

Computing Costs and Schedule



Charles Young, SLAC

Gilchriese Review Committee

October 11-12, 2000

Outline



- Scaling with luminosity.
 - Peak or integrated.
- Projections for CPU and total storage needs.
 - Production activities, e.g. prompt reconstruction.
 - Total data storage.
 - Analysis activities.
- Tape, disk and CPU costs.
- Profile of costs through FY03.
- Contingency plan.
- Schedule
- Summary

Luminosity Assumptions

- Luminosity profile from John Seeman.
 - Peak and integrated.

	<i>Unit</i>	<i>FY00</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Peak</i>	$10^{33} \text{ cm}^{-2} \text{ sec}^{-1}$	2.5	5	8	10
<i>Year Integral</i>	fb^{-1}	25	40	80	115
<i>Total Integral</i>	fb^{-1}	25	65	145	260

Scaling Assumptions

- Production needs.
 - Keep up with data taking.
 - Current capacity matched to current *peak* luminosity.
 - Projections of processing based on *peak* luminosity.
- Projections of total storage based on *integrated* luminosity.
 - Tape storage.
- Analysis needs.
 - Analysis uses all the data.
 - Current CPU capacity matched to current *integrated* luminosity.
 - Disk storage for analysis based on *integrated* luminosity.
 - Modified by decreasing fractions of data on disk.

CPU and Storage Projections

- Production activities.
 - Prompt reconstruction.
 - Disk buffer and CPU.
 - Reprocessing, simulation production, data distribution.
- Tape storage for all data.
- Analysis activities.
 - Disk cache.
 - CPU.

Projections for Prompt Reconstruction

- Scales with *peak* luminosity.
 - Weekly or monthly integral is better.
- Current farm sufficient for $\sim 2.5 \cdot 10^{33} \text{ cm}^{-2} \text{ sec}^{-1}$.
 - No headroom at present.
 - Code speed-up and greater operational efficiency should help.
 - Counteracted by more sophisticated reconstruction.
 - Assume no net change.
- Current farm components.
 - 5 TB of disk of disk buffer.
 - 145 processing nodes.
 - Several small servers.
 - Accounted for in projections as 5 more processing nodes.

Projections for Prompt Reconstruction

	<i>Unit</i>	<i>FY00</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Peak Luminosity</i>	$10^{33} \text{ cm}^{-2} \text{ sec}^{-1}$	2.5	5	8	10
<i>Percentage increase</i>			100%	60%	25%
<i>Disk Space</i>					
<i>Existing</i>	TB		5	10	16
<i>Additional</i>	TB		5	6	4
<i>CPU</i>					
<i>Existing</i>			150	300	480
<i>Additional</i>			150	180	120

From previous page

Prompt Reconstruction Program Speed

- Faster program speed means fewer nodes.
 - Comparison with Belle as reality check.
- Inexact comparison.
 - “Reconstruction” means different things to different experiments.
 - BaBar includes some physics reconstruction and calibration.
 - Physics, such as D^* reconstruction, takes 12-15% of time.
 - Normalizing different CPUs.
- BaBar and Belle comparison in 8/00.
 - BaBar -- 145 CPUs, $\sim 130 \text{ pb}^{-1} / \text{day}$.
 - Belle -- 84 CPUs, $80\text{-}90 \text{ pb}^{-1} / \text{day}$.
 - Comparably rated CPUs.
 - ~ 1 CPU-day per pb^{-1} of data for both BaBar and Belle.

Projections for Other Production Activities

■ Reprocessing.

- Identical but separate setup as in Prompt Reconstruction.
- Currently 5 TB of disk buffer and 145 processing nodes.
- Scale up as in PR.

■ Simulation Production.

- Actual BaBar-wide output so far less than necessary.
- Current *capacity* reasonable match for present data rate.
 - Including new but not fully operational sites.
 - Including recent additional nodes at SLAC.
- SLAC has 1 TB and recently upgraded to 100 nodes.
 - Need to increase buffer space to match CPU.
- Scale up current capacity (at all sites) by peak luminosity.

Projections for Other Production Activities

■ Data Distribution.

- Discrete step in disk space for FY01.
 - Start distributing Mini, Reco, Raw levels to IN2P3.
 - Corresponding network improvements required.
- Scale with integrated luminosity beyond FY01.
- No computing load.

