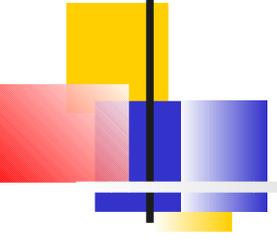


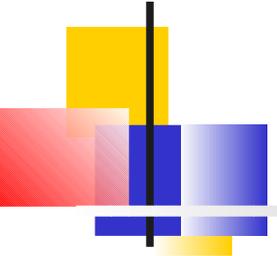
# Metadata for the Common Physicist



*Rick St. Denis, University of Glasgow  
Wyatt Merritt, Julie Trumbo, Fermilab*

---

- Goals of the Presentation
- Use Cases
- SAM in light of use cases
- SAM from 1 to 2, 2 to N – D0, CDF, MINOS, CMS
- Lessons from CDF merger
- Conclusions



# Goals

---

- Introduce: SAM Team, Metadata Working Group
- Describe the Many Faces of Metadata
- Examine metadata HEP Use Cases
- Greater understanding: Benefits of multiple experiment usage (sample)
- What SAM is and the SAM Schema
- Commonality with LHC expressed through use cases
- Support structure for migration: it can be done
- Keyword/Value pairs as a first step in common



# The SAM-Grid Team and the Metadata Working Group



**SAMGrid Project Co-Leaders:** *Wyatt Merritt, Rick St. Denis*

**SAMGrid Technical Co-Leaders:** Rob Kennedy, *Sinisa Veseli*

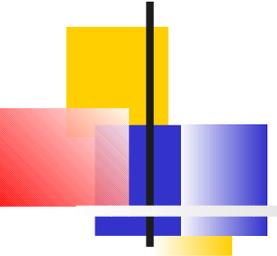
**SAMGrid Core Developers:** *Lauri Loebel Carpenter*, Andrew Baranovski, Steve White, Carmenita Moore\*, *Adam Lyon*, Petr Vokac\*\*\*, Mariano Zimmer\*\*\*, Matt Leslie, Lee Lueking\*\*, Igor Terekhov\*\*, Gabriele Garzoglio, Sankalp Jain\*\*, Aditya Nishandar\*\*

**Support for CDF Migration:** Fedor Ratnikov, *Randolph J. Herber*, Art Kreymer, Valeria Bartsch, Stefan Stonjek, Krzysztof Genser, Fedor Ratnikov, Alan Sill, Stefano Belforte, Ulrich Kerzel, Robert Illingworth

**Database support:** Anil Kumar, *Julie Trumbo*

**Metadata Working Group:** *Tony Doyle*, *Carmin Cioffi*, *Steven Hanlon*, *Caitriana Nicholson*, *Gavin Mccance*, *Solveig Albrand*, *Paul Millar*, *Tim Barrass*, *Morag Burgon-Lyon*

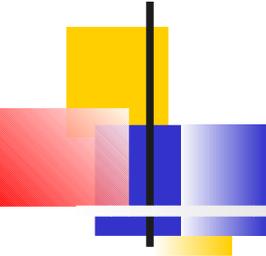
\* Deceased | \*\* Left project | \*\*\* Summer Students



# Outline

---

- Goals of the Presentation
- **Use Cases**
- SAM in light of use cases
- SAM from 1 to 2, 2 to N – D0, CDF, MINOS, CMS
- Lessons from CDF merger
- Conclusions



# Use Cases Summary: HEPCAL, CDF, BABAR, ATLAS

---

## 3 Categories



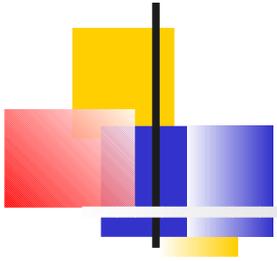
Analysis



Job Handling



Dataset  
Handling



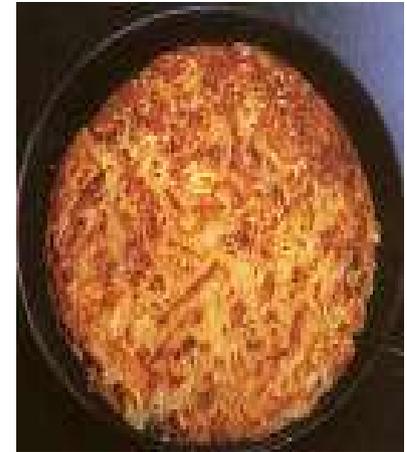
# Analysis

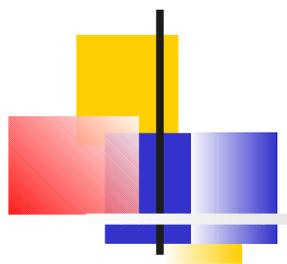
Run a physics  
simulation  
Select a subset of data

Run an algorithm over an input dataset



Ask for File  
Analyze File  
Output File





# Job Handling



**Estimate the system resource cost**

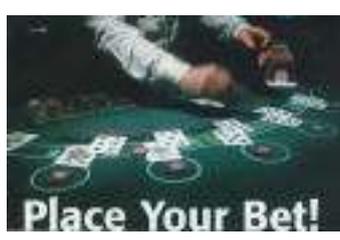


**Monitor the progress of a job**

**Retrieve/Access the output of a job**



**Submit a job to a Grid**



**Repeat a previous job**

**Recover failures in a previous job**



**... with predefined metadata**

# Dataset Handling I

Read metadata for datasets



Update and/or Add metadata for datasets



Resolve physical data



Download a dataset to a local disk



Specify a new dataset



Access a Dataset



Predefine metadata for output dataset

クレープ

新しいパンケーキに改良されたグルメのクレープは「超絶絶」です。

フレッシュレモン	7.25
サワークリーム	7.25
フレッシュバナナ	7.25
ストロベリー	7.25
ブルーベリー	7.25
サワークリームレモン	8.00
サワークリームバナナ	8.00
サワークリームブルーベリー	8.00
サワークリームストロベリー	8.00

Write experiment-specific metadata for the new dataset



# Dataset Handling II

Read all the visible metadata for a specified dataset



Nutrition Facts	
Amount Per Serving	
100g	
Total Fat	10g
Sodium	100mg
Total Carbohydrate	10g
Protein	10g
Dietary Fiber	10g
Sugars	10g
Cholesterol	10mg
Vitamin A	10%
Vitamin C	10%
Calcium	10%
Iron	10%

Merge dataset



Publish a private dataset



"Surely you were aware when you accepted the position, Professor, that it was publish or perish."

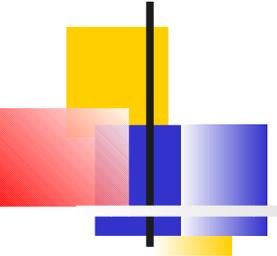
Perform a transform on a dataset

Search for datasets whose metadata match a user query

print



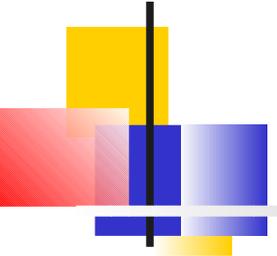
Publish private metadata



# Outline

---

- Goals of the Presentation
- Use Cases
- **SAM in light of use cases**
- SAM from 1 to 2, 2 to N – D0, CDF, MINOS, CMS
- Lessons from CDF merger
- Conclusions

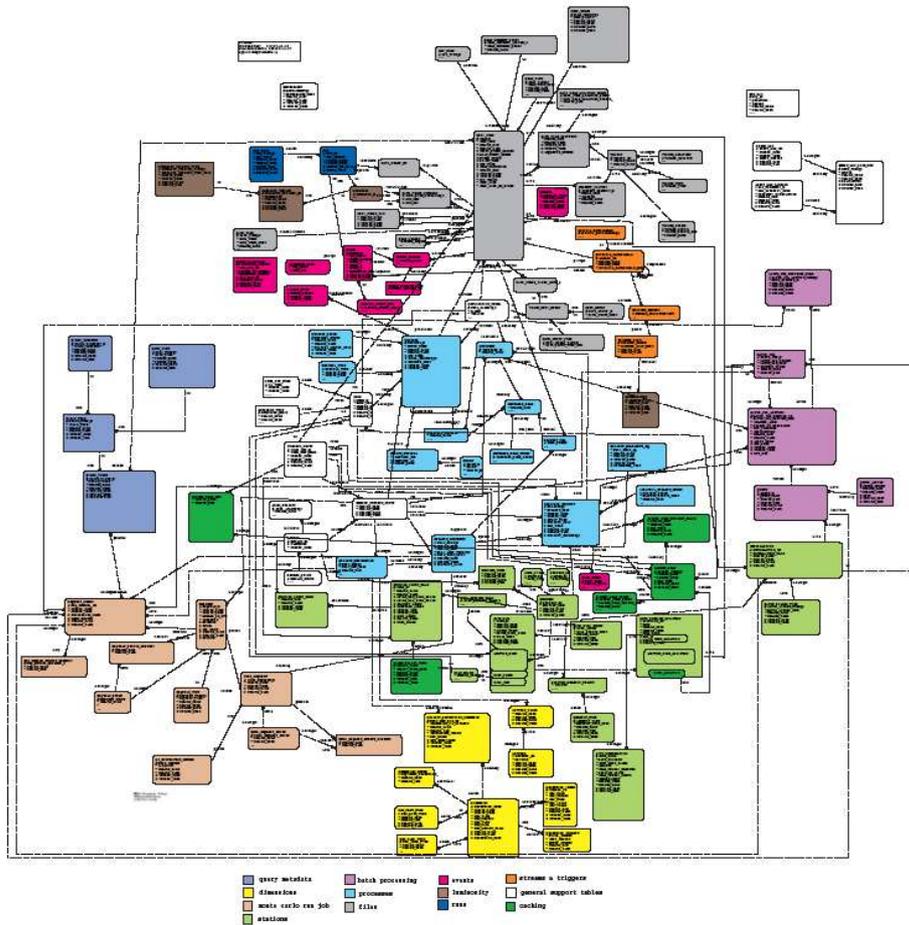


# The SAM Paradigm

---

- A **project** runs on a **station** and requests delivery of a **dataset** to one or more **consumer processes** associated with that station. Consumers perform a transformation on the dataset and output files to store with metadata. Services control optimal delivery and storage.
- File delivery is stateful and a permanent record of data handling is kept for a project.

# Implemented on Relational Database



- DØ, CDF, and MINOS use the same DB Schema
- Relational
  - Matches metadata
- Monolithic
  - Efficient (>360 File/min)
- Flexible
  - Schema updateable *in a controlled fashion*

# File Metadata

- **SAM manages file storage (replica catalogs)**



- Data files are stored in tape systems at FNAL and elsewhere around the world for fast access

- **SAM manages file meta-data cataloging**

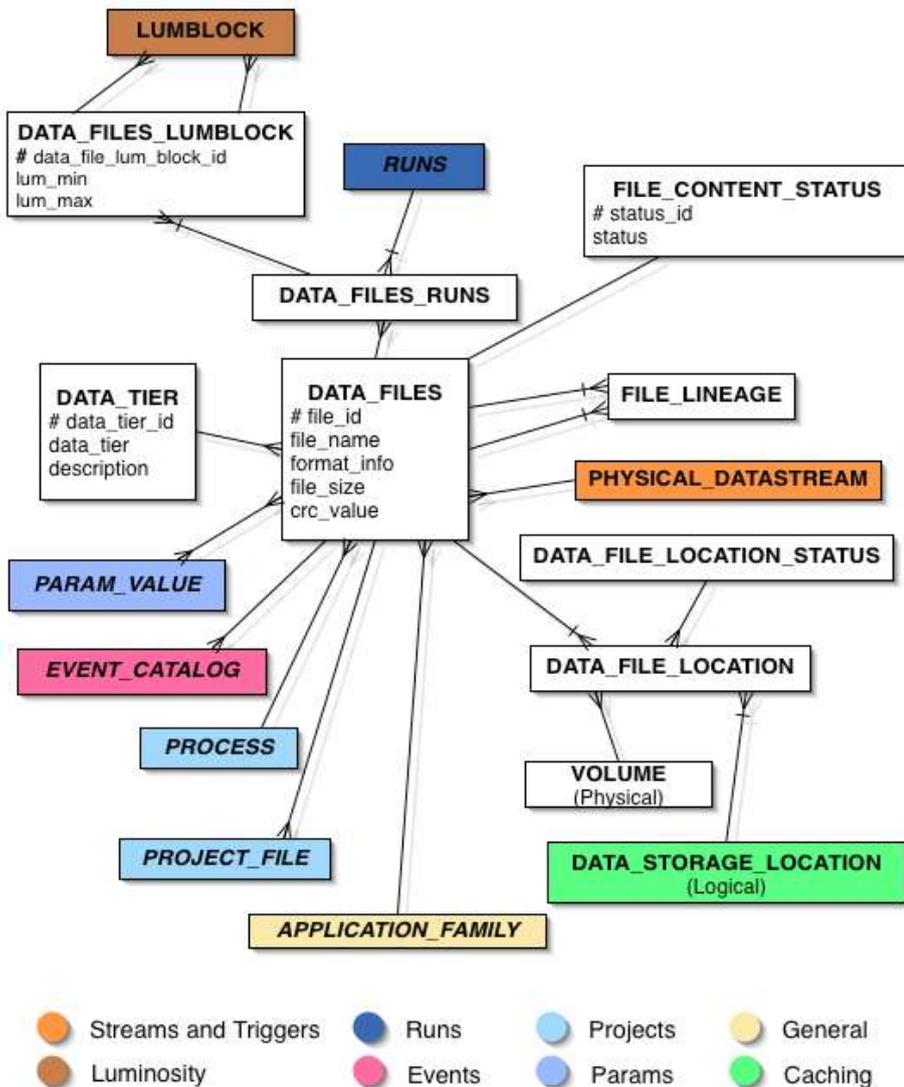
- SAM DB holds meta-data for each file.



