

CDF Computing: Core Technologies - Status and Plans

Robert D. Kennedy

Fermilab, Computing Division
CDF Data Handling Co-Leader

11 September 2003, v1.2

Fermilab Director's Review of Run 2 Computing

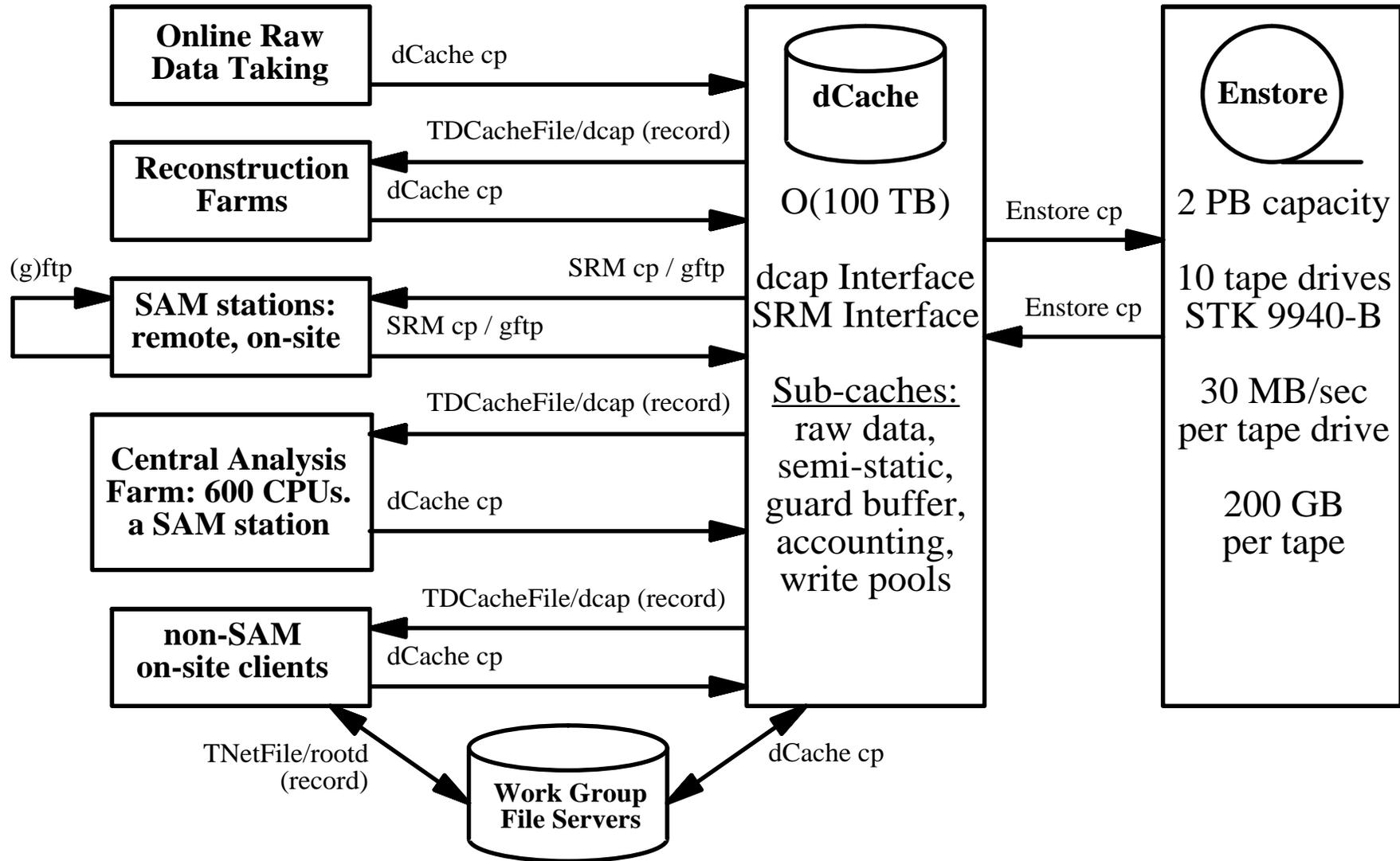
- ⇒ **Focus: Status, Plans, and "Will these technologies scale?"**
- ⇒ **Data Handling (emphasis here): Enstore, dCache, SAM, DFC, PNFS**
Scaling behavior of more concern
- ⇒ **Analysis Facilities (others): Databases, Login Pools, CAF, Global Computing**
Scaling behavior of less concern

