ISPCS 2009 International IEEE Symposium on Precision Clock Synchronization for Measurement, Control and Communication
Brescia, Italy, October 12-16, 2009

White Rabbit: Sub-Nanosecond Timing Distribution over Ethernet

Pedro Moreira, Javier Serrano, Tomasz Wlostowski
Beams Department
CERN
Geneva, Switzerland
{pedro.moreira, javier.serrano, tomasz.wlostowski}@cern.ch

Patrick Loschmidt, Georg Gaderer
Institute for Integrated Sensor Systems
Austrian Academy of Sciences
Wiener Neustadt, Austria
{Patrick.Loschmidt, Georg.Gaderer}@oeaw.ac.at

***Abstract*—White Rabbit (WR) is the project name for a ambiguous project that uses Ethernet as both, deterministic (synchronous) data transfer and timing network. The presented design aims for a general purpose, fieldbus like transmission system, which provides deterministic data and timing to approximately 1000 timing stations. The main advantage over conventional systems is the highly accurate timing (sub-nanosecond range) without restrictions on the traffic schedule and an upper bound for the delivery time of high priority messages.**

In addition, WR also automatically compensates for transmission delays in the fibre links, which are in the range of 10 km length. It takes advantage of the latest developments on synchronous Ethernet and IEEE 1588 to enable the distribution of accurate timing information to the nodes saving noticeable amounts of bandwidth.

I. Introduction

Large control systems have distributed nodes that need to be synchronized. The typical definition is that at a given instant of time all the nodes in a system must agree with a notion of the current time. Synchronization is necessary in a distributed system to establish a global ordering of events and tasks. This may be accomplished by having all nodes using the same external time reference. For example, the Coordinated Universal Time (UTC) can be distributed via telephone, radio, GPS each one giving different levels of accuracy.

This paper presents the design and implementation of an Ethernet based protocol, that accomplishes sub-nanosecond clock synchronization between 1000 stations over a tree structure topology with a maximum of 4 levels of hierarchy to the timing station. The clock synchronization accuracy is defined by a set of measurements relative to the UTC master clock. To achieve this end it was analysed and reduced the error factors to achieve clock synchronization within the defined accuracy.

The delay and jitter in the protocol stack are minimised by hardware assisted timestamping. In addition, the presented protocol implements, at its physical layer, frequency transfer using the carrier data link to ensure that master and slave nodes clocks are synchronous. This is enabled and standardised using Synchronous Ethernet (SyncE) [1].

The protocol also implements IEEE 1588, also known as Precise Time Protocol (PTP), for clock offset and link delay compensation. The combination between SyncE along with PTP results in a relative substantial reduction of PTP messages in the network. Therefore more bandwidth is available to transmit critical messages.

The organisation of the paper is as follows: It starts by presenting related work with respect to clock synchronization. Then, in section III, it is described the protocol based on Ethernet that allows sub-nanosecond accuracy. Section IV includes network simulations of the presented protocol followed by the presentation of the hardware implementation used to get some timing measurements that validates the performance of the proposed timing system. Finally, section V discusses test measurements, followed by conclusions and the plan for future work.

II. Related Work

Timing distribution can be done by standardised protocols for clock synchronization with easily available implementations including the Network Time Protocol (NTP) [2], the Precision Time Protocol (PTP) [3], or receivers for the Global Positioning System (GPS) and even systems based on IRIG-B.

NTP is a wide-spread protocol for clock synchronization. However, it is targeted at the Internet and its accuracy is within a few milliseconds. This protocol is designed to deliver reliable and long-range synchronization for applications with not too high accuracy requirements.

To fulfil the more demanding needs of test & measurement applications, IEEE 1588 (PTP) [3] has been developed, which is able to provide sub-microsecond performance. Many of the research activities concerning IEEE 1588 have been targeted at Ethernet [4]. Prior research shows that clock synchronization between the master node, which provides the UTC reference clock, and the slave nodes, which synchronizes to the reference clock, is possible in the ns-range [5].