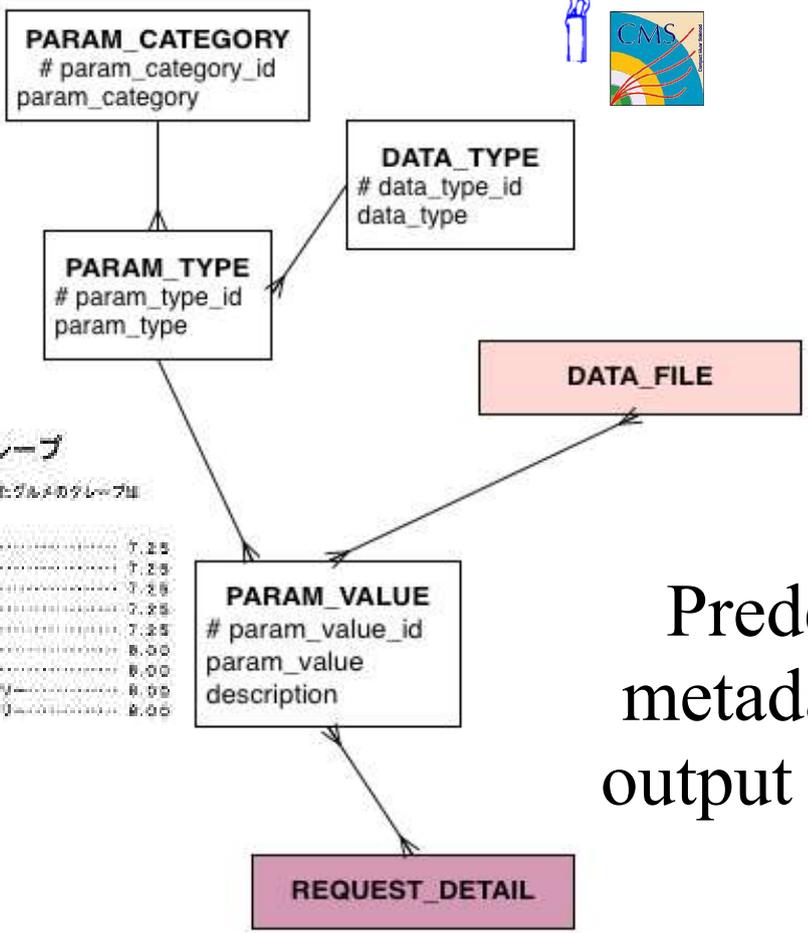
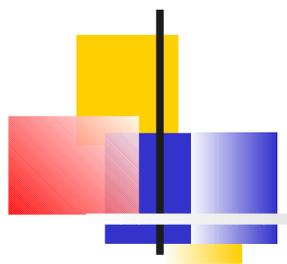
# Data Files Metadata



- Data Files: The heart of SamGrid
- Fixed metadata
  - File name, size, crc
  - Production group
  - Data Tier (Raw, Reconstructed ...)
  - Application, Locations
  - Detector, Runs, Event info
  - Project/Process, Luminosity
  - Stream/Trigger
- Connection to free metadata (Params) ...



# Params (Free file metadata): A common element with ATLAS, LHCb



クレーブ  
同じバインディングに定義されたグルムのクレーブは  
"格納" 可能です。  
フレクチュレーション ..... 7.25  
デタークリム ..... 7.25  
フレクチュレーション ..... 7.25  
ストロベリー ..... 7.25  
ブルベリー ..... 7.25  
サウークリム ..... 8.00  
サウークリム ..... 8.00  
サウークリム ..... 8.00  
サウークリム ..... 8.00

Predefine  
metadata for  
output dataset

- Fixed metadata allows easy and performant querying
- Free metadata for application specific items
  - Categories group parameters (pythia, isajet, ...)
  - Types are the keywords (decayfile, topmass, ...)
  - Values
  - Queries are more difficult

● Data Files    ● MC Requests



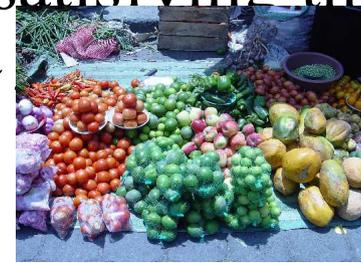
"Thank you very much when you assigned the position, Professor. It was just as I needed it."

# Metadata Definitions

- **SAM manages definitions of datasets based on metadata**

- SAM DB stores definitions based on metadata by group and user. These are resolved to lists of files satisfying those definitions when a user chooses to run a

- “data\_type physics and run\_number 78904”



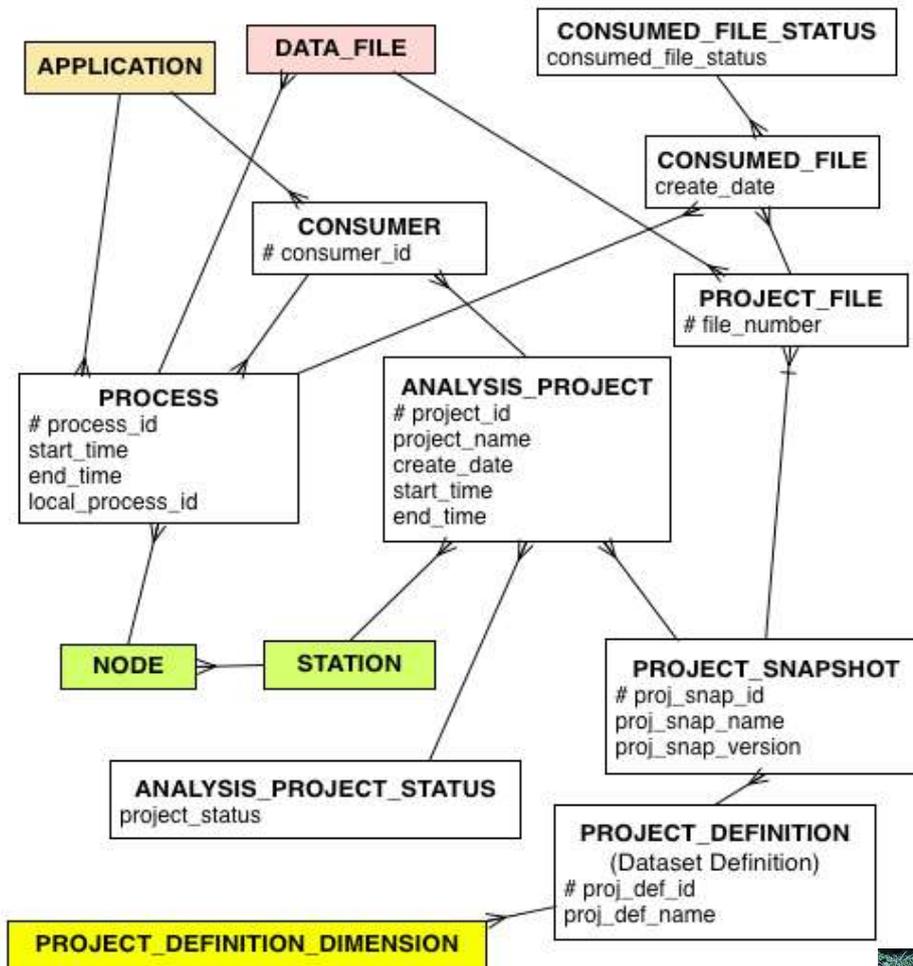
- **SAM manages analysis bookkeeping**

- SAM remembers what files you ran over, what files you processed successfully, what applications you ran, when you ran them and where. Hence it is possible to recover from errors and repeat runs.





# Project Metadata



- General
- Data Files
- Dimensions
- Station

- Projects run by a user in a group on a dataset **Snapshot** with nodes from a SAMGrid station
- A Project has one or more **Consumers** (usually one)
- A Consumer has one or more **Processes**
- A Process is a job on a node. Keeps track of consumed files

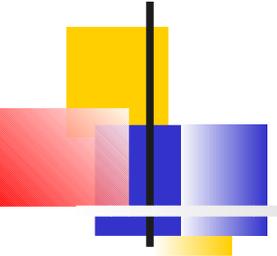


# File Delivery

- SAM manages file delivery by dataset
  - Users at FNAL and remote sites retrieve files out of file storage. SAM handles caching or can interface to other cache systems (See Rob Kennedy's Talk)
  - You don't care about file locations







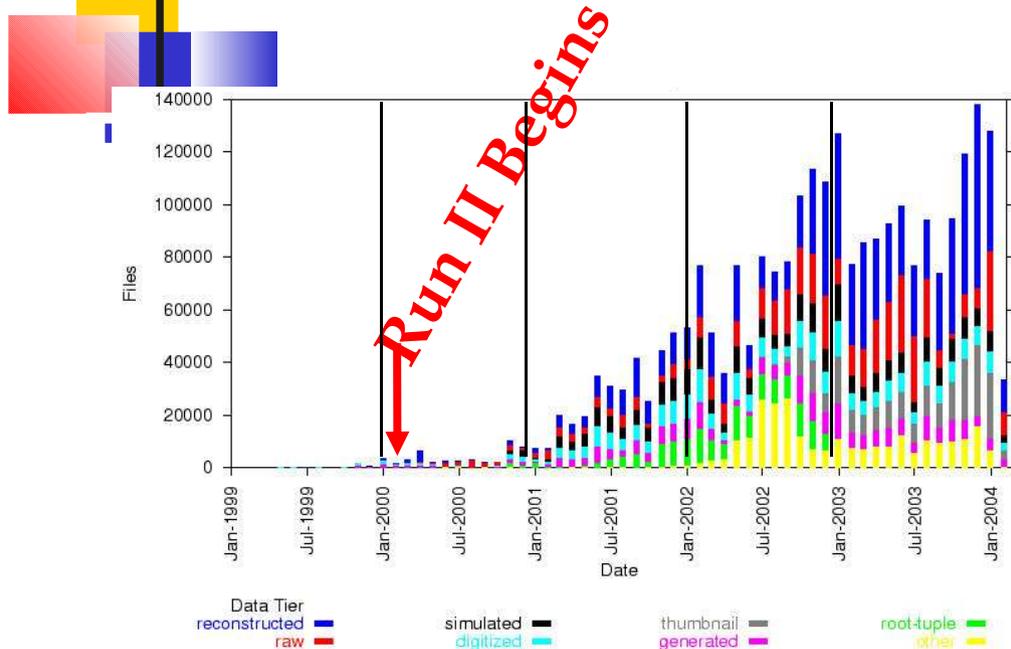
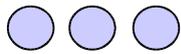
# Outline

---

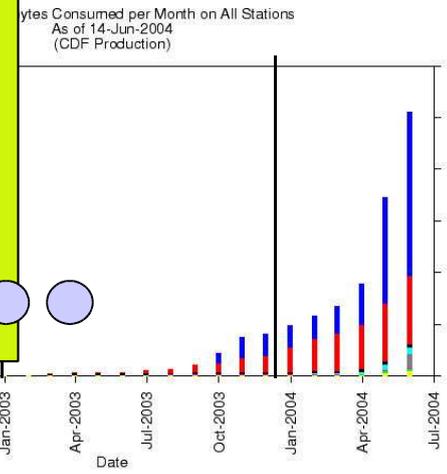
- Goals of the Presentation
- Use Cases
- SAM in light of use cases
- **SAM from 1 to 2, 2 to N – D0, CDF, MINOS, CMS**
- Lessons from CDF merger
- Conclusions

