Update on D0 and CDF computing models and experience

Frank Wuerthwein

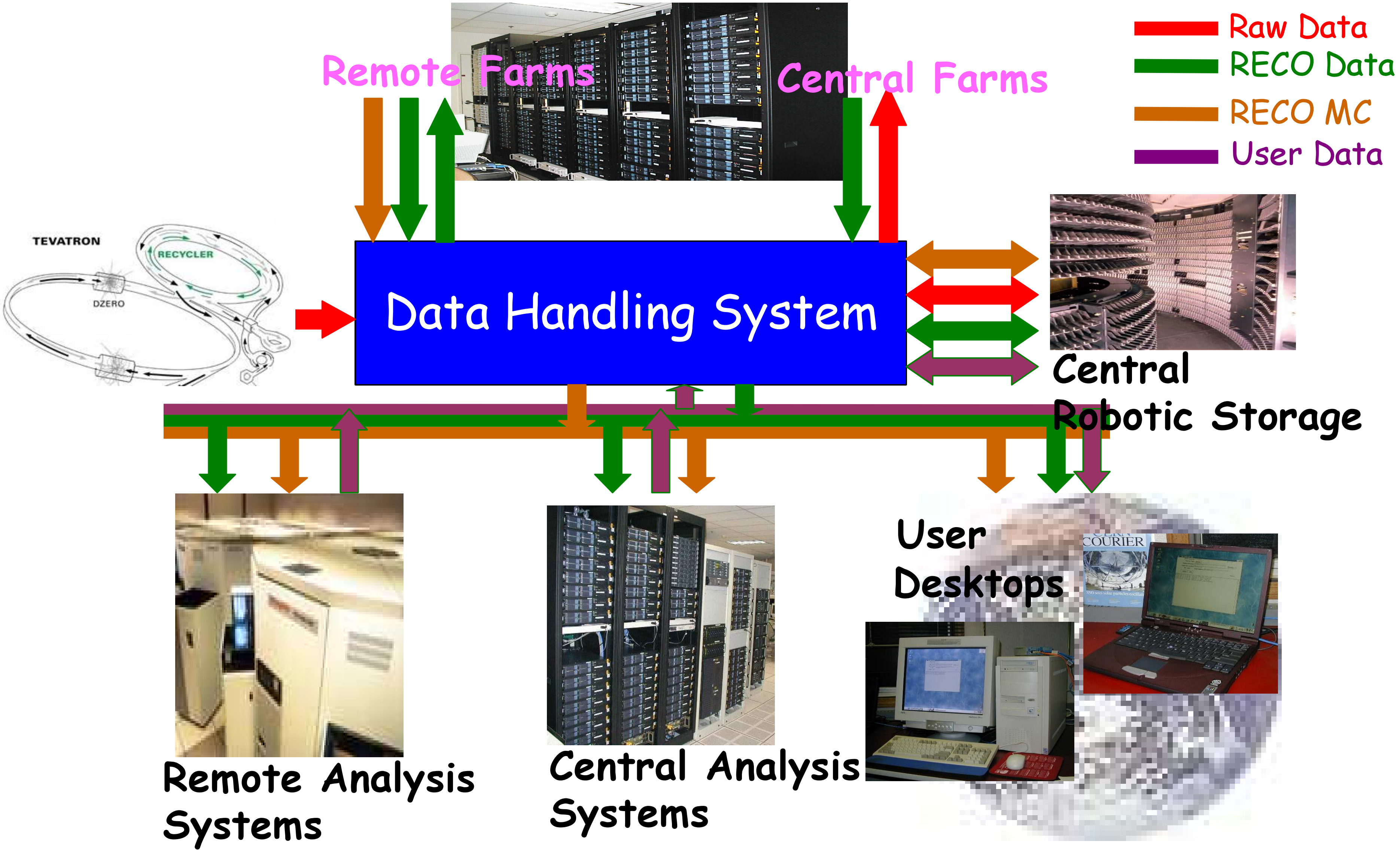
UCSD

For CDF and DO collaborations

October 2nd, 2003

Many thanks to A.Boehnlein, W.Merritt,G.Garzoglio

Computing Model


TEVATRON
RECYCLER
DZERO
COURIER

🟥 Raw Data
🟩 RECO Data
🟧 RECO MC
🟪 User Data

Remote Farms

Central Farms

Data Handling System

Central Robotic Storage

Remote Analysis Systems

Central Analysis Systems

User Desktops

Vital Statistics

Vital stats.	CDF	D0
Raw data (kB/evt)	135	230(160)
Reco data (kB/evt)	50-150	200
Primary User data (kB/evt)	25-150	20
Format for User Skims	Ntuple/DST	TMB
User skims (kB/evt)	5-150	20-40
Reco time (Ghz-sec/evt)	2	50-80
MC chain	param. & geant	param or full geant
MC (Ghz-sec/evt)	15	1 or 170
Peak data rate (Hz)	80-360	50
Persistent Format	RootIO	D0om/dspack
Total on tape in 9/03 (TB)	480	420

Hardware Cost Drivers

- CDF: User Analysis Computing
  - Many users go through 1e8 evts samples
  - Aggressive bandwidth plans
  - 0.8-1.2 M$ CPU farms needed for FY04/5/6
- D0: (Re-)Reconstruction
  - Small # of layers in tracker -> pattern rec.
  - 3month, once a year reprocessing
  - Prorated cost: 0.9-0.6 M$ CPU farms FY04/5/6
  - Purchase cost: 2.4-1.8 M$ CPU famrs FY04/5/6

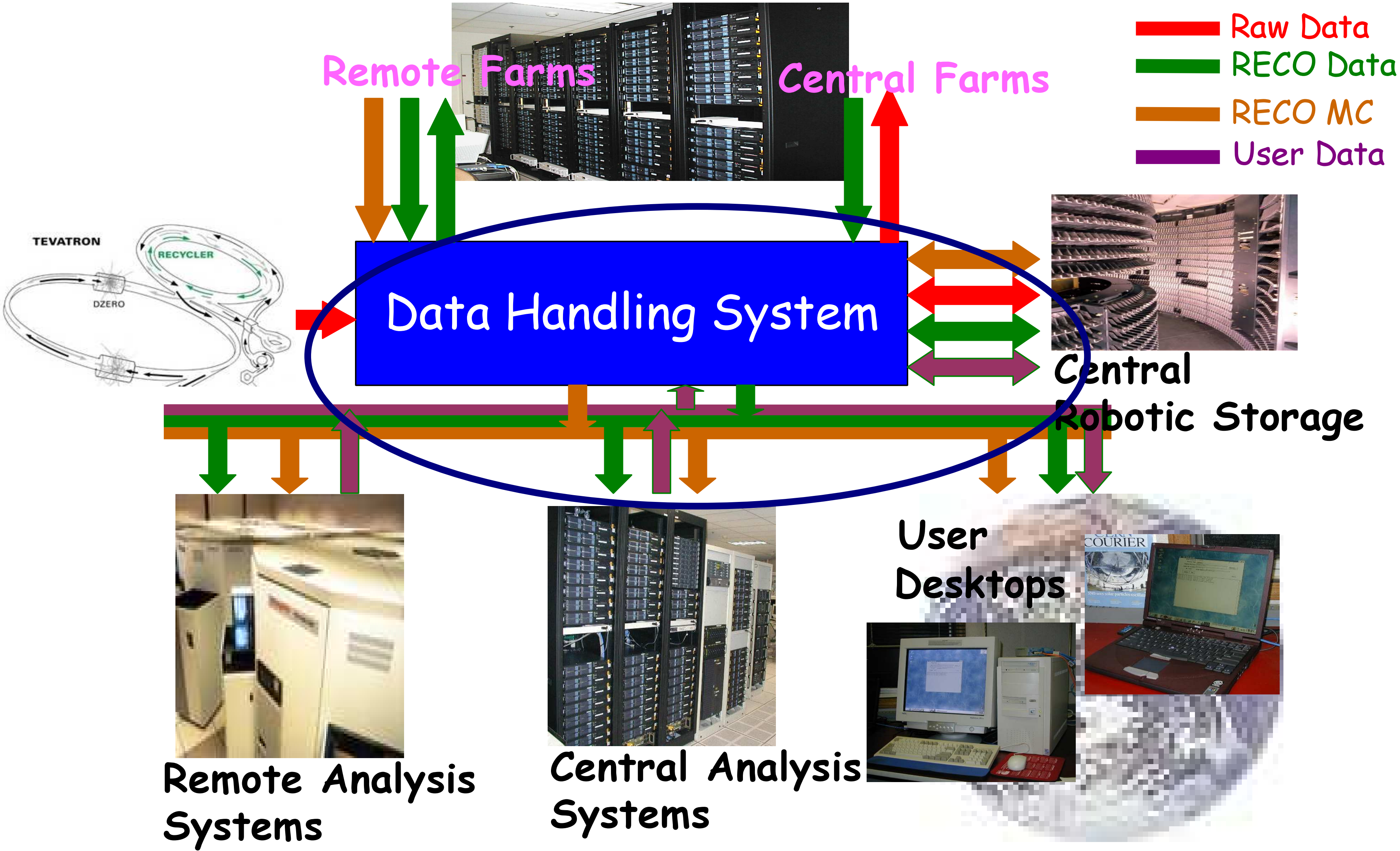
Offsite Computing Plans

	CDF	D0
MC production	today	today
Primary reconstruction	NO	NO
Reprocessing	>FY06	FY04
User level MC	FY04	NA
User Analysis Computing	FY05	~FY05

Needs & fiscal pressure differ.
=> Focus differs as well.

