DO Computing Model and Plans

Amber Boehnlein

Bird Review Committee

Sept 12, 2003

Charge To Committee

The CDF and D0 experiments are asked to together propose and arrange a series of talks from experiment and computing division personnel that present

a) the current status of the computing systems and how they operate, both at Fermilab and worldwide, to enable the experiments to collect, store and analyze the Run II data.

b) the experiment requirements and proposed computing model for the next 3 years, together with the estimated costs at Fermilab in terms of both equipment and manpower.

c) the agreements in place by collaborating institutions to provide either manpower or services that the experiment relies on for some part of the processing and analysis of data.

This talk addresses the operational status of the systems, cost estimates for the computing model and the management of a global model for the DO experiment.

Use current operational understanding to make estimates

Computing Status

DO has a highly successful computing structure in place

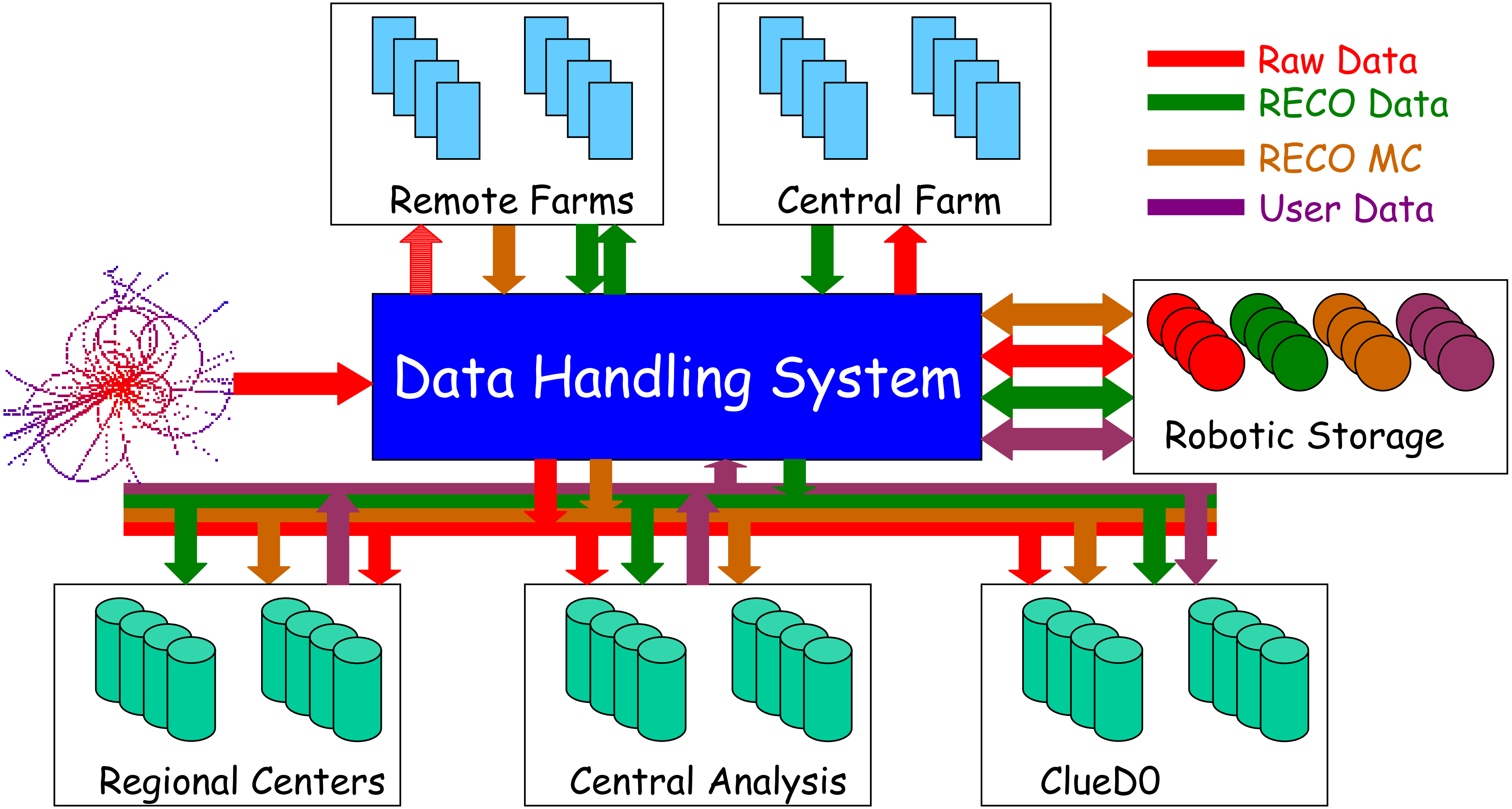
2002 Status 2003 status

- Sequential Access by MetaData (SAM) catalogs and manages data access-Operations are stable, schema evolution, users inserting skimmed data sets.

Robotic storage with reliable drives and media—9940B in production, will migrate to reuse tapes

- Domino provides high I/O capacity and user access to large amounts of data-Reduced to 128 processors
- Commissioning the commodity backend In production
- Basic software infrastructure in place—In need of attention
- Upgrading processing farm:Completed—new purchase
- Fruitful collaboration with the Computing Division on joint projects—Continued SAM improvements, tests of JIM
- MC generation performed at collaborating institutions-Now doing analysis, and reprocessing
- DORECO has basic functionality—Improved tracking, use of calibration constants from data starting
- Basic Filtering at L3—In need of attention
- Online output rate is at design.

DØ Data Flow



Amber Boehnlein, FNAL

Towards a Global Model

- Two planning efforts—Regional Analysis-> Offsite Analysis Task Force
  - Computing resources and infrastructure are only one aspect of effective offsite analysis
- Effectively integrating hardware resources requires structural changes and additional effort (management and technical)
  - The Computing Planning Board now has different composition and is charged with focusing on global issues.
  - We must develop a support model in which developer and operations effort is supplied in conjunction with the hardware to make the global model a success—with support also covered by MOU.
  - We are increasing focus on aligned activities within GRID projects.
  - We are track available resources and determine the deployment as best meets strategic and tactical needs
- Financial considerations have to be addressed—use computing contributions to offset common fund contributions, start from BaBar Model, details are still being settled
- Track total estimated needs and value of contributions—use and extend planning spreadsheet.
- Operational contributions also important and valued.

