lasted 7 hours until user problem was found.**

- Loads of up to 95% CPU use over day, 100% for hours.
  - Causes long delays, connection timeouts, and interferes with farms operations.
  - May be related to simultaneous read and write usage, replication, and other load factors

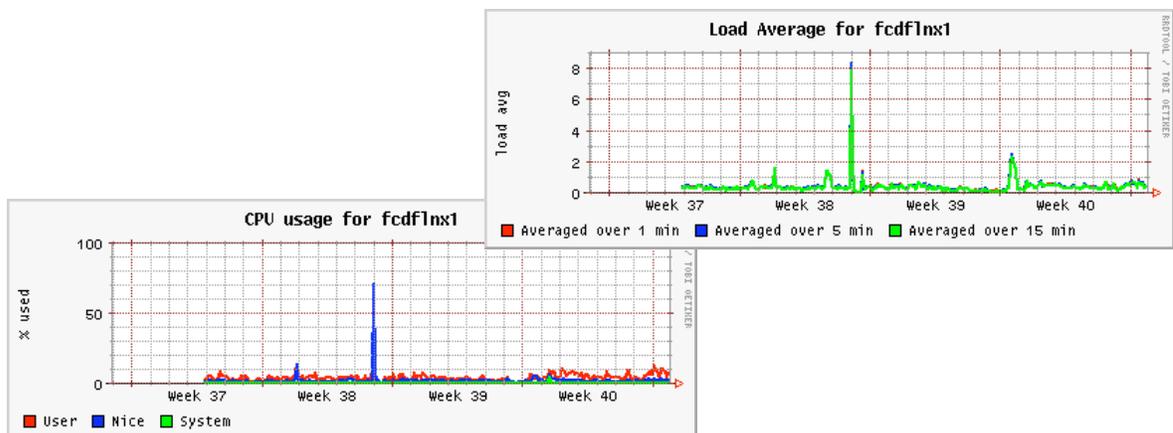


## High cpu load on fcdfora1

- Began to occur in February
- Severe problem May - July, partly remedied by repairs and changes in CDF code
- Typical usage graph in late summer:



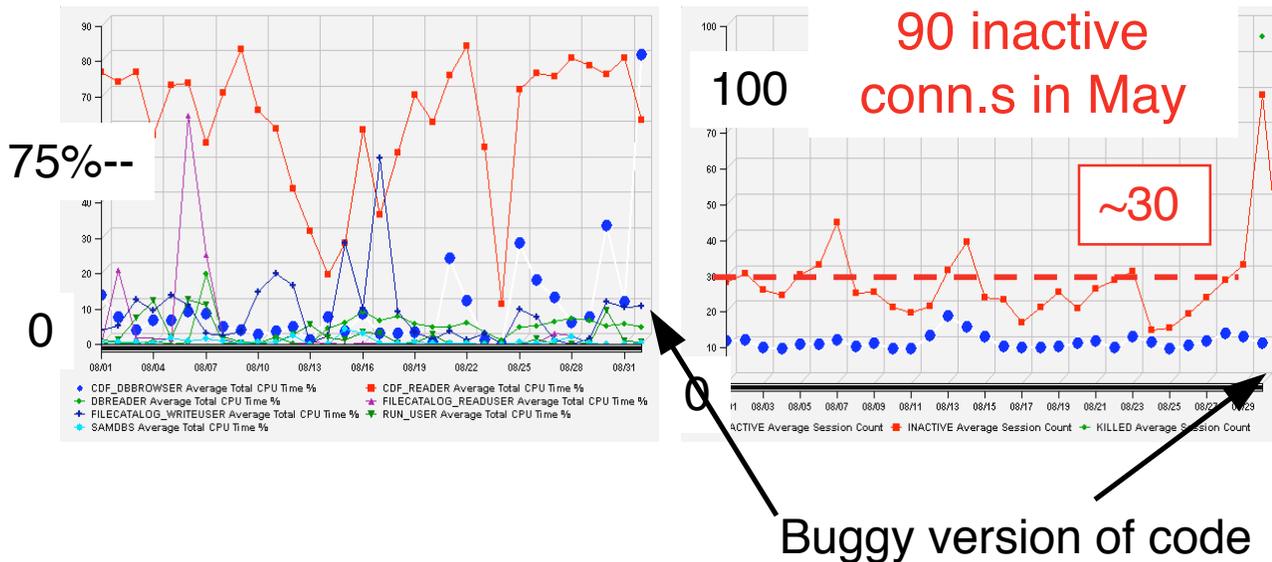
- Recent load has been near 100% for days at a time on fcdfora1, while replica is mostly idle.