# SAM: from One Experiment: DØ

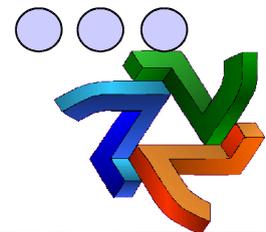
40 active sites



To a second experiment: CDF



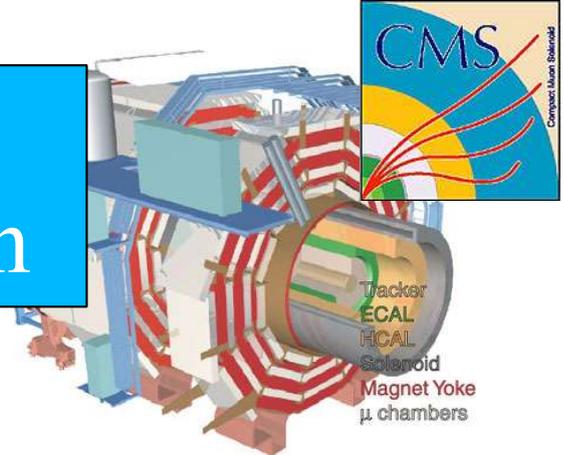
To MINOS



25 active sites

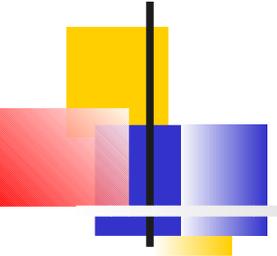


CMS Evaluation



2 sites @fnal

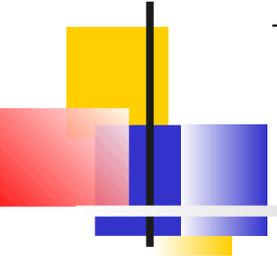




# Outline

---

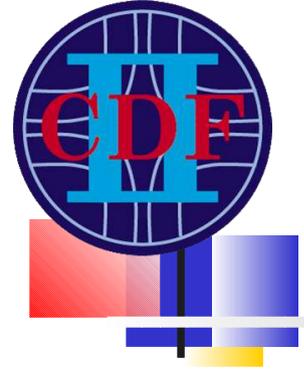
- Goals of the Presentation
- Use Cases
- SAM in light of use cases
- SAM from 1 to 2, 2 to N – D0, CDF, MINOS, CMS
- **Lessons from CDF merger**
- Conclusions



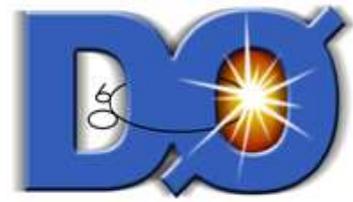
# First:DBA Standards that made CDF adoption of SAM feasible

---

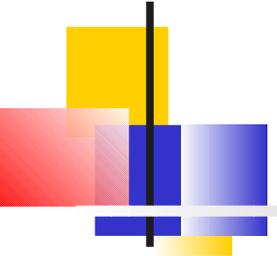
- Centralized Oracle Database at FNAL
- Three tier system ensures DB integrity
  - Development - Newest schema with artificial or special data. Used for testing
  - Integration – Dress rehearsal for modifying schema using a copy production data upon which a test harness is run.
  - Production - The real thing



# Overview of Impact of CDF Involvement



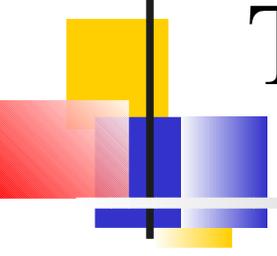
- CDF participation provided opportunity for revisiting of the original D0 Design including D0 experience derived from use in different phases: MC, commissioning, stable running.
- An entirely new user community provided the trigger for a second generation design, the need for which was recognized by the original users.
- Boundries became more clearly defined and natural separation into services occurred.



# Important Features of Schema Change

---

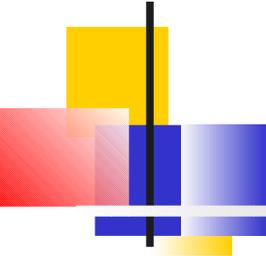
- Many runs in a file; separate luminosity bookkeeping
- Clean separation of file types: Generic, MC, processed
- Keep track of group responsible for file
- Require at DB Level: format, size, crc type/value, file content status id
- Not Required at DB Level: data tier, file partition, process id, stream, event count, first/last event number start/end times
- Removed: MC - min bias no. & type, physics process



# Three Examples: Deeper Implications

---

- Process ID:
  - Change in Paradigm
- Separate Luminosity bookkeeping:
  - Illustration of how to link different database schemas
- File Type:
  - Change in location of business rules



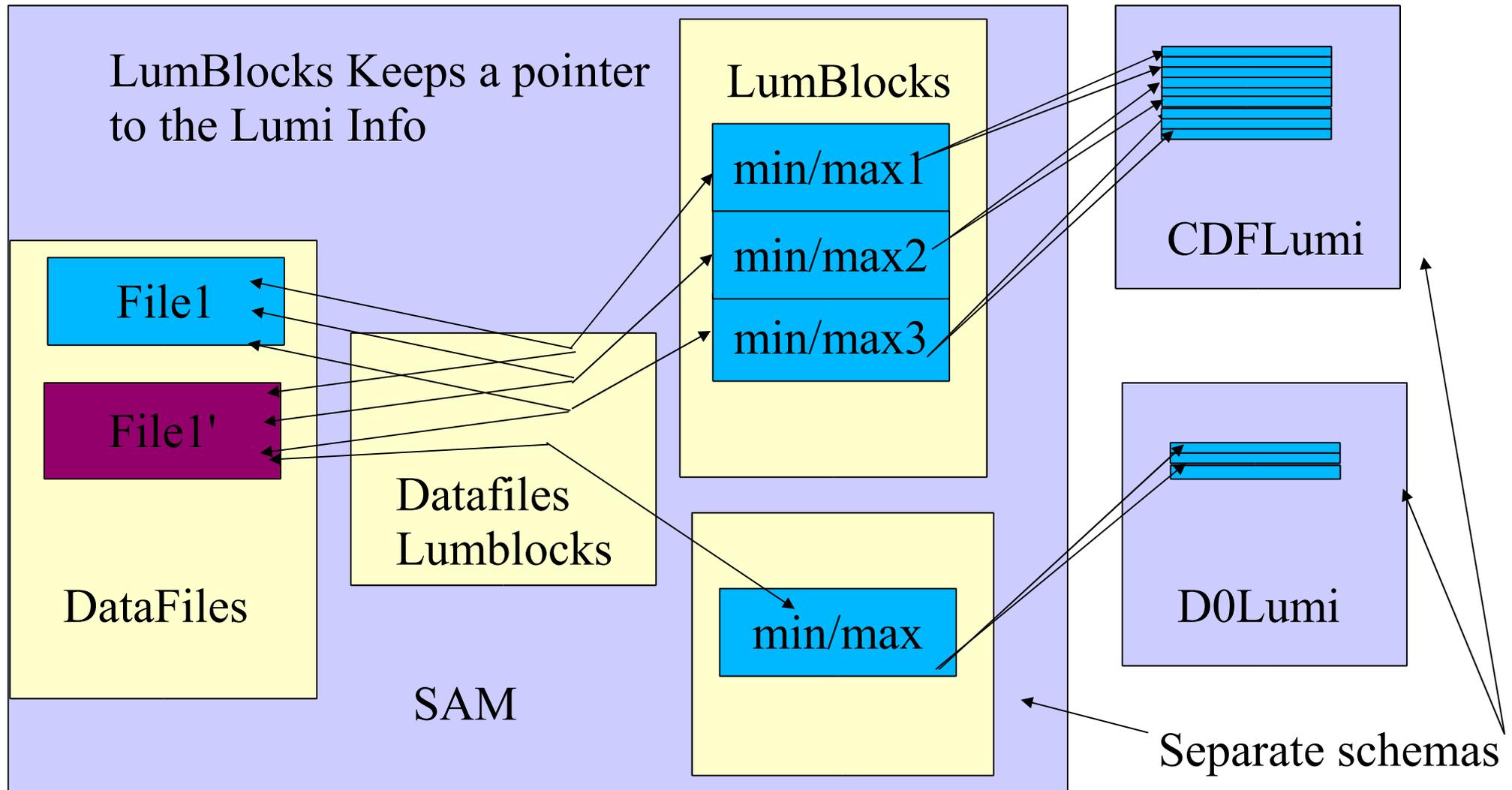
# Process ID

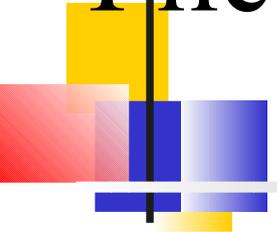
---

- Sam Assumes
  - A **process** produces a **file**.
  - You ALWAYS want a process for a file
  - Therefore ProcessID is required
- Reality says
  - Sometimes files are imported from users not running with SAM to get input and keep track of files

**The Process ID cannot be required**

# Linking Schema: Luminosity Bookkeeping

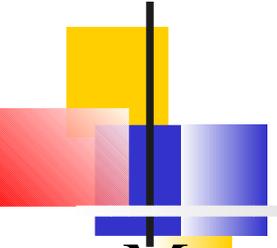




# File Type: Change of location of business rules - Implement Rules in API

---

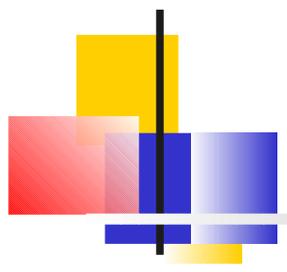
- physicsGeneric
  - **Must have:** Data tier is *unofficial reco* (D0)
- NonPhysicsGeneric
  - **Must have:** File status of *being imported* or *deleted* (CDF)
- Imported detector
  - **Must have:** File status of *available* with Data tier of *raw* and 17 characters.



# Conclusions

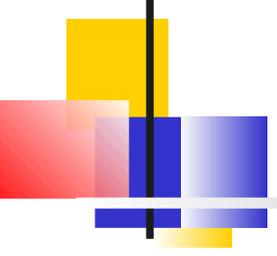
---

- Metadata: Workflow Processing, File/physics, Authorization, Quota
- Greater understanding: Experimental Lifecycle maturation, need for sharp boundaries, natural demarcation of services when experiments join: benefits to both.
- SAM is a system of data handling and work flow services described by metadata modelled on a relational database
- SAM implements the HEPCAL Metadata use cases.
- Migration of schema with running experiments is inevitable and can be accomplished
- Detailed schema and API implementations can be shared across HEP experiments.



---

## Extra Slides

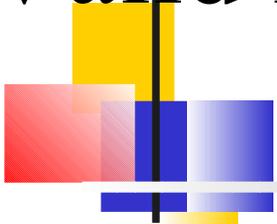


# Interfacing

---

- Interfaces:
  - Batch system interaction
  - Experiment-specific metadata
  - Storage and use of external caching

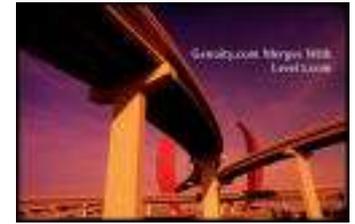
# Valid Data Groups: Workflow-Data handling interaction



Ask for File  
Analyze File  
Output File



Merge dataset



- Workflow Step Transition
- File operations atomic
- Metadata for workflow
- Born of CDF/D0 Joint Effort

Perform a transform on a dataset

## To Be Processed

InFile 1

InFile 2

InFile 3

InFile 4



InFile N

## Processed

InFile 1

InFile 2

InFile 3

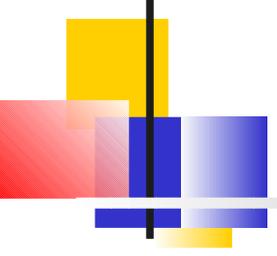


OutFile 1

OutFile 2



OutFile M



# Sam in Operation

---

- Looking at SAM in operation -

SAM TV @ DØ      SAM TV @ CDF

- Currently created from log files
- Version in development is created from MIS database, filled by new MIS server

# CDF SAM Deployment

## SAM Stations:

### Monitor Level: Critical



[cdf-sam](#)

Host	Version	Up Since
fcdfdata016.fnal.gov	v4_2_1_69	19 Jul 2004 18:34:19

### Monitor Level: High



[cdf-cnaf](#)



[cdf-fzkka](#)



[cdf-knu](#)



[cdf-oxford](#)



[cdf-rutgers](#)



[cdf-sdsc](#)



[cdf-taiwan](#)



[cdf-toronto](#)



[cdf-trieste](#)



[cdf-ttu](#)

Host	Version	Up Since
cdfsam.cnaf.infn.it	v4_2_1_63	22 May 2004 07:19:01
cdf.fzk.de	v4_2_1_72	23 Jul 2004 10:58:27
cluster67.knu.ac.kr	v4_2_1_72	13 Jul 2004 03:38:41
matrix.physics.ox.ac.uk	v4_2_1_71	20 Jul 2004 11:34:23
hexsam.rutgers.edu	v4_2_1_63	06 Jul 2004 18:16:06
t2sam01.sdsc.edu	v4_2_1_72	22 Jul 2004 14:21:40
ascaf.sinica.edu.tw	v4_2_1_72	20 Jul 2004 09:43:33
bigmac-cdf03.physics.utoronto.ca	v4_2_1_63	14 Jun 2004 10:57:44
pccdf2.ts.infn.it	v4_2_1_63	19 Jul 2004 13:31:03
pantheon.cs.ttu.edu	v4_2_1_63	23 Jul 2004 08:58:07



[cdf-riken](#)



[cdf-ral](#)



[cdf-rdk-fnal-1](#)



[cdf-sam2](#)



[cdf-scotgrid](#)



[cdf-scotgrid-2](#)



[cdf-taiwan2](#)



[cdf-test](#)



[cdf-ttu-hpcc](#)