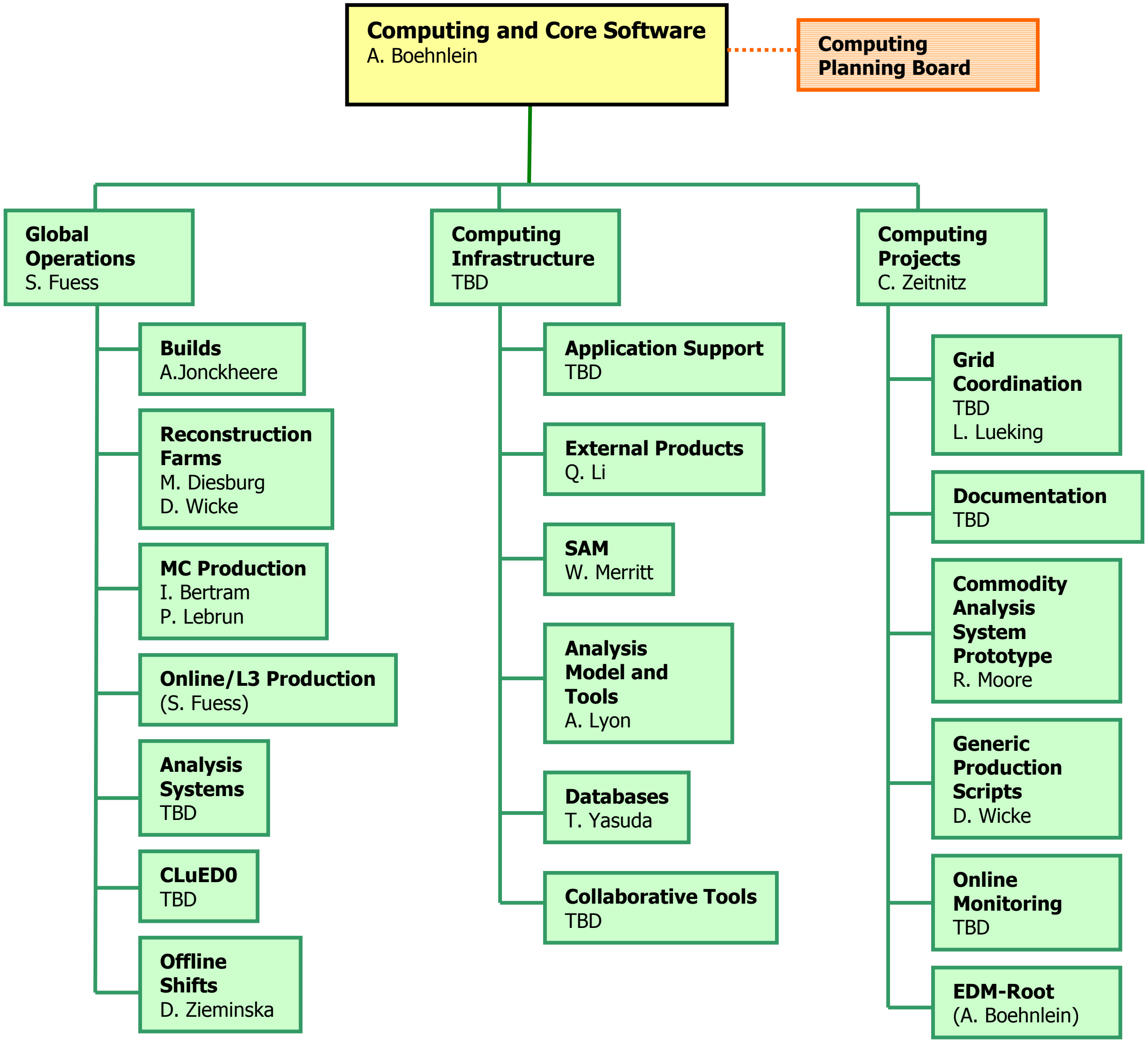
The Virtual Center

- For cost basis, determine the cost of the full computing system to met all needs as if the center were located at FNAL, plus equipment required to support offsite
- Presented today as draft
  - Disk and servers and CPU for analysis
  - Production activities such as MC generation, processing and reprocessing.
  - Infrastructure such as gateway machines and code servers
  - database machines and servers
  - Mass storage
  - Cache machines and drives to support extensive data export
- Not included as a cost estimate, but vital
  - Wide Area Networking
  - Desktop computing

DO Computing Management

- DO Computing and Core Software Management reflects global nature of computing on DO
- Computing Planning Board
  - administers “Virtual Center”, MOUs
  - Serves as point of contact for
    - FNAL CD
    - Centers (which can have their organizational structure)
    - DO Collaboration—large dynamic range of skills and views
    - External agencies
  - Makes strategic recommendations
  - Oversees planning exercises
  - Current Membership
    - Amber Boehnlein, Chip Brock, Gavin Davies, Laurent Duflot (Algorithms), Greg Landsberg (physics), Peter Maettig, Dugan O’Neil, Andy White

Organization Chart



- **Computing and Core Software**
  A. Boehnlein
  - **Computing Planning Board**
  - **Global Operations**
    S. Fuess
    - **Builds**
      A.Jonckheere
    - **Reconstruction Farms**
      M. Diesburg
      D. Wicke
    - **MC Production**
      I. Bertram
      P. Lebrun
    - **Online/L3 Production**
      (S. Fuess)
    - **Analysis Systems**
      TBD
    - **CLuED0**
      TBD
    - **Offline Shifts**
      D. Zieminska
  - **Computing Infrastructure**
    TBD
    - **Application Support**
      TBD
    - **External Products**
      Q. Li
    - **SAM**
      W. Merritt
    - **Analysis Model and Tools**
      A. Lyon
    - **Databases**
      T. Yasuda
    - **Collaborative Tools**
      TBD
  - **Computing Projects**
    C. Zeitnitz
    - **Grid Coordination**
      TBD
      L. Lueking
    - **Documentation**
      TBD
    - **Commodity Analysis System Prototype**
      R. Moore
    - **Generic Production Scripts**
      D. Wicke
    - **Online Monitoring**
      TBD
    - **EDM-Root**
      (A. Boehnlein)

DO Computing and Analysis

- CD department for DO computing
- Responsible for data handling and analysis tools, farm production and some database support
  - 2 full time SAM developers/operation + 1 SAM project manager
  - Associate Scientist and Research Associate—analysis tools and physics
  - 1 full time farm operations
  - 1 full time database applications
  - 1 DO and CD management
  - 1 full time physics analysis (rotating position)
  - 3 people who put research fraction into DO
  - 1 build manager
- DO Computing Systems—four people on pager rotation for the analysis systems
  - 24/7 support for SGI systems
  - Minimum people—will be taxed with Linux fileserver transition, dCache

CD Support

- Management of
  - Networking
  - Farms
  - Storage systems
  - Database machines
  - Building infrastructure
- Software and development support for
  - Online
  - Storage
  - SAM-GRID
  - Database application and DBA support
  - DO specific Infrastructure software such as EDM-root project
  - General product support for compilers, linux, CERN products

