WIDAR Correlator Backend Processing Options

Bruce Rowen
NRAO, Socorro
2001 November 19

Summary:

Careful thought must be given to the design and topology of the WIDAR correlator backend if we wish to allow for flexibility in adopting future processing technologies as they become available. An “ideal” system would allow full designed data rates from the baseline boards to be implemented in the hardware from the beginning. Unless we go through a very expensive upgrade process, we are stuck with the limitations of the baseline boards data pipe for the life of the machine. Computing power and interfacing technology is always in a state of flux. It is to our advantage to stay with generic/ standard data pipes that will be compatible with current and future computing hardware. If higher processing data rates are needed we can take full advantage of the state of the art or pick optimum price/performance targets without concern of extensive modifications to the correlator hardware.

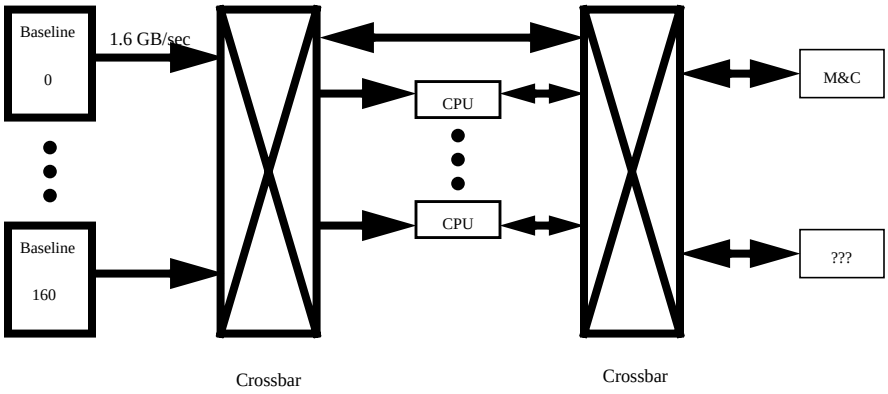
1 Introduction

The data pipes from the baseline boards rival current telcom and large computer network systems in bandwidth requirements. These systems are therefore good models to better understand our data flow issues and better envision hardware upgrade paths. In certain configurations the correlator has the ability to process baseline data over several baseline boards. This data needs to be “stitched” together before most of the post processing for that baseline can be completed. In a perfect world we could merge and split these baseline data flows in real time via a switched fabric topology. Backend computers need the ability to input, process, and output the data flows up to the rates demanded by the observational requirements. This could be accomplished by either one large complex computer or a number of smaller, simple computers organized in a cluster. As computing power continues to increase and prices fall, we should have the ability to scale the backend systems in either direction without totally scrapping the current hardware. The “backend computing” problem poses two separate but inter-dependent goals. We first need to maximize data flow from the baseline boards of the correlator while minimizing or eliminating future hardware modifications and second we need to best structure the backend computing environment to be both flexible and upgradeable.

2 Current Design Intentions

The current design indicates that a form of FPDP (Front Panel Data Port) be employed. This meets many of the goals, it is simple to implement, has data rates slightly beyond 100MB/s (Megabytes per second) per device, and is

“Ideal” correlator backend



inexpensive. Higher data rates indicate that multiple FPDP devices be installed in such a way as to allow additional data paths to be switched in as needed. Current baseline board schematics depict four FPDP drivers yielding a total data bandwidth of 500MB/s. FPDP 2 specifications call for a doubling of bandwidth which would extend baseline data rates into the gigabyte range. Once free of the baseline boards memory, data would then be bussed to a backend processor through a compatible FPDP receiver card installed on the computers bus.

Problems with the choice of FPDP in our ideal world are several, it is a short cable length solution, requiring data receiving devices be physically close to the baseline rack, and the general inability to switch the data flow before computers are inserted into the data stream. Switching the data stream without processor intervention is desirable for the same reason data dumping from the baseline boards should be “processorless”. Computer I/O bottlenecks are the dominant parameter limiting correlator data rates. The distance problem can be mitigated by choice of data medium (i.e. high quality cable) or use of FPDP “serializer” such as Systran's “FiberXtream” product. Cost becomes an issue with the serializer however since one module is required for each FPDP port and at a cost of $800 per port times number of ports times 160 baseline boards we are starting to talk about real cash. The serializer would however solve the switching limitation since COTS hardware is available which provides switch/route capability for serial FPDP.

Computing side drawbacks to FPDP lie in the bus nature of the interfaces. Current PCI bus based computers have 64 bit busses clocked at 33 MHz. This yields a net bandwidth of about 120 MB/s. Initial start-up specifications for the correlator suggest a maximum data rate of 25 MB/s per baseline board. Given these limitations, we see that a minimum of one computer for each set of four baseline boards will be required. Should for some reason the PCI bus be supplanted or otherwise obsoleted, we would have to replace all FPDP receiver cards with new hardware.

