BNL--41847

DE89 003158

Data Acquisition for RHIC
Report of the Working Group

M. Atiya, B. Gibbard, R. Hackenburg, M. LeVine*, T. Throwe, W. Watson
Brookhaven National Laboratory, Upton, NY 11973

J. D. Cole, M. Drigert
Idaho National Engineering Laboratory, Idaho Falls, ID 83415

H. Huang
Massachusetts Institute of Technology, Cambridge, MA 02139

I. Juricic
Columbia University, New York, NY 10027

C. Lourenco
CERN, CH-1211 Geneva 23, Switzerland

* Convenor

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

1. INTRODUCTION

As experimental configurations for RHIC become better defined [1], the requirements for data acquisition for each of the evolving experiments becomes susceptible to detailed analysis. An earlier contribution [2] made it clear that the scale of these experiments makes demands on data acquisition that are at least as severe as some of the large-scale collider experiments being mounted at Fermilab and LEP. In this report, we attempt to answer the following questions:

- What sort of performance is required by each of the experiments?
- Is there a single architecture flexible enough to accommodate all of the proposed experiments?
- What are the costs associated with such an implementation?
- How far in advance of beam does a data acquisition implementation need to be started?

The RHIC accelerator design, with 114 bunches in each ring, allows approximately 100 nsec between crossings. Depending on the physical scale of an experiment, this may not be enough time to distribute a first-level trigger decision to all components of the detectors. It certainly is not enough time to allow for a sophisticated decision. In order to avoid dead time imposed by the first-level trigger, a mechanism must be provided to store [for eventual digitization] the pulses arising from each detector element, for several crossings.

Research has been carried out under the auspices of the U.S.D.O.E.

In carrying out this analysis, we were forced to anticipate the outcome of the development efforts for the readout electronics for RHIC (see [3]). The impact of the readout electronics on the data acquisition system architecture and costs cannot be underestimated. In the following discussion, it is assumed that the digitization of all signals takes place in chips mounted on the detectors and that the resulting digital signals are highly multiplexed before being shipped to the data acquisition crates. We also assume that sparse data scan [suppression of zero descriptors] takes place in the readout crates. This latter assumption has as a consequence that the limiting factor in number of channels per crate is no longer the number which can be physically accommodated [e.g., 96 channels/slot for a Fastbus crate], but, rather, the bandwidth limitation of the crate backplane.

The following discussion is carried out in terms of an architecture based on Fastbus crate segments. Computeıs manufactured by Digital Equipment Corporation are also mentioned. Focusing on specific implementations was necessary to calculate costs and performances; they are not to be construed as recommendations.

2. ARCHITECTURAL REQUIREMENTS

Data acquisition, in this report, is taken to begin at the point where level 2 trigger decisions have allowed the digitizers to run to completion. Event rates, therefore, refer to events which have survived a level 2 trigger decision.

Event rates of 5-50 kHz are possible from the point of view of digitizing hardware. A 50 kHz event rate for a tiny event size [10 kByte] corresponds to 0.5 Gbyte/sec data rate. Most of the experiments being discussed will have event sizes ranging from 100 kByte to a few Mbyte. We see immediately that the event rate quickly becomes irrelevant; it is the input data rate [event rate×event size] which is the limiting factor.

The output of a data acquisition system's front end is the event stream sent to the host for logging. This output stream is limited in data rate by available logging media to approximately 1 Mbyte/sec. Thus another important property of the data acquisition system is its ability to reduce the input data rate to the acceptable output rate via software [level 3] trigger decisions.

3. A COMMON ARCHITECTURE

We present an architecture for discussion in terms of the requirements outlined above.

3.1. The Readout Crates

The readout crates (Fig. 1) contain slaves which are likely to be receivers for the digital data streams, highly multiplexed on to optical fibers, from the detector-mounted digitizers. These modules will have pedestal memories and will suppress channels that are zero after pedestal subtraction. We estimate that the level of multiplexing at each fiber will be 100 and that as many as 24 fibers could be accommodated by a single Fastbus module. Thus a single Fastbus crate full of such slaves might represent as many as 48000 detector elements. The analysis presented assumes the bandwidth of the Fastbus backplane to be 40 Mbytes/sec. Even if only 10% of the detector elements fire in a typical event, a single such crate would produce 20 kByte/event, limiting the system to 2000 level 2 triggers/sec. If a higher rate is desired, fewer receiver modules can be accommodated in a single crate.

