CDF Requirements and Budget for Computing in Run 2

Robert M. Harris

Directors Review of Run 2 Computing

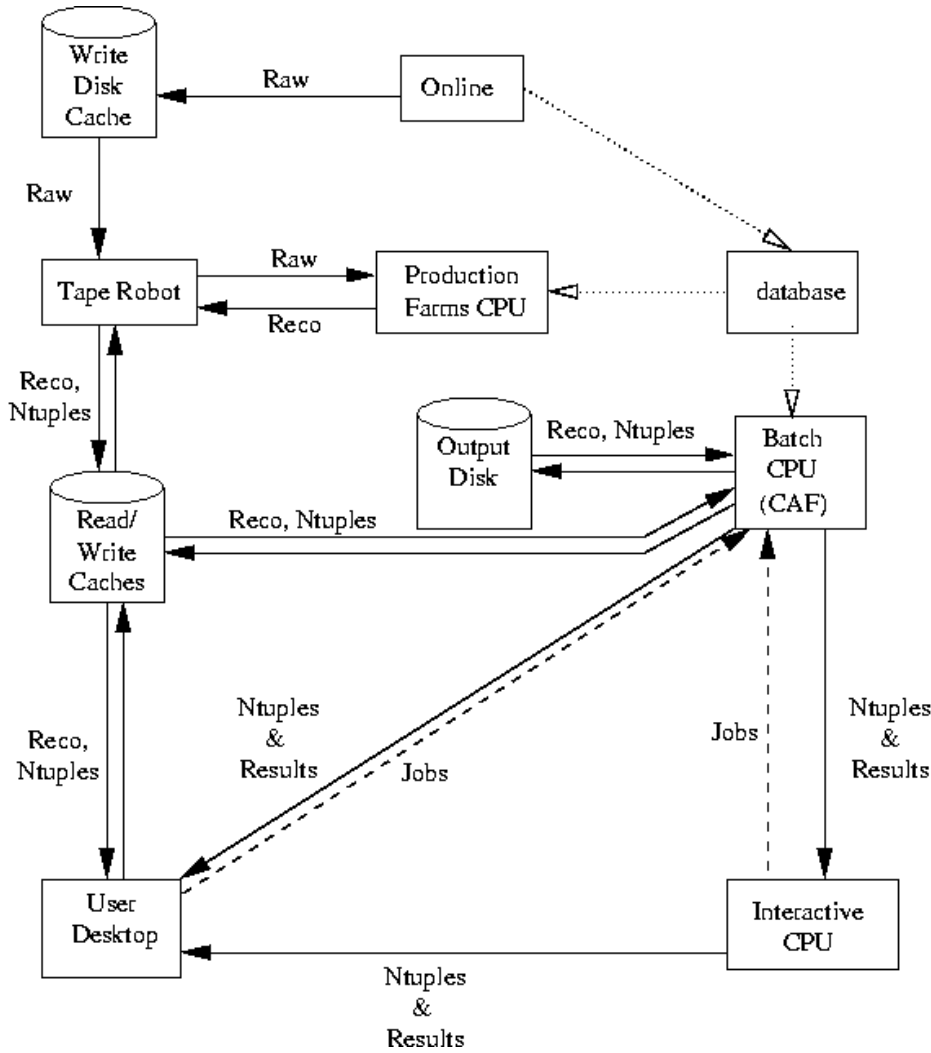
Sep 11, 2003

Outline

- Overview
  - Summary of computing and requirements models.
  - Required computing and equipment procurements FY02-FY06
- Details of procurement plan by sub-system
  - CAF batch CPU
  - CAF Cache Disk
  - Tape drives & robot
  - Networking
  - Farms
  - Databases, Interactive CPU & Miscellaneous.
  - 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 output 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.



Requirements Models Introduction

- Requirements Models
  - CDF 5914 from one year ago assumed analysis requirements scale with luminosity.
    - Based on analyzing high Pt datasets.
  - CDF 6639 and 6640 adds requirements that scale with events logged.
    - Analysis of large datasets, for example for Bs mixing and high statistics physics.
    - The added requirements produce additional costs not included in CDF 5914.

- Increases in raw data logging drives increases in the requirements and the cost.
  - FY04: Drop of event size allows 50% increase in event logging.
  - FY05: Upgrade to CSL will further double the bandwidth capability of data logging.
  - FY06: Upgrade to DAQ will increased the bandwidth capability another 50%.

