

Report of the Beyond 2008 Task Force

C. Hearty (chair), D. Brown, R. Faccini, K. Honscheid, H. Lacker,
S. Playfer, B. Ratcliff, J. Smith, G. Vasseur

Ex-officio:

G. Bonneaud, D. MacFarlane, R. Mount, S. Prell

February 9, 2007

Abstract

The Beyond 2008 task force was formed in October 2006 by Hassan Jawahery with the task of providing management with information needed to develop a detailed model for the shape of the experiment after the end of data taking. This report summarizes the information.

1 Introduction

The Beyond 2008 task force was set up by *BABAR* management in October 2006 with a charge to advise management on the development of a detailed model for the shape of the experiment beyond the data taking phase. (The full charge is included in appendix A). The committee was requested to provide this advice on the time scale of the December 2006 collaboration meeting.

To organize the work, the task force was divided into smaller groups that focused on four areas: physics, personnel, service work, and computing hardware. There are connections between these areas that were discussed at weekly meetings and in work between meetings.

These four areas are discussed below, and are followed by a short description of the issues related to *BABAR* data analysis in the long term (> 10 years). We conclude with a summary of the key points.

2 Physics program beyond 2008

The high priority analyses that *BABAR* needs to publish on the full data set are listed in Table 1. The goal should be to publish these by the end of 2010, although a tail extending into 2011 may be expected. Results for these decays

have been published with current data but the much larger dataset expected by the end of data taking in fall 2008 will afford substantial improvements in precision since most of these measurements are still statistics limited.

Most of these measurements probe fundamental properties of the Standard Model and would be sensitive to deviations that could indicate physics beyond the Standard Model. Given the interest in these decays and the fact that *BABAR* is not the only experiment in a position to make these measurements, we feel an obligation to ourselves and the community to publish results for these analyses in a timely way. Given that a number of *BABAR* collaborators will be reducing their commitments in favor of LHC or other experiments, we anticipate that we would lose the ability to do some of the most important measurements if they are delayed substantially past the end of data taking.

Table 1: Highest-priority analyses which should be published within 1 – 2 years of the end of *BABAR* running in 2008. In many cases there are several charged states for topics listed below.

Charmless:
$\sin 2\beta$ in $b \rightarrow s$ penguin: $K^0 + (\eta', K^+K^-, K_s^0 K_s^0, \pi^0, f_0, \rho, \omega, \pi^0\pi^0)$
BFs of color-suppressed decays for SU(3) limits: $B \rightarrow \eta'\pi^0, \eta\pi^0, \eta'\eta, \eta'\eta', \eta\eta)$
2-body (TD, BFs, \mathcal{A}_{ch}): $\pi\pi$ (also TD for α), $K\pi, KK$
Dalitz plots: $3\pi, K\pi\pi, 3K$
VV decays (BFs, \mathcal{A}_{ch} , polarization): $\rho\rho$ (also TD for α), $\phi K^*, \rho K^*, \omega\rho, \omega K^*$
Direct CP: $\rho^0 K^+, \eta K^+, \dots$
$B \rightarrow$ Charmonium + X :
$\sin 2\beta$ with $B \rightarrow c\bar{c}K^0$, $\cos 2\beta$ with $B \rightarrow J/\psi K^*$
Measurements of γ :
$B \rightarrow DK$ (Dalitz, GLW, ADS), $D^{(*)0}K^{(*)0}, D^{(*)}\pi, D^{(*)}\rho$
B semileptonic:
V_{ub} with both inclusive and exclusive $b \rightarrow X_u \ell \nu$
Radiative B decays:
inclusive and exclusive $b \rightarrow s\gamma, b \rightarrow d\gamma$ (BFs, \mathcal{A}_{ch}), $B \rightarrow K^*\gamma$ TD
Rare B decays:
inclusive and exclusive $b \rightarrow s\ell\ell$ (BFs, A_{FB}), $B \rightarrow K^{(*)}\nu\bar{\nu}$
B leptonic decays:
$B \rightarrow \ell\nu(\gamma), B \rightarrow \ell\ell$
LFV in τ decays:
$\tau \rightarrow \ell + (\ell\ell, \gamma, \pi^0, \eta, \eta', K_s^0, \dots)$
Rare charm decays:
$D \rightarrow \ell\ell$, FCNC in D decays
D mixing and CPV:
Mixing ($D^0 \rightarrow K\pi, K\pi\pi, K3\pi$); CPV ($KK, \phi K_s^0, K_s^0\pi^0$)

The physics program of *BABAR* is much larger than this core set of analyses.

