THE WHITE RABBIT PROJECT

J. Serrano, P. Alvarez, M. Cattin, E. Garcia Cota, J. Lewis, P. Moreira, T. Wlostowski,
CERN, Geneva, Switzerland
G. Gaderer, P. Loschmidt, Austrian Academy of Sciences, Wien, Austria
J. Dedič, Cosylab, Ljubljana, Slovenia
R. Bär, T. Fleck, M. Kreider, C. Prados, S. Rauch, GSI, Darmstadt, Germany

Abstract

Reliable, fast and deterministic transmission of control information in a network is a need for many distributed systems. One example is timing systems, where a reference frequency is used to accurately schedule time-critical messages. The White Rabbit (WR) project is a multi-laboratory and multi-company effort to bring together the best of the data transfer and timing worlds in a completely open design. It takes advantage of the latest developments for improving timing over Ethernet, such as IEEE 1588 (Precision Time Protocol) and Synchronous Ethernet. The presented approach aims for a general purpose, fieldbus-like transmission system, which provides deterministic data and timing (sub-ns accuracy and ps jitter) to around 1000 stations. It automatically compensates for fiber lengths in the order of 10 km. This paper describes the WR design goals and the specification used for the project. It goes on to describe the central component of the WR system structure - the WR switch - with theoretical considerations about the requirements. Finally, it presents real timing measurements for the first prototypes of WR hardware.

INTRODUCTION

CERN started thinking about a suitable successor for the timing system of the LHC injectors in 2006. The main drawbacks of the current system are its limited bandwidth (500 kb/s on a multi-drop RS-422 line) and its lack of bi-directionality. Limited bandwidth results in an otherwise unnecessary proliferation of different timing networks for each accelerator, and this extra complication propagates through low-level software making maintenance harder. Lack of bi-directionality has two main disadvantages:

- Stand-alone timing receivers, though requested by clients, cannot be designed because there is no way to read status information back from the cards remotely.
- Cabling delay compensation cannot be automated. Instead, manual calibration using traveling clocks is used, resulting in manpower-intensive error-prone campaigns.

At the same time, GSI began brainstorming about the timing system for the FAIR facility, and since other collaborations with CERN were already underway it seemed natural to try to come up with a single timing system which served both sets of requirements. The similarities of the two complexes in terms of timing precision and sequencing needs helped in this regard. The requirement for a high-bandwidth full-duplex link quickly resulted in the choice of Ethernet for the physical layer. Indeed, Ethernet is not only a very high-performance and well known solution but also one where long-term support is beyond doubt, and this was an important requirement for both CERN and GSI.

The Ethernet ecosystem has been recently complemented with two standards which make the task of remote node synchronization easier. On one hand, Synchronous Ethernet [1] defines a clock transmission strategy based on recovering a clock from an Ethernet data stream using a Phase Locked Loop (PLL). This is of course not new to accelerator timing systems, but it makes the adoption of Ethernet in this area much more natural. On the other hand, cabling delay compensation can now be done using the Precise Time Protocol (PTP, IEEE 1588). The idea of mixing these two standards and coming up with a strategy to deliver messages in a completely deterministic way to all nodes in the network gave birth to the White Rabbit project.

SYNCHRONIZATION SCHEME

In order to achieve sub-ns accuracy in the synchronization of around 1000 nodes, WR defines a timing hierarchy by naming one of the switch ports the ”uplink” port, whereas all other ports are labeled ”downlink”. The first switch in the hierarchy gets its clock off an external source, such as a GPS Disciplined Oscillator (GPSDO). It then uses it to drive the encoding of all transmitters in each of the downlink ports. These downlink ports are then connected either to a final node or to the uplink port of another switch, therefore generating a tree of switches where all internal clocks are derived from the master source clock, as can be seen in Figure 1.

