The Value of a Millisecond:
Finding the Optimal Speed of a Trading Infrastructure

Viraf Reporter | V06:007 | April 2008 | www.tabbgroup.com

Vision

Capital markets are evolving with ever-increasing volume and unpredictable volatility, across all asset classes and geographies. Hardware and software innovation have enabled a boom in sophisticated strategies through electronic means – once only the domain of the largest sell-side institutions. But innovation favors no sides, as the cost barriers to developing these technologies continue to fall.

In this sea change, most markets are increasingly fragmented and reliant on multi-tasking applications and connectivity to unite them. The analysis of these disparate sources of liquidity is one of several challenges to traders of all stripes. Among those market participants aiming for first place, the ability to quickly process, analyze and react to the onslaught of data is a critical component of their competitive advantage. Whether those market participants are exchanges trying to attract liquidity, hedge funds seeking alpha or algorithms chasing best execution, speed is a necessary trait for survival.

Yesterday's ultra-low-latency is merely today's low-latency. As soon as one market participant ratchets up the quickness of its trading environment, competitive forces dictate that others will play catch-up. Thus, the cycle of competition is never complete, and the constant change in today's markets requires firms to implement technology with great foresight. In such a competitive environment, any viable strategy must emphasize the importance of adaptability and flexibility.

Latency is often discussed, but not easily understood. Its definition will likely remain in debate for some time. Yet its impact on market risk and operational risk is felt in quantifiable terms. This characteristic can occur in various ways and at various points throughout the enterprise. It is generally measured as the period of time it takes a packet to travel from source to destination or the amount of time one part of the system waits for another part to catch up. Longer wait times translate into lower application value, regardless of scale. Therefore, minimizing latency is a primary objective in deploying critical trading applications.

Trading institutions have invested heavily in applications that deliver competitive value. It is paramount that these proprietary solutions receive and transmit the relevant data through the fastest possible means. While the sell side has traditionally played the role of technology vendor for its buy-side clients, their rigor and complexity makes proprietary infrastructures a basic principle for all parties concerned.

Latency, or rumors of it, is a destructive force that can incapacitate a trading firm at any time.

TABLE OF CONTENTS

Introduction

As much as people say "Time is money," there are few jobs and business models that are directly tied to the spinning hands of a clock. Sure, people charge by the hour; but it is not the time itself that is valued, but the work done during that hour. When time is the measure of a person's fees, there is an incentive to work more slowly. However, when time is the measure of value, then there is an obsession with speed.

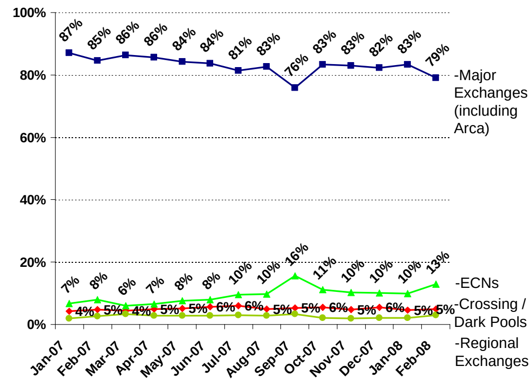
Races – whether the Iditarod or the Indianapolis 500 – are the purest examples of "time-is-money." In 2007, only 0.36 seconds separated first and second place in the Indy 500, but first place won $1.6 million, while second place had to settle for $719,000. Thus, the implicit value of time (from the second place finisher's perspective) was $2,447 per millisecond. We can also observe that the value of time in the Indy 500 (and in most races) is decidedly non-linear, with time being more highly valued toward the beginning of the race, and then rapidly diminishing.

In securities trading, time and the related concept of timing have a long history of significance. The pricing of options contracts has a time-related component. The analysis of market microstructure attempts to discover the duration of the impact a trade has on the price of a security. But most relevant to the issue of speed is the increased velocity of trading across electronic markets and its impacts on traders such as: (1) market makers; (2) black-box traders; and (3) high-frequency quantitative funds whose profits come from various forms of arbitrage, and for whom the value of a millisecond could rival that of the Indy 500.

