CLAS-NOTE 1999-017

CLAS Future Computing Needs

Dieter Cords, Mark Ito, Bernhard Mecking,
Stepan Stepanyan, Dennis Weygand

May 28, 1999

1 Introduction

This document presents an estimate of the computing resource requirements for the CLAS Collaboration for the next few years. Estimating these resource needs is hard to do accurately; estimates are often proven wrong either by unforeseen advances in computing technology or improved ideas for analysis techniques. We are presenting our best guesses based on our year of experience analyzing data from CLAS.

2 Computing Model for CLAS

As a basis for estimating computing resource needs for CLAS we present a model of the analysis process. What follows is intended to be representative of how the bulk of CLAS data will be analyzed over the next few years. The principal activities are taking the data, performing calibration passes on a subset of raw data, first pass processing of the raw data, separation into analysis streams and production of data summary tapes (DST's), generation and analysis of Monte Carlo data, and detailed analysis of particular physics channels. As such these steps are typical of any large nuclear or particle physics experiment. In this section we quantify the parameters of each step which are relevant for our resource needs estimate.

2.1 Taking the data

During normal data taking, events are obtained at a rate of 2500 per second. The average event size is typically 4 kilobytes. For this estimate we will assume ten months of running per year and that the CLAS data acquisition is actually taking events 50% of the time during these ten months (our typical running efficiency).

2.2 Calibration passes

The first analysis step after taking the data is to generate calibration constants. To do this we first run a preliminary reconstruction pass on the raw data, and then run a variety of calibration programs on the resulting reconstructed data. Typically, we can calibrate our detectors by reconstructing about 5% of the raw data. This preliminary reconstruction is usually repeated, as problems are found and improvements are made, so we assume here that it will be done twice for each set of raw data.

2.3 First pass processing

After calibration constants are obtained we run our reconstruction program on the complete raw data set (Pass 1). The reconstruction program takes about 50 milliseconds per input event on one of our 300 MHz single-processor Linux farm nodes (one CPU). Again, we have found that we usually repeat this process. For this estimate we will assume that this step is done twice.

Although the reconstruction eliminates events that fail to reconstruct, it also adds information to the successful one. The net effect has been neutral; the size output data sets are approximately the same as the input data sets.

2.4 Analysis streams and data summary tapes

At various points in the analysis, events will be separated into several data streams based on various event classification criteria. These streams will represent a duplication of some of our processed data. They may or may not contain a subset (or superset) of information from the processed data. They will also be generated at various stages of the analysis, some of the streams being derived from other streams. Data summary tapes are one example of such a stream. For these streams, as the name implies, the information retained for each event will be substantially less than that for the raw data.

For this estimate we assume that the additional CPU power needed to create these output streams is 50% of that needed to do the Pass 1 processing once.

We also assume that a complete traversal of the processed data set will be necessary three times and that the output data size each time is 50% of the input data size. We hope that some streams can be generated in the Pass 1 processing; the additional tape reading referred to here is in addition to any Pass-1-generated streams.

2.5 Generation and analysis of Monte Carlo data

Monte Carlo data sets are an important component of most physics analyses. The number of events, size of events and CPU time required to generate and analyze each set will vary widely, depending on the analysis.

We estimate that each Monte Carlo event takes one second to generate and analyze that the size of each event is the same as that of the raw data, and that the total number of Monte Carlo events needed will be 20% of the number of raw data events.

2.6 Analysis of particular physics channels

The final jobs required to yield numbers that go directly into final physics results will also have huge variation in their demand for computer resources.

We estimate that the total CPU resources for doing physics analysis is comparable to one iteration of the Pass 1 processing.

3 CPU Requirements

CPU usage is probably the easiest category in which to generate estimates. To keep up with one iteration of the Pass 1 processing requires 52 CPU's. If we consider the estimated need for repeated processing for Pass 1, as well as completing calibration reconstruction (with repeats), separation into streams, and all Monte Carlo processing we end up with a need for 450 CPU's to keep up with the incoming data rate. Monte Carlo processing accounts for 210 of these CPU's. We plan to off-load Monte Carlo processing to farms being developed by collaborating institutions. With this assumption, the JLab-resident CPU load is 240 CPU's.

4 Tape Silo Requirements

A look at our current silo configuration give us the following parameters: 30 seconds to robotically load a tape, 10 megabyte/second speed from tape to cache disk per drive, 3 megabyte/second from cache disk to farm node per node. For a single request, in the absence of other requests, the bottleneck is the cache-to-farm transfer, and for our canonical 2 gigabyte data files, the total transfer time is about 15 minutes. In general, there are many such requests working in parallel; 3.3 requests saturate a single drive.

