

Requirements and Budget for CDF Computing in Run 2

R. Snider

for CDF Computing and Analysis Dept.

CD Project Status Meeting
August 27, 2003

Introduction

- Director's review of Run 2 computing
 - Present three year plan for offline computing system
 - Include resource requirements and cost
- This talk is preview of requirements and budget to be presented
- Starting point for analysis
 - Update to budget and plan from last year's review
 - Projections and estimates through 2009
 - Compare to experience over past year
- Significant change from last year
 - CDF will increase event logging rate by 50% next year
 - Additional factor of two increase in data rate in FY05
 - Additional 50% data rate increase in FY06

Introduction

- Goal for this talk
 - Assess impact of event rate increases on offline computing requirements
 - Old baseline requirements
 - New baseline requirements
 - Estimate cost associated with meeting the requirements
 - Present three year procurement plan
- Will not discuss:
 - Details of existing hardware
 - Software issues
 - Non-Fermilab contributions, commitments
 - Plans beyond scope of projected resource requirements

Outline

- Define the estimate scenarios
- Overview of system components
- Assumptions, resource usage models
- Individual system analyses
 - Central Analysis Facility (CAF)
 - Data Handling (DH)
 - Production Farm
 - etc.
- Cost summary
- Conclusions

Context of estimates

- Will present estimates under two scenarios:

- New baseline

| Fiscal year | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|-----------------------|------|------|------|------|------|------|------|
| Logging rate (MB/sec) | 20 | 20 | 40 | 60 | 60 | 60 | 60 |
| Raw data compression | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Peak event rate (Hz) | 80 | 120 | 240 | 360 | 360 | 360 | 360 |

- For comparison, the old baseline (from 2002 run plan)

| Fiscal year | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|-----------------------|------|------|------|------|------|------|------|
| Logging rate (MB/sec) | 20 | 20 | 20 | 60 | 60 | 60 | 60 |
| Raw data compression | No |
| Peak event rate (Hz) | 80 | 80 | 80 | 240 | 240 | 240 | 240 |