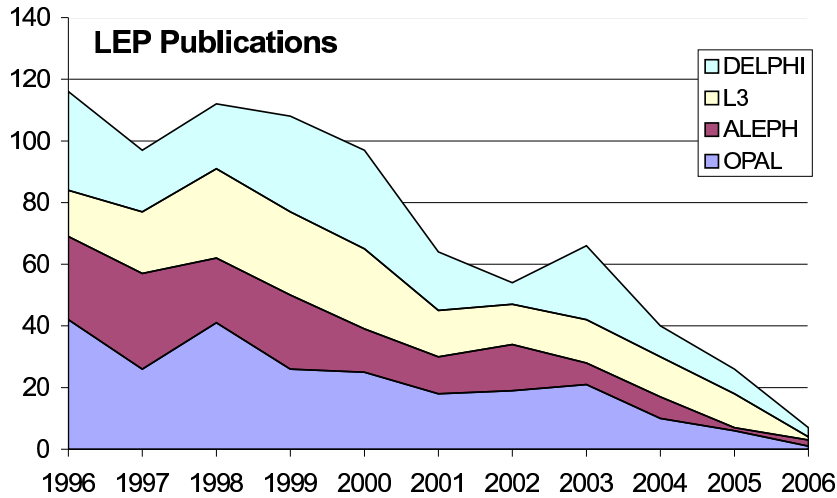


Figure 1: LEP publications as a function of time. LEP2 stopped running in 2001.

For instance, this can be seen in the results presented at ICHEP 2006 when we had about 70 results on topics other than those in Table 1. The additional areas where we expect a significant number of publications are: non- CP B physics (rare decays and tests of strong interactions), charm and charmonium spectroscopy, low-energy physics with initial state radiation, $\gamma\gamma$ physics, and τ physics other than that given above.

In order to better understand the profile of physics result after 2008, we extracted from the SPIRES database information on publications from previous experiments (Fig. 1). In particular, we looked into the number of publications per year for Aleph, L3, and CLEO. While there were only slight drops in the number of publications at the end of LEP1 and CLEOII data taking, there were substantial decreases 2–3 years after the LEP2 and CLEOII.V periods. We conclude that the collaborations stayed mostly intact until the end of LEP2/CLEOII.V and in this respect the *BABAR* situation after 2008 may be more like that of LEP2/CLEOII.V, which further emphasizes the need for doing the analyses in Table 1 within ~ 2 years after the end of data taking. However, we also observe that there is a longer tail of papers published in a period from 3–6 years after the end of the data-taking period. Those are mainly new ideas or complicated analyses that are limited by systematic errors. From the personnel outlook in Sec. 3, we expect to have 30–100 FTE working on *BABAR* in 2010 and beyond.

3 Personnel

We made a survey in which we asked the PIs of *BABAR* institutions:

- if they planned to continue working on *BABAR* after the end of data-taking?
- to estimate the active students, postdocs and staff involved in
 - short term analysis activities in 2008-2010 after the end of data-taking (hereafter referred to as “short”)
 - long term analysis activities after 2010 (hereafter referred to as “long”)
- to indicate how much these people would be present at SLAC
- to comment on any factors that would influence their answers

The numerical results of the survey are summarised in Table 2. From estimates of the FTE fractions given in the returns to our survey, we calculate that an academic is worth 0.5 FTE, an RA/staff member is worth 0.7 FTE and a student is a full 1.0 FTE. Applying these scale factors to the total membership numbers *excluding SLAC*, we obtain approximately 200 FTEs in the “short” analysis period, down from 330 FTEs in 2006. The reduction in manpower by 40% between now and the “short” analysis period is fairly uniform over the different regions, and over the different categories of staff.

Table 2: Summary of Personnel survey for *BABAR* after 2008, *excluding SLAC*. The uncertainty on the “short” institutions reflects the return rate of 80%. The uncertainty on the “long” institutions also includes the groups who responded but are undecided. The number of members in the “short” period is our best estimate extrapolating from the returns we received.

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)

For the “long” period we have only tabulated the number of institutions that are likely to remain active on *BABAR* data analysis. From the returns at least a third of all institutions do expect to be active, but there are many undecided institutions, and it may be as high as a half. Perhaps surprisingly, all regions are still represented at approximately their previous fractions. We make a rough

estimate of a total of about 60 FTEs after 2010, but it could be as low as 30 FTEs or as high as 100FTEs.

From the survey we expect that 35-50 collaborators from outside will be based at SLAC in the “short” data analysis period. These are mostly students plus a few RAs, and they come primarily from the US, Italy and Canada. They will need appropriate office space in a well-located area in order to guarantee efficient communication. About 150 collaborators will make short visits, particularly during collaboration meetings, for which they will need shared office space.

The evolution of the SLAC participation on *BABAR* has been considered separately. The combined PhD physicist and graduate student numbers drop from 43 at the end of the data-taking to about 20 at the end of the “short” data analysis period. The numbers follow an exponential curve with a decay constant of 3 years. They do not include people who are not *BABAR* members, but support operations or computing activities. The SLAC participation seems to scale down in a similar way to the rest of the collaboration.