DH Current Status: Overview

- 1) **Enstore: Mass Storage System. In Production. Runs on dual STK Powderhorn 9310s, 10 STK 9940B drives.**
 - 2) **dCache: Network-accessible Disk Cache. In Production. 68 TB of space in various sub-caches. (Soon: 93 TB).**
 - 3) **SAM: Data Handling Framework. In use at CDF. Long-term schema migration and Framework adaption in progress.**
 - 4) **Data File Catalog: Datafile Meta-data in RDBMS. In Production. DFC schema to be replaced by corresponding SAM schema.**
 - 5) **PNFS: Meta-data underneath Enstore, dCache. In Production. Database that looks like Unix filesystem, from DESY.**
 - 6) **Networking, ...: Overall smooth operations.**
- *) Many thanks to CD, DESY, ROOT, D0 for work/help/support.**

CDF DH Baseline Goal FY2004

CDF Computing: Core Tech
Robert D. Kennedy
11 Sept 2003
page 3



Underlying Meta-data: SAM Schema, PNFS (dCache, Enstore)

Baseline Goal FY2004: Some Tasks

CDF Computing: Core Tech
Robert D. Kennedy
11 Sept 2003
page 4

- ⇒ **CDF Enstore fully migrated to use of 9940B tape drives**
- ⇒ **CDF dCache fully implemented, O(100 TB), with sub-caches**
- ⇒ **All Enstore access goes through dCache (read .and. write)**
- ⇒ **Tapeless data paths for produced (and later raw) datasets**
- ⇒ **Fully adapt to SAM schema; drop DFC schema, keep API**
- ⇒ **CAF adapted to become a SAM station w/direct dCache access**
- ⇒ **Simplified "entry-level" SAM input for users (carrot, not stick)**
- ⇒ **Robust operations, doc'd procedures, user experience issues**
- ⇒ **Work to begin: CAF output catenation, unified data processing (ntuples supported at same level as EDM format).**

Enstore: Status and Plans

CDF Enstore: smooth operations in the past year
depends on PNFS service
maintained by CD, activities in coll. w/CD-ISA

- 1) **9940A to 9940B Format Migration: achieves 2 PB capacity**
Status: O(>50%) done, now running smoothly. Schedule driven by need for recycled tape capacity... not "in a hurry".
 - a) CD-ISA migrates data from A format tapes to B format tapes.
 - b) Then, recycles A tapes whose data have been migrated to B format.
 - c) CDF writes all new data, raw and produced, to B format.

- 2) **Low "technology" risk into FY05, but very noticeable costs**
 - a) Have not yet filled out the existing robots with tape media to achieve 2 PB.
 - b) Tape I/O demands are motivating acquisition of additional tape drives/movers.

- 3) **Plan: "X" technology in FY05 with 2x 9940-B cost effectiveness.**
 - a) Do not have substantive experience with "X", so unsure will work on time in field.
 - b) Fall-back: more robotics and 9940-B media: costly, but low risk.

dCache: Status

CDF dCache: declared in production June/July 2003
relatively smooth operations since then
depends on PNFS service
maintained by CD, activities in coll. w/CD-CCF

dCache = Network-accessible disk cache as front-end to mass store.

- a) Expects reliable network, no integrity checks. Oriented towards on-site access.
- b) TDCacheFile (Root class): easy transition from local file to remote dCache access.
- c) MUCH effort invested to achieve stable operations at scale of CDF CAF load.
- d) Sub-caches: distinct sub-units to separate "cycling" datasets from stable ones.
- e) More product development required to achieve CDF baseline, almost there though.

