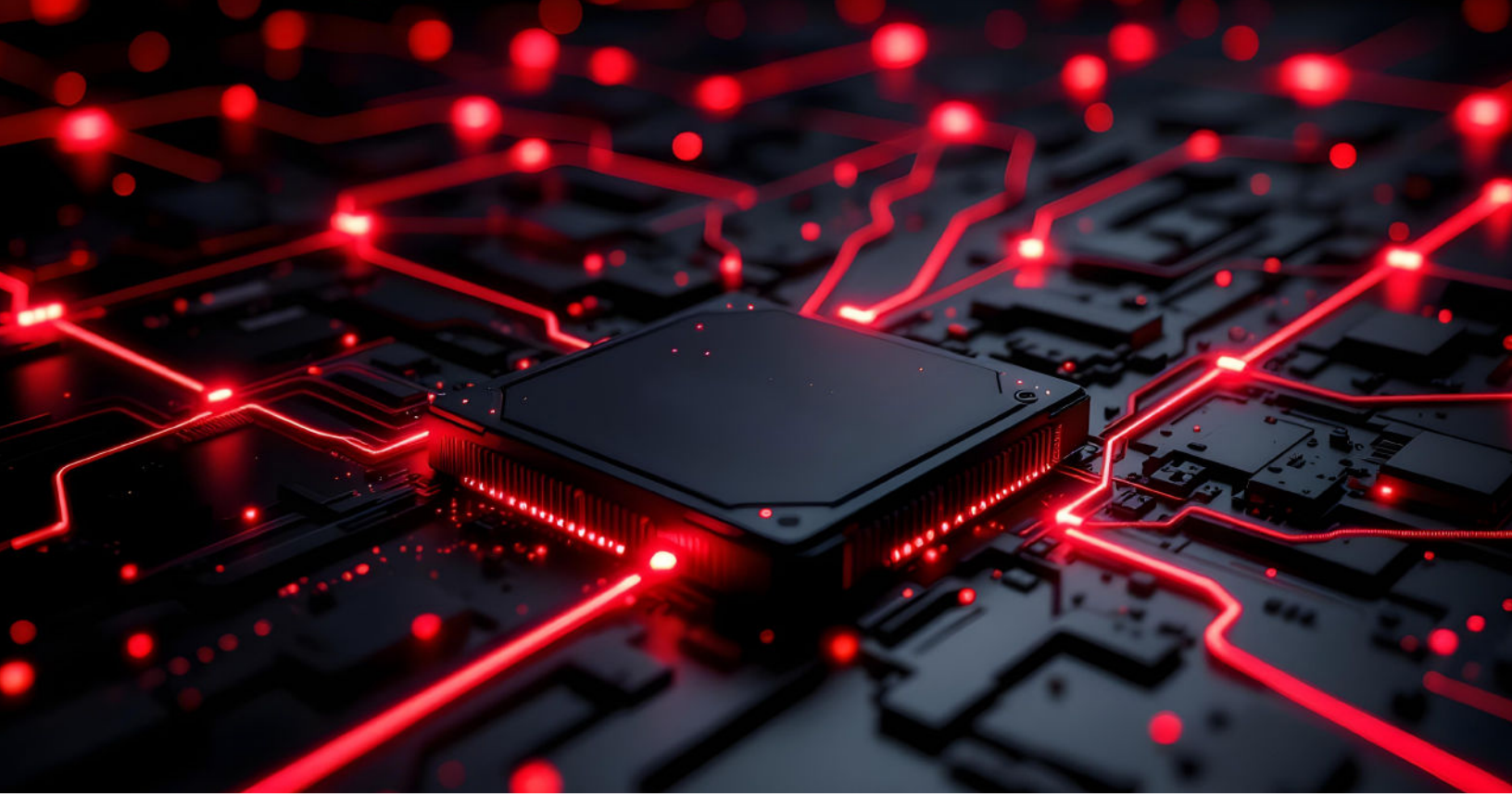
Futurum

SiTime

Precision Timing's Critical Impact on Data Center ROI

How selecting the right timing solutions can unlock higher operational performance, lower TCO, and improve ROI for data centers

AUGUST 2025



Introduction

If GPUs are the brain of AI, Precision Timing is its heartbeat. As compute demand for both AI training and inference workloads increases, so does the complexity of data and compute power orchestration, making Precision Timing all the more critical.

For reference, the number of GPUs in a single training cluster increased from just 800 in 2016 to over 16,000 today. Despite significant improvements in training and processor efficiency, that number is expected to increase to 100,000 in the coming year and to reach 1M by 2030, translating to a cost of $100B per data center.

Managing the enormous compute potential of so many GPUs requires networking them together in a distributed computing architecture. In that type of architecture, AI training efficiency is proportional to how fast GPUs exchange data and how tightly synchronized they perform tasks. As efficiency increases, GPU clusters behave more like single machines that cycle between compute, communicate, and synchronize activities, all of which depend on Precision Timing. Precision Timing helps electronics products become smarter, quicker, and safer by providing dependable clock signals that can be trusted, no matter the environment.

