This full text paper was peer-reviewed at the direction of IEEE Instrumentation and Measurement Society prior to the acceptance and publication.

Timing models for PTP in Ethernet networks

Chandra Mallela (Intel-PSG, Penang), Kiran Tholu (Intel-PSG, Penang),
Mark Bordogna (Intel-SDG, Hudson)

ABSTRACT

The 1588 Precision Timing Protocol (1588-PTP) states that a timestamp event is generated at the time of transmission and reception of any event message and that the timestamp event occurs when the message's timestamp point crosses the boundary between the node and the network (event generation points). The protocol defines the message timestamp point for an event message as the beginning of the first symbol after the Start of Frame (SoF) delimiter. Up to Ethernet link throughputs of 10Gbps, due to the single lane nature of the Phy, this definition of timestamp point is quite easy to comprehend. While this definition is unambiguous even for higher throughputs, it adds considerable engineering challenge for the Ethernet link throughputs above 10Gbps, which can be typically a multi-lane in nature. Higher link throughputs such as 40Gbps & 100Gbps involve multi-lane distribution (MLD) via the virtual lanes in the PCS and the subsequent mapping of these virtual lanes into the physical lanes in the PMA, both introducing the possibility of timestamp event happening on any virtual lane and its subsequent mapping to any physical lane. Most of the 1588 designs in the market rely on timestamping inside a MAC. The delay from the timestamping point to the event generation point is estimated in order to calculate the net timestamp to be inserted in the relevant PTP event messages or to transfer the timestamp to the software stack. Thus, tracking the SoF delimiter from the timestamping point to the exit/entry point on the physical link, becomes utmost important in order to arrive at the timestamps of the highest accuracy required for various mission-critical timing applications. This article tries to bring out a comprehensive timing model detailing various delays to be considered for high-accuracy timestamps for Ethernet links. Such timing models may help organizations run upfront system simulations for the targeted accuracy requirement, saving implementation costs down the line.

Keywords

IEEE 1588; PTP; Timing model; single lane; Multi-lane.

1. INTRODUCTION

IEEE 1588 v2 [1] is a packet-based protocol designed to synchronize real-time clocks known as Time of Day (ToD) counters in Ethernet networks through phase synchronization, frequency synchronization (syntonization), with reference to a master ToD, typically utilizing hardware timestamping methods. The application software running on these networks, with the help of the 1588 synchronization layer, facilitates various audio/video and industrial instrumentation applications. The 1588 v2 protocol with the correction field in the packet header having 16b for fractional nanoseconds, can theoretically achieve a resolution of circa 15 femtoseconds though in practice, nanoseconds accuracy is demonstrated by various systems [2]. Hence, the Ethernet networks with 1588 capability are viewed as a profitable alternative to the costlier Time Division Multiplexing (TDM) networks (such as Synchronous Optical Networking (SONET) and Synchronous Digital Hierarchy (SDH)), which have in-built frequency synchronization, and even to software timestamping protocols such as NTP[10]. Such high precision & accuracy Ethernet networks help the mobile and telecom operators reduce their Capital and Operational expenditures and improve their profit margins even under immense competitive environment. The accuracy requirements vary depending on the network application. The power distribution networks require sub-microsecond phase accuracy [3]. The mobile networks require, at the moment, up to 50 ppb [4] (50 ns/s). The onset of LTE-A and 5G will only increase these accuracy requirements further [5]. These high accuracy applications demand the best timestamp accuracy.

Further, with increasing user demand for higher data rates, the access and aggregation networks along with mobile backhaul networks have started implementing high throughput Ethernet links in the upwards of 1G right at the access node itself. This necessitates even higher throughput links such as 10G in the aggregation networks and thus 40G/100G Ethernet links in the core networks. The increase in content has resulted in large amounts of data to be stored and processed in the data centers, which are thus on the verge of using very high throughput Ethernet links of 400G. As some typical time-sensitive, streaming applications span all the above Ethernet nodes, accuracy loss in any MAC node (MAC associated with Ethernet port) – be it low throughput node or high throughput node, is not an acceptable proposition.

Intuitively, higher throughput links should provide even better accuracies as they operate at high frequencies because the inaccuracies unaccounted in terms of clock-cycles will be less in magnitude due to the reducing time period of the clocks. However, the complexity of the 802.3 MAC node implementation increases as the throughput of the MAC increases from 1G to 100G and beyond, resulting in more variable delays, which when not accounted for, will result in higher inaccuracies at higher throughputs, contrary to our intuition. This complexity can chiefly arise due to the shift from the single-lane link implementation up to 10G Ethernet to multi-lane link implementations for 40G and beyond. Further, the attachment unit interfaces (AUIs) of different forms in the system will introduce delays that can vary from reset to reset, adding to the complexity further.

