

# Proposal for DØ Regional Analysis Centers

*I. Bertram, R. Brock, F. Filthaut, L. Lueking, P. Mattig,  
M. Narain, P. Lebrun, B. Thooris, J. Yu, C. Zeitnitz*

*June 21, 2002*

## **Abstract**

The analysis of data from Run II will be such a significant effort, that in all likelihood it cannot be done by relying on the Fermilab site-based systems alone. Rather, success in this endeavor will require full world-wide participation. This, in turn, requires coordination of software and datahandling. This document motivates the need for and presents proposed specifications for a set of DØ off-site analysis institutions, called Regional Analysis Centers. These institutions would serve local DØ collaborators with computing and analysis resources including data caching, possibly Monte Carlo generation management, database services, and job control management. Initial attempts at specifying requirements for computing infrastructure and other services that these centers might provide are provided. An attempt at identifying the characterization of both a very significant RAC (appropriate to a large computing center) and a minimal RAC (appropriate to a university physics department) has been made. Specific conclusions are enumerated with highlights to needed information and additional studies. This document also lists sites which might evolve to be among the first such centers.

<b>1. INTRODUCTION .....</b>	<b>4</b>
1.1. Assumptions .....	5
<b>2. CHARACTERIZATION OF AN RAC-BASED ANALYSIS .....</b>	<b>6</b>
2.1. An Example .....	6
<b>3. DØ REMOTE ANALYSIS MODEL (DØRAM) ARCHITECTURE.....</b>	<b>10</b>
<b>4. DATA CHARACTERISTICS.....</b>	<b>12</b>
4.1. Raw data – 250KB/evt .....	12
4.2. DST – 150KB/evt .....	13
4.3. Thumbnail – 10KB/evt .....	13
4.4. MC Data Tier .....	13
<b>SERVICES PROVIDED BY REGIONAL ANALYSIS CENTERS.....</b>	<b>14</b>
5.1. Code Distribution Services .....	14
5.2. Monte Carlo Production .....	16
5.3. Batch Processing Services.....	18
5.4. Data Caching and Delivery Services .....	18
5.5. Data Reprocessing Services.....	19
5.6. Database Access Service .....	21
<b>6. REQUIREMENTS OF REGIONAL ANALYSIS CENTERS.....</b>	<b>22</b>
6.1. Location Issues .....	23
6.2. Network Bandwidth .....	23
6.3. Data Storage .....	23
6.4. Database Requirements .....	25
6.5. Summary of Data Storage .....	26
6.6. Computer Processing Infrastructure .....	27

6.7. Support Personnel .....27

6.8. Category B RAC .....28

6.9. Category D RAC.....29

6.10. Category C RAC.....29

**7. POSSIBLE SITES AND CURRENT CAPABILITIES .....30**

7.1. Europe .....30

7.2. United States .....31

7.3. South America and Asia .....32

**8. PROTOTYPE REGIONAL ANALYSIS CENTER PROJECT.....32**

**9. ORGANIZATIONAL AND BUREAUCRATIC ISSUES .....33**

**10. IMPLEMENTATION TIME SCALE .....34**

**11. CONCLUSIONS.....34**

**12. APPENDIX .....36**

12.1. Summary of Conclusions.....36

**BIBLIOGRAPHY .....39**

## 1. Introduction

*The scientific results anticipated from the Tevatron Runs IIa and IIb are of the highest importance for High Energy Physics. The goals for these runs include both sensitive search experiments (such as for the Higgs boson and possible supersymmetric states) and very precise determination of important physical parameters (such as the top quark and W boson masses). Both kinds of measurements are tightly correlated with the broader international program of testing the Standard Model at the level of quantum loops where new physics must make itself known – indeed, the broader program will be driven by the results from DØ and CDF. In order to realize both the potential for discovery and to appropriately extend the reach of the precision measurements these aggressive physics goals require enormous luminosities and hence the resulting volume of data of all kinds will be measured in many petabytes (PB). Further, in addition to high-profile measurements, over the next decade there will be more than a hundred separate analyses resulting in Ph.D. theses for probably many hundreds of graduate students and post doctoral researchers. All of these measurements will tax the collaboration's understanding of the detector to a very fine level and push event simulation to limits of theory and computation which have not been probed before.*

*It follows, then, that the coming decade-long analysis effort will require mobilization of literally hundreds of people: it will have to be truly international. In the past, analyses at the Tevatron have been close to home – local physicists bore the brunt of the effort and local resources were sufficient. This time the situation is different: the size of the data set is significantly greater than FNAL computing resources alone can support and the complexity of the coming analysis will require that the intellectual effort will have to scale both with the data and the magnitude of the problems which will have to be solved. It will not be sufficient to simply spend money at Fermilab for computing power, even if such funds were available.*

*It follows that in order to make full use of the 78 remote institutions in DØ, nearly as much capability for data access, collaborative code development, and intellectual contribution should exist off-site as exists on-site. For data distribution, DØ has a head start: the SAM system currently makes distribution and tracking of significant quantities of data a reality. Managing job processing in a global environment is another matter. Full inclusion of boundary-less job submission and data access will require the incremental deployment of future GRID tools, but SAM will be at the heart of the effort.*

*The DØ experiment is running during a transitional period under which it will be seen whether the GRID can reach the ambitious goals of its proponents. Hence, any offsite capability envisioned for DØ should at least be sophisticated enough for collaborators to be productive with early tools and yet be flexible enough to make use of the future envisioned capability, should it emerge during the experiment's lifetime. This places a burden on planning – to be both aware of possible GRID developments and yet not be totally dependent on them.*

**Conclusion 0.** Remote analysis capability with full access to the data, code, and collaborative analysis is necessary in order to satisfy the physics goals of Run IIa and IIb. A structured environment which systematizes and standardizes these services is the best way to implement this program.

---

This document proposes a particular off-site environment called a Regional Analysis Center (RAC) as a means of helping to organize the next 10 year's worth of analysis and to also best leverage collaborators' abilities to gather resources which can be directed at the DØ analysis project. An RAC is envisioned to be a primary institution with specified resources which serves as a data and computing hub for geographically adjacent and appropriately connected DØ institutions. Possible services provided by and responsibilities of RAC's are the focus of this proposal. Their implementation is anticipated to be incremental in both capability and in their numbers. This report draws a variety of conclusions and lists alternative opportunities where necessary.

Since this document is a first look at this subject, it undoubtedly contains areas which are not fully addressed. The collaboration should consider the technical – and the sociological – requirements and propose suggestions and ideas. The authors are convinced that this opportunity is unique and might change the way we “do business”...for the better.

## 1.1. Assumptions

The tasks that *might* be imagined for off-site analysis centers include: *ab initio* reconstruction of events (i.e., RECO analysis of raw data producing the streamed outputs and individual data tiers), emergency reprocessing at the RECO level due to a possible coding or calibration error, reprocessing of data at the DST level, detector element-level analysis (calibrations, alignments, etc), and physics analysis at the DST, Thumbnail (TMB), and/or ROOTuple levels. In order to construct a picture of what “analysis” might mean in the future, a variety of assumptions have been made, and are detailed below.

### 1.1.1. Are RAC's Off-site Reconstruction Farms?

The above is a question which is often asked in discussions of this effort. The current size of a typical reconstructed event from the DØ detector will be as much as 300KB. The average output rate of the online DAQ system is 25Hz in Run IIa (assumed to be double that for Run IIb), which constitutes a 7.5MB/s average throughput<sup>1</sup>. The number of events in a mean Run IIa year will be on the order of  $7 \times 10^8$ . The evolution of the FNAL DØ reconstruction farm is designed to keep pace with this rate. Also, the FNAL storage requirements for processed data and producing the subsequent tiers of derived data are significant, but likewise expected to keep pace. This leads to an initial primary conclusion:

---

<sup>1</sup> Note that the maximum rate is currently 50Hz and so the maximum throughput is closer to 15MB/s

**Conclusion 1.** It is anticipated that the FNAL processing farm will be sufficient for all of Run II primary reconstruction needs. RAC's are not envisioned for *ab initio* event reconstruction.

This document is organized as follows: A characterization for Regional Analysis Centers might function in a real analysis is addressed in Section 2 by way of an example and a proposed architecture of the DØ Remote Analysis Model (DØRAM) are discussed in Section 3. Section 4 describes the data formats which might be directed to off-site centers. Sections 5 and 6 cover the services and suggested requirements for such centers. Section 7 enumerates currently interested institutions. A proposal for the establishment of a specific project is presented in Section 8. Sections 9 and 10 discuss preliminary thoughts on policies, implementation time scales, and other bureaucratic issues. Section 11 summarizes the conclusions and highlights areas which are incompletely specified and/or need more attention. Section 12, Appendix, collects all of the conclusions. Finally, the last section, 13, is the bibliography.

## 2. Characterization of an RAC-based Analysis

The job ahead in Run II is larger than that of Run I. One clearly noticeable feature is that the collaboration is bigger and the number of off-shore groups is significantly larger. This feature alone has led to an acknowledgement that analyses of Run II data will necessarily involve a larger effort from outside of Fermilab than did the analyses of Run I. As noted, the efforts required in order to reach the levels of precision in top quark and Electroweak, and QCD reactions appropriate to the statistical power of the data will require significantly more sophisticated computing and Monte Carlo study. Suffice it to say, the effort required in order to meet the challenges presented with this gold mine of data will require the whole DØ World's full efforts. Considerable attention will have to be paid to creating an off-site analysis environment which is as capable as that enjoyed by a collaborator who happens to reside at Fermilab. This implies that data delivery, code availability, cross-boundary resource sharing, and database access will all have to be addressed in order that the analysis experience is as negligibly different as possible, regardless of location, or home system idiosyncrasies, or individual institutional resources. If the creation of this environment is successful, what would life be like?

### 2.1. An Example

The ideal circumstance for remote analysis would be the ability for an off-site/off-shore group to be able to make a measurement with minimal on-site presence. This involves, of course, significant improvement in video conferencing capabilities, but more importantly 1) regular and perhaps automated access to versioned analysis software, identical to that maintained at Fermilab and 2) access to those files of the derived data and Monte Carlo data necessary for a particular physics project. This sort of remote, self-contained analysis happened very rarely in Run I.

As noted, such capability is a basic requirement for off-shore institutions and at least a desirable goal for many of the U.S. groups. The purpose of this section is to describe a simple real analysis in terms of what a user might actually do and how that user would rely on the RAC's and the FNAL central site.

The project chosen for illustration is the determination of the  $W$  boson cross section for which one can rely primarily on desktop ROOT tools and storage of only ROOTuples at the user's home site (the "USER"). Roughly, the analysis universe is presumed to consist of the following elements: 1) A set of RAC's, referred to here generally as the "WORLD"; 2) a "USER" which is a physicist or group at a single institution partnered with a specific one of the set of RAC's called the "URAC"; 3) a set of remote institutions which can provide Monte Carlo generation, called the MCWORLD; and 4) Fermilab, the "LAB", which provides raw data and ultimate database services. Roughly speaking, the USER makes use of computing capability, storage and caching volumes, and perhaps database server facilities at the URAC, and through it to the WORLD. The philosophy is that remote sites are used for reduction of datasets into ROOTuples which can be analyzed back at the USER facility.