The requirements for IEEE 1588-synchronized clocks in the higher levels of the LAN eXtensions for Instrumentation (LXI) standard [6] are making the ability to do timestamped measurements readily available in newly introduced, commercially available instruments.

IRIG-B is a popular time code format and is in use throughout the world providing UTC from GPS time and frequency receivers to a wide variety of devices. It is capable of providing microsecond level clock synchronization between devices [7].

978-1-4244-4392-5/09/$25.00 ©2009 IEEE

The GPS gets precise time information from the Navstar GPS satellites and can provide time with an accuracy of better than 100 ns relative to the international time standard, Coordinated Universal Time (UTC), as maintained by the U. S. Naval Observatory. GPS provides the state-of-the-art in accuracy to both the navigator and timing user. IEEE 1588 or GPS are the preferred clock synchronization methods for systems with high demand on the accuracy, e. g. test & measurement applications.

III. White Rabbit

White Rabbit (WR) is the code name for a project being carried out by high energy physics laboratories and academic research labs. The aim is to develop a distributed timing and data network capable of synchronizing up to 1000 nodes with an accuracy less than 1 ns relative to the master timing station. The data network should be able to have deterministic behaviour with very low transmission latency.

The choice to run in Ethernet networks is due to cheap cabling and infrastructure costs, high bandwidth, efficient switching technology, better interoperability, etc. However, industrial applications require the networked clocks to be precisely synchronized. This is much more challenging to accomplish over Ethernet.

Another characteristic of this project is that it follows the open source paradigm, with both software and hardware sources being distributed on a Open Hardware Repository web portal [8].

A. Synchronous Ethernet

While there are several ways to achieve frequency transfer over Ethernet [9], one gaining momentum is Synchronous Ethernet (SyncE) [1]. SyncE uses the physical layer interface to pass timing from node to node in the same way timing is passed in SONET/SDH. This gives confidence that networks based on SyncE will be not only cost-effective, but also as highly reliable as SONET/SDH based networks. The technology has been proven to be able to transfer very accurate timing over long distances [10], [11].

SyncE provides a mechanism to transfer frequency over the Ethernet physical layer, which can be traceable to an external source such as a GPS. Its primary idea is that the input reference clock is not free-running with an accuracy of ± 100 ppm, but rather locked and traceable to a primary reference clock as defined in ITU G.811 (achieving long-term accuracy of ± 10 ppt) [12].

The Ethernet link may be used and considered as apart of the synchronization network. SyncE provides a frequency transmission hierarchy formed on a link-by-link basis. The syntonization performance is evidently immune to variations of the traffic load and packet delay, because clock recovery works on the physical layer independent of data transmission.

Although syntonization exist in traditional Ethernet between the sending side and the receiver on one link, the recovered clock is only used to sample the incoming data. In contrast to SyncE the receiving node does not propagate the clock further, but uses a local oscillator for transmissions. Therefore, the clock of a sending node is only propagated to the directly attached opposite node of the link, but in the traditional approach, there is no way for building larger synchronous network structures.

