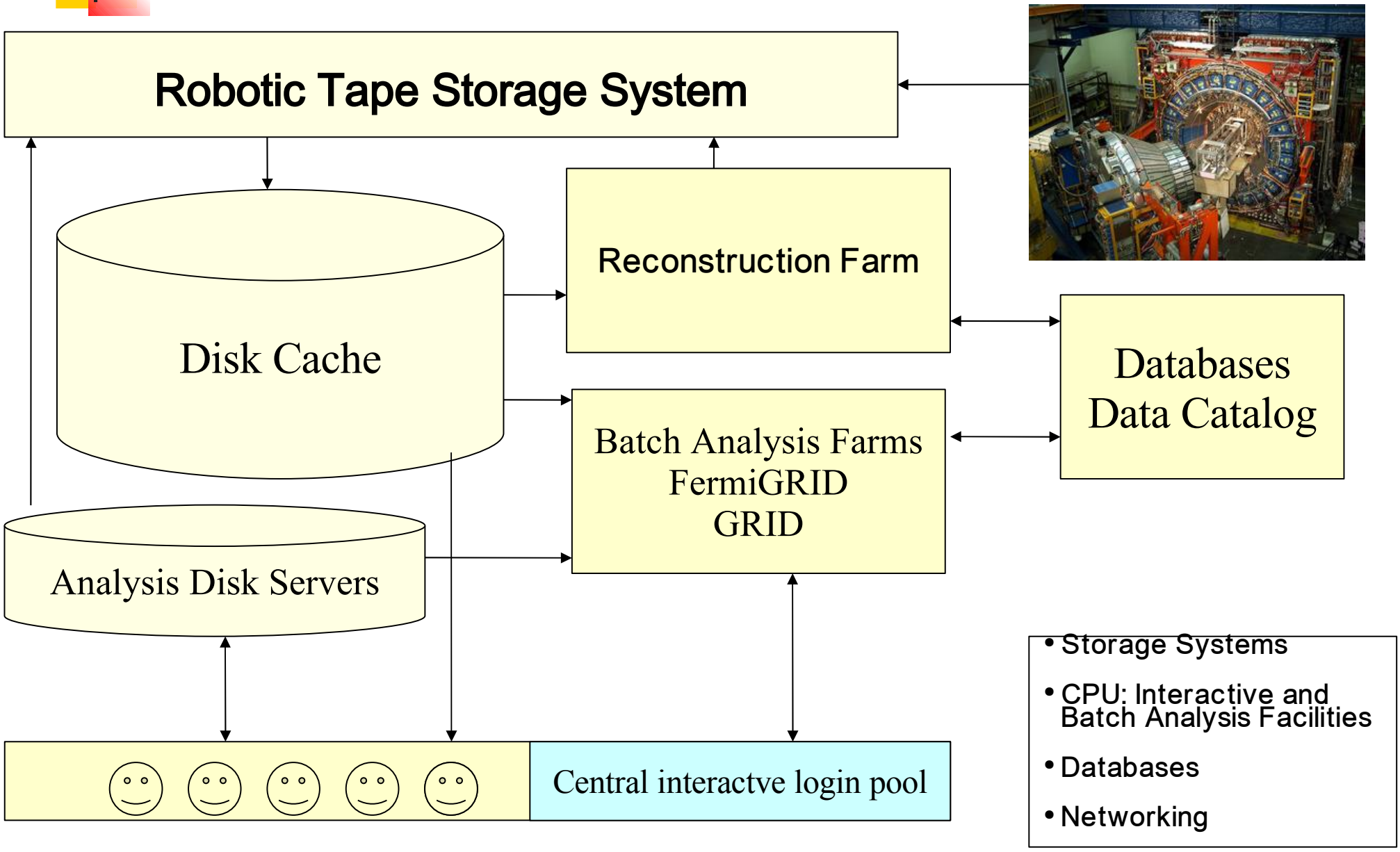
CDF Computing Infrastructure and Budget Request for FY 2006