- Average logging rates = 70% of peak

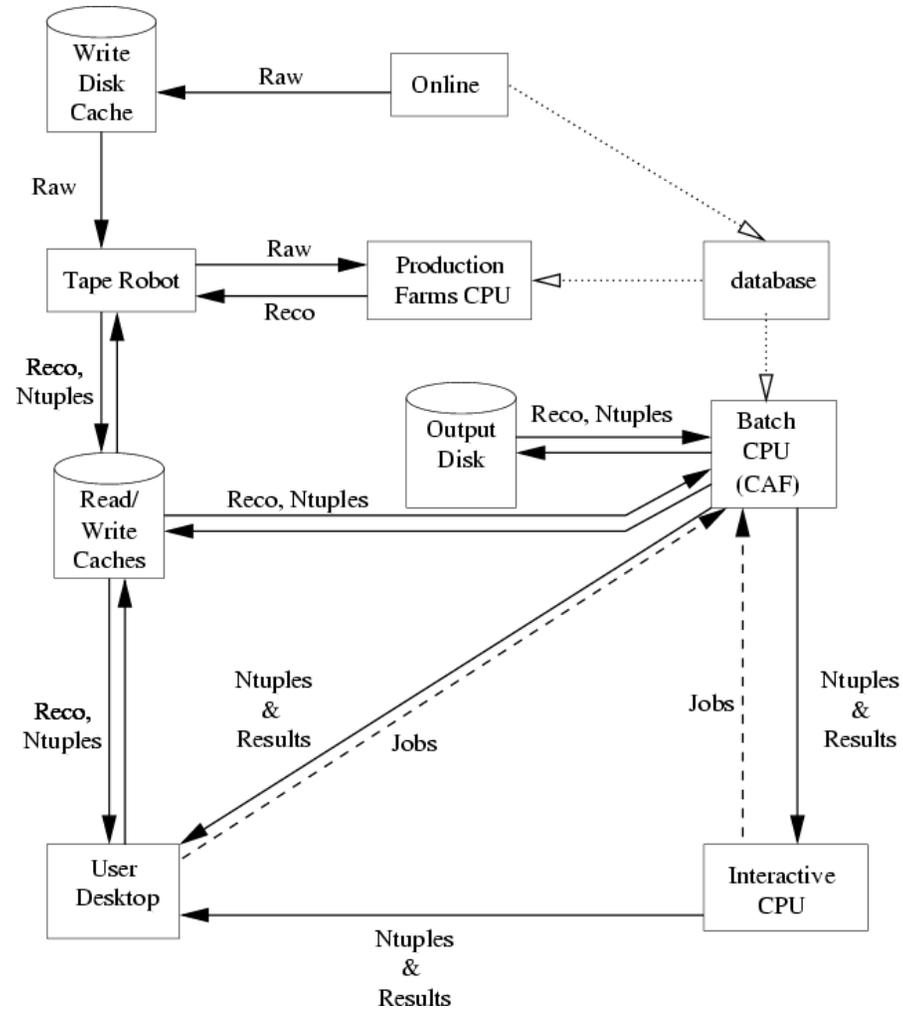
Context of estimates

- Assume the "design" luminosity from June 2003 Project Plan

| | Luminosity by FY (1/fb) | | | | | | |
|---------------------|-------------------------|------|------|------|------|------|------|
| | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| Delivered per year | 0.22 | 0.38 | 0.67 | 0.89 | 1.53 | 2.37 | 2.42 |
| Integrated | 0.3 | 0.68 | 1.4 | 2.2 | 3.8 | 6.1 | 8.6 |
| Integrated acquired | 0.25 | 0.61 | 1.2 | 2.0 | 3.4 | 5.5 | 7.7 |

- Assumes 90% data logging efficiency
- Assume current software performance

Offline computing system overview



Offline computing system overview

- Central Analysis Facility (CAF)
 - 600 CPU farm (largest line item in the budget)
 - 180 TB network attached disk (managed by DH)
 - Interactive CPU
- Data Handling (DH)
 - Data archive
 - Tapes
 - Tape drives
 - Read/write disk cache
- Production Farm
 - 400 CPU farm
- Associated networks, DB servers

Current computing usage model

- Production farms
 - Primary reconstruction of data from detector within days of data-taking
 - Periodically re-processes all logged raw data with new version of software
 - Special projects
 - Wholesale re-processing
- CAF
 - Physics groups create secondary datasets from production output
 - Users analyze secondary datasets
 - Create tertiary datasets, ntuples or other highly compressed output
 - Small MC jobs
- Desktops
 - Analyze tertiary datasets, ntuples or other highly compressed data formats

Current computing usage model

- Off-site
 - Monte Carlo production
 - Large scale production of MC samples coordinated by physics groups
 - Analysis of tertiary datasets, ntuples or other highly compressed datasets

Basic computing resource model

- Demand for computing resources scales in two ways:
 - Number of events / data volume
 - Event / data logging rate
- Scaling with number of events / data volume
 - Archive volume
 - With user analysis model
 - CAF capacity
 - Contributions to archive I/O
 - Contributions to network throughput
- Scaling with event / data logging rate
 - Production farms
 - Contributions to archive I/O rate
 - Contributions to network throughput

- Define the estimate scenarios
- Overview of system components
- Assumptions, resource usage models
- Individual system analyses
 - Central Analysis Facility (CAF)
 - Data Handling (DH)
 - Production Farm
 - etc.
- Cost summary
- Conclusions

CAF CPU requirements

- Batch CPU farm
 - Use analysis model to estimate demand
- Three components to batch CPU analysis model
 - User datasets that scale with luminosity (high-Pt triggers)
 - Fixed cross section of 400 nb
 - User datasets that scale with integrated running time (low-Pt triggers)
 - Balance of (average) bandwidth above 400 nb
 - Assume average event logging rate = 70% peak rate
 - Note: not included in "old baseline" estimates
 - Secondary dataset creation
 - Requires periodically running over all production output
- Other activities (MC) not explicitly included
 - Many will scale similarly