Once clocks have been transmitted and recovered in all nodes, there remains the task of compensating transmission delays, which can come from electronics in switches and nodes or from time of flight in the fibers. The first component is deemed fixed to first order and can be corrected by either manual or automatic calibration. The second component, that due to fiber delay, can show variations throughout the year due to thermal effects. For fibers not deeply buried underground and lengths in the order of 10 km these effects can well go beyond 1 ns. The traditional approach to solving this problem in Ethernet networks is to use a two-way scheme like PTP. However, this has the effect of generating traffic whose only purpose is to maintain synchronism, and this traffic might interfere with the critical application messages exchanged among nodes. WR proposes to use instead continuous measurements of the phase of the bounced-back clocks with respect to the transmit clocks in

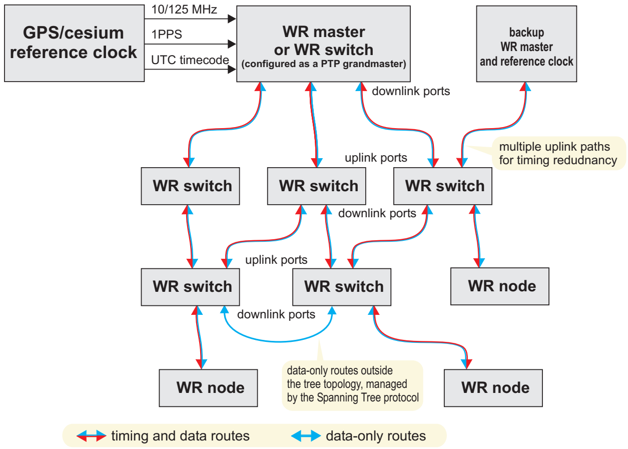
GPS/cesium reference clock
10/125 MHz
1PPS
UTC timecode
WR master or WR switch (configured as a PTP grandmaster)
backup WR master and reference clock
downlink ports
uplink ports
multiple uplink paths for timing redudnancy
WR switch
WR switch
WR switch
downlink ports
uplink ports
WR switch
WR switch
WR node
downlink ports
WR node
data-only routes outside the tree topology, managed by the Spanning Tree protocol
WR node
timing and data routes
data-only routes

Figure 1: Switch synchronization hierarchy.

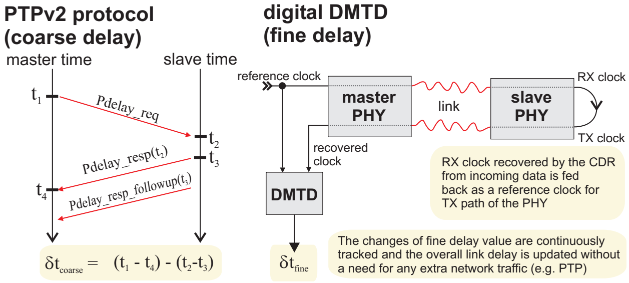
PTPv2 protocol (coarse delay)
master time
slave time
t1
Pdelay_req
t2
t3
Pdelay_resp(t2)
t4
Pdelay_resp_followup(t3)
δtcoarse = (t1 - t4) - (t2-t3)
digital DMTD (fine delay)
reference clock
master PHY
link
slave PHY
RX clock
TX clock
recovered clock
DMTD
RX clock recovered by the CDR from incoming data is fed back as a reference clock for TX path of the PHY
δtfine
The changes of fine delay value are continuously tracked and the overall link delay is updated without a need for any extra network traffic (e.g. PTP)

Figure 2: White Rabbit's PTP scheme with continuous phase measurement.