Fast Money

The financial market that best represents the growing dependence of profits on speed is the US stock market. Regulatory changes aimed at increasing competitiveness have given way to a market structure that favors automated trading. The subsequent increase in volume and velocity of the US equity markets has forced a rethinking of the trading infrastructure. In just a matter of years, the time it takes to process orders has changed from being measured in seconds to milliseconds. The matching engines within newer execution-venue infrastructures have moved into

Exhibit 1
Matched US Equity Flow



Source: TABB Group, "Liquidity Matrix" single-digit microsecond capabilities. This is a competitive advantage that must be equaled by the traditional exchanges, or the 8% loss of market share to other execution venues in the past year will only continue to grow (see Exhibit 1).

Applications addressing market data, liquidity and execution have undergone enormous change. The latest technologies deliver a new, higher standard of capability and sophistication. While this evolution complements the regulatory changes that have brought new opportunities in an ever-more-competitive landscape, it also brings new threats and challenges for all categories of players in the game.

As an example of the velocity of the equities market – the Indy 500 of the stock market so to speak – consider the SPDR (or SPY), an exchange-traded fund representing the performance of the S&P 500 index, and one of the most active equities instruments traded. In the first quarter of 2008, SPY traded at a per-share price of around $130, and had an average daily volume of about 250,000,000 shares. That equates to an average daily dollar volume of over $32 billion. One would also observe that on a volume-weighted basis, SPY moves about $0.50 in either direction, on average, over a 15-minute period; this is a simple representation of its volatility. This equates to almost $1,400 of traded volume per millisecond in this one security alone.

There are speed limits in the trading world. It is commonly understood that a network's ultimate speed limit is the speed of light. However, that is reaching too far in applying any reality to the capabilities of even the fastest networks today. There is inherent latency due to the material composition of the hardware used to make the trades, as well as to the inherent timing of software processes, so that there is always some time that elapses before a single message is even transmitted. But just as a rickshaw has no place in the Indy 500, humans are nearly obsolete in these aspects of trading.

The blink of an eye takes about 300 milliseconds. That means on average, over $400,000 in SPDR volume alone trades in the blink of an eye!

Latency-Dependent Revenue

Latency is a relative term, so let's be clear about the definitions of relative speed: We currently qualify low-latency environments as those with sub-ten (single digit) millisecond round trips. Microsecond measurements are already being introduced in portions of the messaging process -- especially for hardware – which will lower overall readings in the not-too-distant future.

Furthermore, the definition of low latency for one component may be too slow for another component, and vice versa. Because different trading functions and business models value speed in different ways, not every market participant will focus on the same pieces of the trading infrastructure. Exchanges, market makers and agency-only algorithm providers operate in very different ways and will get the most bang for their buck out of speeding up different parts of their infrastructure.

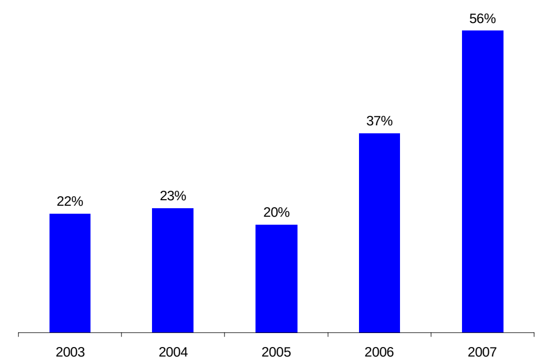
Exchanges and Latency

The business model of an execution venue (either as part of an exchange or an Alternative Trading System) depends, in part, on its ability to receive, aggregate, manage and match orders for a series of securities. The traditional exchange model relied on humans called specialists to carry out a fair and balanced market in order to attract buyers and sellers to come together. The newer model does away with qualitative measurements of market behavior and matches orders according to a very simple methodology: price and time.

In the traditional exchange model, specialists were given certain trading and informational privileges along with their obligations. Although quicker thinkers undoubtedly made better specialists, it does not mean that managing the order book expeditiously was the best way to maximize profits. In fact, there are many cases – e.g., a temporary deviation in the balance between bids and offers—where slowing the pace of execution favors the specialist. However, this slowdown is a negative for investors who must live with the uncertain status of their orders.

