Strategies for Deployment of Accurate Time Information using PTP within the All-IP studio

Thomas Kernen[1], Nikolaus Kerö[2]

[1] Cisco Switzerland and [2]Oregano Systems Austria

ABSTRACT

As the industry transitions from SDI based production to an all-IP studio environment progresses, some of the finer points related to a smooth migration such as Time & Sync are capturing more attention from the early adopters.

Phase & frequency alignment of baseband signals are a critical element in media production. In the IP world, the required functionally is delivered via the IEEE 1588 Precision Time Protocol (PTP) specification. Whilst having enabled many industries to transfer their synchronisation requirements via PTP to the IP centric environment, special care needs to be taken for each specific industry and its specific constraints.

This paper draws on the extensive research the authors have carried out on the use of PTP for the media production industry. It summarises their work on areas such as how PTP aware vs. non-aware networks behave under load, IP Quality of Service for PTP messages and Grand Master redundancy models. Concluding with the impact of design considerations, network architecture constraints and device requirements for a successful all-IP synchronised media production facility.

INTRODUCTION

With the advent of distributed systems where every node has a sufficient amount of local resources to perform given tasks partly independently from all other units, reliable and timely data communication became a mandatory requirement together with the need for a common notion of time or at least a method to convey a common frequency to all nodes. Lacking a unified communication mechanism fulfilling all demands, every application domain independently developed legacy systems tailored to their specific needs. In industrial automation, for example, a number of competing field bus systems were prevalent for more than 20 years. In the broadcasting industry the great majority of studios still operate entirely with an SDI based infrastructure. Common to all those communication systems was their ability to convey both data and some kind of time and/or frequency information to all nodes. These systems were highly optimized thus requiring dedicated hardware and firmware to be developed, maintained, and updated. This was also their most crucial shortcoming, as they could not keep pace with the ever growing demand for bandwidth combined with increasingly stringent requirements on low latency data transfers. Whenever incremental improvements, by replacing only a number of core hardware units, were not a feasible solution, the entire infrastructure had to be replaced in a time consuming and costly process.

In recent years, Ethernet (and IP) based communication systems have significantly matured from their pure office IT based origins. Nowadays, they are being considered a cost effective yet highly powerful alternative to nearly any legacy system in the market, thus replacing these gradually for nearly any application domain. Consequently, the broadcasting industry is moving towards the all-IP studio as well. Leaving aside its many

undisputed advantages over all legacy systems, Ethernet has one important property to consider: It's an inherently asynchronous medium, to be more precise, data transfer is only synchronised on a per link basis between two adjacent nodes, thus Ethernet does not provide a common frequency on its own via the physical layer alone. This turned out to be a shortcoming only at first sight. Legacy systems like SDI generally are limited to provide only a common frequency, as opposed to packet based communication mechanisms which can transport absolute time information highly accurately as well using dedicated protocols: NTP, the network Time Protocol and PTP, the Precision Time Protocol being the two most commonly used.

Version 2.0 of PTP – defined in the IEEE1588-2008 [1] standard – turned out to be ideally suited for highly accurate clock synchronisation over local area networks and even shows remarkable performance of wide area networks. This standard was written with a broad view for a variety of possible use cases without focusing on a specific application. This was accomplished via a simple yet robust design while providing wide operating ranges for all relevant parameters to choose from. As PTP can easily and precisely be tailored to specific requirements via a PTP profile, it was rather quickly adopted by all relevant industries. Telecom [2, 3, 4], the power industry [5], and finally the broadcasting industry [6, 7] all defined their own PTP profiles and subsequently started to deploy PTP on a large scale, most of them as their only means of synchronisation.

The following sections will present a brief introduction of the basic principles of PTP complemented with the main sources of error to reckon with and efficient methods to cope with them. Special focus will be put on deployment strategies for large networks followed by a discussion of PTP redundancy.

BASIC PRINCIPLES OF TIME TRANSFER WITH PTP

The Precision Time Protocol relies on a strict hierarchical principle for time transfer. At any given point in time only one single node (often referred to as Ordinary Clock – OC) within a network acts as a PTP Grandmaster while all other nodes revert to PTP Slave mode. The Master distributes its time periodically by sending out Sync messages. They contain the absolute time of the Master at the moment when the packet is sent. All other nodes denote the time at which they received these messages on their respective local clocks. The difference between these two timestamps equals the offset of the two clocks plus the transmission time of the message. If the latter is known, the Slave can correct the offset of its clock with respect to the Master. In order to do so, the Slave transmits a Delay_Request message to the Master, denoting the send time. The Master draws a timestamp upon receiving and returns this value to the Slave via a corresponding Delay_Response message. The PTP message flow with the underlying equations for delay and offset is shown in figure 1. By default, all PTP messages are transmitted as multicast, however, PTP provisions unicast message transport mechanisms as well.

