



BaBar Beyond-2008

Prepared by BaBar management

G. P. Dubois-Felsmann, A. Jawahery, J. Olsen, S. Prell, W.J. Wisniewski

Version 15

September 7, 2007

BaBar Beyond 2008

This document describes the main elements of a plan that has been developed by the BaBar collaboration for the analysis of its final data set beyond the data-taking phase of the experiment. The plan builds on the report of an internal taskforce (Beyond-2008 Taskforce) that performed an intense study of the physics objectives of the experiment in the era beyond 2008, as well as the required manpower and resources for carrying out the program on a timely basis.

The document is organized in sections 1-6, as follows:

- 1. Introduction and an executive summary of BaBar's physics-analysis-computing plan for the beyond-2008 era;**
- 2. A brief discussion of the "core" physics goals of the experiment;**
- 3. Status of the collaboration, including a time profile of the current and the estimated physicist manpower beyond 2008;**
- 4. Computing for beyond-2008, including a detailed discussion of the analysis that has been performed to formulate the computing strategy for 2008 and beyond. This section is organized as follows:**
 - A brief discussion of the basic elements of the BaBar computing model and the operational realities of the computing system;**
 - The final data sample;**
 - The processing of Run 7 data and the question of reprocessing;**
 - Analysis of computing scenarios for 2008 and beyond;**
 - Proposed computing strategy for 2008 and beyond, and estimates of the required resources and the budget;**
 - Computing manpower and estimates of the required computing FTE's.**
- 5. A brief summary of plans for the state of the BaBar detector beyond the data-taking phase;**
- 6. Concluding remarks.**

1. Introduction

With the approach of the end of the data-taking phase of the experiment, the collaboration has in the past two years been assessing the breadth and the reach of its "core" physics program and the strategy for completing the program on a timely basis.

Some of the major factors in defining this program and its timeliness, are the importance of precision determination of the Standard Model (SM) parameters, the potential of the measurements for revealing effects of New Physics through deviations from the SM expectations, relevance of these precision measurements to understanding the nature of New Physics signals that may emerge from the LHC, the time profile of the interest and the involvement of the BaBar collaborators in the program, and of course the availability of resources to carry out the program.

In the summer of 2005, the collaboration performed a study of the physics reach of BaBar's then-expected "1/ab" data set. This study helped identify and emphasize the key elements of the "core" physics program, provide estimates of the expected accuracy of the measurements, and the potential impact of these measurements in constraining the SM parameters as well as in confronting various models of physics beyond the SM. The results of the study are reported in a BaBar Analysis Document, BAD#1228 (R. Faccini et al.). The list of physics topics in Table 2 reflects the conclusions of this work and the related workshops and studies that followed. In the fall of 2006, the BaBar management constructed a 1st- order plan for the analysis and computing aspect of BaBar in the beyond-2008 era, and formed an internal taskforce charged with evaluating the validity of the plan and its primary assumptions, in view of the available manpower and resources in the collaboration. The taskforce presented its report at the BaBar collaboration meeting in December 2008, which also includes the results of a survey of the time profile of the available manpower in the collaboration, the required manpower and resources to carry out the "core" physics program, some of the potential benefits from a full reprocessing of the final data set and its timing and impact on the required resources. The taskforce report is in BAD#1680 (C. Hearty et al.).

In this document we present an analysis and computing strategy that is built on the model and the recommendations of the taskforce. An executive summary of these recommendations and the emerging plan is as follows:

- Performing the measurements that define the "core" physics program of the experiment (see section 2 and Table 2) requires a period of about 2 to 3 years beyond the end of the data-taking phase of the experiment in September 2008. Both the duration and resource requirements for this program are largely driven by the importance of extracting the results on a time scale commensurate with the time profile of the participation of the postdocs and students currently working on the experiment, as well as the relevance of the results to interpretation of the early LHC results. This would suggest that during this period, hereafter referred to as the "Intense Analysis Period," the mode of operation of the collaboration in analysis and computing activities must remain essentially the same as the current model, relying on the availability of resources across the 5 Tier-A centers, as well as the computing capabilities in the Tier-C centers for simulation production.
 - **An important factor that affects the magnitude and time profile of the required resources is the need for and the timing of a full reprocessing or re-skimming of the data during and after the completion of Run7.**

In spring 2007, the collaboration performed a study to identify potential improvements to the event reconstruction or simulation code and their impact on physics measurements. A decision has been made to deploy these improvements in the release that will be used for processing Run 7. The proposed computing strategy allows for reprocessing or re-skimming of runs 1-6 in 2008, using the same release, aiming for completion by the end of calendar year 2008. However, this scenario does not allow for a full re-skimming of Runs 1-6 in time for summer 2008 conferences, taking into account both the manpower limitations in the BaBar computing system and constraints on the required peak computing hardware resources. The estimates of the computing hardware resources and computing manpower required for this program are presented in Tables 6 and 7, respectively.

- While the studies have shown that the collaboration will have the manpower to perform the core physics measurements, achieving this goal requires a continuation of a well coordinated analysis effort to best utilize the available manpower and resources. This will involve the continuing presence of BaBar collaborators at SLAC during the Intense Analysis Period, occupying a significant part of the space available in the ROB.
- Long term analysis of BaBar data: Guided by the experience of previous experiments at LEP & SLC, we expect that access to the full BaBar data set and a subset of the AWG skims on disk, as well as the ability to generate simulated data, will be required for a period of 3 to 5 years beyond the Intense Analysis Period. The taskforce studies indicate that by the end of this period, the BaBar analysis manpower level would be around 30 FTE. In this era, we expect that a gradual shift will occur from the current distributed computing system to one in which SLAC serves as the central depository of the full data set as well as the AWG skims, and provides the CPU power required for data analysis. We also expect that the BaBar occupancy of the ROB will be reduced by about 1/3 to 1/2 each year, starting in about 2010.
- Very long term access to BaBar data: Interest in the physics potential of BaBar data will likely continue well beyond the analysis phases outlined above. For this era, a model will need to be developed for archiving the data and maintaining its usefulness and accessibility to BaBar members and possibly non-BaBar members.

2. The Physics Goals

At its inception, the primary physics objective of the BaBar experiment was to investigate the breaking of the CP symmetry in weak decays of B mesons and its consistency with the predictions of the CP breaking mechanism in the charged-current sector of the Standard Model, the CKM mechanism. Having established the breaking of the CP

symmetry in B decays in 2001, a major focus of the experiment has been on the precision determinations of the CKM parameters, that is the three angles of the unitarity triangle, α , β and γ , and the magnitudes of the CKM elements, which are in turn extracted from the measurements of the time-dependent and time-integrated CP violating asymmetries and the decay rates for a large set of B decay channels (see Table 2). The final precision of these measurements at the B factory experiments will largely set the ultimate accuracy to which the charged-current sector of the Standard Model is understood, until such time that a much larger data set is obtained at an e^+e^- super-B factory.

Currently, eight years into its data collection phase, with a data set of ~ 400 1/fb and more than 280 papers published (and submitted for publications), the experiment is poised to more than double its data sample with an improved detector by the time it turns off in September 2008. Just as the data sample has grown, so has the breadth of the BaBar physics output, with a number of physics results of far reaching impact both in the primary physics areas of the program, namely CP violation studies, as well as in other areas of physics accessible to the experiment. A brief accounting of some of the major physics achievements of the experiment, thus far, is given below:

- Discovery of the breaking of the CP symmetry in B decays
- Observation of Direct CP violation in the decay $B \rightarrow K^- \pi^+$
- Observation of CP violation in penguin dominated decay mode, $B \rightarrow \eta' K^0$
- Observation of D^0 - $D^0(\text{bar})$ mixing
- Discovery of new Ds states.
- Discovery of new quarkonium states, $Y(4260)$ and the related channels.

