TECHNICAL OVERVIEW

PRECISION TIMING FOR THE NEXT WAVE OF DATA CENTER APPLICATIONS

ABSTRACT

Data center-scale computing has arrived, as distributed modern workloads scale to the size of an entire data center. Now more than ever, these data-intensive application workloads rely on time-synchronous operations to run successfully at data center scale. Having emerged as an alternative to the legacy network time protocol (NTP), precision time protocol (PTP) delivers higher accuracy levels in IP networks. However, the need to provision dedicated hardware and software to support PTP positions it as a niche solution incapable of meeting the requirements of high-scale data centers. NVIDIA® DOCA™ Firefly, a software-defined, hardware-accelerated time precision service, overcomes this challenge, addressing the needs of modern data centers and enabling the next wave of applications.

DOCA Firefly democratizes and operationalizes precision timing for any data center environment at any size by leveraging the PTP support in NVIDIA® BlueField® data processing units (DPU) and providing a turnkey PTP solution with data center-wide visibility. DOCA Firefly is ideally positioned for an increasing number of data center applications, including distributed databases, industrial 5G radio access networks (RANs), video streaming, high-performance computing (HPC) collective operations, and more.

THE DATA CENTER—THE NEW UNIT OF COMPUTING

Fueled by the advent of data science and AI, the rapid growth in compute and data demands has stretched the limits of legacy data centers. Modern workloads that are more complex than ever, together with the rise of cloud-native computing, are redefining the way applications are being developed and deployed. To cope with the complexity and massive data and compute requirements, modern applications are increasingly distributed and based on containerized microservices. These developments have made the data center the new unit of computing, where application processing is performed at data center scale.

Data center-scale computing has increasingly come to rely on CPUs for general-purpose computing, GPUs for accelerating computing, and DPUs for data center infrastructure processing. The emergence of the DPU has had profound implications on the overall architecture of data centers. DPUs offload, accelerate, and isolate infrastructure services that previously ran on CPUs, thereby improving security, boosting application performance and driving down costs.

TIME: THE FOURTH DIMENSION IN THE DATA CENTER

The data center, acting as a single, immensely powerful computing platform, runs various types of workloads; however, not all workloads are created equal. Some applications are compute-intensive, while others are data-intensive. While computing power, data storage, and network speed are crucial to run data-driven applications, more applications today are real-time and delay-sensitive. Distributed databases, video streaming, and telco radio networks demand that all participating compute nodes operate synchronously to deliver optimal performance and efficiency. This positions time synchronization as the fourth dimension in the data center, adding to compute, storage, and network performance.

Traditionally, data center nodes have used the network time protocol (NTP) to synchronize their clock times. NTP is capable of synchronizing thousands of computers within a few milliseconds of timing skew between one another. Although this method of synchronization may have been sufficient for addressing the needs of legacy applications, modern applications increasingly rely on sub-microsecond time synchronization. The need for a unified and precise time scale dramatically increases as data center networks become faster. The lack of an accurate clock synchronization mechanism may negatively impact the performance of existing applications, while also inhibiting the creation of new applications.

PTP EMERGES TO DELIVER PRECISION TIMING

The IEEE 1588 Precision Time Protocol (PTP) is used to synchronize clocks throughout a distributed system. PTP achieves higher accuracy than the legacy NTP, making it suitable for modern applications. PTP was initially published in 2002. PTP v2, which introduced significant advancements, followed in 2008 and is the de facto standard today. Subsequent improvements were published in November 2019.

How PTP Works

The PTP standard describes a hierarchical master-slave architecture for clock distribution. The standard specifies a number of clock types that facilitate the distribution of clocks in a system. In a typical data center network, there are one or more devices acting as a source-timing reference called the grandmaster. In addition, a primary reference clock (PRC) is required to deliver a common timing source to all devices. Typically, this is based on one or more global navigation satellite systems (GNSSs), such as the Global Positioning System (GPS), being fed to the PTP grandmaster.

The grandmaster transmits synchronization information to destination devices, known as ordinary clocks, usually implemented by data center servers. Another clock type—the boundary clock—is a device with multiple network interfaces that can relay clock information on its network segment. Boundary clocks are useful in environments where routers and/or other devices block PTP messages. Data center network switches usually serve as boundary clocks that distribute clock information to computers and/or other interconnected switches.