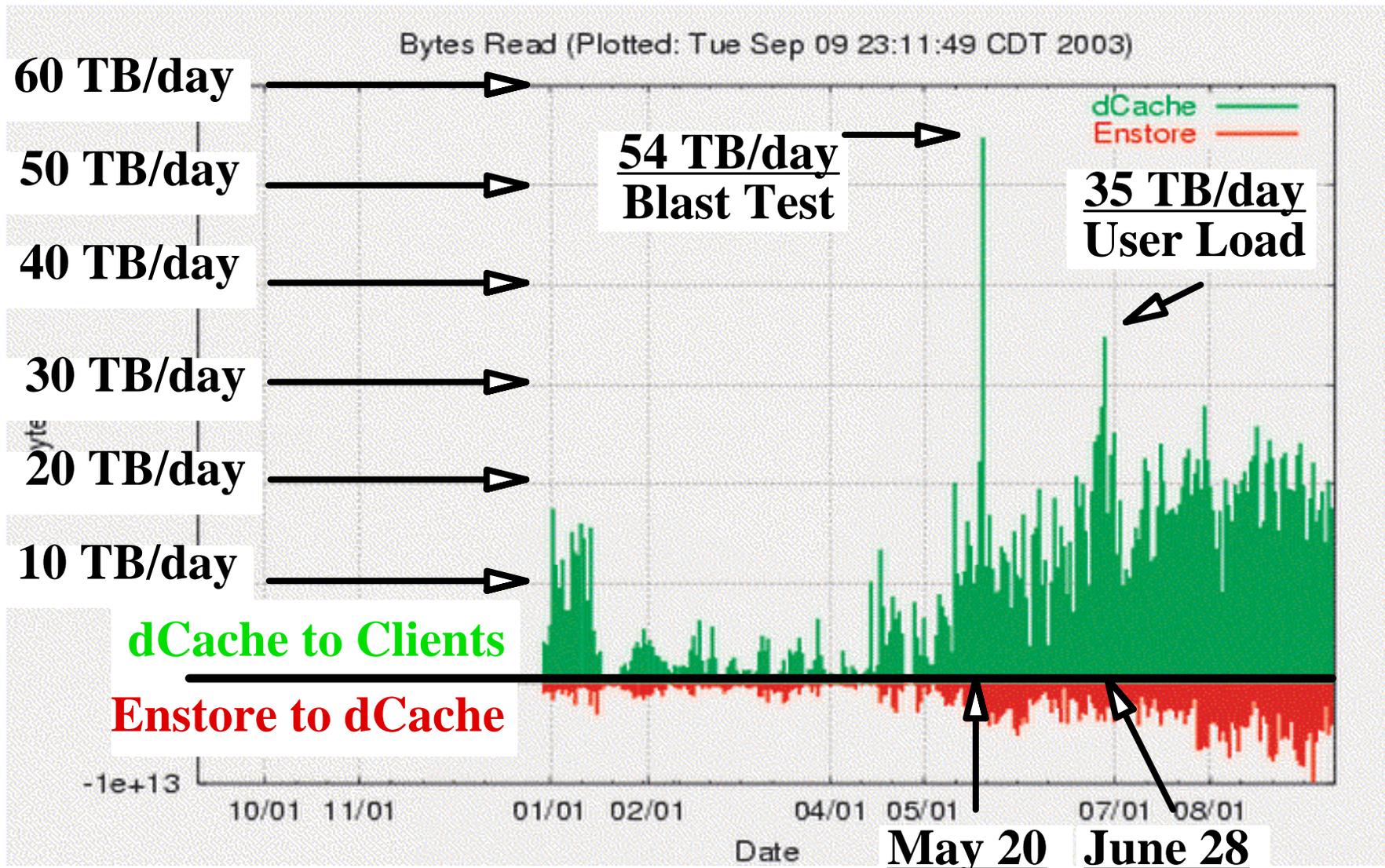
1) All CDF data accessible to all CDF clients via dCache.

- a) Legacy caching systems being absorbed or discontinued soon.
- b) Good experience in past months: 1 interruption (PNFS), 1 logging data drop-out
- c) Cache space, tape access demand: backlog of tape restore requests do occur.

2) dCache can support any data file format, including ntuples.

- a) Need a meta-data system integrated with Root analysis framework to proceed.
- b) dCache has advantages over Root Netfile service and manual file maintenance.

dCache: Bytes Read/day in 2003



<http://cdfcam.fnal.gov:8090/dcache/outplot?filename=billing-2003.05.daily.brd.png>

dCache - Scaling

CDF dCache: Proven at today's (CDF CAF) scale, but work is needed at each ratchet step of scaling to prove again.

