1 Computing Requirements

In this section we present an estimate of the computing resources required by the CLAS Collaboration to acquire, reconstruct, simulate and analyze the CLAS12 data in a timely fashion. We assume that tasks like reconstruction and simulation will keep pace with the data acquisition after the start of data taking for CLAS12 in 2015. The computing enterprise for CLAS12 is divided into stages: data acquisition, calibration, reconstruction, simulation, reconstruction studies and physics analysis. The first four stages represent the process of taking the raw data and turning it into 4-momenta and identified particles. The reconstruction studies are needed to optimize the reconstruction of the data and the simulation while the physics analysis stage represents a broad range of activities using the results of the final event sample (measured and simulated).

We now discuss the computing requirements for the data acquisition and focus on the number of computing cores, disk space, and tape storage. Table 1 shows the assumptions that go into our calculations. Using the data in Table 1 we calculate the data rate, the number of events collected

Event rate	10 kHz	Weeks running	35
Event size	10 kBytes	24 hour duty factor	60%

Table 1: Data Acquisition parameters.

in a year of running, and the volume of that data.

$$\text{Data Rate} = \text{Event Rate} \times \text{Event Size} = 100 \text{ MByte/s} \tag{1}$$

$$\text{Average 24-hour rate} = \text{Data Rate} \times \text{24 hour duty factor} = 60 \text{ MByte/s} \tag{2}$$

$$\begin{aligned} \text{Events/year} &= \text{Event Rate} \times \text{Weeks Running} \times \text{24 hour duty factor} \\ &= 1.3 \times 10^{11} \text{ Events/yr} \end{aligned} \tag{3}$$

$$\text{Data Volume/year} = \text{Events/year} \times \text{Event size} = 1270 \text{ TByte/yr} \tag{4}$$

These results will be used in the calculations below.

We next consider the resources we will need to calibrate our data and keep pace with the incoming data in CLAS12. We have determined the CPU-time to reconstruct an event from previous work with the CLAS12 physics-based simulation *gemc* and the CLAS12 reconstruction code *SOT*. We found that 155 ms is long enough to reconstruct most of the events that CLAS12 will collect. That result and other assumptions are in Table 2. Calculations of the necessary CPU time in

CPU-data-time/event	155 ms	Data fraction	5%
Data passes	5		

Table 2: Calibration parameters.

seconds (CPU-s) for calibration for one year and the number of cores required to keep pace with

the data flow follow.

$$\begin{aligned}\text{CPU time/year} &= \text{Events/year} \times \text{CPU-data-time/event} \times \\ &\quad \text{Data fraction used} \times \text{Data passes} \\ &= 4.9 \times 10^9 \text{CPU-s/year}\end{aligned} \tag{5}$$

$$\text{Cores} = \frac{\text{CPU time/year}}{T_{yr}} = 173 \text{ cores}$$

The quantity T_{yr} is the number of seconds in one year multiplied by the efficiency of an individual core in the JLab computing farm which is close to 90%.

Reconstruction of the CLAS12 data (known as cooking) will be the second-most computer intensive task behind simulation (see below). The time required to reconstruct an event using the CLAS12 reconstruction code *SOT* is mentioned above. The other parameters have been estimated based on experience with the CLAS6 detector and are listed in Table 3. We first estimate the

CPU-data-time/event	155 ms	Output size/input size	2
Data passes	2	Output fraction on work disk	10%

Table 3: Reconstruction parameters.

CPU time (in CPU-s) required for the reconstruction to keep pace with the incoming data and use this to determine the number of cores needed. We also estimate the disk and tape storage and the average bandwidth needed to move these data since these tasks will require considerable resources. The calculation of the bandwidth includes time for reading in each event and writing out the reconstruction results.

$$\begin{aligned}\text{CPU time per year} &= \text{Events/year} \times \text{CPU-data-time/event} \\ &\quad \times \text{Data passes} \\ &= 3.9 \times 10^{10} \text{ CPU-s/year}\end{aligned} \tag{6}$$

$$\text{Dedicated farm cores} = \frac{\text{CPU time per year}}{T_{yr}} = 1387 \text{ cores} \tag{7}$$

$$\begin{aligned}\text{Cooked data to tape} &= \text{Data Volume/year} \times \text{Data passes} \\ &\quad \times \text{Output size/input size} \\ &= 5080 \text{ TByte/yr}\end{aligned} \tag{8}$$

$$\text{Disk storage} = \frac{\text{Cooked data to tape}}{10} = 508 \text{ TByte}$$

$$\begin{aligned}\text{Average bandwidth} &= \text{Event size} \times (1 + \text{Output size/input size}) \times \\ &\quad \frac{\text{Dedicated farm cores}}{\text{CPU-data-time/event}} \\ &= 268 \text{ MBytes/s}\end{aligned} \tag{9}$$

Simulation of the CLAS12 response will be an essential part of the reconstruction and analysis because the precision of many experiments will not be limited by statistical uncertainties, but by systematic ones. Understanding the detector is necessary to distinguish physics effects from possible experimental artifacts. Table 4 shows the parameters used in the calculations that follow. The CPU-sim-time/event is the time required to simulate an event using the CLAS12, physics-based simulation *gemc* and to reconstruct it with the CLAS12 reconstruction package *SOT* on a single core. To estimate the number of simulated events we need in a year we have studied the properties