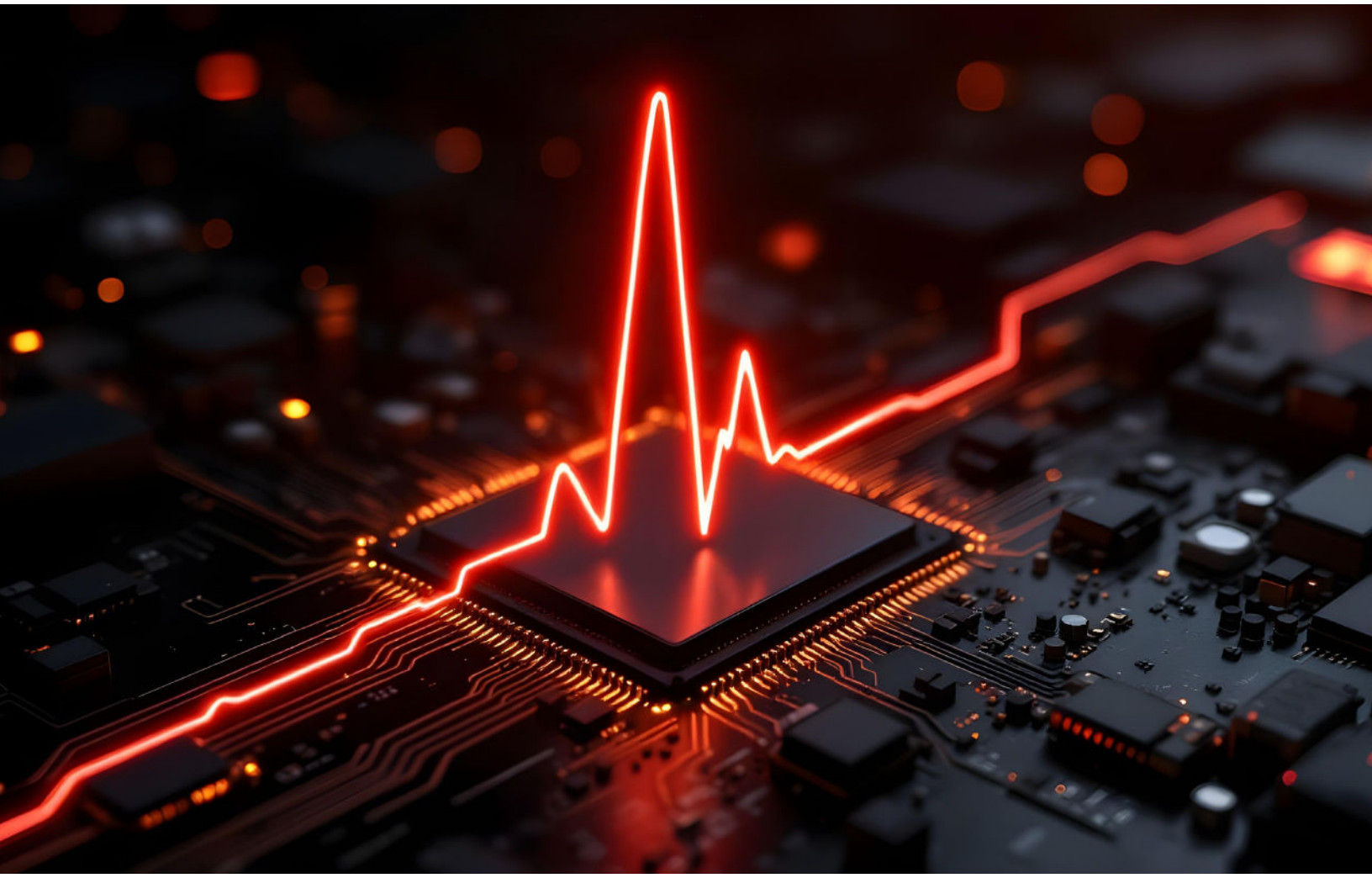
In a data center environment, think of trillions of information packets moving back and forth every second. The more precise the timing of that flow, the more efficient the cluster is. The less precise the timing, the more inefficient it is, costing data center operators precious time and resources, to say nothing of the added cost of running workloads so inefficiently.

Since distributed computing isn't limited to training, the importance of timing extends beyond it: Once an AI model has completed training in a dedicated back-end network, it is then moved to production in a front-end network where online users may interact with it (the process known as "AI inference"). Once again, albeit on a smaller scale, as the AI model size grows, the number of GPUs needed to both contain and inference from it also grows. Here again, Precision Timing is critical to running inference, especially at scale, in an optimized and efficient manner.