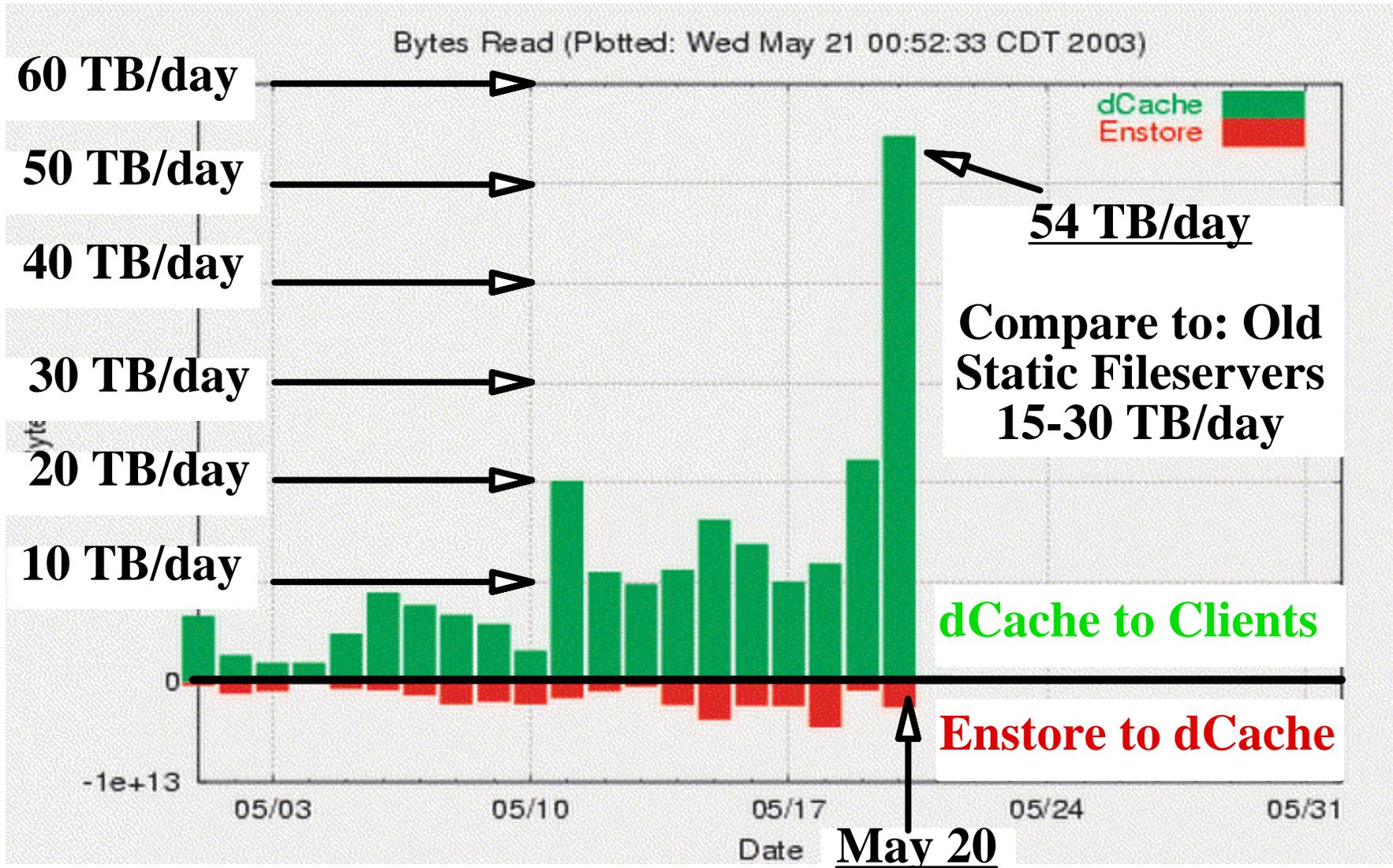
- 1) Blast/Super-blast tests: N clients reading CDF data *fast***
 - a) Stresses number of simultaneous clients handled and data I/O handled.
 - b) Stresses number of file-in-cache queries.
 - c) Does not stress system with many file restore requests.

- 2) dCache interface to PNFS: will need improvement.**
 - a) Super-blast test: PNFS request back-up inside dCache (cfr. CDF 6672)
 - b) File-in-cache queries can sometimes take O(1 minute) per file.

- 3) Experience: CAF CPU idle while waiting**
 - a) Does not happen often due to pre-fetching. Cache hit/miss ration monitored.
 - b) When it does happen, first file can take hours to be delivered: idle CAF CPU.
 - c) Fix == SAM. Sam coordinates data delivery with job execution.