P.Murat, FNAL

- Outline:
  - CDF computing infrastructure and its elements
  - Modelling of the demands
  - Budget for FY'06 and projections for FY'07-08

CDF Computing Infrastructure



Model of the CDF Computing

Inputs

performance of the Tevatron

- Profiles of the integrated and instantaneous luminosity
- use **design** projections by the accelerator division (8.5 fb^{-1} by 2009)

performance of the CDF detector

- Data logging rate , raw event size

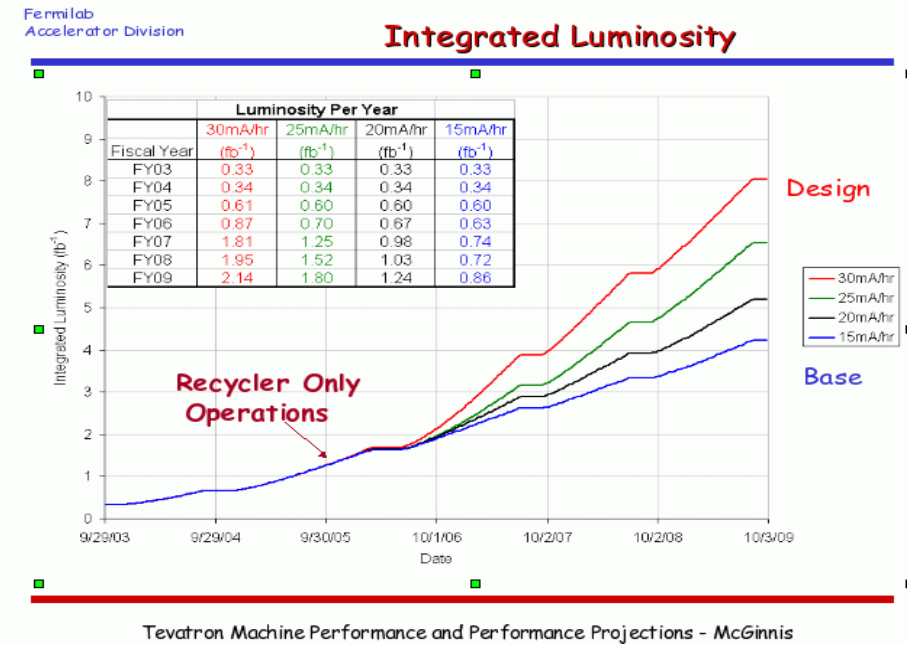
- Data volume scales with the integrated luminosity

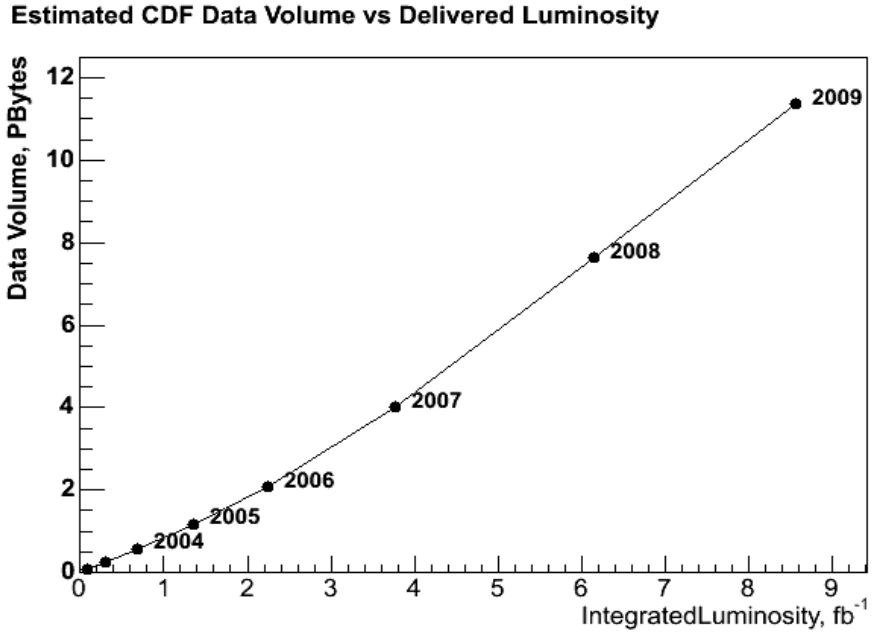
Performance of the reconstruction algorithms