The newer model, best represented by the Electronic Communication Network (ECN) – with the simple mantra of "Who has the best price?" followed by "Who put that best price out first?" – has radically changed the execution landscape. By explicitly rewarding the timeliness of the participants, the ability to quickly receive, manage and relay information about the order book has become critically important to an exchange or an ECN's own competitiveness. In response, exchanges have evolved toward the ECN model. Cash trading and market data revenues, which are the most susceptible to low-latency competition, have been steadily increasing over the last four years, and now represent more than half of the total revenues of US equity exchanges (see Exhibit 2).

Exhibit 2
Percentage of Latency-Related US Equity Exchange Revenues



Source: Exchanges and TABB Group

In the early years of the movement from the traditional exchange model to automated matching engines, there appeared several opportunistic strategies directed at “latency arbitrage” between these liquidity points. Predatory algorithms able to take the opportunity simply executed a trade in one execution venue and offset it in another for instant profit within several millisecond intervals throughout the trading day. This was due to the differing latencies of the respective liquidity pools, a phenomenon that should become extinct as all execution venues move toward lower-latency automated matching engines.

Latency and Other Players

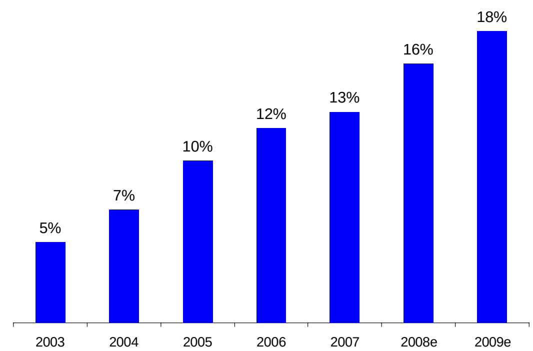
The shift toward automated matching made speed an issue for all market participants. Liquidity providers – those who generate profits by capturing discrepancies between the supply of and demand for a particular security – are also impacted by the shift. The specialist firms who relied on their favored status within the traditional exchanges are being driven to extinction. Market makers – a slightly different breed of liquidity provider – can also be susceptible to faster markets, but according to recent filings, a good portion of their flow is still called in over the phone. This gives the market maker time to monetize the information contained in these orders. But despite the continued existence of phone orders, some market makers have made the necessary infrastructural changes to compete in a low-latency environment.

Another good example of a latency-dependent participant is the agency execution provider. In particular, we are talking about the software – mostly provided by brokers – that intelligently manages a client order’s interaction with the markets, determining when, where and how much to send at any given time.

The part of this software that decides where to route an order, called a smart order router (SOR), makes decisions based on real-time market data. Because the brokers’ clients (the buy side) tend to take liquidity out of the market – by most estimates at least 60% of the time – it is important that those orders reach their destination quickly when a trader decides to aggressively interact with posted liquidity. The greater the latency within the brokerage infrastructure, the greater the chance that the SOR will reach the destination too late – losing out to a competitor’s SOR, which is faster by a hair at executing the quote. For a broker whose value proposition is closely tied to offering best execution to the buy side, that is the kiss of death.

Consistently insufficient execution speeds would be identified by higher spread costs and similar metrics in buy-side Transaction Cost Analysis (TCA) reports, relegating that broker to second-best, and therefore undermining any hope of positive differentiation vs. the competition (see Exhibit 3).

Exhibit 3
Percentage of US Institutional Equity Commissions Exposed to Latency Risk



Source: Exchanges and TABB Group

In 2008, 16% of all US institutional equity commissions are exposed to latency risk, totaling $2 billion in revenue. As in the Indy 500, the value of time for a trading desk is decidedly non-linear. TABB Group estimates that if a broker's electronic trading platform is 5 milliseconds behind the competition, it could lose at least 1% of its flow; that's $4 million in revenues per millisecond. Up to 10 milliseconds of latency could result in at least a 10% drop in revenues. From there it gets worse. If a broker is 100 milliseconds slower than the fastest broker, it may as well shut down its FIX engine and become a floor broker.