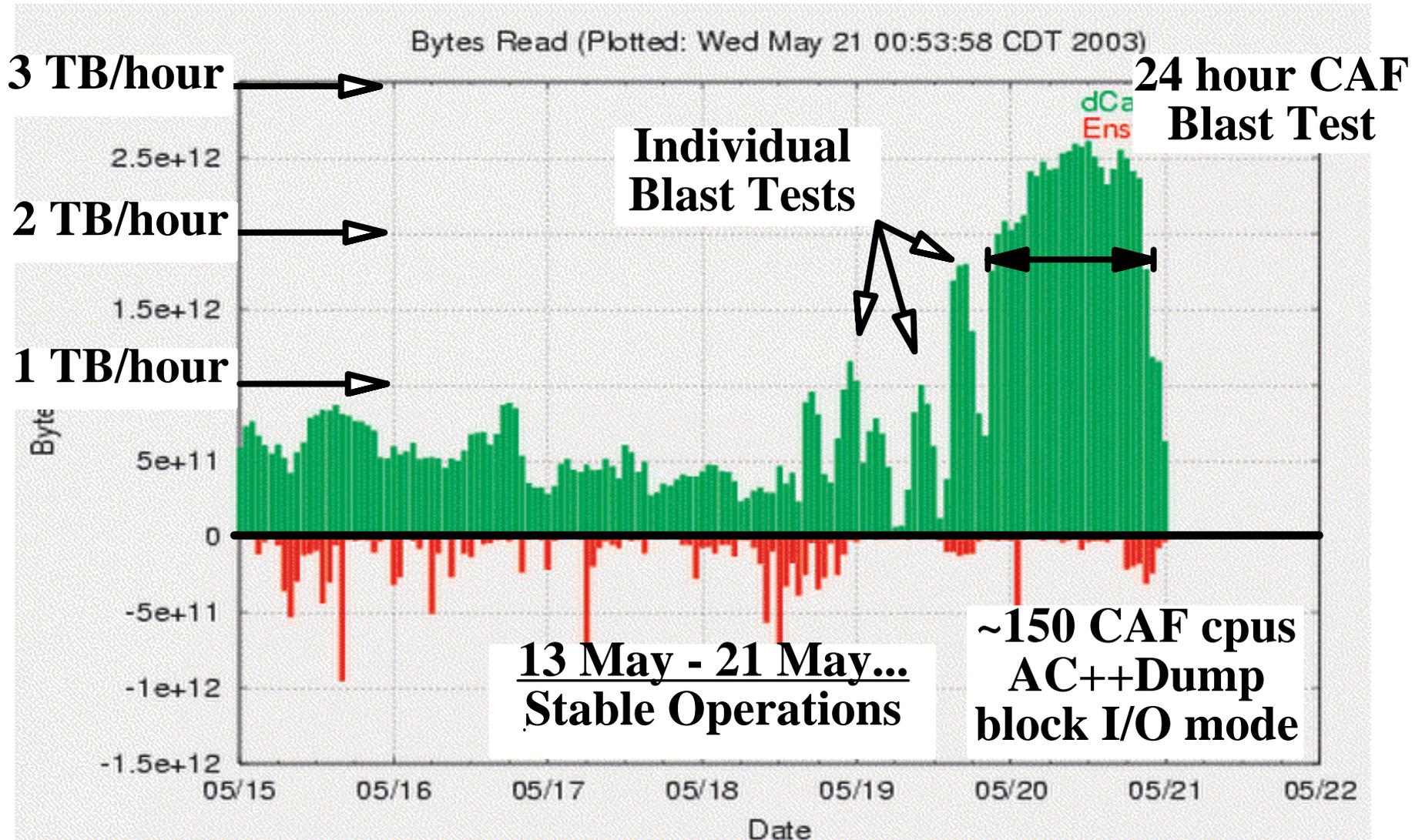
- 4) DH Issue: Large or deprecated datasets, tape access backlogs**
 - a) Management now requires "policing" rather than automated policy.
 - b) dCache sub-caches help, but cannot be adapted quickly as demands shift.