D0: WAN distributed DH
CDF: WAN distributed user analysis computing

Computing Model


TEVATRON
RECYCLER
DZERO

🟥 Raw Data
🟩 RECO Data
🟧 RECO MC
🟪 User Data

Remote Farms

Central Farms

Data Handling System

Central Robotic Storage

Remote Analysis Systems

Central Analysis Systems

User Desktops

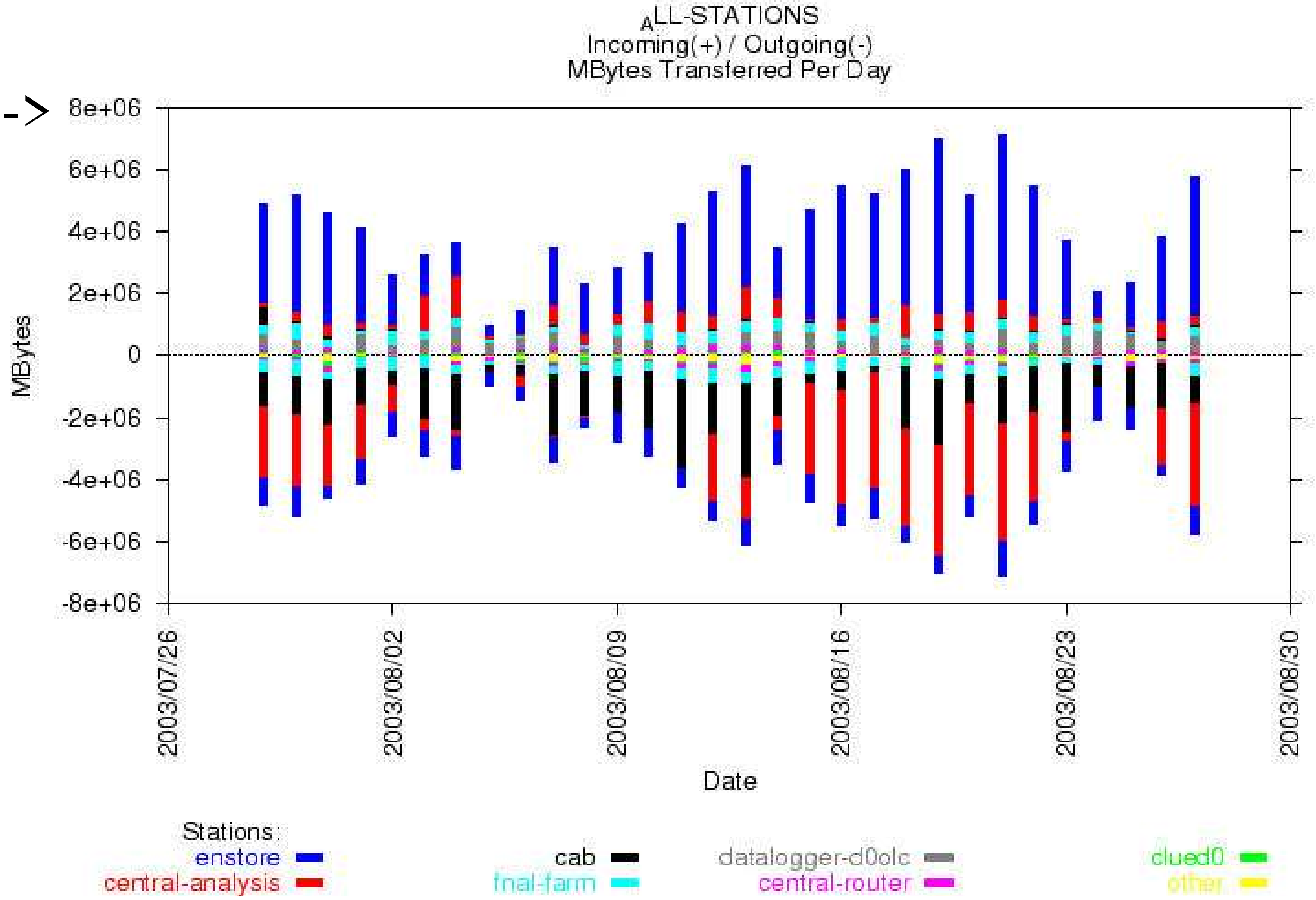
COURIER

Sequential Access via Meta-data

- Flagship CD-Tevatron Joint project. Initial design work ~7 years ago. Pioneered by D0.
- Provides a "grid-like" access to DO & CDF data
  - Comprehensive meta-data to describe collider and Monte Carlo data.
  - Consistent user interface via command line and web
  - Local and wide area data transport
  - Caching layer
  - Batch adapter support
  - Declare data @ submission time -> optimized staging
- Stable SAM operations allows for global support and additional development
  - CDF status: SAM in production by 1/04
  - CDF requires: dCache (DESY/FNAL) caching software

DO SAM Performance

8TB per day ->



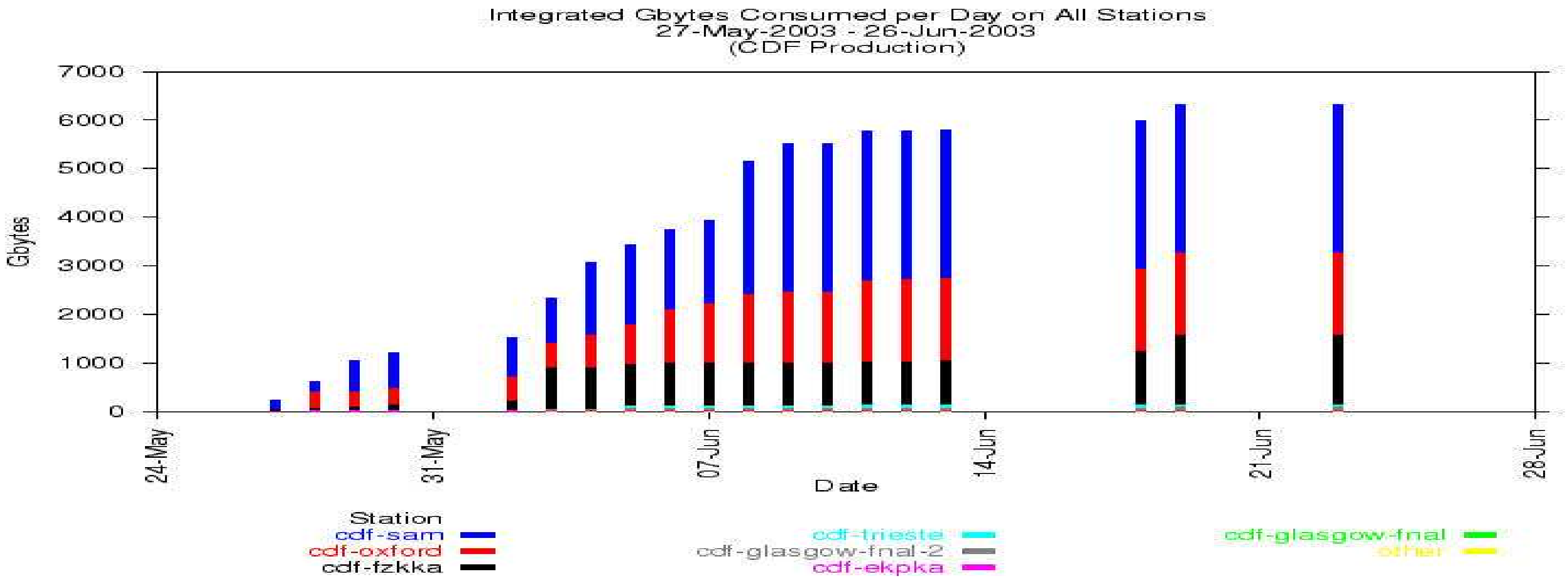
Frank Wuerthwein, UCSD

CDF Data Handling

- dCache
  - 3 caching policies: permanent, volatile, buffer
  - 100TB disk space; 200-600MB/s total read
  - Drawback: data declaration @ runtime
- Remote systems use SAM today.
- Future Vision: SAM->dCache->Enstore
  - Storage & Cacheing via Enstore/dCache
    - Cache for Enstore tape archive
    - Virtualizing "scratch" space without backup
  - Quota, data movement, metadata via SAM

