CDF Plan and Budget for Computing in Run 2

Robert M. Harris

Directors Review of Run 2 Computing

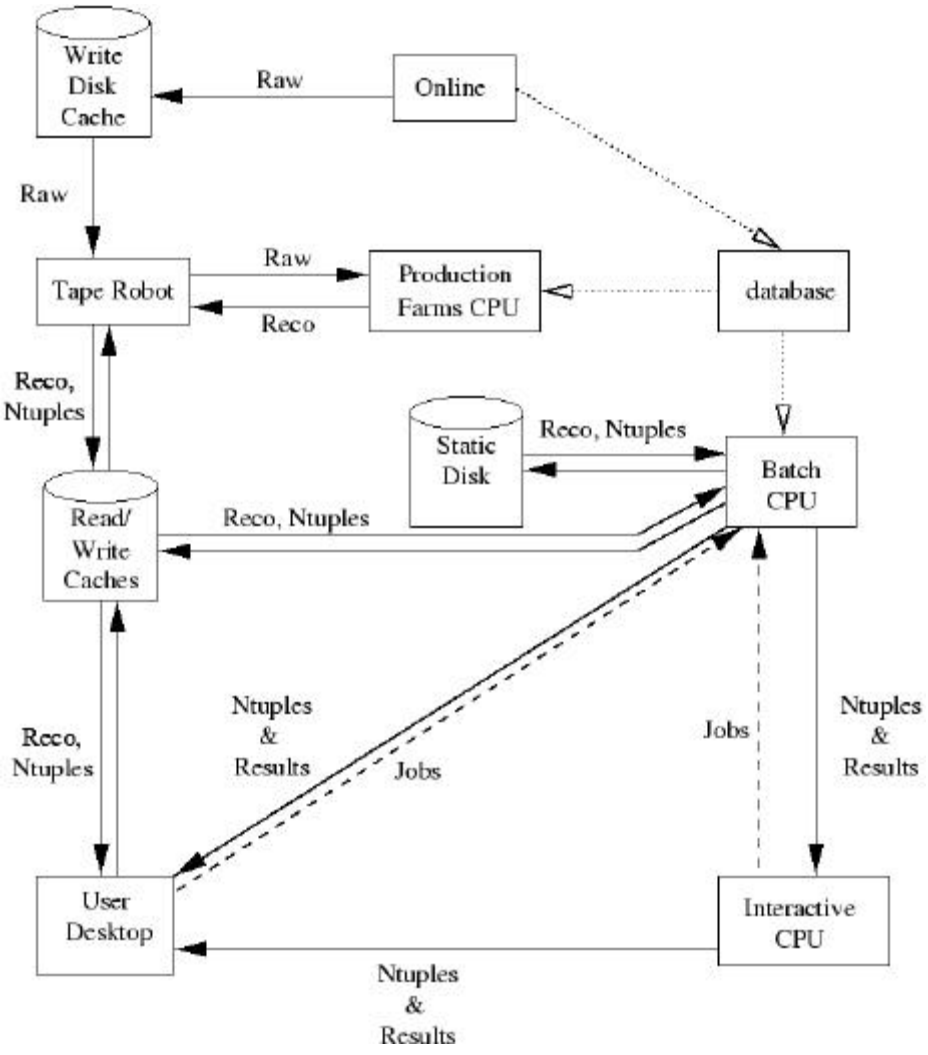
June 4, 2002

Outline

- Overview
  - Summary of computing and requirements models.
  - Required computing and equipment spending FY02-FY05
- Details of spending by sub-system
  - CAF batch CPU
  - CAF static disk
  - Cache disk
  - Tape drives & robot
  - Networking
  - Farms
  - Interactive CPU
  - Databases
  - Tapes and Operating
- Conclusions

Summary of Computing and Dataflow

- Raw Data
  - Written to write cache before being archived in tape robot.
  - Reconstructed by production farms.
- Reconstructed Data
  - Written by farms to tape robot.
  - Read by batch CPU via read cache.
  - Stripped and stored on static disk.
- Batch CPU (CAF).
  - Produces secondary datasets and root ntuples for static disk.
  - Analyzes secondary datasets and ntuples.
- Interactive CPU and desktops
  - Debug, link and send jobs to CAF.
  - Access data from cache and CAF.
  - Write data to robot via cache.
- Database and replicas provide
  - Constants for farms, CAF, users.