"Blast" Test: Bytes Read/day



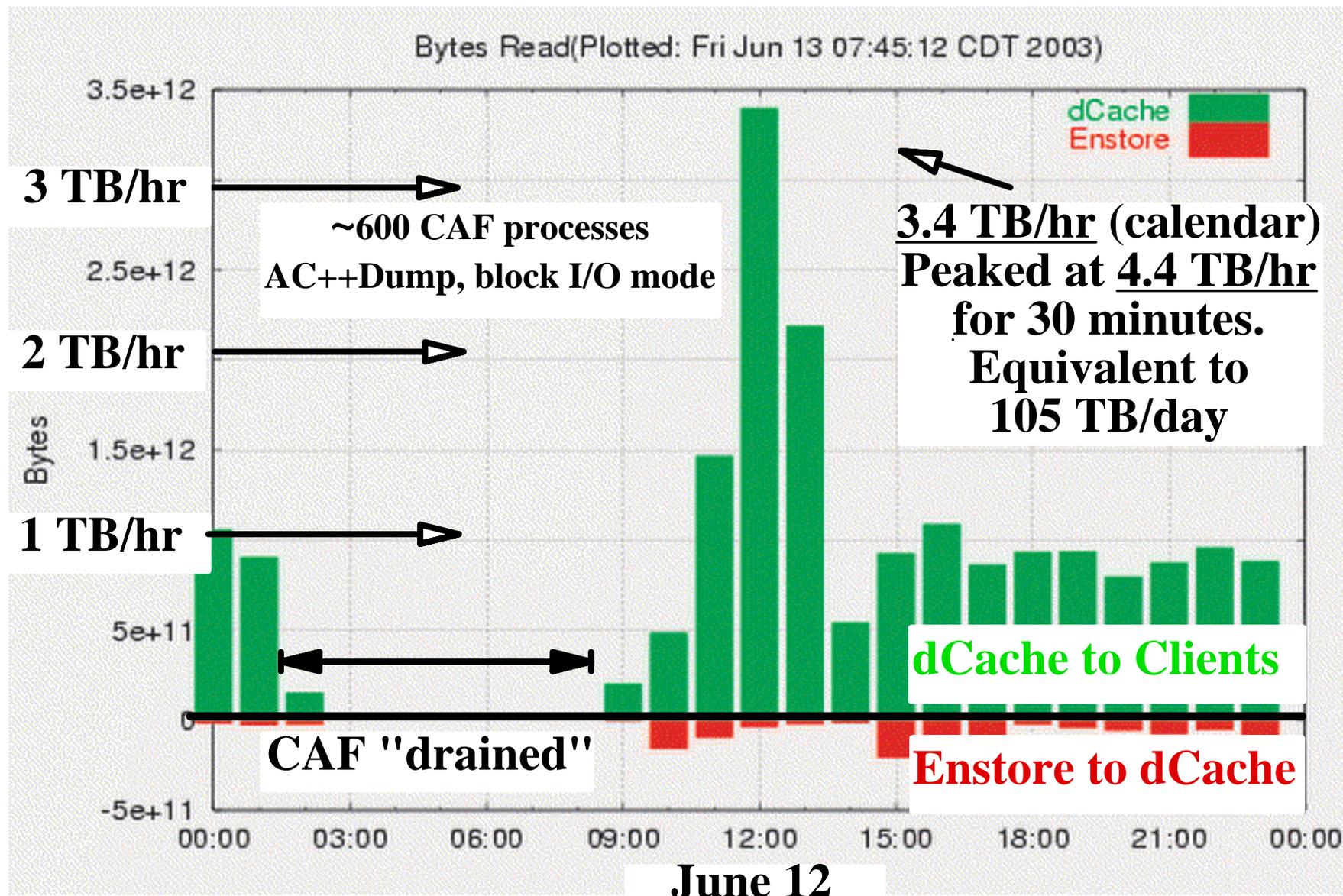
<http://cdfcam.fnal.gov:8090/dcache/outplot?filename=billing-2003.05.daily.brd.png>

