Introduction

This informative contribution describes White Rabbit, a non-telecom application of combined Synchronous Ethernet and Precision Time Protocol which enables sub-nanosecond accuracy and picoseconds precision of synchronization.

Discussion

1. White Rabbit overview

White Rabbit (WR) is a nickname for a project and its expected outcome: an Ethernet-based network with low-latency, deterministic data delivery and network-wide, transparent, high-accuracy timing distribution. The WR project was launched at CERN within the renovation effort of the current control and timing system. The White Rabbit network is required to provide synchronization and convey control messages to thousands of devices within CERN’s accelerator complex spanning over 10km. The strategic decision to base White Rabbit on well-established technologies (Ethernet [1], Precision Time Protocol (PTP) [2] and Synchronous Ethernet (SyncE) [3]) resulted in a considerable interest in the project by other scientific facilities as well as companies. Therefore, WR is now a multi-laboratory, multi-company international effort to create a universal fieldbus for control and timing systems. In addition to CERN also the new Facility for Antiproton and Ion Research (FAIR) at GSI (Darmstadt, German) and the Laboratori Nazionali del Gran Sasso (Italy) will use WR. Some of the potential applications include Cherenkov Telescope Array, The Large High Altitude Air Shower Observatory (Tibet, China) and the Cubic Kilometre Neutrino Telescope (Mediterranean Sea).

WR is based on Ethernet, PTP and SyncE technologies. Compatibility with standards describing Ethernet and PTP has been ensured and verified. SyncE compatibility is for further study.

A WR network can be seen as a standard Ethernet/PTP network with additional features available for WR devices: (1) sub-ns and robust synchronization and (2) reliable, low-latency and deterministic data delivery. This document focuses solely on synchronization in WR.

2. White Rabbit synchronization

The accuracy of the PTP synchronization is implementation-dependent. The standard is foreseen for sub-nanosecond accuracies. However, such performance is not achieved in typical PTP implementations. WR achieves sub-nanosecond accuracy by basing its time distribution on PTP standard and addressing the following issues limiting PTP’s performance:

1. Limited precision and resolution of timestamps
2. Unknown link asymmetry
3. The quality of the PTP-synchronization depending on the exchange rate of PTP messages
4. Synchronization fluctuation during switchover between redundant connections or network reconfiguration
In order to address these issues two further technologies are used and combined with PTP: SyncE and Digital Dual Mixer Time Difference (DDMTD [3]) phase detection. The combination enables for network-wide sub-nanosecond accuracy of synchronization but requires additional (to standard PTP) logic and data exchange. These add-ons are incorporated into PTP with WR extension, called WRPTP [6].

2.1 Timestamps precision and resolution

The common notion of frequency distributed in the entire network over the physical medium enables to cast the problem of timestamping into a phase detection measurement (using DDMTD). The frequency transported from the master to the slave is looped back to the master, as depicted in Figure 1(a). This enables the master to detect the phase shift between the output (reference) frequency and the input (looped back) frequency. The process during which the DDMTD-detected round-trip phase shift (\(\Delta \phi_{MM}\) in Figure 1) is used to enhance timestamp precision and to calculate the precise round-trip delay (\(\text{delay}_{MM}\)) is called Fine Delay Measurement. During the fine delay process only the reception timestamp \((t_2, t_4)\) measurements need to be improved as they are transmitted and timestamped in different clock domains. A basic timestamp (to be enhanced) is obtained by the detection of the Start-of-Frame Delimiter (SFD) in the Physical Coding Sublayer (PCS). In order to acquire a precision-improved timestamp, it is necessary to eliminate the possible ± 1 LSB error (8ns) due to jitter of clock signals and clock domain crossing. Therefore, the Time-Stamping Unit (TSU) produces timestamps on both the rising and falling edges of the clock and later in the process one of them is chosen. The process involves three steps (see [4] for details):

1. Rising/falling edge timestamp choice \((t_r \text{ or } t_f)\)
2. Calculating the picosecond part, checking its sign and adding a clock period if necessary
3. Extending timestamps with the picosecond part \((t_{2p}, t_{4p})\)

The modified timestamps are used to calculate the precise round-trip delay:

\[
\text{delay}_{MM} = (t_{4p} - t_1) - (t_3 - t_{2p})
\]  

Figure 1. Fine Delay Measurement.