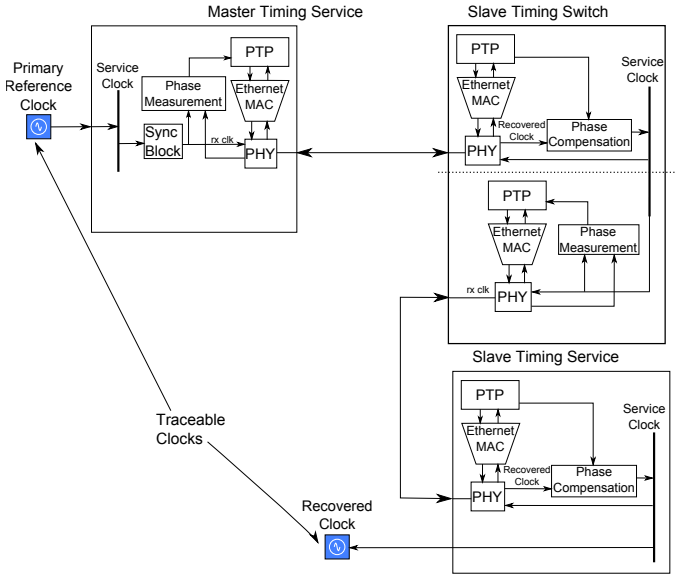

Fig. 1. Synchronous Ethernet in White Rabbit

Network synchronization in WR is based on clock hierarchy with the highest accuracy clock at the top. Slave clocks are PLL-based, locked to an external reference that is being recovered by the data link. The PLL cleans the recovered clock, i. e. it removes jitter generated from the clock recovery circuitry before being feed to the transmitting device. In addition, WR master provides additional functions beyond jitter cleaning which are presented later in this paper.

It is well known that accumulated phase noise degrades the synchronization performance between the master and slaves. Accumulated phase noise is usually the main factor for complete lost of synchronization on the timing station. White Rabbit tackles this effect by doing periodic measurements of the phase difference between the recovered clock and the reference master clock. The evaluated difference is then transmitted to the clock for further compensation in each slave. Consequently, there is almost no accumulation of phase noise in each node, when compared with packet-switching synchronization messaging protocols, which results in much better clock stability and reliability.

B. IEEE 1588

The Precision Time Protocol (PTP), standardised by the IEEE 1588 standard [3] was first published in 2002 and constitutes an evolving synchronization protocol for packet-switching networks (e. g., for Ethernet). The protocol uses messages carrying precise timing information. For high precision, hardware timestamps after the physical layer can be used to enhance the synchronization performance.

For larger networks, intermediate network elements called *boundary* clocks are required to compensate for switch-

ing/routing delays. These elements are adjusting their own clock to the master clock and then they serve as masters for the next network segment. Due to instability caused by a cascade of such elements, the *transparent* clock (TC) concept was proposed by [13] and introduced by PTPv2. These clocks calculate the packet residence time and are therefore transparent to the nodes. Consequently, no control loop in the intermediate element is needed for providing timing information to the next local clock and hence the synchronization at the timing slave is not dependent on the control loop design in the intermediate bridges. The transparent clock concept has been adopted in the new version of IEEE 1588 published in 2008 [9]. The authors of [14] show that it is possible to reach a synchronization precision of 1 µs for topology with up to 30 consecutive slaves.

To expand this limit it is important to study the factors that influence the quality of the synchronization process and encounter methods to minimise the effect of these factors.

The White Rabbit PTP clock implementation relies on two processes: phase measurements and line delay estimation.

C. White Rabbit Delay Model

Figure 2 shows an illustration of the clock synchronization process. This process relies on the propagation of PTP `Sync`, `Pdelay_Req`, and `Pdelay_Resp` messages in a peer-to-peer link.

`Sync` messages are managed in the White Rabbit network not only by the master node but also by the WR switches. The `Sync` message rate is defined by the quality of the recovery clock in the slave node. Measurements show that there is no need to transmit `Sync` except when the data link synchronization is lost. This loss of synchronization can be assumed to occur very rarely in GbE.

The main difference comparing to the IEEE 1588 standard is that `Sync` messages are only transmitted by the upper clock node, and not by the Best Master Clock. This method prevents PTP messages to be exchanged between long links from the master to the far side slave. As a consequence, firstly this reduces the number of `Sync` messages between master and slave which are trespassing the whole timing network and secondly it decreases the quantity of jitter introduced by each switch node.

This scheme enables more freedom in terms of timing to the WR network. In addition, the White Rabbit switches are responsible for arbitrating the IEEE 1588 messages to maximise bandwidth and to reduce latency of critical messages.