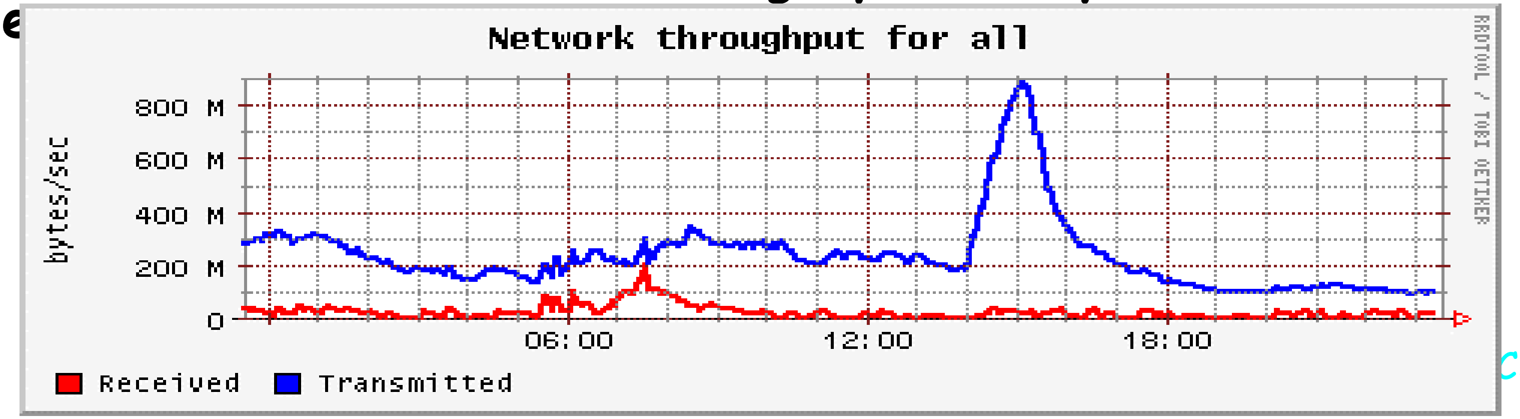
CDF DH performance

CDF Remote SAM Usage for the past month



CDF CAF dCache Usage yesterday

900MB/se
At peak



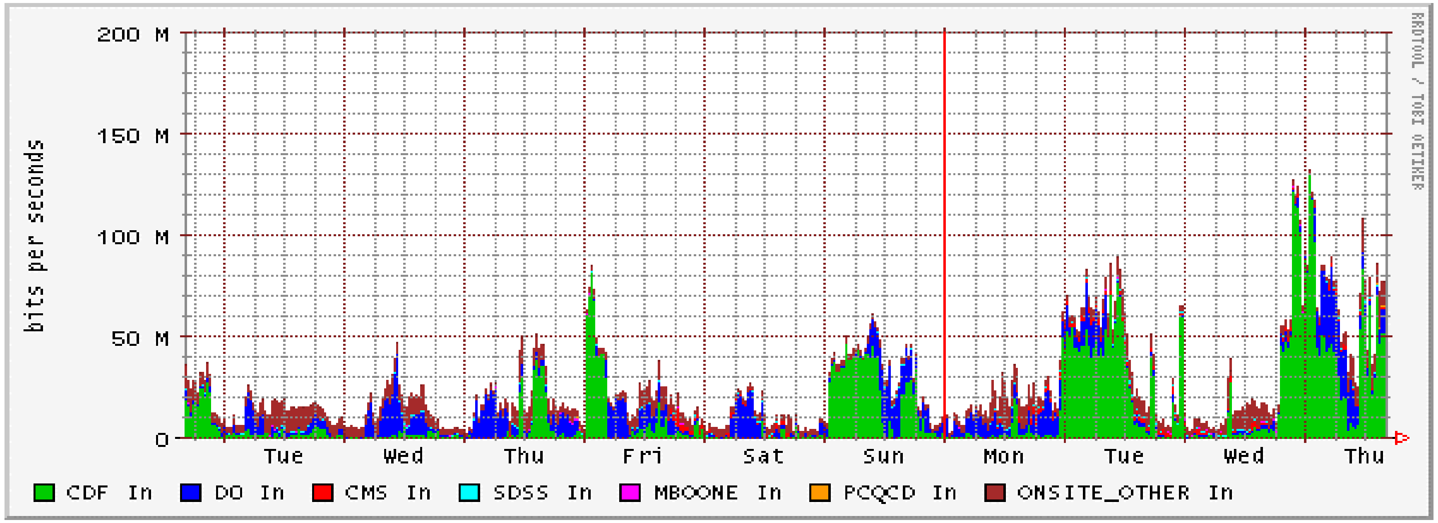
CSD

Wide Area Networking

OC(12) to ESNET, proposed OC(48) to Starlight (~1year)
In/Out Traffic at the border router random week in June.

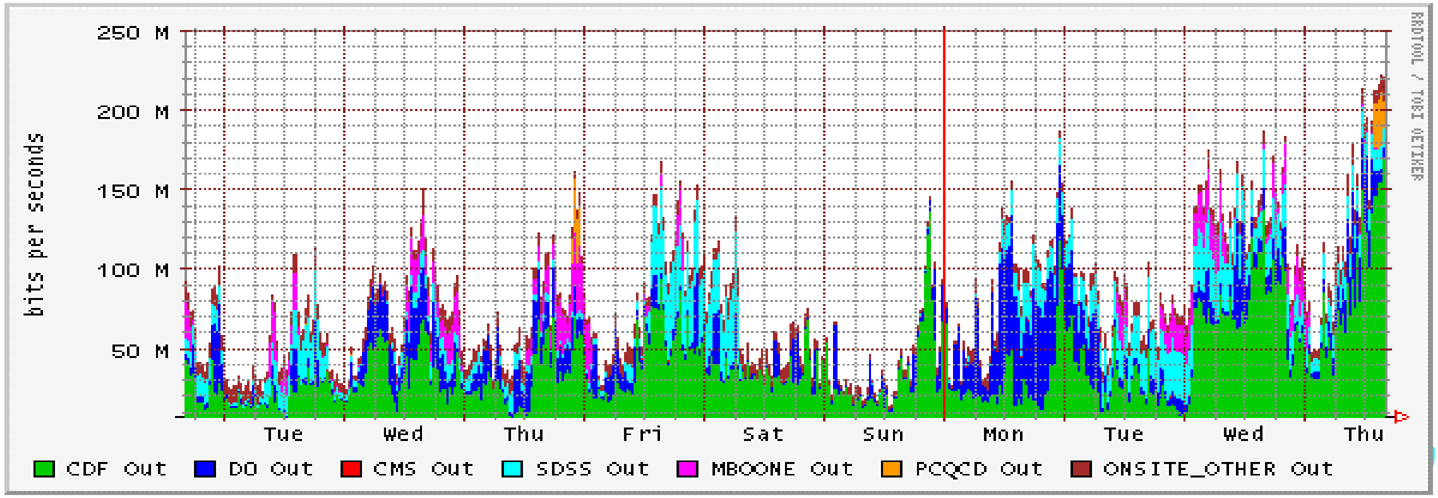
Inbound

100Mb/s->



Outbound

100Mb/s->



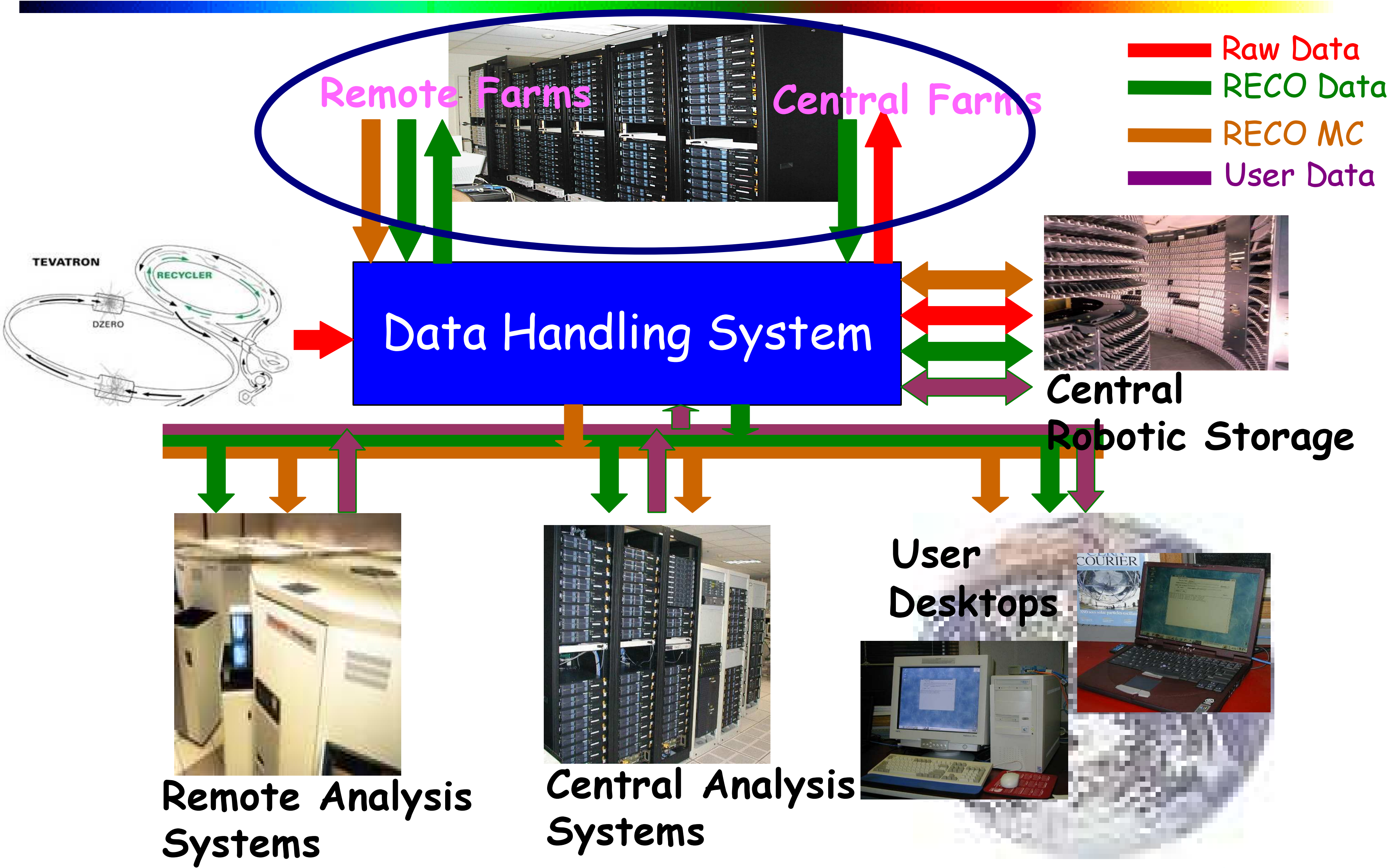
Databases

- Both experiments use Oracle on SUN systems
- Databases used for SAM, calibration, run and trigger table tracking, luminosity accounting
- Different models for each experiment
  - DO uses a tiered architecture of user applications, db servers and databases
  - CDF uses direct db access and replicated databases
  - Neither system is ideal
    - Perhaps not exactly a technical issue
    - Both experiments have severely underestimated requirements, development time, use cases, integration time, and operational burden.
    - Table design usually non optimal for use.
  - A personal opinion as a non-expert: poor table design leads to massive DB needs -> early expert help worth it.