Mendel, D. et al [6] have raised the complexity in multi-lane Ethernet links as ambiguity in the 1588 v2 protocol and tried to address the ambiguity through modified statistical definitions, which might not yield any accuracy improvement, except convenient mathematical expressions. In the process, however, they have shed light on the possible sources of inaccuracy, which one must consider for high-accuracy timestamping. We believe that the protocol is accurate in its definition – 'a timestamp event is generated at the time of transmission and reception of any event message and that the timestamp event occurs when the message's timestamp point crosses the boundary between the node and the network'. This definition is both necessary and sufficient to derive better timing models that yield improved accuracies when the delays identified in the model are measured with appropriate implementation mechanism for Ethernet links of any throughput – be it single lane or multi-lane Ethernet links.

This has motivated us to arrive at timing models that capture the possible delays and the delay variations both in a single-lane link and a multi-lane link of the Ethernet nodes, compliant to the protocols: IEEE 802.3 and 1588 v2. Such timing models will help us to understand the error sources in the MAC node implementation as well as in the network upfront and address any possible unaccounted delay upfront with a suitable mechanism. Such timing models taken in an entirety of the network might also help in simulating the network behavior to estimate the achievable accuracy in the entire network. Further, these models can be improved by adding additional variables observed as the differences between network simulation and validation of the actual network, which eventually expedite the realization of complex time-sensitive systems spanning multiple networks.

This following sections detail the scope of the timing models to begin with, and then bring out the sources for delays and delay variations leading to the timing models for single-lane link for Ethernet throughputs up to 10G and the timing models for multi-lane link for Ethernet throughputs from 40G/100G and beyond. The article ends with conclusions and scope for future work.

2. A generic timing model for 1588 implementation in the MAC

A unit timing model for an entire Ethernet network consists, chiefly, of repeated (a) MAC node timing models covering up to the meeting point with the network (b) link (cable) delays (c) delays of the repeaters, AUI (Attachment Unit Interface) components present between the links, (d) analog delays from the clock generators and PLLs all along the network path for the 1588 packets, (e) delays due to the clock and the data jitter, as shown in Figure 1.

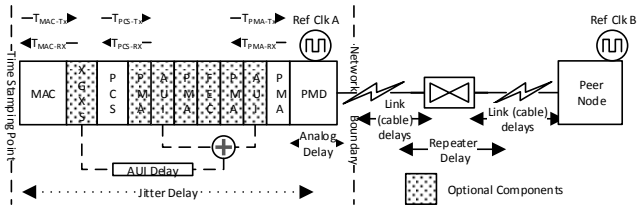

Figure 1: Unit timing model components

The article focuses on the timing models arising from (a) & (c). In general, each delay in the model can be characterized by the mean delay and the standard deviation representing the variable component. The worst-case intrinsic uncertainty of the unit system spaning (a), (b), (c), & (d) is the sum of all the variable components plus or minus any unaccounted delays such as asymmetry and thus represents the maximum accuracy attainable in the system. At a network level, the interplay among these models needs to be simulated during packet flow in the context of ordinary clocks (master/slave), transparent clocks and boundary clocks in order to assess the maximum possible accuracy in the entire network or the multitude of networks, as shown in Figure 2.

We assume that the timestamping is carried out inside the MAC while the protocol does not specify or imply any such mechanism. The key reason is that the complete packet at layer2 is visible, thus easily parsed to check whether it is a PTP packet, modified as required in terms of the ToD and CF fields in the packet header and its CRC on the modified packet can be calculated accordingly – required for 1-step operation that is more efficient from network bandwidth perspective. However, the same is not possible in the PCS (Physical Coding Subsystem) and the PMA (Physical Media Attachment), where the encoded (either 64/66b or 8/10b) packet cannot be parsed to enable PTP 1-step operation without the necessary logic for possibly word-aligner and deskew logic, decoding, parsing, field updates, CRC update and re-encoding CRC-updated PTP packet. Thus, the best place for timestamping a PTP packet is either MAC or a component (be it a PCS or PMA) where the logic components of the MAC - parsing, field updates, CRC update - need to be reused as explained above.

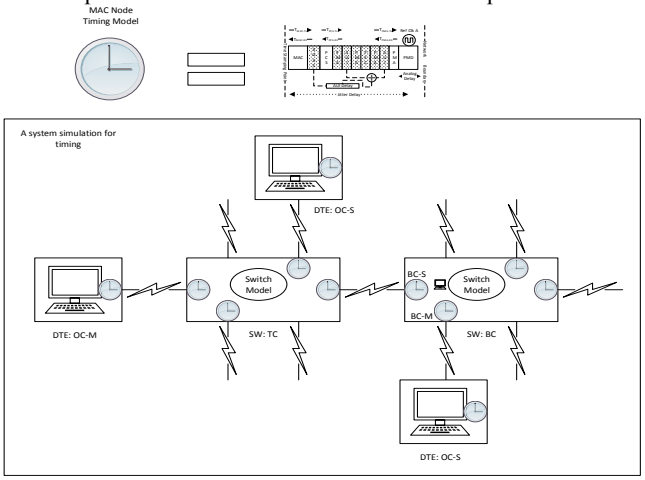

Figure 2: Usage of timing models in the simulation of a system

2.1 Architectural components

As shown in the Figure 3, the key components from 802.3 architecture for the MAC throughputs ranging from 1Gbps to 100Gbps and beyond [7,8,9] are MAC, Reconciliation (RS), PCS, PMA and PMD. The article treats RS as part of the MAC, and PMD as part of the link delay, given their deterministic delay characteristics. Thus, the key components of focus for deriving the timing models are MAC, PCS and PMA across the throughputs of the MAC.