In other words, synchronizing distributed computing, particularly to manage AI workloads at scale, requires Precision Timing. And while the cost benefits of running as efficient a system as possible thanks to Precision Timing are worth spending a good deal of time on, the same gains in efficiency, speed, and positive user experiences can also stimulate more demand for AI solutions. The Jevons Paradox, attributed to economist William Stanley Jevons for having been observed in his 1865 book The Coal Question, applies here: His hypothesis is that as a resource becomes more efficient to use, the cost of using the resource drops, which can stimulate demand. This was similarly echoed by Microsoft CEO Satya Nadella, who explained that "As AI gets more efficient and accessible, we will see its use skyrocket, turning it into a commodity we just can't get enough of." As such, efficiency improvements from Precision Timing can increase AI adoption and the discovery of new applications.

Despite the vast majority of the tech industry's attention focusing on best-in-class silicon, components and software to drive the best possible AI workload performance per watt and per dollar, the timing piece of that overall equation often gets missed. And that needs to be rectified, because if the timing piece of the AI workload technology stack isn't handled right, no amount of spend on GPUs will yield optimal efficiency, performance, and ROI outcomes. Timing is therefore the most critical piece of the AI puzzle still being missed.

Let us dig into some of the most critical benefits that Precision Timing brings to communicate, synchronize, and compute data.



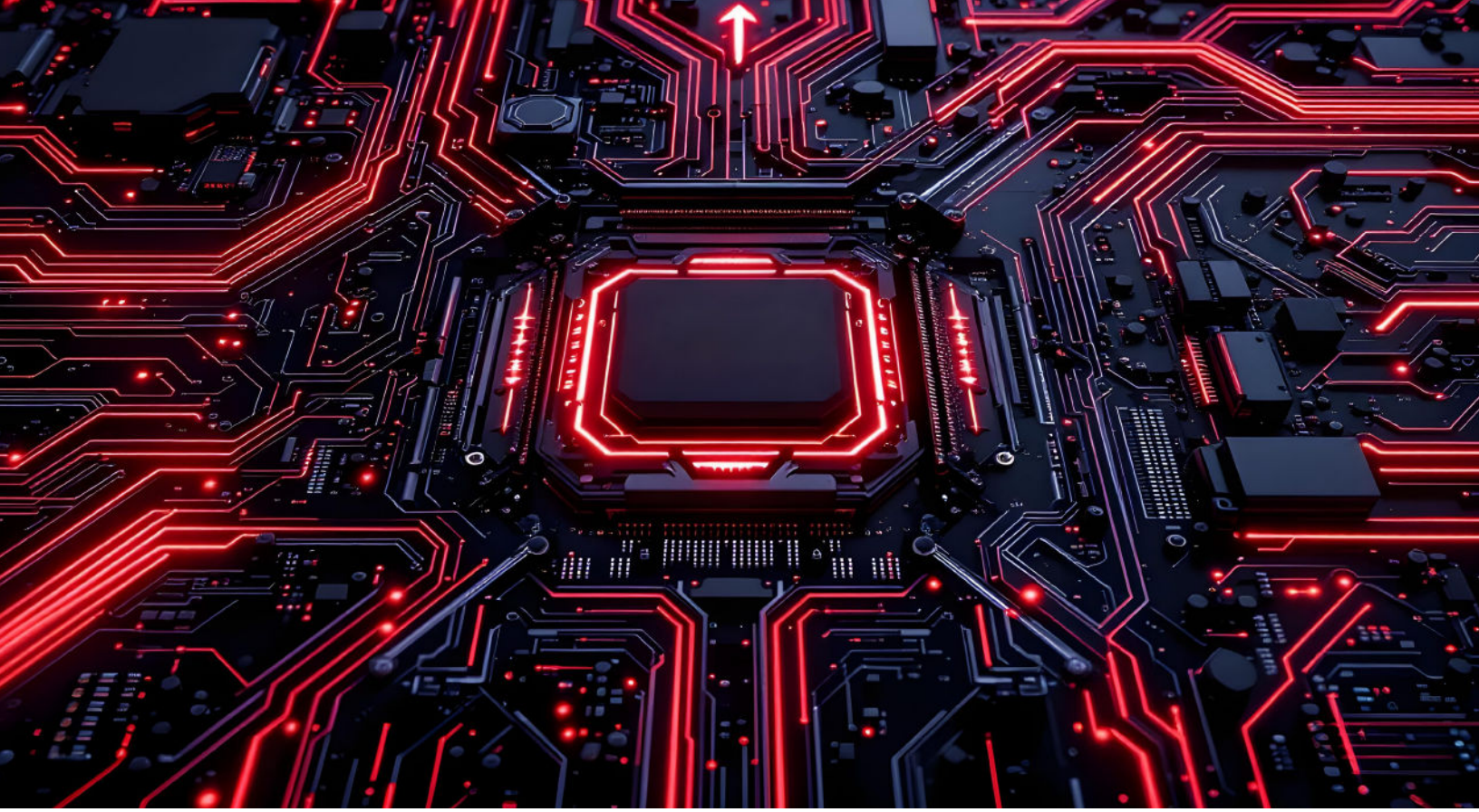
Leveraging Timing to Optimize Data Centers in the Age of AI

As data centers increasingly transition from traditional compute workloads to AI applications, demand for more infrastructure and faster performance is bound to continue to expand, with data center servers evolving from traditionally compute-centric architectures to more high-speed interconnect-inspired designs. This means that timing components have to both support this evolution and become better at facilitating it at scale.