"Blast" Test: Bytes Read/hr



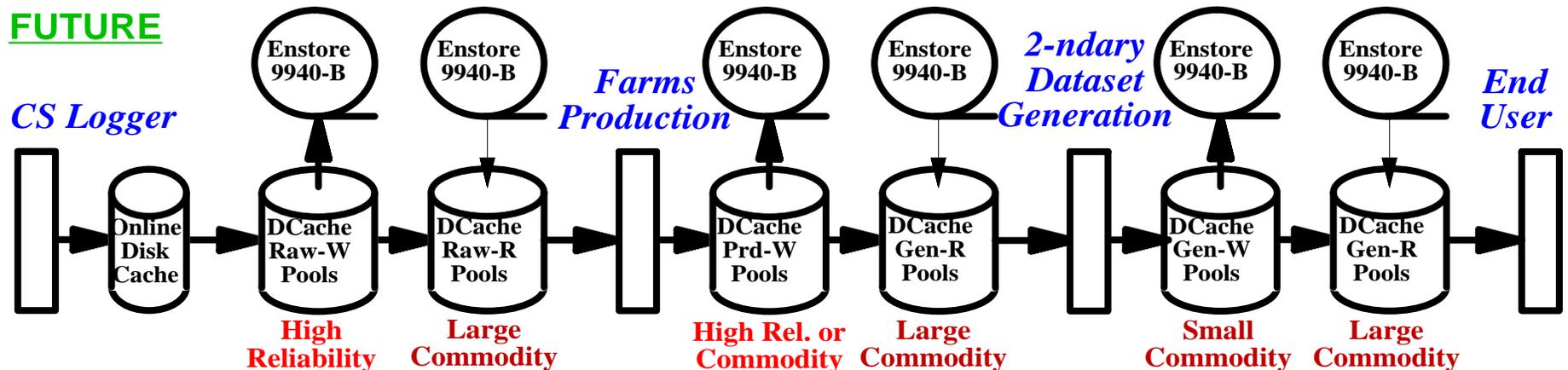
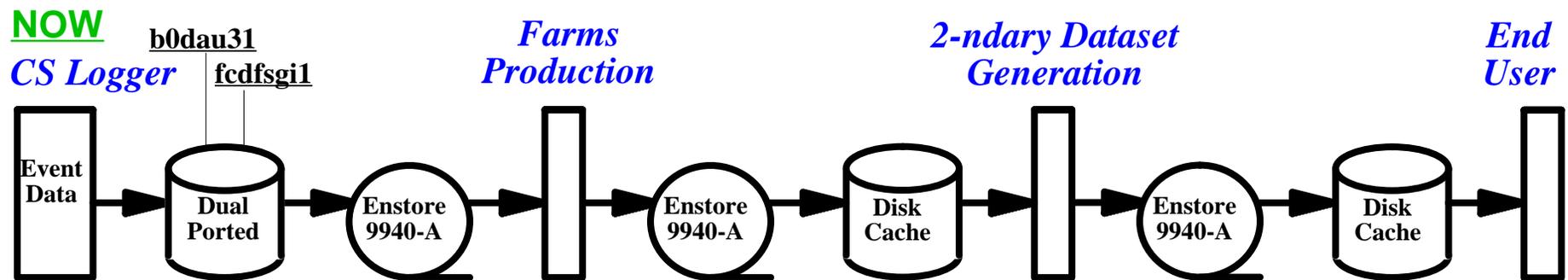
CDFDCA: 24 file servers, non-CAF and 50% of CAF have access. 100% of blast datasets on disk.

"Super-Blast" Test: Bytes Read/hr



dCache: Plans

- **Krb5 authenticated dcap doors: secure writes, track users**
- **Write pools + pool-to-pool copying: tape out of critical path**
(Upgrade of Offline side of Raw Data Logging: write to dCache)
- **Improved error reaction, CDF-oriented monitoring interfaces**
- **Longer-term: "scratch space" for CAF output catenation**



SAM - Status at CDF

CDF SAM: in use, mostly at off-site institutions
joint project with CD, D0, and more recently CDF

SAM = a Data Handling Framework.

- a) Has datafile integrity checks. Means for CDF support of remote data access.
- b) More extensive meta-data catalog functionality. Consistent "write" facilities.
- c) Combined development eliminates redundant solutions, reduces costs.
- d) Clear path to future GRID-supported tools and global computing environment.

1) Existing CDF SAM adaptation is in use for physics.

- a) SAM in regular use at CDF institutions in UK and at Karlsruhe.
- b) Use with MC production now being tested.
- c) Good experience overall. Working on versioned configuration management.

2) Much invested in a common meta-data schema. Migrate to it.

- a) CDF SAM meta-data now updated parasitically from DFC. Maintenance-heavy.
- b) DFC API can be implemented by new SAM meta-data, so swap-out possible.

3) Improvements in CDF Infrastructure interface to SAM.

- a) File handling in SAM and AC++ have subtle differences: can be overcome.
- b) CDF multi-branch datafiles: avoid whole-file cache transfers where possible.

SAM: CDF Migration Plans

CDF Computing: Core Tech
Robert D. Kennedy
11 Sept 2003
page 14

Plan: Re-implement existing DFC API using SAM meta-data

Goal: Early October 2003

- 1) Adopt new joint schema (5.1) in production, test on SAM
- 2) DFC "write" interface to fill both DFC and SAM meta-data.
 - *) Now being done for production farms output
- 3) Switch DFC user API to use SAM meta-data instead of DFC
- 4) More little issues... at the end of a long, complex project.

