



# Trading at the Speed of Light

How Intelligent Network Monitoring Delivers  
the Low Latency and High Speed Required by  
High-Frequency Traders

Sponsored by



## Introduction

Saying that success in today's stock market relies on speed feels somehow insufficient.

When profits and losses come down to nanoseconds, success requires something more than speed. Buy just a split second sooner than another trader and you get a lower price. An instant later, the same money buys less. Time doesn't just matter; it means everything.

When individuals want to buy stock today, they typically log into a portal, specify a stock and how many shares they want and press "buy." A request goes through a series of networks. If the buyer doesn't specify a purchase price, the order is either filled by one of several wholesale brokerages or one of the several dark pools, created initially for big institutions to help keep large trades a secret but now open to everyone. If a purchase price is specified, the buy order is filled at one of the public exchanges.

All of that takes place in less than one second. It's within that tiny span that high-frequency traders make their living – earning profits on fractions of fractions of seconds that add up to millions of dollars.

And, as in much of modern life, automation plays a large role in high-frequency trading. Sophisticated algorithms trigger buys or sells at light speed, enabling the high turnover of shares that makes profitability possible in the HFT world.

Instead of holding stocks for months or years in hopes of gaining 5 or 10 percent, traders rely on algorithms to read market data and turn shares over in seconds or less.

In the simplest terms, speed equals profit, and latency is the great enemy of speed. To achieve the highest possible velocity, engineers and operators must have complete visibility across the entire network, from firewalls to routers to other servers, so they can hammer latency out of every component. They need high availability, high throughput and the confidence that no data is being lost.

It's a high-stakes game where losing even a single packet can cost money. In the following pages, we'll examine how latency affects high-frequency trading and what tools and methods can be used to reduce it.

## Trading at the Speed of Light

Speed has been the most important variable in market trades since at least the late-1860s, when Thomas Edison's Universal Stock Ticker printed telegraph wire signals at a rate of one character per second.

Today, microseconds are a competitive differentiator, both for the demand and the provision of electronic execution services. The most successful high-frequency traders are those that can establish the lowest possible latency between processing environments. Their business model is based on the concept of latency arbitrage, in which traders use automated, electronic trading to capitalize on a small price difference between two or more markets.

In HFT, latency has a direct impact on the effectiveness of their algorithmic strategies and their ability to outperform competitors, especially during peak periods of intense activity. Information Week once cited an estimate that an advantage of one millisecond can be worth \$100 million a year to a major brokerage firm.

Microbursts, peaks in traffic that occur for just a few microseconds, can overwhelm a network that is not properly optimized, causing transactions to be lost or require retransmission. Such events can even result in complete data loss that cancels a transaction entirely.

Just one packet of information dropped because of a microburst that exceeds network bandwidth can dramatically increase trading time. Since an HFT algorithm performs thousands of trades per second, every necessary retransmission can translate into significant revenue loss. The challenge is to minimize trading latency end-to-end, while handling peaks without packet loss.



## HFT Firms Need Speed, Data Access and Capacity

To establish competitive differentiation and profit in the HFT market, firms are focused on maximizing three variables:

- Speed of execution
- Instant access to market data
- Ability to sustain peaks of activity