Common pain points and data center inefficiencies touch on this specific efficiency optimization challenge. For instance, we know that every AI training cycle requires each GPU to share its result with all other GPUs in a synchronized manner to ensure all GPUs update the model consistently. This all-to-all model of AI training requires more and faster interconnects in an AI server than with traditional compute servers. The problem, however, is that if GPU-to-GPU interconnects aren't able to transmit data at the highest possible speeds, GPUs will sit idle, waiting for data from the network. For reference, AI training efficiency drops in proportion to both the number of machines waiting for the slowest machine to complete and the total wait time/cluster. Network congestion can lead to delays and inefficiencies, which can in turn become extremely costly to data centers.

Additionally, speed increases place more demands on timing components, most notably the need to become quieter (express less jitter). For reference, doubling data rates requires half the jitter just to maintain the same timing margin. Lastly, distributed databases can generate unwanted traffic due to timing uncertainties and load imbalances among GPUs can also slow down the entire AI training process. Getting the timing piece of this puzzle right is the key to optimizing network synchronization.

More precise timing can also reduce the timing uncertainty for distributed database writes, eliminating traffic needed to break ties, which would otherwise cause additional traffic and load imbalances.



How Timing Solutions Improve System Efficiency

Broadly, Precision Timing enables efficient communication and synchronization, ensuring that all GPUs in an all-to-all model operate in lockstep. Synchronization creates accurate timestamps to order and coordinate events, which network telemetry relies on for causality in distributed systems. In other words, AI training, because it is built on a distributed computing model, is heavily dependent on accurate synchronization at scale to be efficient.

For starters, synchronization ensures model consistency. To reduce TPTC (time per training cycle), workloads are broken up into segments and assigned to multiple GPUs. Each GPU then processes its own assigned subset of data. Pushing out these data packets, collecting them, aggregating them, and sharing the aggregated results back with all of those GPUs requires complex time coordination. Without precise time synchronization between all of these moving parts, some GPUs could, for example, update different versions of the same model, which could result in errors.

In a similar vein, synchronization also allows AI models to converge properly. By making sure that no GPUs finish either sooner or later than others, every part of a model will converge at the same rate. Because instability in model convergence can also extend total training time, addressing this with proper synchronization is a simple way of eliminating that inefficiency.

Additionally, accurate synchronization improves scaling efficiency: As the number of GPUs per cluster grows, coordinating updates between GPUs helps ensure that all GPUs finish their work at the same time. This matters because if one or more GPUs take longer than the rest to complete an iteration, the entire iteration will take longer, making the training process inefficient.

Compute

While we tend to focus a lot on GPUs because of their importance to AI training, AI servers require a variety of different chips beyond GPUs. These other chips also require timing supplied from a system clock tree made up of a series of timing components, such as oscillators, buffers, and clock generators. This tree supplies all clock frequencies required by the baseboard (or AI compute tray) to operate properly.

Additionally, AI servers generally include CPU motherboards that are likely to lean on real-time clock, buffers, and clock generators to enable a variety of clock trees fulfilling PCIe, DBx, and CK440 requirements, for example. To ensure continuous error-free operation, timing must be stable, reliable, and low-noise (low jitter). A broad timing portfolio is typically needed to construct an agile, complete, and properly optimized clock tree.

How Timing Can Lower TCO and Improve ROI for Data Centers

While efficiency for efficiency's sake can be the reason enough to focus on optimizing timing solutions, the case can easily be made for more TCO and ROI-centric approaches to validating the importance of boosting system efficiency with timing solutions. Case in point: Because network telemetry is derived from timestamps, securing the fastest and most accurate timestamping available on the market is critical to help DevOps and AIOps teams ensure maximum cluster uptime and efficiency – which translate into lower TCO and higher ROI.

The TCO/ROI discussion around timing solutions can be summarized to a simple set of quantifiable benefits (in no particular order):

- **Reducing GPU idle time and improving network speeds and utilization:**

  Combining faster interconnects with Precision Timing can prevent GPUs from waiting idle for data. The cumulative effect of these efficiency improvements can add up quickly for a data center, resulting in improved GPU uptime, compute capacity improvements, lower cost per workload, and more efficient power utilization.

  Sustained GPU uptime ensures that a data center's initial investment in GPU hardware will yield its full potential value, translating directly into a more favorable return on CapEx. Improvements in GPU uptime also means more workload throughput, which in turn can open up more revenue generation potential. AI models can be trained and deployed faster, scientific simulations can yield quicker results, and more customers can be served with GPU-accelerated services.