As part of our survey we asked the groups which factors are most important in determining their continued participation in *BABAR*. While there were some regional differences a clear pattern became apparent. Pressure from funding agencies to move resources to the LHC was the most important factor, followed by potential difficulties in recruiting post docs and graduate students. The ILC and a Super B Factory were mentioned by many groups as the source of uncertainty about their long term plans. It seems to us that in the short to medium term these future projects actually enhance the participation in *BABAR* data analysis, but eventual operation of a Super B Factory would take away most of the long term effort. Commitments from SLAC to support *BABAR* data analysis and to provide facilities for R&D on future projects are needed to retain a particle physics user community at SLAC.

4 Service work requirements

The many tasks needed to produce a data set ready for physics studies and the corresponding simulated events are referred to as “service work”. These tasks include operating, aligning, and calibrating the detector, reconstructing the data, producing massive MC samples, maintaining and validating physics tools, and managing the experiment, its data, and its computing resources. Service work is an individual and institutional obligation, and has typically required about 30% of the collaboration’s manpower.

System managers, tools group leaders, and others within collaboration management were polled to provide estimates of evolution of the service work load. The result is shown in Fig. 2.

The work on the detector systems will drop dramatically when data taking is completed. However, the systems are responsible for alignment, calibration and validation, so will remain active during reprocessing. Afterwards, this effort will be reduced to some modest code review and constants maintenance.

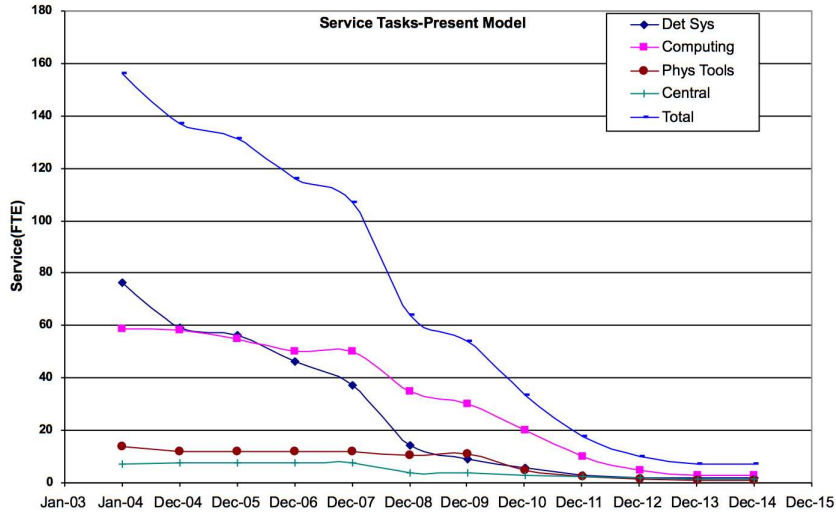


Figure 2: The total service work estimate is given in blue, while component estimates are given for (1) detector systems (black), (2) computing (purple), (3) physics tools (red), and (4) central management (green). Note that the computing tasks include those done by both physicists and computing professionals.

The required computing effort level will remain rather large until the reprocessed data and Monte Carlo samples are in hand. Some effort to maintain access to the large data sets is expected to be required as long as there is a significant amount of analysis underway. In addition, MC production—in particular, signal MC—will continue at a lower level. Some of the final analyses may be quite sophisticated and require large MC samples.

The physics tools groups handle algorithms, constants, and validation for basic analysis tools, such as tracking, PID, and Neutrals. The effort required by these groups will stay nearly constant until reprocessing is well underway. It will then decline, but some support will need to be retained.

Central tasks related to detector operations will be completed at the end of data taking, but a number of computing tasks and overall collaboration management tasks will continue with little diminution until reprocessing is complete. Substantial management is still expected to be important during the analysis phase.

Figure 2 indicates that the service task load after the end of Run 7 will be approximately half of its current level. The drop in the number of FTE physicists over that time is similar, approximately 40%. Therefore, the service load per person will not drop substantially until after the end of reprocessing. Note, however, that computing jobs, which may be harder to fill, become a larger fraction of the total load.

It is worth considering the staffing experiences of other major experiments after the end of their data taking periods. Generally, they have encountered

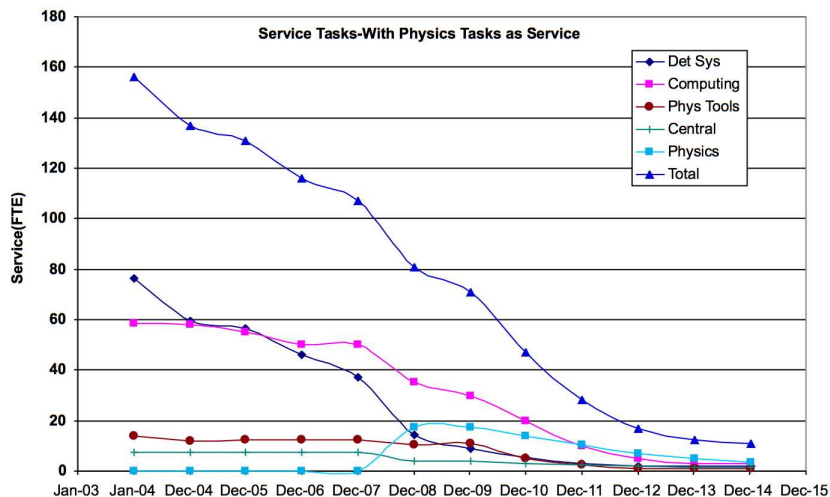


Figure 3: The total service work as a function of time, with the addition of physics service tasks (blue-green line) starting in 2008. Note that these tasks exist prior to 2008, but are not included in the accounting.