each one of the switch downlink ports, as can be seen in Figure 2. A PTP-like exchange can be done initially to figure out a rough estimate of the two-way delay expressed in 125 MHz ticks. From then on, the continuous phase measurement takes over and piggy backs on any traffic without perturbing it. Every 125 MHz tick is put to good use, so the performance in terms of synchronization is that of PTP with a message exchange rate of 125 MHz, and with absolutely no overhead. In addition, measuring clock phases can be done very precisely in a much easier way than measuring time intervals between pulses in a one-shot manner. The phase measurement scheme we chose for WR is a digital variation of the so-called Dual Mixer Time Difference (DMTD) technique [2]. The idea, as depicted in Figure 3, is to generate a frequency very slightly offset with respect to the frequency of the two sources whose phase difference must be measured. Sampling the two signals with the offset frequency results in very slow sweeping, and measuring time intervals between rising edge crossings at the output of the flip-flops can therefore be done easily with a simple counter running at a reasonable frequency. Thanks to the zooming effect of the slow sweeping, the coarse time intervals measured translate into real-life time differences in the ps range. Another advantage of this scheme is that it's completely linear, given enough stability of the frequency sources, and that it can be put in a loop to build linear phase shifters, a very difficult task otherwise.

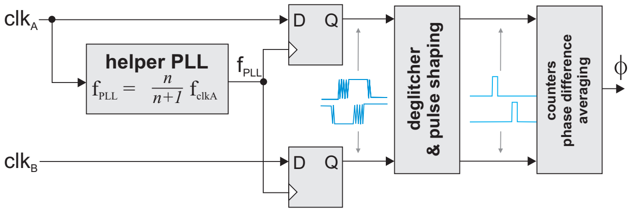
clkA
helper PLL
fPLL = n/(n+1) fclkA
fPLL
clkB
D Q
D Q
deglitcher & pulse shaping
counters phase difference averaging
ϕ

Figure 3: Digital DMTD phase measurement scheme.

DETERMINISM

Contrary to what happens to the clock, the data traffic sees no hierarchy. WR is a fully switched Ethernet network and any node can speak to any node at any time. It is the responsibility of higher layer protocols to keep traffic orderly. For this task, WR provides help in the form of different traffic types in layer 2, with different associated priorities. While the first choice for demonstrating the concept of different traffic classes was defining a "High Priority" (HP) type of frame with a special Ethertype field, current developments go in the direction of adapting the Quality of Service (QoS) prescriptions described in the 802.3q VLAN standard.

In any case, in order to ensure determinism in the latency of some types of traffic between two nodes, WR specifies different types of traffic that the switch needs to be aware of. In the event of an HP frame hitting a switch while Standard Priority (SP) frames are waiting for delivery in a pipeline, the HP frame would take priority and be output first. In order to keep long SP frames from occupying an Ethernet port for too long, automatic fragmentation of these frames and immediate forwarding of the HP frame is also supported by the WR specification. By automatic fragmentation we mean that the SP frame being output would be cut abruptly but with a special termination sequence that would allow the downstream switch to wait for additional fragments of the SP frame once the HP frame has been broadcast. Once all the fragments have been received, the complete SP frame could be delivered to its destination. Preliminary network simulations have shown this concept to work [3]. Fragmentation is specially useful in networks heavily loaded with HP traffic in which long SP frames run the risk of never reaching their destination.

SWITCH DESIGN

The switch is the core element of the WR network, implementing the standard IEE802.1x Ethernet Bridge functionality and WR-specific extensions. The extensions are enabled only after a proper WR handshake has been performed. Therefore if a non-WR aware device is connected, it sees a standard 802.1x switch.

The switch has been designed in μTCA form factor due to the high throughput, compact size and low price of

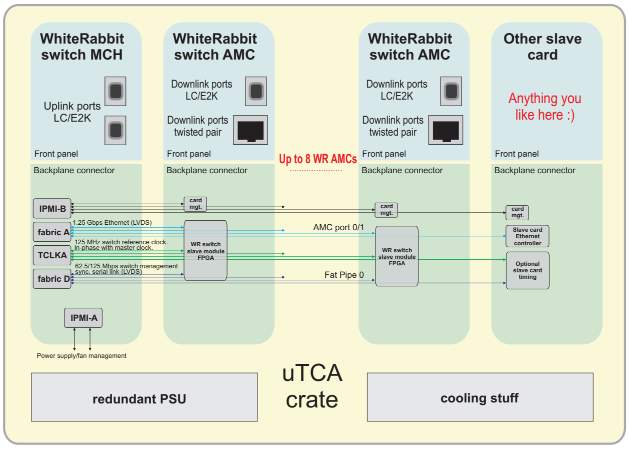

Figure 4: WR Switch implementation in µTCA form factor.