PTP in the Data Center

One of the innovations PTP v2 introduced was the concept of a profile for defining PTP operating parameters and options. Such profiles have been created for diverse markets in various industries: telecommunications, electric power distribution, and media production, to name a few. In data centers, however, the use of PTP has been limited. Until recently, deploying PTP in a high-scale data center environment required specialized and dedicated hardware and software components, adding substantial IT costs and complexities. To distribute clock information to guest virtual machines (VMs) in virtualized environments, PTP must run as a dedicated VM on every hypervisor, adding even more cost and complexity. These overheads have positioned PTP as a niche solution for time-sensitive networks (TSN) that can be applied only to a handful of use cases, unable to meet the scale and efficiency needs of modern data centers.

INTRODUCING NVIDIA DOCA FIREFLY

NVIDIA DOCA Firefly, a DPU-accelerated precision timing service, is designed to accurately synchronize clocks in a distributed computing environment, overcoming the clock synchronization challenges of scale-out data centers. NVIDIA DOCA is a software framework that enables developers to rapidly create applications and services on top of NVIDIA BlueField DPUs, leveraging industry-standard APIs. Built on the state-of-the-art clock architecture and ultra-precise timing capabilities of the BlueField DPU, Firefly is a software-defined, hardware-accelerated precision timing service aimed at addressing the needs of modern data centers and enabling the next wave of applications.

Inspired by the remarkable way in which fireflies in the wild quickly synchronize their blinking, the DOCA Firefly service synchronizes clocks across a network of nodes. Firefly democratizes and operationalizes precision timing for any data center environment at any size.

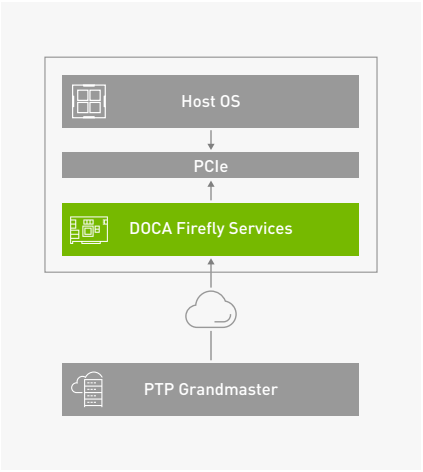
To accomplish these tasks, Firefly leverages the advanced PTP support featured on NVIDIA BlueField DPUs. The engines in BlueField are capable of obtaining the clock time and timestamping data packets in hardware at full wire speed, delivering breakthrough nanosecond-level accuracy. The uniqueness of BlueField's hardware clock architecture is reflected in its ability to maintain and distribute the clock in Coordinated Universal Time[1] (UTC), also known as "wall clock," which is a human-readable format. This hardware capability means that no software elements (firmware or kernel drivers) are required to timestamp packets in high-speed networks, which is key to achieving unprecedented levels of accuracy for a broad range of workloads. Deploying BlueField DPUs in both new and existing data centers is far more cost-effective and simple than specialized TSN equipment.

What's more, the DOCA Firefly service provides a full turnkey solution for PTP time-synchronized data centers at every scale, regardless of the software stack (Linux, VMware, Windows, and so on). Firefly initiates the PTP daemon, monitors its health, and handles port failover/redundancy scenarios for continued operations. It continuously monitors the error boundary at any given time and alerts the user based on a predefined threshold. DOCA Firefly will be expanded in the future to provide data center-wide orchestration and visibility powered by Kubernetes and NVIDIA NetQ.

For application developers, Firefly unlocks a new dimension of time-synchronous operations in the data center. New and existing workloads can now rely on BlueField and DOCA to enable precision timing at data center-scale for boosting performance and bringing new and differentiated digital products and services to market.