CAF CPU requirements

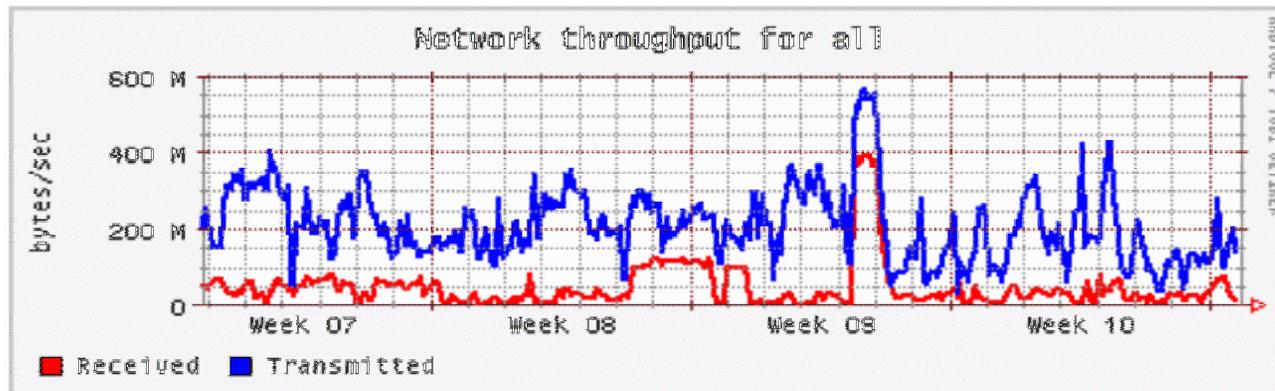
- The analysis model
 - Fixed cross section (high-Pt) analyses
 - Assume 200 users analyzing 5 nb datasets in a single day
 - Typical scale for many high Pt datasets
 - Run-time scaled (low-Pt) analyses
 - Assume 15 users analyzing entire dataset in 25 days
 - Secondary dataset creation
 - Process entire production output three times per year over course of year
- Assume all activities occur simultaneously
- Sum over contributions to arrive at estimated demand, I/O rates
 - Event processing rate = 0.2 GHz-sec (5 Hz on 1 GHz PIII equivalent)
 - Assume 80% CPU utilization (includes latencies)
 - Allow 30% contingency factor

CAF CPU requirements

- Test predictions against usage during past winter
 - About 1/3 of CAF was not utilized
 - Will assume that CAF was not CPU constrained
 - Note, however, that only 1/3 of CAF provided access to certain popular datasets
 - Predicted need for 380 GHz
 - Total of 630 GHz utilized
 - Within a factor of two, assuming winter CAF met the demand

CAF CPU requirements

- Check I/O rate during same period



Predicted read rate
= 210 MB/sec

CAF CPU requirements

- Projected CPU requirements

| Scenario | Required CPU in CAF by FY (THz) | | | | | |
|--------------|---------------------------------|------|------|------|------|------|
| | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| Old baseline | 2.2 | 4.5 | 7.5 | 13 | 21 | 29 |
| New baseline | 3.7 | 9.0 | 14 | 23 | 34 | 46 |

CAF CPU requirements

- Sanity check: simple scaling from current CAF
 - Scale from size of CAF during winter 2003 using scaling extremes
 - Integrated luminosity
 - Number of events logged (run-time scaling)

| Model | CAF requirements by FY (THz) | | | | | |
|-------------------|------------------------------|------|------|------|------|------|
| | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| New baseline | 3.7 | 9.0 | 14 | 23 | 34 | 46 |
| Luminosity-scaled | 4.0 | 8.0 | 13.0 | 22 | 37 | 51 |
| Run-time scaled | 3.0 | 6.0 | 8.2 | 13 | 17 | 22 |

- Estimates typically at the high end of the range
 - Current capacity = 0.94 THz

CAF CPU procurement plan

- Assume:
 - Dual CPU boxes at \$2.2k each (constant dollars)
 - Effective cost of machine purchased mid-FY03 for 2.2 GHz P-III equivalent
 - Speed doubles every 18 months
 - Retire nodes after 3 years
- Current capacity = 0.94 THz
- Procurement plan:

| Fiscal year | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|-------------|------|------|------|------|------|------|
| Needs (THz) | 3.7 | 9.0 | 14 | 23 | 34 | 46 |
| Nodes added | 346 | 525 | 318 | 422 | 384 | 232 |
| Speed (GHz) | 3.5 | 5.6 | 8.8 | 14 | 22 | 36 |
| Cost (\$1k) | 760 | 1200 | 700 | 930 | 850 | 510 |

CAF monitoring

- Comment on CAF model
 - Currently lack monitoring tools needed to precisely determine how CAF is being used
 - New offline release, 5.1.0, will have greatly enhanced capabilities
 - Which datasets/files are being accessed
 - CPU consumption
 - Event rate
 - Type of job
 - Data analysis
 - MC
 - Ntuple analysis
 - User id
 - Should allow better tracking of resource usage

CAF network attached disk: requirements

- Basic plan
 - Put as much processed data on disk as possible + staging / cache space
- Requirements
 - Luminosity-scaled dataset: scale winter capacity on CAF (56 TB) by number of logged events
 - Run-time scaled dataset: space for 7 days of analysis on disk

| Scenario | Fileserver capacity by FY (TB) | | | | | |
|--------------|--------------------------------|------|------|------|------|------|
| | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| Old baseline | 210 | 290 | 420 | 670 | 920 | 1200 |
| New baseline | 290 | 600 | 830 | 1300 | 1700 | 2200 |

