

## Flows and Controls in the Seasons of SAM

Mike Diesburg, Lee Lueking, Kurt Ruthmansdorfer, Heidi Schellman, Vicky White

December 4, 1997

### **Abstract**

The Sequential Access Model is described in terms of the data flows, control mechanisms, and bookkeeping. Data Flows are modeled with a spreadsheet which explores the data rates through the different parts of the system for usage patterns corresponding to five phases of the Run II data taking period.

# Contents

<b>1 Introduction</b>	<b>2</b>
<b>2 Flow, Control and Bookkeeping</b>	<b>2</b>
<b>3 The Five Seasons</b>	<b>5</b>
<b>4 The Layout of the Simulation spreadsheet</b>	<b>5</b>
<b>5 Evolution of each Access Pipeline</b>	<b>9</b>
<b>6 Summary</b>	<b>10</b>

## 1 Introduction

This document describes details of the Sequential Access Model (SAM) including data flows, controls and bookkeeping. In order to understand the magnitude of the data flows for each part of the system a spreadsheet analysis has been assembled which attempts to predict access patterns and bandwidth needed. Five time periods during the experimental operation are chosen which exemplify distinct usage characteristics for the system. Numbers are developed for aggregate bandwidths and total size required for each of the main data repositories. It is assumed that a 1 GB file is employed for most data, although some analysis caches have other sizes used to help optimize the system. In order to understand the various bandwidth numbers, the access of primary data (PD) is broken down into the 5 pipelines defined in the SAM description document (1) Metadata, (2) Thumbnail, (3) Random event picking, (4) On-demand files and (5) Freight train data serving. In addition, to complete the picture, the input to and output from from the farm processing and the storage of user derived data (DD) are also estimated.

## 2 Flow, Control and Bookkeeping

(This section is just being started)

The flow of data, control for the system, and collection of bookkeeping information is shown in Figures 1,2,3 respectively.

The data is stored in eight repositories based on functionality. Some of these may, in fact, share the same physical robot or disk storage areas. The arrows in Figure 1 indicate the flows, and the ovals represent processes which access or produce data.

A rough picture of the control paths needed for the SAM system is represented in Figure 2. The nature of the control is quite different for the different pipelines and these details are yet to be worked out. Files on-demand are completely user driven with the requesting user managing local disk buffer and CPU usage for his processing. Farm processing and Freight Train data delivery are under a centralized control, and the structure for this is not shown but will need to be worked out. In these two cases a type of process control will need to be implemented. There are subtle differences between the process control needed for the farm processing and the freight train activities, nevertheless it is hoped they can share common control software. The control for pick events differs from the other modes since it deals with data at the event level, as opposed to the file level, it therefore needs some additional machinery.

The bookkeeping server structure is shown in Figure 3. The plan is to provide a server for each type of pipeline or processing activity. These servers will provide bi-directional communication