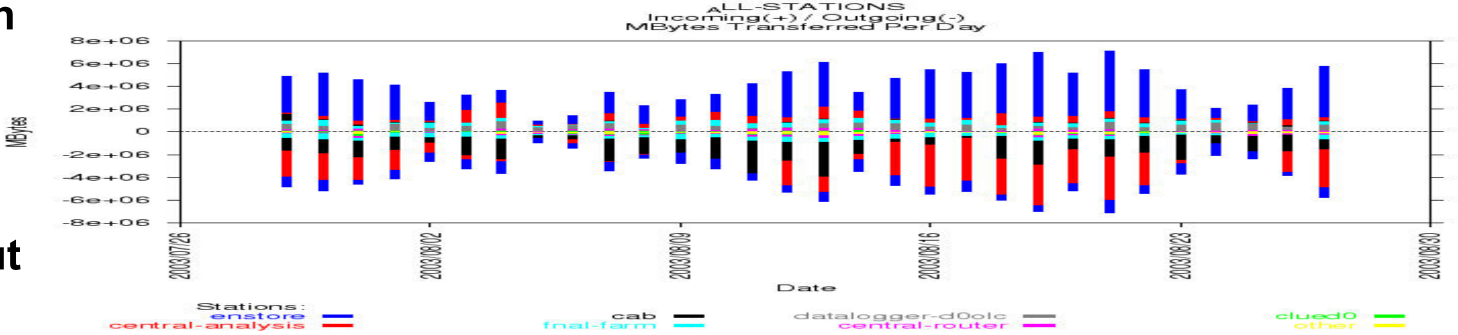
Data Storage & Access

Data to tape as of Sept 8, 2003, ~ 5 GB lost

Library	Data Tiers	Storage
9940A	Raw (May) DST (Feb)	207 TB
9940B	Raw DST	116 TB
LTO	MC Tiers Data TMB	94 TB

Plot shows daily transfers in/out of SAM stations Colors represent stations- Blue is to/from tape and Red is to/from farm

8TB in



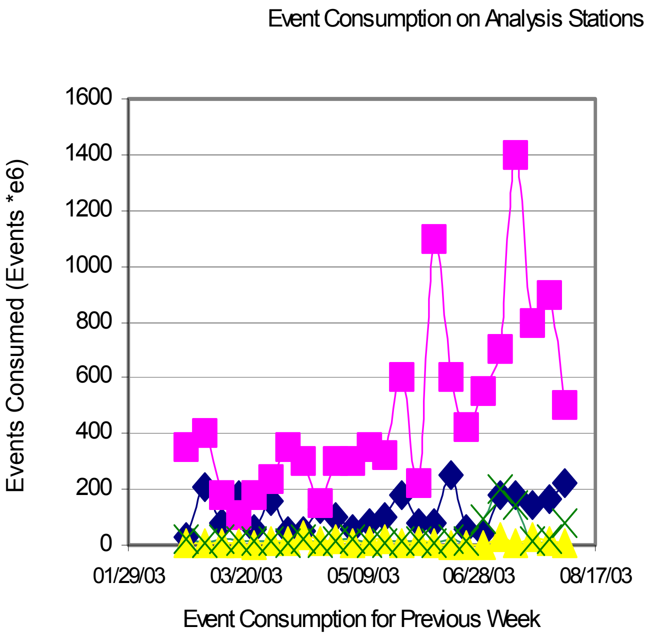
8TB out

August, plotted daily

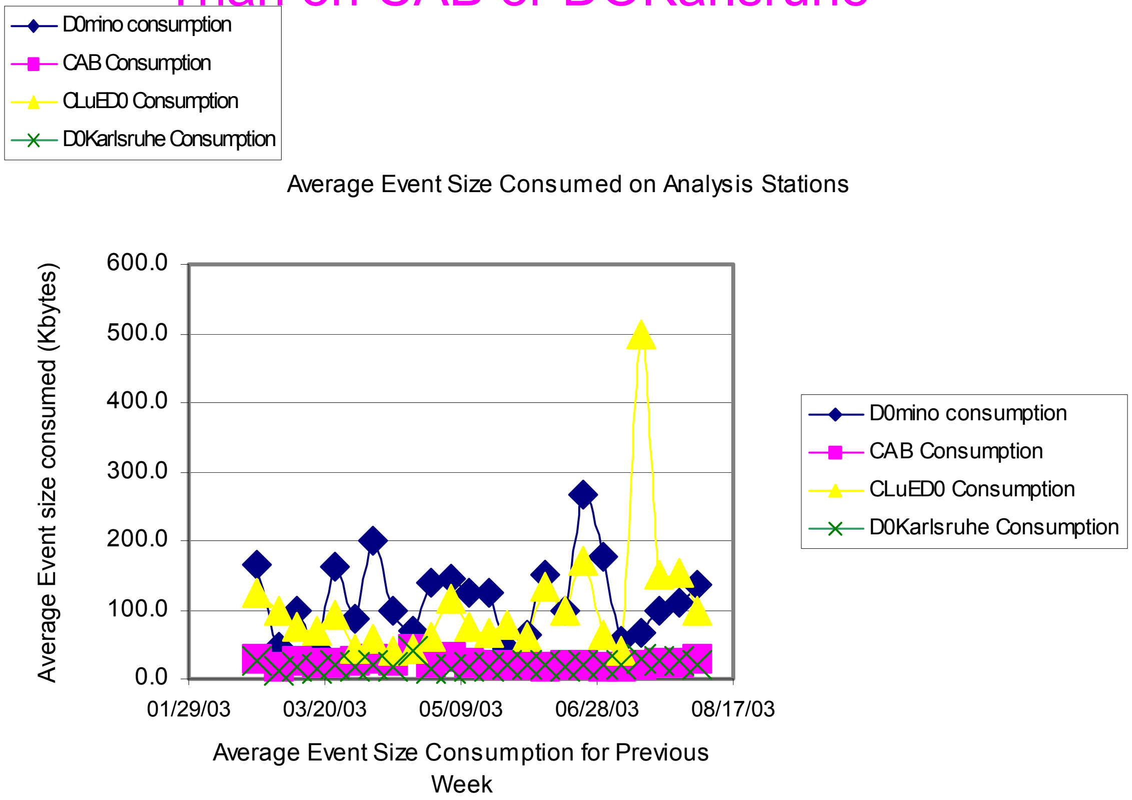
Note: daily transfers doubled Since March