<time/event> , its dependence on L_{inst} (new for this year)

Assumptions about the analysis needs

- Performance of the analysis code
- Disk and CPU usage patterns





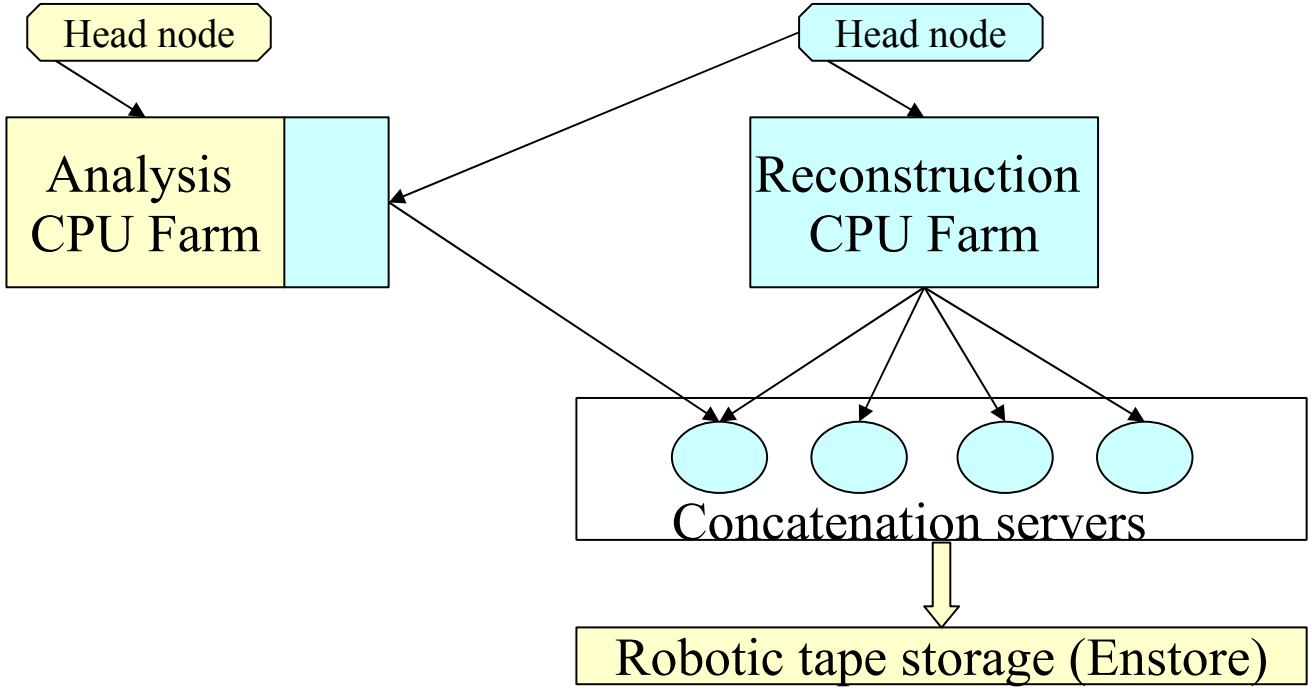
Elements of the Computing Infrastructure

- **CPU :** batch (reconstruction and analysis) and interactive
- **Storage: d**isk, tape
- **The rest:** databases, networking, rest.rest

- CDF approach: unification of the infrastructure elements
  - Design GRID-compatible “building blocks”
  - clone them