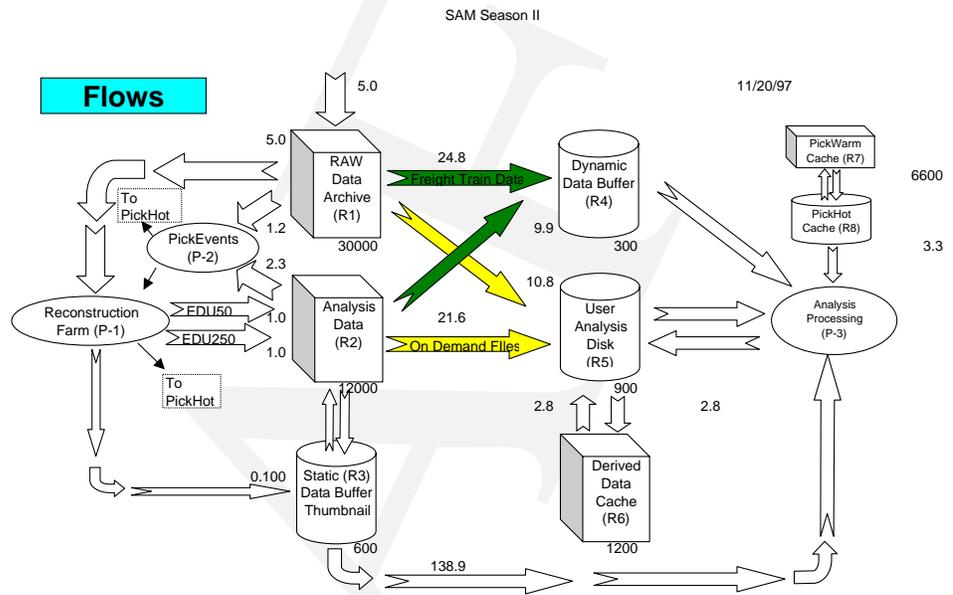


Figure 1: Data flow chart for the Sequential Access Model.

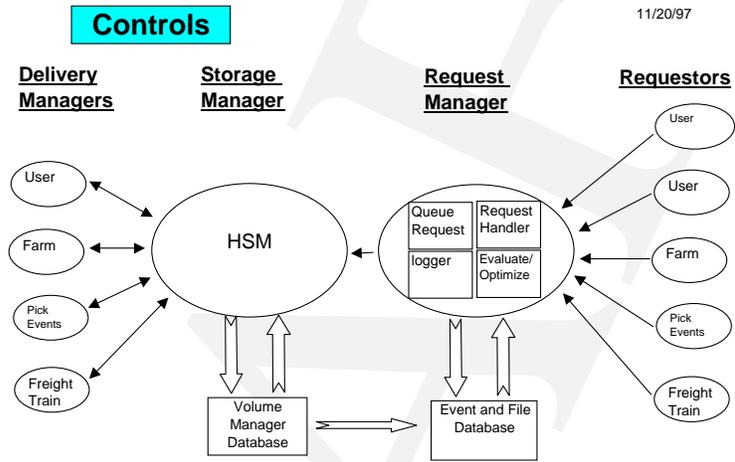


Figure 2: Data access control for the Sequential Access Model.

to a central DB server which talks with the DB system. Obviously, these servers need to be extremely reliable if we are to depend on the DB information for tracking all phases of the data processing and access. Figure 3 shows how information is transmitted from the various clients processing data to and from the central event and file database.

### 3 The Five Seasons

We assume that the patterns of accessing the data in Run II can be generalized for five basic phases, or seasons, of the experiment:

1. **Commissioning:** Before collider data is actually taken, there will be several months of detector and system troubleshooting. There will also be Monte Carlo data floating around used to debug the off-line reconstruction. The amount of data recorded during this period will not be large, and its use is somewhat unpredictable. However, this can be used as a test period for the data management system.
2. **Early Data Processing:** This is the most chaotic of all data handling periods. The data being taken is of inconsistent quality and the off-line processing extremely immature. Some subsets of this data are reconstructed many times. Data selection strategies are put into place and many need to be modified or re-executed. Much of the emphasis is on the RAW data. Integrated luminosity is low.
3. **Mid-term:** Around the middle of the running period, the reconstruction algorithm begins to stabilize, with new versions needed only every month or two. Much of the early data is reprocessed to provide complete and consistent data sets for physics analysis. The accelerator luminosity reaches new highs. Individual events are selected from raw data and cached at about the 10% level.
4. **Late-term Steady-state:** By the last third to quarter of the run, the inertia for changing the reconstruction program will become very large as data accumulates quickly enthusiasm for change fades. The experiment enters a steady state, and the chaos is at a low. Only partial processing of data, or fixing, is attempted due to the long lead-times caused by I/O and processing overheads. Raw events continue to be cached at the 10% level. Record luminosities are recorded.
5. **Post-run:** Some processing is done after the data taking period ends. No new data is added to the input repository. The caches are built and access to the raw data diminishes rapidly.