Amber Boehnlein, FNAL

Data Handling Metrics

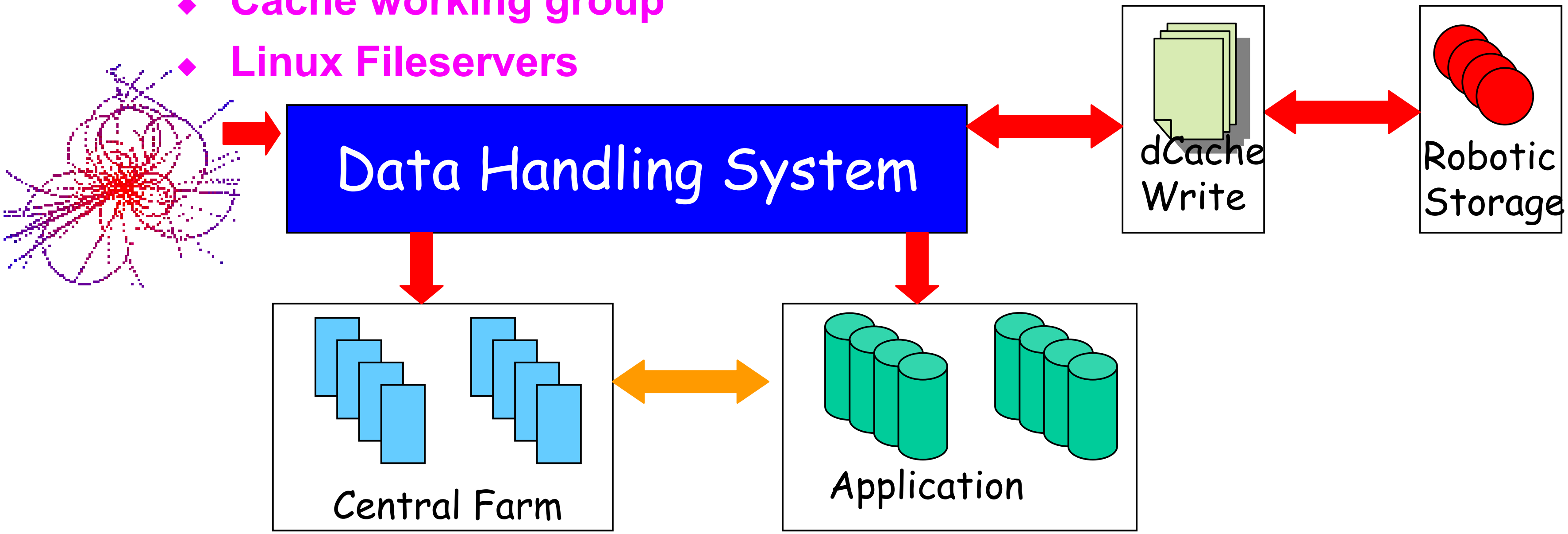


Consumption plot gives information
On analysis uses cases.
Use more varied on desktop, Domino
Than on CAB or DOKarlsruhe



Data Handling, cont.

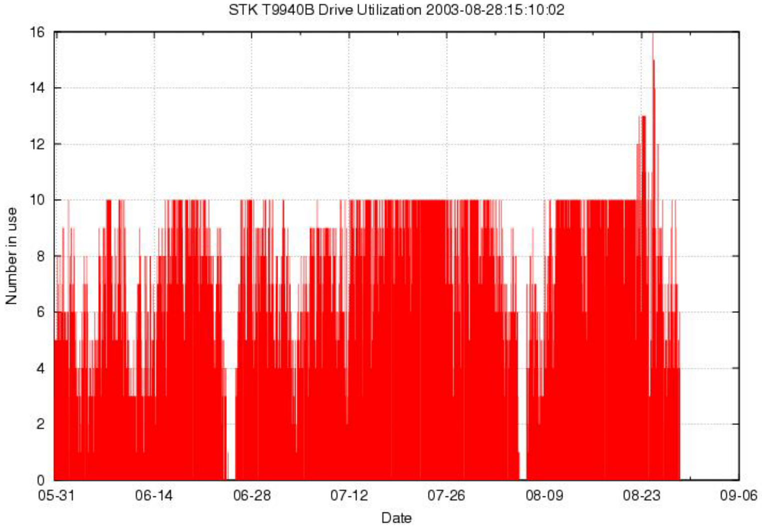
- Improved operations
  - Investigate slow transfers
  - Continue to improve tools, documentation metrics
- Extensions
  - Grid
  - Support for Remote Systems
  - Integrate dCache-first in online for monitor data
  - Cache working group
  - Linux Fileservers



Amber Boehnlein, FNAL

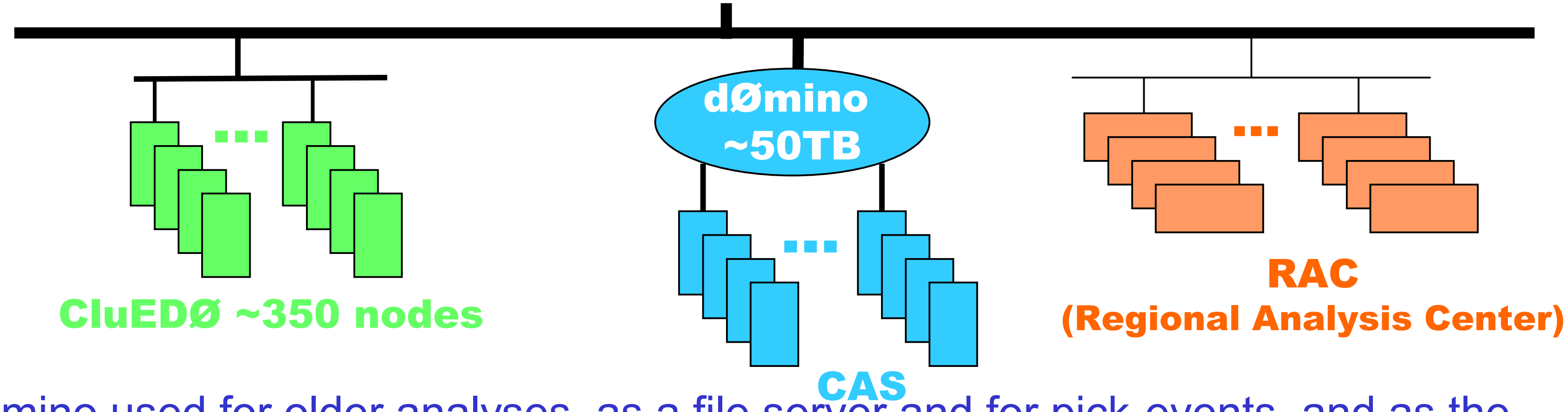
Tape Drives

- In 2004, had estimated that we needed approximately 30 9940B drives
  - 20 in hand—buy 5 more in 2004, use dCache to reduce burden
- In 2005, buy more LTO2 drives and migrate existing LT0 data
- In 2006, replace 9940Bs



Amber Boehnlein, FNAL

Analysis Systems



- Domino used for older analyses, as a file server and for pick-events, and as the central routing station to offsite.
  - In next year, want to transition to linux fileservers, reduce reliance on D0mino
- CLuEDO desktop cluster at DO administered by DO collaborators,
  - SAM station, batch system and local fileservers for analysis.
  - Home areas served from an SGI machine to D0mino and CLuEDO—need faster disk on SGI or replacement system.
  - Management of CLuEDO in transition.
- Central Analysis backend at FCC:
  A PC/Linux dØmino back-end supplied and administrated by the computing division
  160 dual 2GHZ AMD nodes, each with 80 GB disk, works as local SAM cache
- Remote Analysis Centers (RAC):
  Institutions with CPU, disk and personnel resources to serve collaborators