SAM-Grid

- Several aspects to SAM-Grid project, including Job and Information Monitoring (JIM)
  - Part of the SAM Joint project, with oversight and effort under the auspices of FNAL/CD and Particle Physics Data Grid.
  - JIM v1.0 is being deployed at several sites, learning experience for all concerned
  - Close collaboration with Globus and Condor teams
  - Integration of tools such as runjob and CAF tools
  - Collaboration/discussions within the experiments on the interplay of LCG with SAM-Grid efforts
- Tevatron experiments working towards grid environment (interest in OSG).

Computing Model



🟥 Raw Data
🟩 RECO Data
🟧 RECO MC
🟪 User Data

Remote Farms

Central Farms

Data Handling System

Central Robotic Storage

Remote Analysis Systems

Central Analysis Systems

User Desktops

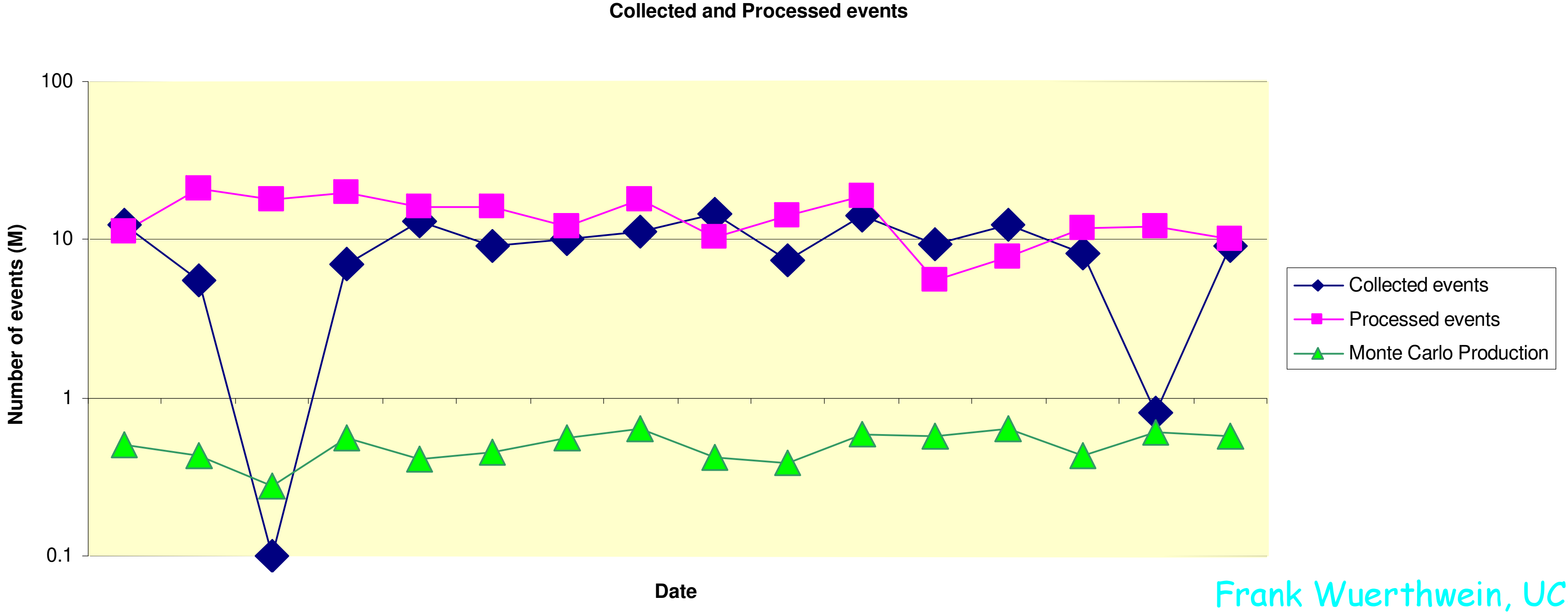
TEVATRON
RECYCLER
DZERO
COURIER

Frank Wuerthwein, UCSD

DO Farm Production

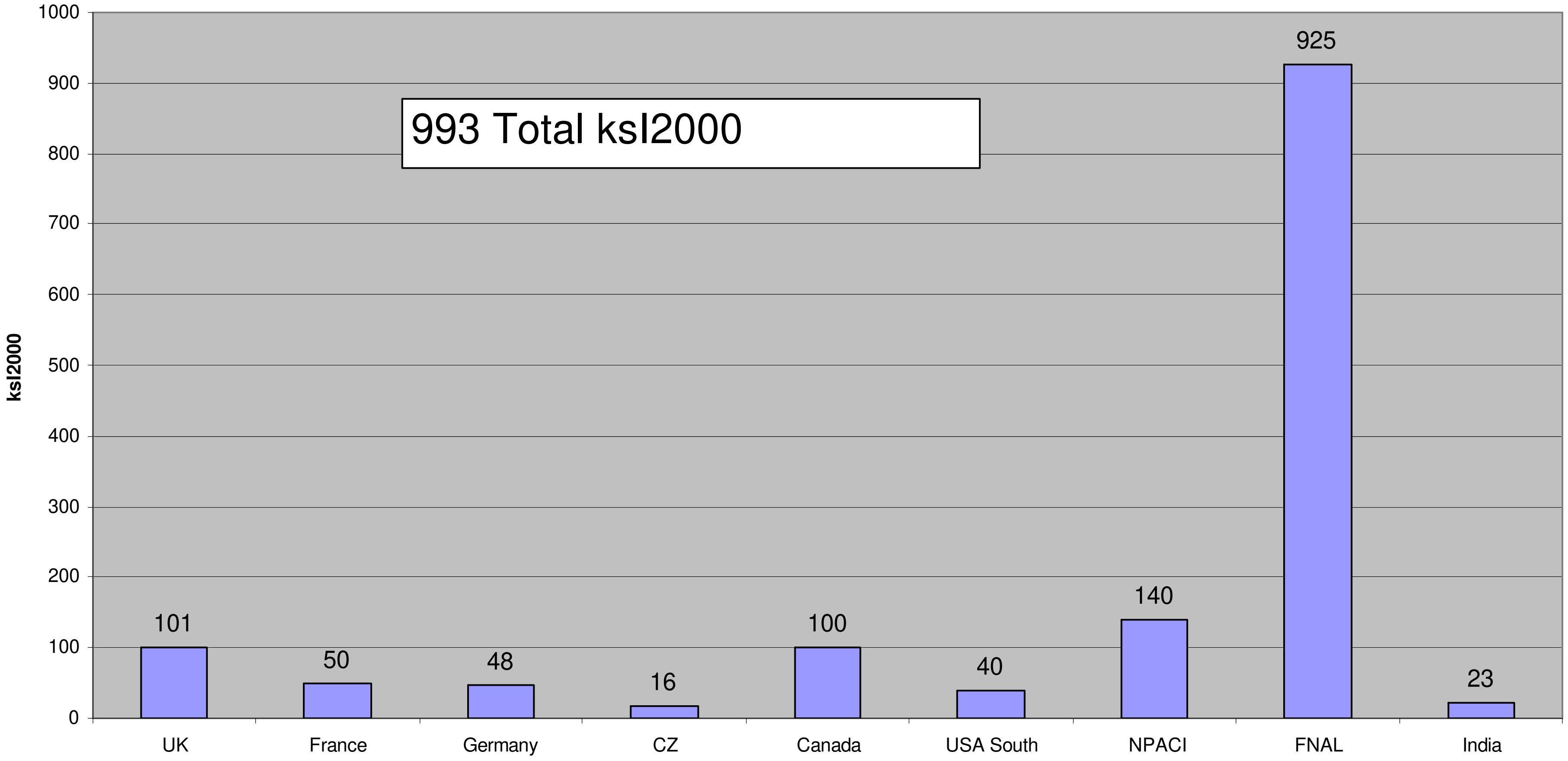
- DØ Reconstruction Farm—13 M event/week capacity- operates at 75% efficiency—events processed within days of collection. 400M events processed in Run II.
- DØ Monte Carlo Farms—0.6M event/week capacity-globally distributed resources. Running Full Geant and Reconstruction and trigger simulation
- Successful effort to start data reprocessing at National Partnership

  Advanced Computing Infrastructure resources at University of Michigan.