3.2. The Readout Controller

Mounted in each receiver crate is a readout controller which is responsible for reading the slaves in that crate and passing the resulting data on to a Fastbus cable segment. Each readout controller proceeds in parallel with the others, limited only by bandwidth in its crate.

It can be seen from the preceding discussion that it is essential to be able to read the contents of the slaves at the maximum possible rate. Thus the readout controller must be a bit–slice engine. [Several Fastbus masters based on bit–slice engines already exist.] The data stream from the crate segment is output directly on a cable segment. Thus each event fragment corresponding to the contents of a single readout crate is present on a distinct cable segment.

The bit–slice engine is controlled by a general-purpose microprocessor which is accessible via an Ethernet port (not shown in Fig. 1). Thus bit–slice instructions may be down–loaded from the host via this Ethernet port. Multiple Ethernet segments connected by bridges may be used here, as well as the 100 Mbit/sec systems which should be readily available in 3 years.

To facilitate subsequent processing, the readout controller will embed word counts for logical detector subdivisions in the data stream. This allows for rapid and efficient construction of pointers so that downstream processing is not required to search for descriptors.

Hardware would be provided to generate cyclical redundancy check words as the data is transmitted; corresponding hardware at the receiving end will check the integrity of the data on the cable segment.

3.3. The Spy

A spy on the cable segment functions as a frame grabber: it stores in its memory the contents of an entire event fragment. Analysis programs running on the host or on workstations can access these fragments via the spy's Ethernet port. Providing these spies at the readout crate level allows for monitoring of detector performance without wasting the Ethernet bandwidth by needlessly transmitting an entire event when only a small part of it is necessary. Spies at the next level provide access to the complete, assembled, event. While the spy is shown as functionally distinct from the readout controller, it might be implemented as part of the readout controller.

3.4. The Event Buffer Memories

Each cable segment is connected to a series of memories in event buffer crates. Each buffer crate has a separate memory, connected by a cable segment, corresponding to each readout crate. If the experiment is input–data–rate limited, then, to preserve total bandwidth up to the microprocessor farm level, there must be as many of these buffer crates as there are readout crates. In an experiment where the number of readout crates is determined by the number of detector cells to be read out, and where the data rate is not the determining factor, the number of event buffer crates may be reduced considerably, with a commensurate savings in cost. The use of multiple cable segments, with data flowing independently in each, allows the assembly of an entire event in one event buffer crate, without compromising the input bandwidth.

A synchronizing processor (not shown) selects an available event buffer crate as destination for the next transfer from the readout crates. A complete event is represented by the event fragments present in a single crate. These whole events are transferred from the buffer memories via a second cable segment to the third level trigger system. A second set of spies residing on these cable segments provides access to integral events for online monitoring.

The bandwidth limit assumed for all segments has been 40 Mbyte/sec, corresponding to a 100 nsec transfer time. This is a conservative estimate, based on some uncertainty about the reliability of the cable segment at higher speeds. Equally important are economic considerations arising from high speed memories required in the event memories; tens of Mbytes of 40 nsec memory will be costly! These are factors to be considered when operating parameters are defined.

3.5. The Third Level Trigger System

The third level trigger decisions are carried out by a farm of microprocessors arranged in banks; each bank serves a single event buffer crate. Events which survive the third level decision are collected by a dedicated microprocessor which transmits them to the host.

The microprocessors will be commercial, board-level products, based on one of several RISC chip sets appearing in the marketplace. These chip sets already have a processing power of 17 MIPS. It is conservative to estimate that 25 MIPS versions will be available in 3 years. A processor equipped with 16 MByte of memory should be available for under $10K. These processors will run code written in Fortran.

3.6. Limitations of the Architecture

Because there is an upper limit (approx. 24) to the number of event memories which fit into a single event buffer crate, and because there must be a buffer memory in each event buffer crate for each readout crate, this architecture cannot easily expand beyond 24 readout crates. This implies an input bandwidth limit of 1 Gbyte/sec.

The limited number of readout crates also implies that an experiment with a very large number of cells would have to provide multiplexing on a large scale. For example, an experiment with 10^6 detector cells, would require 40000 cells to be fed into a single readout crate.