difficulties in staffing essential tasks, finding not only that the number of FTEs becomes small after data taking ceases, but that each FTE becomes less efficient, since it is diluted by many people giving small fractions of their time. It may be important to act proactively, encouraging individuals and institutions to continue contributing to the essential needs of the experiment by expanding the range of tasks receiving service credit. These could include analysis group leadership, activities related to editing and reviewing publications, and activities related to conference talks. Figure 3 adds a set of these tasks to the list of service tasks starting in 2008. It should be emphasized that the peak in “Physics Service Tasks” in 2009 is just a result of a change in accounting. However, putting new set of tasks in at this time does reflect the reality that the collaboration must modify its perspective, take full ownership of the physics analyses work that must come after the large data sets are taken, and accept the obligation to finish the full task of producing the physics.

5 Computing beyond 2008

We do not foresee the need for any major restructuring of computing after 2008. Instead we assume an adiabatic transition of the existing computing hardware, software, operations, and support model, up to and beyond the final period of *BABAR* data taking. The required resources can be computed by scaling from the present according to the foreseen increases in luminosity and decreases in demand. In the next section we briefly comment on the computing issues

associated with use of *BABAR* data long after 2008.

5.1 Computing hardware needs for intense analysis

To quantify the computing hardware needs for the intense analysis period, we assume a total integrated luminosity of 945 fb^{-1} at the end of data taking. Another important input is the number of available Monte Carlo events, which is taken as three times (same as the currently used factor) the number of hadronic events in real data.

The hardware required at the beginning of the intense analysis period is estimated by extrapolating the current needs to the final luminosity. It is summarized in Table 3 in terms of disks for data storage and CPU power for running analysis jobs.

Table 3: CPU and disks required in the intense analysis period, estimated by extrapolating current usage to the predicted final luminosity. The *BABAR* standard CPU measure 30SpecInt95 unit corresponds for example to 1/10 of a dual intel xeon 2.8GHz machine.

	Resources needed
CPU (30SpecInt95 unit)	
Data analysis	6185
MC analysis	12050
Total	18235
Disks (TB)	
Data	490
MC	875
Older processing	410
Fraction of raw data	55
User data	275
Technical reserve	315
Total	2420

During the intense analysis period, *BABAR* will need to maintain its computing hardware for analysis at about the same level as at the end of data taking and to keep using several TierA's.

5.2 Computing professional support needed for intense analysis

It may be difficult to attract capable physicists to take on computing-oriented services tasks, such as the development and maintenance of the skim system, through the end of data taking and beyond. Such tasks may need to be taken on by computing professionals. The overall plan for the evolution of the *BABAR*

support by computing professionals is currently being developed by a separate task force, and so will not be discussed here.

A detailed description of the more general *BABAR* computing support by SCCS can be found in appendix B.

5.3 Monte Carlo Production

As the data sample doubles over the next two years the demand for simulated Monte Carlo events will increase at least proportionally. Monte Carlo production in *BABAR* utilizes distributed computing resources at member universities as well as national and international computing centers. This model has been very successful and it seems reasonable to continue this approach for the remaining two years of data taking and beyond. It is however likely that during this time some of the existing production sites will shift their resources toward the LHC physics program.

Part of this loss can be compensated if other sites increase their production rate. At least some of that increase can come from replacing older hardware, due to the increasing performance of PCs with time. Informal discussions with three of the largest university based production sites (Colorado, Ohio State and Victoria) showed that there is interest to continue this service for the *BABAR* collaboration. Both the University of Victoria and the Ohio State group expressed their interest to work with *BABAR* management to secure the funds required to expand their production facilities. More formal study of this question is beyond the scope of this document.

5.4 Reprocessing

BABAR has reprocessed its data sample 3 times during the course of the experiment, most recently in 2004-2005 (the 18-series reprocessing). The purpose of these reprocessings was to realize the benefits of improvements made to reconstruction algorithms and calibrations for physics analysis, and to apply those benefits uniformly across the data set. The benefits to the analysis community of reprocessing come directly from the improved accuracy of the data, and indirectly through the consistency of the data sample. Data sample consistency also reduces the amount of disk space needed to keep the full data sample accessible, and reduces the manpower needed to support the data. A full reprocessing of the data implies remaking the complete Monte Carlo sample, and a complete skim.