PTP uses the international atomic time TAI as its timescale providing a monotonic flow of time throughout the network as opposed to NTP which relies on UTC where leap second events have to be communicated and handled correctly by all nodes. To facilitate synchronising secondary (wall) clocks to a local time scale based on UTC, PTP does provide information on the current amount of leap seconds together with data on a pending leap second event as a 24h pre-warning.

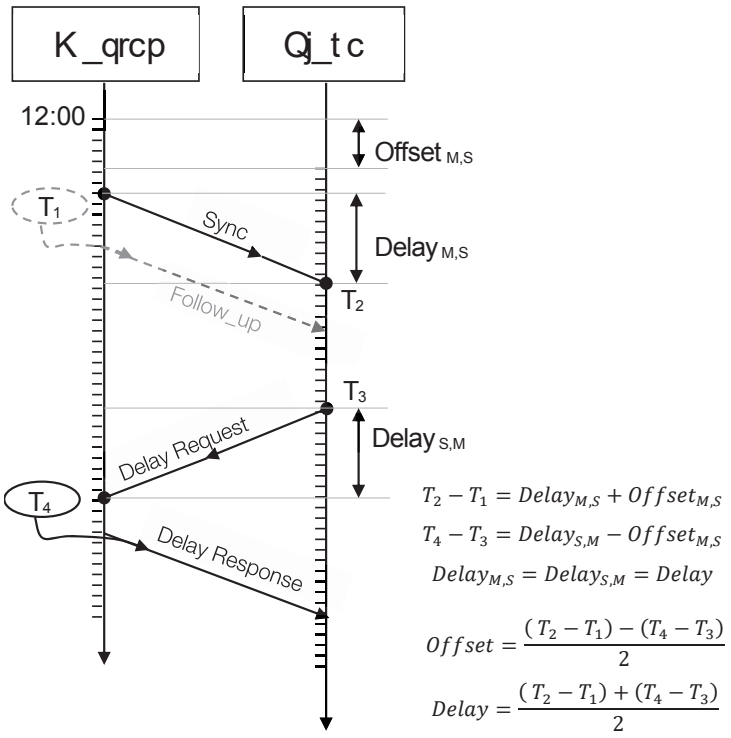

Figure 1 PTP Message flow

Master Election Process

One of the most important features of PTP is its ability to automatically select a node to become a Master without any user interaction and no or minimal configuration efforts. Every PTP Master continuously advertises the quality of its clock by sending PTP Announce messages. These contain a number of parameters specifying the clock, the most important being the Clock Class. If no Master has been elected i.e. during the start-up phase on an entire network or if the currently active master fails, the master election process is triggered on every Slave. The nodes able to assume Master role, start sending Announce messages themselves. All nodes will evaluate these messages by comparing their clock quality parameters according to a defined precedence. As a result, the node with the highest quality is elected to become the new Master, whilst all other nodes start synchronising to it. It is worth mentioning that a PTP network will always elect the best clock to become its Master. If a new node is attached to the network, it first of all listens for incoming Announce messages sent by the currently active master. If it concludes that its clock has a higher quality, it can take over as a master by starting to send Announce messages as well, causing the current master to eventually back off (i.e. revert to passive or slave state) and all other nodes to switch to it. The PTP Master election process is referred to as Best Master Clock Algorithm (BMCA).

PTP Accuracy

To reach sub-µs accuracies, all four timestamps have to be drawn at the precise moment when a PTP event packet is actually sent or received. This requirement rules out any purely software based solutions as used for NTP, where the timestamp is drawn from the host CPU while it assembles a new packet. After having completed this task, the packet traverses through several layers within the operating system and the respective hardware module before it is eventually sent out over the physical medium. Any of these tasks can be interrupted at any time for an unknown duration. These software-induced latencies may well vary in the range of several ms representing a strict limitation on the achievable accuracy. These barriers can only be overcome by drawing timestamps in hardware via

specific modules placed as close as possible to the physical medium. They are usually attached to a high-resolution clock driven by a high quality crystal oscillator and can provide time stamping accuracies well below 10ns. If the master is in not capable of actually inserting the send timestamp into the sync message while sending it, it will forward it in a Follow_Up message sent directly after the corresponding Sync message as shown in figure 1.