This is a straightforward analysis requiring standard packages and capabilities. As such, it constitutes an important target for RAC concept design. *In order to be classified as minimally successful, the RAC concept must be able to cope with this measurement, or something like it.*

### 2.1.1. $W$ Boson Inclusive Cross Section Determination

The analysis chosen is the determination of the inclusive  $W$  boson cross section. The assumptions for this example are:

- The primary USER analysis is at the ROOT level, or equivalent
- The analysis may include TMB files resident at the URAC
- The USER is a SAM site
- The URAC with which it is associated is also a highly capable SAM site
- DST's are 100% disk resident and available from RAC's around the world
- The MC calculations are initiated at MCWORLD farms which are SAM sites

The basic steps that are required in order to make the measurement are deceptively straightforward: count the number of corrected events with  $W$  bosons above background and normalize to the luminosity. In order to do this within the assumptions above, a strawman chain of events has been envisioned as an example.

Some actions are presumed to be automatic, such as the delivery of a complete set of TMB files from FNAL to the RAC. Other actions are initiated by the USER (or a physics group). As represented in Figure 1, requests for some remote action are blue lines with arrows from the workstation to some processor connected to a storage medium. Purple lines represent the creation or splitting of a data set and then the copying of that set. Dashed lines represent a copy, usually a replication, from one GL to another. A black line without an arrow represents a calculation.

In an attempt to establish and discuss the functionality of a possible analysis strategy, a strawman logical path has been constructed. In some steps, there may be differences in taste or eventual implementation and places which say "USER" might actually be steps

performed on behalf of a whole physics group. Please note that the details are not important – different people may come up with different analysis strategies. Rather, the goal here is to picture how a plausible analysis might proceed and then to imagine desirable features for the system as a whole.

So, with that proviso, a progression of events for determining the  $W$  boson cross section could be as follows:

- i) The data accumulate at the URAC as TMB files.
- ii) The USER initiates a request, perhaps by trigger, to the URAC for a  $W$  signal and background sample of TMB records.
  - (1) These records are replicated to the USER's workstation.
- iii) A preliminary analysis of the TMB files is performed at the USER's workstation, leading to rough cuts for sample selection. The product of this analysis is coded TMB algorithms for signal and background.
  - (1) The results of that analysis lead to the ability to select a working dataset for signal and background.
  - (2) The presumption here is that the measurement will require precision or detailed event, trigger, or detector information which is not available from the TMB data tier alone.
- iv) A request to the WORLD is initiated for the stripping of DST-level files for both signal and background<sup>2</sup>.
  - (1) This is readily done since the TMB records are subset of the full DST records: hence, the TMB-tuned selection is directly and efficiently applicable to the DST by executing the precisely identical algorithms developed in iii.
  - (2) These DST sets are cached temporarily to the URAC
- v) The USER analyzes the DST files to produce specialized ROOTuples which will contain the records not available from the TMB data.
  - (1) This analysis can proceed remotely at the URAC OR if sufficient temporary storage exists at the USER's workstation, back at the USER's site.
  - (2) The produced ROOTuples are replicated back on the USER's workstation.
  - (3) The DST's can be discarded or backed up to tape and readily reproduced/retrieved if necessary.
- vi) The analysis of the original TMB files could also initiate a project of producing specific Monte Carlo generation at a MCWORLD farm site.
  - (1) This is initially initiated by hand (ultimately with GRID tools?) and a cached DST-level file set of signal and backgrounds is produced and cached, either at RAC's near the MCWORLD sites OR at the URAC.
  - (2) These MC DST data can be replicated back to the URAC and discarded or backed up to tape at the remote MCWORLD farm site OR RAC's where they were cached
  - (3) The exact details of the MC generation are stored in the SAM database for subsequent use by another analysis project<sup>3</sup>.
- vii) The MC DST data are analyzed and ROOTuples are produced at the URAC
  - (1) These ROOTuples are replicated back at the USER's workstation
- viii) The luminosity calculation is initiated after the selections have been made at the workstation

---

<sup>2</sup> Alternatively, an event list could be generated and used throughout the analysis. The presumption here is that this would constitute a huge pick-event task and not be the best way to proceed.

<sup>3</sup> Note, this may be especially important for theoretical and experimental systematic error determination. For this, typically many different MC samples must be produced and for ubiquitous reactions such as inclusive  $W$  production, they might be useful for many analysis projects. Hence, step vi. might reasonably include first a query of the SAM database for information about other inclusive  $W$  MC generation schemes. Hence, metadata sufficient to describe the details of MC generation, including theoretical parameters and assumptions as well as versioning and cuts, must be retrievable.

- (1) The USER initiates a set of queries to the URAC proxy database server which subsequently queues the requests to the LAB Oracle database.
- (2) This might result in a flattened luminosity file set which is replicated to the workstation
- ix) With all of this information at hand, the cross section calculation can proceed
  - (1) Of course it will be necessary to repeat many or all of the preceding steps
  - (2) This might be facilitated through the replay of history, saved out as scripts when the process was first initiated

As can be seen, the USER acts as a conductor, initiating requests for data movement among FNAL, URAC, RAC's in the WORLD, and perhaps many MCWORLD farms, coordinating the reduction of DST's to ROOTuples, and redoing the steps as necessary due to mistakes or to include new data which may still be continuously coming in. (Note that the actual implementation of streaming will affect just exactly how some steps would play out, like step iv.)

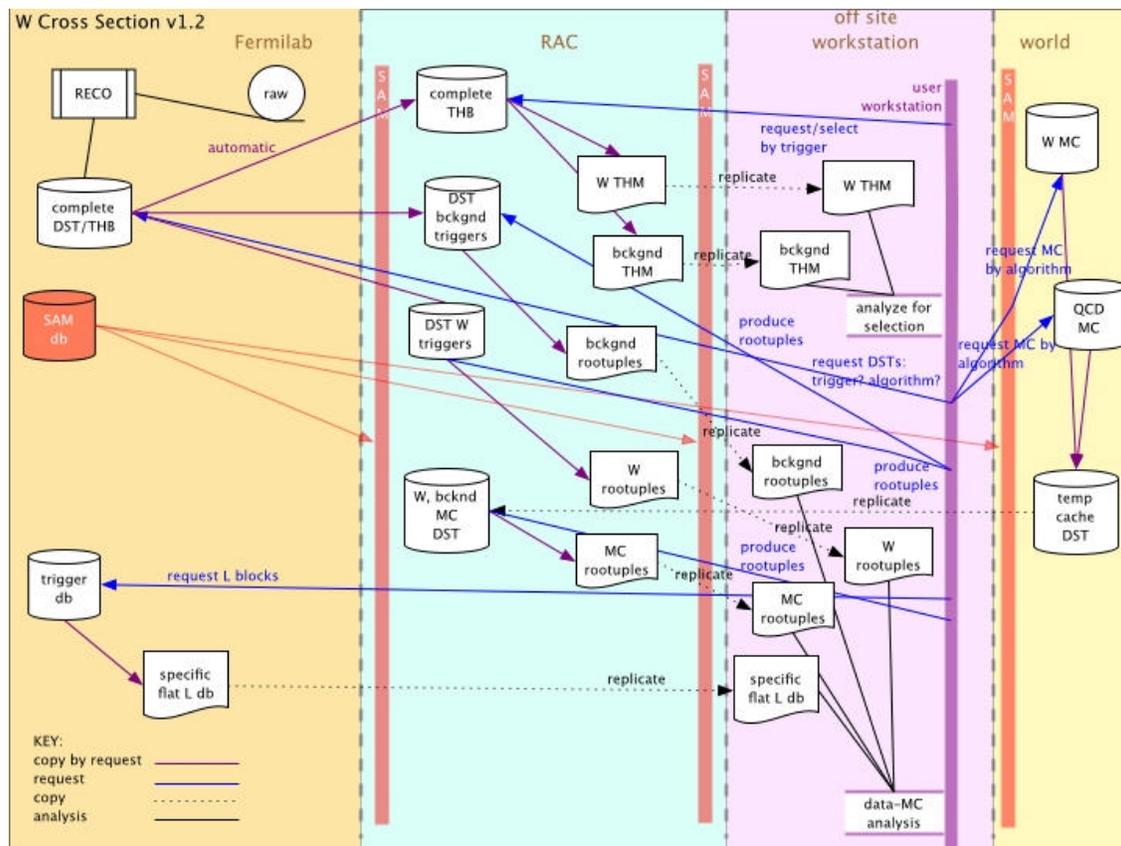
Figure 1 shows the relevant steps in terms of requests, movements of data, and calculations that would either be choreographed by the USER, or actually carried out by that user at the home institution.

So, the requirements for the RAC, in this minimal example might include:

1. Sufficient bandwidth to the outside world for multiple operations of this sort proceeding all of the time.
2. Complete replication of the TMB files at all RAC's.
3. A large amount of temporary storage and the ability for an outside USER to initiate at least staging calculations (e.g., creation of ROOTuples).
4. Access to data in DST form on disk.
5. Siting of a remote proxy database server (see Section 5.6).

An aspect of RAC's which would be desirable, but not necessarily required might include an RAC functioning itself as a MC production facility.

The rest of this document begins the discussion of the details and choices necessary in order to satisfy the requirements and the desirable features for RAC's.