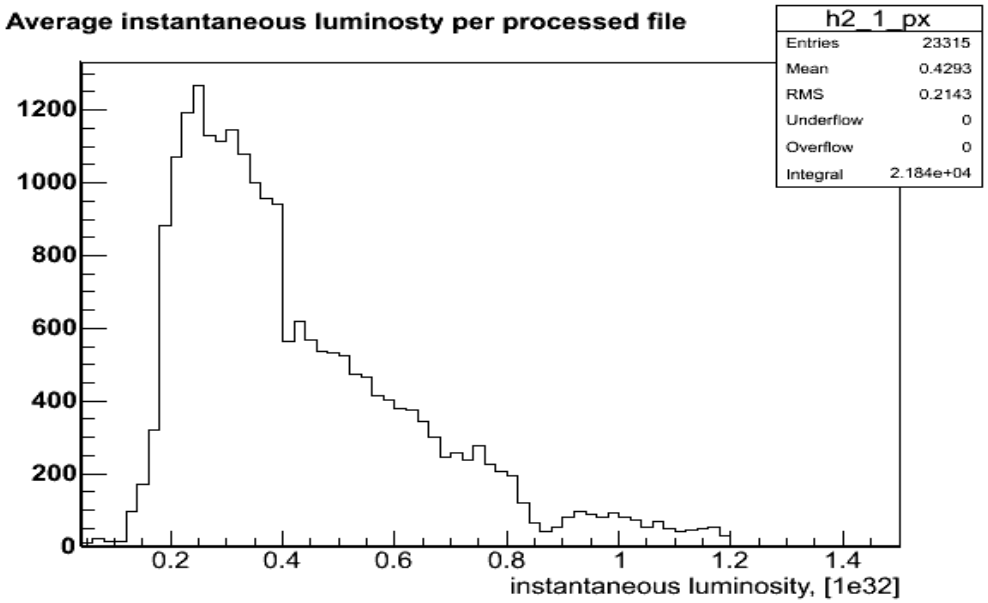
Unified CPU Farm Infrastructure

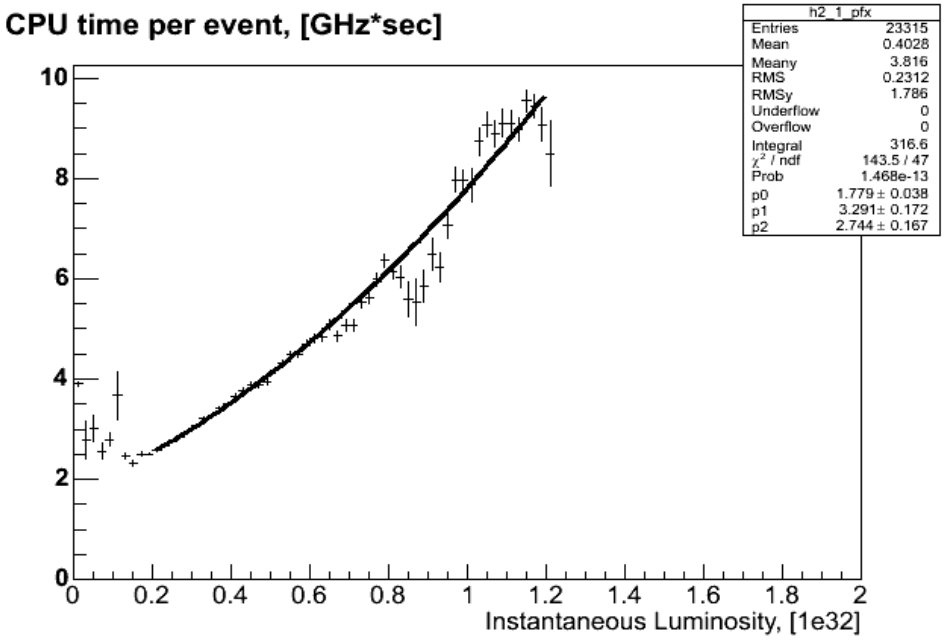
- "Building block" for the CDF batch CPU Farm: **C**omputing **A**nalysis **F**acility (CAF)
  - Batch system: Condor
- **Interoperability**: use fraction of the "analysis" batch farm for offline reconstruction