Quick and intelligent order routing is made more complex by the fact that there are now multiple execution venues, all of which are operating faster than ever before, disseminating order information at faster and faster rates. The other market participants, including the liquidity providers, are reacting to this information with increasing speed, creating more and more updates and thus straining the entire trading infrastructure. Since any SOR analysis requires real-time aggregation and reaction to the volumes of information, speed takes no back seat. The introduction of SORs and algorithms for listed equity derivatives only increases the latency-dependent nature of trading commissions, and makes the challenge even more difficult given the volume of data in the options market. For brokers that define their value proposition by their agency execution quality, the value of a millisecond is at a premium.

Speed of the Enterprise

When trying to achieve faster results, a sole focus on components that are *prima facie* related to speed is a common mistake. From both a technological and business perspective, trading has reached a point where the entire infrastructure needs to be monitored and measured to find and eliminate any signs of slowdown. The most significant improvements in speed will be gained by eliminating the squeakiest wheels.

From an enterprise infrastructure perspective, the core challenge resides within messaging and its related touch points – which we will generically refer to as middleware. Middleware broadly refers to the software composition that connects applications and passes data between them, thereby playing the role of infrastructure plumbing. We will be primarily concerned with this critical component responsible for message-oriented activity.

For example, within an execution venue, middleware connects the many applications that make up the core system. It begins with the applications that receive orders from market participants, commonly referred to as FIX engines, after the ubiquitous Financial Information eXchange Protocol. The order messages are then fed into a database that contains all of the orders as well as a mechanism which matches buys and sells at the same price – the matching engine. When an order is matched, or there is any change in the status of an order, messages are sent back out through the FIX engine. In addition, there is a separate set of applications which disseminate information to the market about the price and size of the orders on its books.

Every application that makes up the core system of an execution venue must be fast, as well as the speed and accuracy of the messages passed back and forth among these applications. Addressing the controllable latency problem within the trading enterprise requires the ability to isolate the issue where it lies. Without a pinpoint capability to "sense" or measure such bottlenecks, it would be difficult to discern whether the problem is that the network does not have enough bandwidth, an application is poorly coded, or if there is simply not enough horsepower in the servers. Thus, there are three areas that typically house latency:

- the network carrying the messages,
- the applications utilizing the messages and
- the hardware processing the messages.

Regardless of how a business derives value from speed, it must address these three areas as a whole. Otherwise, gains in one area could simply lead to lags in another. Indeed, one of the complexities in investigating and solving latency issues is the impact of one area on another.

The Virtuous (or Vicious) Cycle

Making the network carry messages as fast as possible may very well compromise reliability and overload the hardware. Improvements in hardware may necessitate changes to the application layer. The expected benefits from costly hardware and network upgrades alone are often dashed by turbulence caused elsewhere in the system. Problems can surface elsewhere in the infrastructure if all elements throughout the message-oriented path are not tuned or optimized to work together. Dealing with overall performance is like the classic Whac-a-Mole game – hammer down one problem somewhere and, as a result, another pops up somewhere else.

Brokerages and other providers of SORs and algorithms are old hands at dealing with the Whac-a-Mole game. It all begins with the increasing rate of change emanating from the execution venues. If the application that processes these messages – called the feed handler – begins to fall behind the inbound messages, then IT will examine the network, the code and the hardware to assess the cause. Perhaps it appears that a simple increase in CPUs will do. But then it is found that the application is not properly coded to take advantage of the increased CPUs. And once corrected, IT may finally discover that the network cannot handle the increased messaging rates.