Plan: Allow users to easily switch to SAM use on the CAF.

Goal: October 2003

- 1) CDF Framework access to data via SAM - simple user switch
 - *) Largely done. Some more testing required.
- 2) Operate a CAF system as a SAM station - waiting on...
- 3) ESM: Use dCache as SAM cache instead of SAM's cache system.
 - a) Goal: Where network is "perfectly reliable", allow consumers to directly access data files in dCache. Reduces nBytes transferred (ROOT multi-branch)
 - b) Mechanism shown to work with HPSS, to be tested soon with dCache.

Databases: Status and Plans

For the story well-told: See Nelly Stanfield & Lee Lueking's talks

Status: Stable operations in the past year

- a) Extensive monitoring added to facility.
- b) Replica introduced to isolate CAF user load from Farms load on database system.

Approach: CDF uses direct client connection with metering.
(in other words, not a multi-tier architecture)

Scaling: CDF uses replication at present.

- a) Multi-tier architecture not ruled out.
- b) Free-ware databases being investigated.

CDF Analysis Facilities

For more details: See Frank Wuerthwein's talk

1) Linux-based Login Pool as central interactive facility

- a) Replaces central IRIX SMP primarily.
- b) Relatively low-risk commodity technology.

2) SGI IRIX SMPs to be phased out

- a) Legacy disk cache system to be turned off.
- b) Other data services (rootd) easily replaceable with commodity file servers.
- b) Initial reduction of nCPUs in half, then....

3) The CAF... see the next few slides for scaling issues and plans.

4) Off-site Computing Facilities, Global Computing (Frank's talk).

- a) Large (Linux) clusters at several institutions, exploit for MC production now.
- b) DH and/or GRID framework desirable to exploit with little labor
- c) Must exploit to maintain scaling of processing capabilities throughout Run 2.

CAF: Scaling

For CAF Overview: See Frank Wuerthwein's talk

CAF = Batch-oriented Computing based on Commodity PCs.

***) Organized as a "transplantable" product: remote CDF CAFs exist and are in use.**

1) CAF Infrastructure s/w upgrade in 2003 to insure scalability.

a) Multi-process sections. Bmgr handles sections rather than processes.

b) Nprocesses/section grows with nCPUs to insure scalability.

2) CAF User & Software support: scales with nUsers, not nCPUs.

3) CAF Hardware: Not expected to be an issue.

a) FY03 purchases are Intel-based, expected to be more reliable than current AMDs.

b) Expect no problem in FY04 unless hardware much less reliable than existing.

c) Re-evaluate situation yearly!

4) CAF System Failure stats acceptable, may need improvement.

a) Currently 1/3000 user jobs fail due to hardware or system problems.

b) Automatic re-submission or more reliability may be needed by FY05

CAF: Plans

For CAF Overview: See Frank Wuerthwein's talk

1) Condor-CAF: Development started.

- a) Close collaboration with Condor team. Kerberos-aware Condor since last week.
- b) Need significant re-implementation of CAF user monitoring to work with Condor.

2) Usage Monitoring Improvements: Being implemented.

- a) CPU time (per event) per dataset, I/O per dataset, DH response time.
- b) Independent accounting of CPU consumption for MC and data analysis.

3) Hardware DB: Advanced development stage.

- *) Tracking of hardware failures. Deployable off-site as part of CAF system software.

4) Generalization of Admin scripts: Not yet started.

- a) Few FTE-months once design is clear.
- b) Needed to improve off-site CAF operations support.

5) "Gridification of services": Not yet started.

- a) Several FTE-months once design is clear.
- b) Needed for CDF GRID vision.

Conclusion

Core Technologies: Challenges still ahead. Each component must be proven at each new scale of load. The core components of Run 2 CDF Computing are (almost) in place, ready to grow.

Enstore: Stable. Should easily scale.

dCache: Stable. Scaling issues will require work, careful testing.

SAM: Long-term framework poised to be fully adopted by CDF.

Databases: Stable. Scaling issues being addressed.

Interactives: Low-risk solution to be tested soon to replace SMPs.

CAF: Stable. Ready to scale to a much larger system, if DH can.

Global Computing: SAM+CAF+GRID in the works to integrate off-site facilities into one distributed CDF Computing system.