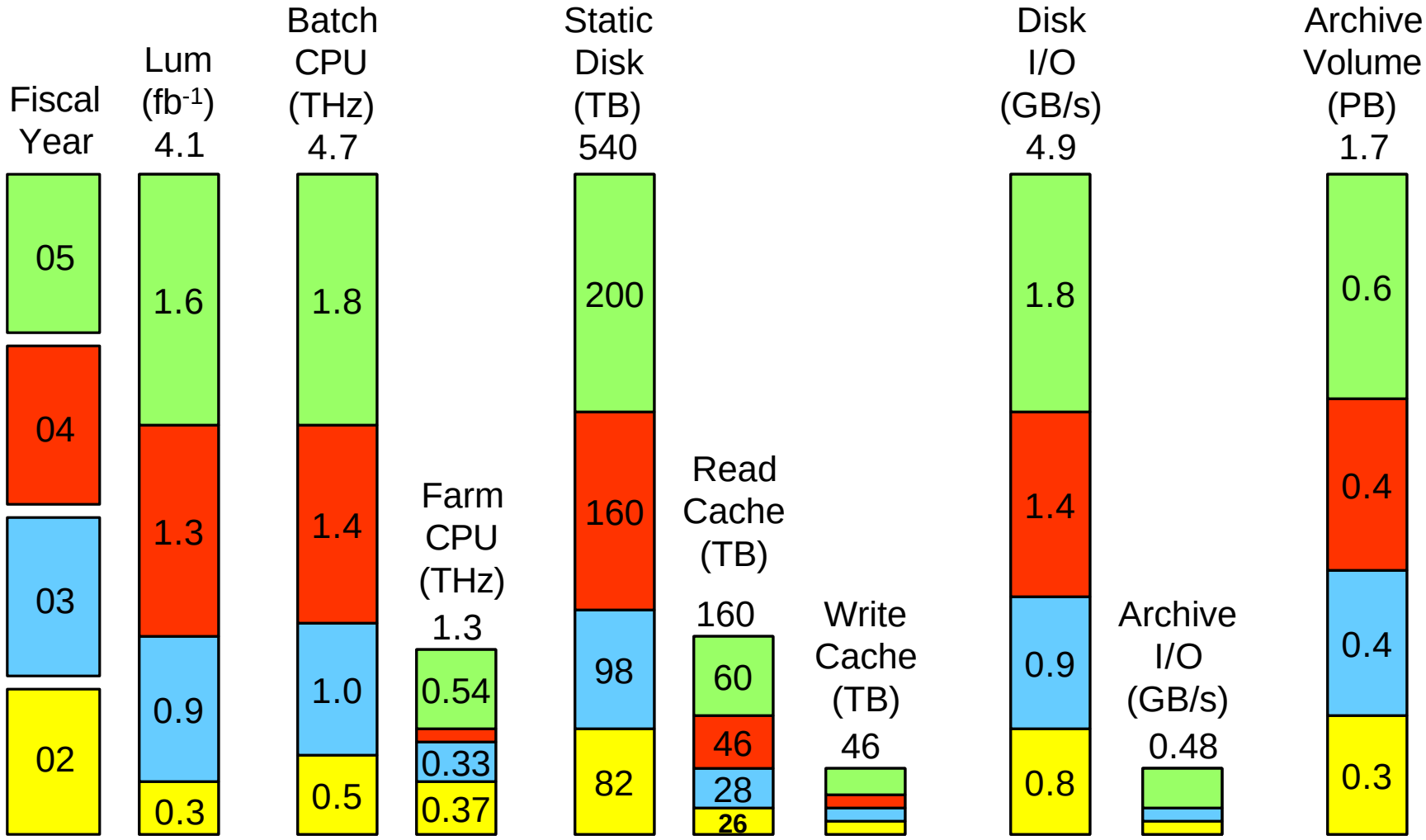
Summary of Requirements Model

- Increases in luminosity will drive computing requirements.
  - Both integrated luminosity and DAQ peak logging rate drive requirements.
  - Model includes offsets to reflect non-scaling needs (commissioning, R&D).