5.4.1 Reprocessing benefits

The specific benefits of a possible reprocessing around the end of data taking are described in detail in appendix C. In summary, there are incremental improvements that require a reprocessing to deploy for all subsystems except the EMC. Improvements are also foreseen for both generation and simulation of the Monte Carlo production that would go with the reprocessing. Some of these

improvements have already been deployed for run 6, while others need work to be ready for run 7. While each of the individual improvements is small (few %), the cumulative effect to the quality of the data could have a substantial effect on the final analysis. A detailed evaluation of that impact is analysis dependent and beyond the scope of this document.

5.4.2 Computing hardware needs for a full reprocessing

A complete reprocessing of the data consists of

- the reprocessing of the complete sample of real data,
- the production of a full Monte Carlo sample using the same reconstruction code as real data,
- the skimming of the real data sample,
- the skimming of the Monte Carlo sample,
- the data distribution to the relevant sites.

The hardware required for each component of a full reprocessing of the data is summarized in Table 4, where the needs at *BABAR* main computing centers (TierA's) and at other sites are quoted separately. Traditionally, data processing and skimming are done centrally in a TierA, while Monte Carlo production is distributed at many sites. We assume here a fraction of Monte Carlo production done in TierA's of 25%.

Table 4: CPU and disks required for a full reprocessing of the data.

	CPU (30SpecInt95 unit)	Disks (TB)
Data reprocessing	2960	45
MC production	11250	170
at TierA's	2810	40
at other sites	8440	130
Data reskimming	1820	20
MC reskimming	5470	55
Data distribution	470	70
Total	21970	360
at TierA's	13530	230
at other sites	8440	130

These numbers assume that the reprocessing of real data will be done in 8 months, the MC production in 12 months and the skimming in 6 months. The amount of disks quoted here is the necessary disk buffer space needed for the production, not the actual disk space to store the data which is given in Table 3.

These numbers assume the entire data sample of 945 fb^{-1} and its associated MC needs to be reprocessed. If we choose a reprocessing scenario where the code is frozen by run 7 (see the next section), run 7 itself does not need to be reprocessed, and the additional resources needed for reprocessing should be scaled by the ratio of luminosities, (nominally 65%).

5.4.3 Reprocessing Scenarios

We consider three reprocessing scenarios in this document. These scenarios are not exclusive, but they span the range of reasonable choices that might be made by *BABAR* management.

Table 5 gives an indication of the evolution with time of the peak CPU usage for the sum of analysis and production under the three scenarios described below. Here we have added all resources at TierA's and smaller sites, and considered that skimming cycles will happen whatever the decision on reprocessing.

Table 5: Estimated peak CPU power (30SpecInt95 unit) required with time for the three proposed scenarios.

Year	2007	2008	2009
No reprocessing	20000	31000	26000
Early reprocessing	20000	40000	26000
Late reprocessing	20000	31000	40000

The first scenario is not to reprocess. The advantages of this scenario are that the final data set will be available immediately after run 7 is complete, and that manpower and other resources which might have been devoted to reprocessing could be used instead on analysis. The costs are the lost gains to analysis from the improvements not made, and in the extra work required to maintain and analyze a final data set with small inconsistencies coming from having been processed with different releases. These inconsistencies could also have consequences for long-term analysis such as forcing *BABAR* to maintain several legacy release structures instead of just one.

A second scenario performs the reprocessing roughly simultaneously with run 7 processing, using the same reconstruction version for both productions. This scenario requires freezing development at the start of run 7. As the reprocessing schedule in this scenario will depend on the run 7 luminosity profile, (possible) software problems, and the availability of hardware and manpower, the results may not be fully available until several months after run 7 ends. The advantages of this scenario are that most of the improvement benefits can be realized, and the final data set will be consistent and available near the end of run 7. In particular, the core analyses which will be the focus of the intense analysis will benefit from these improvements. The costs are the additional manpower and computing resources needed to run the reprocessing simultaneous to run 7 operations and processing, and the loss of additional improvements that might

have been developed during run 7. If resources are frozen at their nominal 2007 level (31000 SLAC units) and the analysis load scales as predicted, the reprocessing of runs 1-6 under this scenario would not finish until well into 2009.

The third scenario starts the reprocessing after the end of run 7. Subject to the usual uncertainties, the reprocessing results in scenario 3 would not be available before summer 2009. The advantages of this scenario are an extended period of development which allows for more improvements, and no need to find new computing resources during run 7. The costs are that the final data set would not be ready until nearly a year after the end of data taking, and so the core analyses of the intense analysis period would not benefit from the improvements. Additional costs come from the transition from the pre-reprocessing sample to the reprocessed sample, which will require an increase in disk and human support during the transition interval. There is also an increased risk with this scenario that the manpower required for a successful reprocessing effort may be impossible to find.