CAF network attached disk: procurement plan

- Model
 - \$20k each
 - 5 TB units available in mid-FY03
 - Density doubles every 18 months
 - Retire servers after 3 years
- Projected costs

| Fiscal year | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|----------------------|------|------|------|------|------|------|
| Needs (TB) | 290 | 600 | 830 | 1300 | 1700 | 2200 |
| Units added | 13 | 32 | 16 | 15 | 8 | 6 |
| Capacity / unit (TB) | 8.0 | 13 | 20 | 32 | 51 | 81 |
| Cost (\$1k) | 260 | 640 | 320 | 300 | 260 | 120 |

CAF interactive systems

- Legacy IRIX systems
 - cdfsga: supports Run 1 analysis and code
 - fcdfsi2: load is falling
 - Will de-commission as early as 2004
- Interactive pool now under development
 - Pool of interactive servers
 - Linux Virtual Server allows user access to members of pool
 - Provides scalable, commodity-based solution to need for large central interactive machine
 - Prototype: CPU power equivalent to fcdfsi2 for only \$60k
 - Deploy by end of 2003
- Estimated cost FY04 through FY06
 - About \$100k per year (less than current maintenance for fcdfsi2)

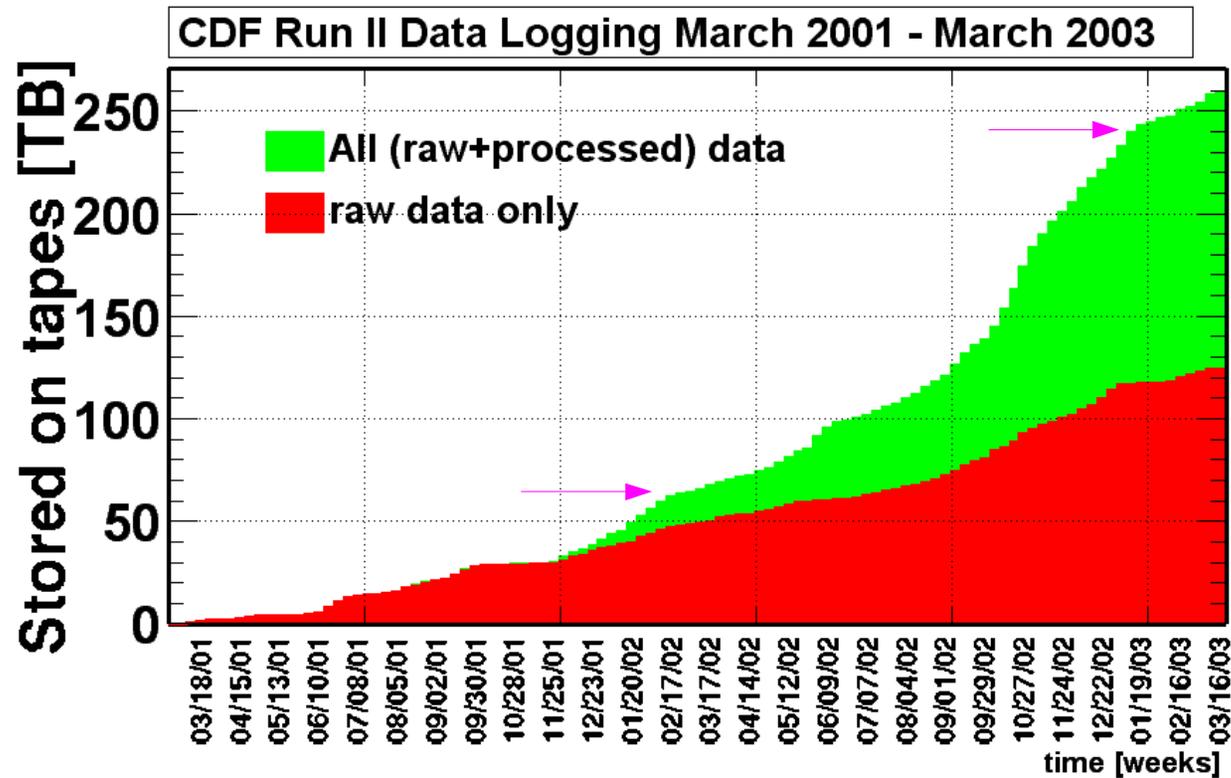
Data handling: data archive requirements