## More illustrative plots:

Adding the replica and fixing CDF code was a good thing:



- Need more control of inactive connections!
- (Already better however than it was in Apr/May)
- Spikes in connections usually caused by bugs not proper usage
- More cpu (and I/O power) to serve user requests in a timely manner is a good thing, and reduces overall license use.



# Current Projects

## (What have we done to get a handle on this?)

---

- DB connection monitoring statistics package:  
(Jim Kowalkowski, Yuyi Guo, Rodolfo Pellizzoni)
  - Based on ErrorLogger
  - Reports results to separate logging server (complete)
  - Changes to CDF code essentially complete
  - Work proceeding on reporting layer
  - User control of detail level possible on a per-job basis
  - Intended to be our primary tool for finding out connection usage patterns
- Calibration API & DBObjects/DBManager support:  
(Jim Kowalkowski, Dennis Box, Yuyi Guo)
  - Not usually listed by the CD as a distinct project, but:
  - Incredibly important to recognize that we spend a lot of our time chasing bugs and features uncovered by, or updates requested by users
  - Need a large amount of programmer time to do all these tasks!
  - Examples:
    - “Get by Process Name,” “Get All Instances Over Run Range” (new functionality, needs design by experts)
    - Connection management support, “metering patch,” etc.
  - Code review on connection code, done results Thurs. 10/17