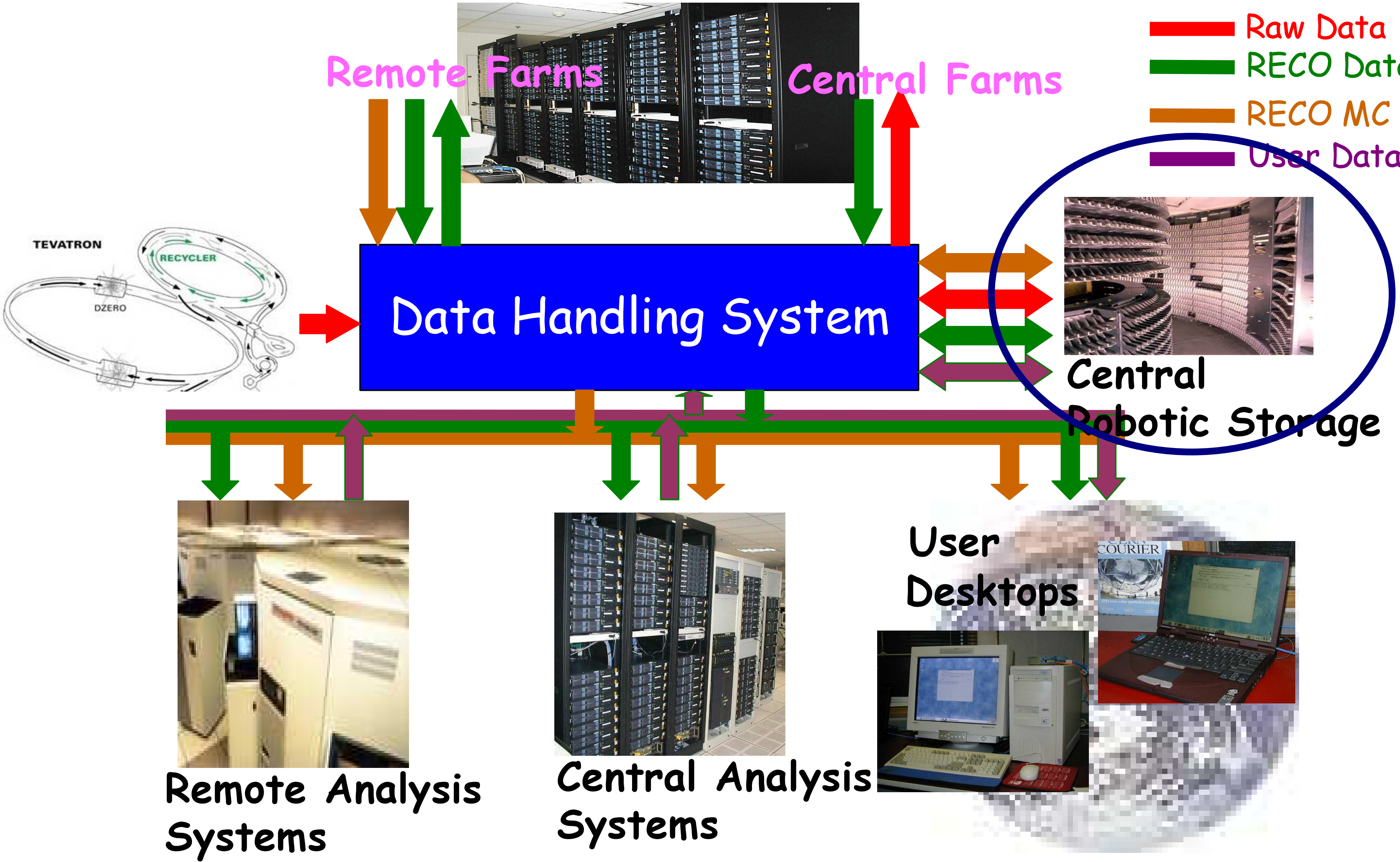
Frank Wuerthwein, UCSD

DO Remote Facilities



Survey of guaranteed DO resources—Work in progress
300M event reprocessing targeted for fall

Computing Model



🟥 Raw Data
🟩 RECO Data
🟧 RECO MC
🟪 User Data

Remote Farms

Central Farms

Data Handling System

Central Robotic Storage

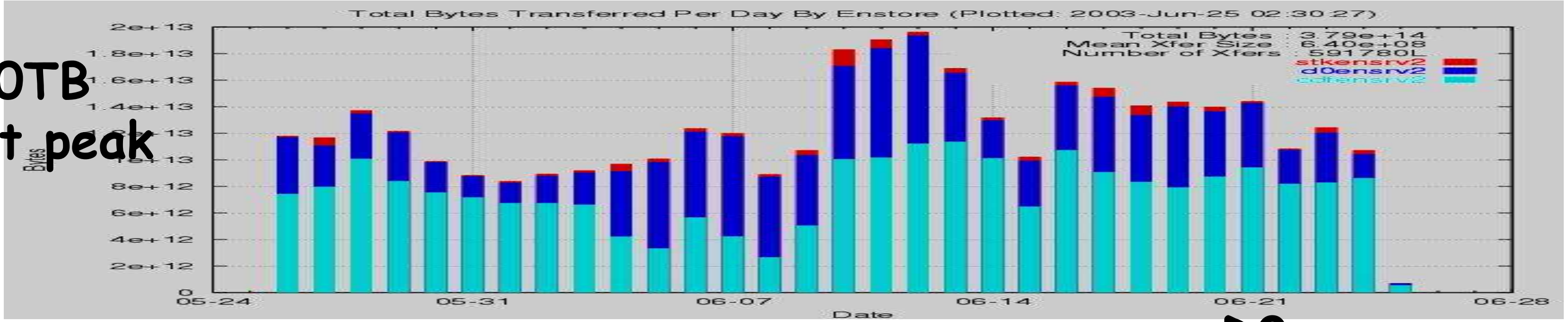
Remote Analysis Systems

Central Analysis Systems

User Desktops

TEVATRON
RECYCLER
DZERO
COURIER

Central Robotics



20TB
At peak

Data to tape, June 25

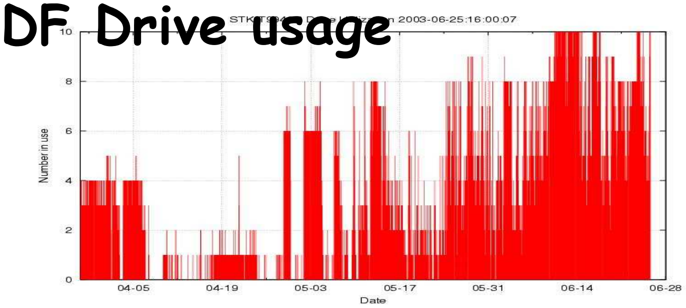
CDF

Library	Stored	#tapes
9940a	302TB	5521
9940b	104TB	1046

D0

Library	Stored	#tapes
STK 9940b	219TB 30.7TB	3780 380
LTO	87.6TB	1099

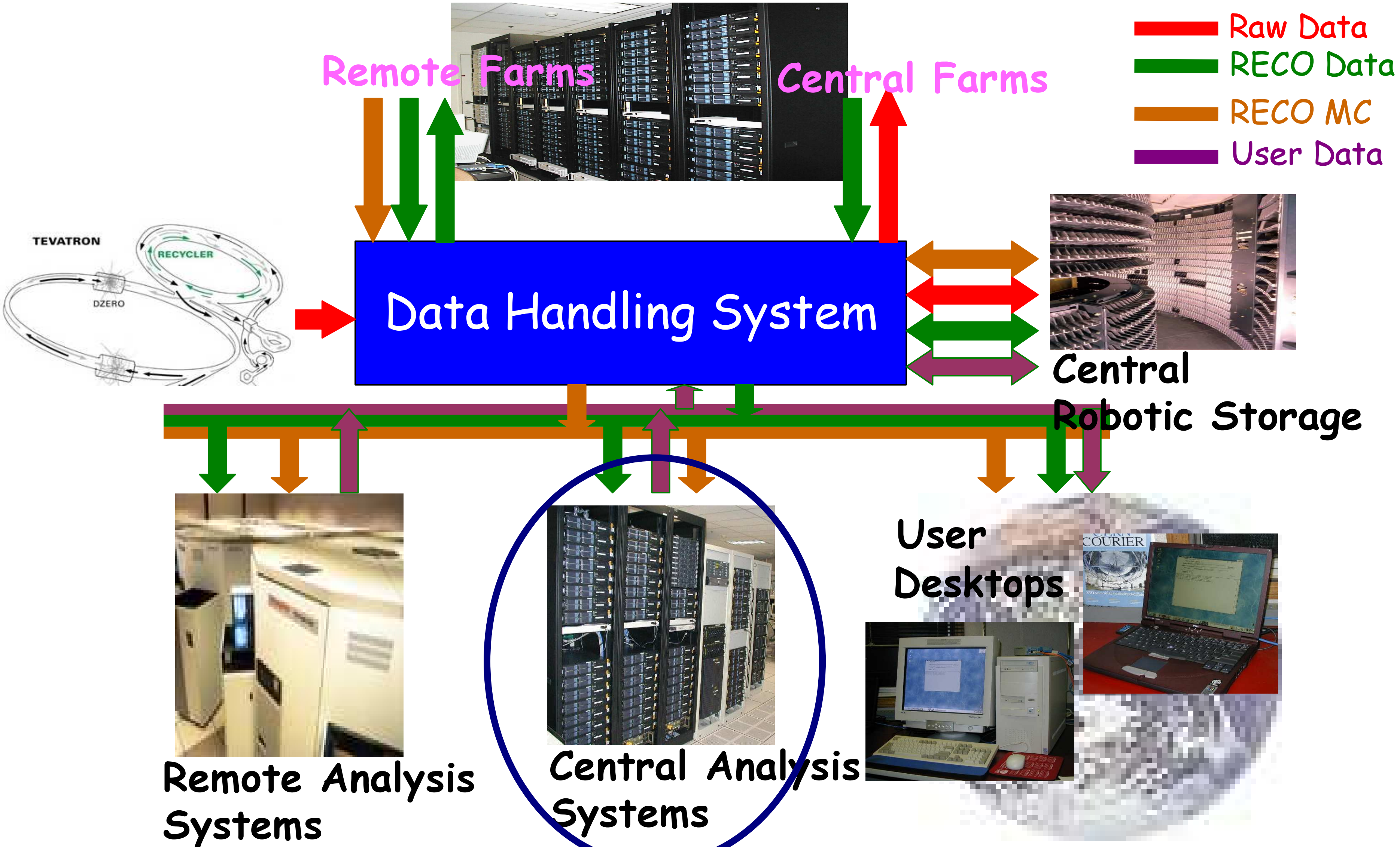
CDF Drive usage



Known data loss due to Robotics/ Enstore for DO 3 GB!

