

# Snapshot of the D0 Computing and Operations Planning Process

Amber Boehnlein

For the D0 Computing Planning  
Board

# D0 CPB And Friends

- Lee Lueking and Chip Brock, Ruth & Wyatt & Vicky, Don & Jon & Bonnie
- Mike Diesburg, Iain Bertraim, Jae Yu, Heidi
- Harry Melanson, Serban Protopopescu, Qizhong Li, Dugan O'Neil
- Alan Jonckheere
- Stu Fuess
- Dane Skow, Dave Fagan
- Nick Hadley, Jianming Qian, ASB

# Data size and Storage assumptions

		size		tape factor	tier disk factor
<b>sizes</b>	raw event size	0.3 MB		1	0.001
	raw/reprocessing size	0.5 MB		0.2	0.001
	data DST size	0.125 MB		1.2	0.1
	data TMB size	0.0125 MB		3	1
	data rootuple size	0.01 MB		0	0
	MC DOGstar size	0.7 MB		0.1	0
	MC DOSim	0.3 MB		0	0
	MC DST size	0.2 MB		0	0
	MC TMB size	0.02 MB		2	0.5
	PMCS MC size	0.02 MB		2	0.5
	MC rootuple size	0.02 MB		0	0

data samples (events)				
	1 day	1 year	phase 1 2 years	phase 2 4 years
event rate	1.90E+06	6.94E+08	1.39E+09	8.33E+09
<b>TAPE data accumulation (TB)</b>				
raw event	0.57	208.14	416.28	2497.65
raw/reprocessing	0.19	69.38	138.76	832.55
data DST	0.29	104.07	208.14	1248.83
data TMB	0.07	26.02	52.03	312.21
data rootuple	0.00	0.00	0.00	0.00
MC DOGstar	0.13	48.57	97.13	582.79
MC DOSim	0.00	0.00	0.00	0.00
MC DST	0.00	0.00	0.00	0.00
MC TMB	0.08	27.75	55.50	333.02
PMCS MC	0.08	27.75	55.50	333.02
MC rootuple	0.00	0.00	0.00	0.00
<b>total storage (TB)</b>	<b>1</b>	<b>512</b>	<b>1,023</b>	<b>6,140</b>
<b>total storage (PB)</b>	<b>0.001</b>	<b>0.51</b>	<b>1.02</b>	<b>6.14</b>
<b>total storage (GB)</b>	<b>1,402</b>	<b>511,672</b>	<b>1,023,343</b>	<b>6,140,059</b>
<b>TIER DISK data accumulation (TB)</b>				
raw event	0.00	0.21	0.42	2.50
raw/reprocessing	0.00	0.35	0.69	4.16
data DST	0.02	8.67	17.34	104.07
data TMB	0.02	8.67	17.34	104.07
data rootuple	0.00	0.00	0.00	0.00
MC DOGstar	0.00	0.00	0.00	0.00
MC DOSim	0.00	0.00	0.00	0.00
MC DST	0.00	0.00	0.00	0.00
MC TMB	0.02	6.94	13.88	83.26
PMCS MC	0.02	6.94	13.88	83.26
MC rootuple	0.00	0.00	0.00	0.00
<b>total storage (TB)</b>	<b>0</b>	<b>32</b>	<b>64</b>	<b>381</b>
<b>total storage (PB)</b>	<b>0.000</b>	<b>0.03</b>	<b>0.06</b>	<b>0.38</b>
<b>total storage (GB)</b>	<b>87</b>	<b>31,776</b>	<b>63,551</b>	<b>381,308</b>

# D0 Institution Contributions

- All Monte Carlo production takes place offsite at Remote Centers
- Expect some analysis to occur remotely
- Investigating compute intensive operations in addition to MC generation for remote centers
- CLueD0 desktop cluster, administered by D0 collaboration members, contributions by institutions
- Institutions can provide project disk on D0mino
- Anticipated that institutions will contribute to CLuBs, the CLueD0 back end

# Access and Analysis patterns

- Current Access and Analysis patterns
  - Much primary analysis done on high level data tier—currently reco based root tuple
  - Physics group coordinated efforts to generate
    - Derived data sets by skimming through root tuple or reco output
    - Picked event samples of raw data for re-reco studies
    - Specialized reprocessing of small data sets