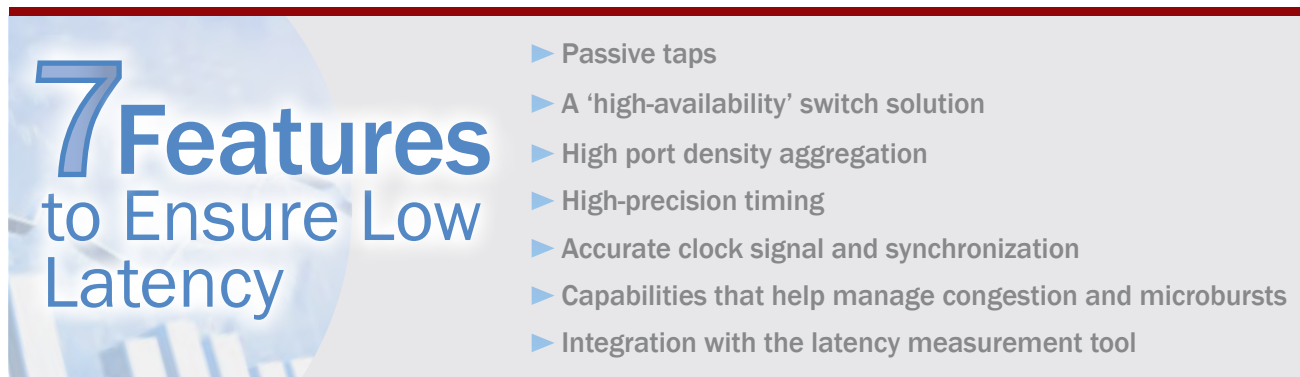
To achieve all those things, firms are making huge capital investment in co-location, networks, market data acquisition and component technologies. They also are acquiring high-performance trading fabric, hardware that delivers ultra-low latency, granular data visibility, intelligent traffic monitoring and precision synchronization.

That fabric guarantees the high level of visibility and control needed to accurately measure the performance of all trading components – firewalls, switches, gateways, feed handlers, matching engines, strategy engines, order routers, order management systems and more. Each component offers an opportunity for even the tiniest performance advantage, but only if you can see and measure it. Traders need to understand the delays and visibility gaps caused by microbursts, for example, or the order fill rate across multiple liquidity providers. Intelligent network monitoring plays a significant role in helping traders to tweak trading strategies – even in real time – while managing or minimizing technology risks to profitability.

Firms also are looking for greater independence between trading and monitoring functions of their networks. By isolating the two from each other, the firms can improve resilience during error or outage conditions while also ensuring that monitoring produces an independent interpretation of trading data that does not rely on derived inputs. That independence is an important regulatory compliance issue.



## Key Features to Ensure Low Latency



### 7 Features to Ensure Low Latency

- ▶ Passive taps
- ▶ A 'high-availability' switch solution
- ▶ High port density aggregation
- ▶ High-precision timing
- ▶ Accurate clock signal and synchronization
- ▶ Capabilities that help manage congestion and microbursts
- ▶ Integration with the latency measurement tool

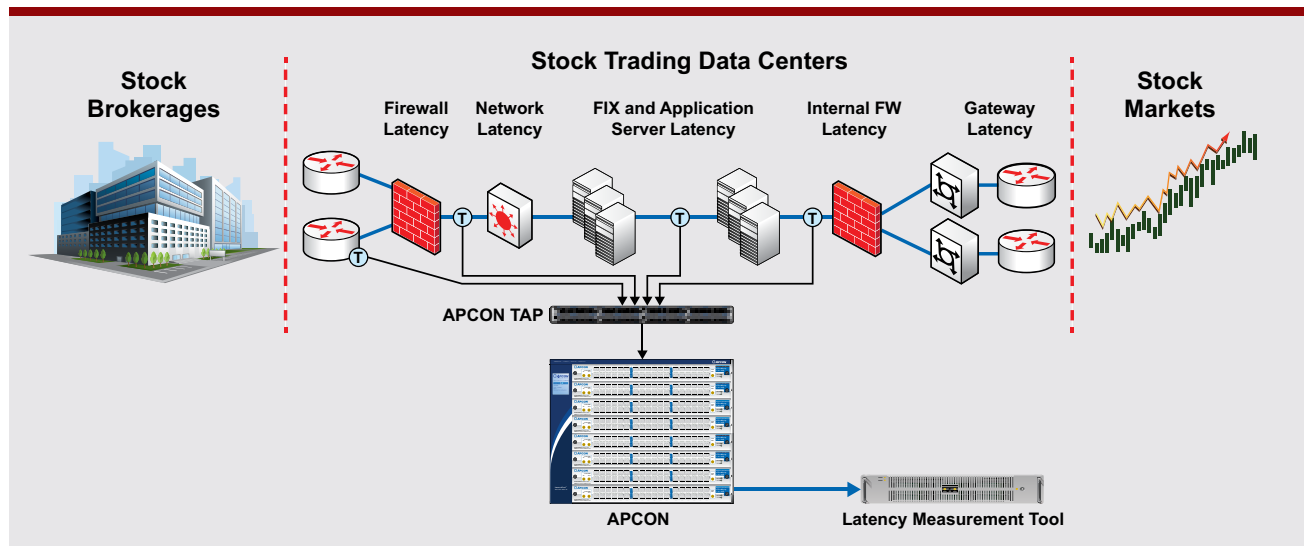
Trading firms looking to design their latency measurement and monitoring platforms should use the following features, capabilities and methods:

- Passive taps – By using passive taps, two important conditions are met. A trading company in a highly regulated industry must record and store everything it does. You never know what you might need to see. Taps use optical splitting to ensure you get a mirror image of in-line data while meeting the most critical requirement – zero latency insertion.
- A 'high-availability' switch solution – When we talk about high-availability, we mean that you don't suffer downtime when components fail, for example when a GPS signal is lost. High availability is a critical part of ensuring a constant ability to measure latency. It is essential to have a high port density aggregation switch with time source holdover capability so the network retains the ability to mark time during a signal outage.

Some latency measurement switches on the market are designed as “network” switches, meaning if something fails, the data is designed to be rerouted. But if there are monitoring tools connected to that switch, when the switch fails, you either lose total tool visibility or you lose accurate timing and latency measurement.

- High port density aggregation – Achieving hop-by-hop latency measurement requires establishing many capture points along the trading path. A latency measurement tool usually has two to four monitoring interfaces and it is not enough to take all the input from the capture points. A high port density aggregation switch provides a platform sufficient to accommodate all needed inputs. It also allows tapping of both primary and standby links, which can prevent monitoring interruption or the need to re-cable during failover events.

## Key Features to Ensure Low Latency



- **High-precision timing** – Accurately calculating latency requires precise time stamping of data packets. For the most accurate measurement, time stamping should be done as close as possible to the source data and not on the latency measurement tool itself. Why? Because latency introduced within the aggregation switch could skew the measurement. Time stamping also should not interfere with production traffic performance in any way. Using a passive TAP to copy traffic and pass it to an aggregation switch for time stamping will achieve the best result without affecting production traffic.
- **Accurate clock signal and synchronization** – Network monitoring equipment that supports protocols including Global Positioning System (GPS), Precision Time Protocol (PTP) and Inter-Range Instrumentation Group time codes (IRIG) is critical to having the most accurate clock source, especially for systems in multiple locations. The most precise method to use is GPS. Regardless of the timing source, you want to be sure all blades can be linked together in a chassis and run off the same source. That will help eliminate time source drift or skew.
- **Capabilities that help manage congestion and microbursts** – Comprehensive and powerful filtering capabilities can filter and reduce the traffic load, reducing the congestion and tool oversubscription that leads to packet loss. The right solution also provides load-balancing capabilities with high availability features that spread the overall traffic load to multiple monitoring interfaces. Again, the goal is to avoid oversubscription and packet loss. Thirdly, buffering capabilities can help accommodate microbursts and prevent packet loss.

## Key Features to Ensure Low Latency

- **Integration with the latency measurement tool** – Network monitoring equipment should be able to read the timestamp inserted for each packet and perform latency calculations correctly. It's also important to note that vendors use different approaches to adding timestamps to packets. Some devices apply timestamps that are calculated from previous timestamps, creating a proprietary packet design dependent on that time-stamping vendor. That can create interoperability issues with measurement tools. A simpler, better approach is to insert raw timestamps, based on GPS or other timing source input, directly on the packet.

Instrumenting many components for hop-by-hop application performance across a complex network can be difficult and expensive. However, the cost can be lowered when the APCON IntellaFlex Series 3000 aggregation switches are deployed with a latency-monitoring tool such as Corvil or TS-Associates. The IntellaFlex Series 3000 switches offer from 36 ports in a 1RU chassis up to 288 ports in an 8RU chassis, and when deployed with IntellaFlex Multi Function Blades, provide all ports with nanosecond time-stamping capabilities.

By adopting an intelligent network monitoring switch with high port density, trading firms gain the advantage of latency measurement precision analytics that can point the network engineer at the best opportunities for latency optimization.