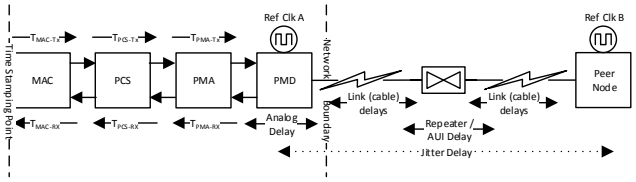

Figure 3: Key components in the 802.3 architecture for PTP

At throughputs of 10G and beyond [8,9], the 802.3 architecture introduces new attachment unit interfaces (AUI), whose delays also need to be accounted as accurately as possible in order to achieve the highest possible accuracy in the entire networking system. The AUIs are dealt with, in their respective sections 3.2.3 and 4.2.3 below.

2.2 Timing model

As shown in the Figure 2, the timing model for a MAC node up to the pin or entry point to the network (without any AUI in between) on the transmit path (Tx) is different from that on the (Rx) path. The differences include exclusive operations on the Rx path such as word aligner, alignment marker logic, deskew logic, and idle insertion/deletion due to the ppm differences (between the recovered clock and the system clock). There is also difference in points of timestamping along with the different delays due to the clock jitter and the data jitter, on the Tx and the Rx paths.

The generic model consists of calculating the net delay from the point of timestamping (where we detect the SoF -the first bit after the SFD) till the PMA pin and adding (Tx path) or subtracting (Rx path) this delay to/from the timestamp obtained from the ToD counter (ToD_{Tx}, ToD_{Rx}) at the point of timestamping, to arrive at the net timestamps (TS_{Tx}, TS_{Rx}) to be used in the protocol computations. The $T_{MiscD\text{-}Tx}$ & $T_{MiscD\text{-}Rx}$ represent any unaccounted delays such as asymmetry, that are measured later. T_{SclkDJ} represents the delay due to the jitter difference between system clock and the transmitted or received SoF, whereas T_{RclkDJ} the delay due to the jitter difference between recovered clock and the received SoF.

$T_{Tx} = T_{MAC\text{-}Tx} + T_{PCS\text{-}Tx} + T_{PMA\text{-}Tx} + T_{SclkDJ} + T_{MiscD\text{-}Tx}$

$T_{Rx} = T_{MAC\text{-}Rx} + T_{PCS\text{-}Rx} + T_{PMA\text{-}Rx} + T_{RclkDJ} + T_{SclkDJ} + T_{MiscD\text{-}Rx}$

$TS_{Tx} = ToD_{Tx} + T_{Tx}$

$TS_{Rx} = ToD_{Rx} - T_{Rx}$

The model due to AUIs is explained as follows. The AUIs form two kinds of components – an extender component such as XGXS and a Phy module that connects to the cable. In case of extender components with AUI on both ends, the XGXS might implement deskew logic on the receive path at both ends as shown in Figure 1. The delays due to the deskew logic vary from reset to reset. However, no layer 2 operations such as packet parsing or modification is performed and thus the delays in the extender component will need to be accounted in the PTP packets in the nearest MAC node or Phy node, where the packet modification is possible for 1-step operation. For 2-step, the followup packet will be sent later taking all these delays into account. In either case, the delay accounting will be in a location different from the extender, where the PTP experiences the actual delay. The extender delay model can be summarized as

$T_{Tx\text{-}AUI} = T_{PCS\text{-}Tx} + T_{SclkDJ} + T_{MiscD\text{-}Tx}$

$T_{Rx\text{-}AUI} = T_{PCS\text{-}Rx} + T_{RclkDJ} + T_{SclkDJ} + T_{MiscD\text{-}Rx}$

Most of the Phys implement PCS and PMA functionality without packet parsing and modifying feature. In this case, like the case of extenders, the delay accounting will be in a location different from the Phy. However, some advanced Phys have implemented the packet parsing & modification feature to accommodate PTP 1-step operation and thus the PTP delays are accounted within the component. The Phy delay model can be summarized as below. Depending on the type of implementation, delays like $T_{MAC\text{-}Tx}$ and $T_{MAC\text{-}Rx}$ can be present.