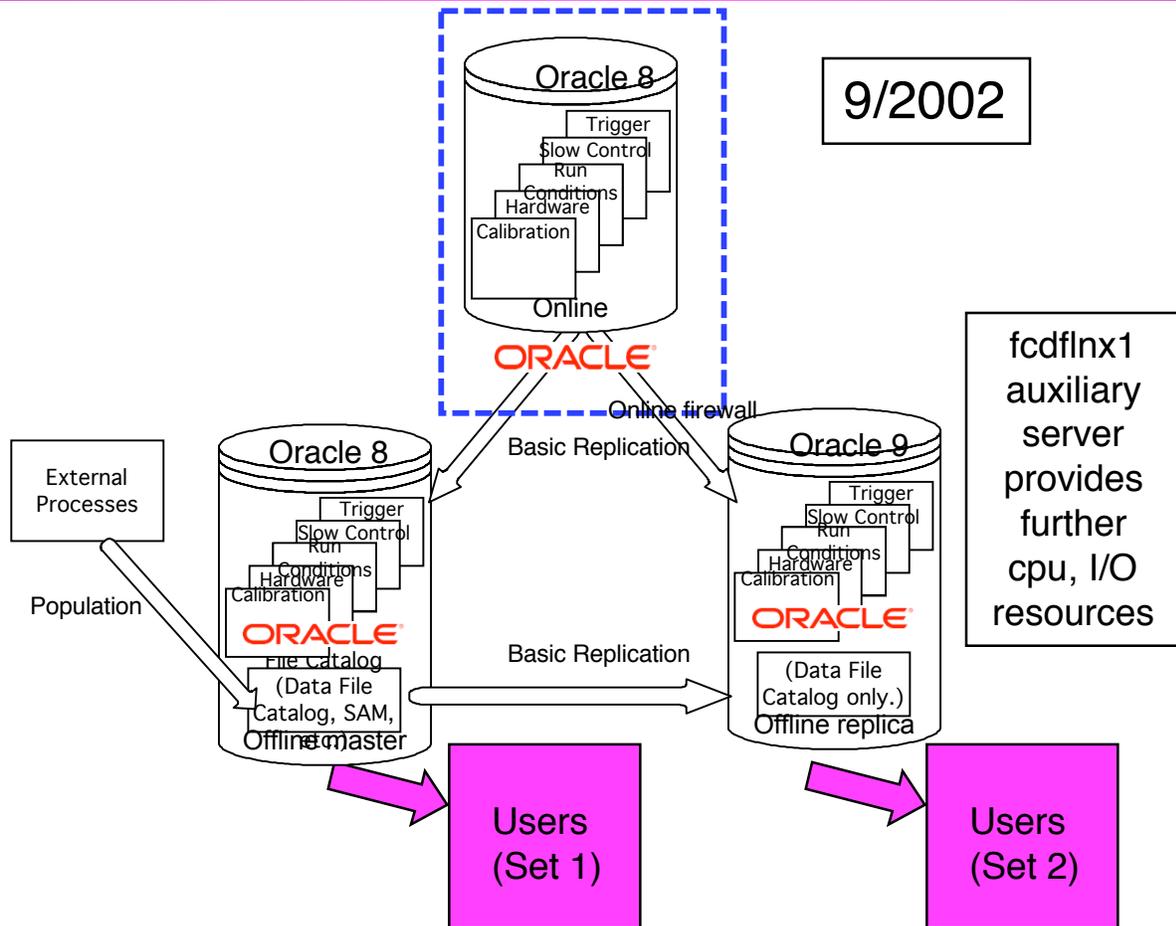
## Current Projects, cont'd

---

- New API development: (No one defined to do this yet)
  - Need true APIs for all other DB applications (Hardware, Run Configurations, Trigger, Slow Controls)
  - Lack of APIs leads to lots of direct Oracle calls in user code
  - Solving this is a necessary predecessor to deploying a third tier or freeware
- Datastreams Replication: (Anil Kumar, Nelly Stanfield)
  - Study implementation of 9iv2 datastreams for db replication
  - Waiting on 9\_2\_0\_2 patch from Oracle
  - Essential part of our future planning (see diagrams)
- Freeware investigation:  
(Svetlana Lebedeva, Richard Hughes, David Waters)
  - New joint CDF/CD project
  - Initial goals: reproduce MySQL calibration-only database done by two CDF collaborators, test, deploy, & support.
  - Move on from this to investigate alternatives (PostgreSQL, etc.) and to study classes of support needed for various user job types.\
  - Testing essential.
  - Partial road map exists, but people are in short supply.
  - Could be very important project in the future.