D. Fine Delay Measurements

In addition to link delay measurements performed by IEEE 1588, there is a phase measurement module that measures phase difference between master reference clock and the recovered clock. This module is based on the Phase/Frequency Detectors (PFD) [15]. The resulting phase difference between the referred clocks is transmitted to a slave node, using the correction field in the PTP message header. This enables the WR network to have the required sub-nanosecond accuracy.

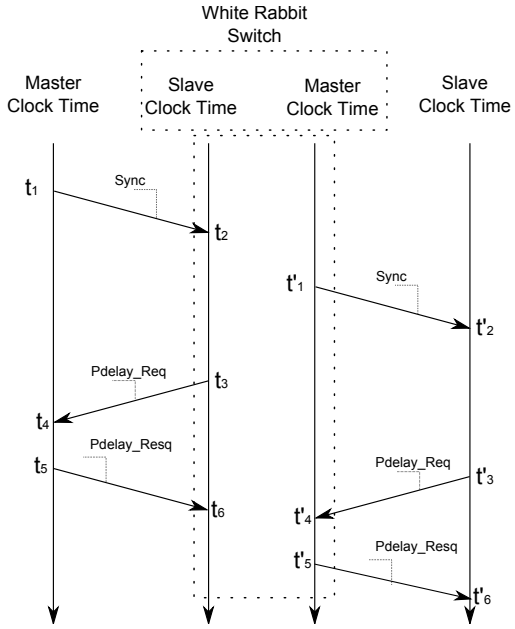

Fig. 2. White Rabbit delay model

IV. Performance Evaluation

Due to the big span of challenges, the evaluation of the proposed system is done using two strategies. First, a simulation gives information about for system and protocol characterisation. Second, hardware tests are performed to find out about the physical behaviour of the transmission links.

A. Simulation

As the White Rabbit system approach affects several layers of the network stack, there is an evident need for system simulation in order to find out about the overall behaviour. The final goal is to investigate the behaviour of the WR switch (which is the central and key element of the system), the traffic scheduling and the required upper layer protocols. Consequently, the simulation shall deliver information about the relevant parameters for traffic flow and the limiting factors and maximum performance values for bandwidth, delay, and packet drops. The advantage of such a simulation approach is that insight into the system internal variables (like buffer levels, traffic patterns or the like) can be gained synchronously, with a lot of further benefits compared to a real-world system. These include the possibility of simulating faster then real-time, setting of breakpoints without affecting the system stability, and easy recording of data of several distributed nodes.

The simulation framework is set-up in OMNeT++ [16] using the INET framework. Since the basic operation of WR is covered by a modified MAC for Ethernet, existing INET modules have been modified to support packet preemption, fragmentation and LT encoding of data.

All these measures are intended to allow for deterministic delivery times of HP (high priority) messages, which can be

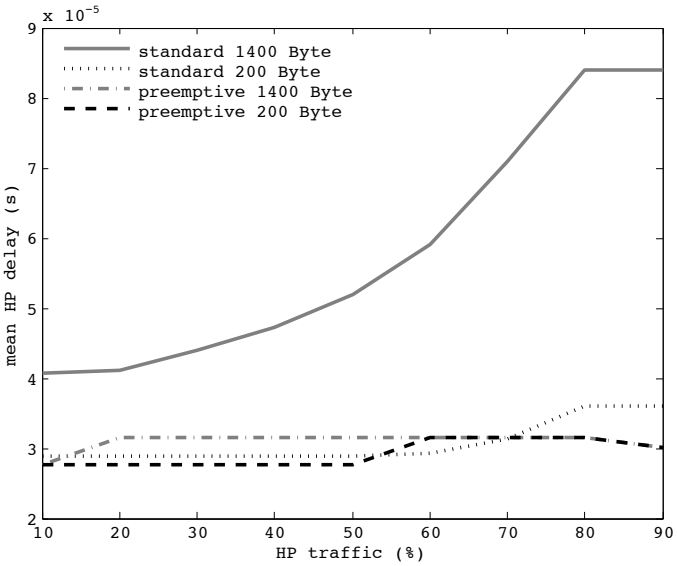

Fig. 3. Simulated delay performance for (non-) preemptive high priority packet treatment with two different lengths for standard priority messages. The load percentage is given in parts of the total link bandwidth.