4. ESTIMATE OF HARDWARE COSTS FOR AN EXPERIMENT

It should be apparent from the above discussion that a wide range of input bandwidth requirements, as well as a range of third level triggering demands, can be accommodated by the architecture outlined; it remains to be determined whether it is economically feasible.

Table I details a cost estimate for the case of:

1. Input bandwidth = 0.5 Gbyte/sec
2. Third level processing requirement = 1000 MIPS

The first item corresponds to readout crates equipped with readout controllers and spies, but otherwise empty. Too little is presently known about the readout electronics to be able to estimate the cost of the slaves. The number of readout crates required is dictated by the input bandwidth.

The number of event buffer crates is equal to the number of readout crates. The number of memory boards for this example is equal to the square of the number of readout crates: for input bandwidths much larger than 1 Gbyte/sec, the cost of this item quickly begins to dominate the total system cost. The cost of the event memory crates includes the necessary readout controllers and spies.

The device which synchronizes the readout controllers with available event buffer memories has not been included; its cost is expected to be small.

The microprocessor farm cost scales independently of the input bandwidth: here the number of MIPS required is given by the computing needs per event for the third level filter algorithm, together with the input event rate. For this example we use a farm with aggregate power of 1000 MIPS. With chip sets now beginning to be available, the power of 1000 VAXes can be compressed into 50 boards!

A host computer is required to control the front end, handle a reasonably large number of users, run database programs, and develop online and monitoring software. For an experiment of the scale expected at RHIC, a machine of the class of a DEC 6240 [12 MIPS] is required, outfitted with magnetic tapes, and an array of disks.

The host computer is meant to be the nucleus of the computing resources at the experiment. Its power is inadequate to serve the needs of a large collaboration; it also lacks the essential bit-mapped graphics required to generate rapidly the experimental views we have all come to find indispensable. These gaps are filled by an array of workstations. Some of the workstations will certainly need to be equipped with color graphics, while most of them will serve adequately with monochrome displays. The costs shown in this example include the file servers necessary to avoid saturating the host with the I/O demands of multiple workstations. The 18 workstations detailed here represent an additional 30 MIPS.

Finally, some third-party software will have to be purchased, notably a distributed database.

5. DEVELOPMENT COSTS AND SCHEDULE

We have estimated the development costs assuming that a single group would develop the hardware and software for all of the experiments. These estimates are not very different, however, from the costs each experiment will have to bear if, for example, four different efforts proceed independently.

The schedule for development is based on the assumption that the system should be essentially finished 2 years before physics running [BEAM–2] is anticipated. This allows the system to be used for detector debugging and calibration, and provides adequate time for resulting problems to be corrected. With this goal in mind, the group would begin to be put in place at [BEAM–5]. Figure 2 shows the expected staffing profile, ranging from five people the first year to 23 people at BEAM–2. An average salary (including overhead) is estimated at $100K. The cumulative cost of manpower over the development cycle would be $10M. If each of four experiments proceeds independently, the total cost will be four times this amount.

An additional capital cost of $1.5M would be required to purchase tools, developmental hardware and prototypes, and a small-scale host computer system. Here again, if each experiment pursues an independent development path, the cost to purchase developmental hardware, etc., for four experiments will be $6.0M.

6. SUMMARY

We have outlined an architecture which is flexible enough to meet the needs of a wide variety of experiments. Use of a single architecture allows for consolidation of development of both hardware and software. A number of arguments can be made to proceed in this direction:

- Duplication of manpower is avoided (saves $30M !!)
- The software development task can be carried out correctly:
  - Documented software
  - Diagnostics for both hardware and software
  - A single set of RHIC standards and formats facilitate adaptation of outside [e.g., CERN] software
- Enhanced buying power, for both hardware and software
- Duplication of development hardware and tools is avoided (saves $4.5M)

REFERENCES

1. *RHIC Workshop, Experiments for a Relativistic Heavy Ion Collider*, April, 1985, P. E. Haustein and C. L. Woody, editors, BNL–51921, and *Proceedings of the Second Workshop on Experiments and Detectors for a Relativistic Heavy Ion Collider (RHIC)*, May, 1987, H.–G. Ritter and A. Shor, editors, LBL–24604.
2. J.W. Sunier, *Proceedings of the Second Workshop on Experiments and Detectors for a Relativistic Heavy Ion Collider (RHIC)*, May, 1987, H.–G. Ritter and A. Shor, editors, LBL–24604, pp. 243–250.
3. Report of Readout Electronics Working Group, contribution to this workshop.