µTCA backplanes. The design consists of two types of cards: the MCH (Management Carrier Hub) card – which has the uplink ports and delivers WR-compatible Ethernet to slots in the µTCA backplane – and the AMCs (Advanced Mezzanine Card) with 4 downlink ports each, allowing for easy upgrades in the number of ports (see Figure 4).

The modular approach gives a possibility of integrating the MCH seamlessly with specialized AMC cards, such as the WR to GMT (General Machine Timing) translator card which interfaces the WR network with the existing CERN timing system. The MCH can also act as a timing receiver, delivering raw timing – reference clock, Pulse Per Second (PPS) and timecode signals – to slave cards which do not use the backplane Ethernet links. It is also possible to use the MCH alone as an independent WR switch with a special adapter board.

The MCH hardware consists of two FPGAs: one implementing the actual switch and another one containing the PLLs and timing hardware. There is also an ARM9-based CPU running Linux, doing the high-level part of PTP protocol as well as handling the configuration and management tasks of the switch via SNMP or SSH [4].

PERFORMANCE MEASUREMENTS

Figure 5 shows the jitter between the corrected clocks of two prototype MCH cards connected through 2 km of fiber. This jitter of 80 ps was measured over 10 minutes during which the fiber link was submitted through severe temperature variations using a hot air gun. Turning feedback off showed these temperature changes would have provoked uncorrected time shifts of several nanoseconds.

CURRENT STATUS AND FURTHER WORK

A preliminary protocol specification has been written and used as a basis for the development of a proof-of-concept MCH design. This card is now able to send Ethernet frames back and forth and compensate fiber delays

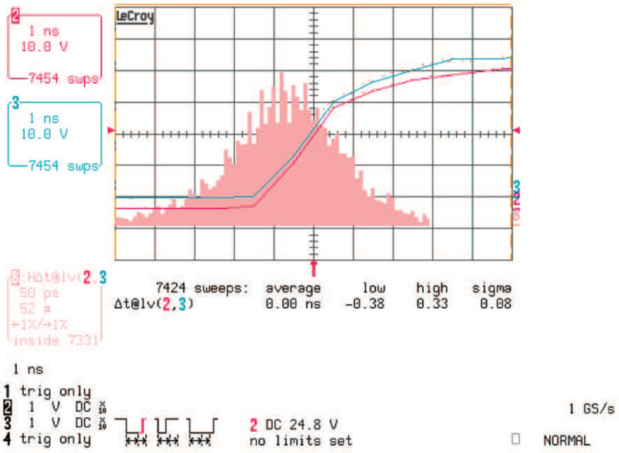

Figure 5: Histogram showing jitter over 2 km fiber link.

using the PTP protocol. Network simulations have shown the preemption idea to be promising, but no real-life tests have been done on deterministic data delivery.

The evolution of the project in the short term will involve finishing the switch prototype and designing end nodes to test point to point communication and delay compensation over multiple layers of switches.

REFERENCES

[1] ITU-T G.8262/Y.1362 standard, International Telecommunication Union.

[2] D.W. Allan, H. Daams. “Picosecond time difference measurement system”, Proc. 29th Annual Frequency Control Symposium, Atlantic City, USA, pp. 404-411, 1975.

[3] P. Moreira, J. Serrano, T. Wlostowski, P. Loschmidt, G. Gaderer. “White Rabbit: Sub-Nanosecond Timing Distribution over Ethernet”, ISPCS 2009, Brescia, Italy.

[4] White Rabbit Switch Technical Specification, see `http://www.ohwr.org`.