2.2 Link asymmetry

The link asymmetry is not detectable by PTP; if known, PTP corrects for it to increase synchronization accuracy. WR enables to automate the evaluation of link asymmetry by the following means:

1. Imposing restrictions on the WR network setup (only WR-compatible devices, single medium for two-ways communication)
2. Imposing requirements on the implementation (deterministic PHY, hardware timestamps of appropriate precision)
3. Providing theoretical means (WR Link Delay Model) and protocol support (WR extension to PTP) to enable measurement and exchange of the necessary parameters

High accuracy of synchronization is achieved in a network consisting of only WR devices connected with single fibers for two ways communication (1000BASE-BX10, 1310/1490nm pairs
with a single LC connector, as described in [5]). Such a network setup eliminates link asymmetries due to:

- different network paths of PTP packets (master-to-slave and slave-to-master)
- different fiber types and lengths in different directions

The reference plane for the link delay (and asymmetry) calculations is set on the SFP’s optical in/out (Figure 2). The effective points where the (enhanced) timestamps are generated are located in the FPGA-implemented phase detector (Figure 2). The delays between the effective points and the reference plane are denoted \( \Delta_{\text{txm}} \) and \( \Delta_{\text{rxm}} \) for egress and ingress, respectively, at the master, \( \Delta_{\text{txs}} \) and \( \Delta_{\text{rxs}} \) for egress and ingress, respectively, at the slave. These delays are called \textit{fixed delays}. The delays between the reference planes of the master and of the slave are denoted \( \delta_{\text{ms}} \) and \( \delta_{\text{sm}} \) for master-to-slave and slave-to-master directions respectively. They are called \textit{variable delays}.

Figure 2. Sources of fixed delays.

The place of the reference plane has been chosen carefully and the names of the delays describe their properties. Figure 3 presents \textit{WR Link Delay Model} which is used in WR calculations of link delays and link asymmetry. In the model, the one way master-to-slave delay is defined as:

\[
\text{delay}_{\text{ms}} = \Delta_{\text{txm}} + \delta_{\text{ms}} + \Delta_{\text{rxs}} = t_{2p} - t_{1}
\]  

The fixed delays (3) have source in the following (details in Appendix B.6 of [6] and [4]):

- Propagation delays of electronic components and PCB traces (\( \delta_{\{T/R\}X_{\text{CIR}}[\text{m/s}] \))
- Optical transceivers (\( \delta_{\{T/R\}X_{\text{SFP}}[\text{m/s}] \))
- Internal structure and clocking of the PHY (SerDes) chips which shall be deterministic between power-ups (\( \delta_{\{T/R\}X_{\text{PHY}}[\text{m/s}] \))

\[
\Delta_{\{T/R\}XJKLM} = \delta_{\{T/R\}X_{\text{CIR}}[\text{m/s}] + \delta_{\{T/R\}X_{\text{SFP}}[\text{m/s}] + \delta_{\{T/R\}X_{\text{PHY}}[\text{m/s}]}
\]  

The components \((a)\) and \((b)\) of the fixed delays can be evaluated during production and are invariant (in constant temperature). The requirement of deterministic PHY imposes its constant delay (component \((c)\)) at least within a Phase-Locked Loop/Clock Data Recovery (PLL/CRD) lock cycle. WR supports measurement of the PHY’s delay upon establishing a link (described in Appendix B.6.3 of [6]), if necessary. Consequently, the values of all the fixed delays are known. The temperature variation of the elements introducing fixed delays is assumed to cause minimal (if non) effect on the link asymmetry (proper design and implementation of hardware is crucial in this assumption).