[cdf-ttu-phys](#)



[cdf-tufts](#)



[cdf-ucsd](#)



[samadams](#)



[sangfarm](#)

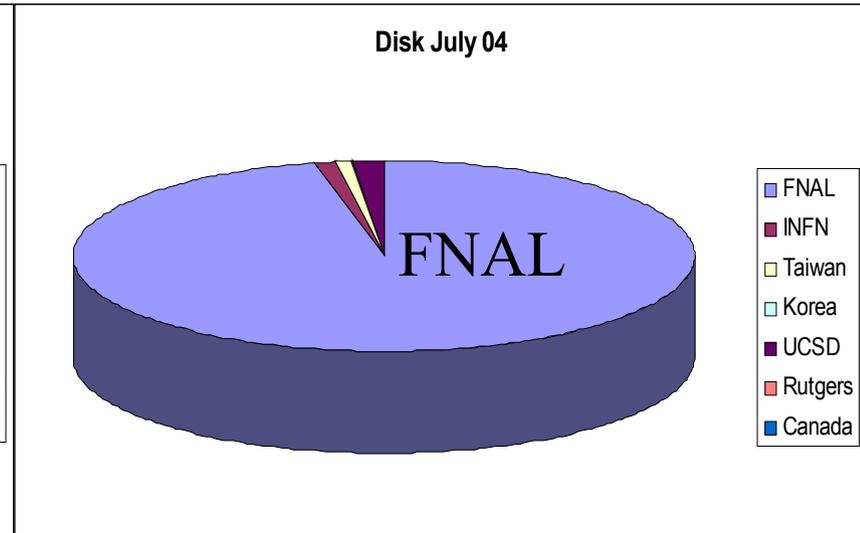
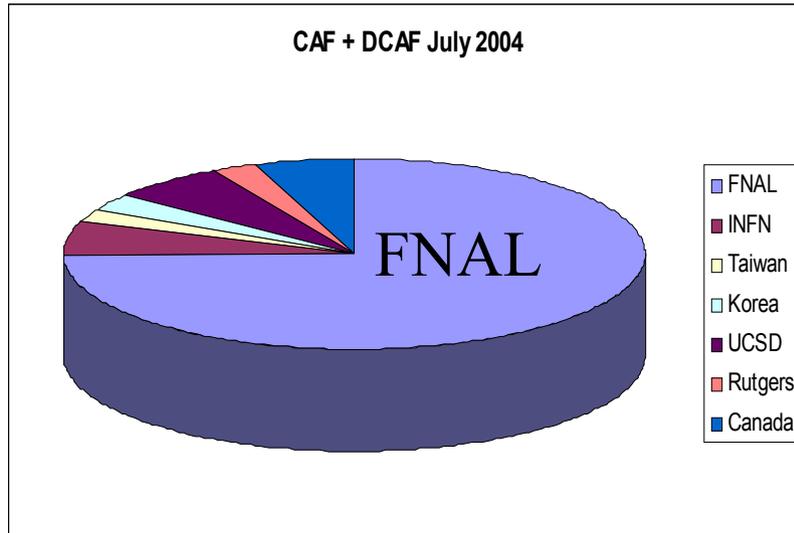
cdfa.rl.ac.uk	v4_2_1_63	08 Jul 2004 09:34:33
cdfpcb.fnal.gov	v4_2_1_72	20 Jul 2004 01:01:02
castor.phys.ttu.edu	v4_2_1_72	19 Jul 2004 22:36:10
samadams.fnal.gov	v4_2_1_63	21 Jun 2004 16:20:32
sangfarm.fnal.gov	v4_2_1_64	16 Jul 2004 17:08:40

# CPU Growth OK, Disk Growth Slower: Need network and/or use offsite for MC

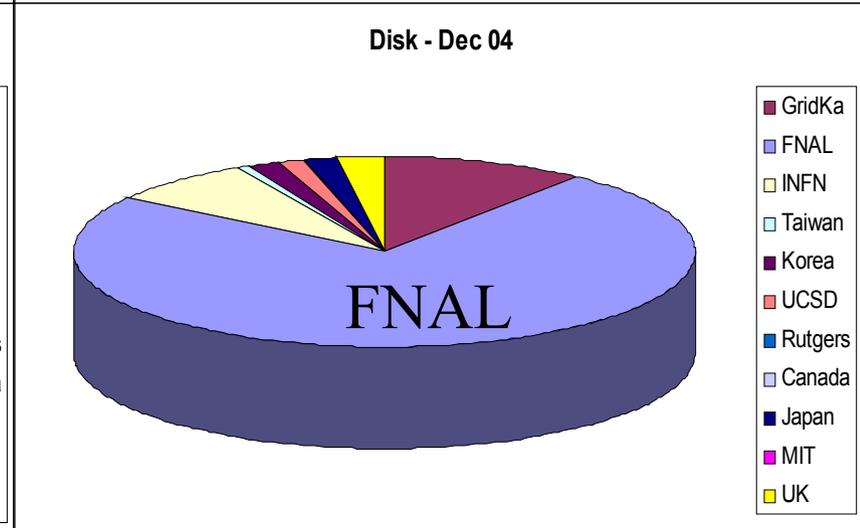
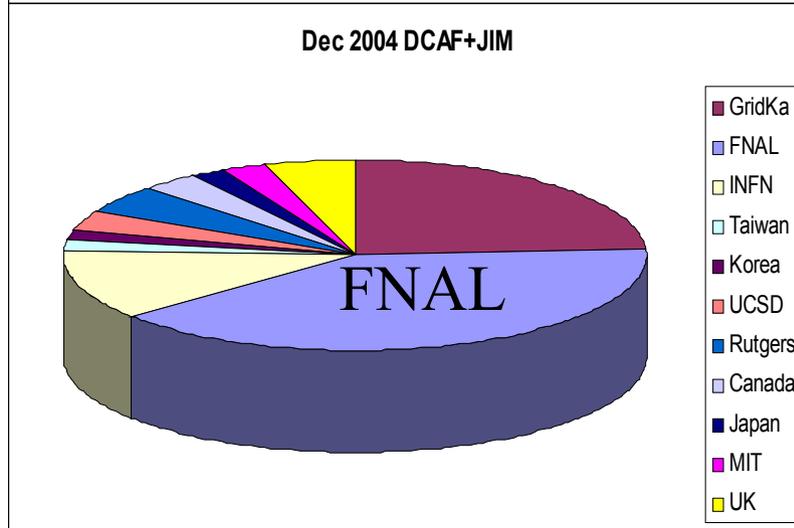
## CPU

## Disk

July  
04

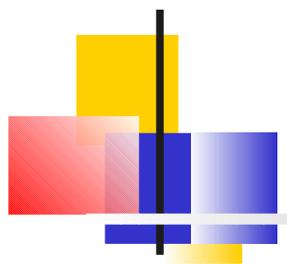


Dec  
04



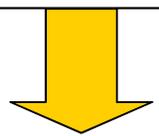
See <http://cdfkits.fnal.gov/DIST/doc/DCAF/>

# CDF Global Task Submission & Execution

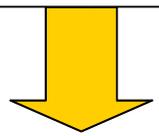


Run a physics simulation  
Select a subset of data  
Run an algorithm over an input dataset

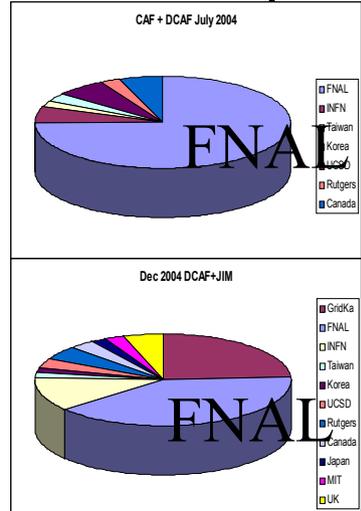
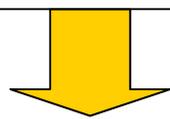
DCAF Gui/CLI