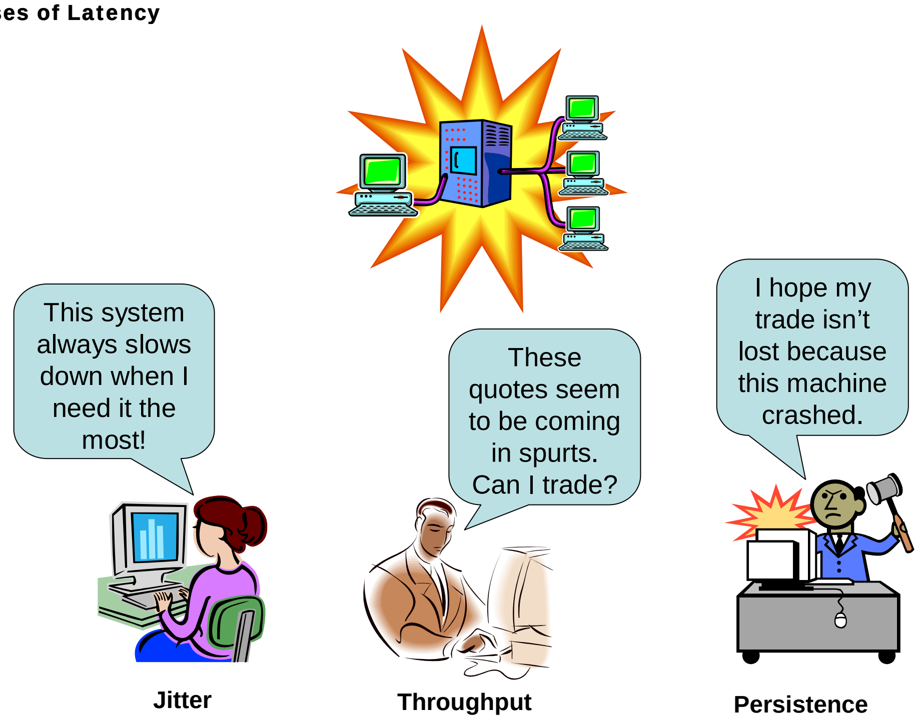
But the issue does not end there. Now that the feed handlers are up to speed, the next application down the line – the quote aggregator – may be too slow, and IT will once again need to re-examine each of the three main houses of latency. The constant improvement (or in cases of neglect, degradation) of the system is a virtuous cycle that leads to enterprise-wide gains in reliability and speed (or vicious cycle that leads to the reverse).

Any failure in networking, applications or hardware leads to latency, typically in the form of three main factors: throughput, jitter and persistence. Throughput is a measurement of the file size or amount of data capable of being transmitted within some delivery timeframe. Jitter is the level of noise or static in the system, likely occurring during the busiest times of the trading day. Persistence is an important characteristic of fault tolerance or reliability within any application environment. It represents the availability of data for an application.

Although there is no cause of latency that is more or less important than another, it is arguable that the most basic cause of latency is throughput. If the network of an execution venue simply does not have the bandwidth to handle its message traffic in a timely fashion, then it would need to be addressed prior to any modification of the applications (e.g., FIX engine, matching engine) or upgrade in server hardware. Basically, if the network can’t handle the load, it would be difficult to spot other shortcomings in the infrastructure.

Of course, an execution venue's ability to process massive amounts of data does little if messages are lost. Persistence is an important characteristic of fault tolerance or reliability within any application environment. It represents the availability of data for an application. Messaging-oriented middleware must provide data to applications regardless of the state of the environment. Automated trading environments are dependent on data accuracy; therefore, the data's survival is critical. If an application arises after the data was sent, it must receive the data as part of the historical stream. If the data source is no longer running, the data must still be made available (see Exhibit 4).

Exhibit 4
The Causes of Latency



Source: TABB Group

When an execution venue sends out an update on a stock – for instance, when its best bid is now a penny higher than before – the brokerage firm's feed handler will receive that piece of data and pass it to downstream applications. But let's say one of the SOR servers has crashed during that timeframe. When the SOR server is revived, the application needs that information to construct an up-to-date picture of the market. And if the feed handler also crashes, then the middleware must be able to get that data from elsewhere. Otherwise, incoming orders will be improperly routed by the SOR.

Jitter is one of the more challenging causes of latency. Applications can handle almost any kind of event – except those they are not programmed to handle. Applications are particularly unsuited to handle exceptions because it is difficult to predict how a network, application or piece of hardware will react during "stress" periods. A stress period can have several causes—a trader could mistakenly enter a very large order (known as “fat fingering”) that throws the market into a frenzy – but the most common stress periods occur at the opening bell, in the closing minutes or following some market-moving event.

As markets continue racing ahead due to the technological improvements in analyzing and executing trades, participants have to have equally fast market and operational risk management capabilities – or the consequences will “net” a dire result at the most inopportune time. Governance is lacking when front- and mid-office latency concerns are addressed without including back-office concerns such as risk management.