A breakdown of the BaBar publications in various areas is presented in Table 1, reflecting both the physics reach of the data and the breadth of the physics interest of the BaBar collaborators. In Figure 1 the time profile of the data collection is shown side-by-side with the BaBar publication rate, revealing an interesting resemblance that is partially a result of the opening of new physics channels with increasing size of the data set. The data set at the end of the data-taking phase of the experiment presents the collaboration with yet another opportunity to further enhance its physics output and impact.

Some of the key measurements that are considered amongst the “core” physics goals of the program are listed in Table 2, followed by brief remarks on each set of measurements in the following subsections.

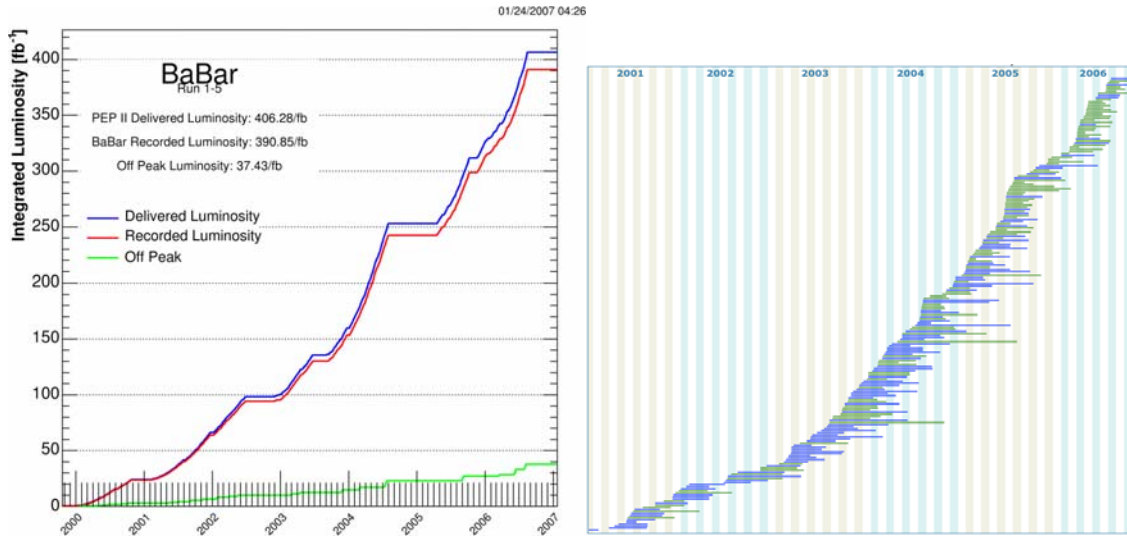


Figure 1: Time profile of BaBar data collection since the start of data taking in 1999 (left), and that of BaBar publications for the same period (right).

Physics Areas		Number of Publications	Number of ongoing analyses	
B Physics	CKM parameters	The angle β	31	14
		The angle α	10	8
		The angle γ	20	6
		V_{ub} & V_{cb}	15	19
	B Meson Decays	Charmless B Decays	53	22
		Hadronics B Decays	44	15
		Radiative B decays	18	11
Leptonic B decays		7	18	
Quarkonium		16	17	
Charm Physics	Charm decays	24	27	
	D0 mixing and CPV	5	5	
Tau Physics	Lepton Flavor Violation	5	15	
	Tau Decays & lifetime	3	5	
ISR Physics		11	19	

Table 1- A breakdown of BaBar publications (submitted or published) and ongoing analysis in various physics areas. Some papers overlap more than one physics topic, while the analyses (last column) are uniquely categorized.

2.1. The CKM parameters

As discussed above, the precision determination of the parameters that define the quark mixing (CKM) matrix has been the centerpiece of the physics program of the B factories and is considered the lasting legacy of the experiment. Having established the breaking of the CP symmetry in B decays, the goal of the program quickly shifted to the test of the consistency of the measurements with the CKM picture, which may reveal deviations from the Standard Model and provide information on the nature of New Physics beyond the SM. The most important aspects of this program that can be reached from the BaBar data are the CKM unitarity angles α , β , & γ and the magnitude of V_{ub} , V_{cb} , V_{td} , & V_{ts} . A list of measurements that connect with the CKM parameters is given in row 1 of Table 2. Together with the data from the Belle experiment, these measurements will set the ultimate accuracy of the knowledge of the CKM matrix.

2.2. The loop-dominated B meson decays

An equally important physics focus of the experiment is precision measurements of the processes that are sensitive to the contributions of new physics through higher-order loop diagrams, such as $b \rightarrow s\gamma$ or its gluonic counterpart. New interactions and new particles that may be found at the LHC can contribute to these processes through additional virtual loops and alter the decay properties, including CP asymmetries, from the SM expectations. The first and most celebrated of these channels is $b \rightarrow s\gamma$, which ever since its discovery by CLEO in 1993 has had a constant presence in the studies of the effects of New Physics and constraints set on parameters of these models. The data from the B factory experiments has enormously expanded the reach of this area of investigation into New Physics through measurements of a large set of decays that originate from radiative- and gluonic-penguin decay processes. In addition, the larger data sets allow for the investigation of these processes through a number of complementary observables, some of which have cleaner theoretical interpretations, including decay rates, time-dependent CP asymmetries and direct CP asymmetries. Table 2, shows a list of the key measurements in this area. Amongst these measurements, time-dependent CP asymmetries in penguin dominated modes have received significant attention because of their relatively clean theoretical interpretation. In the Standard Model (SM), to a good approximation, the time-dependent CP asymmetry in these modes is expected to be $\sin 2\beta$, whereas the current data shows a tantalizing 2.5σ deviation from $\sin 2\beta$. Improving the accuracy of the measurements that contribute to this test is a major goal of the experiment with the “1/ab” data set.

2.3. The leptonic B and D decays

Leptonic B and D decays play an important role in testing the predictions of Lattice QCD and in searching for evidence of New Physics effects. BaBar has measured the leptonic

D_s decay rate, which leads to a determination of the decay constant f_{D_s} . Information on the B decay constant f_B can similarly be extracted from the decay rate for $B \rightarrow \tau \nu$. This process is also sensitive to the contributions of charged Higgs or SUSY particles, and thus can help constrain the range of parameters of the models of physics beyond the SM.

2.4. D^0 Mixing and CP violation in charmed hadron decays, and rare charm decays

The recent observation of D^0 mixing by BaBar, soon followed by Belle, has now opened a whole new area of investigation with the potential to reveal effects of New Physics, owing to the fact that both mixing and CPV in D decays are highly suppressed in the Standard Model. While D^0 mixing is now firmly established, a major effort involving the analysis of a number of decay channels is required to determine its parameters. Rare charmed decays such as $D^0 \rightarrow 1^+ 1^-$ provide additional power for probing New Physics effects.

2.5. Lepton Flavor Violation in Tau decays

The tau pair sample of events recorded by BaBar, which is nearly equal to the B pair sample, provides a unique opportunity for the study of Lepton Flavor Violation (LFV) in tau decays. These SM-suppressed processes are highly constrained in the light lepton system. Constraints on LFV in the tau lepton sector are dominated by the results from the B factories. In fact, in a few channels, the experimental sensitivities are in striking distance of the predictions of certain models of New Physics.