The formulae used by every Slave to compensate for its offset (see figure 1) take two assumptions about the delay implicitly for granted: It has to be constant and identical in both directions. In modern switched networks, packet forwarding is optimized for maximum throughput rather than constant forwarding time. Therefore, the transmission time varies on a per packet basis; this effect is referred to as Packet Delay Variations (PDVs). The second assumption is by far less critical and is usually met to a very high degree (i.e. sub 50 ns or even sub 10 ns) in most cases by modern high- grade network devices. However, asymmetric transmission delays will inevitably introduce a residual offset, which the slave neither can detect nor compensate for. Therefore, they need to be measured using out-of-band methods at least prior to deployment. They are more likely to occur in networks with highly asymmetric loading or in cases where multicast is used for Sync messages while Delay_Request message are sent in unicast (mixed mode PTP). Their influence needs to be analysed under real world operating conditions.

PTP Aware Network devices

Prioritising PTP traffic can mitigate the effect of PDVs on synchronisation [8]. However, this can only be done to a certain extent. Rather than simply requesting the forwarding time to remain constant under any loading condition, IEEE1588 has specified two different types of PTP aware network devices, Transparent Clocks (TCs) and Boundary Clocks (BCs), respectively. The former measure the time it takes a switch to process and forward a PTP packet and relay that information to the slave. This is done by drawing timestamps at ingress and egress. Boundary Clocks, on the other hand, are active PTP devices taking part in the message exchange. The port to which the PTP Master is attached to will assume a Slave role synchronising the local clock of the BC to it. All other ports will assume Master role. Thus BCs can be used to build a hierarchical time distribution architecture. Their port assignment is by no means static or constant; they participate in the BMCA in the same way as all other nodes. As soon as a new Master is selected all their ports will switch to their respective new state building a multi-level hierarchy with the new Master at its top.

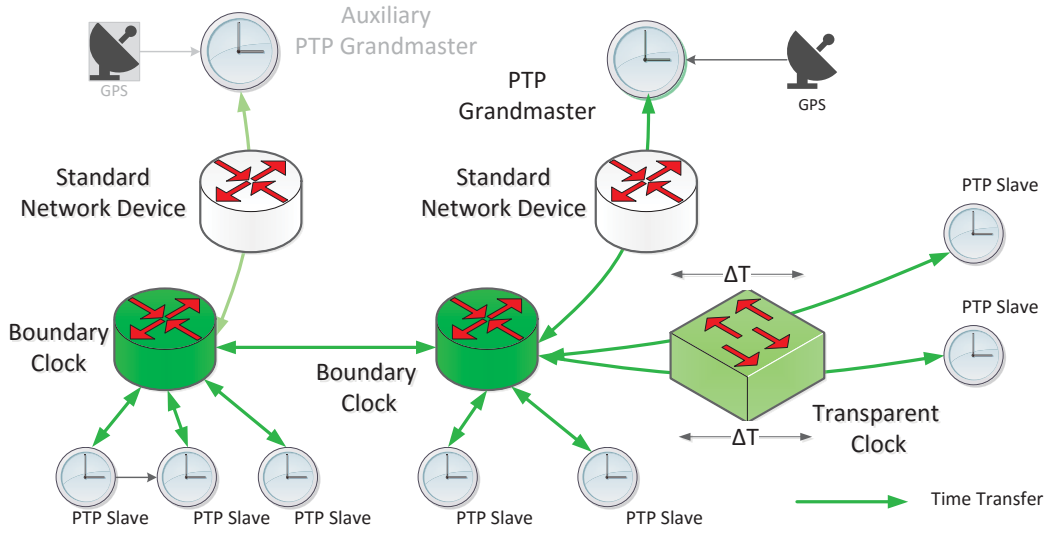

Figure 2 Typical PTP Network Structure

At first sight, TCs seem to be a straightforward solution for minimizing PDV effects on accuracy, because they require less configuration compared to BCs. However, with BCs the PTP traffic is kept local between two adjacent devices reducing the strain on the top most devices in a hierarchical multi-hop network most prominently on the PTP Grandmaster itself. A typical network topology using both standard and PTP network device is shown in figure 2.