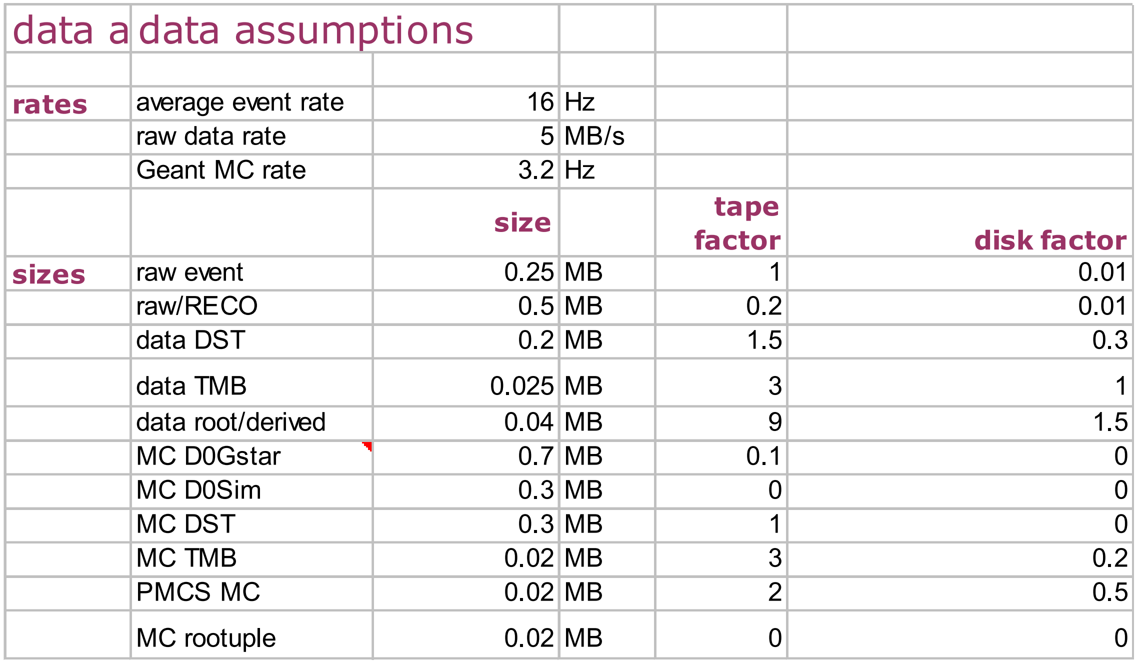
Application & Data Tiers

Parameterized MC time/event	*1 GHz-sec/event*
Full Geant Chain MC time/event	*170 GHz-sec/event*
Reconstruction on collider data time/event	*50 (60,80) GHz-sec/event*
Data DST size/event	*200 Kbytes/event*
Data TMB size/event	*25 Kbytes/event*
MC Døgstar size/event	*700 Kbytes/event*
MC Døsim size/event	*300 Kbytes/event*
MC DST size/event	*200 Kbytes/event*
MC TMB size/event	*25 Kbytes/event*

Assume 16 Hz data collection rate (measured)

Assume that most hardware needs scale with the data collection rate

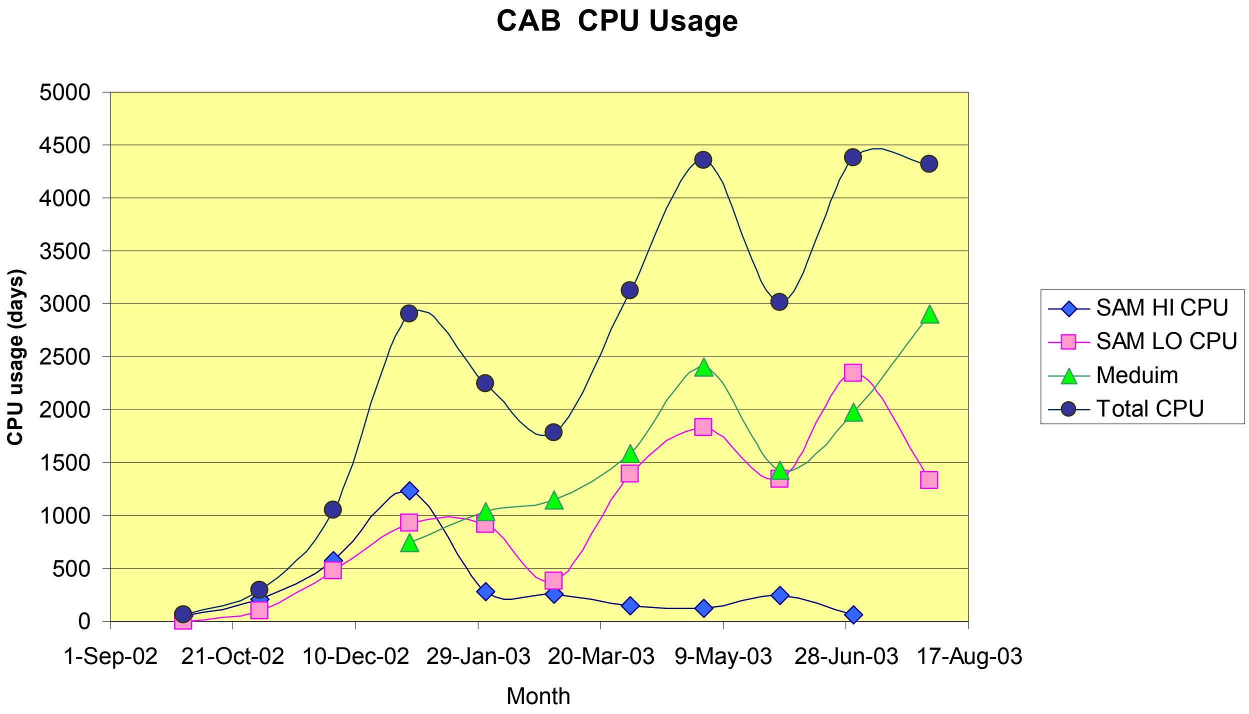
Storage Need Estimate



data a	data assumptions				
rates	average event rate	16	Hz		
	raw data rate	5	MB/s		
	Geant MC rate	3.2	Hz		
		size		**tape factor**	**disk factor**
sizes	raw event	0.25	MB	1	0.01
	raw/RECO	0.5	MB	0.2	0.01
	data DST	0.2	MB	1.5	0.3
	data TMB	0.025	MB	3	1
	data root/derived	0.04	MB	9	1.5
	MC D0Gstar	0.7	MB	0.1	0
	MC D0Sim	0.3	MB	0	0
	MC DST	0.3	MB	1	0
	MC TMB	0.02	MB	3	0.2
	PMCS MC	0.02	MB	2	0.5
	MC rootuple	0.02	MB	0	0