  The ability to consistently deliver these high-value computations without interruption not only speeds up internal projects and innovation but can also allow service providers to handle higher workload volumes, which can fuel revenue streams and further strengthen their competitiveness in the market.

  Because GPU downtime often leads to a loss of progress on long-running computational jobs, which can result in costly restarts and wasted energy, ensuring GPU uptime can significantly lower the instance of IT support hours spent on diagnosing failures, managing repairs, and re-running tasks.

Ensuring higher GPU uptime improves service reliability and customer satisfaction, which are crucial for long-term ROI: Consistent performance and availability can convert into increased trust, stickier loyalty, and reduced churn. Lastly, more predictable GPU performance also allows for more accurate resource planning and better adherence to Service Level Agreements (SLAs).

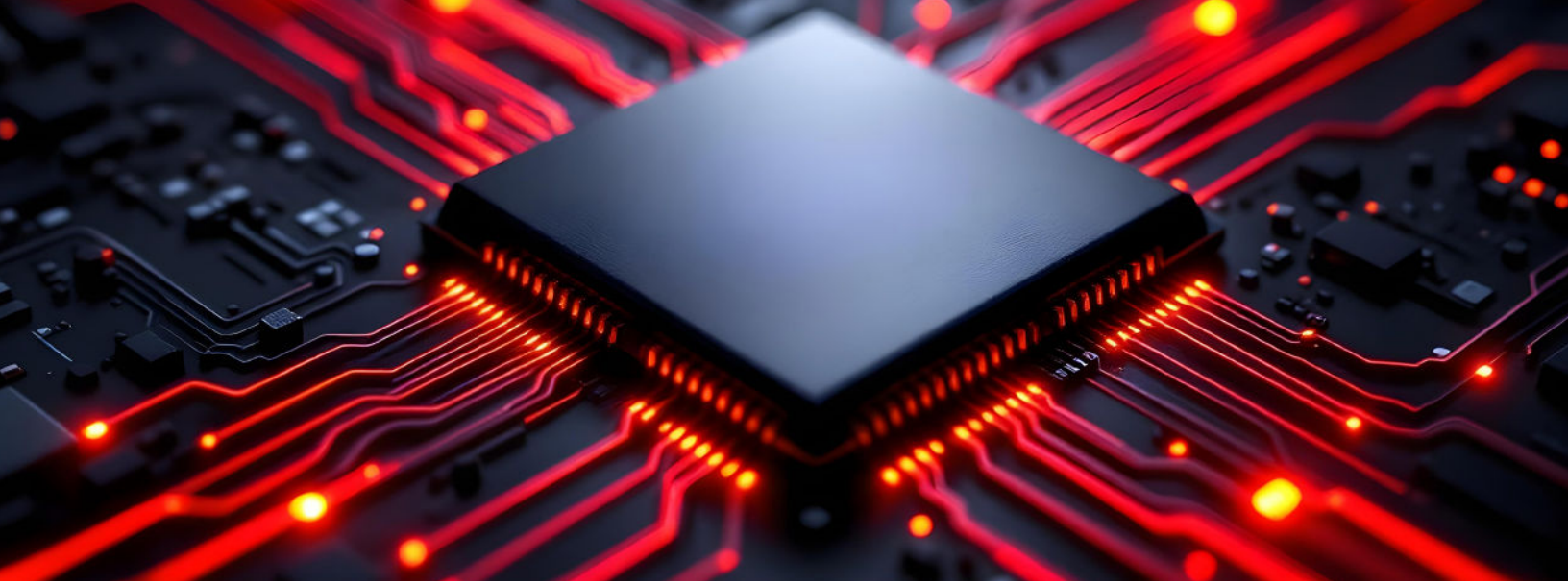
Precision Timing and synchronization also enable higher network speeds and utilization. Faster network infrastructure allows for quicker data transfer between servers, storage systems, and end-users, which translates into more responsive applications and reduced wait times. This heightened application performance can boost the productivity of internal users relying on data center resources and significantly improves the experience for external customers accessing hosted services. As a result, businesses can achieve faster task completion, support more concurrent users effectively, and increase overall satisfaction, all of which contribute positively to ROI.

Increased network speeds and superior utilization can also expand a data center's overall throughput: A well-tuned, high-speed network can move larger volumes of data more efficiently, enabling the support of more demanding applications such as AI model training and high-volume transaction processing without immediate and costly infrastructure overhauls. By maximizing the data handling capabilities of the existing network fabric, organizations can either defer or reduce capital expenditures on new networking hardware, potentially improving ROI by extracting more value and supporting greater scale from their current investments.

Optimizing network utilization with better timing solutions and synchronization can also play a crucial role in minimizing latency and preventing performance bottlenecks that otherwise weaken data center efficiency. This not only ensures consistent application performance but also allows other expensive data center resources to operate at their full potential. By eliminating common inefficiencies, data centers can operate more smoothly, ensuring that all invested capital in compute and storage is effectively leveraged, directly contributing to higher ROI.