CPU-sim-time/event	485 ms	Fraction to disk	2%
Sim-events/year	3.2×10^{11}	Fraction to tape	10%
Output event size	50 kBytes	Multiplicity	1.5

Table 4: Simulation parameters.

of the planned trigger in CLAS12 and the backgrounds associated with those events. We find that of the 1.3×10^{11} events we expect to collect in one year with CLAS12 (see Equation 3) about half will have a good electron that will be reconstructed. Of that sample we expect about half will be background leaving us with about one-fourth of the event rate as good physics events. In order to adequately simulate the properties of this sample we need about ten times as many simulated events so the statistical uncertainty on the simulated events will be much less (about one-third) than the statistical uncertainty of the data. This number, Sim-events/year, is the number of events for a single, high-statistics simulation of the final physics sample and is listed in Table 4. The multiplicity factor in Table 4 is included to account for computer time to optimize the simulation and to study systematic effects, *e.g.* comparing high-statistics simulations using different event generators. The results of the calculations follow.

$$\begin{aligned} \text{CPU-time/year} &= \text{CPU-sim-time/event} \times \text{Sim-events/year} \times \text{Multiplicity} \\ &= 2.3 \times 10^{11}\ \text{CPU-s/year} \end{aligned} \tag{10}$$

$$\text{Dedicated farm cores} = \frac{\text{CPU-time/year}}{T_{yr}} = 8,139\ \text{cores} \tag{11}$$

The simulation of the CLAS12 is resource intensive so we also considered the volume of disk and tape storage required and the bandwidth necessary for transporting the data.

$$\begin{aligned} \text{Work disk} &= \text{Sim-events/year} \times \text{Output event size} \times \\ &\quad \text{Fraction to disk} \\ &= 318\ \text{TBytes} \end{aligned} \tag{12}$$

$$\begin{aligned} \text{Tape storage} &= \text{Events/year} \times \text{Output event size} \times \\ &\quad \text{Fraction to tape} \\ &= 1,588\ \text{TBytes/year} \end{aligned} \tag{13}$$

$$\begin{aligned} \text{Average bandwidth} &= \frac{\text{Output event size} \times \text{Dedicated farm cores}}{\text{CPU-sim-time/event}} \\ &= 839\ \text{MByte/s} \end{aligned} \tag{14}$$

We also expect that in addition to reconstruction and simulation considerable computing resources will be devoted to optimizing the reconstruction of the physics data and the simulated events for particular analysis projects. This task may require studying the reconstruction for a subset of the data (*i.e.*, skim files). The assumptions we make for this part of the CLAS12 computing are shown in Table 5. The calculations of the number of cores necessary to keep pace with

CPU-data-time/event	155 ms	Fraction of desired events	5%
Data passes	10		

Table 5: Reconstruction studies parameters.

the CLAS12 data acquisition follow. The disk storage and average bandwidth were calculated in the same manner as the previous ones and we found requirements of 508 TBytes for disk and 78 MByte/s for bandwidth. The amount of data archived to tape we expect to be small compared to our other requirements.

$$\begin{aligned}\text{CPU time per year} &= \text{Fraction desired} \times (\text{Events/year} + \text{Sim-events/year}) \times \\ &\quad \text{Data passes} \times \text{CPU-data-time/event} \\ &= 3.4 \times 10^{10}\ \text{CPU-s/year}\end{aligned} \tag{15}$$

$$\text{Dedicated farm cores} = \frac{\text{CPU time per year}}{T_{yr}} = 1214\ \text{cores} \tag{16}$$

Once the reconstruction has been optimized for a particular analysis project, we expect there will be considerable computing resources devoted to analyzing the results. These data will not require a full reconstruction so the compute time per event will drop considerably, but most of the data set will usually be studied. The assumptions we make for this part of the CLAS12 computing enterprise are shown in Table 6. The calculations of the number of cores necessary to keep pace

CPU-analysis-time/event	8 ms	Fraction of desired events	50%
Data passes	10		

Table 6: Physics Analysis parameters.

with the CLAS12 data acquisition follow. The disk storage and average bandwidth were calculated in the same manner as the previous ones and we found requirements of 889 TBytes for disk and 279 MByte/s for bandwidth. We expect the amount of data archived to tape to be small.

$$\begin{aligned}\text{CPU time per year} &= \text{Fraction desired} \times (\text{Events/year} + \text{Sim-events/year}) \times \text{Data passes} \times \\ &\quad \text{Data passes} \times \text{CPU-analysis-time/event} \\ &= 1.7 \times 10^{10}\ \text{CPU-s/year}\end{aligned} \tag{17}$$

$$\text{Dedicated farm cores} = \frac{\text{CPU time per year}}{T_{yr}} = 607\ \text{cores} \tag{18}$$

To summarize our estimates we present Table 7. It lists the number of cores, disk, and tape storage required for the different stages of the CLAS12 computing enterprise plus totals for each

item. This is our estimate of the computing resources necessary for the CLAS Collaboration to analyze the data from CLAS12 in a timely and productive manner. Note the number of cores required for simulation is more than the number for reconstruction, reconstruction studies, and physics analysis combined.

	Cores	Disk (TByte)	Tape (TByte/yr)
DAQ	-	-	1,270
Calibration	173	-	-
Reconstruction	1,387	508	5,080
Simulation	8,139	318	1,558
Reconstruction Studies	1,214	508	-
Physics Analysis	607	889	-
Sum	11,520	2,223	7,938

Table 7: Requirements summary.