Our most recent experience with the release 18 reprocessing shows that, from the time the code is frozen, it takes roughly 3-4 months to actually begin production [2]. Most of this time was spent validating the changes, and in tracking down experts capable of fixing the problems uncovered by the validation. As the expertise in the collaboration decays, this situation can only get worse. Past experience also shows that a successful reprocessing requires the dedicated effort of 1 or 2 people willing to spend essentially full time on the project.

6 Long-term access to *BABAR* data

In the absence of a Super *B*-Factory, the *BABAR* data set will not be superseded. It may be important to retain the ability to analyze this data set significantly beyond 2008. For example, it may be interesting to study the lower-energy consequences of discoveries at the LHC or the ILC. However, this will require work on our software system.

BABAR software currently has many dependencies on external software, such as operating systems, compilers, and standard libraries. As the computing support manpower decays with time, it will become increasingly difficult to maintain compatibility with these packages. Other experiments have dealt with this situation by instituting a *freezeout* date, beyond which the code base and its dependencies are not allowed to change. This model requires providing dedicated (frozen) platforms with access to the full data sample. It is not clear this model will function in today's environment of constantly changing cyber security threats or with *BABAR*'s large data sample.

Some software research and development work will be required to allow long term access to *BABAR* data. This could follow the freezeout model, or it could be along the lines of reducing the scope and dependencies of our software. It would be best to start the R&D soon, before the necessary expertise decays away. A partial FTE of a computing professional under the guidance of a physicist is

probably sufficient.

7 Summary

- The core *BABAR* physics program consists of the measurements listed in Table 1. It is important that *BABAR* publish these results on the full dataset, as it will significantly constrain the flavor structure of physics beyond the standard model in a way that is complementary to the LHC. They must be completed in a timely fashion, Summer 2010, with a tail extending to 2011.
- A survey of the collaboration indicates that there will be sufficient personnel to complete this core program through the end of data taking and shortly thereafter. However, the number of people and FTEs will decline to approximately 60% of current levels during the intense analysis period following the end of data taking. Beyond 2010, there are significant uncertainties due to external factors, in particular the progress on other large projects. However, there will likely be sharper drop in personnel levels, particularly among postdocs and graduate students. The decline will be rapid for groups working on LHC or other experiments taking data, and slower for those working on the ILC or other longer-term projects.
- Maintaining a critical mass of *BABAR* collaborators at SLAC will be an important component in ensuring the timely production of the core physics program.
- The broader exploitation of the data is likely to continue for 3–6 years, driven by people not on other running experiments.
- In the absence of a Super *B*-Factory, the *BABAR* dataset will not be superseded in the foreseeable future, and so it is clearly in the interest of our community that it be analyzable in the long term. However, this could require significant resources, particularly if the goal is to allow access to the broader community. The cost/benefit tradeoff requires additional study.
- Reprocessing would incrementally improve tracking and particle identification performance, and the corresponding new MC events will be generated using a new version of GEANT, and with updated event generators and distributions. It will also produce a consistent data set for the final analyses.
- Reprocessing Runs 1–6 during Run 7 would produce a fully reprocessed dataset at the end of data taking, ensuring that it would be available to the largest number of analyses. The corresponding date for freezing code would be in fall 2007, the same date as code aimed at Run 7. Reprocessing after the end of Run 7 would allow approximately one year longer for code development, but there is a risk that many analysts would be unwilling to switch to a new processing a year after the end of data taking.

- To complete reprocessing and Simulation Production in times comparable to previous reprocessings, it will be important for the collaboration to not only maintain, but to expand, the existing resources both at the Tier A sites and the university SP sites. Reprocessing Runs 1–6 during Run 7 would require a substantial increase in computing hardware relative to Run 6, and would place a substantial load on computing personnel. Reprocessing the full data set following Run 7 would require approximately the same amount of computing hardware, but a year later.
- Whether we reprocess or not, there is a significant load of non-analysis tasks, including physics tools (such as particle ID, tracking efficiencies, and neutrals corrections) and computing operations (such as skim production, PR, and data quality). The total service work load per FTE will not substantially decrease from current levels until the end of reprocessing. However, the mix will change, with a greater emphasis on computing tasks. It may be necessary to move computing professionals currently supported through common fund agreements into computing operations in the post-running era.
- A significant effort by the collaboration on committees such as the Publications Board and Speakers Bureau is not currently counted as service work. Service credit awarded to such activities may help maintain high-quality reviews as the collaboration size declines after the end of data taking.