Analysis program



DCAF

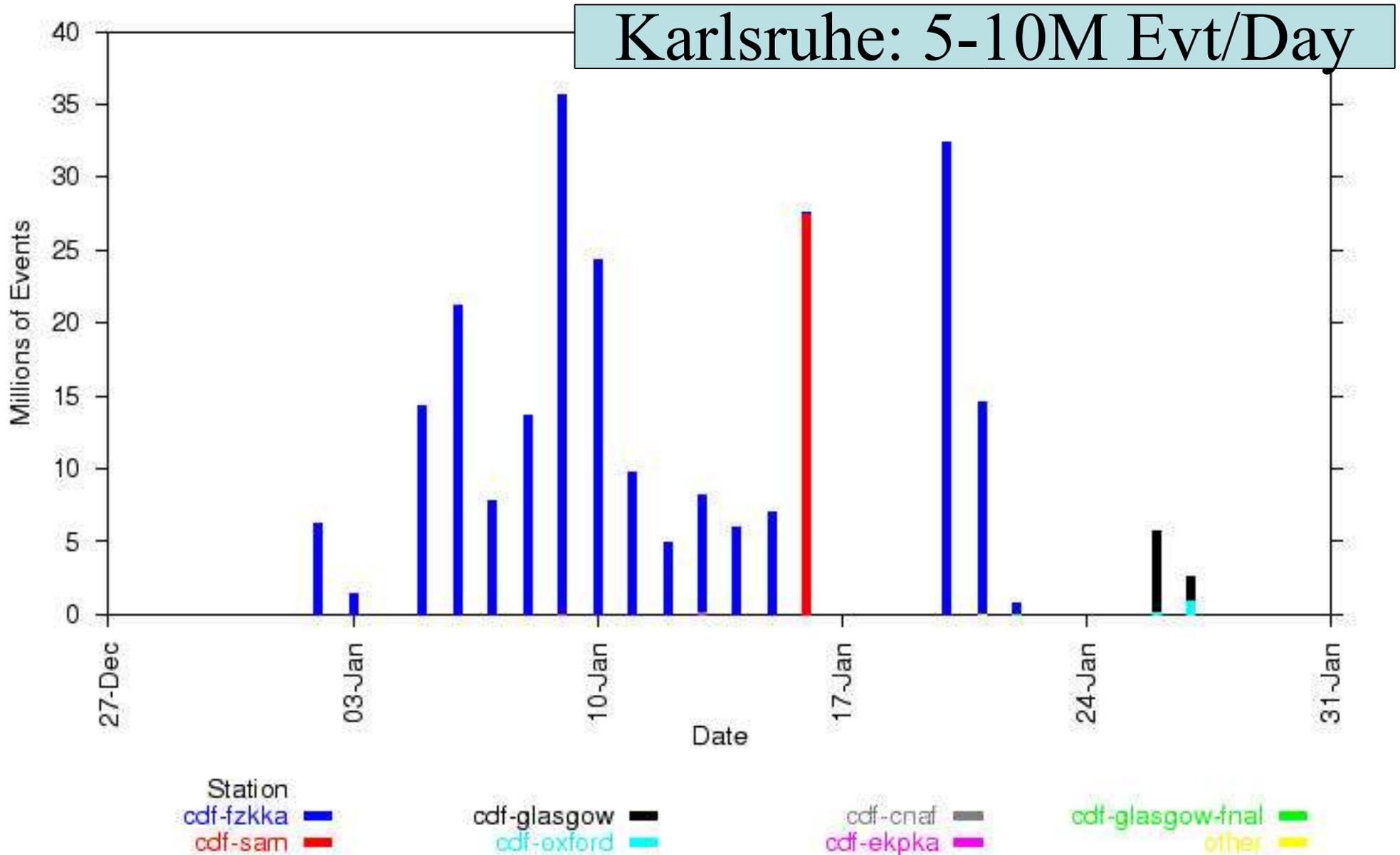


DCAF: 200GHz farm

Sam services on head node

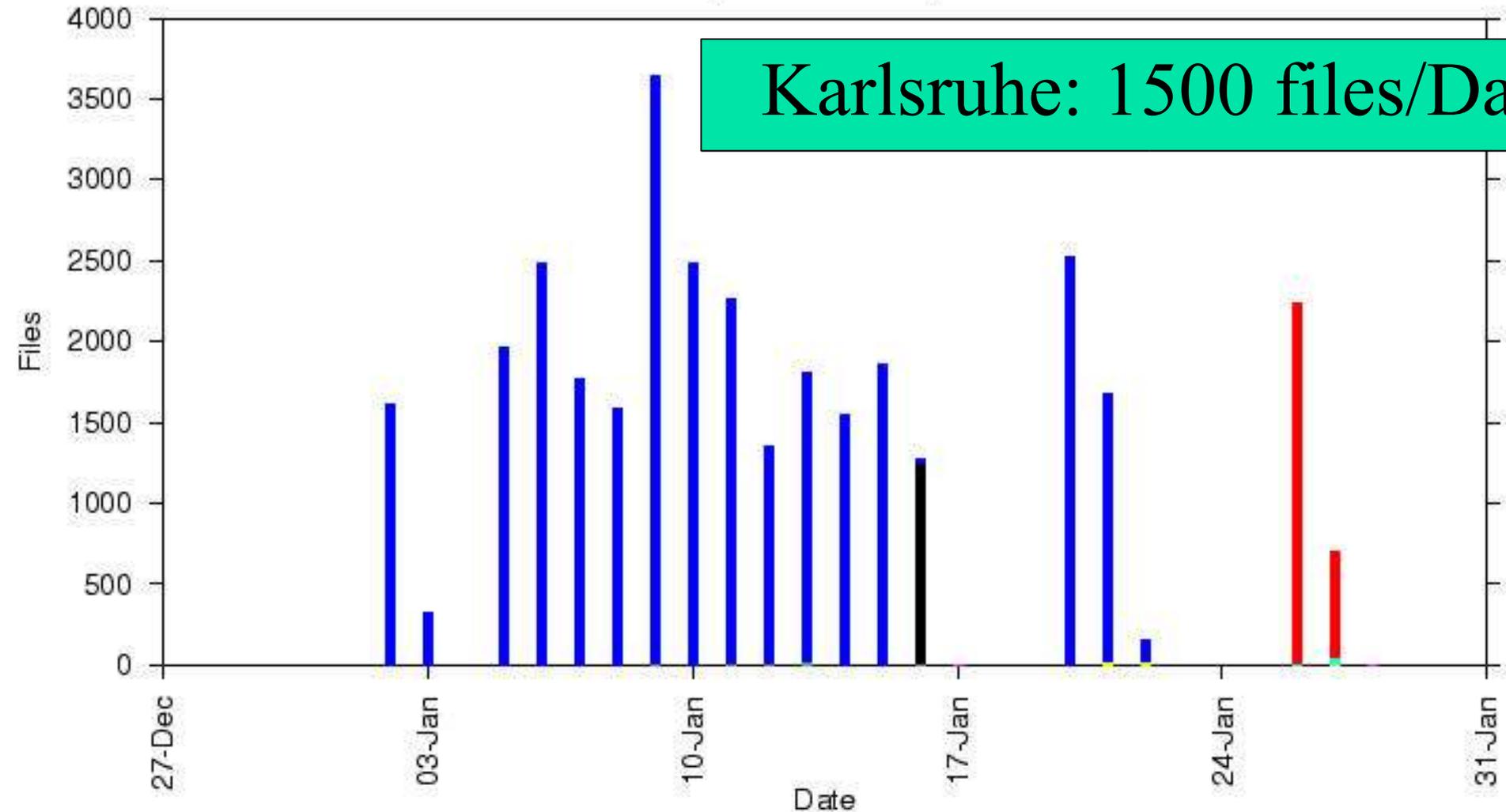


# CDF Events Transferred per Month



# CDF Files in a Month

Karlsruhe: 1500 files/Day



Station

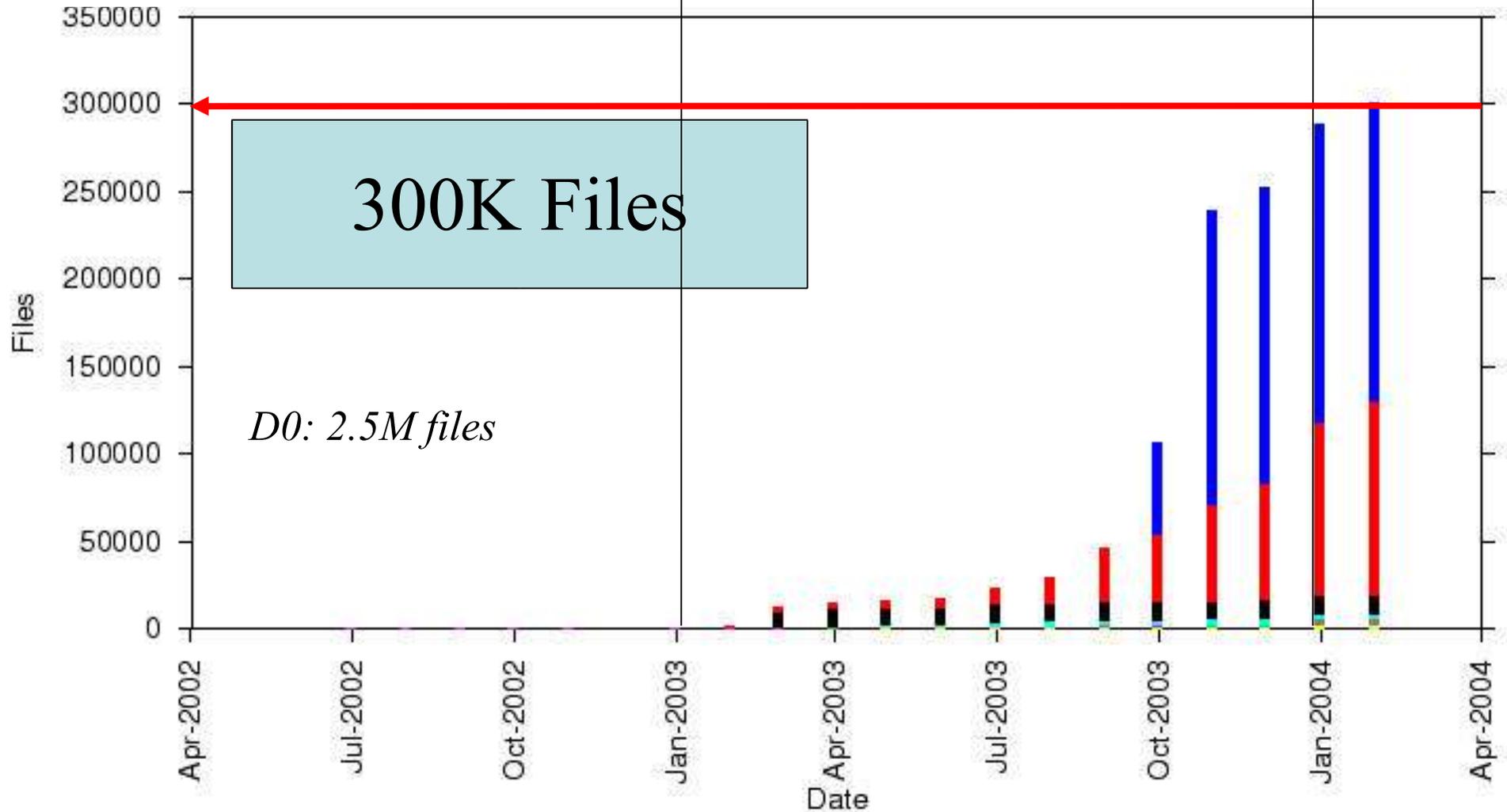
cdf-fzkka cdf-sam cdf-ekpka cdf-cnat

cdf-glasgow cdf-oxford cdf-test other

# All CDF Files Moved by SAM

2002

2003



Station

cdf-sam (blue)  
cdf-fzkka (red)

cdf-oxford (black)

cdf-ekpka (cyan)

cdf-glasgow (grey)

cdf-glasgow-fnal (magenta)

cdf-trieste (green)

other (yellow)

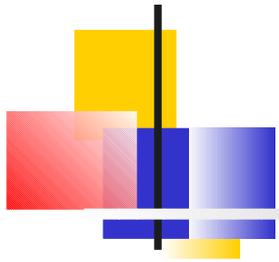
14 Sep 2004

Sam and the Grid at CDF

# Q: What is SAM?

A: Data handling system for Run II  
DØ, CDF and MINOS

---



- Distributable `sam_client` provides access to:
  - VO storage service (`sam store` command, uses `sam_cp`)
  - VO metadata service (`sam translate` constraints)
  - VO replica location service (`sam get next file`)
  - Process bookkeeping service

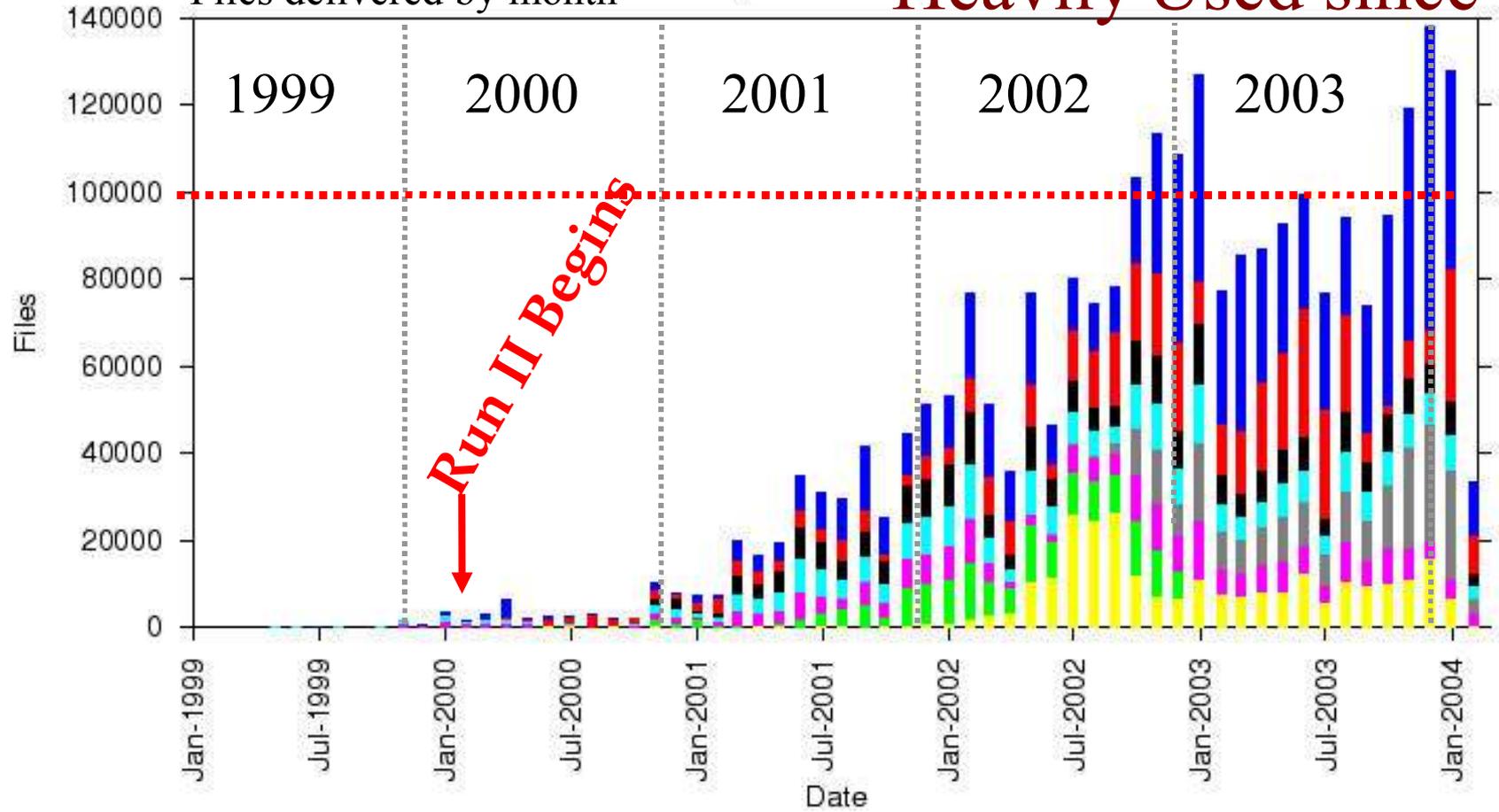
Designed for PETABYTE ( $10^{15}$ ) sized  
experiment datasets