- The CPU and disk needs scale primarily with integrated luminosity.
  - Batch CPU requirements from CAF Review Committee: CDF 5787.
    - 1.1 THz / fb^{-1} allows 200 simultaneous users to process a 5 nb sample in one day.
  - Static disk requirements from physics group requests.
    - 125 TB / fb^{-1} from dataset, MC & other requests or all reco data on disk + 25% extra.
  - Read cache requirements from CDF 5787 assuming a heavy caching model.
    - Large cache is fallback in case static disk requirements are larger than anticipated.

- The tape archive and farms CPU needs scale primarily with DAQ logging.
  - 20 MB/s (80 Hz) upgraded to 60 MB/s (240 Hz) in FY05, and a few details:
    - Uptime of 15%, 30%, 30%, 15% in FY02, 03, 04, 05 (shutdown ½ year).
    - Reprocessings (1→ 0.3) and reconstructed event size (170 → 100 KB).
    - Secondary data, ntuples, and MC data size estimates (~ ¼ of all archive).
    - Processing times for MC and raw data, with room for multiple interactions.
    - Takes into account peak vs. DC needs of the farms.

Required Computing FY02 – FY05



Equipment Spending Plan

FY	CAF batch CPU ($M)	Inter-active CPU ($M)	Farm CPU ($M)	DB ($M)	Robot & Drives ($M)	CAF Disk ($M)	Cache Disk ($M)	Net-Work ($M)	Legacy System ($M)	Total ($M)
2002	0.59	0.07	0.22	0.02	0.77	0.47	0.16	0.25	0.69	**3.24***
Spent	(0.19)	(0.07)	(0.11)	(0.00)	(0.25)	(0.12)	(0.04)	(0.12)	(0.69)	**(1.59)**
2003	0.48	0.15	0.22	0.15	0.35	0.35	0.11	0.25	-	**2.06**
2004	0.48	0.20	0.13	0.10	0.35	0.35	0.11	0.25	-	**1.97**
2005	0.60	0.20	0.19	0.10	0.35	0.35	0.13	0.25	-	**2.17**

- FY02 spending covered by FY02 budget of $2M and $1.2 - 1.4M from misc sources
  - FY01 carryovers of $0.4M, hardware returns of $0.36M, non-Fermilab contributions of $0.27M, anticipated non-Fermilab contributions of between $0.15M and $0.36M.
- FY03 – FY05 spending achieved by Fermilab equipment budget of $2M + misc.
- Not included above is $0.5 M per year operating expenses for tapes & misc.

CAF Batch CPU Spending

Stage	FY	Needs (THz)	Duals Bought	Duals Total	Speed (GHz)	CPU (THz)	Total (THz)	Cost ($M)
1	2002	.1	69	69	1.3	0.18	0.18	**0.19**
2	2002	.5	160	229	1.8	0.58	0.76	**0.40**
3	2003	1.5	192	421	2.5	0.96	1.72	**0.48**
4	2004	2.9	192	613	3.5	1.34	3.06	**0.48**
5	2005	4.7	+240-229	624	5.0	2.40	4.7	**0.60**

- Needs include all batch analysis needs of users and physics groups.
- Each year roughly 200 duals are purchased.
- Dual speed increases with Moore's Law (doubling every 18 months).
- Cost per dual is constant at $2.5K.
- Every 3 years duals are replaced: FY05 duals replace FY02.