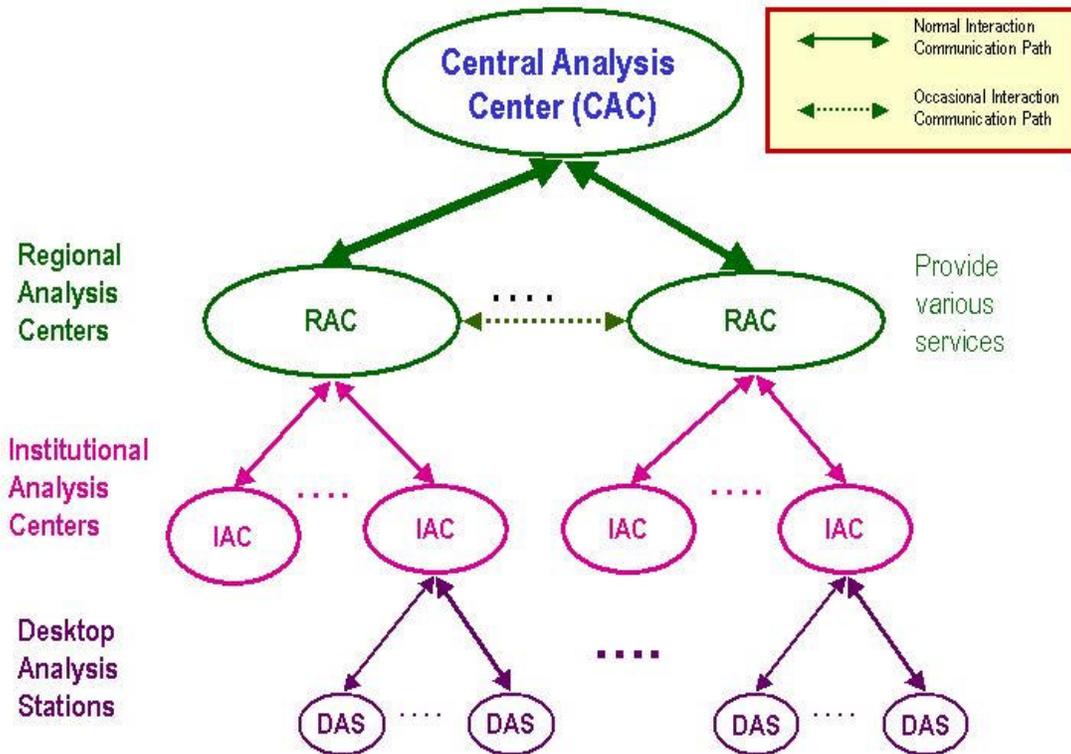
**Figure 1** Simple diagrams of chain of events for *W* cross section analysis. The various geographical locations (GL) are shown as the colored areas: FNAL (brown, left), at least one RAC (green, next), the user’s workstation (pink, next), and at least one MC farm (yellow, right). The vertical purple line in the user GL represents the user’s workstation and roughly the logical (and perhaps temporal) progression of events proceeds from top to bottom along that line. The vertical orange bars represent the SAM system: lines which cross them indicate data transfers which have been managed by SAM.

### 3. DØ Remote Analysis Model (DØRAM) Architecture

As the above example suggests, one could imagine a system of three sorts of analysis centers, each with some dependence on the one above it in the hierarchy. In order of their capabilities they would be: 1) Fermilab, presumably the sole Central Analysis Center (CAC), 2) a set of Regional Analysis Centers, and 3) groupings of Institutional Analysis Centers (IAC), each group associated with a single RAC. The RAC’s act as gateways for their IAC’s both to other RAC’s and to Fermilab and also serve as storage (long and short term) sites supporting their associated IAC’s. Figure 2 shows a sketch of the proposed architecture of a DØ Remote Analysis Model (DØRAM) involving Regional Analysis Centers (RAC) and corresponding Institutional and Desktop Analysis Stations (DAS).

Ideally, resource management should be done at the regional analysis center level. Indeed, within a full GRID environment, when a user requires CPU resources, he/she

would initiate a job to a local server at his/her IAC which would process it and gauge the available resources within the institution to see if the tasks can be sufficiently handled within its own network. If there are not sufficient resources within the institutional network, that request would be propagated to the RAC, and the same kind of gauging will occur within the regional network. This process would continue until the request reaches to the server at the CAC at which time, the priorities will be assigned by a procedure determined by the collaboration. In a complete GRID environment, the fate of the request would ideally be unknown to the user and the completion of the request would result in answers produced at his/her desktop. This is the ultimate scenario.



**Figure 2** A schematic diagram of a DØ Remote Analysis Model (DØRAM) Architecture.

In reality, various institutional priorities and unique queuing strategies make identical IAC activities unlikely. Also, prior to a complete GRID environment, human guidance will be required to forward the requests along the various paths to completion. Just how this interim management of tasks will be handled and how much of it is required is a serious issue for RAC implementation.

**Conclusion 2.** RAC-centered resource management is an important goal. While initially resource management may require considerable human organization, it is desirable to augment and replace that intervention with emergent GRID tools. The priorities assigned to tool deployment remains to be worked out with sufficient Use Case analyses and some real-

world experience. Accordingly, the actual capabilities of the evolving system need to be carefully planned, biased toward smooth running rather than alpha or beta testing of GRID sites.

This architecture is designed to minimize unnecessary traffic through the network. The Central Analysis Center in this architecture is Fermilab which provides the ultimate storage location for the entire data set. The CAC provides corresponding metadata catalogue services and other database services as well. Just what other derived data the RAC's store is a subject for consideration, and a suggestion on this issue will be made below.

As discussed above, the RAC's provide services defined in Section 5 of this document to users in its network of Institutional Analysis Centers (IAC) within its region. The goal is that as many requests from users within a region should be handled within its network without generating additional network traffic beyond.

## 4. Data Characteristics

The kinds of data formats that will be involved in physics analyses have been discussed (see Ref. 1, 2, and 3), but at this time they have not yet been fully implemented nor their designs frozen. The assumptions about the kinds of data formats made available by reconstruction and Monte Carlo (MC) include for data: a rather complete DST, the thumbnail record, and a standard ROOTuple tree (probably augmented by the physics and detector group's individual needs). MC outputs include possible DØGstar files, DØSim files, DST's, TMB's, and associated ROOTuples and ROOTrees.

In this section these data formats are summarized. In order to define the optimal distribution of the data, more and better estimates are needed of the number of users at the various places (e.g. on both sides of the Atlantic) and their data transfer needs. Based on such information realistic studies on data transfer might be informative<sup>4</sup>.

**Conclusion 3.** Continued evaluation of the number of potential off-site users and their anticipated needs should be undertaken very soon. A preliminary census has been done. The follow-up should include more detailed scenarios and/or capabilities for a more realistic assessment.

One may think of various schemes depending on the available resources and bandwidth, ranging from data tiers replicated on disk at all RAC's to tiers which exist on disk only with fractions at each of the RAC's. The solutions realized depend also on the progress of the GRID and related infrastructure. The kinds of the planned data tiers are the following:

### 4.1. Raw data – 250KB/evt

All of the raw data for the experiment will be permanently stored on tape at Fermilab. Rapid access is anticipated through the high volume tape archive system, ENSTORE. At some point, disk archive costs may make tape too expensive, but this will not likely occur

---

<sup>4</sup> Also simulation tools like MONARC might be of help.

until during Run IIb. Raw streams such as cosmic rays or monitoring data might also be useful for a variety of purposes.

#### 4.2. DST – 150KB/evt<sup>5</sup>

This format is expected to contain all information needed to analyze events and be a fraction of the size of the raw data. This format will serve two different purposes. The first is as source of ROOTuples for physics analysis using pre-selected events and the second is for reprocessing some limited aspects of reconstruction before making the ROOTuples. This format will have calibrated and packed “pseudo-Raw Data” from subdetectors (other than central tracking, for which clusters and global tracks are kept), in order to provide the flexibility for redoing most aspects of the reconstruction. It will be in the standard RECO format Refitting tracks is possible, but not with raw track hits.

#### 4.3. Thumbnail – 10KB/evt

The thumbnail file format is primarily expect to be for “...rapid...event selection with rather minimal analysis<sup>1</sup>”. These thumbnail data may be sufficient as a data set for higher statistics analyses or ROOTuple production. There is considerable emphasis on maintaining the small record size. Reference 2 spells out in detail the proposed thumbnail contents.

#### 4.4. MC Data Tier

Due to large volume of MC data being generated, it is currently impossible to store as well as transfer via the network to FNAL all file outputs from various stages of MC production such as generator, DØGSTAR, DØSIM, and DØRECO. The RECO and final output ROOTuple from RECOAnalyze are the only files currently transferred from remote MC production sites to SAM store at FNAL. In the future, as more MC production sites come online, and the MC capacity increases, it may be possible to store only the ROOTuple outputs for MC in SAM. However, it will also be useful to locally store the DØGSTAR outputs as space allows in order that DØSIM can be run on them. In addition, other smaller, intermediate outputs for some samples may also be stored. These are useful for detector performance, algorithm, and triggers studies to name a few, where the full statistics are not necessarily needed.

Table 1 summarizes the assumptions regarding sizes of these records (see *DØ Computing and Software Operations and Upgrade Plan* Ref. 5). The significant amount of overall RAW and derived data only allows for disk storage of complete TMB files at FNAL– the rest of the data will be on tape for Run IIa, accessible only through ENSTORE.

**Conclusion 4.** A complete review should be done of the planned data tiers with special attention paid to potential off-site reconstruction opportunities with DST’s and analysis opportunities with a TMB’s. This should be done before deploying the DST/TMB files.

---

<sup>5</sup> Reference 1 also refers to an EDU250 tier as a “debug format” (DBG). It was supposed to contain enough information to retrace the actual reconstruction processing as well as contain the physics information. It was not expected to be used for all events

Table 1 shows the storage totals for an average Run IIa year.

	size	tape factor	disk factor
raw event	0.25 MB	1	0.001
raw/RECO	0.5 MB	0.2	0.001
data DST	0.15 MB	1.2	0.1
data TMB	0.01 MB	2	1
data root/derived	0.01 MB	8	0
MC DOGstar	0.7 MB	0.1	0
MC D0Sim	0.3 MB	0	0
MC DST	0.15 MB	1	0.2
MC TMB	0.02 MB	3	0.5
PMCS MC	0.02 MB	2	0.5
MC rootuple	0.02 MB	0	0

**Table 1** Storage requirements for each data format for the Fermilab site. The factors in the RH columns model the various data format storage needs for tape and disk as factors (or fractions) of the numbers of RAW events. For example, in this suggestion, only 10% of the RAW events could be stored on disk in DST format.

Table 2 shows the accumulated totals for the experiment, as modeled for the Director's Review (June 5, 2002, Ref. 5). The annual cost of tape media alone (~3/4PB per average RunIIa year) is approximately \$0.5M.

## 5. Services Provided by Regional Analysis Centers

This section considers the details of the varieties of services one might reasonably expect to be provided by Regional Analysis Centers.