# SAM goes from One Experiment: DØ



Files delivered by month

Heavily Used since 2002

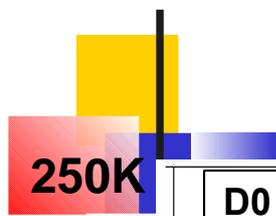


## DØ -40 active sites, 9@FNAL

# Usage Statistics for D0

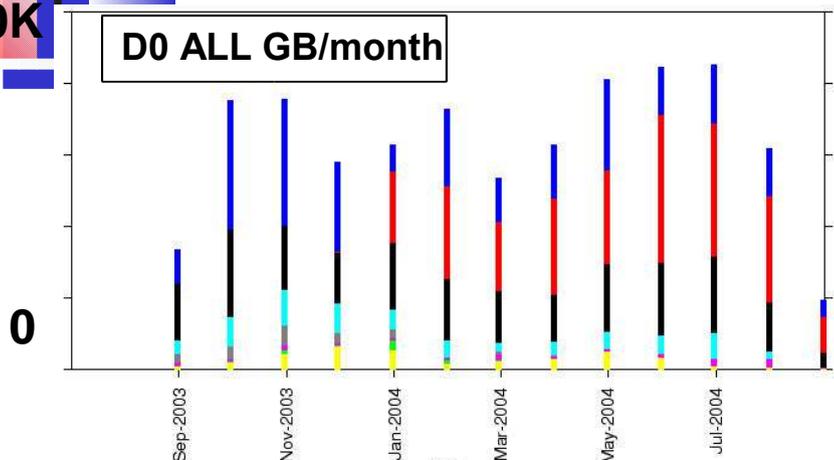
Sum = 2.1 PB; 50B evts

## SAM



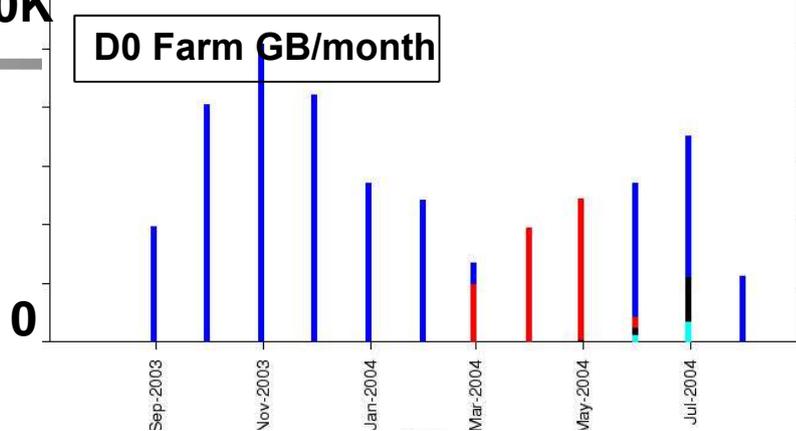
250K

D0 ALL GB/month



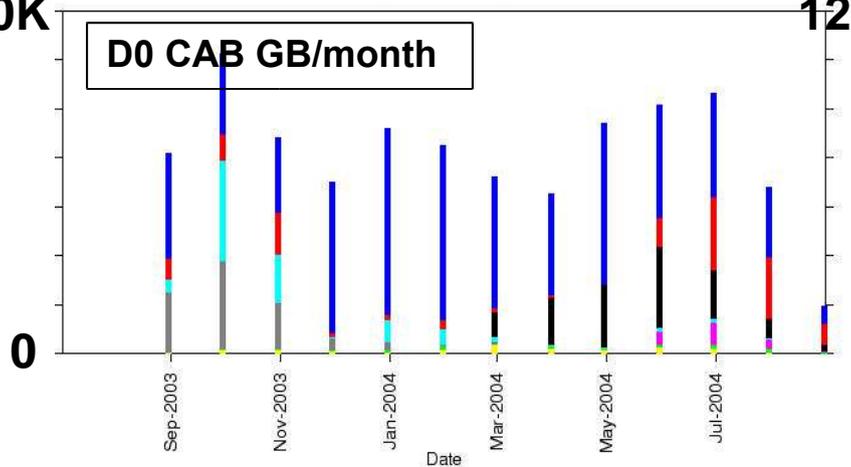
30K

D0 Farm GB/month



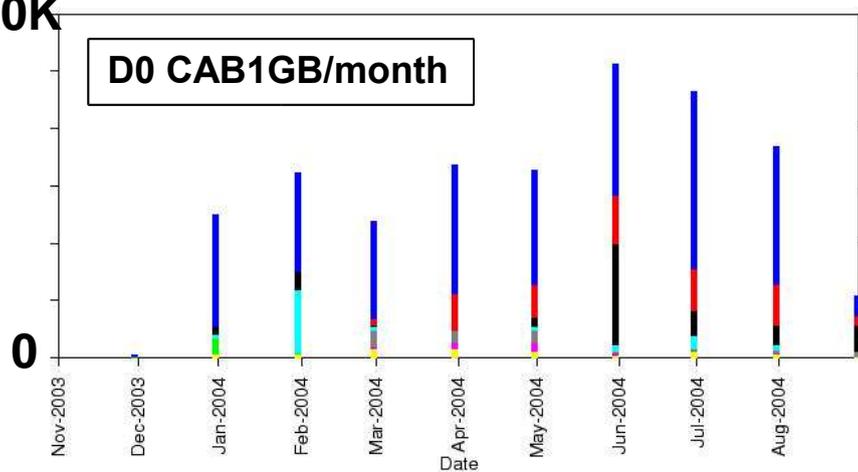
70K

D0 CAB GB/month



120K

D0 CAB1GB/month



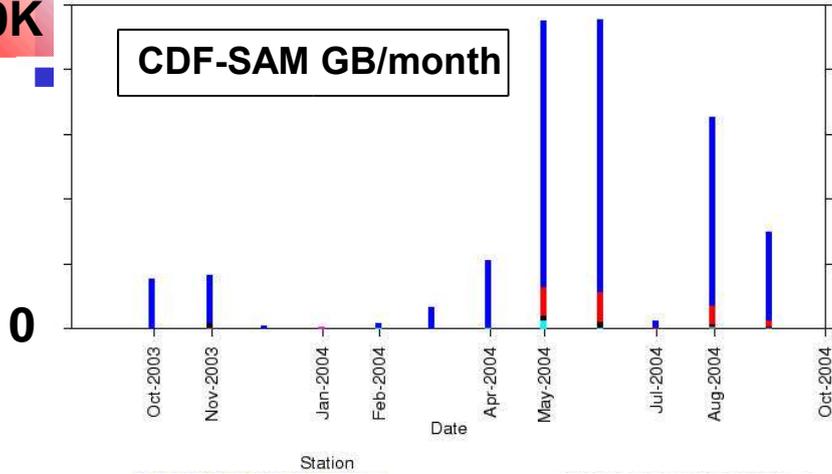
# Usage Statistics for CDF

Sum = 1.5 PB; 12B evts

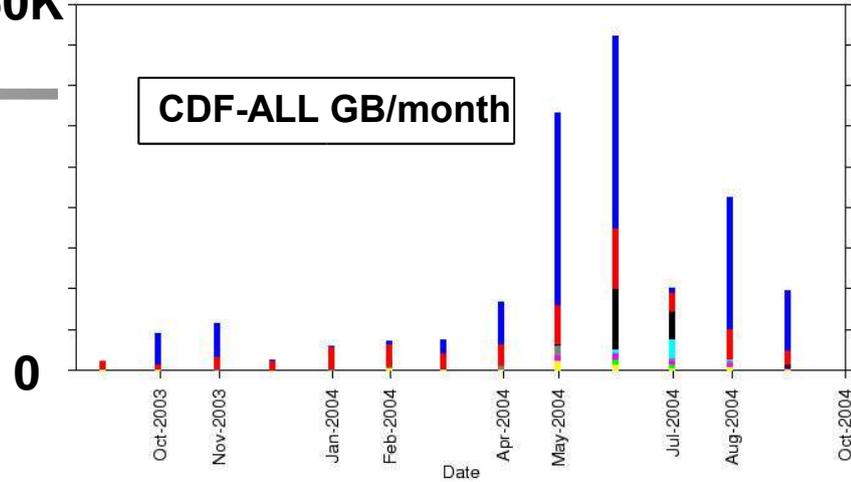
## SAM

450K

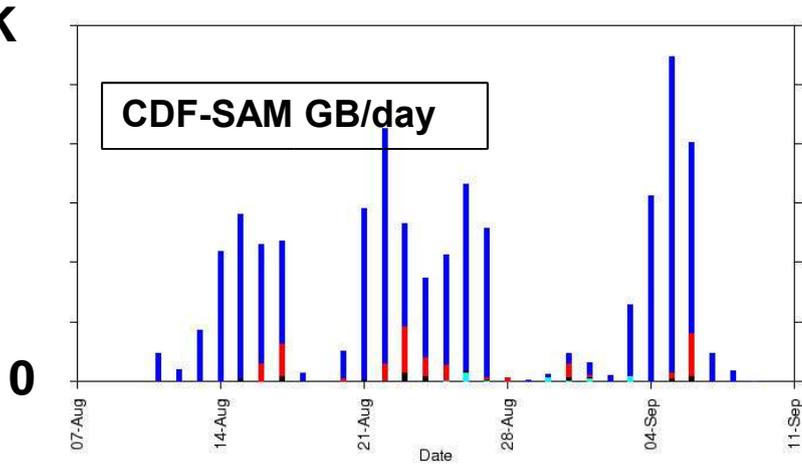
CDF-SAM GB/month



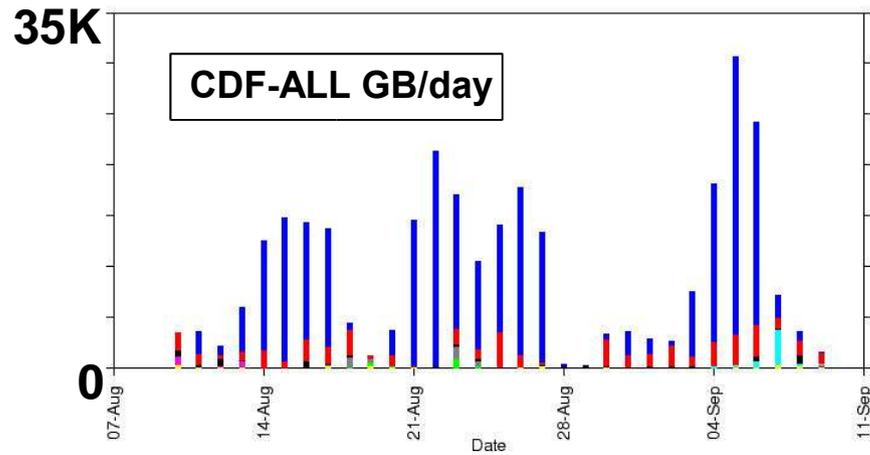
CDF-ALL GB/month



CDF-SAM GB/day



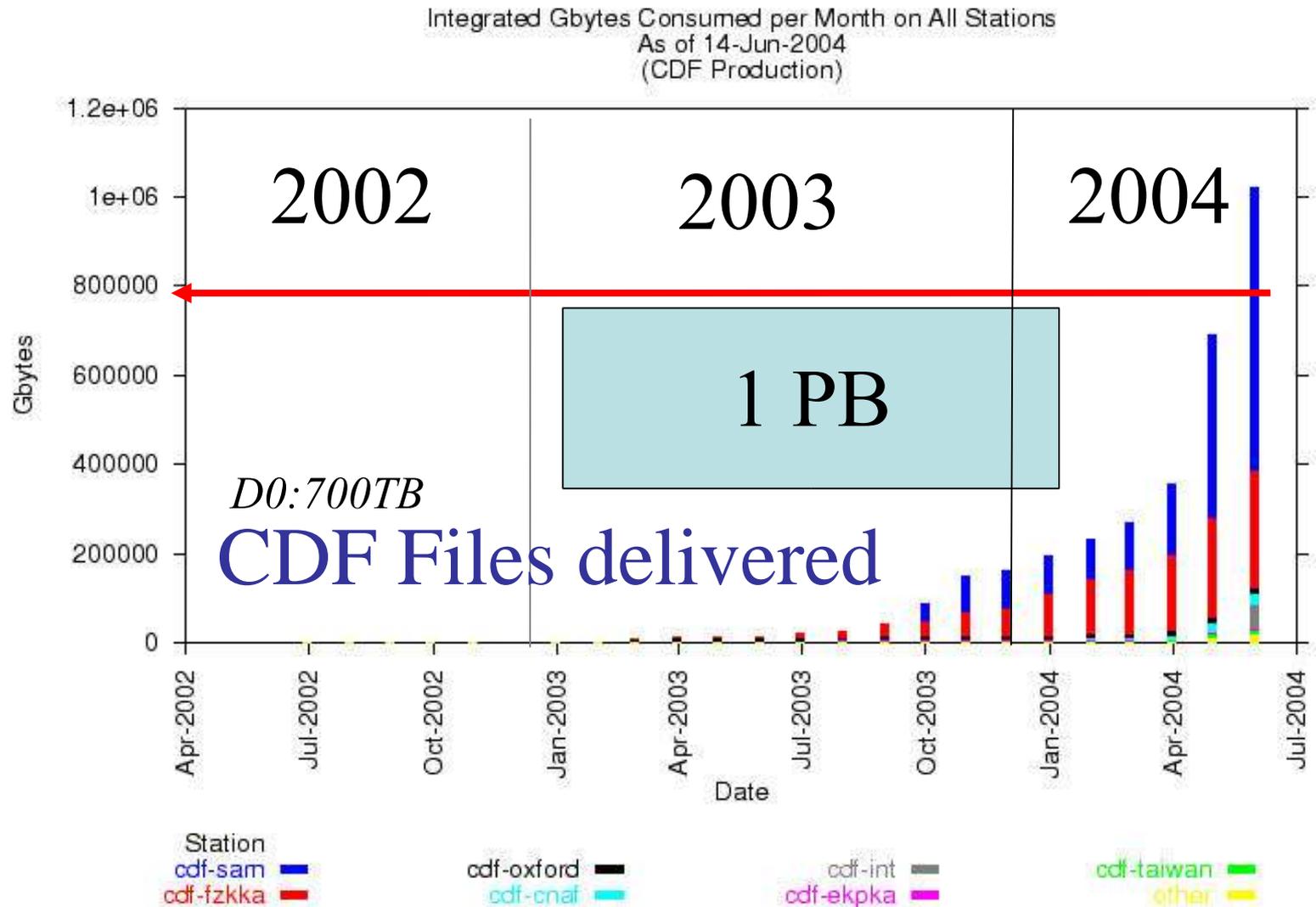
CDF-ALL GB/day



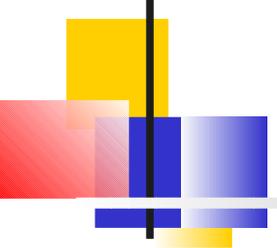
Current Resources			
Cluster Name and Home Page	Monitoring and Direct Information Links	CPU (GHz)	Disk space (TBytes)
<a href="#">Original FNAL CAF</a>	<a href="#">queues</a> , <a href="#">user history</a> , <a href="#">ganglia</a> , <a href="#">sam station</a> , <a href="#">consumption</a>	1200	200
<a href="#">FNAL CondorCAF (Fermilab)</a>	<a href="#">queues</a> , <a href="#">user history</a> , <a href="#">analyze</a> , <a href="#">ganglia</a> , <a href="#">sam station</a> , <a href="#">consumption</a>	2000	~(shared w/CAF)
<a href="#">CNAFCAF (Bologna, Italy)</a>	<a href="#">que</a>		7.5
<a href="#">KORCAF (KNI)</a>			6
<a href="#">ASCAF (Ac Sinica, Taiw)</a>			
<a href="#">SDSC Condor (San Diego)</a>			7.0
<a href="#">HEXCAF (Rutgers)</a>	<a href="#">consum</a>		4.0
<a href="#">TORCAF2 (Toronto CDF)</a>	<a href="#">queues</a> , <a href="#">ganglia</a> , <a href="#">disk status</a> , <a href="#">sam station</a> , <a href="#">datasets</a> , <a href="#">consumption</a>	576	10
<a href="#">JPCAF (Tsukuba, Japan)</a>	<a href="#">queues</a> , <a href="#">user history</a> , <a href="#">sam station</a> , <a href="#">datasets</a> , <a href="#">consumption</a>	152	5.0
<i>Current Totals:</i>		5012	234