Summary of Production Activities

From previous table

	<i>Unit</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Additional Disk Space</i>				
<i>Prompt reconstruction</i>	TB	5	6	4
<i>Re-processing</i>	TB	5	6	4
<i>Simulation production</i>	TB	2	2	2
<i>Data distribution, etc</i>	TB	10	7	4
→ <i>Total</i>	TB	22	21	14
<i>Additional CPU</i>				
<i>Prompt reconstruction</i>		150	180	120
<i>Re-processing</i>		150	180	120
<i>Simulation production</i>		100	120	80
→ <i>Total</i>		400	480	320

Projections for Total Data Storage

- Scales with *integrated* luminosity.
- Tighter event selection to reduce data sample.
 - Implemented as filter in reconstruction stage.
 - Hadronic events at 5 nb, compared with $\sigma_{\text{had}} \sim 4.5$ nb.
 - Fewer non-hadronic events accepted, 3-nb equivalent by FY03.
 - Relatively small impact for simpler events.
 - *Some* non-hadronic events needed.
 - Detector studies, e.g. μ -pair, $2\gamma \rightarrow h^+h^-$.
 - Other physics, e.g. τ .

	<i>Unit</i>	<i>FY00</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Year Increment</i>	TB	200	342	610	772
<i>Total</i>	TB	200	542	1152	1924

Objectivity Micro for Beam Events

- Micro event size reduced from 5 to 3 kB / hadronic event.
 - Expect very little change in content.
 - Mostly through packing and overhead reduction.
- Non-hadronic event size $\sim 1/2$ of hadronic event.
- Need ~ 2 TB per year in FY01 through FY03.

Objectivity Micro for Beam Events

	<i>Unit</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Integrated Luminosity</i>	fb ⁻¹	40	80	115
<i>Hadronic Events</i>				
<i>Cross section</i>	nb	5	5	5
<i># events</i>	10 ⁶	200	400	575
<i>Event size</i>	kB	5	4	3
→ <i>Total space</i>	TB	1.0	1.6	1.7
<i>Non-hadronic</i>				
<i>Cross section</i>	nb	5	4	3
<i># events</i>	10 ⁶	200	320	345
<i>Event size ratio</i>	kB	0.5	0.5	0.5
→ <i>Total space</i>	TB	0.5	0.6	0.5
→ <i>All Beam Events</i>	TB	1.5	2.2	2.2

Decreasing event size

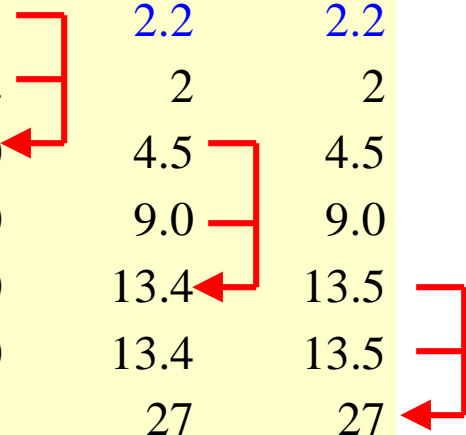
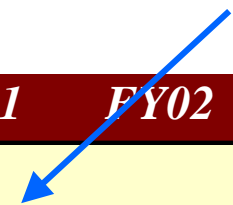
Tighter event selection

Space for All Micro

- Objectivity Micro.
 - Two versions -- reprocessing, tests, etc.
- Monte Carlo sample.
 - 2x space as real events.
- Kanga Micro same size as Objectivity.

From previous table

	<i>Unit</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Beam Events</i>				
<i>Objectivity Micro</i>	TB	1.5	2.2	2.2
<i># versions</i>		2	2	2
<i>Total space</i>	TB	3.0	4.5	4.5
<i>MC Objectivity Mirco</i>	TB	6.0	9.0	9.0
<i>All Objectivity Micro</i>	TB	9.0	13.4	13.5
<i>All Kanga Micro</i>	TB	9.0	13.4	13.5
<i>Total Micro space</i>	TB	17	27	27



Total Storage Space by Year

- Incremental storage.
 - Micro + Mini + Reco + Raw.
- Tape space not reused.