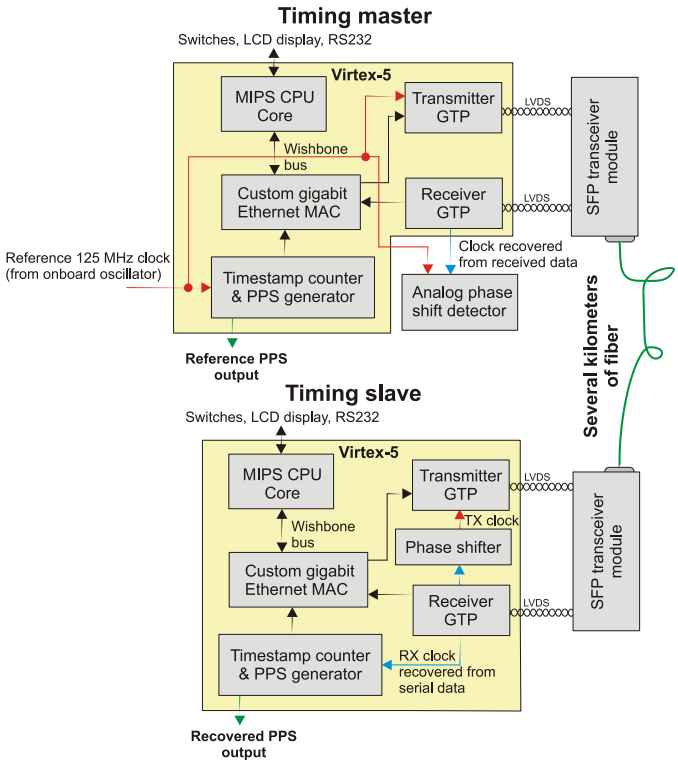

Fig. 4. Test bench block diagram

used for critical control operations. Besides those modifications of standard hardware, all other traffic is sent in standard Ethernet frames called SP messages (standard priority). The simulation of the MAC alone is able to deliver information about the advantage of a preemptive approach over common priority queuing and can therefore predict the behaviour of the overall network under certain load conditions, in case of collisions, and for specific error scenarios. These results give essential input for tuning protocol parameters, like queue sizes, timeouts and error handling.

The simulation itself is designed to support the evaluation of all layer of the network from the MAC upwards. Consequently, several other modules are currently designed to emulate the behaviour of PTP and the WR control message protocol in the proposed technology. The latter is used to announce the special capabilities of WR nodes in contrast to standard Ethernet nodes, which are also allowed to be connected without affecting the network. Using PTP, the simulation also gives essential hints on the expected synchronization performance based on the measurement results obtained from the hardware described in the next section.

One exemplary simulation result for a point-to-point connection in WR is presented in figure 3. The graph shows the delay behaviour for different loads of HP messages with a constant SP message rate of 25 % (measured in parts of the total link bandwidth). The simulation indicates that, in general, preemption on layer 2 delivers a lower delay behaviour than pure priority based traffic queuing. Further, it is independent of the SP message frame size. Especially on overload condition, the interruption of standard traffic allows to keep an upper boundary for the packet delivery time. The difference to priority queuing increases with the size of the SP frames.

B. Hardware Implementation

The proposed timing solution is implemented in a hardware test bench. The goal of the test bench is to analyse the network timing performance in a single node to node transmission link. The test bench is composed of two nodes, connected by 10 km of single mode fibre links. To evaluate the overall automatic link delay control loop of WR, the link is stressed to different temperatures, which changes the transmission delay.