## How Precision Time Stamping is Configured to Measure Latency

As we've already noted, precision time stamping plays a critical role in ensuring both maximum network performance and regulatory compliance by measuring latency and helping to identify opportunities for improvement.

However, time stamping can play that pivotal role only if it is performed properly, with the right access to reliable timing sources and if latency measurement tools are able to properly read the time stamps.

In the time stamping process, an eight-byte stamp is added to the end of the packet, just before the CRC checksum. The first four bytes indicate the number of seconds since 12 a.m. on Jan. 1, 1970. The second set of four bytes indicates ingress time, accurate to 3.2 nanoseconds, as defined by the first set of bytes. Then, a new checksum is calculated for the packet.

Accurate, synchronized time stamping across various nodes of a network requires obtaining an accurate clock signal. There are several ways of obtaining that signal.

The most common method is to set up a receiver for a GPS, which is synchronized worldwide to coordinated universal time (UTC). However, because the latencies involved are so tiny – at the microsecond level – even the length of cable from the network switch to the GPS antenna introduces a few nanoseconds of “clock skew” that can vary from one network node to another. This skew must be accounted for when the physical layout of the data center is known.



Once a single node on the network has reliable access to this time (or even an arbitrary clock maintained locally), the network can use Inter-Range Instrumentation Group (IRIG) time code signals to keep the entire data center synchronized.

Some users will have the GPS antenna installed in the data center and provide the GPS signal to a Network Time Server such as Symmetricom SyncServer S350 with the IEEE1588/PTP Grandmaster option to provide PTP clocking signal to the network.

APCON products support timing protocols that include GPS, PTP and IRIG-B. Timing may be synchronized among many time stamping blades and APCON switch chassis with PPS.

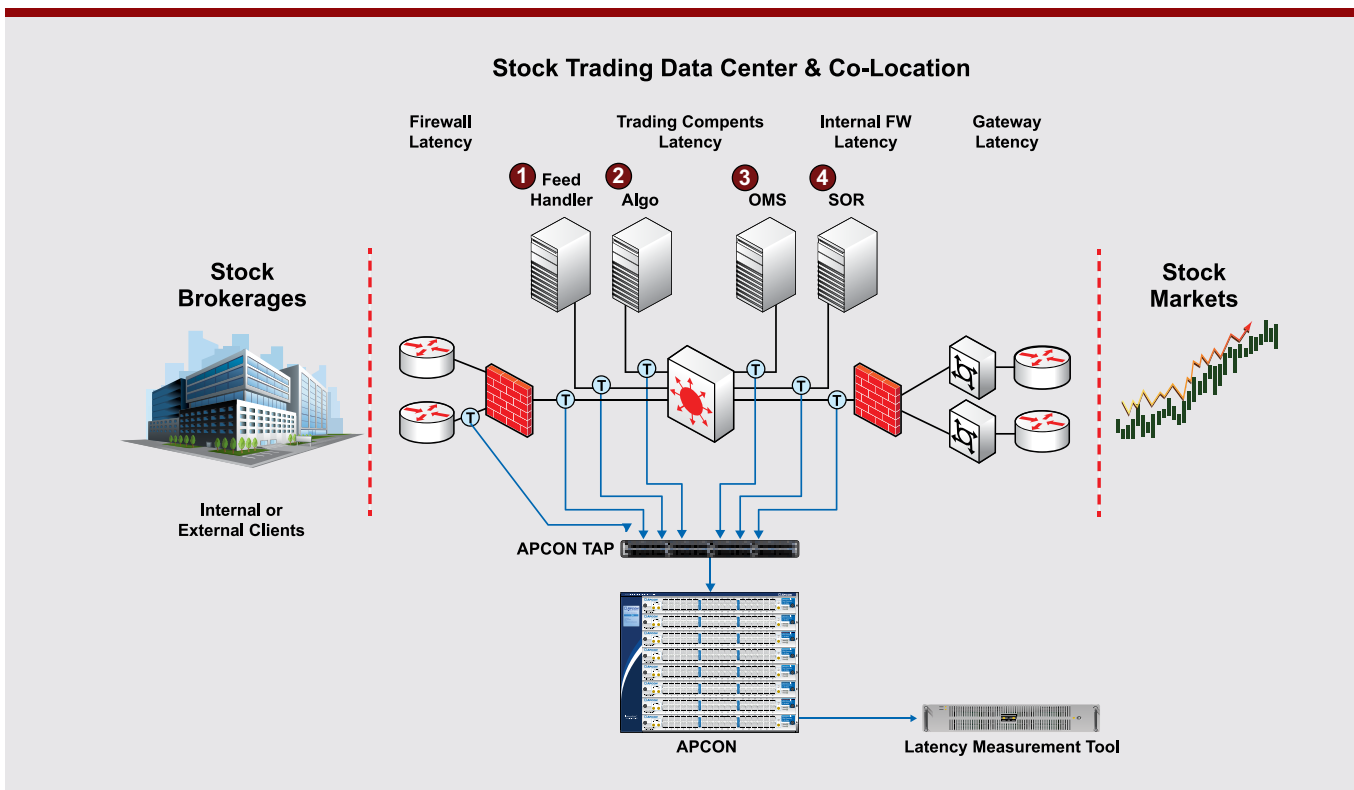
To ensure time stamps can be read accurately by latency measuring tools, APCON also supports integration with multiple tool vendors, including Corvil, TS-Associates, Velocimetrics and WildPackets.