The remaining source of the link asymmetry in a WR link, is the asymmetry of the physical medium – difference between variable delays (\( \delta_{\text{ms}} \) and \( \delta_{\text{sm}} \)). In case of a fiber connection (the only allowed in WR so far), the asymmetry is directly related to the different propagation velocities for the wavelengths used. The relation between \( \delta_{\text{ms}} \) and \( \delta_{\text{sm}} \) is called relative delay coefficient (4) in WR nomenclature.

\[
\alpha = \frac{\delta_{\text{ms}}}{\delta_{\text{sm}}} - 1
\]
It characterizes a given fiber and can be calculated using refractive indexes, or preferably measured for a given fiber type.

Knowing the coefficient (α), the round-trip delay (delay\text{MM}) and the fixed delays (Δ\text{txs}, Δ\text{txm}, Δ\text{rxs}, Δ\text{txm}), the asymmetry (6) can be calculated and further used to correct the mean path delay (5), as defined in PTP [6]. See [4] and [6] for a detailed derivation.

\[
\alpha = \frac{\text{delay}\text{MM}}{2} = \frac{(t_{4p} - t_1) - (t_3 - t_{2p})}{2}
\]

\[
\text{asymmetry} = \Delta_{\text{rs}} + \Delta_{\text{tx}} - \frac{\Delta - \alpha \cdot \mu + \alpha \cdot \Delta}{2 + \alpha}, \text{ where } \Delta = \Delta_{\text{txm}} + \Delta_{\text{txm}} + \Delta_{\text{txs}} + \Delta_{\text{txs}}
\] (6)

\[
\text{delay}_{\text{rs}} = \mu + \text{asymmetry}
\] (7)

The t_{2p} and t_{4p} in (5) are the enhanced-precision values of timestamps t_2 and t_4, respectively, which are obtained during the Fine Delay Measurement.

3. PTP and SyncE

In both PTP and SyncE networks, logic hierarchical topologies have to be created in order to establish the reference source and the direction of timing distribution (UTC/TAI and/or frequency) – the master-slave relationship between nodes. In this process, the nodes’ timing characteristics are exchanged and distributed algorithm decides on the logic topology.

The Synchronization Status Message (SSM) mechanism defined by SyncE to exchange nodes’ characteristics is not used in WR (SSM messages are neither issued nor handled by WR devices). Instead, the logic topology of frequency distribution is aligned with the hierarchy created by PTP. Such alignment is necessary as frequency distribution is used to enhance PTP’s timestamps and to perform the fixed delays measurement.

The Announce messages, used by PTP to exchange nodes’ characteristics, are additionally used to recognize WR devices (capable of frequency transfer). The distributed algorithm (Best Master Clock Algorithm, BMCA) which is used by PTP to establish logic topology, is modified to enable hardware-supported seamless switchover between redundant connections.

4. White Rabbit extension to PTP

A great effort has been made to align (above-mentioned) WR-specific solutions with the PTP standard and stay fully compatible with PTP gear. Consequently, WR can be seen as an extension to PTP (WRPTP) [6]. WRPTP defines WR-specific communication, logic, parameters and datasets as well as defining hardware and implementation requirements.

4.1 WR messages

WR benefits from PTP’s messaging facilities. It defines WR Type-Length-Value (TLV) to exchange WR-specific information. WR TLVs are suffixed to PTP Announce messages and are used to create Signaling Messages.

It is important to mention that the fractional nanoseconds part of timestamps is included in the PTP messages in correctionField, clause 9.5.10 of [2] and Appendix F of [6].

4.2 WR Link Setup

The process of establishing a WR link between two WR-compatible devices is called WR Link Setup. It involves the recognition of WR devices, syntonization over the physical layer, measurement of the fixed delays (if necessary) and exchange of their values across the link. In order to recognize WR devices, the PTP Announce message is suffixed with WR TLV. The standard PTP devices ignore the WR TLV (as defined section 14.1 of [2]), while WR devices read and interpret its content. The content of WR TLV includes (section 6.5.2 of [6]):
• The flag \textit{(wrConfig)} indicating WR-specific configuration of the port sending the message – the port can be configured to default to PTP \textit{(NON\_WR)}, allow to become both master and slave \textit{(WR\_M\_AND\_S)} or allow only specific role \textit{(WR\_S\_ONLY, WR\_M\_ONLY)}.