# Extrapolation of Analysis patterns

- Assume that physics or analysis group generation of derived data sets continues
  - Skim of thumbnails for desktop or club analysis
  - Skim of DSTs for studies of which tmb contains inadequate information, re-reco
  - Pick events of raw data samples, small dst samples
  - Supply freight train to regulate fast DST access.
- Assume that the bulk of the users will do analysis on the TMB, either on DBE, club or remote cpus
- Smaller group does more time intensive analysis as a service running over larger data sets on DBE

- Currently difficult to estimate analysis cpu usage
  - Assume generation of derived sets is relatively quick/per event, but happens often
  - Most DST access implies time intensive operations. Estimate it is at least the order of farm processing
    - Would support 3 simultaneous users,  $\frac{1}{4}$  farm processing time in 3 months each on  $\frac{1}{3}$  data, for initial data set.
  - Reco time per event is expected to increase dramatically as a number of multiple interactions
  - Make overall estimate of 75 seconds/event (500 MHz) for reco and analysis and re-reco—collaboration to weigh relative balance available in 2005, staged in.
  - Institutions can contribute CPU to CLuBs, assume FNAL contribution of \$50K yearly
  - Remote center reprocessing, or dst level reprocessing is under evaluation.

# Farm Processing

Average Rate:	75	CPU	SpecI2000						
Farm Efficiency:	70%	3GHz	960						
Misc. Processing:	10%	4GHz	1280						
Reprocessing:	50%	6GHz	1920						
Cost/node:	3,000	9GHz	2880						
I/O Cost/100 nodes	25,000	14GHz	4480						
		20GHz	6400						
FY05 Target Spending Fraction:	20%	30%	50%	Total					
Execution Time	500MHz CPUs at Beginning of Run	FY03, 3GHz Nodes No. Nodes    Cost	FY04, 4GHz Nodes No. Nodes    Cost	FY05, 6GHz Nodes No. Nodes    Cost	Target No. Nodes    Cost				
30	5143	72    241,000	108    349,000	180    590,000	360	1,180,000			
75	12857	180    590,000	271    888,000	452    1,481,000	903	2,959,000			
100	17143	241    798,000	361    1,183,000	602    1,981,000	1204	3,962,000			

Farm processing capacity in Summer '02 ~50Hz

D0mino backend

16 node, 1 GHz

80 nodes, 2 GHz, summer '02

# Roles of D0mino

- D0mino provided a centralized, stable and uniform work environment
  - Interactive and batch services for on and offsite users
- High I/O provided by 8 Gigabit ethernet connections—The Central Analysis SAM station
  - Interactive
  - To/from robotic storage
  - To/from secondary analysis systems
  - To/from the backend
- Access to project disks and disk cache (30 TB)
- Analysis CPU provided by Linux backend nodes

# Upgrading D0mino

- Replacing D0mino requires identifying which parts of the design can be better served by more cost effective solutions
  - Linux back end to supply compute power for access to large data samples
  - Seeding CLuBs (the Clued0 backend) as solution for intermediate (1TB) samples
  - Continue to evaluate I/O performance
- Upgrading D0mino (O3000) would cost about \$2M, and would have to be phased in. To stay within nominal guidance, would cut analysis/farm capacity by about 1/2
- Fortunately, SAM gives D0 a lot of flexibility.

# Backup facility

- Two primary consumers
  - Project disk archive
  - User driven backups of small samples
- Clearly a need, but not clear how best to accomplish.

# Phase 1, Robotic Storage Plan

- D0 has 1 STK silo with 9 9940 drives
  - Writing raw data and Reco output to mezzosilo1
- We have an option on a second silo
- AML2 with 6 LTO drives
  - Writing MC data
  - Plan to start writing Reco output as a test on May 7
  - If test is successful
    - Will purchase more LTO Drives and a few more 9940x
  - Else
    - Will purchase as many 9940x as is feasible
  - Likely need to purchase a few more drives of each type to get us to decision point.

- The overall estimated need for Robotic storage for phase 1 can be accommodated by the Mezzosilo1&2 and the AML2 even with the current generation of drives/media.

Current Capacity approx  $2 * 300 \text{ TB} + 750 \text{ TB}$ —  
compare to roughly 1 PB needed.