CPU requirements

- 3 components:
  - Offline reconstruction
  - User analysis
  - MC Production
- Straw-man approach to MC Production '2005:
  - ~600 Mln events
  - <time/event> ~ 30 GHz*sec
  - ~2 THz*months, non-uniform load
- Take MC Production out of the budget equation: it is done off-site
- Offline reconstruction:
  - "1-pass" scheme a success
  - **3.8 Ghz*sec per event in average ===>**
  - Default needs very well predicted, unplanned reprocessings add a small factor of ~ x1.1



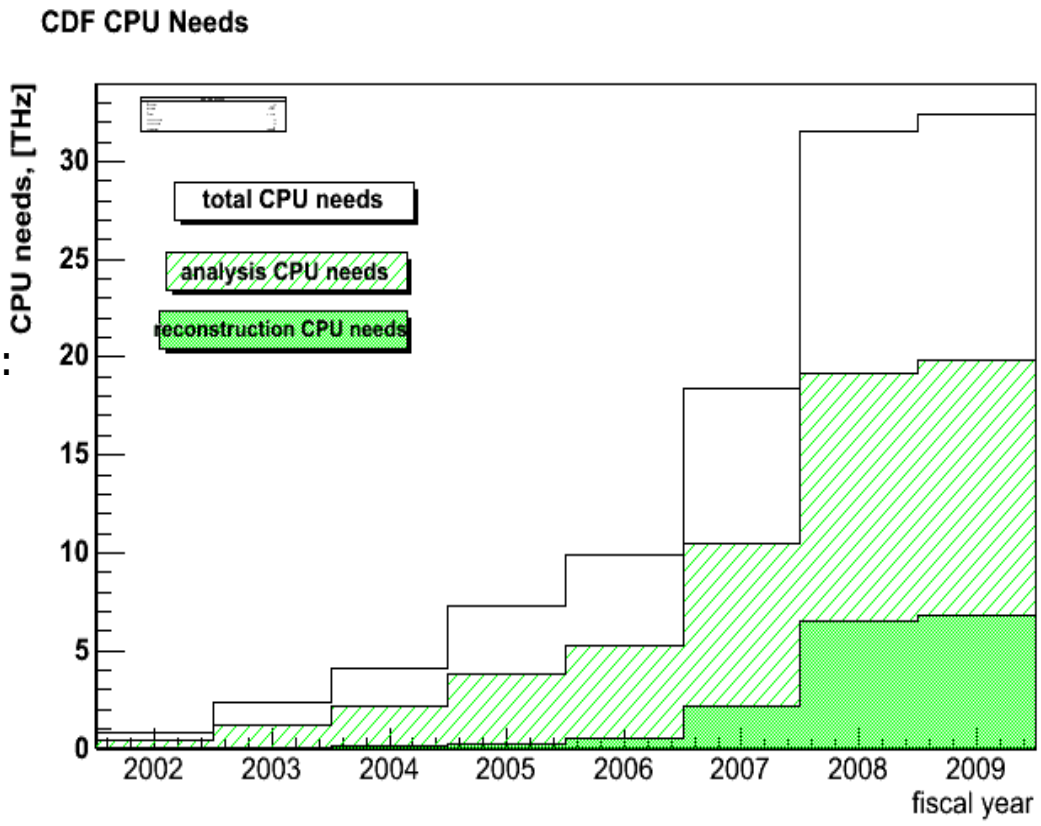


Analysis CPU requirements

- Analysis techniques well established
  - All the physics groups make ntuples
- Highly peaked activity – ntuples need to be made in short time, contingency ~x2 (remake ntuples) is a must
  - physics analysis algorithms are not fast: time per event ~ offline reconstruction