How NVIDIA DOCA Firefly Works

The NVIDIA DOCA Firefly service is a DPU-accelerated service. Running on the BlueField DPU as a containerized process, Firefly is agnostic to the host operating system and doesn't consume any CPU cycles of the host. This deployment model removes software dependencies and can be applied seamlessly in heterogeneous environments. The following diagram describes how Firefly obtains clock information from a PTP grandmaster and relays it to the host. The DPU is fully isolated from the host system, so Firefly doesn't speak directly to the host OS. Instead, Firefly writes the clock information to PCIe and the OS reads the clock information from there. Similarly, If the host runs a VMware hypervisor, the guest VMs can retrieve the clock information by reading it from the PCIe.

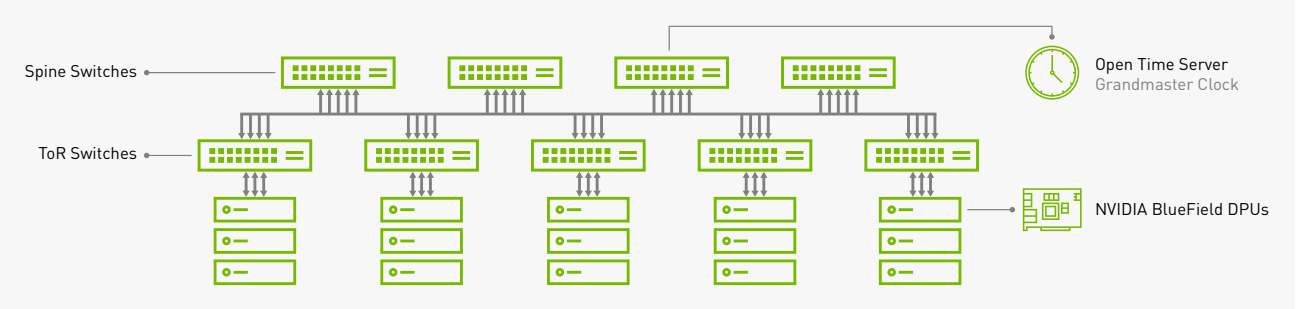


[1] The official abbreviation for Coordinated Universal Time is UTC. It came about as a compromise between English and French speakers.

Coordinated Universal Time in English would normally be abbreviated CUT.
Temps Universel Coordonné in French would normally be abbreviated TUC.

The Open Compute Project (OCP) has published a reference deployment guide under the Time Appliance Project (TAP), which prescribes how NVIDIA BlueField DPUs are used as ordinary clocks in globally PTP-synchronized data centers. The following figure illustrates this deployment, heightening the role of the DOCA Firefly service running on DPUs.

The open time server is the root of the clock tree that distributes clock information via PTP messages, while the role of the data center network switches is to propagate the clocks to destination nodes.



PRECISION CLOCK ACCURACY

Building upon the time-optimal BlueField DPU architecture and the PTP stack, Firefly achieves breakthrough time-synchronization accuracy. The following hardware features are purpose-built into the BlueField DPU for delivering precision timing services:

- A monolithic hardware clock supporting UTC is key for timestamping packets at full wire speed. In addition, the clock architecture embedded in the BlueField DPU enables precision testing and measurement by utilizing the pulse-per-second (PPS) in/out device functions. BlueField's ability to support UTC makes it possible to achieve nanosecond-level accuracy. Respectively, applications demanding clock information can obtain it in UTC format.
- Hardware time stamping of packets allows scheduling transmission according to a precise time period, eliminating congestion events.
- Advanced packet-pacing technology—how a series of packets are scheduled for transmission—avoids traffic bursts and network congestion. The BlueField hardware clock can streamline packet-pacing operations over time, adjusting packet transmissions according to the hardware clock.

- NVIDIA Accelerated Switching and Packet Processing (ASAP^2)™ technology provides time-based flow classification and action. This allows for advanced, flow-level traffic steering and manipulation in a software-defined network environment.

The following table describes a real-world application use case that demonstrates the accuracy that can be achieved with Firefly:

Key application	Distributed database at massive scale
Problem statement	When NTP is used to synchronize all participating nodes, a database transaction takes around ~20 milliseconds (20 x 1/1,000 of a second), in which the database is locked to ensure data correctness
Key performance indicators	When PTP is introduced on NVIDIA BlueField DPUs, the wait time is reduced to 50 microseconds (50 x 1/1,000,000 of a second), which is 400X less wait time.