$T_{Tx\text{-}Phy} = T_{PCS\text{-}Tx} + T_{PMA\text{-}Tx} + T_{SclkDJ} + T_{MiscD\text{-}Tx}$

$T_{Rx\text{-}Phy} = T_{PCS\text{-}Rx} + T_{PMA\text{-}Rx} + T_{RclkDJ} + T_{SclkDJ} + T_{MiscD\text{-}Rx}$

As noted above, the delay models for AUI and Phy can be taken as a subset of the delay model of the MAC node. Thus, henceforth, the article emphasizes on the delay model for the MAC node.

3. Ethernet links up to 10G

The MAC throughputs up to 10G are single-lane single link, except the XAUI(10G Attachment Unit Interface) termination for 10G, which has 4 lanes. However, the first bit after the SDF delimiter is always on lane 0 and thus the timestamping reference point is lane 0 only as the first bit never falls on other lanes. Thus, the timing model up to 10G can be treated as single-lane single link system.

3.1 Architectural components

The key architectural components are MAC, PCS, PMA for the MAC node's timing model up to 10G, with additional XGXS component only for 10G MAC node, as shown in Figure 4.

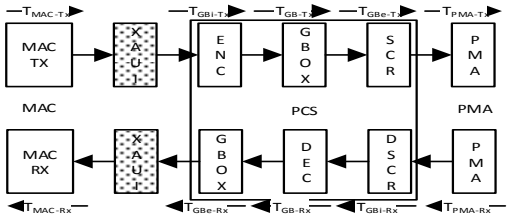

Figure 4: Key components for single lane Ethernet up to 10G for PTP

3.2 Timing model

3.2.1 Tx path

The key design components on the Tx, as shown in the Figure 4, are MAC where the timestamping unit resides, the encoder block, the gearbox which is typically a phase compensation FIFO (the read and the write clocks are derived from the same PLL) and the PMA. Occasionally, the gearbox FIFO can have its write and read clocks from different PLLs, in which case, the FIFO is a clock compensation FIFO.

$T_{MAC\text{-}Tx}$ is the overall delay between the timestamping point and the PCS ingress point. The timestamping clock can be an integer multiple of the MAC clock for more accurate timestamp value or an unrelated clock, in which there might some accuracy loss due to clock domain crossing. In any case, the quality of the timestamping clock should be as best as possible to result in the best accuracy.

$T_{PCS\text{-}Tx}$ is the total delay incurred by the SoF in the PCS block on the Tx path. The major contributing factors are the pipeline or memory delays up to the gearbox FIFO (T_{GBi}), gearbox FIFO (typically a phase compensation FIFO) depth (T_{GB}) and the pipeline delays from the FIFO output to the PMA ingress in addition to the scrambler latency (T_{GBe}). Further, the SoF gets shifted by another 2 bits due to 64/66 encoding, which needs to be accounted in either T_{GBi} or T_{GBe}. The T_{GB} can vary from reset to reset and thus we need to have a mechanism to measure the FIFO depth. In case of multiple GB FIFOs, the T_{GB} represents the sum of all such delays and so is the case with T_{GBi} and T_{GBe}. Thus,

$T_{PCS\text{-}Tx} = T_{GBi} + T_{GB} + T_{Gbe}$

$T_{PMA\text{-}Tx}$ is the constant delay offered to the SoF from the PMA ingress to the egress point to the network, due to serialization. T_{SclkDJ} represents the effective delay of the SoF on its traversal to the egress pin of the PMA due to the differences in clock and data jitter, on the transmit path.

3.2.2 Rx path

The key design components on the Rx, as shown in the Figure 4, are MAC where the timestamping unit resides, the decoder block, the gearbox which is typically a clock compensation FIFO (the read and the write clocks are derived from the different PLLs) and the PMA. Occasionally, the gearbox FIFO can have its write and read clocks from the same PLL, in which case, the FIFO is a phase compensation FIFO. The analysis below follows the flow of SoF propagation.

$T_{PMA\text{-}Rx}$ is the constant delay offered to the SoF traversal from the network ingress to the PMA egress point to the PCS due to de-serialization.

$T_{PCS\text{-}Rx}$ is the total delay incurred by the SoF in the PCS block on the Rx path. The major contributing factors are the pipeline or memory delays up to the gearbox FIFO in addition to the descrambler latency (T_{GBi}), gearbox FIFO (typically a clock compensation FIFO) depth (T_{GB}) and the pipeline delays from the FIFO output to the PMA ingress (T_{GBe}). A mechanism to measure the FIFO depth is required to account for the idle insertion and deletion if the FIFO is a rate-matching or clock-compensation FIFO and account for the depth variation from reset to reset in all the cases. Further, the SoF gets shifted by another 2 bits due to 66/64 decoding, which needs to be accounted in either T_{GBi} or T_{GBe}. In case of multiple GB FIFOs, the T_{GB} represents the sum of all such delays and so is the case with T_{GBi} and T_{GBe}. Thus,

$T_{PCS\text{-}Rx} = T_{GBi} + T_{GB} + T_{Gbe}$

$T_{MAC\text{-}Rx}$ is the overall delay between the PCS egress point to the MAC and the timestamping point inside the MAC. The delay accounting is very similar to that on the transmit path.