PTP PROFILES FOR THE BROADCASTING INDUSTRY

In April 2015, SMPTE published the SMPTE ST 2059-2 [6] describing an IEEE1588 profile for the broadcasting industry, intended to serve all possible use cases. Thus it has to provide support for a variety of different application scenarios ranging from outside broadcast trucks and small studios to large and diverse broadcast facilities with all-IP based production flows. Common to all use cases is the demand for sub µs accuracy and the exclusive use of Layer 3 IP network protocols (IPv4 and IPv6 respectively).

An ST 2059-2 network may be built using PTP aware network devices, in doing so PTP Boundary Clocks as well as PTP Transparent Clocks are supported in any suitable combination. However, PTP network devices are not mandatory thereby permitting to operate PTP either with existing standard network components or mixed networks using PTP network devices at certain critical locations making best use of existing infrastructure and thus limiting the initial investment required to successfully deploy PTP.

Within a typical broadcasting environment, a significant number of mobile devices such as cameras need to be frequently disconnected and reconnected again. They have to be fully operational more or less instantaneously after being connected to the network. Therefore, ST2059-2 specifies a minimum lock time of 5 seconds for every node joining a PTP network, after which its offset to its PTP master has to remain within +/-500ns. The resulting implications on both the network and node design have been investigated in [8].

PTP REDUNDANCY

Generally, all nodes within a PTP network are capable of detecting a failure of the current Master. It is important to point out that there is only one condition to do so: the absence of Announce messages for a user definable time period. Only this event will trigger the BMCA to be executed at every node eventually leading to another node taking over as the new Master. At least two nodes being able to assume Master role have to be present. For many application domains this requirement is fulfilled by default, because all nodes will be capable of both roles. This is an efficient way to maintain a common notion of time even in the absence of an external time reference.

However, in many broadcasting applications, permanent locking to an external time traceable source is equally important. Furthermore, most nodes are not intended to become Masters, such as cameras, due to their limited local resources. They are therefore configured to operate as Slave only devices. Consequently, at least two dedicated Grandmasters have to be installed and configured accordingly. Typically, both of them will be attached to an external time reference like GPS yielding identical clock quality parameters, the BMCA will selected one to be the primary and the other will remain passive.

Any Master re-election takes time during which the whole network is left free-running. Furthermore, in case of a number of BCs being connected in a daisy chain, all local clocks need to re-settle to the new Master as well. This is an iterative process, because the clocks are effectively cascaded. However, the strain on the control loop of every end node is less stringent. It has to cope with a gradual change of the time information of the BC it is directly attached to rather than having to handle large jumps in offset and delay which can occur, especially, if the two Masters are not co-located. For enhanced redundancy reasons

these could well be placed at opposite end of a network.

It has to be kept in mind that the BMCA covers only a sub-set of possible failure conditions i.e. a total failure of the current Grandmaster or a broken network path. Transient or permanent loss of PTP event messages caused by fault conditions at the Master or a network device, while Announce messages are still generated and forward correctly will not trigger the BMCA. Nonetheless, any such error condition will disrupt the synchronisation causing all affected nodes to deviate from each other. Furthermore, sudden changes in the PDV caused by either overproviding of a network or some kind of malfunction will deteriorate the synchronisation accuracy and thus should be treated as error conditions as well. We performed a more in-depth redundancy investigation in [9].

Although there are no native provisions within PTP to detect or counteract these secondary failure conditions, by no means they preclude using PTP even in mission critical applications. As such use cases require in-depth monitoring of the whole network infrastructure to guarantee reliable media transmission in the first place, these tools need to be extended to cover PTP traffic as well. As a second step, extended redundancy methods can be applied for highly critical devices. This, however, requires dedicated PTP implementations on the slave side capable of detecting these conditions as well as setting appropriate counter measures.

EVOLUTION OF NETWORK ARCHITECTURES

Over the course of the last decade, network architectures such as those used for enterprise networks or datacentres have significantly evolved. The traditional network was built upon an Access/Aggregation/Core layer topology (see figure 3) due to the “North-South” data path that was common in between servers connected in the facility communicating with external parties (e.g.: another facility or the Internet).