- Errata
  - Silicon Upgrade
    - CDF 6639 and CDF 6640 assumed a 6 month shutdown in FY06 to upgrade the silicon.
    - This reduces the amount of data and the cost in FY06.
    - On September 2 we learned that the silicon upgrade will not take place.
    - Here I present new requirements and costs for FY06 without a silicon upgrade.
      - **More data taking efficiency in FY06 translates into more computing costs.**
  - The tables presented here supersede those in CDF 6640.
    - I found and fixed a few minor numerical errors.

Requirements Model

- Data logging model
  - Upgrades in each of FY04, FY05 and FY06 as mentioned.
  - Uptime of 30% each FY (No silicon upgrade) and data logging at 70% of peak capacity.
- Analysis CPU needs scale with both luminosity and the number of events.
  - High Pt analysis: ~ THz / fb^{-1} allows 200 users to process a 5 nb sample in one day.
    - Model predicts CPU usage during Winter 2003 conferences within a factor of 2. Model is low.
    - Model accurately predicts I/O used by the CPUs during Winter 2003 conferences.
  - Large dataset analysis: 15 users analyzing the non-high Pt dataset in 25 days.
- Disk needs scale with the number of events.
  - High Pt: ~ 0.1 PB / Giga-events from scaling usage during Winter 2003 conferences.
  - Large Dataset Analysis: Enough disk for seven days of processing on the CAF.
- Tape archive needs scale primarily with the data logging bandwidth.
  - The tape archive I/O requirements are for raw data, reconstruction farms, and CAF.
    - Dominated by the CAF I/O needs for user analysis and secondary dataset production.
  - The tape archive volume includes raw, produced, and secondary data.
- Reconstruction farms needs are from CDF 6640, slightly different than CDF 6639.
  - A CPU time of 5 sec/event on a 1 GHz processor allows for code slowdown with lum.

Required Computing FY02 – FY06

Assumed Conditions					Total Requirements (02-03 old, 04-06 new)				
FY	Int. L (fb^{-1})	Int. Evt (10^9)	Peak (MB/s)	Peak (Hz)	CPU (THz)	Reco (THz)	Disk (PB)	Tape I/O (GB/s)	Tape Vol (PB)
02A	0.08	0.1	20	80	0.5	0.4	0.1	0.1	0.2
03A	0.30	0.6	20	80	1.5	0.5	0.2	0.2	0.4
04E	0.68	1.4	20	120	3.7	0.8	0.3	0.6	0.7
05E	1.35	3.0	40	240	9.0	1.4	0.6	1.4	1.6
06E	2.24	5.4	60	360	16.5	2.0	1.1	2.7	2.8

- FY04-06 data logging upgrades increase events logged each FY.
  - ➔ Numerical coincidence that events scale roughly with integrated luminosity.
- FY04-06 analysis CPU, disk, I/O and tape needs scale roughly with events.

Equipment Plan: Past and Future

FY	CAF CPU ($M)	Inter. CPU ($M)	Farm CPU ($M)	DB ($M)	Tape Drives ($M)	Cache Disk ($M)	Net-Work ($M)	Legacy System ($M)	Total ($M)
01 A	-	-	0.25	-	-	-	-	0.75	**1.0**
02 A	0.39	0.07	0.22	0.02	0.77	0.63	0.25	0.69	**3.0**
03 A	0.31	0.08	0.13	0.14	0.20	0.34	0.23	-	**1.4**
04 E	0.76	0.12	0.19	0.10	0.27	0.20	0.25	-	**1.9**
05 E	1.16	0.10	0.19	0.10	0.57	0.64	0.19	-	**3.0**
06 E	1.03	0.10	0.19	0.10	0.78	0.56	0.12	-	**2.9**

- Spending Drivers
  - FY01 we tasted run 2 needs, changed computing model, held $1M out of $2M.
  - FY02 first buildup of CAF & Enstore allows for surge in spending to meet need.
  - FY03 reduced Fermilab budget prevents a significant increase in CPU.
  - FY04 increase of events logged & large dataset analysis increases CPU costs.
  - FY05 doubling of data logging increases CPU, tape drive & disk costs.
  - FY06 DAQ upgraded increases bandwidth by 50% and costs remain high.

