BNL-94172-2011-CP

Fast BPM data distribution for global orbit feedback using commercial Gigabit Ethernet technology

R. Hulsart, P. Cerniglia, R. Michnoff, M. Minty

Presented at the 2011 Particle Accelerator Conference (PAC'11)
New York, N.Y.
March 28 – April 1, 2011

Collider-Accelerator Department

Brookhaven National Laboratory

U.S. Department of Energy
Office of Science

DISCLAIMER

FAST BPM DATA DISTRIBUTION FOR GLOBAL ORBIT FEEDBACK USING COMMERCIAL GIGABIT ETHERNET TECHNOLOGY*

R. Hulsart#, P. Cerniglia, R. Michnoff, M. Minty
Brookhaven National Laboratory, Upton, NY, U.S.A.

Abstract

In order to correct beam perturbations in RHIC around 10Hz, a new fast data distribution network was required to deliver BPM position data at rates several orders of magnitude above the capability of the existing system. The urgency of the project limited the amount of custom hardware that could be developed, which dictated the use of as much commercially available equipment as possible. The selected architecture uses a custom hardware interface to the existing RHIC BPM electronics together with commercially available Gigabit Ethernet switches to distribute position data to devices located around the collider ring. Using the minimum Ethernet packet size and a field programmable gate array (FPGA) based state machine logic instead of a software based driver, real-time and deterministic data delivery is possible using Ethernet. The method of adapting this protocol for low latency data delivery, bench testing of Ethernet hardware, and the logic to construct Ethernet packets using FPGA hardware will be discussed.

NETWORK OVERVIEW

Data distribution requirements

To globally correct horizontal beam motion, a scheme using 36 BPM measurements to drive 12 dedicated corrector magnets (per ring) was selected [1]. Controller modules in each of the six service buildings provide setpoint signals to the four local corrector magnet power supplies (two for each ring). BPM position data from all 72 locations (36 per ring) is required at each of these six buildings for the correction matrix calculations, where each corrector setpoint depends on all 36 measurements from its respective ring.

Ethernet Broadcast Methodology

Since all six controllers require the same BPM data, the special 'broadcast' Ethernet addressing mode was utilized. On a standard Ethernet network, a destination address with all 48 bits set to '1' causes a packet to be forwarded to all nodes on a network. A daughter card was added to each of the 72 selected RHIC BPM electronics modules which could calculate the beam position within each turn (approx. 13 μs) and then transmit this data via a standard Gigabit Ethernet interface. By sending this data to the special 'broadcast' address, all network switches would automatically forward the packets to all six of the service buildings, for use by the controller modules.

*Work supported by Brookhaven Science Associates, LLC under Contract No. DE-AC02-98CH10886 with the U.S. Department of Energy.
#rhulsart@bnl.gov

Topology

In order to globally distribute BPM data across the 3.8 kilometer RHIC rings, the existing infrastructure of single mode fiber optic cable was utilized. As shown in figure 1, The existing cabling forms a two tier star topology with all equipment rooms within the ring (alcoves) feeding into the closest of six service buildings located at the even clock positions above the ring. These buildings then each have a home run cable to a central hub room in building 1005S.

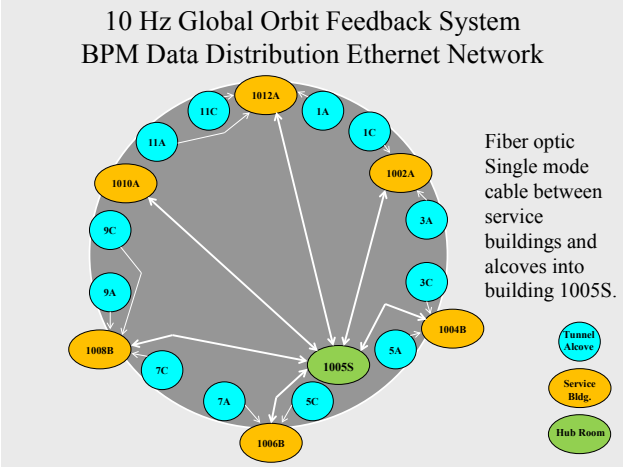

Figure 1: Network topology.