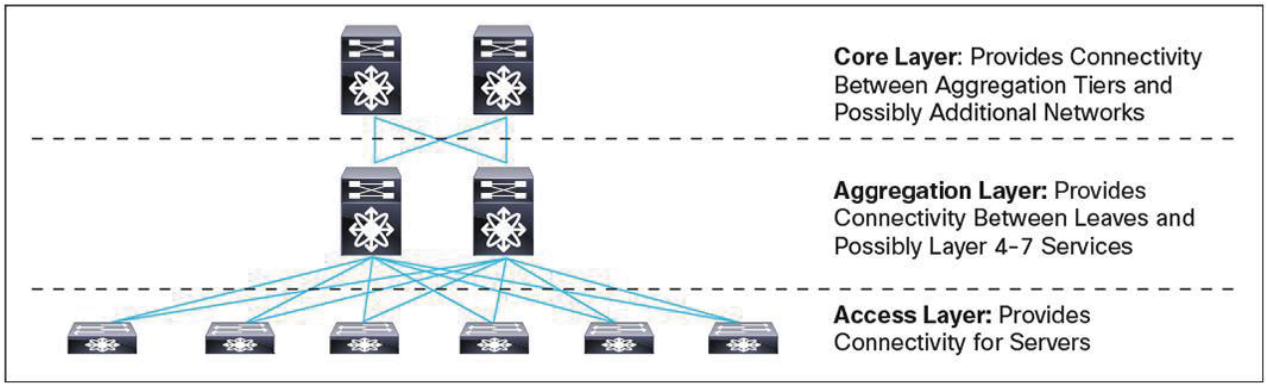

Figure 3: Traditional Core/Aggregation/Access Topology

As we move to the virtualisation of endpoints and applications shift from monolithic processes to distributed ones, the traffic patterns are now following an “East-West” model whereby most of the traffic is actually within the Datacentre itself. Therefore, the network architecture has collapsed from a 3-tier to a 2-tier model known as “Spine-Leaf” (see figure 4).

The Spine layer provides the bandwidth between the Leafs and the required redundancy by interconnecting Leafs to multiple Spines. The Leaf layer provides connectivity for the compute and/or storage elements within a given rack. Hence why Leafs are also referred to as “Top of Rack” (ToR) switches. This allows for distributing the network closer to endpoints thereby limiting the number of dedicated fibres and optics whose properties may be distance dependent (multi-mode vs. single mode optics).

This architecture delivers a short and consistent path between endpoints, thereby reducing packet delay variation and latency for time sensitive traffic such as media essence and PTP messages. Combined with PTP-aware network nodes such as Boundary Clocks, this offers PTP scaling capabilities whereby each node is communicating directly with the next hop network device and therefore is not overloading the Grandmaster with messages from all the endpoint slaves.

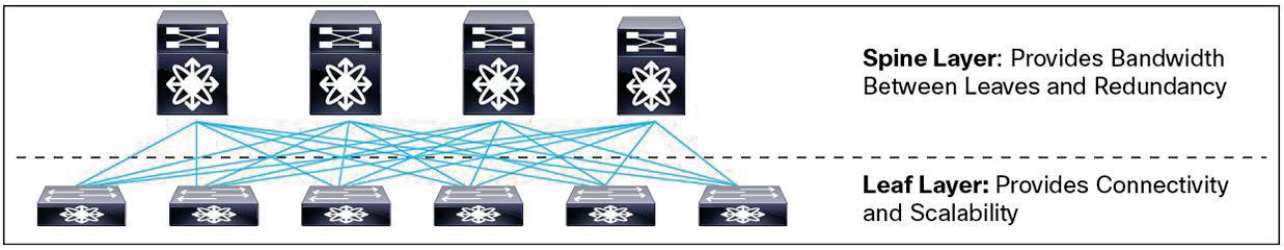

Figure 4: Modern Spine-Leaf Network Topology

Such networks are designed to be non-oversubscribed as required by media production, enabling linear bandwidth scalability with additional spine switches. This allows for planning growth beyond that of a centralised chassis switch approach causing a number of challenges, if migration is required in a live environment.

Large networks

Whilst many media production workflows will be covered by the Spine-Leaf architecture, there are use cases requiring spanning more hops than presented above. In such cases PTP must continue to meet the targets defined in the 2059-2 PTP profile. This has been proven possible in our research on optimising such large networks for high time accuracy. Models have been tested with up to 9 hops between the 2 OC nodes [10]. Figure 5 shows the respective test setup with 3x 3G-SDI traffic added in parallel to the PTP traffic. All nodes were non-PTP aware units except the two switches on the opposite edges of the network were PTP aware units. Figure 6 depicts the offset of the slave with respect to the Grandmaster. The results were obtained using default 2059-2 values for all parameters.

