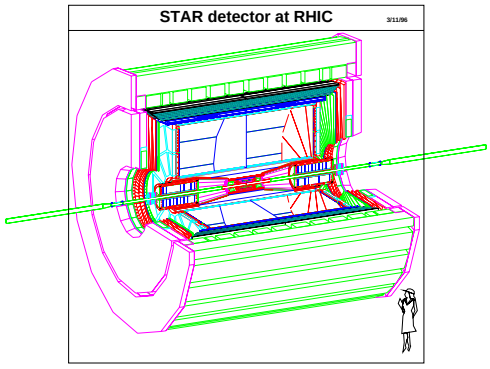
Inter-processor data transfer via SCSI

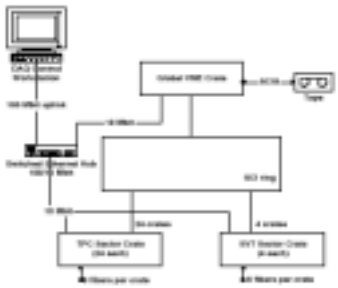
J.M. Nelson
University of Birmingham
Birmingham, England

M.J. Levine, T.J. Ljubucic, M.W. Schulz
Brookhaven National Laboratory
Upton, NY



The STAR experiment has been assembled at the Relativistic Heavy Ion Collider (RHIC) at the Brookhaven National Laboratory to investigate high-energy collisions of heavy nuclei. The ultimate goal is the discovery of the phase transition to a quark-gluon plasma.

The STAR detector consists of a number of sub-units of which the principal component is a large Time Projection Chamber situated in a solenoidal magnetic field. The data rate from the detector will be about 15 MB/s. The design of the STAR Data Acquisition system has been described previously by Ljubicic et al. (Ref: 1)



The diagram illustrates the main components of the data acquisition system. Data are transported over fibres from different detector sub-units for pre-processing and thence via an SCI network to the Event Builder in the global VME crate. Events will then be transferred to a storage medium which must accept an average data rate of at least 16 MB/s. Ethernet is used for system control and monitoring.

The RHIC Computing Facility (RCF) operates a Managed Data Server which is based on a High Performance Storage System (HPSS) - a cluster of computers and an automated tape library. The RHIC experiments, including STAR, will feed data to HPSS over Gigabit links.

Problem.

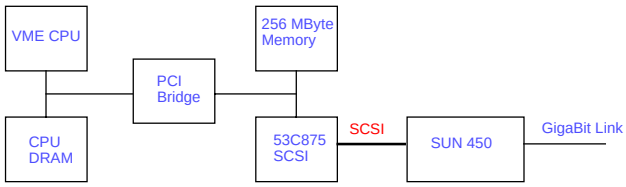
The CPU power required to service a Gigabit link is not available in the Global VME crate without serious interference to event processing. Instead, a dual processor SUN 450 workstation equipped with 1 Gigabyte of memory and substantial disc space, will operate the link and act as a data buffer for DAQ.

Requirement.

The link between the DAQ and the SUN 450 has to satisfy two main requirements:
a) it must support data rates at 20 MB/sec and preferably higher so as to free internal DAQ buffers as quick;y as possible.
b) it must consume not more than 5% of CPU time in the Event Builder processor.

Solution.

Ultra-SCSI has been chosen for this link. Operating with a clock speed of 20 MHz and in synchronous 16-bit (FAST/WIDE) mode, Ultra-SCSI provides byte burst rates at 40 MHz. DMA engines are used on the SCSI controllers on both sides of the link. The CPU demands are negligible.



Implementation.

An abbreviated layout of the Motorola MVME2604 processor used for the Event Builder is shown in the diagram above. Program structures and the vxWorks kernel are maintained in a 32 MB DRAM on the CPU's local bus while other components are on a PCI bus. Event fragments are located in a 256 MB memory which can be directly accessed by a SYMBIOS 53C875 Ultra-SCSI controller on a Cyclone PMC Card.

SCSI Controller

The 53C875 chip has a 4k on-board RAM for table space and code storage. It can be programmed to operate either as an initiator - the usual case - or in target mode. The DMA engine supports scatter/gather mode which is essential given the fragmented layout of STAR events in the Event Builder 256 MB RAM. The 64-bit instruction set allows the programmer to control all aspects of the SCSI bus and the DMA engine. Once triggered, the 53C875 is able to operate autonomously and will interrupt the host processor on completion of a task, if required.

The target mode program for the 53C875 is hand-coded and is down-loaded to the on-board RAM along with all necessary tables and constants. The code supports all SCSI initiator requests appropriate to a sequential device and, once triggered, does not required further intervention from the host - except for the execution of SCSI READ command. The device table in the SUN 450 is configured to support a standard sequential device /dev/star_daq.

Data transfer is initiated by an ethernet link between DAQ and SUN450 indicating that an event is ready by giving the total event size, for example, 20 MB. Actual transfer is executed by the following code sequence: (the open and close are normally executed once only.)

```
fd = open("/dev/star_daq",O_RDONLY)
xfer = read(fd,buffer_address,total_event_size)
close(fd)
```

The code in the 53C875 operates in variable block mode with maximum transfer per block fixed at 256 kbytes. The SUN I/O system manages the transfer by executing an appropriate number of SCSI READ commands each of 256 kbytes followed by a final transfer of the residue.

The 53C875 is loaded with the address of a scatter/gather list for one block and when transfer is completed the code halts the 53C875 and interrupts the MVME2604. At this point, the interrupt handler loads the address of the next scatter/gather list into a location in the 53C875 RAM and re-triggers the 53C875.

Since the 53C875 and the event fragment memory occupy the PCI bus and not the CPU's local bus, the MVME2604 is isolated from the data transfer and the CPU load from the interrupt handler is negligible.

Results

Data rates in excess of 38 MB/s have been measured with CPU loading on the MVME2604 less than 1% and SUN 450 less than 2%.

A separate SCSI controller operating at 20 MHz provides an additional link between the Event Builder and the Online system so that events may be transferred on demand for monitoring purposes. This link operates over 10 metres via differential SCSI line.

References:

1. Proceedings of the Tenth Conference on Real-Time Computing.
IEEE Transactions on Nuclear Science Vol. 45 (1997) 1907
A. Ljubicic et. al. for the STAR Collaboration