1.8 of 5.0 THz is now  
offsite

# To a second experiment: CDF ○○○



Sam Deployed Later to CDF: 25 active sites (2 @ FNAL)

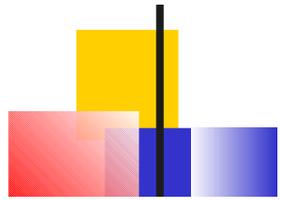


# SAM Terms

---

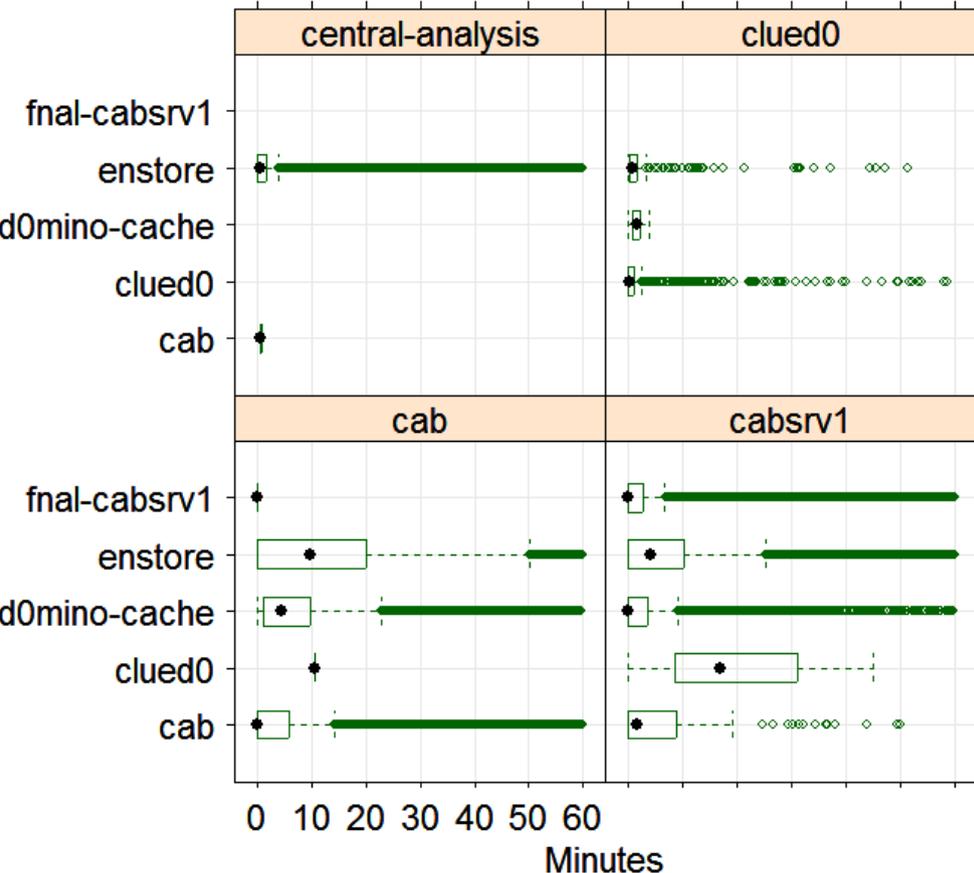
- **Station:** Permanent and transient services that monitor file consumption and make requests to storage resources for more files.
- **Project:** Delivers files to processes and keeps permanent record:  
sam get project summary
- **Dataset Definition:** “data\_type physics and run\_number 78904”
- **Consumer:** User application that consumes and produces data (one or many exe instances)  
Examples: script to copy files; reconstruction job

# SAM Statistics - Operations Data



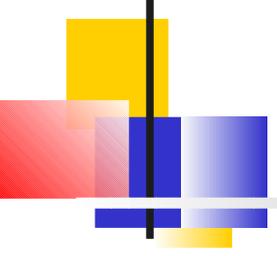
Wait Time for File Delivery (truncated)

0 10 20 30 40 50 60



- Time between *Request Next File* and *Open File*
- For CAB and CABSRV1
  - 50% of enstore transfers occur within 10 minutes.
  - 75% within 20 minutes
  - 95% within 1 hour
- For CENTRAL-ANALYSIS and CLUED0
  - 95% of enstore transfers within 10 minutes

Station	CAB	CABSRV1	CLUED0	CA
% no wait	30%	40%	38%	18%

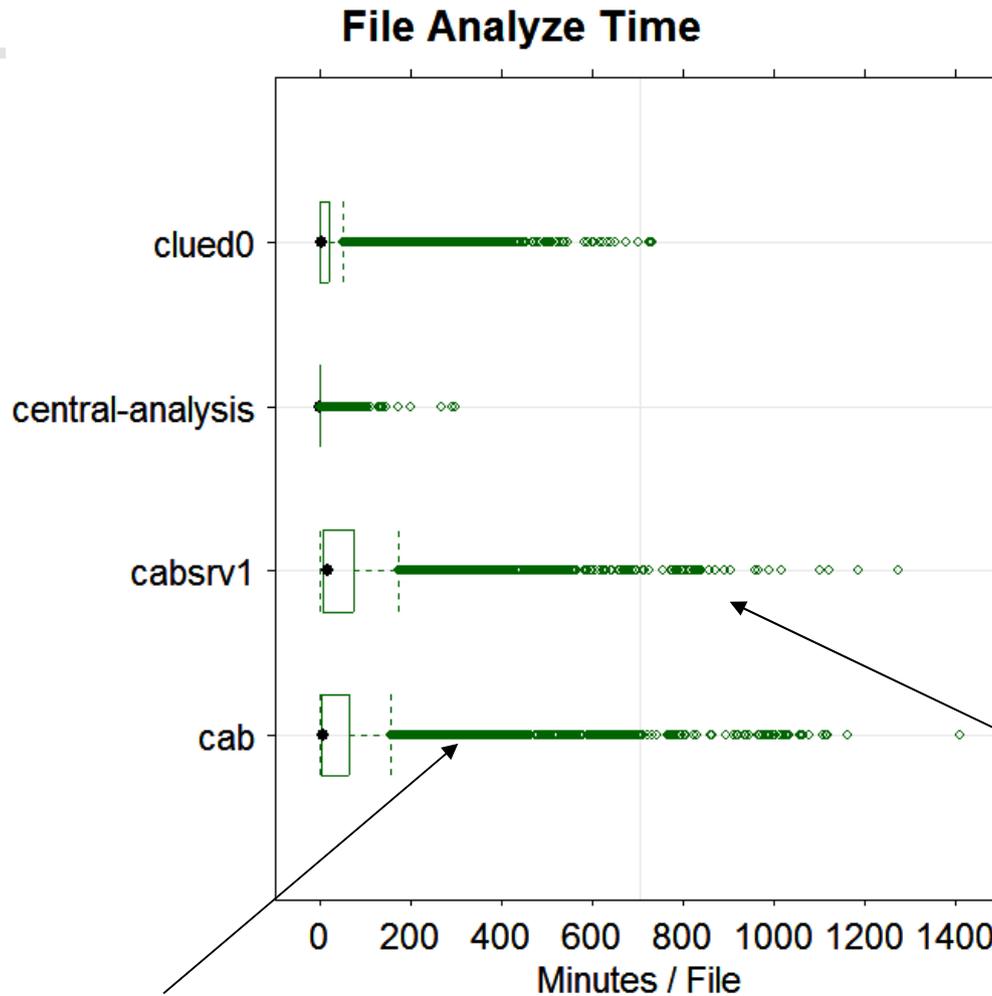
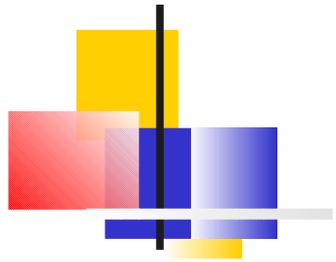


# The Grid part of SAMGrid: JIM

---

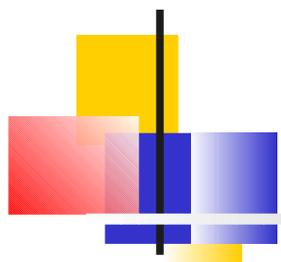
- JIM components provide:
  - Job submission service via Globus Job Manager, augmented by some VO requirements
  - Job monitoring service from remote infrastructure
  - Authentication services

# SAM Statistics - Operations Data



Files from tape  
come later

Cached Files delivered first and fast



# CPU from GridKa

(Biggest present off-site SAM user)

---

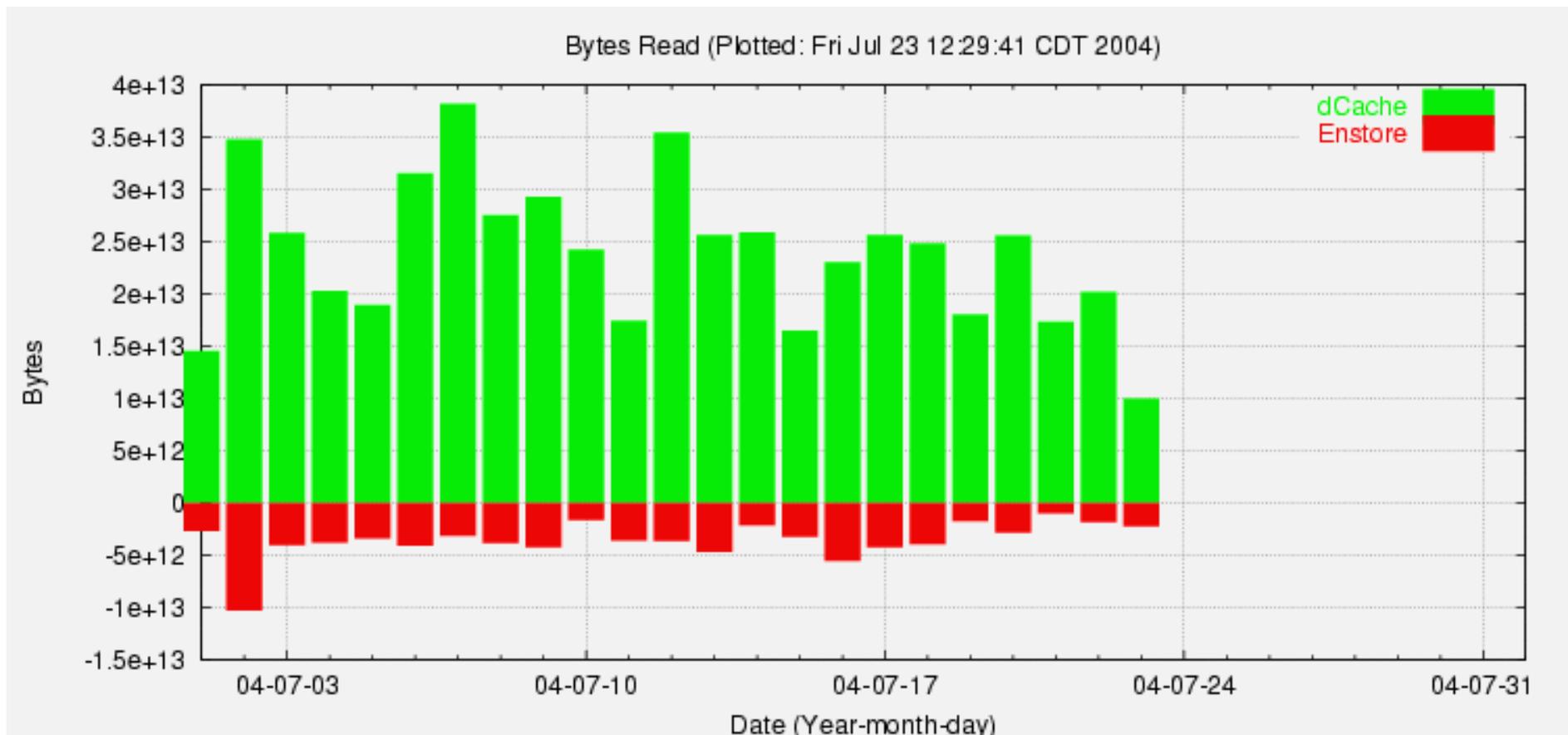
- May 1-6: 650
- May 7-17: 704
- May 18-27: 604
- May 28-31: 710
- May total 492,860 cpu hrs, 1THz roughly
- June 1-7: 740, 8-14 780, 15 power out, 16-30 700
- June total 507,360 cpuhrs, 1THz roughly

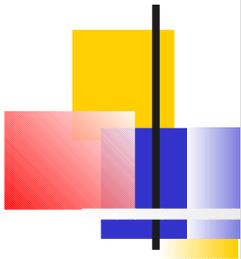
Cluster not CDF-exclusive -  
Need Grid to make this  
resource available  
to full CDF collaboration!