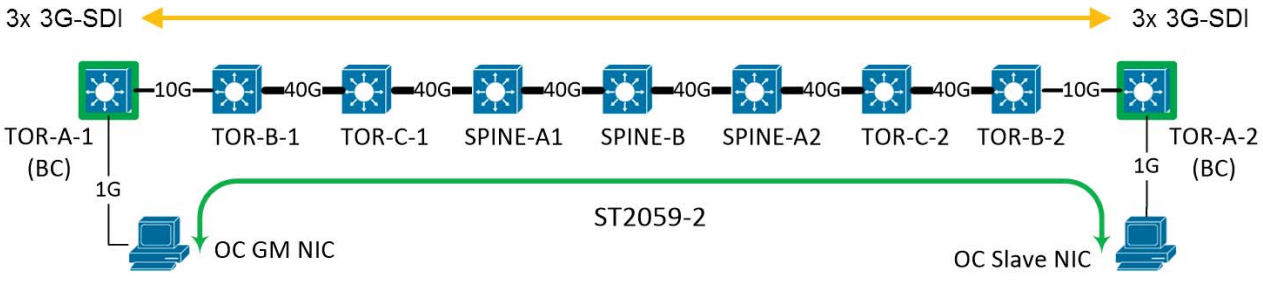

Figure 5: Mixed 9-hop network

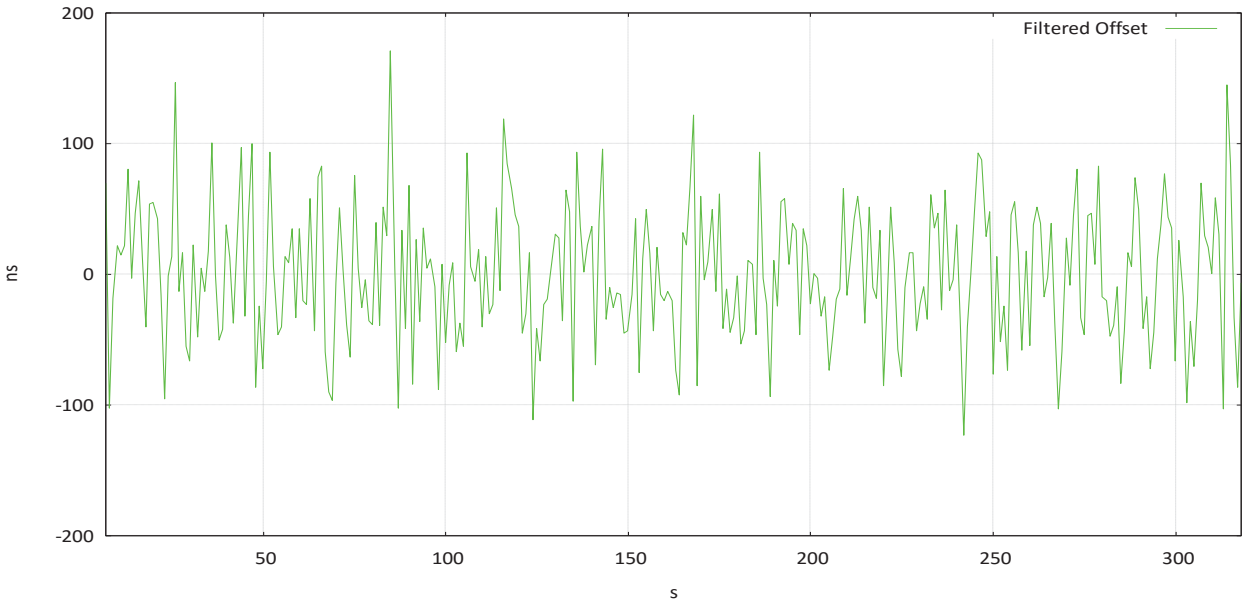

Figure 6: Offset of the slave device

Loaded non PTP-aware networks

As the transport of media essence over IP grows thereby increasing the network load, queuing and scheduling IP packets in order to ensure their real-time delivery in the most deterministic manner possible becomes of the uttermost importance. This statement applies equally to PTP messages. Our research has demonstrated the impact of PTP performance running on 90% and 120% loaded networks and the significant delay in reaching a lock and an accuracy that matches the target value of ± 500ns. [8] This is even more significant for devices that run with a centralised CPU vs. true line rate hardware packet forwarding devices. Tests were made with a 3-hop network topology as shown in figure 7, where PTP traffic flows parallel to a number of independent 3G-SDI streams.