CAF CPU Procurements: Fermilab

FY	Needs (THz)	Duals Bought	Duals Total	Speed (GHz)	CPU (THz)	Total (THz)	Cost ($M)
02 A	.5	179	179	1.3-1.7	0.58	0.58	**0.39**
03 A	1.5	159	338	2.2	0.70	1.28	**0.31**
04 E	3.7	346	674	3.5	2.42	3.70	**0.76**
05 E	9.0	+525-179	1030	5.6	5.88	9.00	**1.16**
06 E	16.5	+466-159	1337	8.8	8.20	16.5	**1.03**

- Cost Calculation
  - Cost per dual is constant at $2.2K for FY04-FY06. Speeds are PIII equiv.
  - Dual speed increases with Moore's Law (doubling every 18 months).
  - Every 3 years duals are replaced.
- Cost Drivers
  - High Pt dataset analysis drove FY02-03 needs, roughly met by Fermilab spending.
  - Increased event logging leads to large CPU costs in FY04-06.

CAF CPU Procurements: Non-Fermilab

FY	Duals Bought	Duals Total	Speed (GHz)	CPU (THz)	Total (THz)	CAF Frac	Cost ($M)
02 A	117	117	1.3-1.7	0.37	0.37	**0.39**	**0.26**
03 A	63	180	2.2	0.28	0.65	**0.34**	**0.12**
04 E	30	210	3.5	0.21	0.86	**0.19**	**0.07**
05 E	+30-117	123	5.6	0.34	0.89	**0.09**	**0.07**
06 E	+30-63	90	8.8	0.53	1.14	**0.06**	**0.07**

- FY02-03 actual non-Fermilab spending boosted size of CAF significantly.
  - Remote institutions were provided with strong resource incentives to contribute.
  - Requirements were covered by Fermilab, but extra capacity was very helpful.
- FY04-06 estimates assume contributions will go down.
  - Emerging grid makes remote computing facilities appear more attractive.

CAF Disk Procurements: Fermilab

FY	Needs (TB)	Servers Bought	Servers Total	Server (TB)	Disk (TB)	Total (TB)	Cost ($M)
02A	82	57	57	2	114	114	**0.63**
03A	180	18	75	5	90	204	**0.34**
04E	285	10	85	8	80	284	**0.20**
05E	604	+32-57	60	13	416	606	**0.64**
06E	1082	+28-18	70	20	560	1076	**0.56**

- Cost Calculation
  - Cost per fileserver is constant at $20K for FY04-FY06. Cost in FY03 scaled to middle FY.
  - Fileserver capacity increases with Moore's Law (doubling every 18 months).
  - Every 3 years fileservers are replaced.
- Cost Drivers
  - High Pt dataset analysis drove FY02-03 needs which were met by Fermilab spending.
  - Increases in data logging leads to large disk costs in FY05 and FY06.

CAF Disk Procurements: Non-Fermilab

FY	Servers Bought	Servers Total	Server (TB)	Disk (TB)	Total (TB)	CAF Frac.	Cost ($M)
02A	35	35	2	70	70	**0.38**	**0.35**
03A	4	39	5	20	90	**0.31**	**0.07**
04E	2	41	8	16	106	**0.27**	**0.04**
05E	+2 -35	8	13	26	62	**0.09**	**0.04**
06E	+2 -4	6	20	40	82	**0.07**	**0.04**

- FY02-03 actual non-Fermilab spending boosted size of CAF significantly.
  - Remote institutions were provided with strong resource incentives to contribute.
  - Requirements were covered by Fermilab, but extra capacity was very helpful.
- FY04-06 estimates assume contributions will go down.
  - Emerging grid makes remote computing facilities appear more attractive.

Tape Drive Procurements

FY	Need (PB)	Need (MB/s)	Drives Bought	Tape (GB)	Rate (MB/s)	Drives in robots	Total (PB)	Total (MB/s)	Cost ($M)
02	0.2	100	10A + 10B	60	10 - 30	10A + 10B	1.1	400	**0.77**
03	0.4	190	3B – 10A	200	30	13B	2.2	400	**0.20**
04	0.7	640	9B	200	30	22B	2.2	660	**0.27**
05	1.6	1432	19X – 11B	400	60	11B + 19X	3.3	1470	**0.57**
06	2.8	2690	26X – 11B	400	60	45X	4.4	2040	**0.78**

- Cost Calculation
  - $30K per drive assumed in each of FY04-06 from B drive cost in FY03.
  - We assume an X drive available in FY05 with twice the I/O and media density of B drives.
- Cost Drivers
  - Driven by storage needs in FY02-03, and mainly by I/O needs in FY04-06.
  - In FY05 we either must upgrade to X as shown or buy more robots for archive capacity.