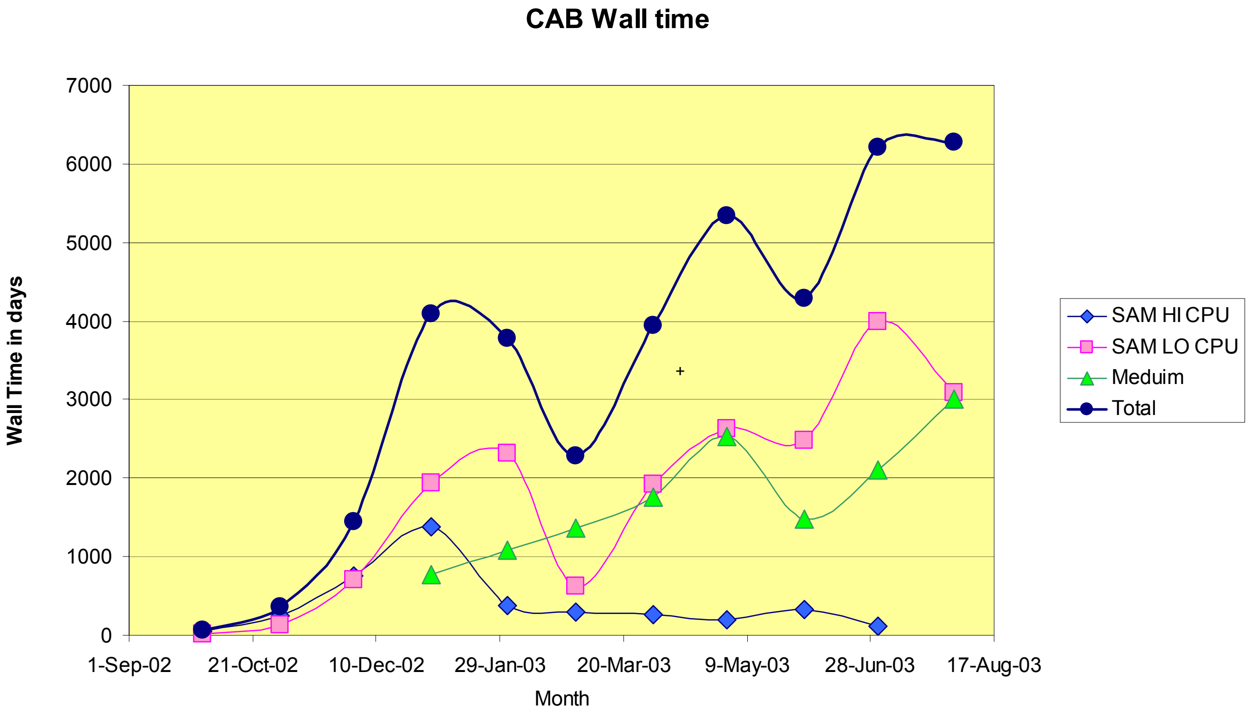
In one data collection year, 800 TB tape storage
85 TB disk for analysis—assume 30% common sample streamed DST on disk <Van Kooten Committee>
Note: Remote Analysis Centers are not included, working to understand use cases.

CAB Performance



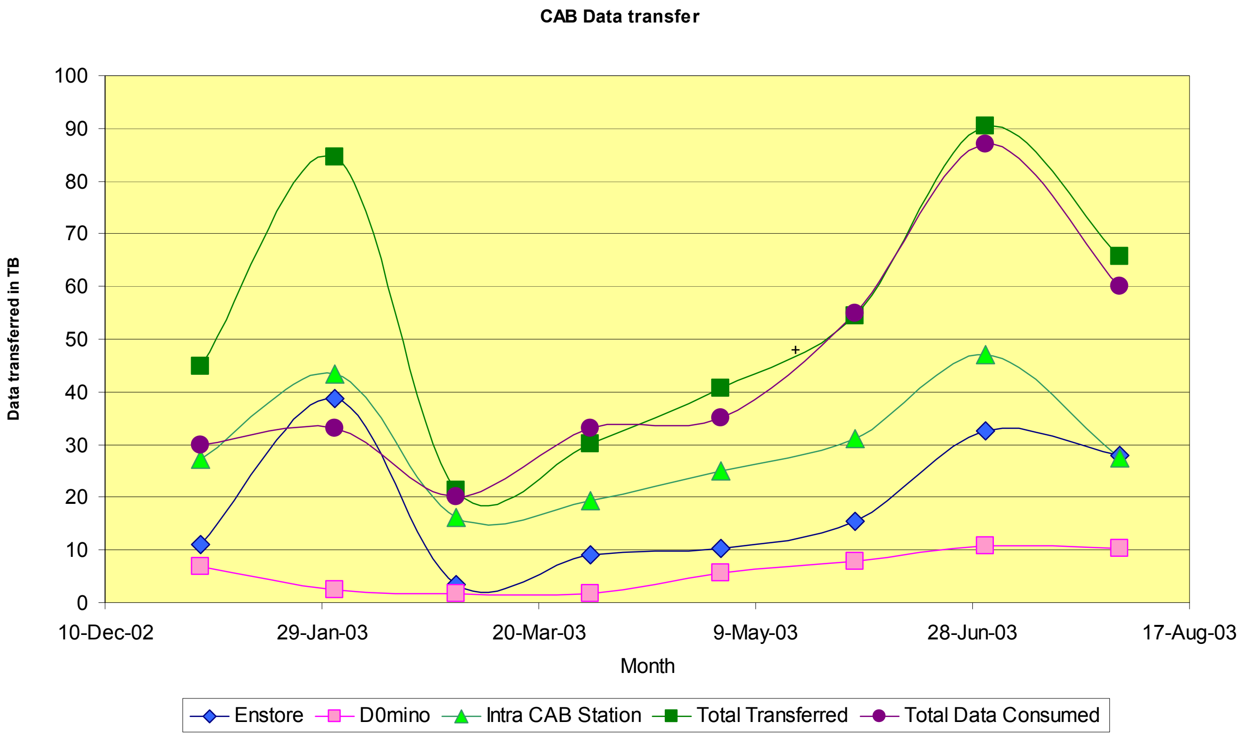
SAM and non-SAM usage shown, for CPU used. SAM queue-6-8 MB/sec
Non-SAM use can be end level analysis, MC tests, and common sample generation.
Most pick-events activity occurs on D0mino
There is a lot of co-ordination within the physics groups, will improve with new Common Samples group

CAB Performance, cont.



SAM and non-SAM usage shown, usage is an estimator for scaling system. At peak times, all processors are in use—add 100% Contingency.