# Architectural Makeup of a HFT Latency Measurement and Monitoring Network

Reliably measuring latency across the HFT network requires an intelligent network monitoring architecture that incorporates high port capacity switching, strategically placed taps, aggregation capabilities and time stamping.

Here's what a typical infrastructure looks like:



## 1 Feed Handler:

Feed handlers receive, normalize, cache, manage, and integrate real-time market data sourced directly from exchange feeds and ECNs. Feed handlers provide access to quote, trade, full depth and breadth of all feed vendor information, including global exchange data, global depth of book, and fields of information published for symbols for many national and international feeds.

## 2 Algo: Algorithmic Trading Matching Engine:

A matching engine is a program that accepts orders from buyers and sellers and subsequently conducts trades. It is an algorithm that operates on an order book and matches and determines prices at which orders are matched. A matching engine will base on certain algorithms such as price-time priority matching algorithm or pro-rata order matching algorithm. Matching engines support different types of orders such as standard orders like Market, Stop and Limits as well as complex type orders like PEG, Trailing Stops and Iceberg OCO.

## 3 OMS: Order Management System

A software-based platform that facilitates and manages the order execution of securities, typically through the FIX protocol. Order management systems are used on both the buy-side and the sell-side, although the functionality provided by buy-side and sell-side OMS's differs slightly. (Typically only exchange members can connect directly to an exchange, which means that sell-side OMS's usually have exchange connectivity, whereas buy-side OMS's are concerned with connecting to sell-side firms.) OMS's allow firms to input orders to the system for routing to the pre-established destinations. They allow firms to change, cancel and update orders as well as access information on orders entered into the system, including detail on all open orders and on previously completed orders.

## 4 SOR: Smart Order Router (or Smart Order Routing)

SOR helps clients route their orders to a preferred destination of stock exchange, based on certain logic under the Best Execution Policy. It allows the trading engines to systematically choose execution destination based on factors such as price, costs, speed, likelihood of execution and settlement, size, nature or any other consideration relevant to the execution of the order.

## Challenge for High Frequency Trading Network Engineers

As we've seen, high-frequency trading is a bit like dancing on the head of a pin – there is extremely little room for error.

To assure the accuracy of trade times requires eliminating latency – latency equals lost opportunity – and providing the capacity to cope with often-volatile networks with tremendous potential for traffic spikes. And, because of regulatory issues, high-frequency trading networks must be able to capture and store every packet.

Achieving a competitive advantage requires complete visibility of traffic between all network components and a high port capacity switching solution that ensures latency measurement tools see all necessary data without losing packets.

APCON's high-availability products offer the reliability and high port capacity to ensure that high-frequency trading houses have the visibility they need to eliminate latency and give their trading strategies a leg up on the competition.