### 5.1. Code Distribution Services

The distribution of code represents a much smaller data volume than any access to data – approximately 4GB now, with anticipation of perhaps a doubling as the analysis matures. From the remote user's standpoint, it is important that all analysis stations worldwide have identical directory paths and that all are aware of the appropriate operating systems which function at the primary point of distribution and maintenance, namely Fermilab. This greatly facilitates debugging installation issues and it rather efficiently and naturally creates useful remote experts who might find use as formal or informal first level triage sites for regionally based problems. Incremental upgrading requires a robust versioning scheme and a means of safely and automatically deleting unused files and directories. The recent DØRACE workshop (Ref. 4) and subsequent documentation and updates constitute a very good start to this.

data samples (events)

	1 day	1 year	phase 1 2 years	phase 2 4 years
event rate	2.16E+06	7.88E+08	1.58E+09	6.31E+09
<b>TAPE data accumulation (TB)</b>				
raw event	0.54	197.10	394.20	1971.00
raw/reprocessing	0.22	78.84	157.68	788.40
data DST	0.39	141.91	283.82	1419.11
data TMB	0.04	15.77	31.54	157.68
data rootuple	0.17	63.07	126.14	630.72
MC DØGstar	0.15	55.19	110.38	551.88
MC DØSim	0.00	0.00	0.00	0.00
MC DST	0.32	118.26	236.52	1182.60
MC TMB	0.13	47.30	94.61	473.04
PMCS MC	0.09	31.54	63.07	315.36
MC rootuple	0.00	0.00	0.00	0.00
<b>total storage (TB)</b>	<b>2</b>	<b>749</b>	<b>1,498</b>	<b>7,490</b>
<b>total storage (PB)</b>	<b>0.002</b>	<b>0.75</b>	<b>1.50</b>	<b>7.49</b>
<b>total storage (GB)</b>	<b>2,052</b>	<b>748,980</b>	<b>1,497,960</b>	<b>7,489,800</b>

	1 day	1 year	phase 1 2 years	phase 2 4 years
event rate	2.16E+06	7.88E+08	1.58E+09	6.31E+09
<b>TIER DISK data accumulation (TB)</b>				
raw event	0.00	0.20	0.39	1.97
raw/reprocessing	0.00	0.39	0.79	3.94
data DST	0.03	11.83	23.65	118.26
data TMB	0.02	7.88	15.77	78.84
data rootuple	0.00	0.00	0.00	0.00
MC DØGstar	0.00	0.00	0.00	0.00
MC DØSim	0.00	0.00	0.00	0.00
MC DST	0.06	23.65	47.30	236.52
MC TMB	0.02	7.88	15.77	78.84
PMCS MC	0.02	7.88	15.77	78.84
MC rootuple	0.00	0.00	0.00	0.00
<b>total storage (TB)</b>	<b>0</b>	<b>60</b>	<b>119</b>	<b>597</b>
<b>total storage (PB)</b>	<b>0.000</b>	<b>0.06</b>	<b>0.12</b>	<b>0.60</b>
<b>total storage (GB)</b>	<b>164</b>	<b>59,721</b>	<b>119,443</b>	<b>597,213</b>

**Table 2** Total storage requirements for each data format for the Fermilab site for tape (top) and disk (bottom). The assumptions of numbers of events are from Table 1.

It may be useful some day to make use of RAC's as receive and forward sites for the binaries. But, generally, it is felt that code delivery need not be handled exclusively by the RAC's, but that individual IAC's can UPS/UPD access the code directly from Fermilab.

**Conclusion 5.** Generally, RAC's need not be the sole sites of code distribution to their IAC's. Rather, at least for the early days, individual installation and updating can be done directly from the Fermilab site.

**Conclusion 6.** Robust versioning and a scheme for guiding or automatically initiating stale file and directory deletions should be designed as soon as possible.

Given the fact that currently "test" releases take place on a weekly basis, unless a high degree of automation takes place this will require a non-trivial, and potentially critical

administrative load. Hence, the current recommendation is that RAC's not be asked to distribute software on a regular basis as a common service.

**5.2. Monte Carlo Production**

MC Production is perhaps the simplest service of all as it is essentially self-contained and does not require database access, unlike actual data from the detector (unless minimum bias overlaid events are added). It is not necessary that an RAC be itself a MC Production facility. Currently there are several remote production sites with the capabilities shown in Table 3.

Site	CPU's	CPU Type	Disk Storage	Mass Storage
<b>Current Sites</b>				
BU	10			
CCIN2P3	X% of 200			
Lancaster	200	750 MHz	1.5 TB	30 TB plus
Nikhef	150	750-800 MHz		TB
Prague	30	700 MHz		
UTA	100	750 MHz	0.4TB	
Total Current	440			
<b>Planned Future Sites</b>				
University of Michigan	100	?	?	?
Manchester University	64	1.9 GHz	2.5 TB	?
Oklahoma	25% of 256	2 GHz	500 GB	
UCD Ireland				
Tata Institute				
LTU				
Karlsruhe				
JINR				
University College				
Total Future	~600			

**Table 3** Current and planned Monte Carlo farm sites.

The minimum required capability of these production facilities must be sufficient in order to meet all of the needs for MC production. The planned rate of MC production over the next five years is planned to half of the DC data-to-tape rate (which are 25Hz and 50Hz assumed for Runs IIa and IIb). For a standard 500MHz CPU the time per event of the various stages of MC processing for WW inclusive pairs are shown in Table 4. The average time per event is approximately 216s on a 500MHz PIII processor.

Process/Time per Event (s)	Generation	DØgstar	DØsim	DØREC O	Analyze	Total
<b>0.5 Events Overlaid, Plate Level GEANT</b>						
WW inclusive	0.8	280	20	19	4.5	325
Technirho	0.8	300	20	21	5	345

**Table 4** Times per event for current processing with a 500MHz CPU for plate level samples.

The plan is to devote 25% of the total MC generation to plate level calculations and the remainder among parameterized MC. Since each MC event will be generated at 6 different luminosities, the total time per GEANT event is estimated to be approximately 550s. If 500MHz CPUs are used, then 2500 such CPU's will be required in order to meet

the Run IIa goal of 12.5Hz, assuming a 70% duty factor. If 2GHz CPUs are assumed to be the norm in a few years, then a factor of four reduction in the number of CPUs can be realized, to a minimum of 625 – which is conceivable given the number of institutes interested in MC farm production. But, once analysis of systematic uncertainties begins, the need for MC production can only increase. Likewise, preparation for and running of Run IIb will add to this total with double the rate required, but presumably faster and faster CPU's. It is therefore conceivable that the experiment's need for MC generation will grow and require a thousand or more CPU's within the next five years. Therefore, there is an obvious need to expand the current resources considerably through RAC and IAC involvement.

It is important to understand what enhancements in SAM capability will be required to accommodate such an increase. In order for RAC to provide MC production services, each remote production facility must be able to run a minimal set of DØ software. These minimal requirements are as follows:

1. A fully functional SAM station, capable of storing and retrieving data from any location.
2. mc\_runjob, the DØ Monte Carlo job submission software
3. The full MC suite of software, including but not limited to: the generators Pythia, Herwig, and Isajet; DØgstar; DØsim; DØRECO; Trigsim; and the equivalent of RECOAnalyze (released as tarballs, the full software distribution is not required).

In order to reconstruct these MC data the production facilities may not require database access, however this deserves further thought. For example, it might be necessary in order to overlay data events on MC data and/or might be useful for keeping track of random number seeds, MC generation parameters, etc..

**Conclusion 7.** It may be important to precisely assess the degree of database access required for Monte Carlo production capability which includes overlaid events.

The most immediate need is a means of automatically fielding user requests for MC generation and scheduling the processing of those requests to the available farms. Currently, this is handled by a human being, but with increasing numbers of farms and increasingly detailed needs, this will quickly become untenable. Relief in the form of tools implementing a two-step process would be useful now: 1) tools to automate the location of available MC resources and then 2) tools to actually refer the MC job submission to those resources. This could conceivably become a one step automated process with incremental GRID implementation and hopefully the design of the two-step tools can be engineered to be a part of a preliminary and final GRID solution. A request system is in testing at this writing.

**Conclusion 8.** Early generation tools for interrogating the MC farm sites for available capability and tools for referring the actual job submission to those waiting sites are required now. Evolution of this system into a single step with full GRID deployment should be envisioned when it becomes available.

The next thing that is required is sufficient bookkeeping in order to keep track of all of the various modeling and simulation parameters used in any given generation project. Sufficient storage space will be required for maintaining all generated data for release upon request. It is not necessary to have these data stored on disk. Metadata search capabilities will be required of SAM in order for users to identify archived MC data generated with their desired characteristics.

**Conclusion 9.** MC generation at RAC and/or IAC sites will be necessary in an increasingly wider scale as systematic uncertainties become a focus of measurements. Sufficient bookkeeping capabilities with the flexibility to imbed and modify the MC generation details will reduce the re-generation of already existing scenarios.

### 5.3. Batch Processing Services

The differences among computer center infrastructures, batch systems, and rules frustrate a worldwide, distributed job submission capability. So, having batch processing at RAC's might reduce at least the complications by first serving that batch resource to only those IAC sites which are within the regional cluster. Hopefully relationships can develop so as to minimize bureaucracy with authentication and maximize physicist and resource efficiency.

It is clear that computing centers involved in DØ are generally running different batch system (LSF, BQS,...), different data storage systems (HPSS, ENSTORE,...) and maintain different authentication policies. Running blindly dispersed jobs in such an environment is a challenging goal, part and parcel the project of the GRID effort.

**Conclusion 10.** For the meantime, sufficient batch processing capability should be at each RAC to serve the needs of only the regional IAC's. What is uncertain is the amount of processing that this will entail and an effort to quantify this should be undertaken. Estimates will be made below.

### 5.4. Data Caching and Delivery Services

As can be seen from Table 2, the storage requirements for the CAC are significant. Indeed, the demands on the Fermilab central resources are enough that there will be only tape storage of DST's (roughly 240TB for RunIIa), like that for raw data. Some other data formats will be best used if disk-resident and it should be a responsibility of RAC's to process and manage:

1. Updates of data sent from Fermilab to all RAC's
2. The requests from its clustered IAC users for data which are stored elsewhere
3. Requests from other RAC's for data which are stored at its site

Specific estimates of data storage requirements are made for a high-capability RAC center as described in Section 6.3.

#### 5.4.1. MC Data Storage Services

For Monte Carlo production it is assumed that until the steady state mode is reached that all MC data are disposable and will need to be fully regenerated after each major software release. For this reason no intermediate stages of the MC production will be saved.

The steady state will be reached once there is a stable detector simulation when all detector components are fully simulated and a close to final simulation of the magnetic field is in hand. Once this stage has been reached the limiting process is obviously the GEANT simulation of the detector. At this point in time the experiment will wish to accumulate a standard MC event sample for comparison with the data. Since the processing after the GEANT simulation only takes 10% of the total processing time, and due to the requirement of modeling several different luminosity bins, both the DØgstar output as well as the final output of each event will be stored. This output will be reprocessed after each improvement in reconstruction, or to simulate different luminosities.

#### 5.5. Data Reprocessing Services

Some significant fraction of raw data may be transferred to RAC's and permanently stored in their cache systems for eventual reprocessing. Initially, it is reasonable to expect to have to reprocess a fraction of data once every six months, as the detector groups try to understand the calibrations and alignments, and once a year when stable operations are achieved.

While Conclusion 1 states that initial reprocessing of raw data is not a requirement for RAC's, it is imagined that any major reprocessing of data will likely require resources beyond those of Fermilab. This is due to the reasonable assumption that the DØ farm may very well be fully employed with new data and not be able to double the load to reprocess old data. An example of a reprocessing need which would require raw data access may be a cell-level calibration or channel-level tracking alignment problem.

Earlier, the ideal model analysis included the presumption of DST disk access through off-site storage. The (presumably more frequent) reprocessing scenario also requires this. Just how to accommodate off-site raw data-level reprocessing is complicated. One model is to regularly store a subset of the DØ collider raw data at the RAC's during running. An estimate is that this would require sustained 14MB/s transfers from Fermilab to a variety of off-site and off-shore locations. Another model is to only undertake off-site reprocessing under dire circumstances and plan for mounting a significant emergency effort in order to disperse the data on tape, on disks, or over the net.