Network Procurements

FY	FCC Cost ($M)	Trailer Cost ($M)	Total Cost ($M)
02	0.25	-	**0.25**
03	0.23	-	**0.23**
04	0.14	0.11	**0.25**
05	0.09	0.10	**0.19**
06	0.06	0.06	**0.12**

- FCC Network Spending
  - Driven primarily by CAF size expansions.
  - Moore's law drop in cost by factor of 2 every 18 months.
- Trailer Network Spending
  - 3 year staged upgrade for gigabit to desktops in trailers.
    - New trailer switches, modules, and 10 Gb between all switches and FCC.

Farms Procurements

FY	Needs (GHz)	Duals	Total Duals	PIII Speed (GHz)	Total (GHz)	Cost ($M)
02	370	64	201 (185)	1.3	403	**0.25**
03	480	+64 - 73	176	2.2	525	**0.19**
04	800	+64 - 64	176	3.5	805	**0.19**
05	1400	+64 - 64	176	5.6	1264	**0.19**
06	2000	+64 – 64	176	8.8	2146	**0.19**

- Cost Calculation
  - Cost per dual is constant at $2.2K for FY04-FY06. Speeds are PIII equiv.
  - Dual speed increases with Moore's Law (doubling every 18 months).
  - Every 3 years duals are replaced.
- Cost Drivers
  - Reprocessing decreases steadily from 1 in FY03 to 0.2 in FY06.
    - Compensates for increase in events logged in FY04-FY06, keeps costs constant.

DB, Interactive CPU & Miscellaneous

FY	DB CPU Added	Cost ($M)
02	0	**0.02**
03	3	**0.15**
04	2	**0.10**
05	2	**0.10**
05	2	**0.10**

FY	Int. CPU ($M)	Misc ($M)	Total ($M)
02	0.07	0.00	**0.07**
03	0.06	0.02	**0.08**
04	0.07	0.05	**0.12**
05	0.07	0.03	**0.10**
06	0.07	0.03	**0.10**

- Databases: following replication strategy.
  - In FY03 used two existing Linux nodes as replicas. Bought SUN production DB.
  - In FY04 add 2 machines: replace aging fcdflnx1 replica, add a new replica.
- Interactive CPU and Miscellaneous
  - Decommission fcdfsgi2 in FY04: first halving in January.
  - In FY03 began purchasing for Linux login pool that reuses fcdfsgi2 FC disk.
  - Plan to scale up CPU/disk each fiscal year as necessary.
  - Miscellaneous costs as well, like new code build node in FY04.

Tapes and Operating

FY	Archive (PB)	AIT-2 (PB)	T9940A (PB)	T9940B (PB)	X (PB)	Tape ($M)	Misc ($M)	Cost ($M)
02A	0.2	0.1	0.23	-	-	**0.42**	0.15	**0.57**
03A	0.4	-	0.22	0.24	-	**0.18**	0.18	**0.36**
04E	0.7	-	-	0.90	-	**0.00**	0.18	**0.18**
05E	1.6	-	-	0.90	1.9	**0.35**	0.18	**0.53**
06E	2.8	-	-	-	3.4	**0.30**	0.18	**0.48**

- Tapes
  - Cost per GB for tape media is (AIT-2, A, B, X) = $(1.3, 1.3, 0.4, 0.2) + 20% contingency.
  - In FY03 we wrote A media, migrated to B a year early, copied half archive from A to B.
  - In FY04 we will complete recycling of A into B tapes and won't spend on tapes.
  - In FY05 we will begin employing cheaper X tapes: can't recycle B media, and lots of data.
- Another $0.18 M per year for racks, installs, FNAL desktops, consultants, etc.
  - Total operating averages to $ 0.4 M / year for FY04 – FY06
  - With large year to year fluctuations due to tape recycling and new media introductions.

Conclusions

- Computing requirements will scale with the size of the run 2 dataset.
  - Increased data over next 3 years will require ~10 times more computing.
  - Moore's law should prevent the cost from exploding.

- Computing procurements required to meet CDF needs
  - $2M in FY04: driven by increased CPU to analyze extra events logged.
  - $3M in FY05: from an additional doubling in data logging (CSL upgrade).
  - $3M in FY06: from 50% increase in data logging (DAQ upgrade).
  - Additional operating expenses of roughly $0.4 M per FY.

- Budget weighted towards analysis CPU
  - ~ 40% analysis CPU
  - ~ 20% tape drives
  - ~ 15% disk
  - ~ 10% networking
  - ~ 10% reconstruction farm CPU
  - ~ 5% miscellaneous