One of the important features of middleware is the ability to tune the interaction of the applications to find an optimal balance. Advanced middleware acts as a traffic cop of sorts, making sure that each part of the system is running as smoothly as possible. Tuning essentially involves tradeoffs; e.g., it could mean more persistence but less throughput. However, some advanced middleware features also allow you to make these tradeoffs at the message level itself. This allows for optimization to occur not simply between applications, but within the content of the messages—everything from order types to specific securities could be used to fine-tune the trading applications.

Solving for Speed

One variable common to these challenges is the capability to respond instantly, at every level of the trading enterprise. It has been stated that in the current, highly automated trading environment, "a latent decision is a bad decision." In fact, making an accurate call on buying or selling securities is no longer sufficient to guarantee trading profits. Instead, capturing that fast-moving opportunity before the competition does is the greatest influence. Indeed, the appetite for speed seems to have no end in sight. The points of failure are not only numerous, but also complex.

When one reviews latency solutions, there are some seemingly simple ones: increase bandwidth, streamline code and boost processing power. However, there is another relatively simple way to reduce time to execution: shrink the length of the message path.

One popular way trading shops are minimizing the message path between their production servers and the execution venues is to be in the same physical space—also known as co-location. Co-located facilities are the ultimate "physical" solution. A trading engine's proximity to the liquidity source is just a practical approach to the physics of receiving and sending messages over varying distances. Such services are becoming part of the value-added offerings from execution venues—adding dollars to the latency market. Connectivity providers that host an exchange infrastructure have opened their facilities – much like infrastructure hotels – such that a trading firm's technology will reside at the same facility, in close proximity to that of the exchange. Such partnerships will become commonplace between liquidity pool and connectivity provider.

While these hosted infrastructures help to deal with the external communication latency issues to a great extent, the brokerage will not be the only co-located service, and thus must still vigilantly fine-tune its internal network, applications and hardware configuration so that it can digest the information without adding latency to the process.

Once the message path has been shortened as much as physically possible, the first place to start looking for further latency reductions is throughput. Traditional Ethernet network infrastructures have already been pushed to their limits. More costly InfiniBand solutions are being considered within large sell-side enterprises to answer this demand.

Another way to solve throughput problems is to shorten the actual size of the messages. In fact, market data feeds used to enhance their throughput measurements by clipping the message and providing only a subset of the critical bits of information. Nowadays, one way to try and boost throughput, at least at the external part of the messaging path, is utilizing advanced compression methodologies. One such protocol is FIX Adapted for Streaming

(FAST), which is currently being applied to US options but is evolving for use in other asset classes and regions.

Co-location and InfiniBand won't garner much profit unless the applications themselves can do the heavy lifting. Underlying many of the applications in the trading infrastructure is Complex Event Processing (CEP). At its simplest level, CEP complements middleware, taking in two or more input streams, and enables filtering, aggregation and analytics. By offloading some of the work from the applications, CEP helps speed up critical decision-making for event-driven functions. The shift toward CEP is being driven by the need for speed in the areas related to automated trading such as smart order routing, algorithms and risk analytics.

The glue that's composed of the message-oriented middleware truly joins front- to back-office needs, spanning the pre-trade to post-trade processes and keeping the firm's overall governance in sync with its transactional speed.

The number of messages in market data feeds has been exploding. Equity options growth is straining the capacity limits of market feeds. This is further exacerbated by competitive algorithmic needs in a sector where speed is a necessary component of analysis within a market structure that remains, for the most part, fractured. In order to handle this immense crush of data, every part of the trading value chain – including the buy side, brokers, exchanges, execution venues and liquidity providers – are looking at each of the solutions above.

Pinpointing the Problem

Messaging infrastructures have long been the domain of a few large software providers. Their installations have been in place throughout the financial sector for decades. Whether or not oncoming evolutions of those legacy products are able to match the superiority of the upstarts remains to be seen. But as in any technology, one thing is clear: the innovation in financial data acceleration is arriving from all areas of software and hardware research. Financial services institutions are always willing first adopters, as competitiveness affects profitability. Given the enormous investments they have already made in these legacy installations, overall spending on messaging infrastructure will remain flat, at around $1.8 billion a year through 2010. However, within this space, low-latency expenditures will nearly double, from under $100 million currently to about $170 million by 2010.