Every node is implemented in an independent Xilinx Virtex-5 FPGA, each containing multiple multi-standard gigabit transceivers (GTP) that are used for data transmission. The clock from the master node is transmitted to the slave, by embedding it in the data transmission link. The recovered clock is then used as the reference clock for the slave. A block diagram of the test bench is shown in figure 4.

The master node is clocked using the 125 MHz on-board oscillator locked to a GPS receiver. This oscillator further generates the reference clock to the Ethernet transceivers, time counters and the pulse-per-second (PPS) outputs.

On the slave side the GTP clock recovery extracts the reference clock from the incoming link. When no data is being sent 8b/10b control characters, i. e. idle characters, are being transmitted to maintain the clocks synchronized. Both, slave and master clocks might not be in phase but they have the same rate. Nevertheless, their phase difference can be defined as constant during short time intervals.

Both nodes implement a simple custom made Ethernet MAC with packet timestamping capabilities. The timestamping oc-

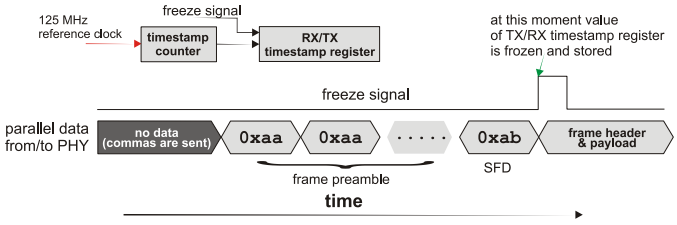

Fig. 5. Hardware timestamping

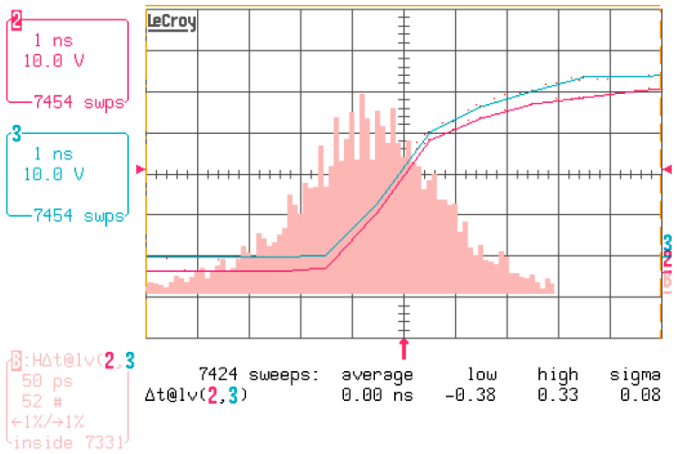

Fig. 6. PPS measurements & histogram

curs on the rising clock edge after the Start-Frame-Delimiter (SFD) symbol as it is shown in figure 5.

C. Measurements

To analyse the timing performance of the White Rabbit system, a PPS output is generated at the master and at the slave node. The difference between the rising edges of both PPS outputs will dictate the performance of a single point-to-point link in a WR network. The results of the measurements are shown in figure 6, where the difference between master and slave PPS output is given as a histogram. The timing link was left running for a couple of hours while, changing the link temperature by blowing a hair dryer to it. The histogram shows that both PPS outputs always correspond to each other with a difference in the range from -0.38 ns to 0.30 ns with a standard deviation of 80 ps.

V. Conclusion

The timing measurements show that the synchronization between the master and slaves variates at maximum ± 380 ps over the optical Ethernet transmission link. The line delays in optical fibres with lengths in the order of 10 km were automatically compensated, even during changing delay drifts caused by harsh temperature changes. This gives confidence that the proposed method can also be used in larger network structures to build a fully synchronous hierarchy which can be also used for applications with demand for high-accurate timing (e. g. accelerator sites).