2.6. Initial State Radiation processes

The study of Initial State Radiation (ISR) events in BaBar allows access to electron-positron annihilation at center-of-mass energies well below the 10 GeV region. Measurements of the e^+e^- hadronic cross section at $\sqrt{s}=1-4$ GeV, which is accessible to BaBar, is an important component in the understanding of the vacuum polarization correction in muon $g-2$ measurements. Analyses of ISR events have also resulted in the observation of a number of new states.

Core physics areas	Analysis Channels
CKM: Angle β Measurements of Time-Dependent CP Asymmetries and direct CP asymmetries	$\sin 2\beta$ from $B \rightarrow c\bar{c}K^0$
	$\cos 2\beta$ from $B \rightarrow J/\psi K^{*0}$
	β from $B \rightarrow Dh$
	β from $B \rightarrow D^{(*)+}D^{*-}$
CKM : Angle α Measurements of decay rates, TDCP asymmetries and direct CP asymmetries	$B \rightarrow \pi\pi$ [$\pi^+\pi^-$, $\pi^+\pi^0$, $\pi^0\pi^0$], $K\pi$, KK
	$B \rightarrow 3\pi$ Dalitz analysis
	$B \rightarrow \rho\rho$ [$\rho^+\rho^-$, $\rho^+\rho^0$, $\rho^0\rho^0$]
	$B \rightarrow A_1\pi$
CKM: Angle γ Measurements of Rates, Direct CP asymmetry and Dalitz Analysis	$B \rightarrow D^{(*)+}K^{*-}$ [Dalitz analysis, GLW, ADS]
	$B \rightarrow D^{(*)0}K^{*0}$
	$B \rightarrow D^{(*)}\pi$
	$B \rightarrow D^{(*)}\rho$
CKM: V_{ub}	Inclusive $B \rightarrow X_u l \nu$
	Exclusive $B \rightarrow X_u l \nu$ [$B \rightarrow \pi l \nu$, $B \rightarrow \rho l \nu$, ...]
Loop Dominated Processes As probes of New Physics	Radiative B decays: Inclusive and exclusive $B \rightarrow s\gamma$ [Rate, A_{ch} ,...] <ul style="list-style-type: none"> TDCP in $B \rightarrow K^{*0}\gamma$ [Probe of helicity of γ] Inclusive and exclusive $B \rightarrow d\gamma$ [Rate, A_{ch} ,...] Inclusive and exclusive $B \rightarrow sl^+l^-$ [Rate, A_{ch} , A_{FB} , ..] Search for $B \rightarrow s\nu\bar{\nu}$
	TDCP in Gluonic Penguin Dominated Channels: $B \rightarrow K^0\phi$, $K^0\eta'$, $K^0K^+K^-$, $K^0\pi^0$, $K^0K_s^0K_s^0$, $K^0\rho$, $K^0\omega$, $K^0\pi^0\pi^0$
	Charmless Decay Properties: Decay Rate, Direct CP, Polarization $B \rightarrow VV$ Decays ($\rho\rho$, ϕK^* , ρK^* , $\omega\rho$, ωK^* , ...) $B \rightarrow \eta\pi^0$, $\eta\pi^0$, $\eta\eta$, $\eta\eta'$, $\eta\eta$ (for $SU(3)$ analysis) $B \rightarrow \rho K$, ηK , $K\pi\pi$, $3K$, 3π
Leptonic B and Charm decays: B and D decay Constant (LQCD) Probe of New Physics	$B^+ \rightarrow \tau^+\nu$, $B^+ \rightarrow l\nu(\gamma)$, $B \rightarrow ll$, $D_s^+ \rightarrow l^+\nu$
Charm Physics	D^0 mixing and CPV Rare Charm Decays ($D \rightarrow ll$, FCNC in charm decays)
Tau Physics	Lepton Flavor Violation: $\tau \rightarrow \mu\gamma$, $e\gamma$, $\tau \rightarrow ll$, $l\pi^0$, $l\eta$, $l\eta'$, lK_s^0 ,

Table 2: Some of the key measurements on the “core” physics program of BaBar.

3. Status of the Collaboration

The BaBar collaboration currently consists of 543 physicists from 77 institutions in 10 countries. A breakdown of the collaboration membership in various categories is shown in Table 3, indicating a very strong base of young physicists, 154 graduate students and 89 postdoctoral research associates, committed to the operation of the experiment in this final phase of data collection and the physics analysis of the data. In order to estimate the strength of the collaboration in the years beyond the data-taking phase, the Beyond-2008 taskforce carried out a survey of the principal investigators (PI) with specific questions related to their plans for participation in the analysis activities in the period 2008-2010 (“short”) and the period after 2010 (“long”). The results of the survey, which reflects a return rate of 80% of the institutions, are presented in Table 4, indicating a strong continuing commitment and interest in completing the physics analysis of the BaBar data in the first 2 to 3 years beyond the data collection phase. The taskforce also studied the required “service tasks” to support the on-going activities of the experiment, which in the era beyond the data-taking phase are dominated by computing and physics support functions. Figure 2 shows the expected evolution of the number of FTE’s required for service tasks, indicating a drop from the current level of 120 FTE’s to around 40 by 2010. Overall, these data support the conclusion that the collaboration has the interest and the manpower to carry out its “core” physics program during the Intense Analysis Period, including the required service tasks.

Regions	Faculty	PhD Staff	Postdoctoral Res. Assoc.	Graduate Student	Non-Ph.D	Total
Canada	10	1	4	8		23
France	13	14	3	10	4	44
Germany	7	5	3	16	1	32
Italy	29	30	11	20		90
Netherlands	1			2		3
Norway	2			2		4
Russia	2	7		1	1	11
Spain		2	2	3		7
United Kingdom	21	2	16	19	2	60
United States	79	54	50	73	13	269
Total	164	115	89	154	21	543

Table 3: The distribution of BaBar physicists in the collaborating institutions.

Region	Returns	Institutions			Members “short” (2006)		
		2006	“short”	“long”	Academics	RAs+Staff	Students
Canada	100%	4	4	2-3	9 (10)	2 (5)	12 (8)
France	80%	5	4-5	3-4	13 (13)	6 (17)	5 (11)
Germany	100%	6	5	3	6 (7)	3 (8)	8 (16)
Italy	67%	12	8-11	5-8	18 (27)	25 (42)	9 (19)
UK	80%	10	7	2	14 (23)	8 (19)	8 (21)
US(West)	87%	15	11-13	3-5	16 (38)	11 (25)	18 (28)
US(East)	68%	19	11-15	5-10	23 (37)	17 (27)	21 (35)
Other	75%	4	2-3	1	6 (7)	6 (9)	8 (8)
Total	81%	75	52-63	24-36	105 (162)	78 (152)	89 (146)

Table 4: Summary of personnel survey for BaBar after 2008 and in 2006 (in parentheses), excluding SLAC. The uncertainty on the estimates for “short” period reflects the return rate of 80%.