T_{SclkDJ} and T_{RclkDJ} represents the effective delay of the SoF from the network ingress pin of the PMA till its timestamping point on the receive path due to the differences in clock and data jitter in both the clock domains of recovered and system/user clocks. In case of only recovered clock being used till the timestamping point, only T_{RclkDJ} is considered.

3.2.3 Impact of AUIs

$T_{MiscD\text{-}Tx}$ or $T_{MiscD\text{-}Rx}$ can account for other measured delays such as asymmetry in the network and/or the XGXS and the Phy delays on the Tx or Rx path, when the XGXS and the Phy do not have the capability to update the 1-step PTP packet on its way. Another key aspect here is that the delay of de-skew logic and the delay due to the FIFO may vary from reset to reset and we need to have an appropriate system mechanism to include these delays in the overall path of the SoF to the network. The advanced XGXS and Phy components might be PTP-1588 friendly in the sense that they might either update the 1-step PTP packet with the respective delay incurred or at the least have a mechanism to measure these delays variable from reset to reset and for the system to read these delays to incorporate them later in $T_{MiscD\text{-}Tx}$ or $T_{MiscD\text{-}Rx}$. However, in case of older components, suitable measurement techniques may be combined with system's capability to assimilate the measured delays from these techniques.

4. Multi-lane Ethernet links

The 1588 timing models become more challenging in multi-lane Ethernet because of the introduction of PCS virtual lanes and their mapping to PMA lanes.

4.1 Architectural components

Multilane architecture calls for transmitting the data on multiple lanes though the data belongs to single link. It is highly likely that the skew experienced by each lane is not the same. This results in multilane skew and calls for mechanism to de skew the data. The 802.3 protocol uses Alignment markers to achieve the de-skew. As the data rate increases the physical lanes are more prone to errors and hence the Forward Error correction(FEC) mechanism is recommended for higher data rates though the FEC is out of the scope for the article. Thus, the key components of the Multilane Ethernet are MAC, PCS, and PMA as shown in Figure 5. The Alignment maker insertion and de skew logic are part of the PCS.

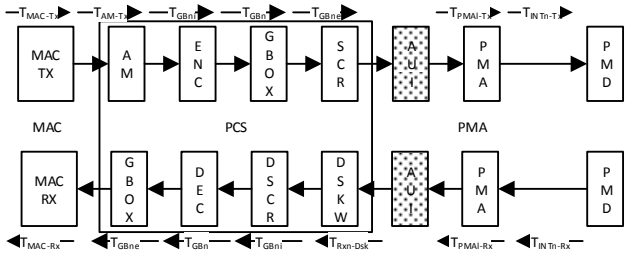

Figure 5: Key components for multi-lane Ethernet of 40G/100G and beyond for PTP

4.2 Timing model

4.2.1 Tx path

Typically, the timestamping is done in the MAC and thus$T_{MAC\text{-}Tx}$ delay accounting is as explained in 3.2.1.

$T_{PCS\text{-}Tx}$ accounting is more elaborate due to the PCS lanes. With the first bit after the SoF, to be considered for timestamping, falling on any PCS lane, we need to consider each PCS lane's physical delay instead of any average delay mechanism for very high accurate timestamping, given that all the PCS lanes may not offer identical delays to the SoF, despite overall alignment within the skew allowed by the protocol. We have 4 PCS lanes for 40G and 20 PCS lanes for 100G and thus have 4 and 20 instances of $T_{PCS\text{-}Tx}$ respectively though the actual implementation might reduce the number of instances due to the same physical path being run at high frequencies to service multiple lanes.

$T_{PCSn\text{-}Tx} = T_{GBni} + T_{GBn} + T_{GBne}$, where n = 0 to 3 for 40G and n = 0 to 19 for 100G.

The multilane Ethernet is subjected to inter-lane skew and the problem is addressed by the introduction of alignment marker (AM) logic. However, the AM introduces complexity in the timestamps as explained. The data from the MAC is 64/66bit encoded and 66b PCS blocks are formed. The blocks are distributed into PCS lanes. It is possible that each PCS lane may correspond to a physical lane or multiple PCS lanes can form a physical lane.