Or

The overall estimated need for Robotic storage for phase 1 can be accommodated by Mezzosilos 1&2 with 9940bs

Assume the purchase of Mezzosilo 2 in 2003

Assume purchase of Drives for Mezzosilo 2 in 2003

Assume additional purchase of drives in 2004

# Drive Estimates

- Support Online operations plus buffer drain
  - 3 10mbyte/sec drives
- Support Farm operations
  - 3 10 mbyte/sec drives
- Support incoming MC
  - 2 drives
- Support Central analysis
  - Freight train for spooling through the DST sample in short interval (3 months) would consume 8 drives
  - MC and other
  - Pick events could consume infinite number of drives
- Buy LTO and STK in 2002—distribution to vary
- 20 drives for new Mezzosilo
- 2004 expect to add drives

# Fixed Infrastructure Costs

- Database
  - machines
  - Database disks and controllers (assumed cost 10X cots)
  - DB Mirrors
  - Software
- Networking
  - Expand links between buildings, FCC
  - Additional switches for DAB, farms
  - D0 to FCC upgrade to 10 Gb backbone upgrade '06
  - Rewiring D0 for Gb to desktop in '07
- Linux build machines and disk
- Additional SAM servers

# Disk Estimates

- Aim for sufficient cache, TMB storage on D0mino
  - All 2002 D0mino project disk additions supplied by the Institutions
  - Assume that model continues for project space
  - Supply additional 18 TB cache per year

## Summary of infrastructure costs:

<u>Infrastructure Costs</u>	2003	2004	2005	2006	2007	<i>Total</i>
<u>Databases:</u>						
DB Hosts, Sun, then Linux	\$60,000	\$60,000	\$25,000	\$25,000	\$25,000	
non COTS disk and controllers	\$60,000	\$20,000	\$10,000	\$10,000	\$10,000	
Mirrors	\$30,000	\$15,000	\$25,000	\$15,000	\$15,000	
Software	\$50,000	\$0	\$50,000	\$0	\$50,000	
DB totals	\$200,000	\$95,000	\$110,000	\$50,000	\$100,000	\$555,000
Networking	\$80,000	\$50,000	\$100,000	\$200,000	\$400,000	\$830,000
Build Machines	\$50,000	\$50,000	\$50,000	\$50,000	\$50,000	\$250,000
Additional SAM servers	\$50,000	\$50,000	\$50,000	\$50,000	\$50,000	\$250,000
Total, fixed cost	\$380,000	\$245,000	\$310,000	\$350,000	\$600,000	\$1,885,000

# Rate assumptions

<b>rates</b>	average event rate	22	Hz
	raw data rate	22.5	MB/s
	Geant MC rate	11	Hz

Average rate assumes an accelerator and experiment  
Duty factor applied to a peak rate of 50 Hz

<b>rate increase assumptions</b>			
	rate factor		3
	phase_1		2
	phase_2		4
	last year		2009
	total years		6
	down year		2005

# Full Cost Estimate, No I/O replacement

Extremely preliminary D0 C&S cost estimate								
		2002	2003	2004	2005	2006	2007	<b>Total(2003-2007)</b>
Fixed Infrastructure Costs		\$400,000	\$380,000	\$245,000	\$310,000	\$350,000	\$600,000	\$1,885,000
farm + analysis cpu		\$800,000	\$640,000	\$938,000	\$1,531,000	\$500,000	\$500,000	\$4,109,000
disk cache		\$0	\$150,000	\$100,000	\$50,000	\$150,000	\$150,000	\$600,000
robotic storage		\$400,000	\$150,000	\$0	\$150,000	\$150,000	\$150,000	\$600,000
tape drives		\$200,000	\$600,000	\$300,000	\$300,000	\$600,000	\$600,000	\$2,400,000
D0mino upgrade		\$150,000	\$0				\$0	\$0
Backup facility			\$350,000					
Sum		\$1,950,000	\$2,270,000	\$1,583,000	\$2,341,000	\$1,750,000	\$2,000,000	\$9,944,000

# Questions to D0

- Is it possible to make a better estimate of analysis CPU needs?
- Role of D0mino
- Relative weighting of tmb and DST analysis—better to have a larger tmb as trade off for less DST usage?
- Role of remote centers

# Questions to CD

- How should we cost mover nodes?
- Relative role of disk vs robotic storage as time goes on
- Where will we put the phase 2 robots?
- Interaction between networking and remote centers
- Suggestions for backup facility