The ability to switch FPDP becomes a performance issue when an observation mode requires higher data rates from a smaller number of baselines. With the previously described configuration, each post processing computer would be configured to handle a net data flow of 100 MB/s from four baseline boards. If the observation mode desired 100 MB/s from each baseline board, we would have to physically swap cables so there was only one processor per FPDP stream.

This would require physical cable swapping and the possibility of moving processors closer to the baseline board racks due to the cabling length limitations. The serializers would overcome these problems but costs involved would quickly overwhelm any advantages.

3 Serial Data Pipes

An alternative to FPDP is choosing among the mix of SERDES (serializer/ deserializer) technologies available. The aforementioned FiberXtream module is one example of this, parallel data is buffered and then shifted out serially over some medium to a receiver which reverses the process. There are a number of high speed serial data transmission formats available today. Among the most applicable are infiniBand, FiberChannel, SONET, and the current host of ethernet formats (10BaseT, 100BaseT, Gigabit, 10Gigabit).

FiberChannel is a popular format currently used in large SAN (Storage Area Network) and other mass data movement environments. FiberChannel is a physical layer on top of which other protocols can be run (SCSI, etc.). Though much like Ethernet, FiberChannel is a point to point protocol. The current FiberChannel standard operates at 100-200 MB/s with a 2x version in the standards pipeline. Prestandard hardware is expected to achieve 1.25 GB/s in 2002.

InfiniBand is a modestly complex (software wise) standard for in-chassis communication that provides full duplex 250 MB/s in its minimum configuration. This requires two transmit and two receive wires (we don’t care about receiving). The InfiniBand specification provides for x4 and x12 links by using parallel implementations of the basic link.

SONET (Serial Optical NETwork) is a standard used by the telcom industry. Physically it is very similar to other optical networks, but it has its own unique framing data transport encoding. Data rates are in the 250MB/s realm.

More commonplace are the ethernet IEEE 802.3 protocols that are standard with virtually any COTS computer system and have the advantage of extremely competitive hardware pricing. Gigabit Ethernet (802.3z) is an extension of the older 802.3 10base and 802.3u 100base standards. except it runs (as its name implies) at 1 Gigabits/s. IEEE 802.3ae is the future 10 Gigabit standard for which several chip vendors have already produced prestandard chip sets. The final polishing and official standard designation is expected by May of 2002.

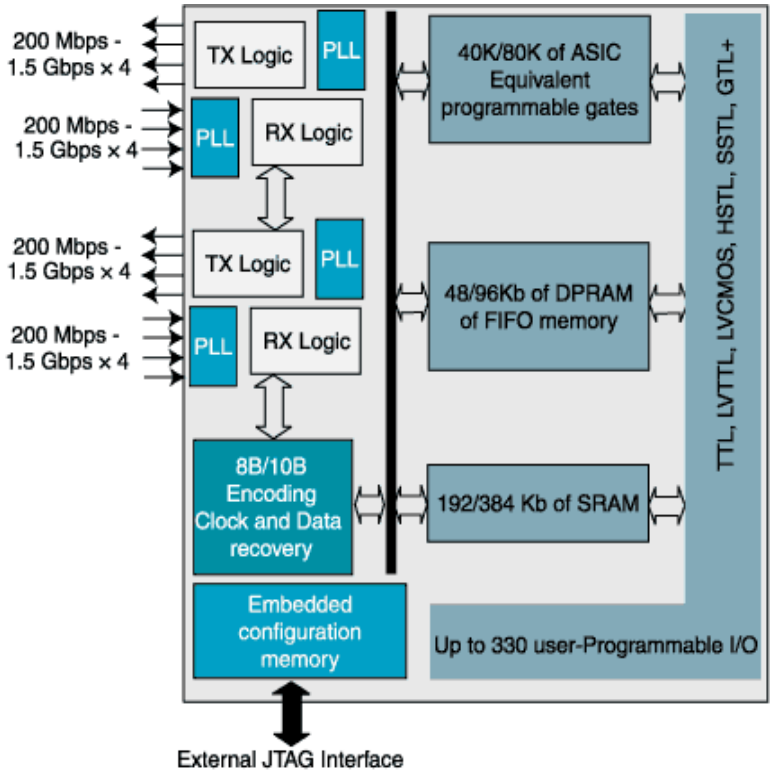
Choosing between one of the more proprietary protocols or the more prevalent 802.3 would have to be made for the backend computers, but should not become a wired in standard for the correlator’s baseline boards for obvious reasons of obsolescence. The wide variety of protocols and multiple vendor sourcing makes these products appealing.