**Conclusion 11.** In general, it seems reasonable to presume that in view of the limited computing resources available at FNAL, and of the improvements that are likely to be made to the algorithms used at present for the reconstruction of the collider data, some measure of reprocessing is expected to be an essential ingredient of the RAC's.

A priori it is impossible to predict at what level, and with what frequency, these improvements will take place. Therefore two scenario's can be envisaged in the extreme: one where all data reprocessing starts from the raw data, and one where reprocessing starts at some point of the reconstruction stage – to be determined for each processing pass, but with the greatest gain if only small parts of the reconstruction have to be redone. Both approaches offer advantages and drawbacks. The second approach seems the most flexible and CPU-friendly. However, it can only work if *all* the required input information is available. For the DST, a good candidate as a basis for event reprocessing, can this be expected to be the case in the near future? Probably not! More critical is the understanding of to what degree this can be expected to be the case in the medium and far future.

The most straightforward option is to make each processing pass start from the raw data. If algorithmic improvements occur at relatively low levels of the reconstruction (and this is to be expected for a fairly long time to come), the processing time required is not likely to be substantially more than that starting from the output of any reconstruction stage. For the time being, it is therefore the preferred alternative; but all that is possible should be done to allow to migrate to the other alternative.

While perhaps an issue independent of the RAC question, this brings a focus to another aspect of a mid-course reconstruction. If data volume is an issue, it should also be questioned (in the long run) whether reconstruction output is kept at all. Clearly this aspect is addressed automatically by the first of the above-mentioned options. In the second case, several options exist to ensure that calibration of the detector remains possible: ensure that sufficient information is available at the DST, retain the reconstruction output for some fraction of the data, or take special calibration runs.

So, the issues which must be understood include:

1. How can the DST be ensured of serving as a useful basis for reprocessing? This is a timely design issue.
2. How will DST's be distributed for this reprocessing?
3. What strategies for raw data-level reprocessing can be designed?
4. If such strategies require RAC participation, how will the raw data make their way to the RAC's?
5. For any reprocessing operation, for the first time original data will reside at a location (many locations!) away from Fermilab. Do these derived data sets get transferred back to the central Fermilab facility? Certainly, the answer is "yes" for the TMB, by design.
6. Obviously, such a scheme involves significant networking and bookkeeping resources at the RAC's and the affordability of this should be understood.
7. How can the consistency of the reprocessing be guaranteed? This is not a matter only of code distribution, but also of hardware, OS, etc.

**Conclusion 12.** Answers to questions 1-7 need to be in hand before reaching a conclusion about how to fully characterize generic RAC's.

### 5.5.1. Data Replication

In case of data replication it is of utmost importance to guarantee that the same reconstruction versions are used for all RAC's and that the results are identical (this requires some kind of certification procedure e.g. on a well defined subset). Also this clearly implies that reprocessing has to be done in a centrally organized manner. For example, no RAC by itself could be allowed to replace an official data set with a privately (even if it is improvement) reprocessed version.

In case of an official reprocessing of the data sets it should be clearly defined by a central institution which RAC is to reprocess which data set. More than one reprocessing on any event has to be avoided.

Question 6 above actually involves some significant collaboration soul-searching – is DØ ready to have its data reside away from Fermilab and spread among many sites? If the answer to that is “no”, then after any future remote reprocessing exercise, an effort has to be mounted of transferring all of the RECO results and the subsequent data tiers back to Fermilab, simultaneously with the concurrent regular data transmission to the RAC's *from* Fermilab. Obviously, this will be a significant network, management, and bookkeeping problem which SAM should handle. The consequences of mistakes could be disastrous.

### 5.6. Database Access Service

DØ has implemented a database architecture employing a central, high availability database instance at Fermilab, and specialized servers that deliver information to clients. This has satisfied the existing needs of the collaboration. The central database is used for two types of information storage within the DØ experiment: 1) file metadata and processing information, and 2) calibration constants and run configuration information. The actual delivery of this data to clients is done through middle tier server applications called “Database Servers” (DBS). These servers impart several important qualities to the system including 1) reducing the number of concurrent users of the Oracle database, 2) enhancing the ability to make schema changes to database tables while minimizing the effect on clients, and 3) overall system scalability.

The most important aspect of the three tier architecture when considering the deployment of the system to the global DØ collaboration for data analysis is scalability. Several enhancements are being built into the calibration DBS's to provide the level of operation needed as we move into the era of remote analysis and data processing. First, two levels of caching are being built in, a memory cache and a persistent server side cache. The memory cache reduces the number of accesses to the central database when large parallel farms are processing events for a limited range of detector runs. The server side disk cache will allow persistent storage of large ranges of calibration and run configuration information at local processing sites and if the network to FNAL is disrupted, or the central DB at FNAL is unavailable, processing will continue unaffected. The new DBS's are multi-threaded and many clients can be serviced simultaneously. If no clients are present, the DBS will disconnect from the database.

Another feature that will provide scalability is the ability of the DBS's to run in a proxy mode. This will allow the establishment of a network of DBS's at remote sites with a configuration similar to the RAC model itself. Each analysis site will have local caching servers which connect to corresponding DBS's at the RAC's. These, in turn, will connect to DBS's at Fermilab. These servers are instrumented so it will be possible to monitor the usage and detect hot spots in the system. Additional DBS's can be deployed to increase the heavy load areas of the deployment. It is anticipated that this will satisfy the needs of the collaboration.

**Conclusion 13.** Reasonable database services could be achieved with the development of proxy database servers at each RAC.

**Conclusion 14.** Full testing of the performance of the new DBS implementation should be performed at the soonest available time.

**Conclusion 15.** A test installation of the proxy server idea at a remote site should be done in the near term.

There are a few technical issues that will need to be addressed. First, remote installations that are running within a Virtual Private Networks behind switches/firewalls will require solutions initiated at the operational site. Second, if it is determined that central database at FNAL is a serious and unacceptable single-point-of-failure for the operational of the overall system, it may be prudent to consider deployment of one or more additional databases that will have "snapshots" of the central database information. This is a viable solution because this is read-only data, and the issues of synchronization for a network of databases do not need to be addressed.

## 6. Requirements of Regional Analysis Centers

One of the main jobs for a RAC is to act as a regional caching and storage intermediary between remote users and the full DØ dataset. Questions of computing power, software updating, and database support are likely to be different from site to site, with some IAC's perhaps having significant computing capabilities either inherently, or in conjunction with other projects located at that institution. Nonetheless, it is expected that institutions which wish to serve the collaboration as RAC's should be prepared to make available significant computing resources.

The Fermilab site is likely a unique tier in most respects – having "A level" capabilities. The following descriptions in Sections 6.1- 6.8 characterize the RAC tier which involves a considerable investment and consequent capability. One can imagine that not all RAC's will be created as equal. What will first be described will be a very well-endowed center, referred to as a Category B RAC. Later, these conclusions will be scaled back in an attempt to represent what would be the minimal capability, a Category D RAC, leaving the designation Category C for a site with capabilities greater than the minimum.

**6.1. Location Issues**

Distribution should follow more or less the geographical density of collaborators (in practice, should be able to “serve” a sufficient number of people). Locations should be defined preferably to avoid unnecessary network traffic overlap. Inevitably, the choices of location will be driven by considerations which are delineated by political boundaries (i.e., language, funding agency, common non-DØ projects, etc.). It is anticipated that existing or planned general use computing institutes will seek to qualify as Category B sites.

**6.2. Network Bandwidth**

Stringent requirements are needed for bandwidth to FNAL: of order 1 Gbit/s, which will involve upgrading the current FNAL connections to the Abilene Network backbone. High bandwidth to other RAC’s is desirable, as also fair amount of data have to be shipped between centers. High bandwidth to IAC’s is also desirable (for those institutes that want to be able to look at data “at home”), but perhaps somewhat less urgent (if IAC’s needs can also be served by processing at the RAC’s).

**Conclusion 16.** An evaluation of networking needs for remote analysis should be done for FNAL, U.S. RAC’s, and overseas sites.

**6.3. Data Storage**

As described above, substantial data storage responsibilities will be housed at the RAC’s collectively and individually. Here, estimates of the kinds and amounts of storage requirements are modeled. Table 5 shows the possible allocations used for this document for Category B and is in the same format as Table 1.

	size	tape factor	disk factor
raw event	0.25 MB	0	0
raw/RECO	0.5 MB	0.001	0.005
data DST	0.15 MB	0.1	0.1
data TMB	0.01 MB	1	2
data root/derived	0.01 MB	0	1
MC DØGstar	0.7 MB	0	0
MC DØSim	0.3 MB	0	0
MC DST	0.3 MB	0.025	0.05
MC TMB	0.02 MB	0	0
PMCS MC	0.02 MB	0	0
MC rootuple	0.02 MB	0.3	0.1

**Table 5** A model for the storage requirements for each data format for a Category B RAC. The meaning of the columns is the same as that of Table 1.

The product that makes any of this ambitious datahandling possible is a robust SAM system. So, adequate support for SAM operations and near-term development is essential.

**Conclusion 17.** A robust and reliable SAM system at every worldwide site is essential. This means that adequate support for both development and operations must be provided.

### 6.3.1. Thumbnail Storage

Thumbnail files, approximately 8TB/y (~16TB/RunIIa), are expected to be disk-resident always. Since they serve as both a low-level analysis format as well as an efficient tool for selection, they should be readily available at all times to all collaborators in both current and previous release versions.

**Conclusion 18.** All TMB records should be disk resident at all RAC sites twice. Total for Run IIa for TMB storage of 16TB disk per RAC.

### 6.3.2. ROOTuple/Derived Data Formats

There will inevitably be a variety of derived formats developed in order to make true desktop analysis feasible. This was the case across the whole of Run I analyses. In order to estimate this, an additional tier has been defined as the size of a TMB file, but nonspecific in content or origin (ROOT, or PAW or ?).

**Conclusion 19.** Significant storage for project formats should be available at RAC's of the same order as the total TMB cache. Total for Run IIa for DERIVED storage of 16TB disk per RAC.

### 6.3.3. DST Storage

A significant benefit to the analysis would be disk-resident availability for DST data. The full set of DST's corresponding to the physics streams of interest at the RAC should be available at the RAC site. Because there will be a considerable emphasis on desktop analysis, the source of ROOTuple production should be readily available. Full DST storage at any single RAC of nearly 150TB per year is excessive. However, if the data are split among, say 10 potential sites, then the load is considerably reduced. It is important to note that this is a service which both greatly enhances the overall analysis effort and is unique to the RAC concept: it compliments and does not duplicate any Fermilab service.

**Conclusion 20.** Complete DST data formats should be disk resident within the sum of the RAC sites: Total for Run IIa for DST data storage of 24TB disk per RAC, which presumes 10% of the total at each RAC site.