TABLE I: Experiment Data Acquisition

Budget Example

Assumption: 0.5 Gbyte input rate

• 12 Fastbus crates (empty) including readout controller @$20K		$250K *
• Event Buffer Memories 12 Crates, 12 Memories each @$3K		$450K **
• 12 Event Buffer Crates including Readout controller @$20K		$250K *
• Microprocessor Farm overhead		$ 75K
• Microprocessor Farm 1000 MIPS		
• 50 Nodes [20 MIPS/16 Mbyte] @$10K		$500K
• Host Computer [DEC 6240] Disk, Tapes, etc.		$500K
• 6 High quality color work stations @$30K includes file server overhead		$150K
• 12 monochrome, less powerful stations @$13K		$150K
• Purchased software		$100K
	Total:	$2425K

* Scales as (input bandwidth)

** Scales as (input bandwidth)^2

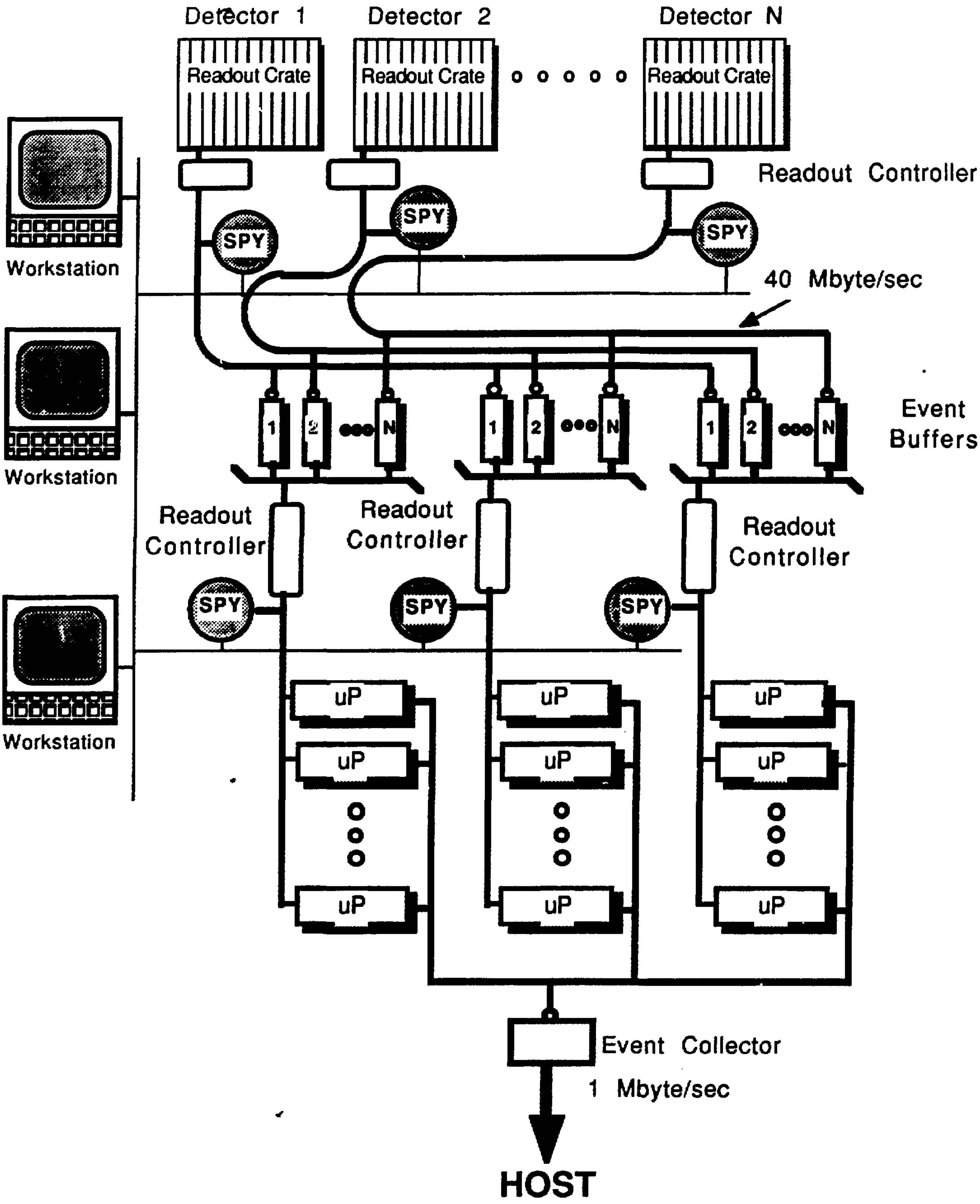

Figure 1. A data acquisition architecture for RHIC experiments

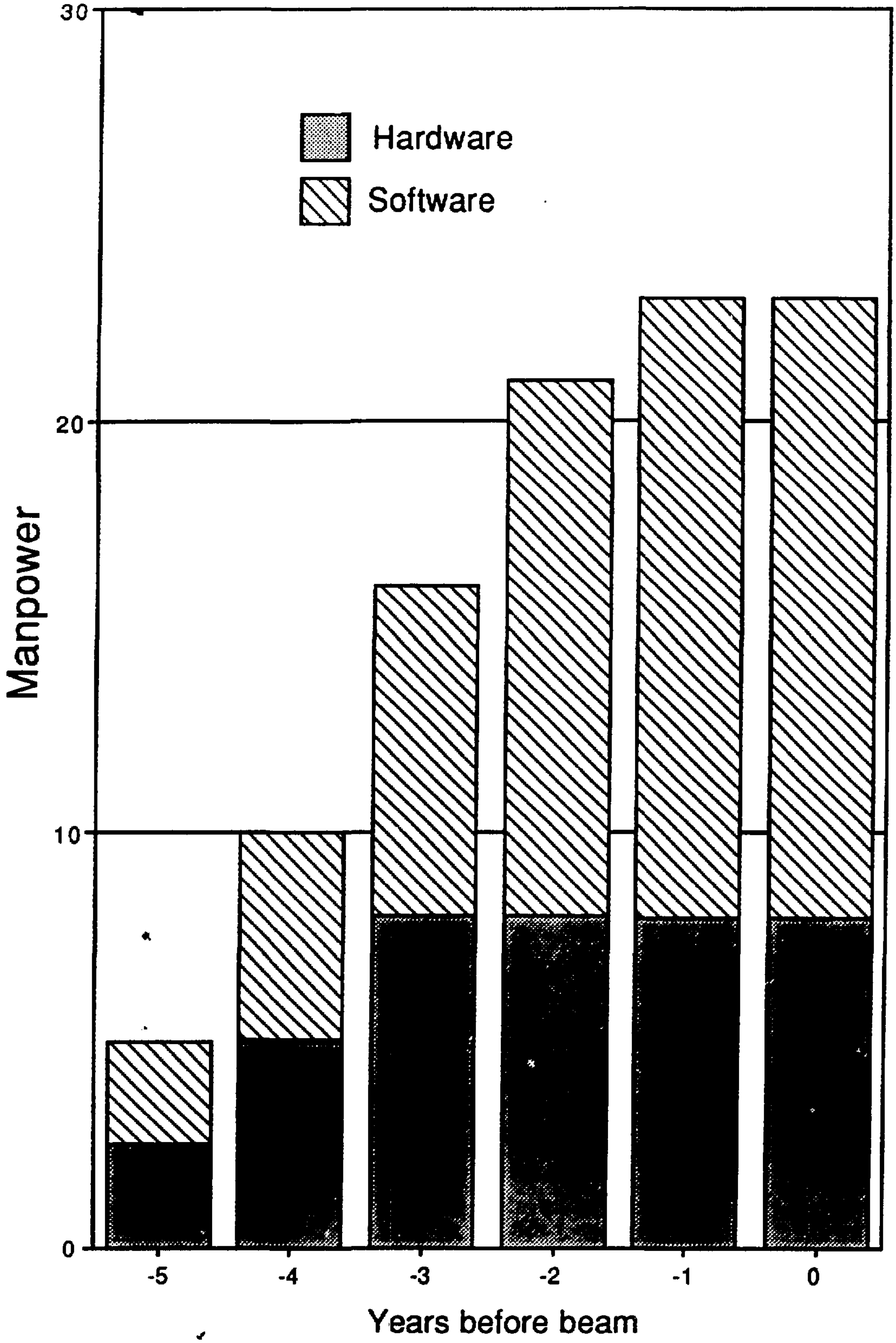

Figure 2. Required staff profile to implement data acquisition for RHIC experiments