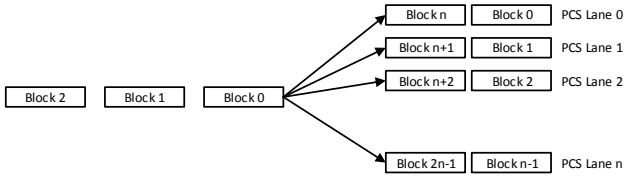

Figure 6: Block distribution to PCS lanes

The data from the 64/66b encoder is arranged into PCS lanes using Round-Robin distribution. First n blocks are transmitted out of the system followed by the next n via n PCS lanes. For the de-skew operation at the receiver, the AMs are inserted between the data blocks across all the lanes as shown below, after deleting equal number of idle control characters or sequence ordered sets. The PTP packet's timestamp will incur a delay of 66b if the alignment marker is inserted after the packet is timestamped. In fact, multiple PTP packets need to be adjusted for this 66b delay depending on the pipeline capacity between the MAC's timestamping point and AM logic insertion point. Similarly, if the alignment marker logic is deleting idle characters or sequence

ordered sets for AM insertion later, the PTP packet's timestamp needs to be advanced by as many bits deleted. A combination of 66b delay due to AM insertion and time advancing due to idle character/sequence ordered sets deletion is possible and thus a close coordination between AM logic in the PCS and the timestamping point in the MAC is called for. Given that the AM logic is common across the lanes, $T_{AM\text{-}Tx}$ characterizes the delay due to the AM logic.

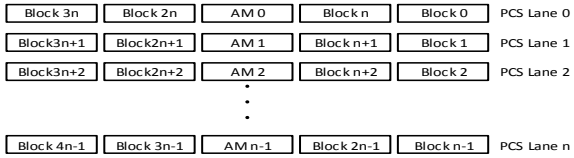

Figure 7: Alignment Marker insertion in the PCS lanes

Thus, more accurate modelling of the PCS delay in case of mult-lane links is $T_{PCSn\text{-}Tx} = T_{GBni} + T_{GBn} + T_{GBne} + T_{AM\text{-}Tx}$, where n = 0 to 3 for 40G and n = 0 to 19 for 100G.

Unlike the single-lane link, the SoF can leave on any of the multiple lanes and thus delay due to the PMA would be specific $T_{PMAl\text{-}Tx}$ where l is the lane on which the SoF is transmitted, given the extreme difficulty in the hardware to maintain symmetrical paths for all the lanes due to configuration capability of mapping any PCS lane to any PMA lane. However, when the number of PCS lanes is equal to that of PMA lanes, the delays due to the PMA lanes can be treated equal. In case where the number of PCS lanes is not equal to PMA lanes, each PMA lane carries multiple PCS lanes, resulting in interleaving of blocks from multiple PCS lanes. For high accuracy, we need to factor in the delay experienced by SoF due to interleaving. The delay due to PMA becomes $T_{PMA\text{-}Tx} = T_{PMAl\text{-}Tx} + T_{INTn\text{-}Tx}$ (Delay for PCS lane n due to interleaving at PMA).

T_{SclkDJ} analysis remains the same as 3.2.1 though there may be multiple instances of the T_{SclkDJ} due to multiple clocks (because of multi-lane Phy) and the respective datapaths.

4.2.2 Rx path

The key design components on the Rx, as shown in the Figure 5, are MAC where the timestamping unit resides, PCS - the gearbox which is typically a clock compensation FIFO (the read and the write clocks are derived from the different PLLs), the decoder block, PCS lane demultiplexing, alignment block where lane alignment and deskew across the lane occur, descrambler, and the PMA. Occasionally, the gearbox FIFO can have its write and read clocks from the same PLL, in which case, the FIFO is a phase compensation FIFO. The analysis below follows the flow of SoF propagation.

The SoF can enter on any of the PMA lanes and thus delay due to the PMA would be specific $T_{PMAl\text{-}Rx}$ where l is the lane on which the SoF is received, given the extreme difficulty in the hardware to maintain symmetrical paths for all the lanes due to configuration capability of mapping any PCS lane to any PMA lane, and receive nature of latching onto any bit to begin with – not necessarily in the transmitted order, both of which can lead to reordering of the PCS lanes as compared to the transmit path; and clock & data recovery. Even for the implementation where the number of physical lanes is equal to the number of PCS lanes (for example 40G) the PMA delay for each PCS lane can be different due to the possible individual nature of clock data recovery. However, in case of the implementation where the number of PCS lanes is not equal to number of Physical lanes, each physical lane receives multiple PCS lanes. When recovering the ingress interleaved data, the PMA may reorder the data received and this may result in additional delay to the SoF inside the PMA. More importantly, this bit reordering delay might vary from reset to reset. For high accuracy modelling the delay due to the reordering need to be accounted for. Thus, $T_{PMA\text{-}Rx} = T_{PMAl\text{-}Rx} + T_{INTn\text{-}Rx}$ (Delay due to reordering for PCS lane n).