DATA TRANSMISSION PROTOCOL

Ethernet with TCP/IP

Although the relatively small amount of data which is needed from each BPM can be easily accommodated with the available bandwidth on a Gigabit Ethernet network, the use of the TCP/IP protocol usually necessitates a software driver stack at each end of the communications path. In addition the extra header information such as IP addresses and ports adds more bytes to the frame. Therefore using a layer 3 protocol (such as TCP/IP) introduces processing delays and unnecessary bytes (and additional non-deterministic delays if not using a real-time operating system). All of this additional overhead increases the latency of data delivery, which adds unwanted phase shift to the data. This is why most orbit feedback implementations have not used Ethernet, but some other custom protocol. Of course, this then necessitates the use of all custom switching and routing equipment that must be designed to transmit and receive packets using the custom protocol.

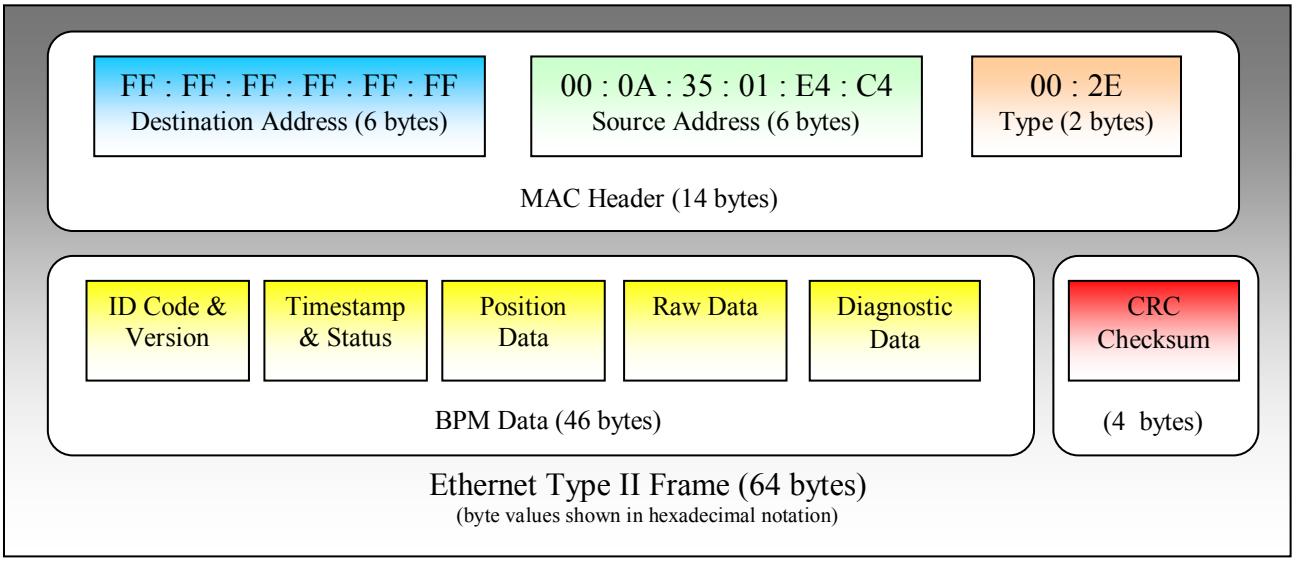

Figure 2: Raw Ethernet packet structure.