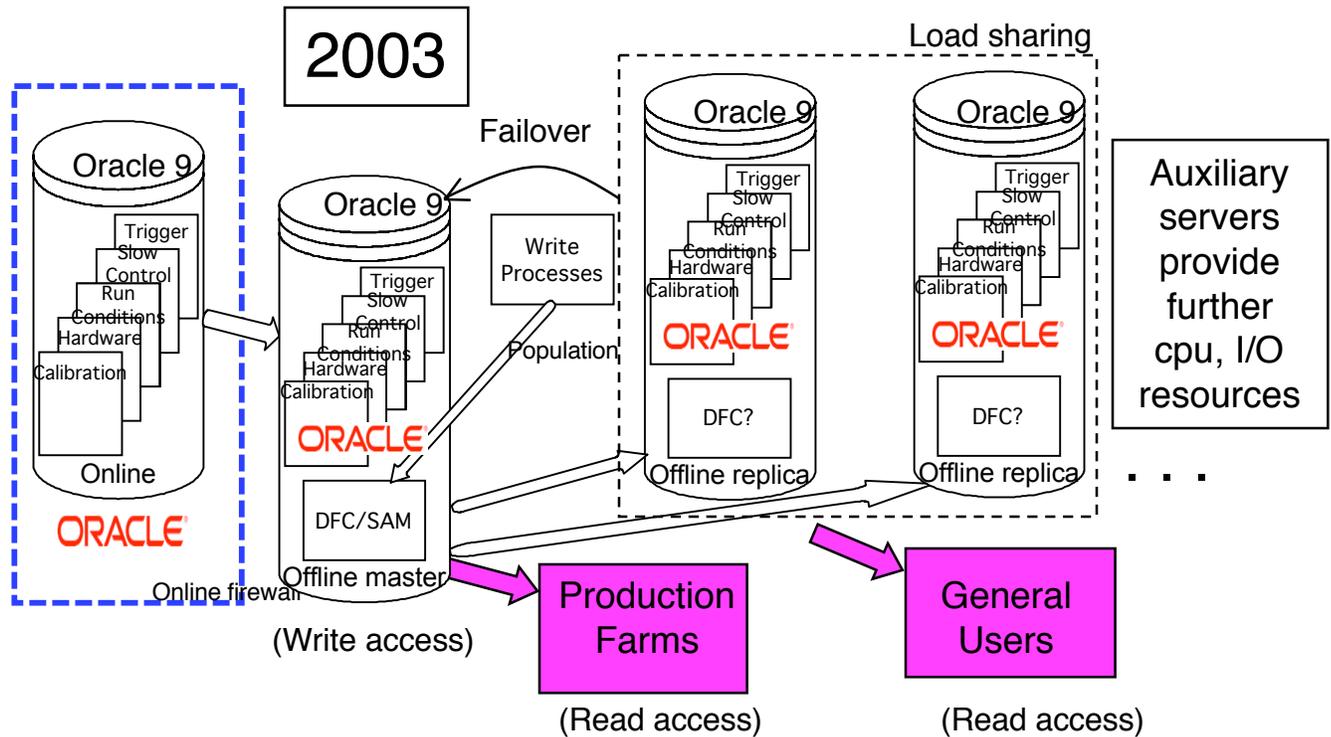
## Current distribution scheme



- 5 main database applications originate online, replicated twice (separately) to offline
- Data File Catalog, SAM added to and reside on primary offline server
- DFC only replicated from primary to replica offline server by basic replication



# Distribution via Oracle



- Oracle 9iv2 “datastreams” replication allows many choices with potentially improved replication scenarios.
- SAM takes over DFC. (One copy, or many? One in present scheme.) Need to decide on read-access portion.
- Connection broker may be needed in the future to assign servers to user jobs, but for now tnsnames-based load sharing and failover capabilities can be used to do this.
- Our current plan is to replace fcdflnx1 with more modern, more robust hardware populated by 9iv2 datastreams.
- Fate of Sun servers not decided. For now, add more disk.





# Task List

---

- **Database monitoring - connections, tables, durations, etc. (CDF/CD)**
  - Analyze usage patterns to help set designs and policies.
  - Develop validation for DB access in new versions of CDF code
  - Profile programs of various sorts using new tools (simulation, etc.)
- **Freeware port of database and replication (CDF/CD)**
  - Participate and help with comparison of freeware choices
  - Scripts to help with population and updating of freeware db
  - Validation of contents to check that they are the same as main db
- **API design, coding and testing (CDF/CD)**
  - Extend calibration-API-like features & design to other databases
  - Use “physicist insight” to determine what we need for analysis
- **SAM/Database/CAF/Grid test stand (CDF)**
  - Small-scale array to test grid concepts for distributing & connecting to database in various ways
- **Consider online needs; possible caching for Level 3**
  - Reduce license usage from present levels (~60 for online!)
  - Improve startup time for Level 3



# Task List, Cont'd

- **Connection broker / negotiator design & testing**
  - Eventually, some kind of broker or negotiator is likely needed.
  - Could be useful for online systems (DAQ & Level 3).
  - Balance latency, ease-of-use, license and other considerations
  - Help design this, test, and make sure that it works off-site as well as on.
- **Study of slow controls and monitoring system**
  - Tables need to be classified and redesigned with eye toward analysis
  - Decide what to save & store for archival vs. replicate (see below)
  - Accessors and methods to use data in analysis programs (B field, etc.)
- **Replication procedures and design**
  - Do we want or need “pull on demand?” (Secondary sourcing)
  - Prefetching? Fetching in batches at begin job rather than by run?
  - Should all data be sent to all servers, or only a subset?
- **Migrate towards the Grid.**
  - How will we broker true global database access in the future?