- **Improving power consumption:**

  The more efficient a process is, the less energy it requires. Most, if not all, of the above areas of lower TCO and higher ROI touch on ways that a more efficient data center can improve its performance per watt. This can manifest either as lower energy consumption and expenditures, or as more revenue per watt, as a data center can operate more GPUs on the same amount of power (boosting Flops per Watt).



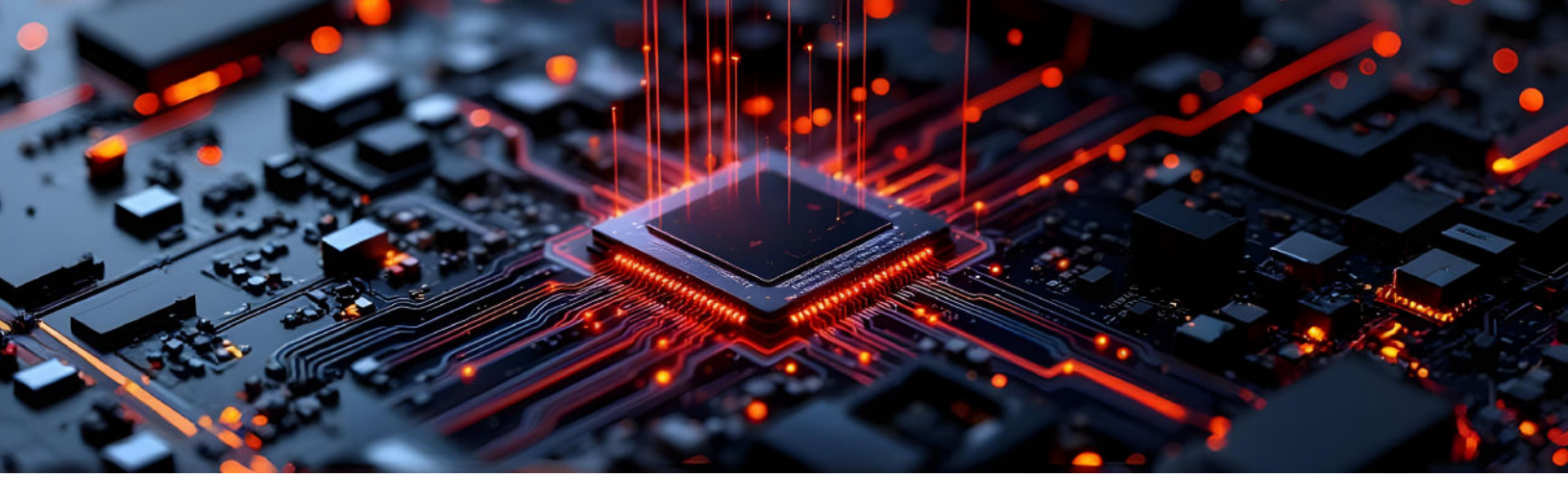
From Theory to Practice: How SiTime Is Disrupting the Timing Solution Market for Data Center Applications

With AI adoption accelerating, optimizing data center utilization and workload efficiency are becoming increasingly critical to data center operators. Large AI clusters, comprising thousands of powerful servers and tens of thousands of AI accelerators, need maximally efficient direct connections between compute nodes and their accelerators. This direct connectivity model, if timed properly, can boost AI workloads by distributing tasks and increasing productivity.

Improved time synchronization can also provide data center operators with more accurate telemetry, highlighting underperforming compute nodes. Precise synchronization enables architects to decrease the so-called "window of uncertainty" (the guard band around time stamps between data packets). A smaller window of uncertainty for distributed databases reduces the likelihood of time stamp guard bands overlapping in time, which lowers the frequency of data re-transmissions and ultimately increases data center efficiency. To that end, SiTime recently released its new ultra-stable, low-jitter SiT5977 Super-TCXO, a compact single-chip solution designed to provide the most resilient timing for distributed applications (including AI) with high bandwidth and better network synchronization.

Designed to streamline AI system architecture and optimize network efficiency, the SiT5977 Super-TCXO improves time synchronization for Smart NICs, acceleration cards, switches, compute node applications, and more, with 3x the synchronization precision compared to previous generations. Despite its small package (5.0 x 3.5 mm footprint compatible with the more common 5.0 x 3.2 mm package), its 156.25 MHz output frequency enables high-speed SerDes and 800G links, and it is capable of driving 800G and higher links via 80 fs phase jitter and LVDS differential output. The first TCXO to offer high-speed differential output for high-speed data communication, it also supports precise digital tuning of output frequency (DCTCXO) up to ±400 ppm pull range, a frequency pull resolution down to 0.05 ppt (5e-14), and eliminates link flaps from activity dips or micro jumps associated with quartz devices. The solution's low-noise PLL integration secures a clean clock edge and exceptionally low jitter, lowering bit error rates.