For future work, detailed investigations using simulation will deliver information about the further influencing parameters and the break-even points between priority queuing and preemption. The gained results will then influence the detailed implementation for layer 2 of WR. Additionally, the inclusion of higher level protocol models will allow to verify the applicability to the desired use cases.

References

[1] ITU-T Study Group 15, *Timing and synchronization aspects in packet networks*, International Telecommunications Union, Geneva, Switzerland, Apr. 2008. [Online]. Available: http://www.itu.int/rec/T-REC-G.8261-200804-I/en

[2] D. L. Mills, "Internet Time Synchronization: The Network Time Protocol," *IEEE Trans. Commun.*, vol. 39, no. 10, pp. 1482–1493, Oct. 1991.

[3] "IEEE Std. 1588 - 2002 IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems," *IEEE Std 1588-2002*, pp. i–144, Nov. 2002, replaced by 61588-2004.

[4] J. Eidson, J. Mackay, G. M. Garner, and V. Skendzic, "Provision of Precise Timing via IEEE 1588 Application Interfaces," in *Proc. IEEE International Symposium on Precision Clock Synchronization for Measurement, Control and Communication ISPCS 2007*, Oct. 2007, pp. 1–6.

[5] P. Loschmidt, R. Exel, A. Nagy, and G. Gaderer, "Limits of Synchronization Accuracy Using Hardware Support in IEEE 1588," in *ISPCS 2008, International IEEE Symposium on Precision Clock Synchronization for Measurement, Control and Communication*, Ann Arbor / U.S.A., Sep. 2008, pp. 12–16.

[6] *LAN eXtensions for Instrumentation (LXI)*, LXI Consortium Inc. Std., Rev. 1.3, Oct. 2008. [Online]. Available: http://www.lxistandard.org

[7] P. Meinhardt, "Time Synchronised End to End Testing Using IRIG-B," March 2008, pp. 611–614.

[8] "Open Hardware Repository," 2009. [Online]. Available: http://www.ohwr.org

[9] "IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems," *IEEE Std 1588-2008 (Revision of IEEE Std 1588-2002)*, pp. c1–269, Jul. 2008.

[10] M. Calhoun, P. Kuhnle, R. Sydnor, S. Stein, and A. Gifford, "Precision time and frequency transfer utilizing SONET OC-3," in *Proceedings of the 1996 International Meeting on Preceise Time and Time Intervals (PTTI 1996)*, Dec. 1996, pp. 339–347.

[11] M. Weiss, S. Jefferts, J. Levine, S. Dilla, T. Parker, and E. Bell, "Two-way time and frequency transfer in SONET," in *Proceedings of the 50th IEEE International Frequency Control Symposium*, Jun. 1996, pp. 1163–1168.

[12] ITU-T Study Group 13, *Timing characteristics of primary reference clocks*, International Telecommunications Union, Geneva, Switzerland, Sep. 1997. [Online]. Available: http://www.itu.int/rec/T-REC-G.811-199709-I/en

[13] J. Jasperneite, K. Shehab, and K. Weber, "Enhancements to the Time Synchronization Standard IEEE-1588 for a System of Cascaded Bridges," in *Proc. IEEE International Workshop on Factory Communication Systems*, Sep. 2004, pp. 239–244.

[14] R. L. Scheiterer, C. Na, D. Obradovic, G. Steindl, and F. J. Goetz, "1 µs-conform line length of the Transparent Clock Mechanism defined by the Precision Time Protocol (PTP Version 2)," in *Proc. IEEE International Symposium on Precision Clock Synchronization for Measurement, Control and Communication ISPCS 2008*, 22–26 Sept. 2008, pp. 92–97.

[15] B. Razavi, Ed., *Monolithic Phase-Locked Loops and Clock Recovery Circuits*, 1st ed. IEEE Press, Apr. 1996.

[16] "OMNeT++ Community Site," 2009. [Online]. Available: http://www.omnetpp.org