CAB File Transfers



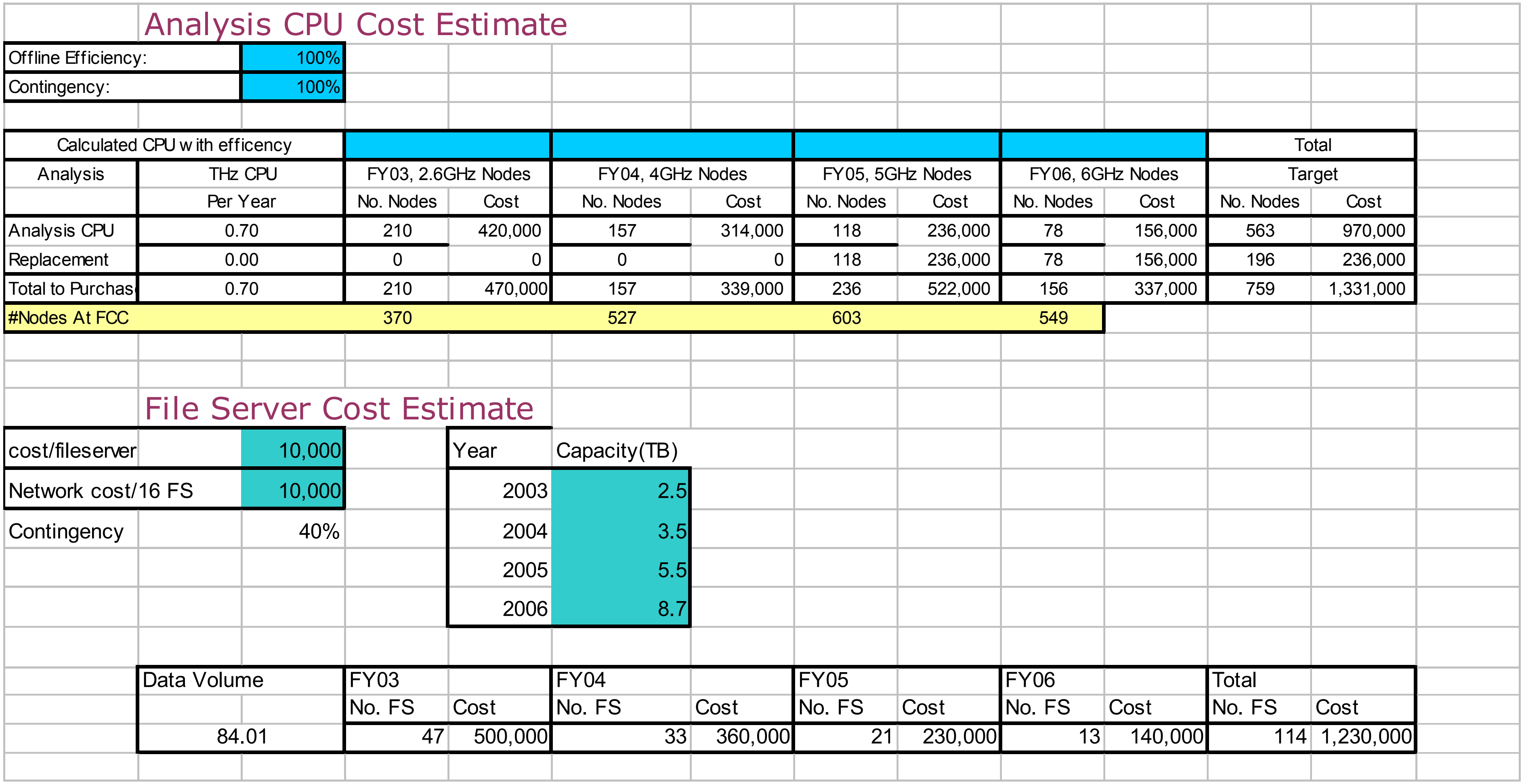
New Plot!—Major components of data transfer on CAB

Slight excess transfers is fine—jobs crash while SAM delivers

Station problem in Feb now fixed.

Amber Boehnlein, FNAL

Estimated Analysis Cost



Analysis CPU Cost Estimate

Offline Efficiency:	100%
Contingency:	100%

Calculated CPU w ith efficency										Total	
Analysis	THz CPU	FY03, 2.6GHz Nodes		FY04, 4GHz Nodes		FY05, 5GHz Nodes		FY06, 6GHz Nodes		Target	
	Per Year	No. Nodes	Cost	No. Nodes	Cost	No. Nodes	Cost	No. Nodes	Cost	No. Nodes	Cost
Analysis CPU	0.70	210	420,000	157	314,000	118	236,000	78	156,000	563	970,000
Replacement	0.00	0	0	0	0	118	236,000	78	156,000	196	236,000
Total to Purchas	0.70	210	470,000	157	339,000	236	522,000	156	337,000	759	1,331,000
#Nodes At FCC		370		527		603		549			

File Server Cost Estimate

cost/fileserver	10,000
Network cost/16 FS	10,000
Contingency	40%

Year	Capacity(TB)
2003	2.5
2004	3.5
2005	5.5
2006	8.7

Data Volume	FY03		FY04		FY05		FY06		Total	
	No. FS	Cost	No. FS	Cost	No. FS	Cost	No. FS	Cost	No. FS	Cost
84.01	47	500,000	33	360,000	21	230,000	13	140,000	114	1,230,000

Includes dCache Read Pools in file server estimates

Amber Boehnlein, FNAL

Primary Production

- **Depends on speed and memory consumption (See H. Melanson talk)**
- **Assumed that 2002 nodes replaced in 2005, etc.**

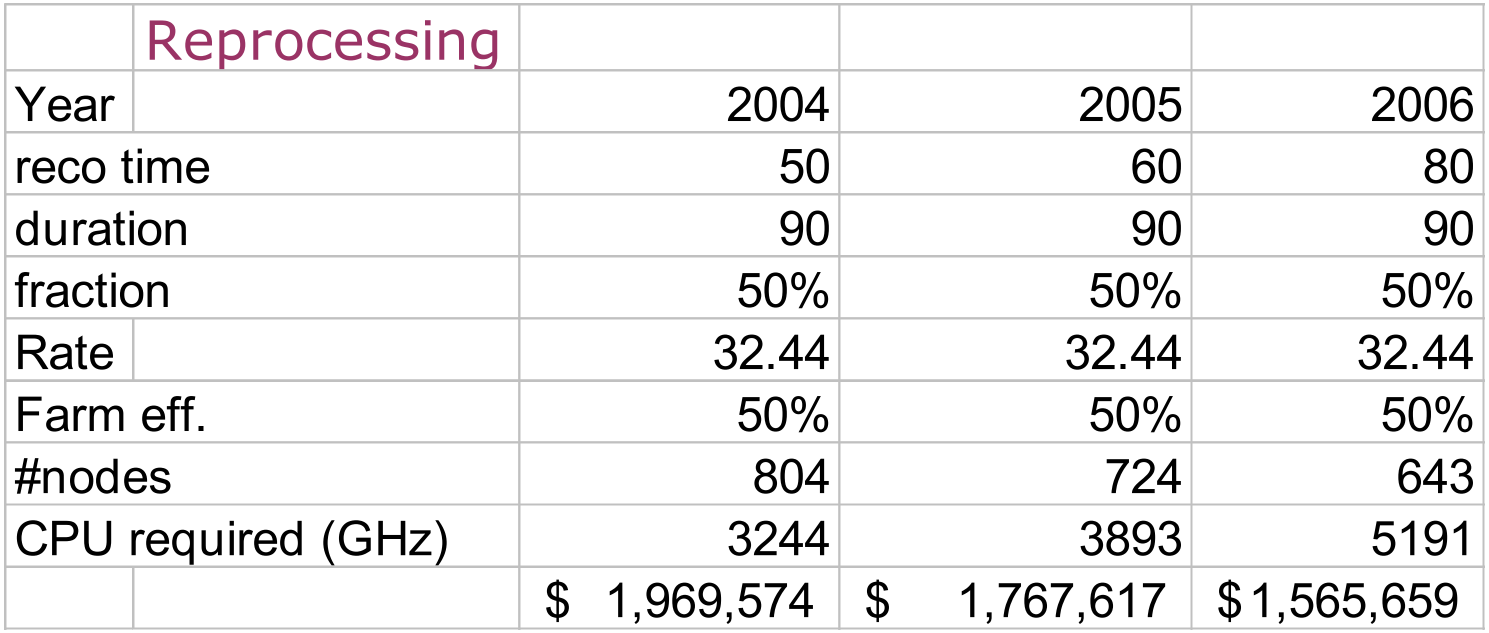
	Primary Reconstruction Cost Estimate		
Year	2004	2005	2006
Reco time	50	60	80
Required CPU	1371	1646	2194
Existing system	670	672	1173
Nodes to purchase	174	181	127
Cost	$407,507	$423,464	$303,573
#Nodes at FCC	530	451	482