A Charge to the Committee

Dear Colleagues,

With the expected completion of the enormous undertaking that began on August 21, 2006, to upgrade PEP-II and IFR by the end of 2006, the experiment will enter its “1/ab” era, poised to raise the size of our data sample a factor of $\sim 2\ 1/2$ by the end of the data taking phase of the experiment in September 2008. While we expect to continue the *BABAR* tradition of analyzing the data on time scales set by major conferences in the next two years, the full exploitation of this enormous data set to achieve the scientific mission of the program will take several years beyond 2008. As affirmed by numerous internal and external reviews of our program in the past few years, the outcome of the primary scientific goal of the experiment, that is precision tests of the CKM paradigm and search for effects of physics beyond the Standard Model through measurements of properties of rare decays of B mesons, charmed hadrons and the tau lepton, is of significant importance to understanding the nature of the new physics, which is likely to emerge from the LHC experiments in the later part of this decade. Clearly, timely achievement of this goal requires a careful planning of the shape of the *BABAR-Beyond-2008* and its available resources and manpower, including a fresh look at the management of analysis effort and the distribution of computing resources in the collaboration. Up to this point, a 1st-order model of the program, outlined below, has been used by the *BABAR* management for the purpose of broad planning and as a guide in responding to inquiries from funding agencies and other interested bodies. Clearly, the time has come to prepare for a more detailed plan.

An outline of the 1st-order plan of the *BABAR* management for *BABAR-Beyond-2008*:

The physics reach of *BABAR* data is very broad, and its exploitation will take many years beyond the data collection phase of the experiment. Beyond the discoveries and surprises that may yet emerge, taken as a whole, the legacy of this breadth is likely to provide a set of constraints on the Standard Model and the effects of New Physics that will remain unmatched for decades. Based on the experience of the previous experiments, and the magnitude and the physics reach of the data set, we expect a period of 6 to 8 years for the analysis phase beyond 2008. Given the fact that some of the key measurements from this data have a major impact in constraining the flavor structure of physics beyond the standard model and are complementary in this respect to the results that are likely to emerge from LHC, it is critical that the collaboration be able to perform these measurements soon after the end of the data taking. For these reasons we characterize the first 2 to 3 years after the data taking period as an “intense-analysis-period”, during which we expect to require the full capabilities of the collaboration resources and manpower to carry out its primary measurements, to be followed by a long term analysis period of 3 to 5 years for a much broader exploitation of the data. A breakdown of the key assumptions in developing this plan is as follows:

- Intense analysis period: Following the end of the data taking phase of *BABAR* in September 2008, a period of 2 to 3 years will be required to perform the analyses of the key channels, which are both critical to achieving the primary scientific mission of the experiment in a timely way, and complementing what is learned at the LHC.
- Make provisions and identify a decision point for a full reprocessing of the data sample in this period.
- Required computing resources: Achieving the goals of the intense-analysis-period, requires the full resources and functionality of the distributed *BABAR* computing system, including the Tier-A and Tier-B centers.
- In the intense-analysis-period, the collaboration will continue to need a highly coordinated effort to best utilize its available manpower and resources. This will involve the continuing presence of *BABAR* collaborators at SLAC, occupying most 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. 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.

On behalf of the *BABAR* management, I would like to thank you for agreeing to serve on this committee of *BABAR-Beyond-2008*, whose charge is to advise the management on developing a detailed model for the shape of the experiment beyond the data taking phase of the program. We ask you to begin your work by examining the validity of the main elements of the 1st-order plan outlined above, if necessary expand on these assumptions, and propose additional criteria and constraints to more accurately define the model. Clearly, demographical data on the collaboration, e.g. expected numbers of students, Ph.D. theses, postdocs etc, is an important input to your evaluation and we expect that you will need help from the management in gathering some of these data, and in some cases you will need to perform your own survey and seek input from the collaboration at large. We also expect that you will consult with various knowledgeable colleagues within the collaboration or in previous experiments,

who may have relevant experience. We hope that the final report on your studies can be presented to the *BABAR* management and executive board on the time-scale of the December 2006 collaboration meeting.

Your report will form the basis for the next steps, which include planning for the implementation of the model by *BABAR* management. The model and the implementation plan will be presented to the laboratory management, the IFC and other funding agencies to justify our requests for funding support and resources. We also foresee that the *BABAR* collaboration council may, upon examining the implication of the proposed model, see the need to consider discussion on collaboration governance in the era of beyond-2008.

Thanks again for your willingness to serve on this important taskforce.

With best regards,
Hassan

Members:

David Brown., LBNL
Riccardo Faccini, Rome
Chris Hearty, UBC (Chair)
Klaus Honscheid, Ohio State
Heiko Lacker, Dresden
Steve Playfer, Edinburgh
Blair Ratcliff, SLAC
Jim Smith, Colorado
George Vasseur, SACLAY

Ex-officio:

Gerard Bonneaud
David MacFarlane
Richard Mount
Soeren Prell

B SCCS *BABAR* Support

Two years ago, over half the SCCS staff were considered to be working on 'B-Factory Operations' a.k.a. *BABAR*. Looking forward to a future where SLAC computing will be driven by several sciences, SCCS now has devised a new and very different breakdown. This new approach was endorsed by the recent BES review. The SCCS effort relevant to *BABAR* is now:

1. Indirect funds: 39 FTE, supports the site network, Windows and Unix account infrastructure and home directories, phones, email . . .
2. Site-wide Science Support: 22 FTE, Landlord responsibility for baseline ability to run clusters, batch system, MPI, mass storage, visualization, . . .
3. PPA-Specific Science Support: 5 FTE, PPA-specific activities such as Geant4, HEP visualization, . . .
4. *BABAR* Support: 10 FTE.