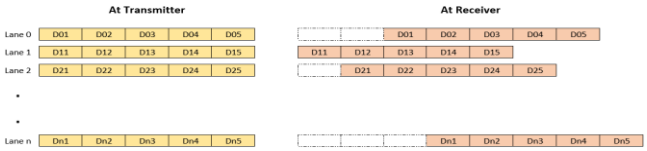

Figure 8: Variable skew among the PCS lanes on Rx

As evident from the diagram shown above, it is possible, after the individual lane alignment, that there is variable skew among the individual PCS lanes in a MLD system when observed at the receiver.

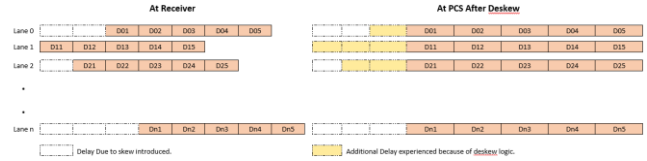

Figure 9: De-skew logic for the PCS lanes on Rx

So the delay experienced by the SoF after entering the system is determined by the lane on which the SoF has fallen and the tap position of the particular lane in the deskew logic. For example, as shown in Fig 9, if the SoF falls on lane0, the delay due to de-skew logic is one clock cycle, whereas if the SoF falls on lane1 the delay due to de-skew logic is 3 clock cycles. The deskew tap position and thus the effective delay can vary from reset to reset or for every trigger of the deskew operation. In some implementations, the data width of the transceivers (PMA) may differ from the data width of the PCS, which may in turn differ from the data width of the MAC. Thus, multiple gearboxes may be implemented as needed, resulting in additional delay and necessitating the gear box (FIFO) depth measurement logic for each of the gearbox in each PCS lane. While, the alignment block replaces the alignment markers with idles, the gear box may act as a clock-compensation FIFO inserting, deleting idles to rate-match ingress and egress data rates and the FIFO measurement logic will take care of it. The descrambler delays can be absorbed in the ingress delay to the relevant gearbox. In addition to this there is a delay introduced by the FEC logic in the RX direction which is not described in this paper. The timing equation for the Rx PCS, thus, is as below:

$T_{PCSn\text{-}Rx} = T_{Rxn\text{-}Dsk} + T_{GBni} + T_{GBn} + T_{GBne}$ where

$T_{Rxn\text{-}Dsk}$ = Delay due to the de-skew operation in Rx.

$T_{MAC\text{-}Rx}$, T_{SclkDJ} and T_{RclkDJ} remain the same as explained in the section 3.2.2. There may be multiple instances of the T_{SclkDJ} and T_{RclkDJ} due to multiple system clocks and multiple recovered clocks (because of multi-lane Phy as explained in the section 4.2.4) and the respective datapaths.

4.2.3 Impact of AUIs

Unlike XAUI where the SoF is expected on lane 0, the XLAUI for 40G and CAUI for 100G can have the SoF on any lane. In such cases, wherever the CAUI interface has different number of lanes from ingress to egress, the data from the ingress PMA lanes need to be first segregated into the PCS lane buffers first and reorder

them into the egress PMA lanes. Thus, these components need to have the relevant buffer depth measuring logic in addition to the logic that can calculate the delay from reset to reset for the possible variation in the depths. Further, only a few vendors provide components with AUI that are PTP 1-step friendly in the sense that they parse the packet and update the packet with the new timestamp and the new CRC before the packet is bit-muxed onto egress PMA lanes. A few of the other components might not offer PTP 1-step support but have built-in logic for measuring delays. For components without any measurement logic, external measurement techniques may be used, preferably reset to reset, for delay calibration and later usage by incorporating them in $T_{MiscD\text{-}Tx}$ or $T_{MiscD\text{-}Rx}$. Thus, for non-PTP friendly components, which can be many due to possibly many AUI interfaces (XLAUI, CAUI) in between, we need to have a proper mechanism at the system level to account these delays in the nearest MAC node as $T_{MiscD\text{-}Tx}$ or $T_{MiscD\text{-}Rx}$, where the packet's timestamp is updated.

4.2.4 Impact of PMDs

Unlike 10G Ethernet links, the 40G, 100G Ethernet links can be 4 lanes of electrical backplane (KR4), 4 or 10 lanes of shielded balanced copper cabling (CR4 or CR10), 4 or 10 lanes of multi-mode fiber (SR4 or SR10), one lane of single mode fiber (FR), and 4 WDM lanes on single mode fiber (LR4, ER4). In case of multi-lane Phys, any delay variations (the delay is calculated as part of the protocol calculations) due to mapping from PMA lanes to the Phy lanes, may be considered in $T_{MiscD\text{-}Tx}$ or $T_{MiscD\text{-}Rx}$ for achieving improved accuracy. In fact, the implementation of PCS, PMA and PMD might be within the tolerable limits of skew in such a way that the deskew logic might not require to deskew the lanes (i.e., overall skew is within a clock-cycle of the deskew logic). Still, the skew delays, if known, need to be accounted at the system level for the highest PTP accuracy. The unknown skew will contribute to the intrinsic inaccuracy in the system.