### 4 The Layout of the Simulation spreadsheet

The work is done in Excel spreadsheets and attached in the Appendix for Seasons II and IV (the other experiment phases will be filled in soon, when these are more fully understood). The pipelines, each with several modes of use, are shown as the row names and the columns represent the parameters for each pipeline sub-category. There are several EDU types shown characterized by their level of processing and event size. The dominate types are RAW250, EDU250, EDU50 and Thumbnail (or EDU5). These sets of data will be accessed through the pipelines with varying degrees throughout the 5 seasons. In an attempt to reduce the number of parameters which are needed to establish the tables for each period, several factors are introduced and are included as the first line, above the column headings, in each season's table. These parameters are:

**Bookkeeping**

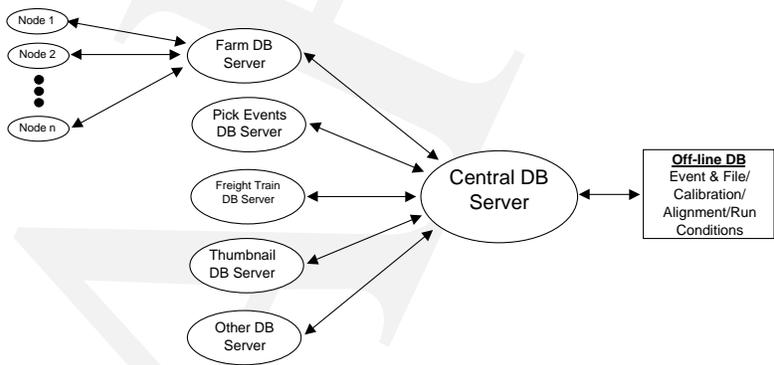


Figure 3: Bookkeeping diagram for the Sequential Access Model.

1. **Total Events:** The total number of RAW data events for the entire two year run.
2. **RAW fraction:** The RAW fraction is simply the fraction of RAW data which is accumulated for the particular season, normalized to the total RAW data expected for the entire run.
3. **Chaos factor:** Chaos is a multiplicative factor representing the fact that the early data is examined and processed many times while the detector and off-line processing are being debugged. This makes a small data set effectively look much larger in terms of bandwidth.
4. **Effective Events:** This is simply the first three parameters multiplied together and estimate how the data and processing fill the system.
5. **DAQ Rate:** The event rate at which data is coming from the Data Acquisition System.
6. **Pick Fraction:** The fraction of events which are being selected to be extracted by the pick facility.
7. **Pick Hot Days:** The number of days picked events will be maintained in the hot disk cache. This will determine the size of disk storage needed for this cache.

There are eight basic storage locations for data called *repositories* and three process categories. The “Flows” diagram associated with each season shows how these are related and give an overview of the system with the repositories shown as R1 through R8 and the processes P-1 through P-3 . The locations are defined as follows:

1. **RAW Data Archive (R1):** The location of RAW data from the DAQ.
2. **Analysis Data (R2):** Location of the Reconstruction Farm output.
3. **Static Data buffer (R3):** Disk storage for the thumbnail data which will be under central administration. Its contents will be changed only when a new thumbnail event is defined, which may be on the order of weeks to months.
4. **Dynamic Data Buffer (R4):** Disk storage for Freight Train analysis. Data on this disk will be updated frequently, on the order of minutes to hours.
5. **User Analysis Disk (R5):** Storage for user and group analysis data. Controlled and updated by users.
6. **Derived Data Cache (R6):** Data which is derived from all sources of primary data analysis (except event picking) which users or groups need to store and retrieve.
7. **Pick Warm Cache (R7):** Picked events which are concatenated into small files ( 100 MB) and stored on tape.
8. **Pick Hot Cache (R8):** Most recently picked events each stored in separate file in this disk cache. This area will be large enough to store events selected over a one or two month period.