- Data archive
 - Contents of archive stored on tape
 - Raw data
 - Production output
 - Secondary datasets
 - MC data
 - Archive I/O via fixed number of tape drives
 - Contributions from:
 - Production farm
 - CAF: user analysis
 - CAF: secondary dataset creation

Data handling: data archive requirements

- The model
 - Event sizes
 - Raw data event size
 - Old baseline: 220 kB
 - New baseline: 220 kB, decreasing to 150 kB starting FY04
 - Production event size = secondary data event size = 180 kB
 - Ignore luminosity dependence in event size
 - Other assumptions:
 - Data logging rate = 70% of peak
 - 30% average machine operating efficiency
 - 10% of CAF I/O via archive (cache misses)
 - Sum over all contributions to volume and I/O
 - To get requirements:
 - 20% contingency for volume
 - 100% contingency for I/O

Data handling: data archive requirements



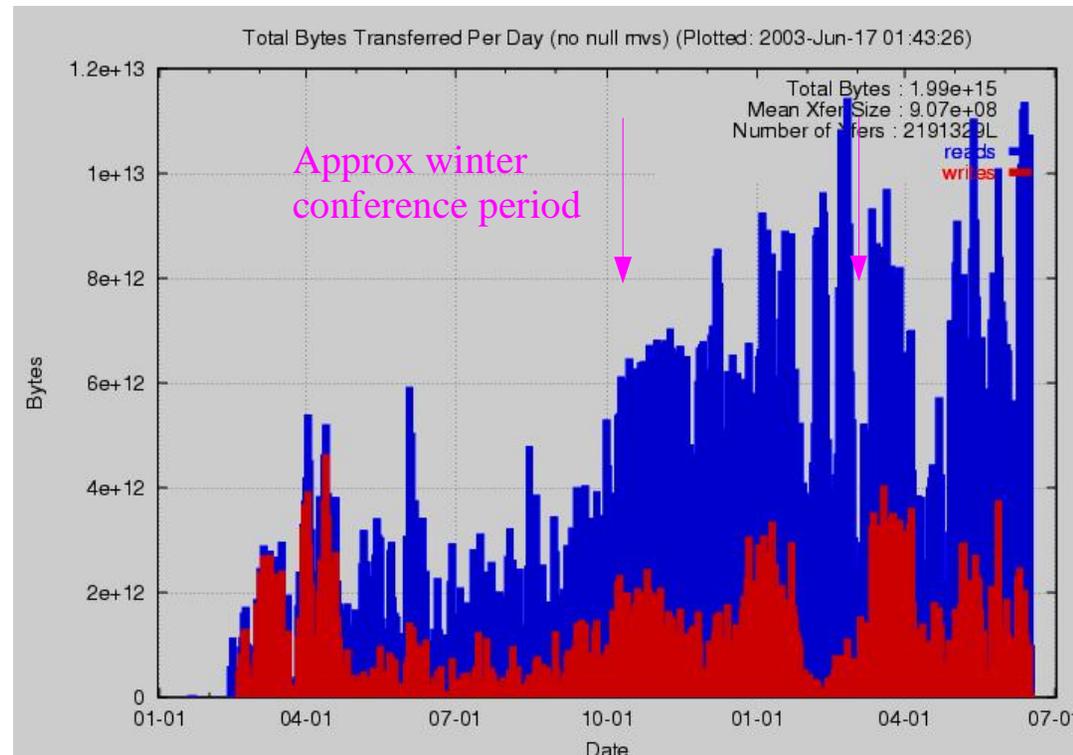
Predicted volume for winter conferences: 180 TB

Observed: 170 TB

Through mid-August, logged total of 550 M events
Predict 580 M through FY03

Data handling: data archive requirements

- Winter conference archive I/O
 - Predict 56 MB/sec = 5 TB/day



Data handling: data archive requirements

- Archive capacity and I/O estimates

| Fiscal year | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|-------------------------|------|------|------|------|------|------|
| Event logging rate (Hz) | 85 | 170 | 250 | 250 | 250 | 250 |
| Raw data (TB) | 120 | 240 | 180 | 350 | 350 | 350 |
| Production output (TB) | 140 | 290 | 210 | 420 | 420 | 420 |
| Secondary datasets (TB) | 140 | 290 | 210 | 420 | 420 | 420 |
| Volume/year (TB) | 410 | 810 | 600 | 1200 | 1200 | 1200 |
| Volume (TB) | 750 | 1600 | 2200 | 3400 | 4600 | 5800 |
| Raw data (MB/sec) | 12 | 24 | 18 | 36 | 36 | 36 |
| Farms I/O (MB/sec) | 26 | 41 | 33 | 71 | 83 | 95 |
| CAF I/O (MB/sec) | 280 | 650 | 980 | 1600 | 2300 | 3100 |
| Archive I/O (MB/sec) | 320 | 770 | 1000 | 1700 | 2500 | 3200 |