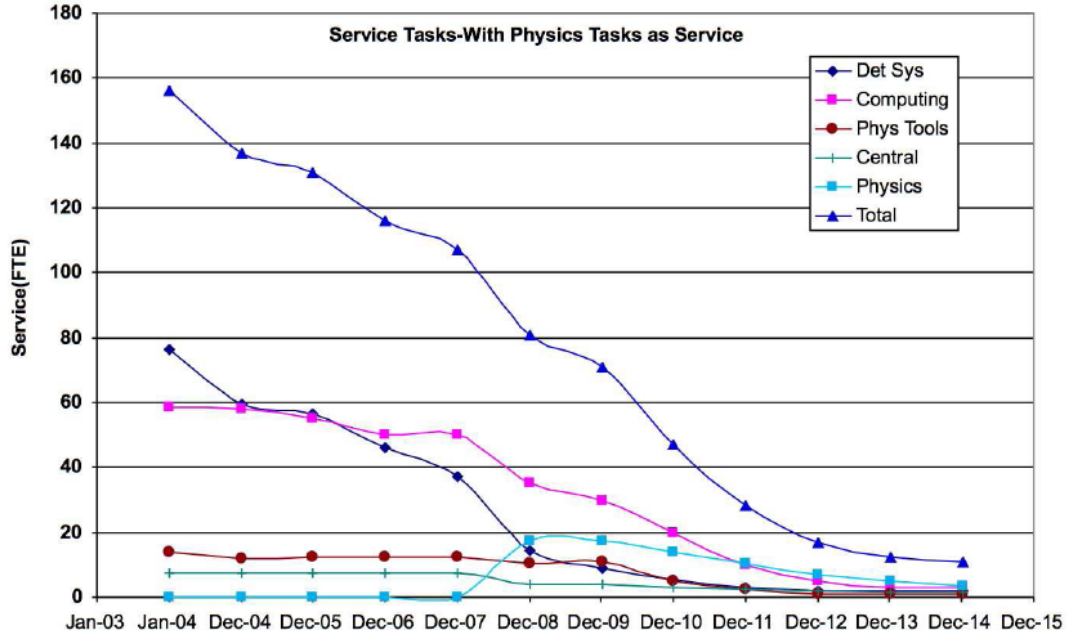


Figure 5: The total service work as a function of time, with the addition of physics service tasks (blue-green line) starting in 2008. Note that the physics tasks exist (such as AWG convenerships and publication board membership) prior to 2008, but are not included in the “service” accounting.

4. Computing for BaBar-Beyond-2008

The BaBar computing model rests on planning for the following key functions:

- Processing (calibration and reconstruction) of the live data;
- Generation of simulated data corresponding to the detector conditions of new live data;
- Skimming of new live and simulated data for physics analysis to match the skims of previously accumulated data;
- Physics analysis of the full dataset at SLAC and Tier-A centers;
- A full re-skim of the accumulated data each year, with 1-3 additional opportunities for partial reskims each year (to allow the introduction of new skims or corrections to existing ones); and
- The capacity for a full reprocessing and resimulation of the data each year.

The BaBar computing system is distributed across the BaBar collaboration, with the bulk of the resources and activities concentrated at the 5 Tier-A centers in France, Germany, Italy, and the UK, and at SLAC. In the BaBar computing model, reconstruction, skimming, and analysis are distributed among SLAC and the non-SLAC Tier-A centers; simulation is carried out primarily (75%) at the Tier-C centers (universities and other laboratories) and 25% at the Tier-A sites and SLAC. In particular, the computing center at Padova carries out most of the computing required for the processing and reprocessing of the BaBar data, and maintains a mirror of the XTC file (raw data) archive at SLAC.

With typically ~150 on-going major analyses in the experiment, analysis-related computing has at this point grown to be the largest component of BaBar computing, responsible for something more than half of the CPU requirements, and driving the requirement for a large pool of disk storage to ensure efficient access to data. Satisfying the needs of the physics analysis effort, which each year peaks around the major winter and summer conferences, has been at the heart of the design of the BaBar computing model.

Central to the chosen solution is the concept of “skimming”, the central production of subsets of the data selected according to loosely defined pre-analysis criteria covering the events of interest to specific analyses. Skimming benefits the overall computing system’s performance by reducing the need for simultaneous access to the same data by multiple analysts, and by allowing analysts to perform and re-tune subsequent tighter selections and further analysis more quickly and with reduced resource requirements. By 2007, a typical full round of skimming produces 180-200 parallel output streams. Because multiple skim event selections do often select some of the same events, and because the skimming process also persists for re-use the results of certain CPU-intensive computations (typically combinatoric analysis of the particles in an event), the aggregate

output of skimming is approximately five times larger than the input and dominates the storage requirements for BaBar's processed data.

The use of skimming facilitates the use of the collaboration's distributed computing resources. Each of BaBar's physics AWG's is assigned to one of the various Tier-A center. The skim streams relevant to the analysis activity of each AWG are then assigned to the AWG's Tier-A center. (A very small number of skims of widespread interest are resident at more than one center.) Collaborators are therefore required to perform their event data analysis computing at the appropriate center.

The full, unskimmed BaBar data sample is also maintained at SLAC and the IN2P3 and Italian (CNAF) Tier-A centers, allowing for the analysis of channels not foreseen in the design of the skims, and supporting "pointer skims", constructed by reference and thus relying on the availability of the original event data.

The Beyond-2008 taskforce endorsed the appropriateness of continuing the current computing model for the Intense Analysis Period of 2009 through 2011. As indicated in the previous sections, the high priority "core" physics topics alone would correspond to ~70 concurrent analyses during this period, a level comparable to that of the current analysis activities in the experiment.

In addition to the constraints set by the physics analysis requirements, the following parameters determine the magnitude and the time profile of the necessary resources for BaBar computing in the beyond-2008 era:

- a) The final integrated luminosity accumulated, and the physical size of the corresponding final data set;
- b) The timing of the final processing and any reprocessing and re-skimming of the data; and
- c) The roles and the number of computing professionals required for supporting the BaBar software and computing activities. This is in addition to the personnel required to operate the Tier-A centers.

4.1. The Final Data Sample

As of this writing, the total integrated luminosity accumulated has been 432 fb^{-1} , with 492 fb^{-1} expected by the end of Run 6 and a further 274 fb^{-1} in Run 7, for a final total for BaBar of 767 fb^{-1} . The analysis herein has been based on these numbers, which are based on John Seeman's presentation to the collaboration on 5 June. As is usual for the spreadsheet, they take into account actual recorded luminosity to date together with estimates of delivered luminosity for the future.

4.2. The Processing of Run 7 and the Question of Reprocessing

An initial processing of both the Run 6 and 7 data will be performed as the data are collected, with minimal latency (except for a “bootstrap” procedure of 6-10 weeks at the start of each run). As in previous running cycles, it is expected that the results of this processing will be usable for high-quality physics analyses, and in fact there is a long BaBar tradition of making data available to analysts within days of its acquisition, in the runup to major conferences.

However, it is prudent to assume that the ultimate measurements in some of the key channels may need to rely on the best software and calibrations possible. Development of improvements to reconstruction and simulation software and to calibrations and detector alignments has been an ongoing process throughout the lifetime of BaBar, and it continues even now. Further improvements could be applied just to future data – that is, to Run 7 alone – or to the entire data set, by means of a reprocessing. Simulation and reconstruction are sufficiently tightly coupled in BaBar that a full reprocessing of the beam data implies a full resimulation as well, and vice versa.

Note that many improvements to the BaBar software (*e.g.*, to PID algorithms or neutral-cluster energy calibrations) can be applied without reprocessing or resimulation. If they affect the event selection or data output of skims, then the skimming of the full data sample may need to be repeated, but it can still rely on the existing reconstructed and simulated data.