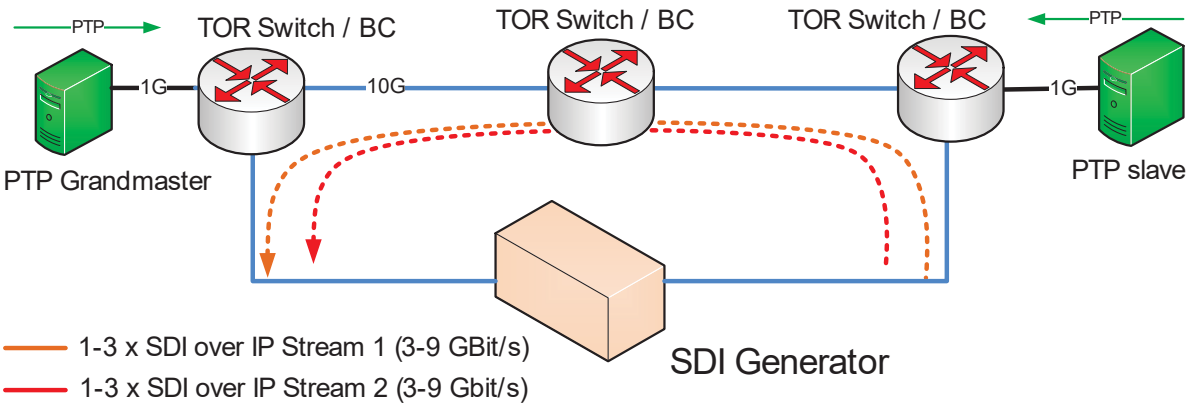

Figure 7: 3-Hop network with PTP and SDI traffic

These tests were performed with ToR network devices, capable of line rate switching. PTP end nodes were compliant to ST 2059-2, providing both fast locking as well as robustness against PDVs.

The following tests were performed:

- Three 3G-SDI streams (90% load), running as non PTP-aware network devices
- Four 3G-SDI streams (120% load), running as non PTP-aware network devices
- Six 3G-SDI streams (180% load), running as PTP-aware BC network devices

When loading the non PTP-aware network up to 90% using three independent 3G-SDI streams, the synchronisation accuracy still stays well below 50 ns. However, if this network is overload by adding a fourth 3G-SDI stream (120% load), the synchronisation accuracy deteriorates significantly to more than ± 1µs as seen in figure 8.

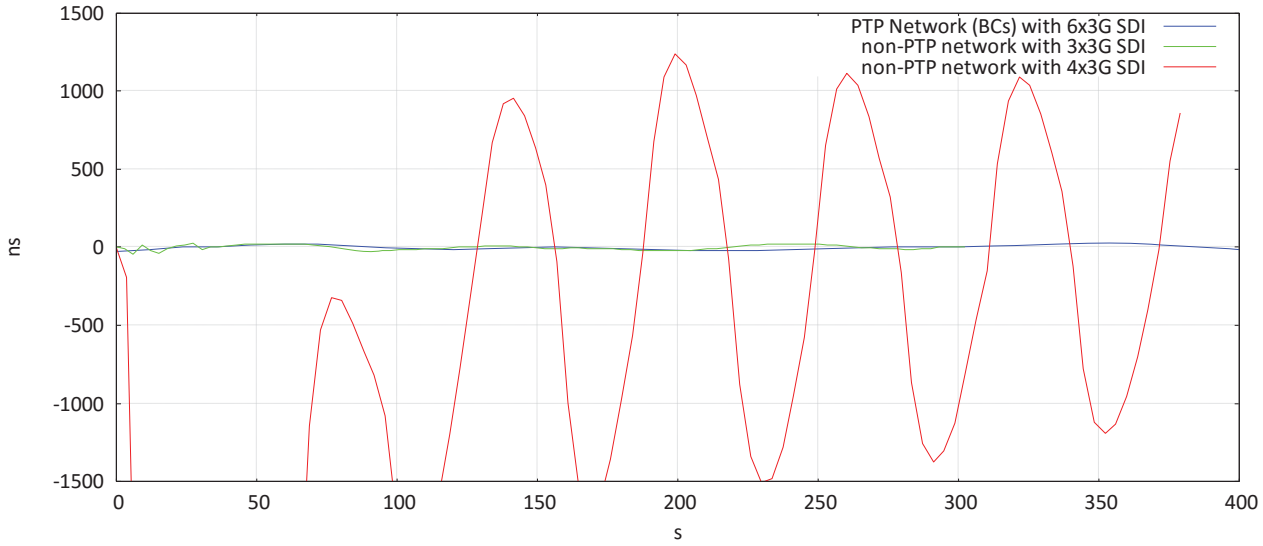

Figure 8: Offset of the slave device under varying load conditions and network capabilities