For this estimate, we assume that each tape drive in use gives the full 10 megabyte/second transfer rate. We ignore the robot tape loading time. Further we assume that the cache-to-farm transfers can be parallelized so that they are not a limiting factor. Then in steady-state, Pass 0 requires 0.1 drives, Pass 1 uses 1.7 drives, streaming uses 1.4 drives, and copying Monte Carlo output back to the silo uses 0.8 drives. This includes repetition factors and copying output from analysis passes and stream creation back to the silo. For other analysis tasks we assume 200 requests per day for 2 gigabyte files. This will occupy another 0.5 drives. The total number of drives needed is 4.3.

5 Disk Resources

For this discussion, we will only include disk resources needed in addition to those already (or soon to be) allocated to CLAS. These include the CLAS work disks, the CLAS cache disks and the personal account disk usage of CLAS collaborators. The assumption is that this amount will be adequate for the work of individual collaborators to proceed.

Turning to additional disk requirements, our effort would be made much more efficient if we had a quasi-permanent set of data available on disk, bypassing the need for access the silo. This would be most useful for the "interactive" component of our analysis effort. One way of estimating this need is to assume that each of the CLAS run groups will want to store about 250 gigabytes of data on disk. We would most likely let each run group decide on the form and mix of data that should be available on disk. There are eight run groups in the CLAS Collaboration so we would estimate a need for an additional 2 terabytes of disk space.

6 Non-JLab Institutional Issues

6.1 Data Export/Import

Collaborators resident at non-JLab institutions have a keen interest in doing some part of their analyses at their home institutions. Several have or will soon have CPU resources that make this feasible, mainly in the form of PC-compatible Linux farms. The major obstacle they face is getting the data from JLab. There are two obvious options:

Ship the Tapes The tried and true UPS method has the advantage of very high bandwidth and a relatively reliable connection. However, the logistical problems are numerous. Tape handling is labor intensive, tapes introduce a storage, retrieval and inventory problem, and tape format has to be compatible between Jlab and the collaborating institution.

FTP the Files Transferring over the network sidesteps a lot of these problems. For example, there is no problem with hardware compatibility since non-JLab institutions can store the data in the form of their choice. However, the bandwidth requirement is steep. A calendar year's worth of summary data is about 13 terabytes assuming a compression factor of 10 over the raw data. This would take about four months to transfer over a dedicated 100 megabit line.

Our current thinking is that the data analysis tasks that involve large volumes of input and output data will be done at JLab, using the tape silo. We do not plan to request additional resources to facilitate large-scale tape copying.

6.2 Contribution of CPU Resources

We note that the combined CPU resources of all outside institutions is equivalent to some substantial fraction of the JLab CPU resources, and using them will be necessary to meet our needs. As mentioned above, Monte Carlo generation and analysis is a natural way for us to take advantage of these resources, since this activity will use tape resources less intensively than analysis of real CLAS data.

This document is available in HTML form from
http://claspc2.cebaf.gov/~marki/docs/offline/comp_res/
and in postscript form from
http://claspc2.cebaf.gov/~marki/docs/offline/comp_res.ps

`$Id: comp_res.tex,v 1.22 1999/10/21 21:50:44 marki Exp $`

CLAS computing resources		
parameter	value	units
event rate	2500.00	events/s
annual running	10.00	months
running efficiency	0.50	
effective event rate	1.04E+03	events/s
	3.29E+10	events/year
CPU time per event	0.05	CPU-s/event
single Pass 1 CPU need	52.08	CPU's
raw event size	4000.00	bytes
raw effective storage rate	1.31E+14	bytes/year
	131.49	TB/year
Pass 0 event fraction	0.05	
Pass 1 repetition factor	2.00	
Pass 0 repetition factor	2.00	
Pass 0 CPU need	5.21	CPU's
Pass 1 CPU need	104.17	CPU's
Stream/Pass 1 CPU ratio	0.50	
Stream multiplicity factor	3.00	
single stream CPU need	26.04	CPU's
Stream CPU need	78.13	CPU's
MC CPU time per event	1.00	CPU-s/event
MC/raw data event rate ratio	0.20	
MC event size	4000.00	bytes
MC CPU need	208.33	CPU's
Analysis/Pass 1 CPU ratio	1.00	
Analysis CPU need	52.08	CPU's
Total CPU need	447.92	CPU's
Total CPU need exclusive of MC	239.58	CPU's
data rate, tape to cache disk	10.00	MB/s
data rate, cache disk to local disk	3.00	MB/s
Pass 1 processing disk read rate	80000.00	bytes/s/CPU
Pass 1 output to input size ratio	1.00	
Pass 1 processed event size	4000.00	bytes
single pass 1 tape need	0.83	drives
Pass 1 tape need	1.67	drives
Pass 0 tape need	0.08	drives
Stream output to input size ratio	0.50	
Stream processing disk read rate	160000.00	bytes/s/CPU
single stream tape need	0.42	drives
Stream tape need	1.25	drives
Run group count	8.00	groups
Disk usage per run group	250.00	GB
Total additional disk usage	2.00	TB
example data compression factor	10.00	
Ethernet transfer speed	100.00	Mb/s
example data time	12.00	months
example time of transfer	4.00	months

Figure 1: Parameters used in this estimate.