A series of three workshops were organized in February, March, and May of 2007, with the goal of identifying areas where BaBar event reconstruction and simulation could potentially be improved, and of assessing the potential impact of these improvements on various physics channels and the timing of the start and end of the reprocessing effort.

Many potential improvements in BaBar reconstruction code have been identified in a broad range of areas. Not surprisingly, the main effort has focused on optimizing track reconstruction in both the SVT and DCH, with the goal of improving track-finding efficiency and minimizing systematic errors associated with SVT alignment. Additional work is ongoing in recovery of energy resolution in cases where photons convert in the DRC bars before entering the EMC, and in a variety of other areas.

In simulation, benefits are anticipated from improvements to the detector physics simulation (*i.e.*, GEANT 4), to the detector model (geometry and material), and to the generator-level particle physics simulation. The workshops identified as key objectives improvements to the generator model in various processes, including semileptonic B decays and Dalitz decays of B and D mesons, and the move to GEANT 4 v8.x. For the simulation of Runs 1-5 BaBar has been using GEANT 4 v6.1, while we recently moved to v7.1 for the simulation of Run 6 data. The goal for Run 7 is to use v8.x, so without a full reprocessing and resimulation we will have simulated events with three different versions of GEANT with very different performance covering Runs 1-5, 6, and 7.

Quantitative evaluation of the impact of these improvements is ongoing, and we have not yet had sufficient experience to draw definitive conclusions on the scale of the improvements we expect in all cases. However, it is clear that a large number of analyses could benefit from a full reprocessing and resimulation. The expected increase in track-finding efficiency will amount to free luminosity for any analysis using low- p_T tracks, while improvements in SVT alignment will benefit precision time-dependent CP-violation analyses, including $\sin 2\beta$ and the dilepton asymmetries.

A broad range of analyses will benefit from the simulation improvements. The important changes in the newer versions of GEANT relative to v6.1 are:

1. kaon interactions with material
2. multiple scattering
3. electromagnetic shower model
4. corrections to pion cross sections

We have fully validated v7.1, which is now the default version used in the simulation of Run 6. We found significant improvements in data/MC comparisons for shower shape in the EMC and the modeling of K_L mesons in the IFR. We expect these changes to impact positively the π^0 systematic uncertainty (currently 3% per π^0), which is the dominant uncertainty in many critical analyses, including the determination of the strange quark mass and the CKM element $|V_{us}|$ in tau decays. In addition to GEANT, we also expect to benefit from improvements in the generator model for semileptonic B decays by feeding the latest measurements of form factors back into the simulation. The use of more realistic Dalitz modeling at the generator level will improve the systematic uncertainties in a large number of very important analyses, including the search for new physics in multi-body $b \rightarrow s$ penguin decays.

Although many measurements can benefit from the reconstruction improvements outlined above, constraints on the computing resources and manpower would not allow and justify an early reprocessing, begun before the start of Run 7. Therefore, we have decided on a plan that would first deploy any improved reconstruction or simulation code in Run 7 processing. The results, together with the current processing of Runs 1-6, would constitute the data sample aimed for the summer 2008 conferences. In order to support that subset of analyses that would be expected to benefit from extending these improvements to the full data set, we propose to plan for the resources for either a full reprocessing or re-skimming of the Run 1-6 data, aiming for completion before January 2009 and usability for the summer 2009 conferences.

4.3. Analysis of Computing Scenarios

We have taken the computing resource estimation model that has been used to compute the BaBar computing budget in recent years, and extended it to allow the investigation of several scenarios and to allow their visualization as a function of time over the next 2-3 years.

The scenarios we consider are based on the following common assumptions:

- A final integrated luminosity of 767 fb^{-1} ;
- A limited set of skims to be run this fall on the Run 1-6 data, completing in time for the 2008 conferences, and allowing for the introduction of new skims or the correction of problems with existing ones, with this skim taking 35% of the CPU resources of a full set and producing a negligible increase in the requirements for disk-residence of data;
- Run 7 data-taking beginning in December 2007, with the usual “bootstrap reprocessing” of the first weeks of the run starting in February 2008;
- Run 7 simulation following one month behind the data;
- A full (~200 streams) skim of the Run 7 data and simulation as it becomes available, with varying assumptions concerning the skimming options for Run 1-6 data in 2008;
- A full skim of the Run 1-7 data starting in the fall of 2009, completing in time for the 2010 conferences.

Note that the assumption that the skim pass aimed at the 2008 conferences will not contain all ~200 skims is a major compromise from the default BaBar computing model, recognizing the excessive workload that would result from preparing the release for a full skim pass in parallel with preparing and validating the Run 7 reconstruction and simulation.

The alternatives considered vary in the assumptions made:

1. A “minimal” model in which, in addition to the above, there is a limited set of new and modified skims of the Run 1-7 data performed in the fall of 2008, in time for the winter 2009 conferences (the first ones at which results from the full BaBar data would be presented). The Run 7 data are included in this in order to allow for new ideas or bug fixes to skimming to be introduced later than February 2008, which is when the main skim pass over Run 7 will begin. This limited skim pass is assumed to require 35% of the CPU of a full pass with all ~200 skims.
2. A “full reskim” model in which the fall 2008 skim cycle is extended to cover all skims, over all of Runs 1-7, and completes in time for the 2009 conferences, in January. This model essentially allows for the possibility of applying an improvement towards all analyses and on the entire dataset that does not, however, depend on a full reprocessing – *e.g.*, a change to EMC cluster calibrations or to PID algorithms. The skim is assumed to require six months to apply.
3. A true reprocessing and resimulation of Runs 1-6, beginning in April 2008 once the Run 7 bootstrap is complete and stable operations have been established, and completing in December 2008. As a normal part of a reprocessing, a full skim will also be performed in parallel with it. The start of this skim will need to be several months earlier than in alternative 2 in order to keep the peak processing load – of skimming in parallel with reprocessing – within reasonable limits.

Time profiles of computing requirements by scenario

Estimates of the monthly requirements for CPU cycles in these three alternatives are presented in Figures 2-4.