The processes are defined as follows:

1. **Reconstruction Farm (P-1):** The reconstruction farm receives RAW data from the RAW Data Repository (R1) and produces three output flows: (1) EDU5 (thumbnail), (2) EDU50, and (3) EDU250. The thumbnail data flows immediately to the Static data buffer disk, and the other outputs are placed in the Analysis data repository (R2).

2. **Pick Events (P-2):** The pick events process is sent data which it sifts through extracting single events for which it is looking. The input for this process could be any of the repositories where data is stored, but in practice would most likely be R1, R2, or R7. The input could simply be a byte-stream of data coming through the HSM directly from a tape drive into the pick events socket. When the event or events have been found, the stream could be switched off, avoiding wasted bandwidth and processing. The output from the pick process would be single events each in their own file, sent to the Pick Warm Cache (R8).
3. **Analysis Processing (P-3):** The analysis processing consists of any of a number of analysis-type processes. These can be either individual user tasks or coordinated group activities. They would run on “centralized” CPU machines closely networked to the storage.

The spreadsheet is made of 10 input, and 16 output columns. The ideas contained in each input column are as follows:

1. **Mode of Access:** Various modes of access for each of the major pipelines, reconstruction farm, and derived data.
2. **Concurrent Access :** The number of individual accesses in the elapsed time given in the next column.
3. **Elapsed Time in Days:** The elapsed time to provide the accesses indicated in column 1. In some cases, this assumes 24 hr/day processing for several days or weeks, e.g. in Freight Train processing of an entire primary data stream. In other cases, e.g. On-demand file-set, these accesses are assumed for only part of each day as peak usage may diminish at night or on weekends. On-demand file access, meaning single files, are needed promptly when requested and it is assumed that no one will have to wait more than 20 minutes for one of these. These areas are a first attempt to address the load-leveled and peak-demand needs of the system.
4. **Elapsed Time in Hours:** Same as column 2, but in hours.
5. **Data Accessed in GB:** The total amount of data read in to satisfy the need.
6. **Data Delivered in GB:** The actual amount of data delivered. This may be different from the data accessed (column 4) in some modes, e.g. pick events, where an entire file may be read in, but only a single event delivered.
7. **Size per file in GB:** The size of each file in which data is stored.
8. **Size of EDU in GB:** The size of each event for the particular Event Data Unit.
9. **Source Repository index:** The repository from where the data is accessed.
10. **Target Repository index:** The repository to where the data is delivered.
11. **Next 8 Columns; Repository 1-8 Bandwidth in MBps:** Bandwidths into and out of each repository.
12. **Final 8 Columns; Repository 1-8 Space in GB:** Space required in each repository.

In some cases, the model assumes 24 hr/day processing for several days or weeks, e.g. in Freight Train processing of an entire primary data stream. In other cases, e.g. On-demand file-set, these accesses are assumed for only part of each day as peak usage may diminish at night or on weekends. On-demand file access, meaning single files, are needed promptly when requested and it is assumed that no one will have to wait more than 20 minutes for one of these. These areas are a first attempt to address the load-leveled and peak-demand needs of the system.

## 5 Evolution of each Access Pipeline

As described in the SAM document, we have identified 5 major access pipe-lines used for access:

1. **Metadata:** This access is to the 50 to 100 Bytes of information stored for each event which allows the data to be accessed. This includes trigger and other data on a per event basis, as well as file and processing information.
2. **Thumbnail:** This is the 5 to 10 kB/event of information stored in on-line disks. In seasons I and II, this disk storage will have a flexible use for RAW, EDU250 and EDU50 used for troubleshooting the system, there is a chaos factor included at these stages since it is not exactly how the access will appear. Rapid turn around is important and in season II, as many as 30 people may expect to read through this data within a period three days. Later, around season III, this data will become more stable and the form of the thumbnail will be understood, but it will still be changing frequently and users will desire to go through the entire set in about 2 days. It will be updated often and accessed frequently. As the data size grows, season IV, it may take as long as a week to go through the entire set. After the run is over, the number of concurrent accesses should go down to just a few.