### 6.3.4. Monte Carlo Data Storage

Not all RAC's will necessarily be Monte Carlo production sites themselves. Significant MC generation will undoubtedly occur at a number of IAC's. For RAC's which are themselves MC production sites or for those which serve MC production sites within their cluster, significant temporary storage will be necessary. The original plan is to store only ROOTuple outputs, and so the storage requirements are minimal. However, should there be a need to store these results for future analyses requiring the same running

conditions, substantial tape storage might be useful. There is sentiment for storage of some MCDST files and so this has been built into the estimate.

**Conclusion 21.** MC storage for per-demand generated events is primarily limited to the ROOTuple needs, with a nominal MCDST compliment as well: Total for Run IIa MC ROOTuple data storage of ~5TB disk per RAC and 10TB tape per RAC; MCDST disk storage of approximately 25TB, presumes 5% of the dataset on disk as shown in Table 5.

### 6.3.5. Temporary Data Caching

Caching for ROOTuple production or filtering or even short analyses would be necessary for both data and Monte Carlo results. Also, staging for MC results requested from other clusters might be required while derived formats are produced for eventual delivery. This is roughly estimated to be of the order of 10% of the total permanent storage.

**Conclusion 22.** Temporary storage needs for staging of data and Monte Carlo analysis and ROOTuple generation are estimated to be 10% of the total of each data format: Total for Run IIa for temporary cache of ~11TB disk per RAC. A more accurate estimation of this need is required.

## 6.4. Database Requirements

In the model in which database proxy servers are located at each RAC, the requirements for storage are minimal in size. At the Fermilab DB home site, storage is non-commodity media and ~40x more expensive than commodity production disks. It is not clear that similarly high-quality storage is required for remote databases and whether much at all is required as devoted to storage in the proxy mode. If a second Oracle database is replicated due to concerns of single-point-of-failure prevention, even then commodity disks may be sufficient. The Run IIa estimate of 1.1TB for Run IIa from Ref. 5 is used to support databases and SAM.

**Conclusion 23.** Storage needs for serving db setups at RAC will not likely exceed a few TB for all data taking. Total for Run IIa for database/SAM needs of 1TB.

data samples (events)

	1 day	1 year	phase 1 2 years	phase 2 4 years
event rate	2.16E+06	7.88E+08	1.58E+09	6.31E+09
<b>TAPE data accumulation (TB)</b>				
raw event	0.5400	0.000	0.00	0.00
raw/reprocessing	0.0011	0.394	0.79	3.94
data DST	0.0324	11.826	23.65	118.26
data TMB	0.0216	7.884	15.77	78.84
data root/derived	0.0000	0.000	0.00	0.00
MC DOGstar	0.0000	0.000	0.00	0.00
MC DOSim	0.0000	0.000	0.00	0.00
MC DST	0.0162	5.913	11.83	59.13
MC TMB	0.0000	0.000	0.00	0.00
PMCS MC	0.0000	0.000	0.00	0.00
MC rootuple	0.0130	4.730	9.46	47.30
<b>total storage (TB)</b>	<b>0.6242</b>	<b>30.748</b>	<b>61</b>	<b>307</b>
<b>total storage (PB)</b>	<b>0.001</b>	<b>0.03</b>	<b>0.06</b>	<b>0.31</b>
<b>total storage (GB)</b>	<b>624</b>	<b>30,748</b>	<b>61,495</b>	<b>307,476</b>

	1 day	1 year	phase 1 2 years	phase 2 4 years
event rate	2.16E+06	7.88E+08	1.58E+09	6.31E+09
<b>TIER DISK data accumulation (TB)</b>				
raw event	0.0000	0.000	0.00	0.00
raw/reprocessing	0.0054	1.971	3.94	19.71
data DST	0.0324	11.826	23.65	118.26
data TMB	0.0432	15.768	31.54	157.68
data root/derived	0.0216	7.884	15.77	78.84
MC DOGstar	0.0000	0.000	0.00	0.00
MC DOSim	0.0000	0.000	0.00	0.00
MC DST	0.0324	11.826	23.65	118.26
MC TMB	0.0000	0.000	0.00	0.00
PMCS MC	0.0000	0.000	0.00	0.00
MC rootuple	0.0043	1.577	3.15	15.77
cache	0.0139	5.085	10.17	50.85
db/SAM		0.500	1.00	2.00
<b>total storage (TB)</b>	<b>0.1393</b>	<b>50.852</b>	<b>102</b>	<b>509</b>
<b>total storage (PB)</b>	<b>0.000</b>	<b>0.051</b>	<b>0.10</b>	<b>0.51</b>
<b>total storage (GB)</b>	<b>139</b>	<b>50,852</b>	<b>101,704</b>	<b>508,518</b>

**Table 6** Run IIa and IIb assumptions for tape and disk requirements for a Category B RAC. The assumptions about running include Run IIa running through 2004, off 2005, and then Run IIb running 2006-2009 with two times the event rate and 25% larger event records as compared with Run IIa. The data assumptions include the issues discussed in the text plus reasonable assumptions for tape backup of derived data.

6.5. Summary of Data Storage

Table 6<sup>6</sup> summarizes the data storage requirements for both Run IIa and IIb following this model and the Computer Policy Board’s plan for data rates (see Ref. 5). Approximately 50TB per average Run IIa year per Category B RAC seems a reasonable target for this most capable RAC. This result may be modified considerably if particular

<sup>6</sup> In order to maintain consistency, it is a worksheet from the same spreadsheet used by the CPB for overall analysis and therefore contains more general entries which appear to be zero for this application.

sites do not choose to store MCDST's. Tape storage was also estimated, but would likely vary considerably with RAC's mass storage system capabilities.

## 6.6. Computer Processing Infrastructure

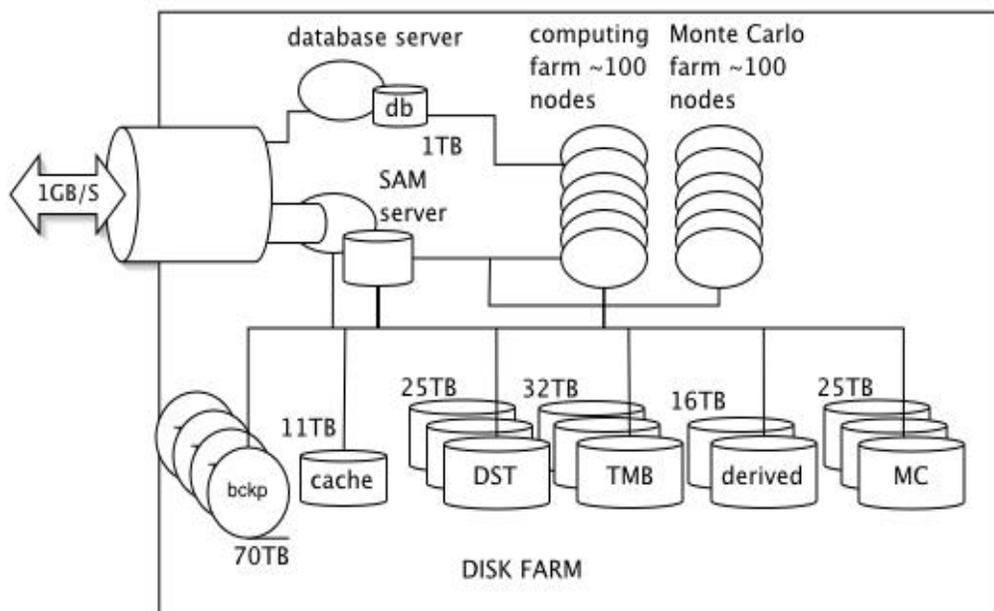
It is difficult to accurately model the actual amount of analysis CPU resources which will be required. Certainly, it scales with the number of people and the number of physics topics (think "theses") which can be addressed...which in turn roughly scales with luminosity: for example new detection capability results in new analyses in B physics, legitimate discovery potential for the Higgs bosons and searches for New Phenomena. A model for predicting the requirements for analysis computing has been devised by the CPB for Reference 5. There, some of the concern was for the creation of the derived datasets, such as the TMB's and DST's. The latter is expected to be rather computationally intensive. For RAC purposes, however, this is not an issue. The model used was to generalize analysis computing into three kinds: a) high resource, but few users (5 seconds per event), b) medium resource with medium number of users (1 second per event), and c) very low resources but with many users (0.1 seconds per event). For the purposes of RAC's, high resource tasks include personal derived data set creation and Monte Carlo event generation. Medium resource tasks include background studies and efficiency determination, detector performance studies, algorithm development and optimization, and trigger studies. Finally, low resource tasks include generation of PMCS Monte Carlo test samples and end-level user analysis on derived data sets.

The CPB model presumes that the number of jobs of low resource computing is 150. It is not unreasonable to presume that a Category B RAC might scale by as much as a factor of  $1/10^{\text{th}}$  of a CAC load and so an estimate of computing at a Category B RAC has been chosen to be a high-side total of approximately 10% of the Fermilab investment in numbers of CPU's. The result of that analysis suggested that 1790 CPU's would be needed, spread over 5 years from 2003-2007. Hence, for these purposes, a rough estimate for a high-side RAC might be 200 nodes spread over that timeframe, or approximately 40 CPU's per year and hence a reasonable expectation for a new RAC would be computing at approximately 100 nodes of modern Pentium class computers running Linux for analysis for a 2-3 year Run IIa analysis. MC production has been estimated at approximately 100 CPU's per new farm.

**Conclusion 24.** Batch resources per Category B RAC should be on the order of 50 nodes per year of modern PC nodes running Linux.

## 6.7. Support Personnel

Once there is convergence on the hardware requirements, a relevant study must be done in order to estimate the support personnel in each RAC, in order to support the infrastructure. While code distribution is not considered a priority RAC responsibility, RAC's should be the primary center where any and all new code should reside. In that spirit,, the RAC must have a mirror copy of CVS database for any required synchronization of updates between RAC's and CAC. It must also keep the relevant



**Figure 3** Sketch of a Category B RAC site with assumptions developed in the text.

copies of databases, and act as a SAM service station to provide data and ROOTuples to IAC's and even DAS.

Each RAC and/or its partner IAC's should have personnel sufficient to provide triage support for installations within their "branch" of the DØ-SAM network. This includes help in DØ code distribution and use, SAM station installation and support, and help with general DØ software applications. Problems arising that are not solvable at the RAC site, can be relayed to other RAC sites for assistance, and finally to FNAL for resolution when required.

**Conclusion 25.** Planning should begin early for constructing and maintaining a sophisticated electronic helpdesk and FAQ for DØ software installation, implementation, and use issues. Triage strategies should be planned.

## 6.8. Category B RAC

As a generalization, a Category B RAC can be now summarized. Figure 3 attempts to do so with the assumptions that it is a MC generation site and that it does provide a significant computational service to its regional IAC's and DAS's. The totals in each category presume resources approximately appropriate to the entirety of Run IIa, as described above. It does not assume capability for raw-data reprocessing. Table 5 summarizes requirements for both Run IIa and IIb. The assumptions used are those from

Ref. 5 document prepared for the DØ internal review and Director's Review from June of 2002.