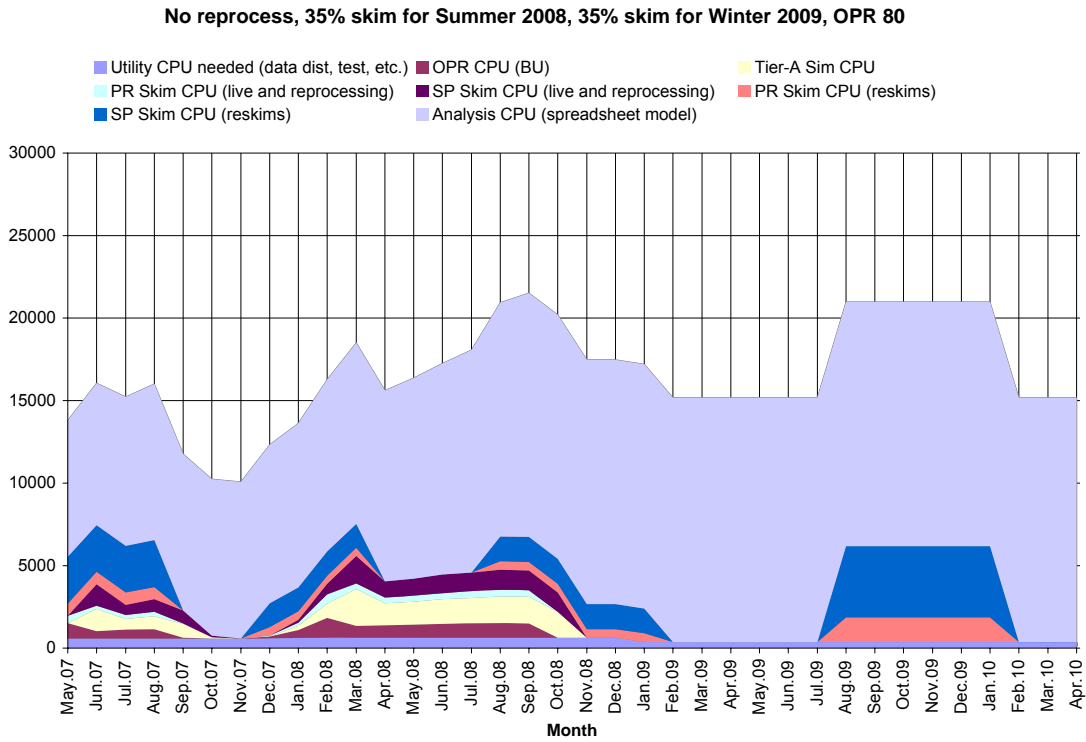


Figure 1 - CPU requirements by month for “minimal” scenario (alternative 1 as described in the text)

No reprocess, 35% skim for Summer 2008, 100% skim for Winter 2009

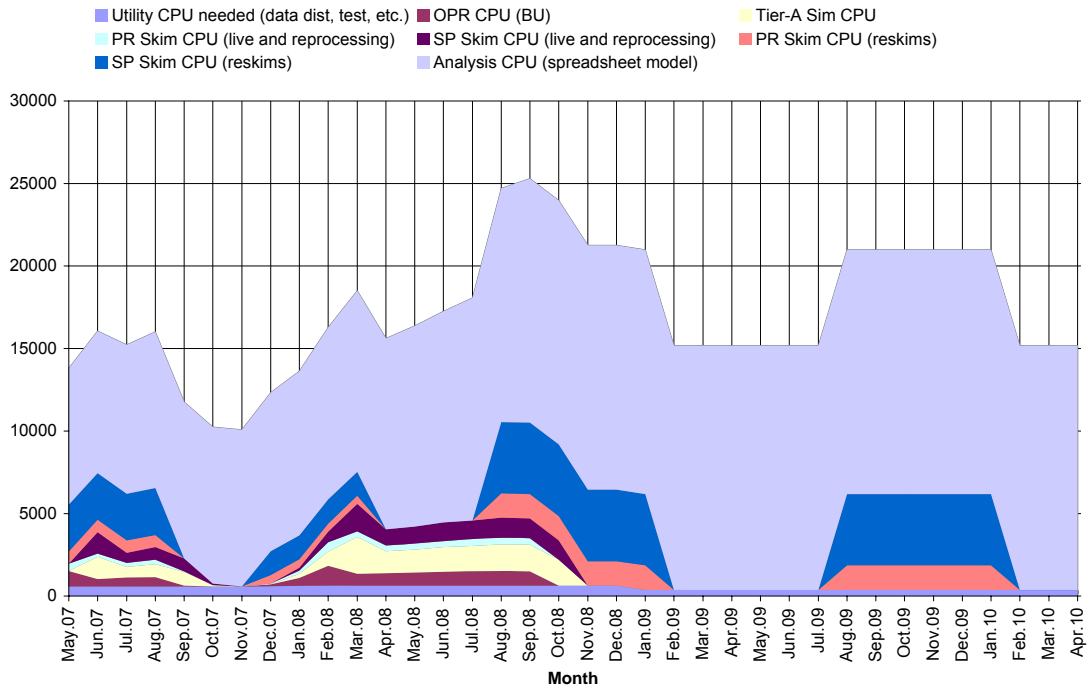


Figure 2 – CPU requirements by month for “reskim” scenario (alternative 2 as described in the text)

Reprocess, 35% skim for Summer 2008, OPR 115, SP 1800

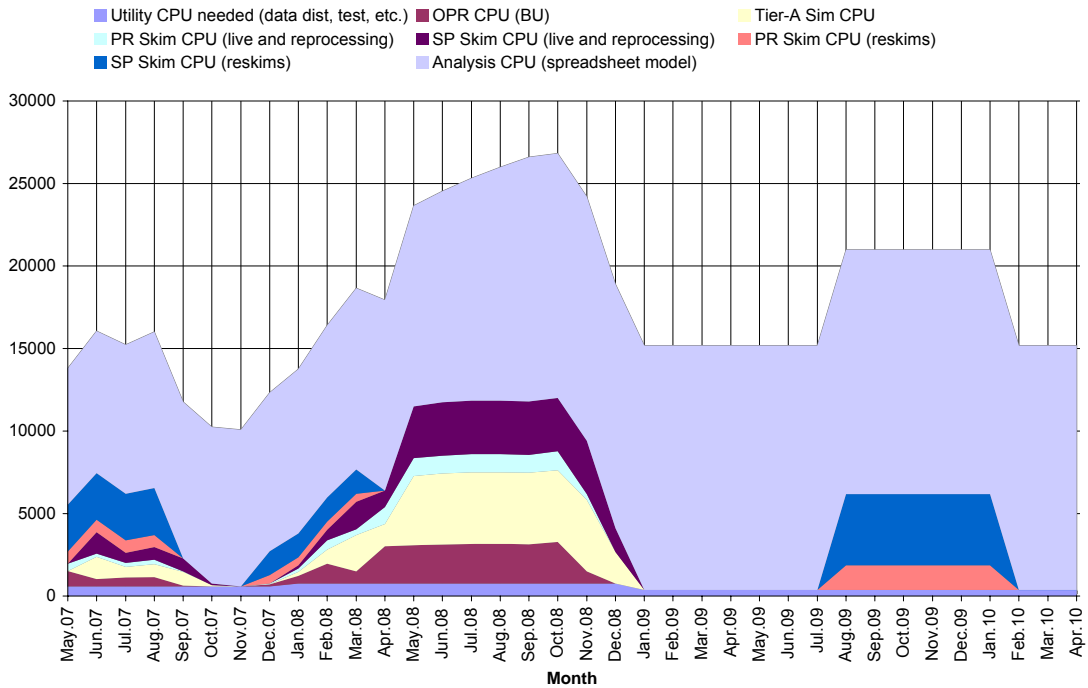


Figure 3 - CPU requirements by month for “reprocessing” scenario (alternative 3 as described in the text)

The CPU requirements were computed according to the projected integrated luminosity month-by-month through the end of Run 7, using the CPU-per-unit-luminosity values for processing and analysis from the latest version of the CSC budgeting spreadsheet (botup_may07_v10*). Note that in this model the “analysis” CPU requirement is an envelope that is intended to enclose the peak analysis load during the runups to conferences; the load does not always reach this boundary at other times of the year. The analysis envelope scales with integrated luminosity, which is why it can be seen to grow during Run 7. It is expected that the analysis load will peak in 2008 not only during conference run-ups, but also in the fall, when it will be possible to release the first publications based on the full BaBar data sample.