Frank Wuerthwein, UCSD

Computing Model


TEVATRON
RECYCLER
DZERO

🟥 Raw Data
🟩 RECO Data
🟧 RECO MC
🟪 User Data

Remote Farms

Central Farms

Data Handling System

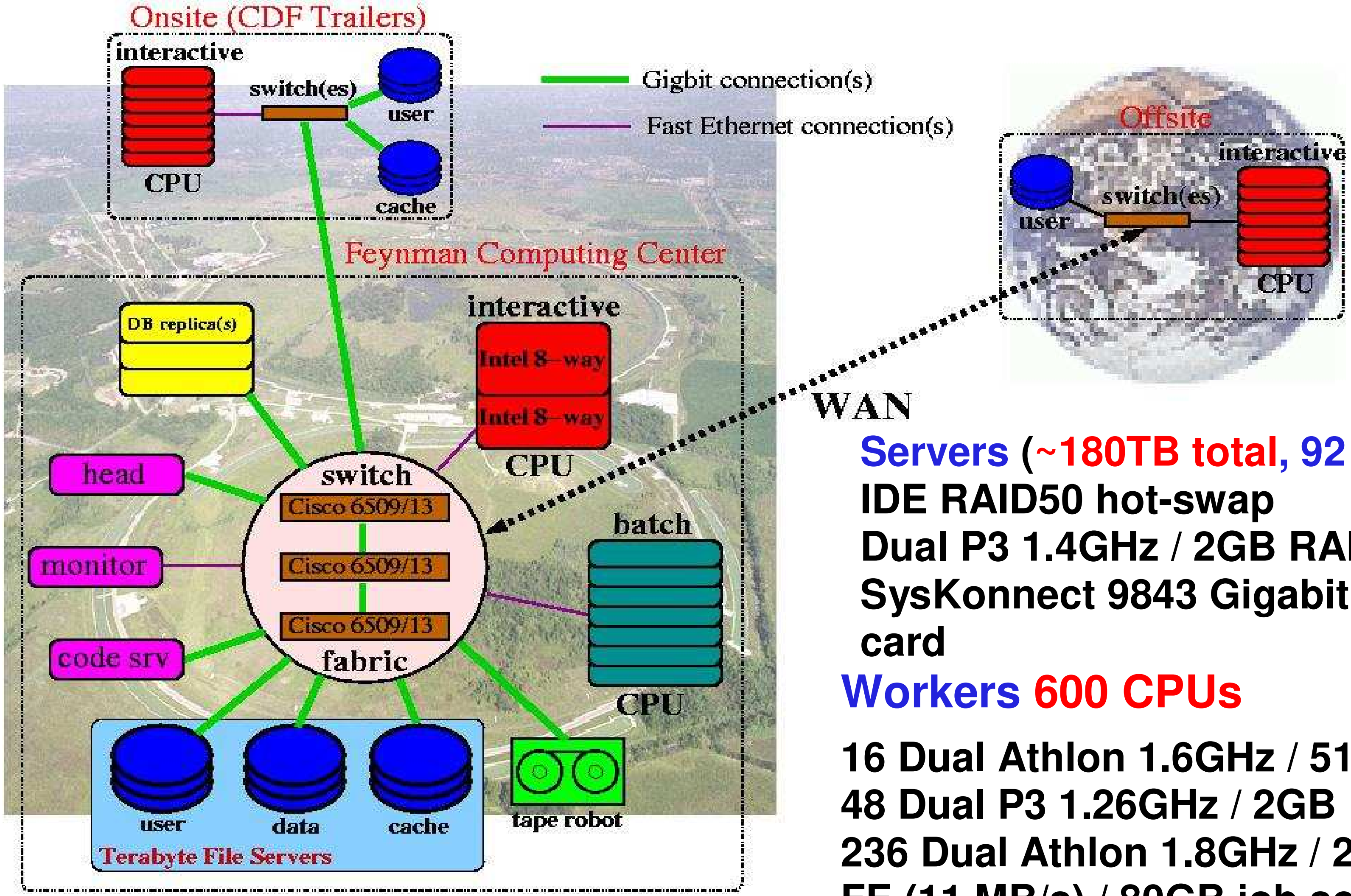
Central Robotic Storage

Remote Analysis Systems

Central Analysis Systems

User Desktops

CDF Central Analysis Facility


Onsite (CDF Trailers)
interactive
switch(es)
user
CPU
cache
Gigbit connection(s)
Fast Ethernet connection(s)
Offsite
interactive
switch(es)
user
CPU
Feynman Computing Center
DB replica(s)
interactive
Intel 8-way
Intel 8-way
CPU
head
switch
Cisco 6509/13
Cisco 6509/13
Cisco 6509/13
fabric
batch
monitor
CPU
code srv
user
data
cache
Terabyte File Servers
tape robot
WAN

Servers (~180TB total, 92 servers):
IDE RAID50 hot-swap
Dual P3 1.4GHz / 2GB RAM
SysKonnect 9843 Gigabit Ethernet card

Workers 600 CPUs

16 Dual Athlon 1.6GHz / 512MB RAM
48 Dual P3 1.26GHz / 2GB RAM
236 Dual Athlon 1.8GHz / 2GB RAM
FE (11 MB/s) / 80GB job scratch each

Use Model & Interfaces

- Useage Model:
  - Develop & debug on desktop
  - Submit "sandbox" to remote cluster
- Interface to Remote Process:
  - Submit, kill, ls, tail, top, lock/release queue
  - Access submission logs
  - Attache a gdb session to a running process
- Monitoring:
  - Full information about cluster status
  - Utilization history for hardware & users
  - CPU & IO consumption for each job

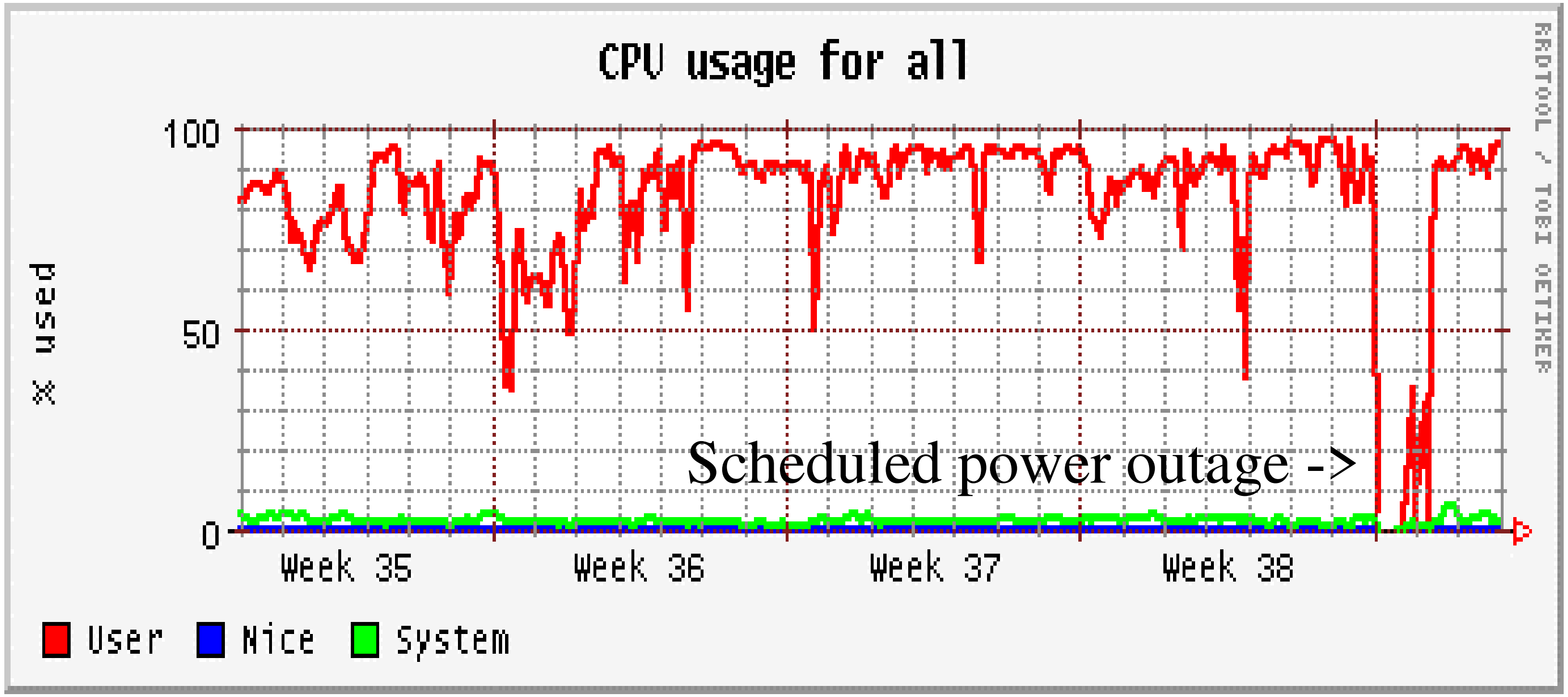
CAF utilization

User perspective:

10,000 jobs launched/day
600 users total
100 users per day

System perspective:

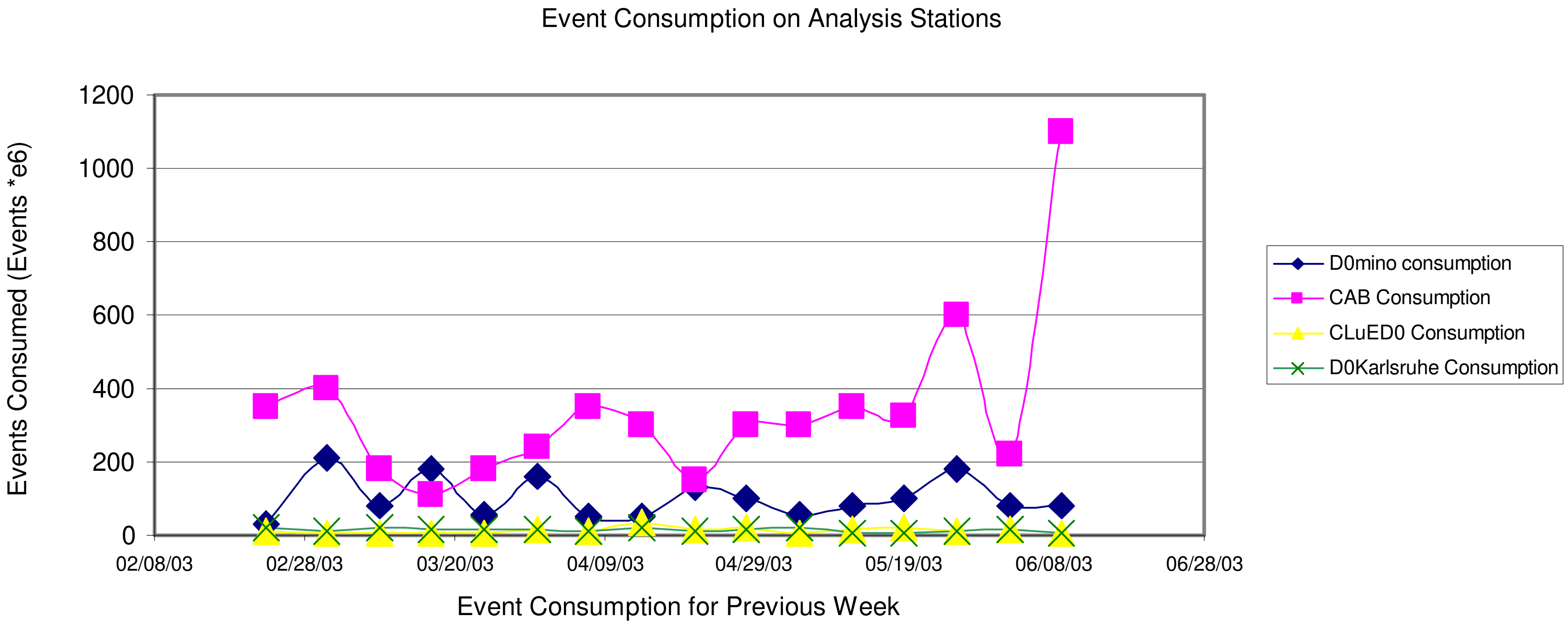
Up to 95% avg CPU utilization
Typically 200-600MB/sec I/O
Failure rate ~1/2000



Scheduled power outage ->

DO Central Analysis Systems

SGI Origin 2000 (D0mino) with 128 300 MHz processors used as 30TB file server for central analysis and as the central router. Central Analysis Backend 320 2.0 GHz AMD machines. Desktop Cluster CLuED0 is also used as a batch engine And for development



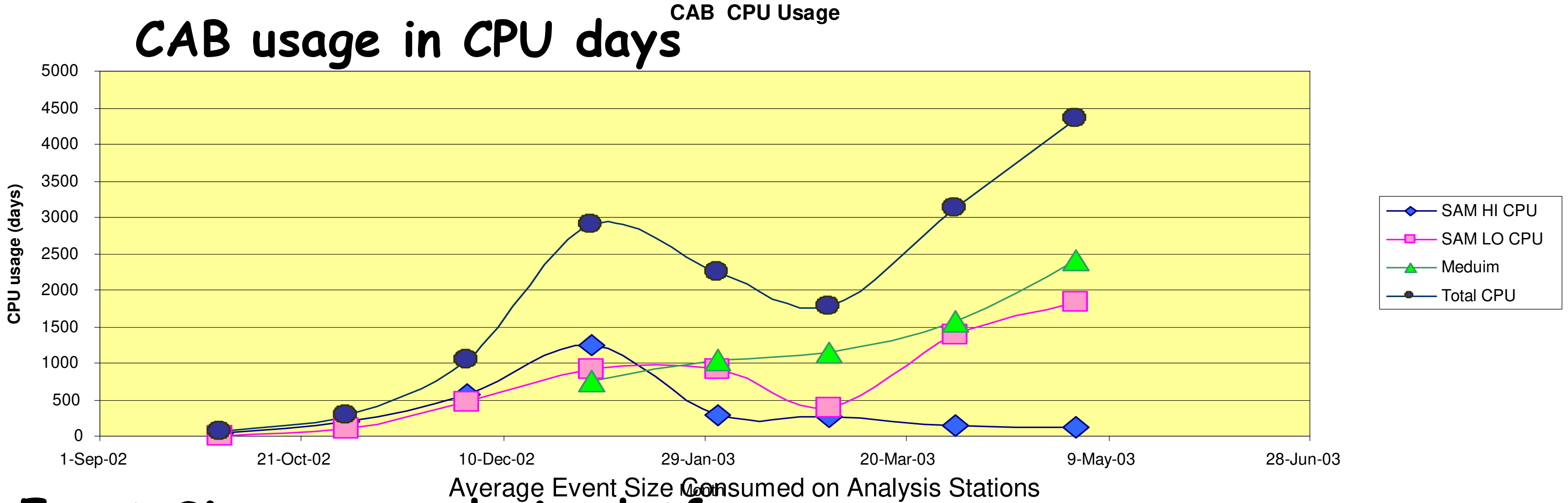
Analysis Effectiveness (includes Reco and all delivery losses)