5. Timing Model

From the sections 3 and 4, the generic delay calculation at a MAC node, can now be defined as below. Statistically, the model denotes the mean values. However, by including a matrix of standard deviation values for each parameter below for the associated conditions (such as Process, Voltage, & Temperature), more exact computation of the timestamps is possible upfront in simulations. These values can later be correlated with the values from the system validation in order to improve the timing models with enhanced parameters for better convergence between simulation and validation down the line.

$$T_{Tx} = T_{MAC\text{-}Tx} + \boxed{T_{GBnt} + T_{GBn} + T_{GBne} + T_{AM\text{-}Tx}}^{T_{PCSn\text{-}Tx}} + \boxed{T_{PMAt\text{-}Tx} + T_{INTn\text{-}Tx}}^{T_{PMA\text{-}Tx}} + T_{SclkDJ} + T_{MiscD\text{-}Tx}$$

$$T_{Rx} = T_{MAC\text{-}Rx} + \boxed{T_{Rxn\text{-}Dsk} + T_{GBnt} + T_{GBn} + T_{GBne}}^{T_{PCSn\text{-}Rx}} + \boxed{T_{PMAt\text{-}Rx} + T_{INTn\text{-}Rx}}^{T_{PMAn\text{-}Rx}} + T_{RclkDJ} + T_{SclkDJ} + T_{MiscD\text{-}Rx}$$

"n" represents the PCS lane number and "l" represents the physical lane number.
n = 0 to 3 for 40G and n = 0 to 19 for 100G and n = 0 for 10G.
The values $T_{AM\text{-}Tx}$, $T_{INTn\text{-}Tx}$, $T_{INTn\text{-}Rx}$, $T_{Rxn\text{-}Dsk}$ are 0 for 10G

Figure 10: Timing Model Equation

6. CONCLUSIONS AND FUTURE WORK

The article details mechanisms to arrive at timing models for single-lane or multi-lane PTP-1588 systems. Given the possible proliferation of multi-lane Ethernet systems down the line, these timing models are hoped to show the path for the highest accuracy for multi-lane systems, helping the companies and the vendors to exchange accuracy information for better decision making upfront and thus improved cost-effectiveness.

The future work may focus on standardizing format for the timing models. These timing models may further be enhanced to include additional architectural components such as FEC and hitherto unknown delay-contributing factors. Additional resolution of the timing parameters in the models, in terms of their contribution to the constant and dynamic time errors [11], will help us assess their impact in the system with improved lucidity.

7. ACKNOWLEDGEMENTS

We sincerely thank Si Xing, Seng Kuan, David Mendel, Nigel Gulstone, Rajiv Kane & Hoss Rahbar, all from Intel, for allowing us to engage with them for different 1588 requirements.

8. REFERENCES

[1] IEEE Standards Association. 2008. "1588-2008-IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems". [Online]. Available: http://standards.ieee.org/findstds/standard/1588-2008.html . [Accessed 3 April 2016].

[2] Chandra Mallela, Yu Ying Choo, and Vince Bridgers. 2016. "Building a 1588 System Solution – Key learnings". ISPCS: Stockholm, Sweden.

[3] T. Mayurama et al. 2015. "Time Synchronization for Real-Time Control Systems without PTP-enabled Switches". ISPCS: Beijing, China.

[4] Small Cell Forum, 075.03.01. "Synchronization for LTE small cells. ". [Online]. Available: http://www.smallcellforum.org . [Accessed 3 April 2017].

[5] Infinera. "Evolving Moble Backhaul to support LTE-A and 5G". [Online]. Available: https://www.infinera.com/wp-content/uploads/2017/02/infinera-an-mobile-backhaul-support_lte-a_5g.pdf . [Accessed 3 April 2017].

[6] David Mendel, Herman Schmit, Divya Vijayaraghavan. 2013. "Packet Arrival Time in 1588 for 40GE/100GE". ISPCS: San Francisco, USA.

[7] IEEE. "IEEE standard for Ethernet Section Three: 1000Mb/s". [Online]. Available: http://www.ieee802.org/3/ . [Accessed 3 April 2017].

[8] IEEE. "IEEE standard for Ethernet Section Four: 10Gb/s". [Online]. Available: http://www.ieee802.org/3/ . [Accessed 3 April 2017].

[9] IEEE. "IEEE standard for Ethernet Section Six: 40Gb/s, 100Gb/s". [Online]. Available: http://www.ieee802.org/3/ . [Accessed 3 April 2017].

[10] D. Mills, J. Martin, J. Burbank, and W. Kasch (June 2010). *Network Time Protocol Version 4: Protocol and Algorithms Specification*. IETF. RFC 5905. Retrieved August 01, 2016.

[11] International Telecommunication Union. Network limits for time synchronization in packet networks. Online: https://www.itu.int/rec/T-REC-G.8271.1-201308-I/en . Retrieved Aug 01, 2016.