Estimation of OPR and simulation capacity requirements

Meeting the month-to-month requirements of the processing schedules described below will require increases in two specific processing capacities within BaBar (as part of the overall increases shown in the diagrams). First, in the scenarios (#1 and #2) without full reprocessing, the total OPR capacity will need to be increased from the present $\sim 50 \text{ fb}^{-1}/\text{month}$ to $\sim 80 \text{ fb}^{-1}/\text{month}$ (in order to process the Run 7 bootstrap in February 2008 in a timely way), while in the reprocessing scenario (#3) it will need to be increased to at

least $\sim 115 \text{ fb}^{-1}/\text{month}$, in order to complete the reprocessing in November 2008 and so allow the simulation and skimming to complete in December.

It is not clear whether it would be possible to extend the Padova ER production capacity to this level. If it were to prove impossible, then the additional capacity would have to be provided at SLAC; it does not appear reasonable to make the operational investments required to host XTC files and ER at a third site. In addition, current bottlenecks in the per-stream PC capacity at SLAC will have to be relieved in order to allow PC to maintain operational headroom at the highest expected PEP-II luminosities.

Second, in the reprocessing scenario (#3), the total BaBar simulation capacity will need to be increased to at least ~ 1.8 billion events/month, and this will need to be achieved without performing more than 25% of this at Tier-A centers (as the standard BaBar computing model requires). The previous peak rate of ~ 1.0 Gevents/month was achieved only by performing over 45% of the processing at Tier-A centers. In other words, the Tier-C simulation capacity will need to be increased to at least ~ 1.35 Gevents/month on its own, a significant increase.

Annual resource requirement estimates

The same assumptions may be used in the context of the CSC resource estimation spreadsheet to compute CPU and disk requirements in 2008, 2009, and beyond, and to perform budget estimates based on available unit cost models. In Table 5 we show the estimates for the resources required in 2008 to implement the three alternative scenarios set forth herein, summed across all Tier-A sites.

Table 5 - 2008 resource requirements for the three study scenarios

	Jan. 07 IFC	Minimal	Reskim	Reprocessing
Production CPU (BaBar-units)	13150	6994	10844	13676
Analysis CPU (BU)	17624	14812	14812	14812
Disk (TB)	2546	2475	2514	2568
Budget (K\$)	3204	2114	2292	2504

The three new budget estimates are based on an updated unit cost model reflecting recent trends in CPU and disk costs. In addition to providing for building up the Tier-A resources to the specified levels, the analysis also takes into account necessary replacement costs for CPU and disks.

The CPU estimates are somewhat higher than the peak values in the monthly profiles shown above, in part because of the limitations of the annual analysis used and in part because they include some provision for contingencies. The monthly analysis, of course, is somewhat fragile in its dependence on assumptions for specific dates for the start, stop, and overlap of the various types of processing.

The CPU requirements for analysis are somewhat lower than those in the January 2007 draft budget for 2008, largely because of the assumption of a lower final integrated luminosity, 767 fb^{-1} instead of 913 fb^{-1} . The requirements for disk are somewhat higher because the model used in January contained an outdated value for the effective total cross-section used for the generation of simulated data.

Note that the luminosity estimate will be revisited by the PEP-II team one final time before the July 2007 IFC meeting. The CPU and disk requirement models will be compared to the actuals for R22 on the same time scale (they are currently based on the R18 production cycle). As noted above, the unit cost model, including “Moore’s Law” effects, for CPU and disk is presently being re-evaluated for consistency with recent purchasing experience within BaBar, and an update to this will be included in the final version of this document and presented to the IFC.

The key observation is that the requirements for the three model scenarios differ only modestly. The disk requirements differ only minimally, driven largely by estimates of buffer space required for the different assumed production activities; they are in practice indistinguishable within the inherent systematic uncertainties in the model. The total CPU requirements show the larger effect: increasing by 18% in the reskim scenario vs. the minimal scenario, and by a further 13% in the reprocessing scenario. Including the combined CPU and disk requirements, and computing the cost for acquiring the new resources required, the total effect on the 2008 budget of choosing the reskim scenario vs. the minimal scenario is approximately 8%, increasing by a further 9% in the reprocessing scenario.

There is some upside risk in all three cost estimates. The CPU unit cost model is based on very recent experience at Brookhaven National Laboratory (BNL) with dual-quad-core Intel Xeon 5335 processors. They performed benchmarks using real physics code (mostly reconstruction) that produced performance just ~15% below the nominal SPECint 2000 ratings for the processors, and no significant penalty for running eight jobs on a host. This is a dramatic change from the behavior of the previous generation of Intel CPUs. They also got an unusually good per-box price. We are relying on this analysis in projecting costs for 2008 and beyond. In order to reflect the uncertainties involved in using this unique data point, we have taken their conservative side in predicting the price/performance for 2008.

SCCS is planning to begin immediately to do a SLAC- and BaBar-specific evaluation of a similar dual-quad-core Intel server, with the intent of having results available to validate the CPU unit costs in time for the IFC meeting in July.

4.4. Proposed Computing Strategy for 2008 and Beyond

Choosing the “minimal” scenario effectively precludes the application of any across-the-board improvements to the BaBar data processing or analysis before late 2009. Given the modest increase in cost in 2008 required to support the reprocessing or reskim scenarios,

we judge that abandoning the ability to apply any major improvements that might be developed presents too great a risk to the quality of the final physics output from BaBar.

We therefore intend to request that the IFC provide the necessary support to assemble the resources required for either reprocessing or reskimming in 2008, and we propose the following strategy:

1. Encourage the BaBar reconstruction and simulation software teams to produce the best processing and calibration they can for the Run 7 data (making sure that their work is always maintained to be readily applicable to earlier data as well, something which is at a minimum essential for its full validation).
2. Pursue an aggressive development and validation strategy for this processing, under the leadership of a BaBar physicist with strong software ability, aiming at code integration this September, followed by an intensive validation exercise based on processing of substantial samples of existing data and involving physics AWGs. This exercise will yield concrete evidence of the physics benefits of the final set of improvements achieved.
3. Begin the processing of Run 7 with this release.
4. Commit to reskimming or reprocessing based on the results of the validation exercise and other conditions at that time.

Computing resources beyond 2008

The annual resource analysis can be extended beyond 2008, under the assumption that a uniform level of effort will apply throughout the Intense Analysis Period, nominally 2009-2011. The results of this analysis are shown in Table 6.

Table 6 - Computing resource estimates in reprocessing scenario beyond 2008

	2008 (reproc)	2009	2010+
Production CPU (BU)	13676	6649	6649
Analysis CPU (BU)	14812	14812	14812
Disk (TB)	2568	2459	2136
Budget (K\$)	2504	[663]	[253]

The model assumes a full reskim in 2009 and each subsequent year, reflecting the assumption of a continuation of intense analysis efforts that would trigger new and changed skims and possible continued improvements to PID algorithms and other software applicable without a full reprocessing. The drop in CPU from 2008 to 2009, of course, reflects the end of live processing and reprocessing. The drop in disk from 2008 to 2009 reflects reduced need for production buffer space, while the drop from 2009 to 2010 reflects the end of substantial support for access to older processing versions (for analyses completing their publication cycle that were based on data processed a year or more earlier). In 2009 it is assumed that some analyses would be completing publication based on the pre-reprocessing data while others would be starting from the new, final

data sample. In 2010 and beyond it is assumed that virtually all analyses would be based on the final reprocessed data sample.