• The flag \textit{(calibrated)} indicating whether the measurement of fixed delays is necessary.

• The flag \textit{(wrModeOn)} indicating whether the WR Link is already established.

The conditions which need to be fulfilled to attempt establishing WR Link and enter WR Slave/Master mode are defined in sections 6.7.1 and 6.7.2 of [6]. Among others, the state \textit{(BMC\_MASTER or BMC\_SLAVE)} determined by the mBMCA needs to agree with the role allowed by WR configuration \textit{(wrConfig)}. If the conditions are fulfilled, the WR Finite State Machine (WR FSM) is started to control the process in which the frequency of the slave is locked to that of the master, the fixed delays are measured (if necessary) and the values of the fixed delays are exchanged. The transitions between WR FSM’s states are triggered by the reception of WR Signaling Messages or hardware events. The execution of each state is time-constrained (timeout). Therefore, a communication disturbance or misbehavior of one of the ports taking part in the process, results in quitting WR FSM and defaulting to the standard PTP.

\subsection*{4.3 Modified Best Master Clock Algorithm (mBMCA)}

The original BMCA allows only a single port (slave) of a Boundary Clock (BC) to be synchronized to a single grandmaster – a time source failure requires re-synchronization and might introduce fluctuations in the notion of time. The modified BMCA (mBMCA) allows for more than one best clock in a single domain, enabling the creation of a logic topology with multiple roots. A BC running the mBMCA can have more than one port in the PTP SLAVE state (slave). This means that timing information is exchanged between a BC and more than one source of time (i.e. Ordinary Clock or BC). At any time any of these sources can be used to perform synchronization (including a weighted average from all slave ports). WR provides special hardware support (Appendix B.2) to enable locking to frequencies recovered on multiple ports and seamless switching of the reference frequency.

The modification of the BMCA applies to the State Decision Algorithm (SDA): the BMC SLAVE recommended state is enforced instead of the BMC PASSIVE state for the clocks with clockClass greater than 127 (Appendix B.1). A port which becomes a slave as a result of the modification, is considered a Secondary Slave. A slave port resulting from the unchanged part of the SDA is considered a Primary Slave and its number is stored in the \texttt{primarySlavePortNumber} Data Set field defined by WRPTP. The best qualified Announce messages (\(E_{\text{best}}\)) from all Secondary Slave ports are compared using the Data Comparison Algorithm (DCA) to determine the "second best master" and the lower order masters. The results are stored in the \texttt{backupParentDS} defined by WRPTP.

\subsection*{4.4 WR PTP mapping}

WR uses the mapping over IEEE802.3/Ethernet to convey the protocol messages. Such mapping seems simpler and more plug&play:

• No network configuration required

• No address resolution required – WRPTP’s stack is also run on embedded CPU in FPGA ([8]).

Moreover, the WRN is meant to be Layer 2 network where WR devices are connected directly therefore UDP mapping does not seem appropriate.

In order to ensure proper functioning of mBMCA with the Spanning Tree Protocol, we currently consider whether to use multicast 01-80-C2-00-00-00 address for all the PTP messages or multicast 01-1B-19-00-00-00 address and appropriate default static entry in the Forwarding Database of all the WR devices.
References

[1] IEEE Std. 802.3-2008
[9] IEEE 802.1D-2004
Appendix A. – WRPTP Tests and Performance

A.1 Daisy chain of cascaded WR Switches

In order to test the performance of time and frequency transfer, the test setup depicted in Figure A1 was assembled. The system consists of 4 switches connected with 5km bare fiber rolls in a daisy chain (15 km total). Varying operating conditions were simulated by heating the fiber with a hot air gun. The results are depicted in Figure A2 and Table 1. A detailed description can be found in [7].