  High-Pt analyses: <time/event> ~ 3GHz*sec

  B-analyses: <time/event> ~ 10 Ghz*sec
- Ntuples made at Fermilab
- Scale analysis CPU needs proportional to the dataset size
- **the total CPU needs are dominated by MC and analysis, reconstruction adds a small fraction**



CPU procurements

	2005	2006	2007	2008	2009	
CPU needs total, (Thz)	7.3	9.9	18.4	31.5		
MC CPU needs (Thz)	3.5	4.6	7.9	12.3		
CPU on-site (Thz)	4.4	6.5	8.4	10.6		
Reco CPU needs	0.52	0.53	2.2	6.5	6.8	
CPU off-site (Thz),exp	0	0	2.1	8.6		
Single CPU clock (Ghz)	3.6	4.6	5.9	7.5		
CPU clock(estimate'2004)	3.9	6.2	9.9			
New nodes	320	300	300	400		
Retired nodes		256	242	244		
Cost per node ($K)	2.2	2.2	2.2	2.2	2.2	
Cost ($K)	704	660	660	880		

- Batch Compute Element infrastructure unified, any GRID site for CDF looks like CAF
- Immediate benefits:

  Reassigning resources of the analysis farm to the production farm (demonstrated)
- Large deviations from Moore's scaling (~x1.6 /year), use x1.3/year
- Used 2004-2005 data to tune the model of the CDF computing needs
  - **estimated needs ~30% lower than last year (data logging rate didn't increase as fast as we expected)**

Storage : Disk and Tape

- **2 very different types of disk storage:**
  - Disk frontend to the robotic tape storage: DCache (**a great success**)
  - Analysis disks : multiple (~10^2) disk servers (hard not even to maintain, but to account for)
- **Is there a unified solution?**
  - Scalable disk pool workshop in March'05 at FNAL:
    - the most promising solution for Run II: DCache-based
- **DCache**: a network-distributed file system with limited access mode
  - extended functionality (emulation of POSIX I/O): can use UNIX commands to work with the files stored in DCache pools
- CDF is prototyping Dcache-based analysis disk pool
  - **Initial success:** logistics works, load tests under way
- **if works, CDF will have a unified GRID-compatible solution for the disk infrastructure**
  - **SRM interface exists**
  - frontend to the tape storage and analysis disk pools based on the same technology
- Risks:
    - performance of the DB, emulating the filesystem
    - Maturity of the Java-based infrastructure

Disk Procurements

	2005	2006	2007	2008	2009	
Disk Needs Total (TB, est)	490	710	1170	2230	3320	
Disk Available	340					
Fileserver size, TB	14	20	35	56	72	
New fileservers	26	23	30	19		
Price per fileserver ($K)	20	20	20	20	20	
Cost ($K)	520	460	600	380		

- Model: required disk volume scales linearly with the dataset size
- Use Moore's scaling (x2 in 18 month)
  - 500 GByte disk drives (Hitachi) just hit the market
  - holds better than for CPU over the last years
- Expected retirement of the 3-year-old fileservers is not happening – will keep them as far as there is enough space and they continue to work
- Retirement – expect ~10-15% effect

Tape Drives Procurement

	2005	2006	2007	2008	2009	
Data volume, PB	1.2	2	4	7.6		
Tape capacity, TB	0.2	0.2	0.4	0.4		
Tape Drives	18B	18B	10X+13B	18X		
New Drives			10X-10B	8X-13B		
Drive bandwidth, MB/sec	30	30	60	60		
Cost, $M		0	0.3	0.24		

