STAR Note #375

The Link between STAR-DAQ and RCF

Motivation for an active buffer between the real time domain of the STAR-DAQ system and the the RCF data logging facility.

1.0 Motivation

We expect there to be latencies in the connection between RCF and STAR-DAQ of between 10sec and 2 hours. The data buffer described in this proposal will accept data from STAR-DAQ at 20MB/s and will accommodate this latency.

In normal operation, RCF will log the data into HPSS, which is not a system with real time behavior. The latencies quoted above are in agreement with reports from other sites using similar configurations.

In addition to these latencies, down times of the system due to failures have to be expected. These are hard to predict. Experience of other sites employing HPSS in data logging tasks suggests that unexpected behavior of the system is not uncommon. The recovery time from incidents seems to depend on two factors: the time it takes to locate the problem and the time the internal data bases require to recover. Taking into account the time for an expert to get access to the system an estimated down time of 1-2 hour is probably reasonable.

The DAQ system has been designed as a real time system. The management of of a large fast disk-based buffer will interfere with the operation of DAQ due to the chosen operating system (VxWorks) and hardware (VME based systems) .

1.1 Why an active buffer?

The connection proposed by RCF from the experiments to the central storage system will be via Gigabit-Ethernet. Benchmarks conducted by interface makers show that for the transport of 25MB/s, approximately 54% of an Ultra Sparc CPU running at 300MHz is needed. In these tests a Sun E450 with 4 CPU's was used.

The STAR-DAQ system uses VME based CPUs as building blocks. From benchmarks performed on these systems we know that handling 25MB/s of TCP/IP traffic would require more than 2 CPUs. Thus the impact of this additional CPU load on the DAQ would be intolerable.

In addition a suitable network controller in a PMC form factor isn't available.

2.0 Attributes of an active buffer

An active buffer should have sufficient CPU resources to handle the traffic and allow the management of the buffers.

For latencies of the order of seconds, the buffer should reside in the memory of the machine.

For longer latencies a disk buffer would have to support writing rates of at least 20MB/s and subsequent reading at 25MB/s to allow fast recovery of the buffer.

The connection between the real time part of STAR-DAQ and the buffer has to have the required bandwidth and must use less than 10% of the real-time CPU.

To ease the burden of maintenance and the cost of providing spare systems the machine should be of a type that is already employed by STAR-DAQ.

3.0 Required hardware

We propose to use a Sun E250 with two 250MHz CPUs, and at least 512MB memory operated under Solaris.

The E250 has to be equipped with two dual channel Ultra-SCSI PCI controllers (single ended) and an Alteon Gigabit Ethernet NIC. In addition, a set of converters from single ended to differential SCSI is required to be able to place the buffer system in a convenient location.

The E250 should be equipped with two internal disks for the system, its mirror image and swap space. These need not be larger than 9 GB each. The buffer space should consist of 18GB Seagate Barracudas. Striped writing will be employed using nine disks connected to three Ultra-SCSI channels. The disks will be located in an external disk tower.

An Ultra-SCSI bus will provide the connection between STAR-DAQ and the active buffer.

To avoid down time of the system due to power glitches we propose a UPS that has sufficient capacity to shutdown the system in an orderly manner in case of power failures.

3.1 Justification

We have tested striped disk I/O using SUN’s SDS (Solstice Disk Suite) on a SUN-E450 with three disks on two Ultra-SCSI controllers. This software package is bundled with SUN’s server level machines and available at no additional cost. Writing was possible at 25.7 MB/sec, while the rate for reading was 46.4 MB/sec. In both cases the file transfer size was 500MB. These rates are sufficient.

The connection via SCSI between our VME based systems and a SUN-E450 has been tested using the built-in Fast-Wide SCSI controller of the DAQ-CPU. The measured rate was 19MB/sec. The CPU load on the DAQ-CPU was of the order of 1% and the CPU load on the SUN was about 2% of one CPU (250MHz). Using one of the four Ultra-SCSI channels will provide ample bandwidth.

Of the requested memory of 512MB, at least 416MB can be used for buffering event data. This provides up to 20sec of data storage at STAR data rates.

The nine 18GB disks provide a total of 162GB of unformatted space which corresponds to about 140GB of usable space. This is sufficient for buffering of up to two hours of data taking.

The UPS is needed to avoid the time penalty involved in restarting a UNIX system that has been stopped in a non-synchronized way. This penalty is proportional to the file system size and thus large for the proposed system.

Two 250MHz CPUs are specified since the expected 25MB/s traffic over the Gigabit Ethernet link will consume about 65% of one of the CPUs. (See product performance results of Alteon Networks' Gigabit adapter.) To keep the system responsive, a second CPU is therefore required. For the current application two CPUs are preferred over one with a higher clock rate since the current price ratio between the 300MHz and the 250MHZ CPUs is 2.4.

4.0 Outline of the usage of the active buffer

Data are read from the DAQ SCSI connection and stored in the main memory. At this stage we will apply traffic shaping to insure that the input to the buffer will on average not exceed 20 MB/s.

The data are transfered from main memory to RCF via Gigabit Ethernet. If the memory buffer fills up due to connection related problems, the system will start writing data to the disk buffer. If the disk buffer is exhausted before the connection to RCF is re-started, the system will switch to record to our local tape drive. When the connection to RCF is available, the system starts writing from the memory directly to RCF. The disk buffer is emptied using the difference between the actual rate and the specified 25MB/s. The tapes will be handled by a second tape drive and transfered to RCF after the disk buffer has been cleared.

This design avoids writing and reading to the disk buffer at the same time and subsequently reduces the requirements of the disk system.

5.0 Responsibility for the system

Purchase of the system components listed in paragraph 6, is the responsibility of RCF.

The STAR-DAQ group is willing to design, implement and maintain a software system as described above. In addition the DAQ group is willing to do the system management needed for the proposed system.

6.0 Costs

TABLE 1.

Item#		**#**	**Price [$]**
1.	SUN E250 without CPU, memory,disks, graphics card or monitor	1	4,917
2.	UltraSPARC-II 250MHz/1MB CPU modules (two are required)	2	2,459
3.	512MB Memory (2*7004a) from a second source	2	578
4.	Alteon ACEnic Gigabit Ethernet Adapter	2	1,000
5.	Dual SE UltraSCSI PCI Card (SFSEWP1533P)	3	503
6.	Seagate Ultra SCSI 9.1GB Barracuda	2	500
7.	Ultra SCSI single ended 18GB Seagate Barracuda	9	930
8.	Small UPS	1	300
9.	Adapters from single ended SCSI to differential	2	640
10.	Case for housing the disks, including 300W power-supply	1	200
11.	Cables	1	100
sum			**25,750**

Items 2 and 5 each include one spare part. Neither spares for the buffer machine nor for the core components of this machine have been listed. In case of a failiure another SUN server can be setup as a short term replacement with limited performance. Nevertheless access to spare parts in a timely manner has to be guaranteed.