Data handling: data archive requirements

- Archive capacity requirements
 - 20% contingency over volume estimates

| Scenario | Archive capacity requirements (TB) | | | | | |
|--------------|------------------------------------|------|------|------|------|------|
| | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| Old baseline | 740 | 1100 | 1600 | 2500 | 3500 | 4500 |
| New baseline | 900 | 1900 | 2600 | 4000 | 5500 | 6900 |

- Archive I/O requirements
 - Factor of 2 contingency over I/O rate estimates

| Scenario | Archive I/O requirements by FY (MB/sec) | | | | | |
|--------------|---|------|------|------|------|------|
| | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| Old baseline | 530 | 820 | 1400 | 2200 | 3300 | 4400 |
| New baseline | 640 | 1500 | 2100 | 3400 | 4900 | 6400 |

Data handling: archive procurement plan

- Archive currently supports two tape densities
 - STK 9940A
 - 60 MB tape cartridges
 - 10 MB/sec I/O rate
 - STK 9940B
 - 200 MB tape cartridges
 - Interchangeable with 9940A's
 - 30 MB/sec I/O rate
 - Currently migrating from 9940A to 9940B tapes
 - Assume needs through FY04 met by 9940B technology

Data handling: archive procurement plan

- Future migration plan
 - Adopt new technology in FY05
 - 400 MB tapes (not interchangeable)
 - 60 MB/sec I/O rate
 - No candidates currently available
 - Copy to higher density tapes at rate of 5500 per year
 - Contents of one tape robot
 - Retire old drives over period of two years
 - One half of inventory each year

Data handling: archive procurement plan

- Media procurement
 - \$75 per tape

| Fiscal year | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|----------------------|------|------|------|------|------|------|
| Needs (TB) | 900 | 1900 | 2600 | 4000 | 5500 | 6900 |
| Units added | 0 | 2400 | 1800 | 3600 | 3600 | 3600 |
| Capacity / unit (TB) | 0.2 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 |
| Cost (\$1k) | 0 | 350 | 140 | 270 | 270 | 270 |

- Cost of migration is about \$200k in FY05

Data handling: archive procurement plan

- Tape drive procurement
 - \$30k per drive

| Fiscal year | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|-----------------------|------|------|------|------|------|------|
| Needs (MB/sec) | 640 | 1500 | 2100 | 3400 | 4900 | 6400 |
| Units added | 9 | 21 | 13 | 15 | 20 | 20 |
| I/O per unit (MB/sec) | 30 | 60 | 60 | 60 | 60 | 60 |
| Cost (\$1k) | 270 | 630 | 390 | 690 | 750 | 750 |