W/Z skim sample/Raw data available =98.2%

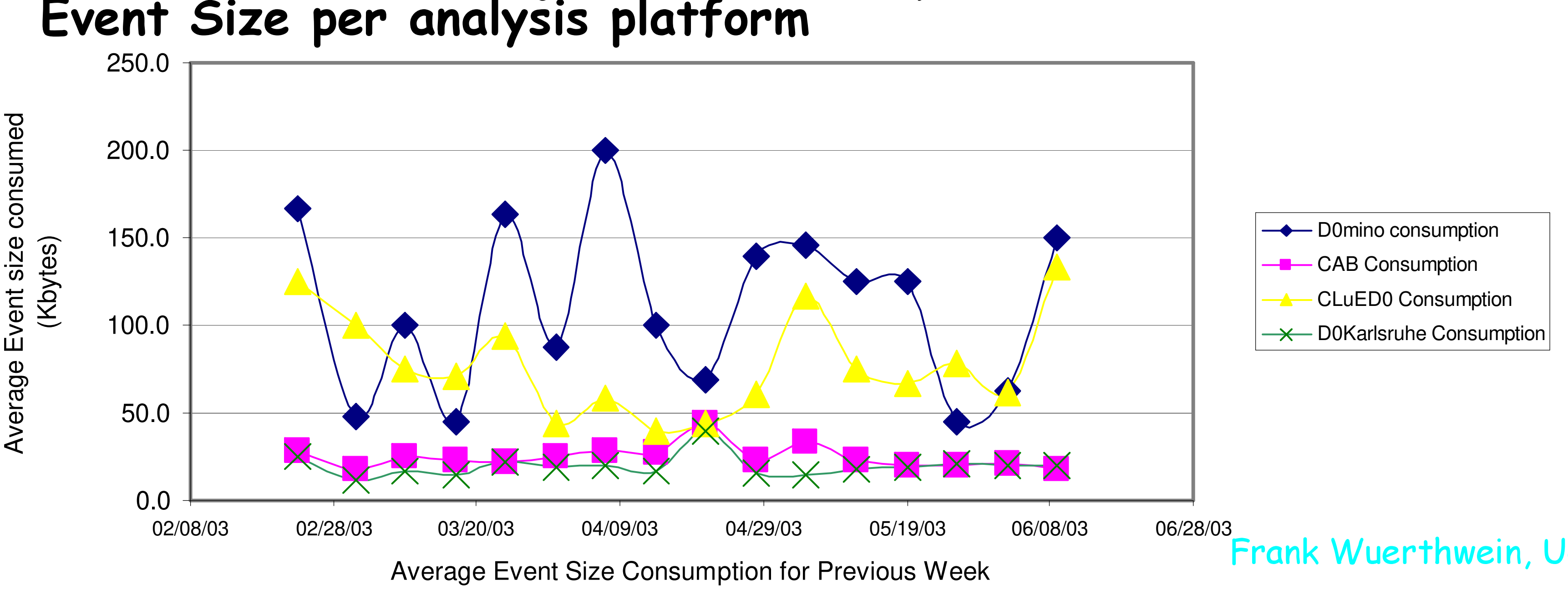
Frank Wuerthwein, UCSD

DO Central Analysis Systems

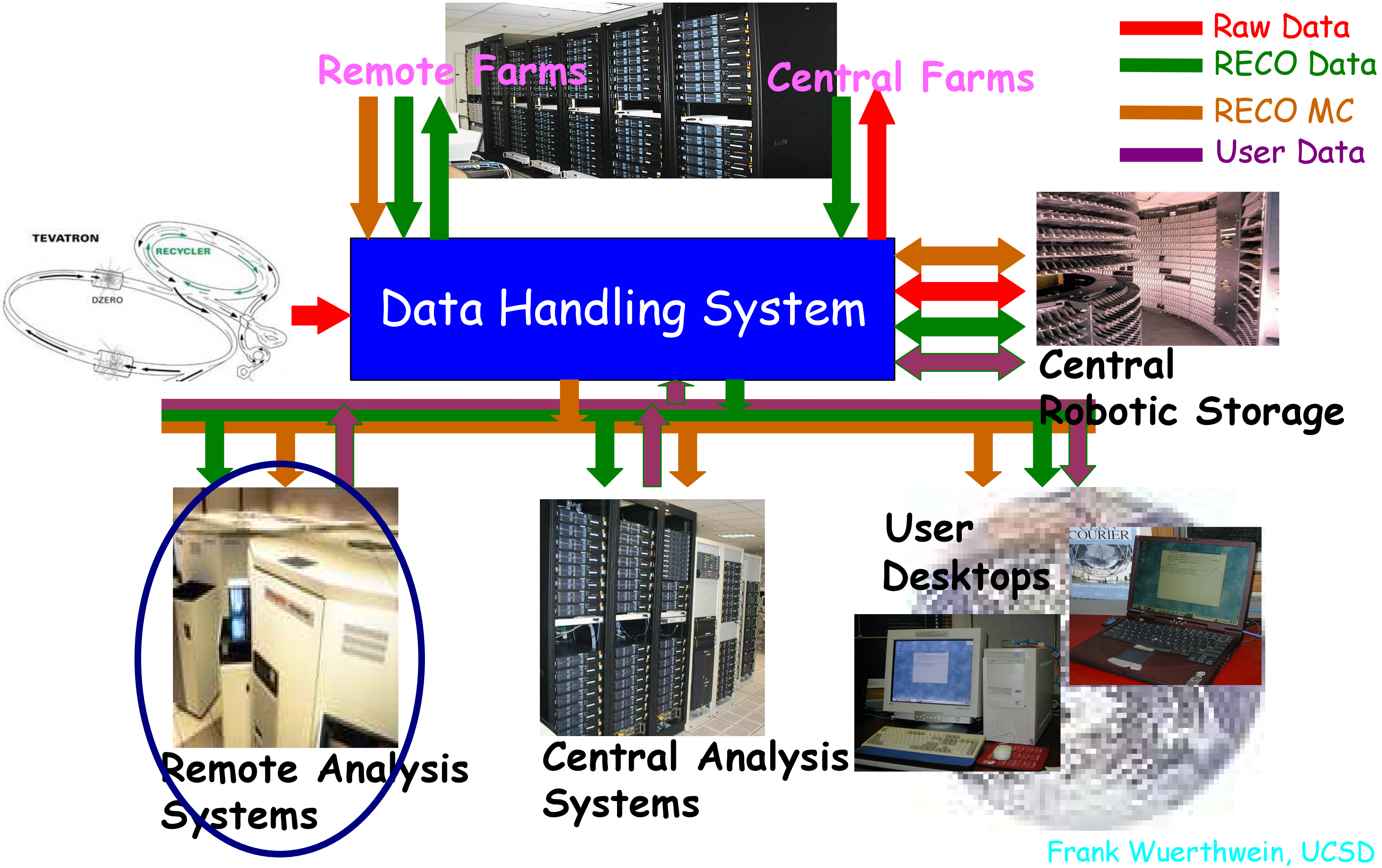
CAB usage in CPU days



Event Size per analysis platform



Computing Model



TEVATRON
RECYCLER
DZERO
COURIER

🟥 Raw Data
🟩 RECO Data
🟧 RECO MC
🟪 User Data

Remote Farms

Central Farms

Data Handling System

Central Robotic Storage

Remote Analysis Systems

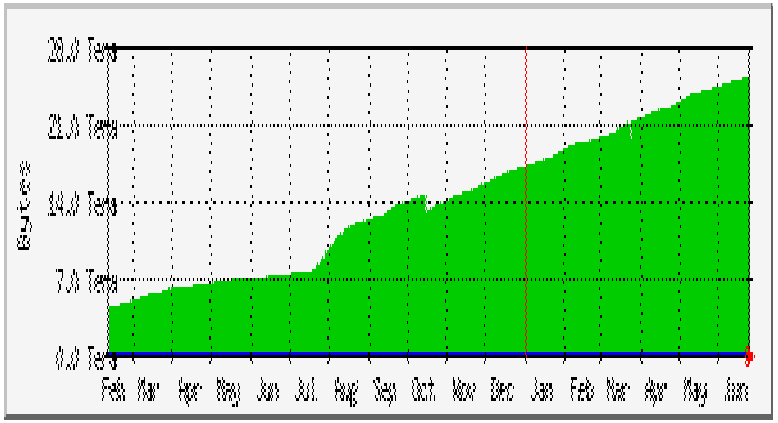
Central Analysis Systems

User Desktops

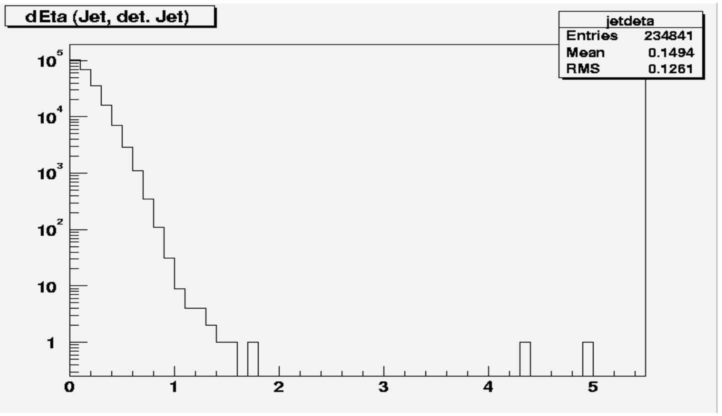
DO Remote Analysis

- Projects run at offsite stations Jan-Apr 2003: 2196 (~25000 projects run on Fermilab analysis stations)
d0karlsruhe,munich,aachen,nijmegen,princeton-d0,azsam1,umdzero, d0-umich,wuppertal, ccin2p3-analysis

- No accounting as yet for what kind of analysis-SAM is the tracking mechanism.



TMB data at IN2P3

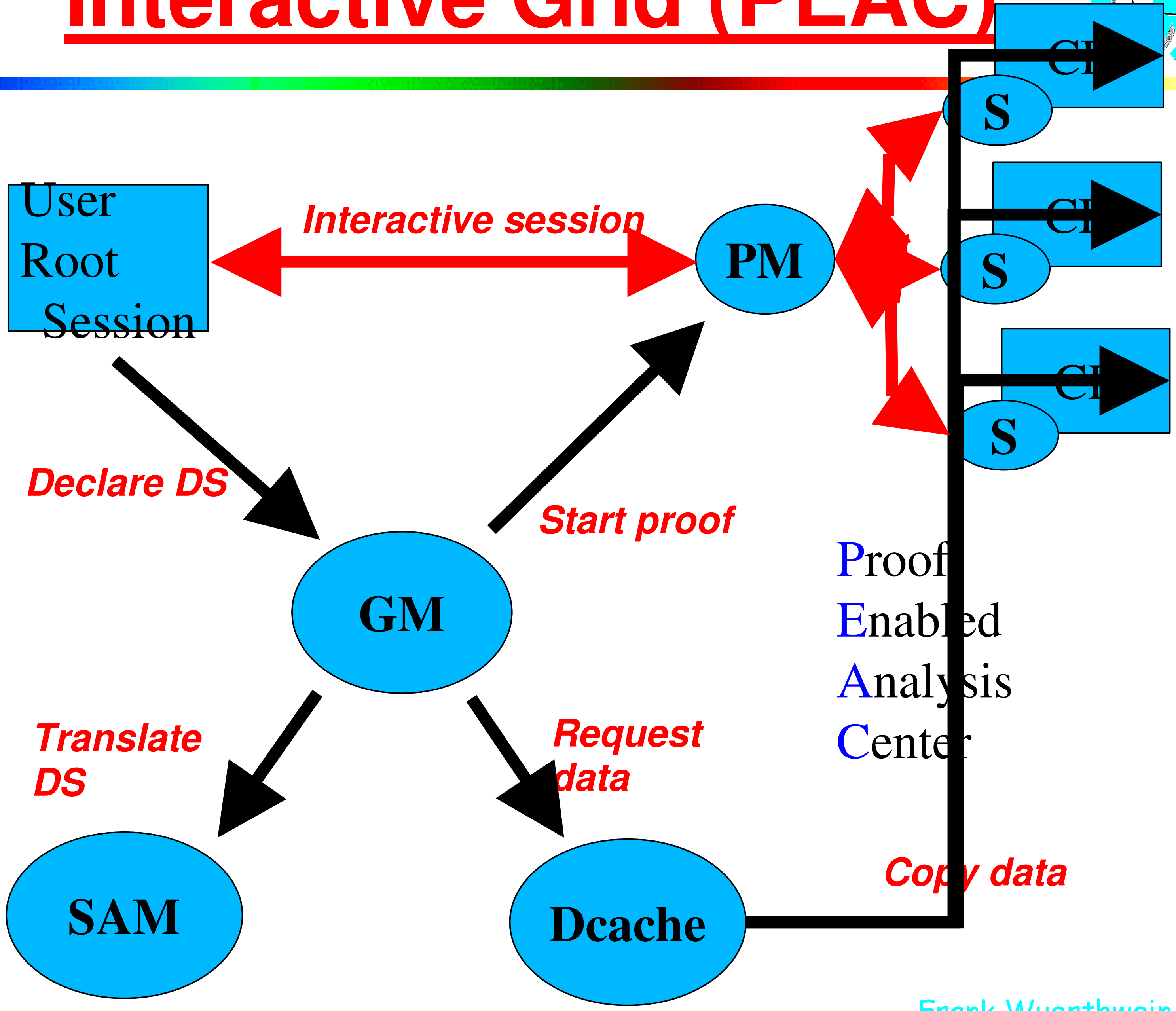


Analysis Verification Plot, GridKa

CDF Remote Analysis

- CAF is packaged for offsite installation.
- 15 CAF : 7 active, 2 test, 6 not active
- Not the only offsite option: GridKa, ScotGrid, various University clusters.
- Still work to do before offsite installations are fully integrated.
- Biggest issues are in support, training, operations.

Interactive Grid (PEAC)


User Root Session
Interactive session
PM
S
S
S
Declare DS
Start proof
GM
Translate DS
Request data
SAM
Dcache
Copy data

Proof
Enabled
Analysis
Center

Summary and Outlook

- Both experiments have a complete and operationally stable computing model
  - CDF focused on user analysis computing
  - DO focused on WAN DH & reprocessing.
- Both experiments refining computing systems and evaluating scaling issues
  - Planning process to estimate needs
  - DO costing out a virtual center to meet all needs
- Strong focus on joint projects: SAM, dCache
- Medium term goals: convergence with CMS.
- Open Science Grid