By contrast, a PTP enabled network, in our case consisting of BCs, can be overloaded with up to six 3G-SDI streams (180% load) while still maintaining excellent synchronisation accuracy below 50ns as shown in figure 7.

Strategies for mixed networks

In order to prepare for the insertion of PTP-aware infrastructure, the use of mixed networks whereby parts of the network nodes are PTP-aware is a common deployment scenario. Since the location of these PTP-aware nodes, such as TCs and/or BCs, within the chain of network elements between 2 endpoints has a direct impact on the precision, stability and recovery performance of the overall time & sync infrastructure, it must be carefully studied ahead of time. Drawing upon measuring the characteristics of such network topologies [10], the optimal location in a mixed network is to place Boundary Clocks at the leaf, (i.e.: the closest network node to the endpoint). Nevertheless, if the network is overloaded at some point in the network path, due to congestion caused by an equipment failure, the performance of the mixed network is degraded compared to a fully PTP-aware network.

CONCLUSIONS

PTP is perfectly well suited to meet all broadcasting requirements for delivering accurate time transfer for the All-IP studio. The ST 2059-2 profile provides the required framework defining all necessary PTP parameter ranges, message transport mechanisms, and network capabilities as well as a channel to transport application specific data via PTP.

It has been adequately well proven, that existing non-PTP aware network devices that are line rate capable can be used as a starting point for deployment. When using such network devices, even heavily loaded with media traffic, PTP will provide sufficiently accurate time transfer. If such networks are overloaded, even transiently, the accuracy drops significantly. Mixed networks with partial PTP support mitigate any network related issues even further, enabling a cost aware transition towards full PTP network support, which, of course, is ultimately the most accurate and reliable solution.

For mission critical applications and large production network environments careful planning is mandatory. This should include pre-deployment measurements preferably

using out-of-band methods as well as thorough analysis of PTP log data. Continuous detailed health monitoring during normal operation as well as provisioning for Master failures will further improve the robustness of PTP, as it provides only a somewhat limited native support for redundancy.

REFERENCES

1. IEEE Standard 1588, IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems, IEEE Instrumentation and Measurement Society, 2008.
2. G.8265.1/Y.1365.1, Precision time protocol telecom profile for frequency synchronization, ITU-T 10/2010
3. G.8275.1/Y.1369.1, Precision time protocol telecom profile for phase/time synchronization with full timing support from the network, ITU-T 07/2014
4. G.8275.2/Y.1369.2, Precision time protocol telecom profile for phase/time synchronization with partial timing support from the network, ITU-T Draft version.
5. C37.238-2011, IEEE Standard Profile for Use of IEEE 1588™ Precision Time Protocol in Power System Applications. IEEE Power & Energy Society, July 14th 2011
6. SMPTE ST 2059-2:2015, SMPTE Profile for Use of IEEE-1588 Precision Time Protocol in Professional Broadcast Applications, E-ISBN 978-1-61482-864-8
7. AES67-2015, AES standard for audio applications of networks - High-performance streaming audio-over-IP interoperability, Audio Engineering Society, 21st Sept 2015
8. Nikolaus Kerö, Thomas Kernen, Tobias Müller, and Mickael Deniaud, Analysis of Precision Time Protocol (PTP) Locking Time on Non-PTP Networks for Generator Locking over IP SMPTE Motion Imaging Journal, 123(2), 37-47, March 2014.
9. Nikolaus Kerö, Thomas Kernen, Fault Tolerant Clock Synchronization for the all-IP studio using PTP, Proceedings of the 2016 NAB Technical Conference, Las Vegas, April 2016
10. Kerö N., Kernen T. Müller T., Schimandl M., Optimizing Large Media Networks for Highly Accurate Time Transfer via PTP, SMPTE Motion Imaging Journal, 125(2), 30-44, March 2016, ISSN 1545-0279