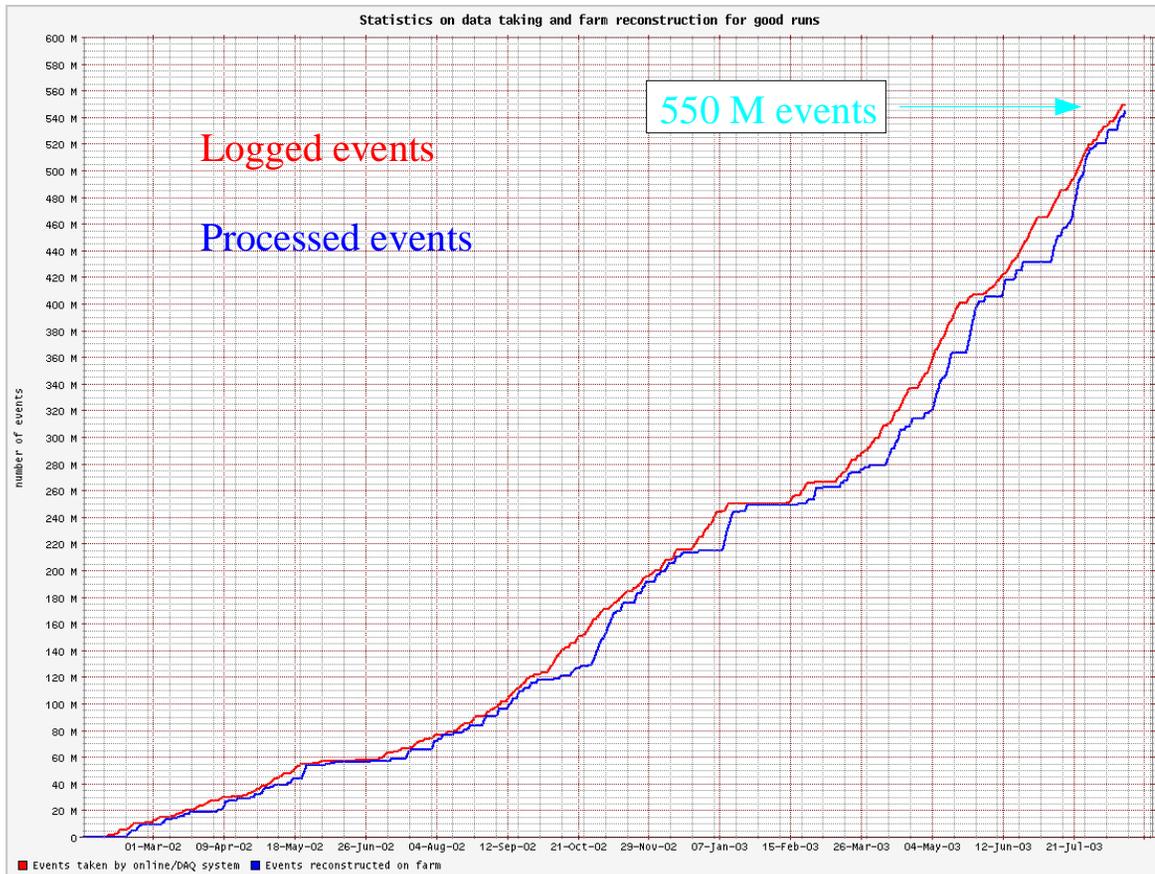
- Define the estimate scenarios
- Overview of system components
- Assumptions, resource usage models
- Individual system analyses
 - Central Analysis Facility (CAF)
 - Data Handling (DH)
 - Production Farm
 - Networks, database servers
- Cost summary
- Conclusions

Production farm requirements

- Design goal
 - Provide capacity to keep up with data-taking in real time
 - Allow simultaneous re-processing of raw data
 - Build in contingency for operational delays
- The model
 - Take sum of event from data logging and re-processing
 - Assume:
 - Event processing time = 3.3 GHz-sec
 - Allow 50% contingency for increases in processing time \Rightarrow 5.0 GHz=sec
 - Farm utilization = 75%
 - Accelerator operating efficiency during stable operations = 60%
 - Process events only when machine is operating
 - Average efficiency = 30%

Production farm requirements

Number of events



Average 7.5 M events
per day over periods
of a week

Production farm requirements

- Re-processing plan

Re-processings by FY

| 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|------|------|------|------|------|------|
| 0.5 | 0.3 | 0.2 | 0.3 | 0.3 | 0.3 |

- Note on re-processing model

- By design, events processed on time scale of stable machine operations
 - 6 months / year
 - Strong motivations for large re-processing fractions in coming years
 - This problem will need to be addressed

Production farm requirements

- Projected farm CPU requirements

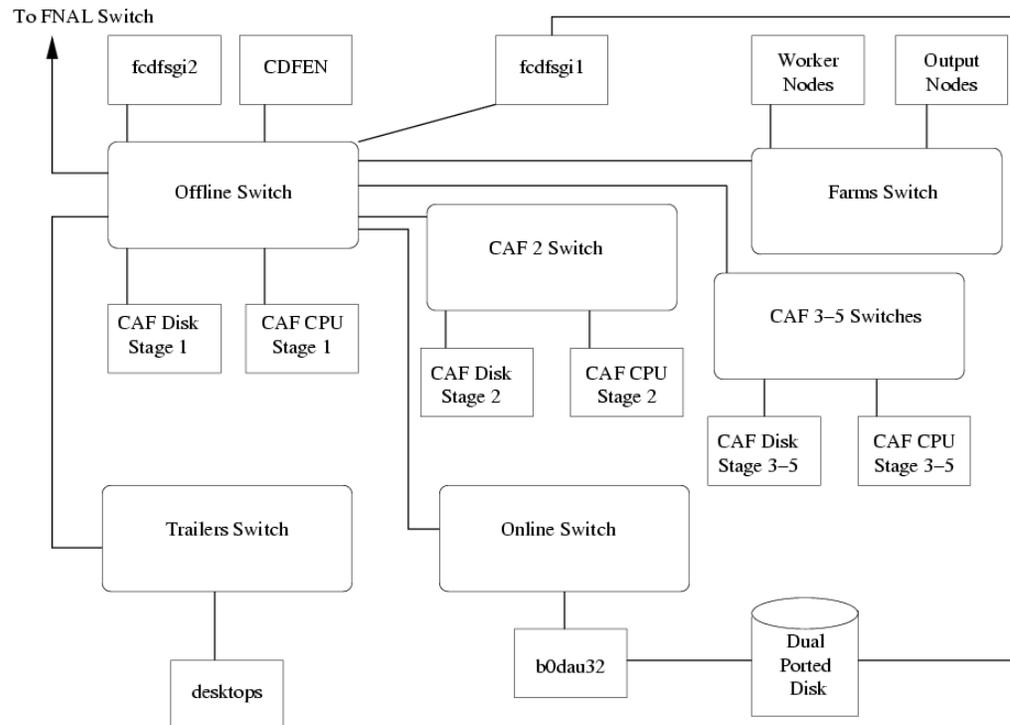
| Scenario | Farm requirements (GHz) | | | |
|--------------|-------------------------|------|------|------|
| | 2003 | 2004 | 2005 | 2006 |
| New baseline | 480 | 800 | 1400 | 2000 |

