Abstract: We present the architecture of the Level-1 trigger for the future LHCb experiment at CERN. The system, a network farm of about 400 CPUs, has an input rate of 1.17 MHz and performs pattern recognition on the input data stream of 4 GByte/sec. The performance results of our prototypes which are based on a two-dimensional SCI network is shown. We demonstrate that we are capable of sending data blocks of less than 200 Byte with more than 1 MHz. Additionally we present a scheduling network to provide flow control in our system.

The Problem

Ever wondered where all the anti-matter from the big bang has gone? Since matter and anti-matter have been created to equal amounts when it all began physicists wonder what happened to the anti-matter. The future LHC experiment will look at proton-proton interactions occurring at a rate of 40 MHz producing a total data rate of about 40 TByte/sec.

An answer to the question above can be given by studying the asymmetry between heavy particles and their anti-particles. This asymmetry is called CP-violation. 

CP being simply a mathematical technique that removes a particle into its antiparticle.

A closer look takes us from a physics to a computer science problem!

Data coming from ODE (off-detector electronics) is merged to subevents by the RUs, the Readout Units, and enters a COTS (commercial off the shelf) computer cluster. Each compute node performs pattern recognition tasks on the entire event data in software and sends out a result. We can summarize:

- Synchronous input rate of 1.17 MHz (≈ 200 Byte/subevent, about 20 sources).
- Asynchronous output rate of 1.17 MHz (128 Byte/event).
- Maximum latency for an event < 2 ms.
- Scalability: total bandwidth and number of required compute nodes may increase.
- Use as many standard components as possible.

A Solution

TagNet Basics

- Using XY-routing we have to make sure that each column is used by only one sender at any given time. This avoids congestion and retry traffic.
- Thus, TagNet has to synchronize the RUs (Readout Units).
- The TagNet scheduler knows about pending events, CPUs available, and columns in use.
- A tag containing a suitable destination is formed and sent to the topmost RU. This sends its subevent and passes the tag to the next RU. The last RU hands the Tag back to the scheduler.
- TagNet transports target ID for the next event, error flags (to indicate that a token has been corrupted), ID of free CPUs, and error recovery messages (allows flushing of events).

Data Transport and current Implementation

I. Compute nodes export a memory region into the global SCI address space. Memory regions are imported by the source nodes. The software infrastructure is provided by the SISCI API. The imported memory region can be accessed by writing to an address, both virtual (user space) and physical (PCI).

II. FPGA's with a PCI interface are used to initiate a DMA transfer to a remote node. The hardware design implements PCI Master/Target functionality, a descriptor buffer to store the target address (see III), block size of the data to be transferred, and optionally an offset into the RAM where the data is located. A basic TagNet implementation is used to schedule the data transfer. The basic version of TagNet uses a one bit wide tag that signals to start the next transfer and a one bit wide busy logic which signals that the node is ready to accept a new tag.

III. The big picture: Data is transferred using device-to-device copy. The transactions are initiated by an FPGA connection. On the PCI bus and putting data into the SCI card's input buffer. We investigated scenarios using two FPGAs communicating via 2 Luench ORCA (33 MHz) and one 40 MHz bus and using a single FPGA (Altera APEX) on a 66 MHz PCI bus.

Results and current Work

1. Transferring 64 Byte data packets on a 33 MHz/32 Bit PCI bus (SCI 33/64, 2 ORCA interleaved): $f_{max} = 1.51$ MHz

2. Transferring 128 Byte data packets on a 66 MHz/64 Bit PCI bus (SCI 66/64, 1 ALTERA APEX): $f_{max} = 1.96$ MHz

Summary and current work:

- We showed that it is possible to transmit 128 Byte packets via SCI links with a rate of almost 2 MHz by using FPGAs.
- TagNet implementation has been tested locally. Basic implementation is sufficient in terms of frequency and reliability for current tests which transfer data to a remote node.
- We did extensive testing on a six node SCI farm in terms of minimum packet latency, reliability, and scalability. A 32 node cluster will be used to repeat the tests in a larger scale environment.

The solution presented is an architecture of a real-time trigger compute farm running at 1.17 MHz. The system consists of about 400 CPUs and can handle an input rate of 1.17 MHz while performing pattern recognition on the input data stream of 4 GByte/sec. The performance results of the prototypes, based on a two-dimensional SCI network, are shown, demonstrating the capability to send data blocks of less than 200 Byte with more than 1 MHz. A scheduling network is also presented to provide flow control in the system.