Further advantages of the serial formats include their ability to be routed through COTS networking switches provided sufficient TCP/IP header information is grafted onto the baseline boards data frames to allow packeting. The switch isolates backend computer bus topology and cluster dimensions from the correlator. Various configurations of these switches can be used to optimize for raw data throughput with large number of processing nodes or they could support several flavors of post data processing by having the data stream be both processed and also stored onto disk as lag sets. Layered switching could also be used to reduce the net bandwidth requirements per switch with a trade-off of increased latency. The flexibility is there, we would not be forced into any single configuration for the life of the instrument.

4 Baseline Board SERDES Hardware

To achieve a baseline board data pipe that is both high speed and protocol insensitive has been to a large part already been solved. Several vendors make CPL (Complex Programmable Logic) devices that support all the popular high speed serial data formats through firmware personality programs. An example device that supports data rates up to 1200 MB/s and can also trickle data out at 802.3 10Base is the Cypress Semiconductor CYP15G and CYP25G family (http://www.cypress.com/psi/index.html). What makes this device attractive for use on the correlator backend is that the chip can support nearly the maximum designed baseline data rate (1600 MB/s) and can also interface with garden variety computers over ethernet via twisted pair copper. This device exists as a PSI (Programmable Serial Interface) core with enough unassigned cells to construct a simple parallel interface to the baseline boards LTA memory. Cypress development software for these series of chips, *Warp*, is listed with a price of $99(USD). LTA dumping could be scheduled without processor intervention. Setup and control of this device with the baseline M&C interface processor would allow full on-line system control of backend data pipe scaling. One device, near full spec baseline data rates, multi-protocol support, upgradeable, all in all a lot of the features desired for a long lived correlator data interface.

Installing full speed data capability on the baseline board up front has the advantage of longer hardware obsolescence cycles. Having a multi-protocol output allows us to scale the backend processing hardware as prices of networking equipment and computers decline.



Cypress Semiconductor “Frequency Agile” PSI with 4 or eight serial channels operating from 0.2 to 1.5 Gbps. *http://www.cypress.com/psi/index.html*

5 Backend Processing Scenarios

Example 1. Observation requires 160 baseline boards at 25 MB/s per board.

FPDP proposed system: Each baseline board dumps at 25 MB/s over single FPDP channels in sets of four into a PCI bussed PC (total of 160/4 or 40 PCs required). Should a fault in one of the PCs occur, human intervention would be required to swap out the PC with a shelved system.

SERDES proposed system: 25 MB/s baseline data pipes configured as Gigabit Ethernet protocol are fed into a switch fabric layer and directed to an appropriate number of PCs for processing. If we assume these PCs have Gigabit Ethernet interfaces with good implementations of motherboard DMA (for instance the current crop of G4 towers), 40 PCs would be engaged. Should a PC fail, the switch could be directed to a warm spare and processing continue with minimal interruption and human intervention.

Example 2. Observation requires 40 baseline boards at 100 MB/s per board.

FPDP proposed system: Each baseline board dumps at 100 MB/s over single FPDP ports. Cables are physically swapped and computers moved over closer to the appropriate racks so one computer handles one baseline board (40 PCs required). Should a fault in one of the PCs occur, human intervention would be required to swap out the PC with a shelved system.

SERDES proposed system: 100 MB/s baseline data pipes configured as Gigabit Ethernet protocol are fed into a switch fabric layer and directed to an appropriate number of PCs (40) for processing. Should a PC fail, the switch could be directed to a warm spare and processing continue with minimal interruption and human intervention.

Example 3. Observation requires 40 baseline boards at 10 MB/s per board. Cross baseline stitching will be required.

FPDP proposed system: Each baseline board dumps at 10 MB/s over single FPDP ports. Stitched baseline data is passed through the computer over it’s network port into the processing computer. The processing computer merges the FPDP data over its PCI bus with the incoming data over it’s Ethernet port. Processed data is set out the same or a secondary ethernet port for storage. (40 PCs required). Should a fault in one of the PCs occur, human intervention would be required to swap out the PC with a shelved system.

SERDES proposed system: 10 MB/s baseline data pipes configured as Gigabit Ethernet protocol are fed into a switch fabric layer. baseline data is merged into appropriate data streams and directed to an appropriate number of PCs (4) for processing. Should a PC fail, the switch could be directed to a warm spare and processing continue with minimal interruption and human intervention.

As can be seen, the use of a data switch greatly enhances the flexibility and robustness of the back end processing. FPDP can be switched, but at a cost of over $1000 per link to encode and decode each end of the line. Without switching there are a number of correlator data flows that either cannot be physically routed or they require extra processor overhead to channel the data. Upgrading computers with the SERDES hardware requires only that they support one of the several standards broadcast by the baseline boards transceiver. FPDP computer upgrades require similar bus standards with the old

computers or the FPDP receiver cards (or FiberXtream receivers) must be replaced.