From previous table

	<i>Unit</i>	<i>FY00</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Micro</i>	TB		17	27	27
<i>Mini</i>	TB		24	43	55
<i>Reco</i>	TB		240	432	552
<i>Raw</i>	TB		60	108	138
<i>Total incremental storage</i>	TB		342	610	772
<i>Total storage needed</i>	TB	200	542	1152	1924

Projections for Analysis Activities

- Data in disk cache for analysis access.
- Simple extrapolation based in integrated luminosity.
 - Faster than Moore's Law can reduce cost.
 - Unaffordable.
 - Dominated by disk costs.
- Need to go beyond simple extrapolation.
 - Even less affordable.
- (Disk) cost reduction.
 - Tighter event selection and smaller event size.
 - Staging of Micro.
 - Cheaper disks.

Simple Extrapolation

- 100% of Tag and Micro on disk.
- Mini under development and not used.
- Staging space for Reco and Raw.
- Not sustainable in any plausible budget.
 - Luminosity increase beats Moore's Law.
- Dominated by disk cost.

	<i>Unit</i>	<i>FY00</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Yearly Integrated Luminosity</i>	fb^{-1}	25	40	80	115
<i>Luminosity Ratio</i>		1.0	1.6	3.2	4.6
<i>Moore's Law (18 months)</i>		1.0	1.6	2.5	4.0

Expected Usage Changes

- More detailed detector studies.
- More analysis topics.
- More sophisticated analyses.
- Systematic effects become more important.

Need to go beyond Micro level.

- Implementation and use of Mini.
- More extensive use of Reco, Raw, MC Truth.
- Faster than simple scaling with integrated luminosity.
- More expensive.

Staging of Micro

- Physics driven resource allocation.
 - Top priority physics stream may be 100% disk resident.
 - Low priority stream proportionately less staging space.
 - Allow both controlled and ad-hoc staging.
 - Controlled staging for production-like activity.
- No need to stage Mini, etc, if Micro not available.
- No analysis disk cache for Kanga after FY01.
 - May be produced for distribution to remote sites.

Incremental Analysis Disk Space for Micro

From earlier table

	<i>Unit</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>All Objectivity Micro</i>	TB	9.0	13.4	13.5
<i>Fraction on disk</i>		100%	70%	50%
<i>Disk space</i>	TB	9.0	6.4	6.7
<i>All Kanga Micro</i>	TB	9.0	13.4	13.5
<i>Fraction on disk</i>		50%	0%	0%
<i>Disk space</i>	TB	4.5	0.0	0.0
→ Total Micro disk space	TB	13.5	6.4	6.7

Other Disk Space for Analysis

- Event sizes for other data levels, e.g. Mini and Reco.
 - From Computing Model Working Group (CMWG) reports.
- Disk resident fraction.
 - From CMWG reports.
 - Decreasing with time.
- Other uses of disk space, e.g. ntuples.
 - Additional 50% based on past usage.

Incremental Disk Space by Year

From earlier table

	<i>Unit</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Micro disk space</i>	TB	13.5	6.4	6.7
<i>Mini</i>				
<i>Fraction on disk</i>		30%	20%	10%
<i>Disk space</i>	TB	7.2	8.6	5.5
<i>Reco</i>				
<i>Fraction on disk</i>		5%	3%	2%
<i>Disk space</i>	TB	12	13	11
<i>Raw</i>				
<i>Fraction on disk</i>		5%	3%	2%
<i>Disk space</i>	TB	3.0	3.2	2.8
<i>Sum of Micro through Raw</i>	TB	36	34	26
<i>Miscellaneous, e.g. ntuple</i>		+50%	+50%	+50%
<i>Total disk space</i>	TB	54	51	39

Disk Cost

- Using relatively expensive RAIDs at this time.
 - Hot swap and hot spare.
 - Vendor = Sun.
- Ordered two cheaper RAIDs for testing.
 - IDE, hot swap but no hot spare.
 - SCSI, hot swap and hot spare.
 - Other vendors.
- Unit cost in k\$ / TB.
 - Start from purchase price.
 - Add server and network costs.
 - Account for RAID and file system overheads to get usable space.

Disk Unit Cost Assumptions

- Disk cost = \$50K / TB in 10/00.
 - “Average” between Sun and other vendors.
 - 60% above disk cost to reflect server and network costs.
- May use different disks in different applications, e.g.
 - Production needs more reliable disk as data not yet on tape.
 - Greater risk for analysis disk cache is acceptable.
 - Tag and Micro need high throughput.