# What is the ODS/DBA “DBA” role in all of this?

---

- **Keep us running!**
  - Replication, table analysis, indexing reviews, 24x7 support of main production servers, etc.
  - Answer questions from users, grant roles and update permissions upon request, etc.
  - Train new people (application coordinators, etc.)
  - Call attention to problems and run statistics and performance monitors at the database level.
- **Help move towards the future:**
  - Get datastreams running!
  - Advise on architectural deployment issues, hardware, usage, etc.
  - Help define scenarios and scripts for replication both to Oracle and freeware, 3-tier approach, etc.
  - Not a substitute for CDF participation, but an essential complement and component of making the above work.
- **Interact with Oracle**
  - Again, not a substitute for CDF, but most logical to have dbas interact with Oracle on technical issues
  - Advise on license and other issues of product use.



## We also need ODS/DBA programmers!

- The ODS and CD programming staff has played an essential role in getting the basic database access code working
  - DBManager, CalibrationManager API, and low-level routines for connection and disconnection control code.
  - Most CDF physicists (even the careful ones) don't have a clue how this code works
  - Changes to the functionality of this code under different use scenarios has been a major source of operational difficulties we have seen with the database, only curable by changes from the original authors
  - Not just programming support from Dennis and Yuyi, but also design support from Jim and others (with input from CDF) has been required to keep things going.
- Any attempt to do this in the future with less than about 3 FTE people is probably doomed to failure
  - I remind you that we have lost 3 database coordinators within the past 2 years over this issue.



# Conclusions

---

- Overall we are right now in pretty good shape:
  - Replica is up and working (4x the power of fcdfora1 at least!)
  - Usage patterns are beginning to be understood
    - New tool to study these has great power for application both now (results already seen) and in the future
  - We have some room to grow
  - DB design / review process working
- We have got to move further, though:
  - Continue to press on our most important problems
  - We have fewer people than before, so have to work efficiently
  - DB replication and architecture need to be studied, understood, and tuned
  - Space report needs to be updated and completed
  - Several serious design issues need to be addressed
  - Need to recruit more people to be able to take on advanced projects such as freeware &/or the grid.
- Thanks to the ODS/DBA group for your support!
  - Don't relax yet, though -- we have lots more to do!!



## Schedule for CDFDB Design meetings Oct. - Dec. 2002

(Meetings Thursdays 3:00 - 5:00 pm FCC2A unless otherwise announced.)

- Oct. 3 Survey of present projects (completed).
- Oct. 17 Report of connection code review.  
DB broker and metering design criteria.  
Summary of "taking stock" meeting.
- Oct. 31 Slow controls overview and schema review.
- Nov. 14 Calibration code, codegen discussion.
- Dec. 5 Replication review: full remote copies vs. partial  
copies, copy-on-demand, or multiple tier.
- Dec. 19 Online system needs (hardware database, run  
configurations database, L3 trigger, etc.).

### Topics that can be brought up anytime:

- Need for emergency code repair or maintenance
- Project prioritization
- Use cases related to the above projects
- Ideas for schema improvements

It is assumed that people are working on various projects throughout this period. Communication of ideas and use cases through the [cdfdb-design@fnal.gov](mailto:cdfdb-design@fnal.gov) list is encouraged. We will most likely begin each meeting with a brief review of overall usage as summarized with a variety of tools, but operational issues in general will be reserved for the Wednesday 11 am database operational meetings.