CAF Disk Spending

Stage	FY	Needs (TB)	Servers Bought	Servers Total	Server (TB)	Disk (TB)	Total (TB)	Cost ($M)
1	02	~30	11	11	1.9	21	21	**0.12**
2	02	82	32	43	2.2	72	93	**0.35**
3	03	180	32	75	3.5	112	205	**0.35**
4	04	340	32	107	5.5	176	381	**0.35**
5	05	540	+32-43	96	8.7	278	558	**0.35**

- Needs are disk requests from the physics groups.
  - Enough to hold all reconstructed data on disk plus 25% contingency (100KB/ev)
- Each year 32 servers are bought (except 43 in FY02)
- The server capacity grows with Moore's law (doubling every 18 months).
- The server price remains roughly $10K
- Every 3 years servers removed: servers bough in FY02 removed in FY05.

Cache Disk

	STG	FY	Needs (TB)	File Servers	FS Size (TB)	Added (TB)	Total (TB)	Cost ($M)
READ CACHE	1	02	12	4	1.9	8	8	**0.04**
	2	02	26	8	2.2	18	26	**0.08**
	3	03	54	8	3.5	28	54	**0.08**
	4	04	100	8	5.5	44	98	**0.08**
	5	05	160	+10 -12	8.7	60	159	**0.10**
WRITE CACHE	2	02	9	4	2.2	9	9	**0.04**
	3	03	17	3	3.5	11	20	**0.03**
	4	04	30	3	5.5	16	36	**0.03**
	5	05	46	+3 -4	8.7	17	53	**0.03**

- Needs include caching of all data from and to tape robot.
- The server capacity grows with Moore's law. cost remains roughly $10K.
- Every 3 years servers are removed: servers bought in FY02 removed in FY05.

Tape Drives & Robot Spending

Stg	FY	Data (PB)	Drives Bought	Tape (GB)	Rate (MB/s)	Drives in each robot	Total (PB)	Total (MB/s)	Cost ($M)
1	02	0.1	10A	60	10	10A	0.33	100	**0.25**
2	02	0.3	10A	60	10	10A 10A	0.66	200	**0.52**
3	03	0.7	10B	200	30	20A 10B	1.43	500	**0.35**
4	04	1.1	10B –20A	200	30	10B 10B	2.2	600	**0.35**
5	05	1.7	10C	400	60	20B 10C	3.3	1200	**0.35**

- Driven by storage needs. I/O needs, not shown, exceeded by factor of 2-3.
  - FY2002 we purchase 2nd STK robot. Each year add 10 drives and R&D of $50K.
  - Copy data from older to newer drives for larger archive capacity (conserve footprint).
- The drive performance follows the STK long range roadmap
  - The B drive (T9940B) triples the A drive in performance as early as FY02.
  - The C drive is assumed to only have twice B's performance.
  - Decommission drives when they become 2 generations old.

Network Spending

Stage	FY	FE Ports Buy	FS	GigE Ports / FS	GigE Ports Total	Switch Buy	Cost ($M)	Misc ($M)	Total ($M)
1	02	96	16	1	16	up	0.04	0.07	**0.11**
2	02	192	44	1	60	1	0.14	0.00	**0.14**
3	03	192	43	1	103	1	0.10	0.15	**0.25**
4	04	192	43	2	189	1	0.11	0.14	**0.25**
5	05	240	+45-60	2	219	1	0.07	0.18	**0.25**

- Roughly half the purchases will be for the CAF, driven by connect needs.
- 1 FE port per worker node and 1-2 GigE port per fileserver
- Cost $0.25K / FE, $1K / GigE and Moore's law decrease.
- Large fraction of miscellaneous cost will be for trailers and B0 to FCC link.