Unit Cost of Disk vs Time

- Moore's Law factor of 2 every 18 months.
 - Long term trend.
 - Consistent with recent experience.
 - Sun disks \$100K / TB last year.
 - \$65K / TB now.

From last page

	<i>Unit</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Moore's Law</i>		1.0	1.6	2.5
<i>Disk Space</i>	k\$ / TB	50	32	20

CPU Cost

- Sun Netra T1 more expensive.
 - One CPU in 1-RU package.
 - Remote management.
 - \$3.3K / CPU.
- New Dell Linux machines.
 - Two CPUs in 2-RU package.
 - No remote management.
 - \$4K per node, i.e. \$2K / CPU.

CPU Unit Cost Assumptions

- CPU cost \$2.5K / CPU in 10/00.
 - Target purchase price of \$1.6K.
 - 60% for network and other overheads.
- Cheaper units not manageable in large numbers.
 - Space constraint for low density non-rack units.
 - Remote management.
- Moore's Law factor of 2 every 18 months.

Summary of Costs by Year

	<i>Unit</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Production</i>				
<i>Disk space</i>	M\$	1.1	0.70	0.29
<i>CPU</i>	M\$	1.0	0.76	0.32
<i>Sum</i>	M\$	2.2	1.5	0.61
<i>Tape</i>	M\$	0.51	0.65	0.58
<i>Analysis</i>				
<i>Disk space</i>	M\$	2.7	1.6	0.78
<i>CPU</i>	M\$	1.8	2.3	2.1
<i>Sum</i>	M\$	4.5	3.9	2.8
→ Total	M\$	7.1	6.0	4.0

Summary of Costs by Year

	<i>Unit</i>	<i>FY01</i>	<i>FY02</i>	<i>FY03</i>
<i>Tape</i>	M\$	0.51	0.65	0.58
<i>Disk</i>	M\$	3.8	2.3	1.1
<i>CPU</i>	M\$	2.8	3.0	2.4
<i>Total</i>	M\$	7.1	6.0	4.0

Contingency Plan

- Mismatch between budget and needs.
 - Lower budget.
 - Higher luminosity.
- Production side.
 - Data not reconstructed cannot be analyzed.
 - Keep production side intact.
- Cut back on analysis side.
 - Physics driven prioritization.
 - Reduce space for low priorities streams.
 - Perhaps to 0% staging space for some.
 - Proportionately greater impact.
 - No detailed plan at this time.

FY01 Example

- Plan calls for \$7.1M.
 - Production + Tape -- \$2.7M.
 - Analysis -- \$4.5M.
- Reduce overall budget by factor of 2.
- Impact on analysis is factor of 5.

	<i>Unit</i>	<i>Current Plan</i>	<i>Lowered Budget</i>	<i>Impact</i>
<i>Production + Tape</i>	M\$	2.7	2.7	--
<i>Analysis</i>	M\$	4.5	0.9	5
<i>Total</i>	M\$	7.1	3.6	2

Purchase Schedule

- 4-6 months for hardware to be available to end-user.
- A few rounds of large purchases in a year.
 - Maximize negotiation power.
 - Relatively large quantities of same model for easier maintenance.
- Start purchase *now* for FY01 needs.
 - Next run starts in February.
 - Expect 2x design luminosity by summer 2001.

Overall Schedule

- Decided on technical direction and strategy.
- Working on implementation plan.
 - First round by 10/20.
 - Finalize by 10/31.
 - Detailed schedule and milestones.

- Currently re-skimming 1999+2000 data.
- Planned to reprocess 1999+2000 data starting 12/00 - 1/01.
- Reconstruct and skim 2001 data as it rolls in.
- Serial reuse of same resources.
- Tight and coupled schedule.

Summary

- Reduce resource needs.
 - Tighter selection, smaller event sizes.
 - Greater granularity.
- Cost ~\$7.1M in FY01.
 - ~\$6.0M in FY02.
 - ~\$4.0M in FY03.
- Gradual performance degradation in case of budget shortfall.
 - Physics based priority allocation.
- Budget impact on analysis is magnified.
- Tight schedule allowing for consistent data sample by summer 2001.