The costs in 2009 and beyond, shown in brackets, are totally dominated by estimated replacement costs. These are currently computed in the CSC spreadsheet from an attempt to keep track of actual CPU and disk acquisitions by year and to account for replacements accordingly, typically four years later. This produces a highly irregular profile of estimated replacement costs, which is not particularly helpful for long-term budgeting. The assumption of a rigid replacement cycle is also at variance with actual experience, in which some equipment is replaced before its nominal life span, and some after, based on details such as actual in-service reliability and power consumption.

Further, the present budget model fails altogether to account for replacement costs in the common fund rebate mechanism for non-SLAC Tier-A sites in the 2009+ era, when no net increases in resources are required. It thus fails to provide an incentive mechanism to member states to perform this maintenance.

We will therefore be recommending to the IFC that for 2009 and beyond the collaboration and the IFC adopt a new model for replacement cost accounting, in which the replacement cycle is flattened out, assuming that a constant fraction (*e.g.*, one fourth) of all hardware still required in each year will be replaced. A budget for this will be computed, based on an updated unit-cost model. It will be proposed, then, that member states hosting a Tier-A site receive a rebate proportional to their usual share of this budget, based on evidence that they are successfully maintaining the reliability and availability of the computing resources at their site. They will be free to accomplish this goal however they see fit: *e.g.*, by replacement, by aggressive maintenance, or by the initial purchase of unusually reliable equipment. If a site fails in whole or in part to maintain their assigned level of computing in a usable state, it will be assumed that those resources will need to be re-established at another site, and the common fund will be used to achieve this.

4.5. Computing Professionals

BaBar computing depends on enormous contributions of both computing professionals and BaBar physicists located at SLAC, Tier-A centers, and Tier-C and other BaBar collaborating institutions. Already during the development of the experiment it was recognized that some of the necessary computing tasks required such highly specialized skills, or such full-time investments of effort, that they could not reasonably be filled from within the pool of collaboration physicists. A mechanism was established to allow 24 FTEs of computing professional effort to be provided by the collaboration's funding agencies: 8 from DOE-supported groups (principally SLAC), 8 from non-DOE, non-US groups, and 8 (later reduced to 6.5) supported by the BaBar Common Fund (employed at SLAC and other US institutions). Although this mechanism was quite successful in providing an important part of the necessary support for BaBar computing, additional computing professionals were still required. SLAC provided the majority of the

additional support required for core computing, while each of the Tier-A sites provided personnel to support users and operation at their site.

Very substantial further contributions to BaBar computing then came from within the collaboration, through individuals who rose to provide leadership in areas of technical expertise, and through the collaboration service mechanism for routine production tasks. The collaboration was fortunate in having a substantial pool of physicists with interest and skills in computing.

Naturally, the list of computing tasks and the roles of individuals have evolved in the past several years, and the original computing professional support mechanism was showing signs of strain, with some tasks going unfilled for a year or more and others not matching the evolving needs of the collaboration. It has also become increasingly difficult to fill the collaboration service computing tasks that require specialized skills, as those physicists with suitable skills have been attracted to newer projects where they can have a more visible impact on the design of a new system.

A task force (Gregory Dubois-Felsmann, Richard Mount, and Mauro Morandin) was commissioned to perform a comprehensive study of the computing tasks required to support BaBar currently and during the Beyond-2008 period. They enumerated the necessary functions and the required FTEs for each role and presented their recommendations to the BaBar management.

The task force examined all of the central computing tasks, including both those performed by physicists and by computing professionals. They found that the tasks now fall more naturally into three categories: those which can be filled by collaboration service effort from the general pool of BaBar physicists, those which require a computing professional, and those which are best filled by a physicist but require one with unusual computing skills, or require a time commitment incompatible with the normal progress of a physics research career – and are therefore increasingly difficult to fill. These categories were denoted P (physicist), CP (computing professional), and SP (special physicist), respectively. It was recognized that filling SP positions may require offering incentives beyond collaboration service, such as the funding of laboratory staff positions.

A summary of the findings of the task force is presented in Table 7.

Table 7 - Summary of requirements for computing experts and professionals

	FTE's in 2007	FTE's needed during Run 7 and through end of (re)processing of data	FTE's needed during intense- analysis period	FTE's needed >2010
Computing professional	17.85	16.35	8.10	4.50
Special Physicist	8.55	7.55	5.80	2.05
Physicist (Collab. Service)	10.20	10.20	6.20	1.70

Note that the amount of “physicist collaboration service” effort is only a subset of the total effort which goes into collaboration service support of BaBar computing, as it does not include site-specific positions such as the Tier-C simulation production operators. The task force restricted its tabulation to the common core of computing support.

The total amount of “computing professional” plus “special physicist” effort required, 23.9 FTE in 2008, is slightly greater than the 22.5 FTE recently intended to be provided by the old computing professionals mechanism. However, it includes a number of tasks (*e.g.*, Computing Coordinator, online system coordinator) that had never been included in the old mechanism, amounting to ~10 FTE, and excludes some no-longer-needed and site-specific tasks that had entered the old mechanism.

The task force and BaBar management are presently discussing the best way to ensure that all these tasks are filled and that appropriate incentives for this are provided. It is anticipated that a mix of common-fund and in-kind support will still be needed, but it is recognized that there must be a system in place to deal with in-kind commitments that are not fulfilled. A core requirement of the new system will be that BaBar management explicitly and annually approve all in-kind contributions of CP and SP effort to BaBar computing, certifying that the person suggested has the necessary skills for the task and that he or she will be able to perform the task at their location (in many cases it will be necessary for the person to come to SLAC for the duration of their work).

The task force is also cataloging the site-specific computing support requirements across the Tier-A complex and the collaboration. While this sort of computing support is not eligible to be covered by the central mechanism, it is important that it be recognized and that collaboration service credit for it be provided as appropriate.

The full report from this task force will be attached as an appendix to the final version of this document.

5. Plans for the BaBar detector beyond the data-taking phase

After data taking is complete in 2008, the BaBar Detector will be moth-balled. The detector assets will be preserved for future reuse in other applications. During the first half-year after completion of data taking, they will be put into a minimal maintenance state. The cooling systems for DCH, DIRC, IFR, and EMC (water cooling) will be drained and dried out. The EMC Fluorinert cooling system, which maintains the photodiode-crystal interface for the 6580 crystals at a constant temperature, will remain in operation. The DCH and IFR gas mixes will be replaced with inert gases, reducing the maintenance burden and cost. The DIRC stand-off-box will be drained of water and dried. The magnet coil will be brought to room temperature and the cryogenic plant will be mothballed. However, vacuum will be maintained in the magnet cryostat by pumps. A robust monitoring system of minimal size that does not require frequent updating will be developed to assure asset preservation. Minimal technician and experienced-physicist effort will be needed to maintain the detector from the end of this transition phase until final disassembly, estimated to be in FY2015.

6. Conclusion

The BaBar collaboration has drawn up a plan for the analysis of its final data sample, following extensive studies within the collaboration of its physics priorities, in view of the timeliness of the results and the availability of people and resources to carry out the program. The collaboration aims at completing the analysis of its “core” physics channels, a list of more than 70 topics, in an Intense Analysis Period of 2-3 years following the end of the data taking phase of the experiment. A computing strategy has been designed to allow for presentation of preliminary results on some of the key physics channels based on the close-to-final data sample at summer conferences of 2008. This model, which contains resource planning for a full reprocessing or re-skimming of the full data set in time for use in summer 2009 conferences, provides the facility for completion of the “core” physics program of the experiment during the Intense-Analysis period.