Farms Spending

FY	Needs (GHz)	Duals	Total Duals	Type (GHz)	Total (GHz)	Cost ($M)
02 (may)		32	169	1.3	294	**0.12**
02 (end)	370	32	201	1.8	410	**0.1**
03	700	+68 - 50	219	2.5	700	**0.22**
04	760	+32 – 23	228	3.5	888	**0.13**
05	1300	+54 – 64	218	5.0	1300	**0.19**

- Needs include raw data processing, reprocessings (1→ 0.3) and MC.
  - Takes into account peak vs. DC needs and slowdown due to multiple interactions.
- Each year roughly 50 duals purchased.
- Dual speed increases with Moore's law.
- Cost is $2.5K/node + networking and misc.
- Every 4 years duals are removed (farms started in 1999).

Interactive CPU

FY	Legacy Processors	Interactive CPU ($M)	Local Disk ($M)	Total ($M)
02	128	0.07	0.00	**0.07**
03	64	0.10	0.05	**0.15**
04	32	0.10	0.10	**0.20**
05	0	0.10	0.10	**0.20**

- The plans for interactive CPU are still being developed.
  - Perhaps an interactive login pool of duals with a method of sharing disk.
  - Perhaps fewer numbers of n-ways (8-ways or more).
- Plan to decommission fcdfsgi2 over roughly 2 years
- Will require additional Linux interactive CPU as fcdfsgi2 decreases.
- The budget for CPU is roughly constant and smaller than batch CPU.
- Costs of local disk increase that users share interactively increase.

Database

FY	DB CPU (n-ways)	DB Disk (TB)	Cost ($M)
02	Reuse	2	**0.02**
03	3	6	**0.15**
04	2	6	**0.10**
05	2	9	**0.10**

- In FY02 re-using a Linux 4-way as our first Linux DB replica.
  - To serve the needs of the CAF, trailers and universities.
  - Suns remain to serve the production farms.
- Plan to add three 4-ways with 2 TB of disk in FY03
  - DB replica dedicated to the trailers and universities.
  - Replacing the aging Sun DB machines
- More replicas anticipated for FY04 and FY05

Tapes and Operating

FY	Archive (PB)	AIT-2 (PB)	T9940A (PB)	T9940B (PB)	T9940C (PB)	Tape ($M)	Misc ($M)	Cost ($M)
02	0.3	0.1	0.3	-	-	**0.52**	0.15	**0.67**
03	0.7	-	0.2	0.7	-	**0.54**	0.15	**0.69**
04	1.1	-	-	0.4	-	**0.16**	0.15	**0.31**
05	1.7	-	-	0.3	1.7	**0.46**	0.15	**0.61**

- Cost per GB for tape media is (AIT-2, A, B, C) = $(1.3, 1.3, 0.4, 0.2)
- Cost includes old datasets copied to higher media densities.
  - In FY02 we copied 0.1 PB from AIT-2 to T9940A. 0.1 PB on both media.
  - In FY03 we plan to copy 0.5 PB from T9940A to cheaper B densities
  - In FY05 we copy 1.4 PB from T9940B to cheaper C densities.
  - Old datasets can be stored for data security or reused to save media costs (A→B).
- Another $0.15 M per year for racks, cards, installs, FNAL desktops, etc.
  - Total operating averages to $ 0.54 M / year for FY03 – FY05
  - With large year to year fluctuations depending on tape copying activity.

Conclusions

- Increases in luminosity will drive computing requirements.
  - The CPU and disk needs scale with integrated luminosity.
  - The tape archive and farms needs scale with DAQ logging capability.
- Computing procurements are planned to meet the needs.
  - Hundreds of PCs analyzing ~200 TB of data on network attached IDE disk.
  - Fast network attached tape drives writing hundreds of GBs to each tape.
- About $2 M equipment and $0.5 M operating needed per FY.
  - This is the anticipated budget coming from Fermilab.
- Budget not dominated by any single system.
  - ~ $0.8M / FY tapes, tape drives and tape robots.
  - ~ $0.7M / FY CPU of all kinds.
  - ~ $0.6M / FY disk of all kinds.
  - ~ $0.25M / FY networking.
  - ~ $0.15M/ FY miscellaneous operating.