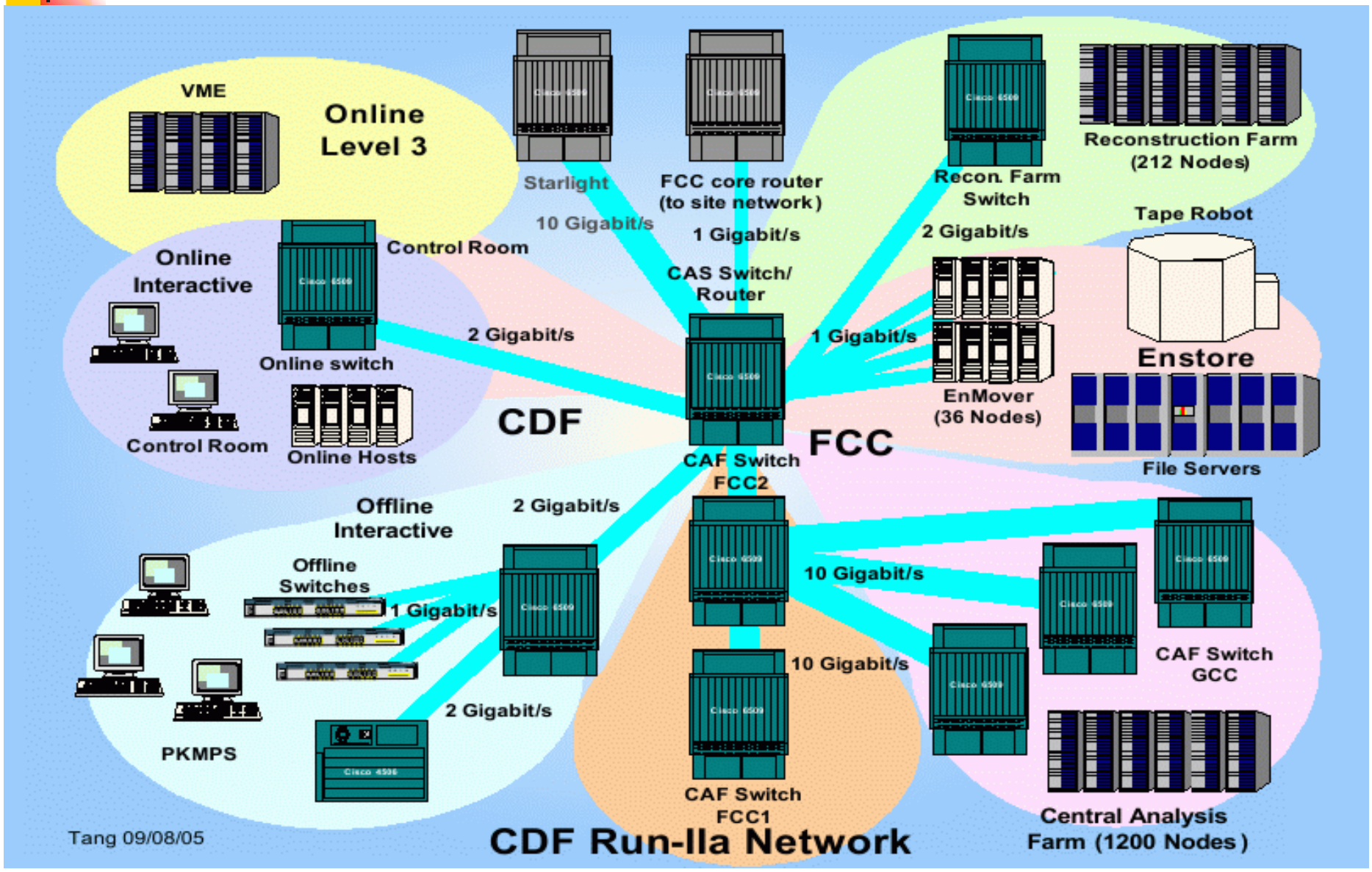
- B = STK 9940B, B*: used STK 9940B;
- X = new tape drives, 400GB/tape, 60 MB/sec (assume $30K/drive)
- By the end of FY'06 expect to fill up both currently used tape libraries (~10900 slots in 2 STK silos)
  - Recycling and compactification of the existing tapes in progress
  - if use additional SILO, need additional tape drives (budgeted)
  - Or migration to new tape technology is needed - **use LTO** ?
- Expect that one-time doubling the tape I/O bandwidth will cover the needs of the experiment through 2009