### 6.9. Category D RAC

The Category B level of investment is considerable, although need not appear all at one time. From the other end, it is important to estimate a reasonable minimum capability. This has been approximated based on subtractions from the Category B calculations:

**Monte Carlo.** A Category D RAC might not be a MC Production site and would not have the requirement of being a data storage center for any IAC in its cluster which is a MC Production site. This means that the computing platforms can be significantly reduced in number (subtract 100 nodes) and that the disk storage requirements can be significantly reduced as well (subtract 20TB).

**Batch Processing.** A Category D RAC would presumably have a reduced batch processing assignment (subtract approximately 50 nodes?) and consequently reduced disk storage required: less derived data cache (subtract 8TB?) and less temporary cache (subtract 5TB?). Finally, it may be possible for particular regions to satisfy their physics analysis needs with TMB files limited to particular streams, rather than 100% of the data (subtract 8TB?).

What should not be reduced in scope is the commitment to the appropriate fraction of DST storage on site at each RAC, sufficient networking capabilities, full SAM functionality and expertise, and a database proxy server.

Reductions in scope by these amounts result in an ultimate, minimal disk storage capacity of approximately 60TB of disk storage and ~50 processors, which is a reasonably sized facility, manageable from within a university department.

### 6.10. Category C RAC

A Category C site, in this calculus, would be an RAC which is more than minimum (for example, perhaps with MC Production capability), but not as fully endowed as the Category B sites.

In evaluating this characterization, it is important to keep in mind two items:

1. Any commitment to becoming an RAC of any size carries with it the obligation of professional support for the equipment and general system support.
2. No single RAC need be fully-formed at birth: all can start small and grow with the need, perhaps over 3 years at least for Run IIa analysis. The investments can then be tailored to funding profiles which fit the opportunities which become available.

## 7. Possible Sites and Current Capabilities

There are a handful of interested institutions currently on only two continents. Obviously, once characteristics are better defined and perhaps after some success is in hand with working centers, more will become interested.

**Conclusion 26.** It might be useful to get a non-binding expression of interest from potential institutions just to see what the maximum might be, and to determine whether interest is not sufficient to support the concept. Too few sites will make the burden on that small set perhaps too significant for their viability.

The institutions currently expressing interest are listed below. Also listed with them are their current resources and commitments.

### 7.1. Europe

#### 7.1.1. Karlsruhe

The computing center at the *Forschungszentrum Karlsruhe* started last year and will be used by all four LHC experiments, Babar and both Tevatron experiments. It is foreseen as a Tier-1 center for the LHC computing. Currently some experiments (ALICE, ATLAS, BaBar, CDF, DØ) have already an individual software server available and BaBar is actively using the center for data analysis as well as MC generation.

In total 36 computing (dual machines) nodes are available now and an additional 32 soon. This gives 136 CPUs (1.3 GHz PIII). The plan is to increase this by another 50 CPUs until fall '02 and 200 CPUs in '03.

A tape system with a capacity of 53TB is available now. In terms of disk space the German DØ group will have 15TB available this year (1.6TB already in place). In terms of network the current link is only 34Mbit/s, but will be increased to 155Mbit/s soon. A first test with 622Mbit/s is foreseen for fall 2002.

#### 7.1.2. NIKHEF

High bandwidth connections exist within the Netherlands in the form of the Surfnet academic network, with a 10 Gbit/s backbone at present and 80 Gbit/s foreseen at the end of 2002. This network extends in fact all the way to Chicago. In addition, NIKHEF is the center for Dutch HEP research, and is also a tier-1 Grid site. A Linux farm consisting of 50 dual CPU, 800 MHz machines is available and is used for the production of Monte Carlo events. Access exists to a 9310 PowderHorn tape robot with a maximum capacity of 250 TB (shared with other users). The farm is being run at an approximately 1 FTE manpower basis.

#### 7.1.3. IN2P3, Lyon

The Computing Center of IN2P3 at Lyon is national French center used by Nuclear and High Energy Physics community. It can provide a large CPU power and Mass Storage.

The CPU is essentially provided by PCIII Linux machine (about 17K Spec Int95) but shared with other experiments. The mass storage is based on HPSS (High performance Storage Services). The actual limitation is given by 36000 cartridges accessible by the robot (about 1000 TB). The amount of disk is more than 20 TB and the IO is using RFIO protocol. CCIN2P3 has a good network connection with Fermilab. The actual bandwidth with FNAL is 20 Mb/s and has to be improved in the future. The CCIN2P3 is also involved in the European DATA GRID project.

## 7.2. United States

### 7.2.1. University of Texas at Arlington

UTA HEP group has two distinct clusters of farm machines. The HEP farm consists of 24 dual Pentium III 866 MHz machines, running under the Linux operating system. Since UTA has been one of the major offsite Monte Carlo production centers for DØ they have substantial expertise on production of simulated events and detectors. The second farm consists of six dual Pentium III 850 MHz processors. This farm is managed in collaboration with UTA computer science engineering department. This farm works as an independent entity but is controlled by the server running on the HEP farm. This farm can be used for various testing protocols for DØ Grid software development. This farm is also equipped with 100 Mbits/sec network switches which will be upgraded to gigabit switches in the spring 2002.

The group has recently added a dual Pentium processor analysis server and completed the deployment of DØ software environment. The server has a total storage space of 350GB. This machine is housed in the main office area of the Science Hall and is equipped with a 100 Mbits/sec network capability. The building is equipped throughout with 100 Mbits/sec network and will be upgraded to gigabit in the middle of year 2002.

The UTA off-campus network is equipped with one Internet II and two IT network connections, each providing up to 45Mbits/sec, totaling approximately 140Mbits/sec aggregated bandwidth.

### 7.2.2. Boston University

The BU DØ group has an analysis server which consists of eight 1.13GHz Pentium II processors and 640Gbytes of storage and is running under Linux Red Hat 7.2. This entire BU physics building is quipped with 100 Mbits/sec switches. Upgrading these to 100 Gigabit switches is being considered. At the moment, BU DØ group has access to the BU ATLAS computing facilities. The BU ATLAS group is equipped with a 10 processor Linux P-III farm, and a 128 processor Linux farm (shared). The BU campus network is equipped with OC12 (665 Mbits/sec) networking, Internet2 to Abilene and vBNS. The networking infrastructure is maintained and owned by the University. In terms of software, the BU ATLAS nodes have Globus 1.1.4, Condor 6.2, PBS and Pacman installed. Due to low usage by BU ATLAS group of these facilities they have agreed to let the BU DØ group utilize their farm for the near future. A clear path of acquisition of computing nodes for BU DØ group needs to be defined. The expertise developed in

conjunction with ATLAS group will be available to the DØ group to install and maintain an RAC.

### 7.2.3. Fermilab

Fermilab will provide processing, storage, and bandwidth to support DØ data taking, reconstruction, and Monte Carlo activities. All Raw detector data will be stored in mass storage throughout the duration of the Run IIa and IIb phases of the experiment. Output from the FNAL DØ reconstruction farm will be kept at FNAL. Final stages of MC processing output, in particular DST and thumbnail tiers, will be sent from processing centers to FNAL for permanent storage. FNAL will supply processing CPU and upgrades for Central Analysis station, DØ -Farm, and several compute server analysis stations. Disk cache will be provided to store one complete set of thumbnail data on the Central Analysis station along with substantial amounts of MC and other data samples.

Bandwidth to deliver data to RAC's will be provided, and overcapacity to allow for special processing projects requiring additional bandwidth to certain centers for particular group activities. Support for DØ code and SAM server deployment and maintenance will be provided to RAC's when needed. Global resource management that enables tuning data delivery rates to RAC's will be controlled by configuration of servers operating at FNAL. Storage robot tape mounts and tape drive usage will be monitored and allocated to provide needed file storage and delivery for all local and remote needs.

### 7.3. South America and Asia

DØ collaborators in South America and Asia are also expected to express interest in forming regional analysis centers.

## 8. Prototype Regional Analysis Center Project

It is important that many of the technical aspects of developing the RAC concept be understood as quickly as possible. While there are not anticipated to be any in-principle killer issues, confidence would be gained and interest would grow with a working site. It should also not be underestimated that there may be a significant sociological shift with the introduction of this sort of analysis structure. Again, the sooner that the collaboration confronts these issues, the sooner its members (and funding agencies) will become comfortable with this way of proceeding. Finally, there must be a significant "buy-in" from the whole collaboration to this concept *and* the particular model espoused: if only an handful of sites are developed, then the burden on them will be considerable and maybe impossible. Therefore, it is essential, again, that the RAC concept be seen to be a workable one and that the clear benefit to the experiment as a whole be obvious. When this is true, then a proper implementation can go forward and various institutions can make their plans.

For all of these reasons, it is deemed to be a useful exercise to start the development of a Prototype RAC (PRAC) very soon. The capabilities of such a site will, at first, be limited, as the goal of such an exercise should be to explore the technical, sociological, administrative, support, etc. aspects of the concept, learn from the experience, and

prepare Fermilab personnel for the organizational tasks ahead. Therefore, the initial technical implementations and gymnastics should be kept to a minimum – say, passing TMB's on a regular basis. Not only should a prototype site be formed, but potential IAC's should also be identified and be ready to demonstrate that they too are capable of making use of the services of this initial RAC. A serious project, managed in the now traditional manner, should be mounted with clear goals, milestones, and responsibilities identified. This will require identified leadership on both ends and the commitment to try something new and make it work. Ultimately, the PRAC will begin to augment its minimal capabilities, based on the accumulated experience and evolve into a real RAC. There should be an evaluation at an appropriate time of how well things worked and what needs to be done in order to take the next steps. With that experience, and hopefully fruitful use of the RAC capabilities, a schedule for deployment of other RAC's can be made.

**Conclusion 27.** A **Prototype RAC Project** should be mounted to establish a working RAC by October 1, 2002. "Working" should be minimally defined to be a) the RAC site accepting continuous TMB files; b) and identified IAC sites in a new cluster using them with relative ease to do DØ analysis.

Hopefully, this Prototype RAC can be established in Europe in order to stress the network as significantly as possible. A PRAC Team consisting of a physicist from the remote location, a systems professional from that location, a physicist from a representative IAC, and a responsible Fermilab DØ physicist should be identified to guide this project through to its final milestone this winter.

## 9. Organizational and Bureaucratic Issues

It is recognized that there is a host of issues of organization, bureaucracy, commitment, privilege, etc which will need clarification and specification before a large scale deployment of RAC's can occur. There are issues of authentication which will occur before the GRID solutions take place and also some measure of authority for remote site management will inevitably be ceded to Fermilab, as far as the interests of the experiment are concerned.