One of its more outstanding features is its high degree of environmental resiliency, with ±1 ppb/°C frequency slope (dF/dT) for optimum performance even under airflow and thermal shock. Also resistant to vibration and board bending, it eliminates external LDOs via on-chip voltage regulators, delivers ±100 ppb frequency stability over temperature with DualMEMS architecture, and lists a -40°C to 105°C operating temperature range. This translates into no resource-draining performance throttling and fewer costly workload restarts, even in the most demanding AI-focused data center environments.



Hardware Optimization Needs the Right Software

Data center operators have a lot of options when it comes to optimization and timing software, and legacy software solutions have been adequate enough to run legacy systems efficiently. As data centers shift to the specific demands of AI workloads at scale, and the table stakes of network efficiency dramatically increase, choosing software that delivers the most precise timing performance becomes paramount.

Until now, timing solutions customers generally had to shop for hardware and software separately (oscillator vendors were experts in oscillators, while clock vendors were experts in clocks and software). SiTime is streamlining the process by also providing software which, by design, is already perfectly matched to the company's hardware, and uniquely effective at both complementing and optimizing it while simplifying procurement for customers.

Broadly, SiTime's IEEE1588 certified TimeFabric software suite delivers an integrated stack and servo for the best possible system performance and faster time to market. Its holdover extension software also incorporates a reference clock with minimized error that, remarkably, exceeds 24 hours of holdover with no need to modify any hardware.

This addresses three specific customer challenges: Limited time to develop a cohesive system solution with disparate components, less-than-optimal performance for disparate hardware, and difficult upgrade cycles. The first is simple: By investing in a fully-integrated system, customers can accelerate their time to market. And because they only need support from one expert in the entire clock tree as opposed to several, the second challenge is also a fairly straightforward fix because SiTime is the expert in every system component. This translates into a number of performance advantages:

- **9x** better synchronization than quartz
- **5x** better sync than spec
- **2x** better holdover under constant temperature (ensuring better service continuity) than when relying on disparate vendors
- **25x** less volume and **3x** lower power than a legacy 1PPSDO, which tends to be large and power-hungry
- No need for an external XTAL/BAW, thanks to its smaller, power-efficient integrated OCXO (over controlled crystal oscillator).

For reference, holdover is the duration over which an oscillator maintains a defined time error for a given temperature profile (e.g., 1.5 µsec error over 8 hours within a 10°C window at 0.5°C/min temperature change). It is dependent on three oscillator specifications: Frequency Stability over Temperature Slope (ΔF/ΔT), Aging (1-day aging), and Wander (HDEV).

Holdover occurs when a node loses time reference (GNSS, PTP, SyncE) and relies on the local OCXO to provide time and continue downstream network operation. The problem is that infrastructure nodes (think 5G DUs, data center switches, and core routers) are all potential points of failure: GNSS Time is susceptible to weather, spoofing, and jamming. Network time can be impacted by the cumulative effect of errors at each node. Local Time and OCXO, for their part, clean up network time and backup to GNSS. For those reasons, network nodes require redundant time sources to ensure proper service continuity. One critical advantage of SiTime's holdover solution is that it addresses this need better than most legacy solutions.

As for the third challenge – upgradeability, SiTime's stack is designed to ensure longer product and system lifecycles, with its software enabling updates and upgrades without requiring board changes. This obviously not only makes life easier for engineering teams, but also helps maximize ROI and lower TCO.