PRIMARY APPLICATIONS

The need for precision timing is becoming ubiquitous for a broad range of data center applications. Below are some of the prominent use cases for the NVIDIA DOCA Firefly service.

Distributed Databases

Distributed databases are a field-proven use case for DOCA Firefly. Scale-out databases involve a large number of compute nodes, interacting with database instances in real-time. Commit-wait is an advanced commit protocol—each time a database transaction is committed, the database is locked to prevent later transactions from reading old data. This external consistency refers to the idea that later transactions can see the outcome of previously committed transactions. With Firefly, every transaction is timestamped in microseconds (1/1,000,000 of a second), which both enables a significant reduction in commit waiting time and increases database throughput. Similarly, any distributed transactional application should have the potential for time accelerations.

Another long-lasting challenge is handling database caching at scale. Modern databases use caches to reduce data access time. Data that is accessed frequently by users and applications is copied into distributed memory for caching purposes. When this data is changed in the database, the common practice is to invalidate all cached data copies. If there's no cached copy, a new query is initiated to the main database to retrieve the data. Invalidating the cached data before committing the new value has significant performance penalties.

With DOCA Firefly, all participating nodes are time synchronized, an efficient time-bound practice to handle cache coherency. Data values are written to cache with an expiration date and time. When the timer exceeds its limit, the cached data is purged even if the value hasn't changed. At a system level, this method allows for new data to be committed in the database, eliminating the need to wait for all cached copies to be flushed. The cache data timer is configurable and can be fine-tuned for optimal performance.

Industrial 5G RAN

5G radio networks are a catalyst for the digital transformation in the manufacturing/industrial space, integrating machine-to-machine (M2M) communication and the internet of things (IoT) for increasing automation of traditional manufacturing and industry practices. Industrial 5G is the communication fabric for advanced machines and robots in the smart factory of the future. The Third-Generation Partnership Project (3GPP) has standardized PTP time synchronization and Synchronous Ethernet (SyncE) in every 5G system to enable industrial IoT. NVIDIA BlueField and DOCA Firefly support the International Telecommunication Union (ITU) G.8273.2 standards, making them ideally positioned to address the stringent time-synchronization requirements of industrial 5G systems.

Video Streaming

As video broadcasters worldwide transition from serial digital interface (SDI) to reliance on IP technology, accurate and reliable time transfer over Ethernet networks has become key. The Society of Motion Picture and Television Engineers (SMPTE) body has specified in its ST 2059 standards the use of PTP for time synchronization. Professional IP video broadcasting workloads, however, often run in heterogeneous computing environments that aren't optimized for PTP clock accuracy. Deploying NVIDIA BlueField DPUs with the DOCA Firefly service addresses the needs of video broadcasting facilities for precision timing, removing dependency on the host OS and supporting virtualized cloud environments. In addition, Firefly can be used to synchronize a number of GPUs that are processing video frames.

HPC Collectives

HPC environments involve hundreds of compute nodes working together on a job. Also known as collective operations, this typically involves nodes sending and receiving variable-sized data from all nodes (all-to-all). In cases of many-to-one,that is a group of nodes sending data to a single node, the network becomes congested, causing retransmissions and packet drops. The DOCA Firefly service can eliminate network congestion events by introducing time-based transmission. Firefly accelerates HPC collective operations by providing precision time services, controlling congestion on the network fabric.

CONCLUSION

Modern workloads are pushing existing data center resources to their limits. Data center-scale computing increasingly relies on accurate time synchronization between all participating nodes. As the fourth dimension in the data center, the need for precision timing has never been greater. NVIDIA DOCA Firefly is the software-defined, hardware-accelerated precision-time-synchronization service that boosts application performance and enables the next wave of modern applications.

READY TO GET STARTED?

- Register to gain early access to the **NVIDIA DOCA** software framework.
- Learn more about **NVIDIA BlueField DPUs**.
- Learn why **NVIDIA DOCA** is the key to enabling seamless adoption of BlueField DPUs.
- Learn more about the benefits of **NVIDIA-Certified™ Systems**.
- **Find an NVIDIA-Certified System** available through the world's leading server manufacturers.

NVIDIA