Possible ReReco Scenarios

- Resources needed will vary as a function of
  - Amount of data to process
  - How quickly it needs to done
  - Speed of Reco
- Constrained by release cycle, analysis timescales

Application	20 GHz-sec/event	50 GHz-sec/event	7 GHz-sec/event
100% data, 6months	1.5 THz/data-year	3.1 THz/data-year	0.5 THz/data-year
30% data, 3 months	0.87 THz/data-year	2.8 THz/data-year	0.3 THz/data-year

Estimated Reprocessing Costs

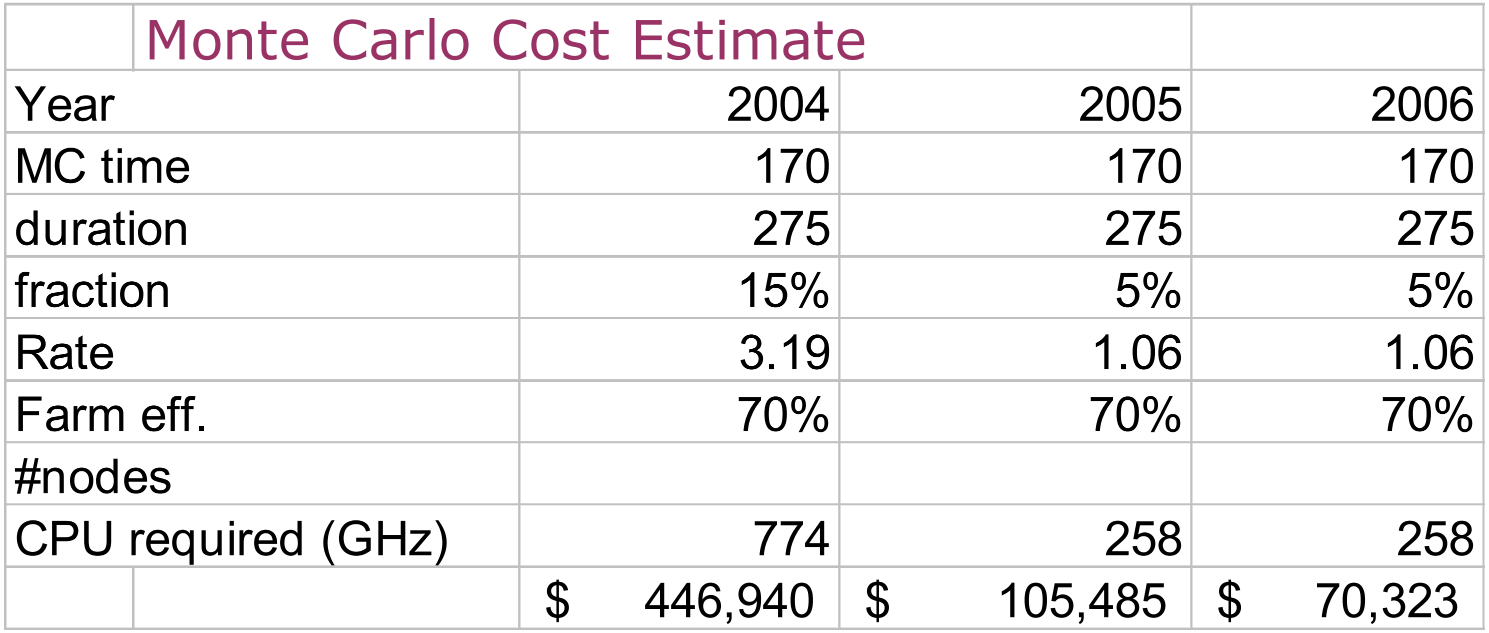


	Reprocessing			
Year		2004	2005	2006
reco time		50	60	80
duration		90	90	90
fraction		50%	50%	50%
Rate		32.44	32.44	32.44
Farm eff.		50%	50%	50%
#nodes		804	724	643
CPU required (GHz)		3244	3893	5191
		$ 1,969,574	$ 1,767,617	$1,565,659

- Assume
  - •50% of data collected yearly is reprocessed and used as the cost estimator
  - •50% efficiency for the farm production
  - •Assume 90 day duration—could be made longer or do more events—cost equivalent to 100% of the data in 6 months
  - •This cost assigned to the virtual center, but pro-rated

Estimated MC Costs

- Legacy machines not counted—many are reaching end of life cycle
- Assume we purchase bulk of needed capacity next year
- Assume overlap with Re-reco machines



Monte Carlo Cost Estimate			
Year	2004	2005	2006
MC time	170	170	170
duration	275	275	275
fraction	15%	5%	5%
Rate	3.19	1.06	1.06
Farm eff.	70%	70%	70%
#nodes			
CPU required (GHz)	774	258	258
	$ 446,940	$ 105,485	$ 70,323

Contributions

Use the FNAL equipment budget to provide very basic level of functionality

- Database and other infrastructure
- Primary Reconstruction farm
- Robotic storage and tape drives
- Disk cache
- Basic analysis computing
- Support for data access to enable offsite computing

Institutional Contributions

- All Monte Carlo production takes place at remote centers
- Secondary reprocessing
- Analysis at home institutions
- Contributions at FNAL to project disk and to CLuED0
- Eventually collaboration wide analysis

Infrastructure Estimates

Infrastructure				
Year		2004	2005	2006
databases				
	servers	$30,000	$30,000	$30,000
	disk	$30,000	$30,000	$30,000
Home Areas		$50,000	$10,000	$10,000
Networking		$120,000	$80,000	$100,000
Machines		$60,000	$60,000	$60,000
Totals		$290,000	$210,000	$230,000

Home areas—either keep SGI and buy faster disk or
Buy replacement system: took average cost.

Networking cost under-estimated-Phil ~$260K in 2004

Cost Estimate-Sept 2003

	2003	2004	2005	2006
FNAL Analysis CPU	$505,400	$339,000	$522,000	$337,000
Primary Reconstruction	$200,000	$407,507	$423,464	$303,573
Re-Reco	NA	$1,969,574	$1,767,617	$1,565,659
Monte Carlo	NA	$446,940	$105,485	$70,323
File Servers/disk	$262,000	$360,000	$230,000	$140,000
Mass Storage	$280,000	$230,000	$100,000	$500,000
Remote Analysis				
Infrastructure	$244,000	$290,000	$210,000	$230,000
FNAL Basic	$1,491,400	$1,626,507	$1,485,464	$1,510,573
Virtual Center Total		$2,454,105	$2,006,482	$1,954,730

Reconstruction is a cost driver—selective reprocessing, speeding up Reco

File servers and farms are not generous—no reprocessing at FNAL in most Basic plan.

Global Remote Analysis in preparation

Very Little flexibility in this plan.

Amber Boehnlein, FNAL

Conclusions

- The DO computing model is successful
  - Having an integrated data handling system enables flexibility in the allocation of resources and effective use of disk and robotic storage and is our path into the GRID era
  - Most performance tracking metrics shown today come from the SAM database
  - TMB format extremely valuable
- Use Virtual Center Concept to calculate all costs.
- DO is shifting our thinking towards a more global model—and making structural changes and plans accordingly.
- We will need increased effort in order to make good use of all available hardware resources