![Image of test setup](image1)

**Figure A1**

![Image of measurement results](image2)

**Figure A2**

<table>
<thead>
<tr>
<th>Switch (slave)</th>
<th>Integrated jitter [ps]</th>
<th>Offset (from master) [ps]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10Hz - 40MHz</td>
<td>mean</td>
</tr>
<tr>
<td>1</td>
<td>1.6637</td>
<td>161.86</td>
</tr>
<tr>
<td>2</td>
<td>2.4887</td>
<td>24.67</td>
</tr>
<tr>
<td>3</td>
<td>2.3025</td>
<td>-135.25</td>
</tr>
</tbody>
</table>

*Table 1.*

A.2 Temperature compensation

In order to evaluate the accuracy and precision of synchronization provided by the WR PTP Core – WRPTP stack running on embedded CPU in FPGA[8], a test setup depicted in Figure A3 was build. A detailed description can be found in [10].

![Image of test setup](image3)

**Figure A3**

The measurement lasted for 2.5 hours and the link was tested at temperatures between +12.5°C and 85°C. The resulting temperature-induced delay change of 17.5 ns was compensated such that the measured offset drift was below 100 ps over the entire temperature range (< 1 % of the change in the roundtrip delay), while the short-term (< 1 minute) rms master-to-slave jitter was smaller than 11 ps. The 3.3 ns constant bias is attributed to the systematic error due to the unmatched cable lengths.
Appendix B – Modified Best Master Clock Algorithm

B.1 Modified State Decision Algorithm Figure

Figure B1. Modified State Decision Algorithm

B.2 Modified BMCA and its hardware support

A seamless switch-over between redundant sources of timing (uplink ports) is heavily supported by the Clock Recovery System (CRS, Figure B2) of the switch and the WR extension to PTP (WRPTP). Figure B3 presents an example where a switch (timing slave) is connected (by its uplinks 1 & 2) to two other switches (primary and secondary masters) – the sources of timing. On both uplinks the frequency is recovered from the signal and provided to the CRS. Similarly, WRPTP measures delay and offset on each of the links and provides this data to the CRS. The mBMCA decides which of the timing masters is “better” and elects it the primary, the other is considered secondary (backup). The information from uplink 1 (primary) is used to control the CRS and adjust the local time. However, at any time all the necessary information from the uplink 2 is available and a seamless switch-over can be performed in case of primary master failure.

Figure B2. Clock Recovery System.

Figure B3. Timing Master redundancy.
Appendix C – WRPTP Profile proposed in WR Specification [6]

C.1 Identification

<table>
<thead>
<tr>
<th>PTP Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>profileName</td>
</tr>
<tr>
<td>profileVersion</td>
</tr>
<tr>
<td>profileIdentifier</td>
</tr>
<tr>
<td>organizationName</td>
</tr>
<tr>
<td>sourceIdentification</td>
</tr>
</tbody>
</table>

Table C1: Profile print form (clause 19.3.3 of PTP)

C.2 PTP attribute values

All nodes shall support the ranges and shall have the default initialization values for the attributes as follows:

- portDS.logSyncInterval: The default initialization value shall be 0. The configuration range shall be -1 to 6.
- defaultDS.priority1: The default initialization value shall be 68.
- defaultDS.domainNumber: The default initialization value shall be 0. Only the default domain is allowed.

C.3 PTP Options

All options of 15.5.4.1.7 and clause 17 of PTP are permitted. By default, these options shall be inactive unless specifically activated by a management procedure.

The node management shall implement the management message mechanism of the IEEE1588-2008 standard.

The best master algorithm shall be the algorithm specified in section 6.4 of [6] (Modified BMC).

The delay request-response mechanism shall be the only path delay measurement mechanism.

The TLV mechanism described in section 6.5 of [6] shall be supported.

C.4 Implementation-specific additions to PTP required by WRPTP

WRPTP shall implement the following implementation-specific features:

- issuing PTP messages with a WR TLV (i.e. suffixed Announce and Signaling Messages), as described in section 6.5 of [6]
- proper handling of PTP messages with a WR TLV as described in 6.5 of [6]
- a non-preemptive WR state machine as defined in 6.7 of [6]
- WR data set and additional WR-specific fields to PTP data sets as defined in 6.3 of [6]
- SYNCHRONIZATION FAULT state transition as defined in 6.6.2 of [6]
- MASTER CLOCK SELECTED state transition as defined in 6.6.1 of [6]
- communication with hardware as described in 6.9 of [6]