The growing size of latency-related risk in the trading world has attracted numerous vendors to the space. The complexity of the trading infrastructure allows for a diverse set of solutions that are appealing to various types of IT departments and suitable for different business models.

The mainstream status of automated trading has brought many of the latency-related vendors together as packaged or holistic solution partners, each of them optimizing their complementary benefits for both buy- and sell-side configurations. The next development will be from single vendors who have developed these disparate components to work together from one proprietary solution.

In order to track down all the factors of latency, enterprises need to be able to narrow the points within the infrastructure where latency is occurring: is it the network, the applications or the hardware? More often than not, investigations of messaging bottlenecks arrive at the home-grown applications as a source of most latency-related problems. However, we can assume nothing.

In the case of exchanges and other execution venues, it would seem that a great opportunity to allow enterprises to address and tune their respective latency issues is lost from the outset. There is obvious hesitation by exchanges to expose possible vulnerabilities, but it is possible for exchanges to publish a synchronized time-stamp of incremental accuracy for a particular message they have published. Such publication would allow consumers of that data to benchmark all subsequent downstream uses of the data where latency requires investigation. This would establish a true “time zero” for all downstream purposes thereafter – a capability that could be accessible within the publishing infrastructure from the exchanges in question. The current hesitation directs the debate internally rather than externally, where the value far outweighs the insecurity to deploy.

Any of the newly-launched exchanges is likely to see this as a differentiating opportunity and claim verifiable bragging rights. The kimono will open for competitive reasons, but will also yield large new revenue opportunities for these exchanges. Many eager developers in the larger automated trading communities will be prepared to pay for such an innovation in furthering their ultra-low-latency objectives.

The importance of pinpointing bottlenecks has led to the demand for troubleshooting and tuning messaging environments. This has brought several variations of measurement tools into popularity. These solutions are available in software and hardware varieties, and given the task at hand, they require passive capabilities in their monitoring. Otherwise, their purpose would be defeated in that they would contribute to the problem they are measuring. Some of these solutions originated as transport-layer dashboards but have since evolved with more-robust benchmarking features, allowing for pinpoint end-to-end monitoring in exposing trouble spots.

These include visibility into the performance parameters at any point of the enterprise configuration. The degree of measurement accuracy can reach into the microsecond (10^{-6}) to nanosecond (10^{-9}) range. This is well beyond the practical limits of current enterprise configurations, but not out of the question for capabilities just around the corner.

The challenge for any of these providers is to preserve the message stream from existing applications. Preserving the message stream reduces or even eliminates the need to recode these critical areas of the business – at least from an initial standpoint – through the exposure of specialty or custom application programming interfaces (APIs). Only later in the implementation cycle, when exposing bottlenecks, can vendors excuse or justify any such intrusion to the clients' status quo.

Conclusion

Since the transactional process requires visualization of – and intelligent order routing to – the disparate liquidity pools, incoming market data can be a primary point of trouble. While many asset classes are not exchange-traded, the majority of equities are. Off-exchange pools of liquidity are still dependent on analyzing more transparent data points provided by exchange data feeds. Therefore, the message handoff from liquidity pool to enterprise middleware must be fluid.

Recent mandates in equity options are releasing a deluge of new data due to the reduction of pricing spreads and a continuation of the decimalization process we have already undergone in the underlying equities markets. Message rates currently averaging 500,000 messages per second (one-minute peak rates) are expected to reach 1 million mps within months for equity options alone. The application of the FAST protocol to these messages currently reduces the messages to about 30% of their prior size via compression.

Given all of these challenges, messaging middleware will need to deliver an equally comprehensive set of tools allowing for the tuning of responses to the different processes we have discussed. Latency needs to be managed just as much as it needs to be reduced. So the visibility of the system's behavior will need to match the system's speed.

A large portion of current spending is directed at maintaining existing legacy installations, and those budgets will experience a redirection toward the need for greater speed in all areas – from front to back office. Therefore, a greater emphasis on latency issues from risk management areas in the near future will not come as a surprise.

▲

TABB GROUP

www.tabbgroup.com

New York
+1.646.722.7800

London
+(0) 207 368 3377

Westborough, MA
+1.508.836.2031