# CDF Data Handling: Dcache on CAF

CDF Reads 25  
TB/Day on CAF

NonGrid Running





Analysis Farm:  fcdhead1.fnal.gov:8000

Specify SAM dataset? SAM Dataset ID:

Process Type:

Initial Command:

Original Directory:

Output File Location:

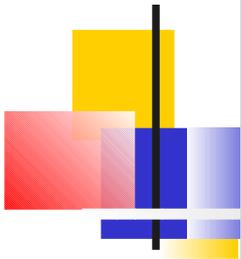
Email? Email Address:

---

(2004-01-29 12:29:30) Specifying of SAM dataset enabled

Easy Use  
of SAM

Originally  
Fermilab  
only



Analysis Farm:  fcdhead1.fnal.gov:8000

Specify SAM dataset? SAM Dataset ID:

Process Type:

Initial Command:

Original Directory:

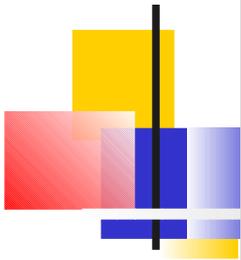
Output File Location:

Email? Email Address:

```
{2004-01-29 12:29:30} Specifying of SAM dataset enabled
{2004-01-29 12:31:58} toronto analysis farm selected
```

Easy Use  
of SAM

Now Works the  
Same  
Inside or Outside  
Lab



Analysis Farm:  fcdhead1.fnal.gov:8000

Specify SAM dataset? SAM Dataset ID:

Process Type:

Initial Command:

Original Directory:

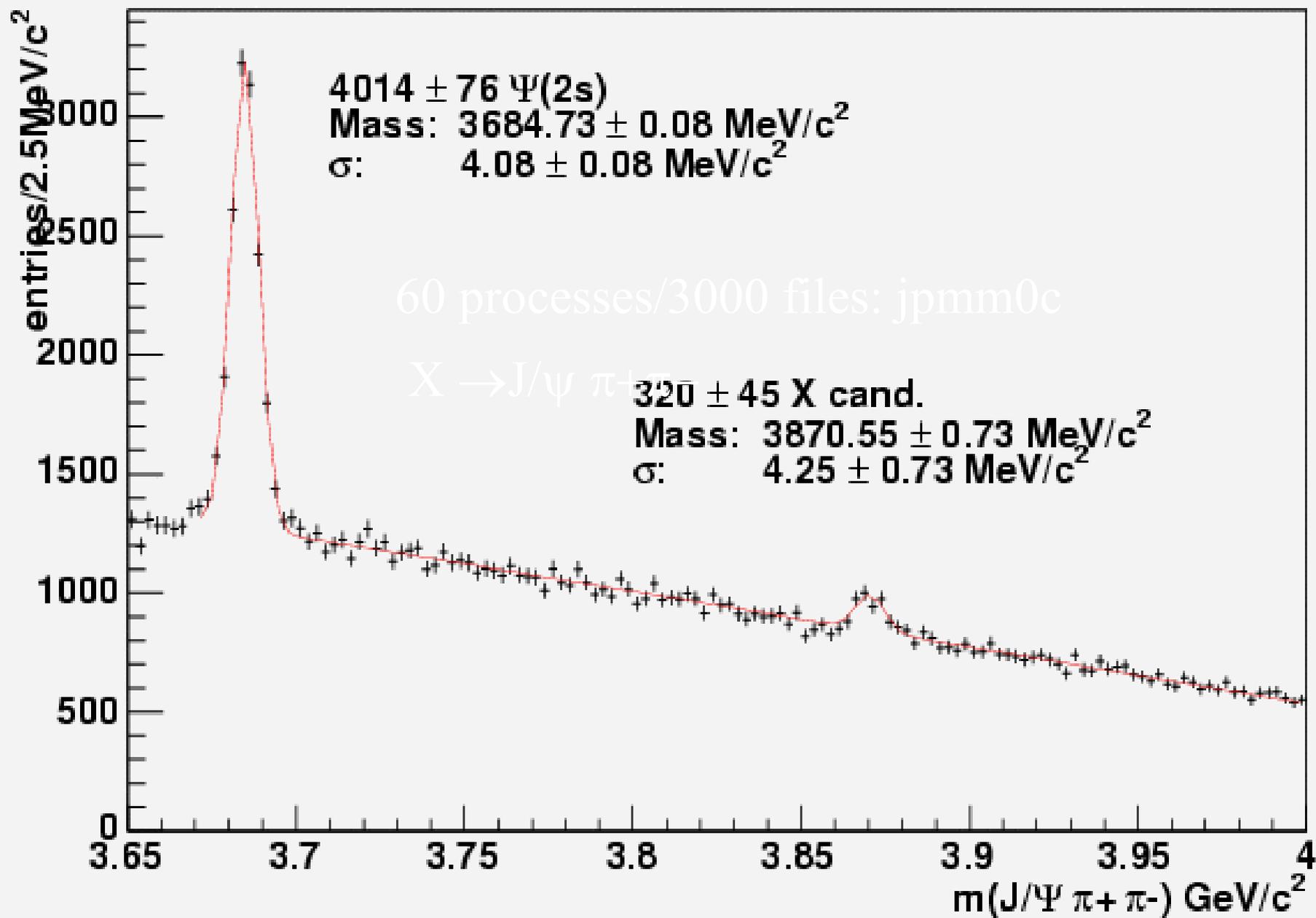
Output File Location:

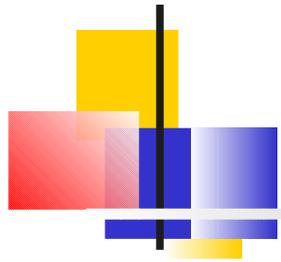
Email? Email Address:

```
(2004-01-29 12:33:39) jim analysis farm selected
(2004-01-29 12:33:44) Specifying of SAM dataset enabled
```

Uses SAM  
In Same  
Way

Example  
Use of  
Grid  
Resource!





# Screen Shot of Web page

[http://hexfm1.rutgers.edu/DATA\\_INFO/sam\\_data/](http://hexfm1.rutgers.edu/DATA_INFO/sam_data/)

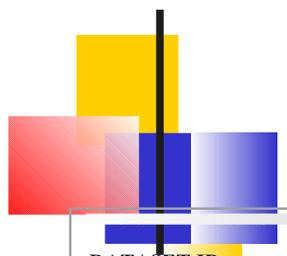
## CDF Datasets on SAM stations

- [cdf-cnaf](#)
- [cdf-fzkka](#)
- [cdf-knu](#)
- [cdf-rutgers](#)
- [cdf-sdsc](#)
- [cdf-taiwan](#)
- [cdf-toronto](#)
- [cdf-ttu](#)



Click on cnaf...

## Datasets Stored Locally on cdf-cnaf: Locked (Still testing dynamic movement of files)

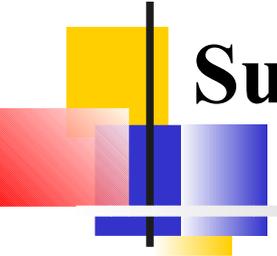


DATASET ID	GBYTES	EVENTS	FILES	CACHED	LOCKED
xbhd0d	158.00	12560062	27848	408( 1%)	325( 1%)
hbhd0c	158.00	12560062	27848	109( 1%)	109( 1%)
hbhd0d	158.00	12560062	27848	351( 1%)	351( 1%)
jbot0h	649.09	3240403	690	1( 0%)	none
gmbs09	1224.65	7676037	1330	17( 1%)	17( 1%)
bpel0d	138.00	10700000	2194	all	all
gpjj08	524.53	16019542	70	1( 1%)	1( 1%)
xpmm0d	524.53	16019542	2675	all	all
xpmm0c	524.53	16019542	2675	all	all
jpmmm08	575.70	27928	10	1( 1%)	none

Expert Usage!  
(testing dynamic  
movement)

All in Cache

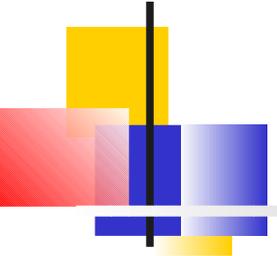
And  
Locked  
Via SAM



# Summer 2004 Goal: Expand Resources, More Efficient Operations

---

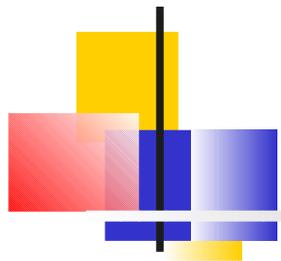
- ✓ SAM on (D)CAFs
  - Reduce DH operations load: EMAIL/Fair Tape Share
- ✓ Pin Datasets Remotely via SAM
- ✓ MC Data Import:
  - Automate to reduce workload
  - Replace DFC with SAM
- 04 Goal was >25% offsite computing load
- Met this goal (35% of CDF collaboration-wide cpu capacity is now available offsite)



## 2004 Goals: Achievements So Far

---

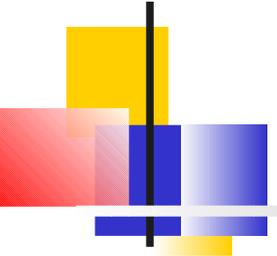
- MC Data Import: will be in 5.3.4
- SAM on (D)CAF:
  - stress testing/fix bugs: need Beta Testers to do real analysis: used 20% of CAF reading golden Datasets (20TB/Day)
  - V6 schema adopted, product deployment now underway
- Datasets Pinned and available
  - [http://hexfm1.rutgers.edu/DATA\\_INFO/sam\\_data/](http://hexfm1.rutgers.edu/DATA_INFO/sam_data/)
- DCAF utilization: few high-intensity users so far but no problems in principle
  - Provided useful cpu capacity for summer conferences
  - Now need next phase of data handling and grid submission



# CDF Grid Strategy: Outlook and Goals

---

- Currently 35% of CDF collaboration-wide open computing capacity from external resources.
  - Utilizes only resources fully controlled by CDF so far: Kerberos/fbsng/CDF Condor dCAF
  - SAM used and available on ALL resources
- December 15, 2004: JIM/Grid3-OSG/LCG comparison ends (Mainly MC)
- By end of 2005: 50% of computing resources from external sources, broader use of Grid



# Conclusions

---

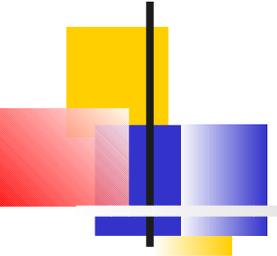
- CDF making good progress toward providing increased off-site computing and DH capacity.
- Can capture many more resources using Grid to achieve physics mission.
- SAM is working now for CDF and will reduce operational loads, improve user experience.
- To make progress, add new software tools and move to capabilities like those supported for/by the LHC and other global grid efforts.

# SAM: The work plan for the next 2 years



---

- Evaluate technology changes/upgrades
  - Improvements for installation/config management
  - CORBA to Web Services
  - XML based logging
  - Distributed database
  - Merge SAM catalog w/ other replica schemas
  - Working with SRM
  - Interaction of tools with data handling: Workflow, local and global job management
  - VO Organisation/security: file transfer



# Problems

## Encountered/Solved/Unresolved

---

- CDF Contentious design issues Sep 03 – Sep 04
  - installation difficulties
  - file name as GUID **no change to model**
  - interface into experiment framework **work in SAM**
  - communication with dcache **work in SAM, future work**
  - use of dimensions and parameters **proposed work in SAM**
  - process bookkeeping **future work in SAM**
- MINOS – file delivery ordering & grouping **no change to model**