- Same procurement assumptions as for CAF

| Fiscal year | 2003 | 2004 | 2005 | 2006 |
|-------------|------|------|------|------|
| Needs (GHz) | 480 | 800 | 1400 | 2000 |
| Nodes added | 64 | 64 | 64 | 48 |
| Speed (GHz) | 2.2 | 3.5 | 5.6 | 8.8 |
| Cost (\$1k) | 0.19 | 0.19 | 0.19 | 0.18 |

Networking

- Network-based computing model
 - CAF and associated file servers driving overall network demand
 - Will require continual upgrades to network capacity throughout system
 - Changes to infrastructure required in many places



Simplified diagram
of CDF LAN

Networking

- The model
 - Demand scales with dataset size
 - Assume price drops by factor of two every 18 months
 - Fails if major technology change / infrastructure upgrade is needed
 - Stage trailer LAN upgrades from FY04 through FY06
- Estimated networking expenditures

| Fiscal year | 2003 | 2004 | 2005 | 2006 |
|---------------------|------|------|------|------|
| FCC cost (\$1k) | 230 | 140 | 90 | 60 |
| Trailer cost (\$1k) | 0 | 110 | 100 | 60 |
| Cost (\$1k) | 230 | 250 | 190 | 120 |

Databases

- Projected needs based upon growth in demand from CAF, farms, trailers
 - Add about two servers per year will keep up with demand
 - \$100k per year from FY04 through FY06

Cost summary

| Fiscal year | Actual expenditures | | | Projected costs | | |
|-----------------------|---------------------|------|------|-----------------|------|------|
| | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
| Batch CPU (\$M) | 0 | 0.39 | 0.31 | 0.76 | 1.16 | 0.70 |
| Interactive CPU (\$M) | 0 | 0.01 | 0.08 | 0.10 | 0.10 | 0.10 |
| Farm CPU (\$M) | 0.25 | 0.22 | 0.13 | 0.19 | 0.19 | 0.18 |
| DB (\$M) | 0 | 0.02 | 0.14 | 0.10 | 0.10 | 0.10 |
| Tape robot (\$M) | 0 | 0.77 | 0.20 | 0.27 | 0.57 | 0.54 |
| Cache disk (\$M) | 0 | 0.63 | 0.34 | 0.26 | 0.64 | 0.32 |
| Network (\$M) | 0 | 0.25 | 0.23 | 0.25 | 0.19 | 0.12 |
| Legacy systems (\$M) | 0.75 | 0.69 | 0 | na | na | na |
| Total (\$M) | 1.0 | 3.0 | 1.4 | 1.9 | 3.0 | 2.1 |




Off-site computing

- Large computing demand suggests off-site computing as a possible resource
 - A priority of CDF to better utilize off-site computing
- Current initiatives
 - MC production (Toronto)
- Require SAM, SAM/GRID to make full use of off-site computing
 - The plan
 - Migrate to SAM for production use at CDF (FY04)
 - SAM/GRID in production at CDF (FY05)
 - Rely on additional off-site contributions to computing (FY06+)

Conclusions

- Expanded data rates will increase demand on computing resources
 - Will require increases over FY03 expenditures in short term to meet demands
 - Budgets of \$2M through FY06 will produce significant shortfall in FY05 when logging rate doubles
 - Decreases in event size, improvements in software performance in principle mitigate the problem
 - Effect of additional increases in logging rate in FY06 offset reduced running time due to shutdown