Raw Ethernet – Layer 2 Only

In a raw IEEE 802.3 Ethernet frame [2] without the additional TCP/IP protocol headers, the minimum packet size can be as small as 64 bytes. Figure 2 shows the structure of the custom packet that was developed. A packet of this size will still be forwarded by any off-the-shelf Ethernet switch which is commercially available. In addition, many modern FPGAs include transceivers and controllers that work with layer 2 Ethernet frames. This gave us the idea to implement a network which only uses the layer 2 Ethernet packet structure, without any TCP payload or IP addresses. Custom state machines were written in a hardware design language to be used by a FPGA at each end of the network to encode and decode these low-level packets at a very high throughput and low latency. With the Xilinx Virtex-5 FPGA and a Marvell physical layer controller IC, a data packet could be constructed and transmitted (or received and decoded) within 2.5 μs.

Data Transmission Rate

As shown above, the minimum packet size is 64 bytes for an Ethernet frame. Anything less than this will have zeros padded in by the hardware until this minimum size is reached. In addition, there are another 8 bytes added as a preamble by the hardware. The specification also calls for a minimum 12 byte gap between packets. Including these, the effective minimum packet size is 84 bytes, or 672 bits. At a rate of 1 bit per nanosecond (1 Gbps), it will take 672 ns to transmit a single packet. Table 1 lists the various data rates achievable using this scheme. The second row shows that 19 BPM measurements can be transmitted in the 12.8 μs RHIC revolution period. For our implementation using 72 BPMs, data could be transmitted every four turns, at a 20 KHz rate. We have decided to reduce this rate by a factor of two for our first operational tests in order to provide overhead room for additional nodes on the network. Therefore the current system transmits data every eighth turn (with an eight turn running average of data acquired each turn), which yields a data rate of approximately 10 KHz.

Table 1: Data Transmission Rates for Full BPM Dataset

Number of BPMs in Set	Time to Send Full Set (μs)	Time in RHIC Turns	Max Rate
1	0.672	0.0525	1.5 MHz
19	12.7	1	78 KHz
36	24.2	1.9	41 KHz
72	48.3	3.8	21 KHz

SYSTEM LATENCY

Acquisition and Cable Delays

Even though a daughter card module performing ultra-fast FPGA based position calculations was added to the existing BPM module, the analog-to-digital conversion still takes place on the older system board, which has about a ten microsecond delay between the sample trigger and the digital data delivery. The calculation can then be performed in hardware in a few microseconds. The other major contributor to latency is that some of the fiber optic cables are over a kilometer in length, giving light-time delays of tens of microseconds.

Ethernet Switch Latency Testing

The biggest question we faced in determining the feedback loop latency was the delays that commercial off-the-shelf Ethernet switches would introduce. Since the purpose of these products is not intended for real-time deterministic data delivery, but for maximum throughput, most manufactures did not include specifications for

latency. Those that did (such as Cisco) would only specify an upper bound (e.g. <20 µs). We realized we would have to measure this delay ourselves in the lab.

A test setup was constructed using the same Xilinx hardware that would be used in the final system, but with different firmware. One module would act as a transmitter, and another identical one as the receiver. A state machine was written for the transmitter that would send a single packet at a desired rate, and just as this packet was sent, an external pulse would be triggered that was connected to a scope. Similar logic was placed at the receiver, and this would generate a pulse if the correct packet was received, which was displayed on the scope as well. The total delay between packet generation and receipt could then be accurately measured. A baseline of 2.5 µs was measured with a direct connection between transmitter and receiver electronics. When a switch was installed in between for testing, this delay could then be subtracted from the total delay to yield the delay in the switch.