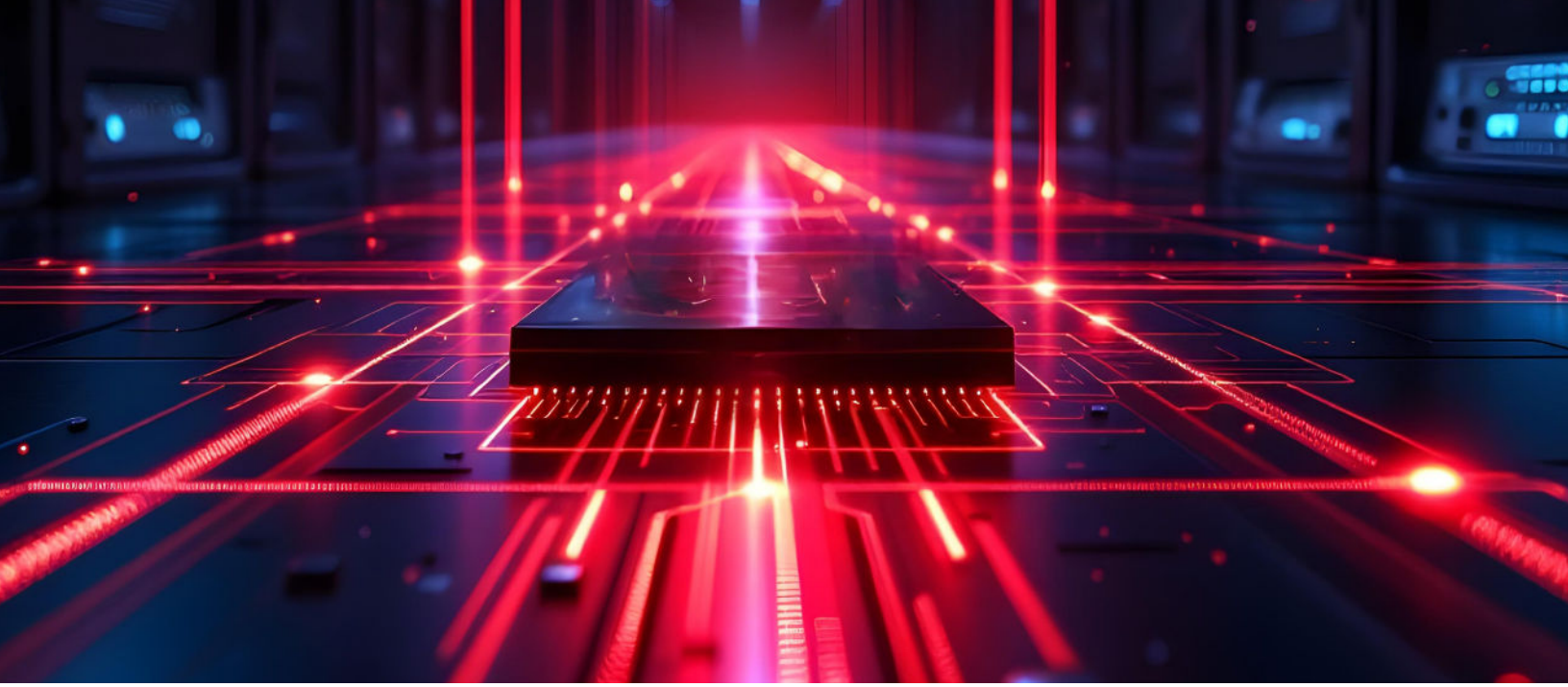
Timing's Importance Beyond the Data Center

Timing is also critical to everyday technologies outside of the data center: Mobile devices, PCs, tablets, connected wearables, wireless audio, drones, GPS devices, and smart vehicles – any device that uses modern RF, 5G, and GNSS chipsets, or any combination thereof, depends on precise timing in order to both synch properly to networks and critical data, and to know precisely where they are at all times. This latter part is especially critical in automotive applications, which benefit from precise geolocation and timing at speed, even when network access is limited or inconsistent.

Back in May, SiTime announced its first mobile clock generator with an integrated MEMS resonator. Dubbed "Symphonic" (or SiT30100), the solution delivers accurate and resilient clock signals (with a system stability as low as ±0.5 ppm) meeting the requirements of 5G, RF, and GNSS chipsets at ultra-efficient power consumption for mobile devices, IoT devices, and asset trackers.

With its integrated MEMS resonator, the 10-pin chip-scale, tiny 2.22 mm^2 Symphonic mobile clock generator combines the functionality of up to four discrete timing devices to reduce board footprint and simplify system design – a critical differentiator for OEMs constrained by increasingly small device form factors and the need to maximize power efficiency.

As with SiTime's data center solutions, this mobile solution is designed to be resilient and operate with maximum precision even in harsh environments and conditions: Aside from the solution's robust construction, an integrated temperature sensor feeds compensation algorithms that help deliver both improved GPS accuracy and faster lock time at the system level, even under the harshest environmental conditions (-30°C to +90°C operating temperature range).



Conclusion

Faster networks and tighter synchronization rely on Precision Timing to drive efficiency gains. This trend is expected to continue. For example, Ethernet speeds will double every 24 months and synchronization, which ends at the server today, will expand to GPUs within a rack and from rack-to-rack. Ongoing efforts to improve synchronization will benefit many applications. For example, tighter synchronization enables optical circuit switching (OCS) to increase switching speeds from milliseconds today to nanoseconds or better to support live traffic, bringing with it higher bandwidths and lower latency and power. Tighter synchronization also benefits time-sensitive security and authentication algorithms to prevent man-in-the-middle and other types of attacks. Quantum computing, just starting to take off, also relies on distributed computing and, as such, becomes more efficient with synchronization. It may also combine with AI to create quantum AI for solving even more complex problems than possible today. All of these advancements require precision timing to become reality.

Important Information About This Report

AUTHORS

Olivier Blanchard
Research Director & Practice Lead,
Intelligent Devices | The Futurum Group

PUBLISHER

Futurum Research

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

SiTime

ABOUT SITIME

SiTime is a leading provider of precision timing solutions based on MEMS (Micro-Electro-Mechanical Systems) technology. Unlike traditional quartz-based components, SiTime's MEMS oscillators offer superior reliability, resilience to environmental stressors, and programmability. Their products are widely used across industries such as automotive, industrial IoT, communications, and aerospace, where precise and robust timing is critical. SiTime's innovations enable smaller, more power-efficient, and more accurate systems, helping drive performance improvements in next-generation electronics.

Futurum

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.