On the one hand, protection of collaborators from insecure or irresponsible behavior (or mistakes) will be required in the form of controls to protect bandwidth, CPU cycles, and/or disk space. On the other hand, since most RAC's will be sites with more than one large scale computing initiative – some with non-HEP initiatives and projects – enhancement of DØ resources through assistance across project boundaries within RAC's and IAC's may be warranted to respond to particular needs or catastrophes. (Of course, the opposite is also true, that loaning of DØ resources to partner projects within and across boundaries may also be necessary.)

Surely, Memoranda of Understanding (MOU) should be negotiated and signed among representatives of the RAC, FNAL and DØ collaboration to plan the resources and

services which will be provided by each party. The following are thought to be preliminary issues for agreement:

1. **Operating Systems:** OS upgrading should be coordinated in order to insure that at least a same version is installed on all RAC's and at CAC. Conflicts with other site responsibilities will have to be negotiated on a case-by-case basis.
2. **Storage:** The amount of storage (tapes, disks) has to be defined.
3. **Computing power:** The CPU capabilities and scheduling, should they be shared machines.
4. **Emergency response:** Agreements for emergency resource sharing should be addressed: for example, in the event of a reprocessing need by DØ, pre-arranged loans of resources should be made available for a specified period at a specified intensity.
5. Most significantly, **professional support personnel** sufficient to maintain 24/7 sites of these magnitudes.

In short, this is a very large project and will be a management nightmare if these and other issues of resources, collaboration, and competition are not thoroughly contemplated in the beginning.

## 10. Implementation Time Scale

The pressure to implement RAC sites may become significant as the luminosity increases. However, a deliberate pace in the proliferation of sites will be necessary and in that sense the Prototype RAC Project will be very instructive and act as a barrier to randomness. By establishing a working site, working out the bugs, and watching it be successful, it will be possible to implement the next steps, which will surely be more sites, with more capabilities. A medium range goal might be to be the ability to perform the  $W$  cross section measurement off site as outlined (or something functionally similar) by 2004. The steps along the way to this goal require considerable thought, but a strawman deployment might include these aspects in roughly this order:

1. Implement the Prototype RAC site and associated IAC cluster – December, 2002
2. Schedule a workshop to evaluate the Project and the model and to plan the next steps – January, 2003.
3. Establish and initiate site selection process – March-August, 2003.
4. Establish and negotiate MOU agreements with RAC institutions – March-October, 2003.
5. Full deployment – January, 2004.

## 11. Conclusions

This document has been a first effort at specifying the requirements of Regional Analysis Centers for the Run II data analysis in DØ. There is considerably more detail required

before embarking on a Global effort of this magnitude. Among the next steps, are the following issues:

- 28 “Conclusions” have been identified in order to put a spotlight on outstanding issues. They each deserve attention and are collected in the Appendix, in Section 12.1.
- A start at a single Use Case has been made, done in narrative form. Additional narratives should be developed and each of them should be deconstructed into legitimate Use Cases for consideration of the DØGRID and SAM groups.
- Only a sketch of the considerable governance and bureaucracy has been hinted at. The state of this effort within the Experiment’s structure needs consideration by the CPB, IB, and spokespeople during the development phase, which will be years. This should take into account the considerable correlations which exist among already established groups.
- A quantification of the amount of resources required has only been hinted at with an effort at bracketing the most capable (Category B) and minimally sufficient (Category D) investments.
- The funding agencies need to be made aware of the similarities and differences between RAC’s and IAC’s and the Fermilab facility. A deliberate attempt has been made to identify needs and subsequent capabilities which are different from Fermilab capabilities, such as 100% MC generation off site and disk-resident DST’s: so as to not duplicate FNAL around the world, rather to compliment it.
- Along the same lines, the contrast between this effort – which is real – and the LHC off-site computing plans – which are still virtual – requires consideration. One can clearly learn from the other, and both can benefit. This RAC effort could be a very significant component of the LHC implementation, if they are allowed to develop together and intermingle experiences and resources.

This is an exciting prospect for DØ and for HEP. Only with this capability will this experiment take advantage of all that Run II promises. Designing this correctly deserves the full attention of the collaboration.

## 12. Appendix

### 12.1. Summary of Conclusions

**Conclusion 0.** Remote analysis capability with full access to the data, code, and collaborative analysis is necessary in order to satisfy the physics goals of Run IIa and IIb. A structured environment which systematizes and standardizes these services is the best way to implement this program.

**Conclusion 1.** It is anticipated that the FNAL processing farm will be sufficient for all of Run II primary reconstruction needs. RAC's are not envisioned for *ab initio* event reconstruction.

**Conclusion 2.** RAC-centered resource management is an important goal. While initially resource management may require considerable human organization, it is desirable to augment and replace that intervention with emergent GRID tools. The priorities assigned to tool deployment remains to be worked out with sufficient Use Case analyses and some real-world experience. Accordingly, the actual capabilities of the evolving system need to be carefully planned, biased toward smooth running rather than alpha or beta testing of GRID sites.

**Conclusion 3.** Continued evaluation of the number of off-site potential users and their anticipated needs should be undertaken very soon. A preliminary census has been done. The follow-up should include more detailed scenarios and/or capabilities for a more realistic assessment.

**Conclusion 4.** A complete review should be done of the planned data tiers with special attention paid to potential off-site reconstruction opportunities with DST's and analysis opportunities with a TMB's. This should be done before deploying the DST/TMB files.

**Conclusion 5.** Generally, RAC's need not be the sole sites of code distribution to their IAC's. Rather, at least for the early days, individual installation and updating can be done directly from the Fermilab site.

**Conclusion 6.** Robust versioning and a scheme for guiding or automatically initiating stale file and directory deletions should be designed as soon as possible.

**Conclusion 7.** It may be important to precisely assess the degree of database access required for Monte Carlo production capability which includes overlaid events.

**Conclusion 8.** Early generation tools for interrogating the MC farm sites for available capability and tools for referring the actual job submission to those waiting sites are required now. Evolution of this system into a single step with full GRID deployment should be envisioned when it becomes available.

**Conclusion 9.** MC generation at RAC and/or IAC sites will be necessary in an increasingly wider scale as systematic uncertainties become a focus of measurements. Sufficient bookkeeping capabilities with the flexibility to imbed and modify the MC generation details will reduce the re-generation of already existing scenarios.

**Conclusion 10.** For the meantime, sufficient batch processing capability should be at each RAC to serve the needs of only the regional IAC's. What is uncertain is the amount of processing that this will entail and an effort to quantify this should be undertaken. Estimates will be made below.

**Conclusion 11.** In general, it seems reasonable to presume that in view of the limited computing resources available at FNAL, and of the improvements that are likely to be made to the algorithms used at present for the reconstruction of the collider data, some measure of reprocessing is expected to be an essential ingredient of the RAC's.

**Conclusion 12.** Answers to questions 1-7 need to be in hand before reaching a conclusion about how to fully characterize generic RAC's.

1. How can the DST be ensured of serving as a useful basis for reprocessing? This is a timely design issue.
2. How will DST's be distributed for this reprocessing?
3. What strategies for raw data-level reprocessing can be designed?
4. If such strategies require RAC participation, how will the raw data make their way to the RAC's?
5. For any reprocessing operation, for the first time original data will reside at a location (many locations!) away from Fermilab. Do these derived data sets get transferred back to the central Fermilab facility? Certainly, the answer is "yes" for the TMB, by design.
6. Obviously, such a scheme involves significant networking and bookkeeping resources at the RAC's and the affordability of this should be understood.
7. How can the consistency of the reprocessing be guaranteed? This is not a matter only of code distribution, but also of hardware, OS, etc.

**Conclusion 13.** Reasonable database services could be achieved with the development of proxy database servers at each RAC.

**Conclusion 14.** Full testing of the performance of the new DBS implementation should be performed at the soonest available time.

**Conclusion 15.** A test installation of the proxy server idea at a remote site should be done in the near term.

**Conclusion 16.** An evaluation of networking needs for remote analysis should be done for FNAL, U.S. RAC's, and overseas sites.

**Conclusion 17.** A robust and reliable SAM system at every worldwide site is essential. This means that adequate support for both development and operations must be provided.

**Conclusion 18.** All TMB records should be disk resident at all RAC sites twice. Total for Run IIa for TMB storage of 16TB disk per RAC.

**Conclusion 19.** Significant storage for project formats should be available at RAC's of the same order as the total TMB cache. Total for Run IIa for DERIVED storage of 16TB disk per RAC.

**Conclusion 20.** Complete DST data formats should be disk resident within the sum of the RAC sites: Total for Run IIa for DST data storage of 24TB disk per RAC, which presumes 10% of the total at each RAC site.

**Conclusion 21.** MC storage for per-demand generated events is primarily limited to the ROOTuple needs, with a nominal MCDST compliment as well: Total for Run IIa MC ROOTuple data storage of ~5TB disk per RAC and 10TB tape per RAC; MCDST disk storage of approximately 50TB, presumes 5% of the dataset on disk.

**Conclusion 22.** Temporary storage needs for staging of data and Monte Carlo analysis and ROOTuple generation are estimated to be 10% of the total of each data format: Total for Run IIa for temporary cache of ~11TB disk per RAC. A more accurate estimation of this need is required.

**Conclusion 23.** Storage needs for serving db setups at RAC will not likely exceed a few TB for all data taking. Total for Run IIa for database/SAM needs of 1TB.

**Conclusion 24.** Batch resources per Category B RAC should be on the order of 50 nodes per year of modern PC nodes running Linux.

**Conclusion 25.** Planning should begin early for constructing and maintaining a sophisticated electronic helpdesk and FAQ for DØ software installation, implementation, and use issues. Triage strategies should be planned.

**Conclusion 26.** It might be useful to get a non-binding expression of interest from potential institutions just to see what the maximum might be, and to determine whether interest is not sufficient to support the concept. Too few sites will make the burden on that small set perhaps too significant for their viability.

**Conclusion 27.** A **Prototype RAC Project** should be mounted to establish a working RAC by October 1, 2002. "Working" should be minimally defined to be a) the RAC site accepting continuous TMB files; b) and identified IAC sites in a new cluster using them with relative ease to do DØ analysis.

## Bibliography

---

- 1 U. Heintz, et al., “Data Tier Proposal”, un-numbered DØ Note, April 21, 2000.
- 2 U. Heintz, “Proposal for Thumbnail Contents, for the Data Tier Committee”, version3, un-numbered DØ Note, December 6, 2000.
- 3 Other documents on data tiers can be found at <http://www-d0.fnal.gov/~serban/>
- 4 Proceedings of the workshop can be found at:  
[http://www-hep.uta.edu/~d0race/workshop\\_feb\\_2002/d0race-workshop-program-feb-11.html](http://www-hep.uta.edu/~d0race/workshop_feb_2002/d0race-workshop-program-feb-11.html)
- 5 A. Boehnlein, et al. (for CPB), “DØ Computing and Software Operations and Upgrade Plan”, in preparation, May, 2002.