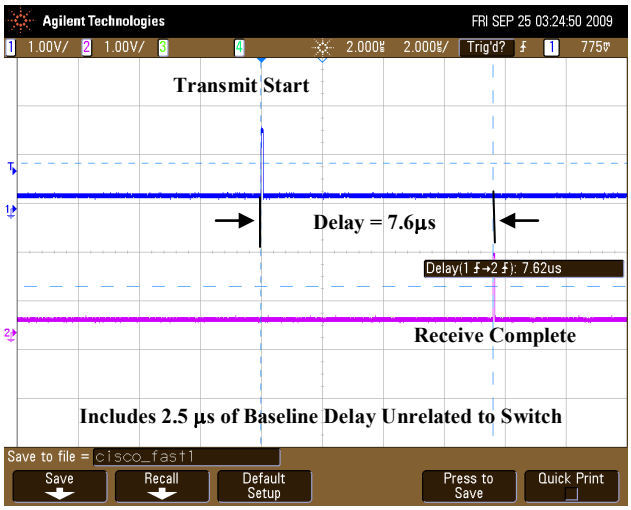

Figure 3: Cisco 2600 latency.

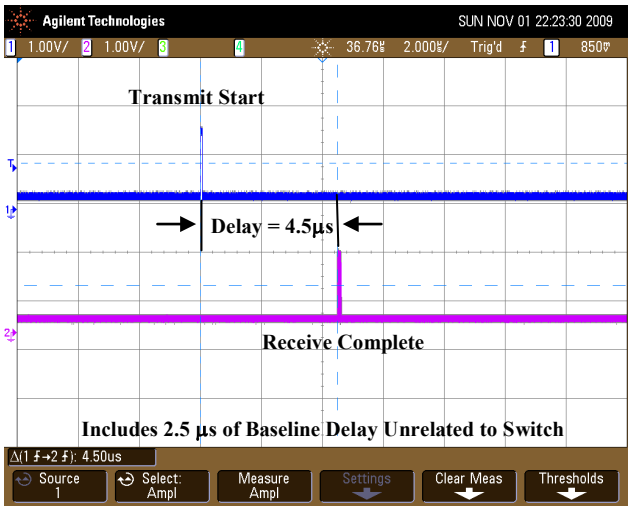

Figure 4: Netgear GS724AT latency.

Testing Results

Initially a Cisco 2600G 'managed' Ethernet switch (commonly used throughout the RHIC facility) was tested. These switches have a vast set of options for security and monitoring of network traffic. It was hypothesized that these management functions would add unnecessary overhead and increase the latency of the switch, compared to some other products that do not include such functions. Figure 3 shows the latency of a single switch of this type – about 5.1 µs (7.6 µs minus the 2.5 µs baseline). A few of the 'unmanaged' switches commonly available were also tested, and the fastest one found was the Netgear GS724AT, which measured less than half the above delay at 2 µs (baseline subtracted). This was also advantageous to know because the slower Cisco product was almost ten times the cost of the much simpler Netgear switch. With over 18 switches required for the complete network, this resulted in a much cheaper but still faster solution.

The data path from a single BPM module to a corrector controller module includes either three or four 'hops' across switches (depending if the BPM is in a tunnel alcove or service building – see figure 1). The above mentioned test setup was modified to add four switches of similar type in series, to simulate the four 'hops' (without the kilometer long cables used in the ring). For the case using all Cisco switches, the total delay was 30 µs. Using four Netgear switches, the total delay was 14 µs. We therefore selected the Netgear switches to be used in the final system.

CONCLUSION

A robust communications system using almost all commercial off-the-shelf equipment was developed in under a year which enabled retrofitting of the existing RHIC BPM system to provide 10 KHz data delivery for a global orbit feedback scheme using 72 BPMs. Total latencies from data acquisition at the BPMs to delivery at the controller modules, including very long transmission distances, were kept under 100 µs, which provide very little phase error in correcting the 10 Hz oscillations. Leveraging off of the speed of Gigabit Ethernet and wide availability of Ethernet products enabled this solution to be fully implemented in a much shorter time and at lower cost than if a similar network was developed using a proprietary method.

REFERENCES

[1] R. Michnoff et al., "RHIC 10Hz Global Orbit Feedback System," Proc. PAC 2011, New York.

[2] "IEEE Standard for Information technology - Specific requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications," pg. 49. http://standards.ieee.org/getieee802/download/802.3-2008_section1.pdf.