C Reprocessing Benefit Details

Below we list the existing and foreseen benefits a final reprocessing, with the expected impact of these improvements on physics analysis. While we have tried to be comprehensive, this list should be considered provisional, as new projects relevant to an eventual reprocessing may not have been started or even thought of at the time of this report.

- Svt

Svt is working on a new pattern recognition algorithm for finding standalone tracks. The existing standalone pattern recognition algorithms rely on 4 of 5 layers being fully functional, while the new algorithm is more tolerant of detector failures. The new algorithm would be deployed in the event of a large number of module failures, or if the radiation damage gets worse, during the next two years. The new pattern recognition algorithm could be deployed sometime during runs 6 or 7, depending on what happens to the detector.

Svt is also working on a more accurate material model of the modules that should reduce the roughly 500 KeV shift observed in the reconstructed Ks mass (not seen in MC), and might also reduce the delta-E shift. This work could be finished in time to deploy for run 7.

Svt is also working on an improved local alignment which will reduce vertex position systematic errors. A new procedure to fit simultaneously for the svt local alignment and the beam boost is also being considered, which should eliminate the time-dependent Mes fluctuations we currently see. This work might be ready in time for run 7.

Improved calibration constants of Svt dE/dx have recently been calculated [1]; This calibration promises roughly 1.5σ $\pi - \mu$ separation and between 2 and 4 σ $e - \pi$ separation for slow pions (rejecting gamma conversion electrons). This same calibration can be used to improve the momentum resolution of slow pions by roughly 50%, via inverse Bethe-Bloch constraint. The software development for deploying the new constants in production has begun and should be ready in time for run 7.

- Dch

Dch has recently released improvements to their pattern recognition algorithm which recovers some 3% of otherwise lost low momentum ($P < 300$ MeV) tracks. These tracks increase the V_0 reconstruction efficiency by between 10 and 25% (10% for K_s , 20% for A and 25% for γ conversions), and the Mes peak at the B mass increases by roughly 5% in the BSemiExcl skim. These improvements don't seem to significantly improve slow pion efficiency. The improved pattern recognition algorithm is fully developed, but probably cannot be fully validated and deployed before run 7.

Dch just deployed a new dE/dx calibration for run 6 that improves the resolution and uniformity by roughly 20 % [1]. This improves PID for tracks with momenta below 2 GeV.

The Dch hit resolution model is currently inaccurate, causing reduced quality track fits and poor track fit chisquared. One consequence is that the track parameter covariance matrix computed by the track fit underestimates the momentum error (coming mainly from Dch hit resolution) by roughly 30%. An improved hit resolution model would improve the momentum resolution and (more importantly) track parameter covariance matrix, yielding better and more reliable kinematic fits, and probably reducing the tails of many distributions. It is not currently known how to improve the Dch hit resolution model, and no-one is actively working on it. If begun now, this work could be conceivably be ready in time for run 7.

- Trk

Several improvements to TrkFixup algorithms are forseen. One constrains low-momentum pions using the Svt dE/dx measurement, which provides a roughly 50% momentum resolution improvement. Another performs a fit of bremsstrahlung electrons incorporating the photon, resulting in better resolution. The algorithms for these are essentially done, they are missing calibration and validation. In principle these improvements could be applied at the mini level, but it would be natural to include them in a recalibration.

- Emc

As the CM2 micro contains all the digis for all good clusters, and new calibrations at any level above the digi can be applied when reading the micro, the Emc forsees no improvements that would require reprocessing.

- Dirc

The dirc recently discovered a bug in their material model that causes a factor of 2 over-estimate of the error on Θ_C for roughly 2% of tracks. This bug is fixed for run 6. No other reconstruction improvements are foreseen for dirc.

- Ifr

Ifr has recently deployed an improved chisquared correction which should slightly improve muon ID. This will be applied in runs 6 and 7. A reprocessing could apply improved alignment parameters to runs 1 and 2, and improved calibration constants for runs 1, 2 and 3.

- Simulation

The simulation group foresees moving to a new version of GEANT4 either for run 6 or 7. The improvements include better shower models for neutral hadrons and electrons, a better MIP model, plus several bug fixes. The physics impact should be a more realistic simulation of Emc quantities for different particle types, which should reduce the ad-hoc Neutrals and PID corrections.

A new generation of Monte Carlo production would also provide the opportunity to update the generator models and branching fractions used in decay models. While these effects can in principle be handled by event reweighting, reweighting more than a few variables is impractical due to missing correlations, lost statistical precision and general complexity. An up-to-date, uniform generic MC sample would make analysis of the final data sample simpler and more efficient.

References

- [1] Alexander Telnov, BAD 1500 .
- [2] Private communication with David Lange.