The exact nature of these accesses depends on the physical layout of the data on the disks. The numbers in the spreadsheets assume that the entire data set is read, however if the data is laid out in a more efficient way this could be reduced by a factor of 10 or more.

3. **Pick Events:** This data consists of events chosen by users based on thumbnail or other analysis experience. Caching selected events makes them more accessible for subsequent requests and there are three levels of caching: (1) hot cache - one event per file on disk, (2) warm cache - 500 to 1000 events per file stored in the analysis robot, and (3) cold cache - uncached events in their native files as they are stored in RAW or processed states. It is assumed that about 10% of all raw events will migrate into warm cache. A simple calculation assuming  $6E8$  events per year ( $5E7$  events per month), gives about  $1.6E6$  events per day total. One-tenth of this is 160k events picked each day. If data is coming in at 50 Hz, this may be a factor of two higher, or 320k events picked per day. However, the total number of files per day will only be around 800 (assuming 4k events / file), so this would indicate that there will be between 800 and 320k files needed per day from the raw repository depending on whether the selection is coordinated or completely chaotic.

One factor which affects the assumptions enormously in this area is the distribution of size for the various trigger streams. The largest anticipated streams (jet and possibly low pt B) will make up more than half of all the data and these data may require a more dedicated, less random, method of access e.g. the freight train server. The larger number (320k) of files per day does not account for pooling requests or for data in warm cache and is thus a complete over estimate. The number used is 1000 events picked per day, which seems to be a reasonable guess for the middle seasons of the run. It is not yet clear how many tape mounts this would imply, but this number sets a sort of upper limit. The number is lower

Season	Aggregate Bandwidths (MBps)							
	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5	Rep 6	Rep 7	Rep 8
I.Commissioning								
II.Early Data	41	36	139	35	38	5.5	0.35	1.5
III.Mid Data								
IV.Late Data	78	106	197	108	38	5.5	3.4	15
V.Post Run								

Table 1: Aggregate Bandwidths needed for all data access modes at various phases of the experiment into and out of the eight repositories of the model.

for seasons I and II, when much of the RAW data will be in warm cache or on disk, and will be much lower in season V when new data stops coming in.

For randomly picking events, obviously the optimum file size would approach one event per file. The optimum file size can assume an equal time for the load-mount-seek, and the actual file read. Also, on average half of the file read time can be saved by only reading the part of the file up to where the event is found. So, for example, if it requires 10 s to get the tape ready for reading, and one can read at 10 MBps, this criteria would indicate that the file size be around 20 MB. This may be the right formula to use for determining the size of the files in the warm pick cache.

4. **Files on-demand:** These are files requested by users for processing. There are two categories: (1) single files, and (2) file-sets, or groups of 10 or so files which are processed together. The single on-demand files are needed to be staged to the user's disk within 20 minutes during an 18 hour part of the day from 6 am to mid-night. The file-sets are staged in a just-in-time fashion so that batch processing can proceed efficiently, there are no time constraints imposed and they are lower priority than the single file requests.
5. **Freight train:** Much of the work of large scale access to EDU50 and EDU250 will be provided through this facility. For this reason, the event picking for these data types may be overestimated. It is assumed that throughout the run, two or three of these projects will be running in parallel and going through all of the accumulated data for a particular stream in a two week time frame.

## 6 Summary

A summary of the aggregate bandwidths and repository sizes to grow the system are shown for the five seasons of SAM in Tables 1 and 2.

## References

- [1] SAM document

Season	Storage Size (TB)							
	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5	Rep 6	Rep 7	Rep 8
I.Commissioning								
II.Early Data	30	12	0.6	0.3	0.9	1.2	6.6	0.03
III.Mid Data								
IV.Late Data	297	119	5.9	0.3	6.2	12	65	0.03
V.Post Run								

Table 2: Aggregate storage space needed for all data access modes at various phases of the experiment in the eight repositories of the model.