Interactive systems, DB, miscelaneous

	2005	2006	2007	2008	2009	
DB cost $M	0.05	0.03	0.03	0.03	3320	
Interactive Systems , $M	0.1	0.09	0.02	0.02		
Misc. spendings, $M	0.05	0.05	0.05	0.05	72	
Total, $M	0.2	0.17	0.1	0.1		

- Interactive systems:
  - New code servers : fcdfcode1, ncdf209, ILP disk, new ILP nodes
- Databases
  - after the Frontier servers were bought, just maintenance
- Miscellaneous:
  - Not predicted hardware procurement

CDF Networking'2005



Networking Procurements

	2005	2006	2007	2008	2009.000	
FCC/GCC cost , $M	0.090	0.120	0.040	0.040	3320.000	
CDF B0/Trailers , $M	0.220	0.092	0.040	0.040		
Total, $M	0.310	0.212	0.080	0.080		

- <u>FY'06 networking upgrades:</u>
  - FCC/GCC : 3rd 6509 switch + Starlink upgrade
  - CDF B0/trailers : switches and fabric upgrades

Computing Requirements Summary

	2005	2006	2007	2008	2009	
Integrated luminosity, fb-1	1.35	2.24	3.77	6.14	8.56	
Total N events, (x1e9)	2.02	3.36	5.66	9.21	12.84	
Peak L3 rate (Mbytes/sec)	35	60	60	60	60	
CPU needs total (Thz)	7.3	9.9	18.4	31.5		
MC CPU needs(Thz)	3.5	4.6	7.9	12.3		
Projected CPU on-site (Thz)	4.4	6.5	8.4	10.6		
Offsite analysis CPU needs (Thz)	0	0	2.1	8.6		
Disk (TB)	490 (est)	710	1170	2230		
Tape I/O (GB/sec)	0.54	0.63	0.99	1.08		
Tape volume (PB)	1.3	2	4	7.6		

- MC production done off-site (45-30% of the total needs)
- Offline reconstruction always on-site , using CDF dedicated resources
- Starting from 2007 significant fraction of the analysis jobs need to run off-site (including FermiGRID)

Equipment Summary

	2005	2006	2007	2008
CPU ($M)	0.76	0.69	0.68	0.9
Disk ($M)	0.51	0.5	0.6	0.38
Tape Drives ($M)	0	0	0.3	0.24
Databases and misc. ($M)	0.1	0.05	0.05	0.05
Networking ($M)	0.14	0.21	0.1	0.1
Foreign contributions ($M)	0.23			
Total Cost ($M)	1.28	1.45	1.72	1.66

- Expected aquisition of the new tape drives determines increase in FY'07

Tapes and Operating Budget

	2005	2006	2007	2008	2009	
Data volume, TB	1.2	2	4	7.6		
Collected Data	0.63	0.8	2	3.6		
Tape size, TB	0.2	0.2	0.4	0.4		
New Tapes	3150	4000	5000	9000		
Tape cost ($)	75	75	75	75		
Total Cost ($M)	0.24	0.3	0.38	0.68		
Reference point	8750					
N(tapes) total, est	8560	11500	10000	19000		

- Capacity of 2 CDF STK silos - 10900 tape slots - expected to be exceeded in FY'06
- Contingency: available slots in non-CDF tape library
- The issue of the tape storage will need to be addressed during the next year
- Compress the tapes while migrating to new technology (400 GB tapes ?)

Summary

- FY'06 CDF computing budget request : $1,45 M
  - Improved modelling
  - lower (~20% less than request for FY'05)
- Looking towards the 'end game' scenario. Several important conclusions:
  - CDF primary reconstruction will be always done at FNAL
  - Next tape technology must be chozen to last through 2009
  - Starting from 2007 significant fraction of the CDF analysis jobs will be running on the CPUs not owned by CDF (GRID)
    - Learning how to use the large datasets at the remote GRID sites