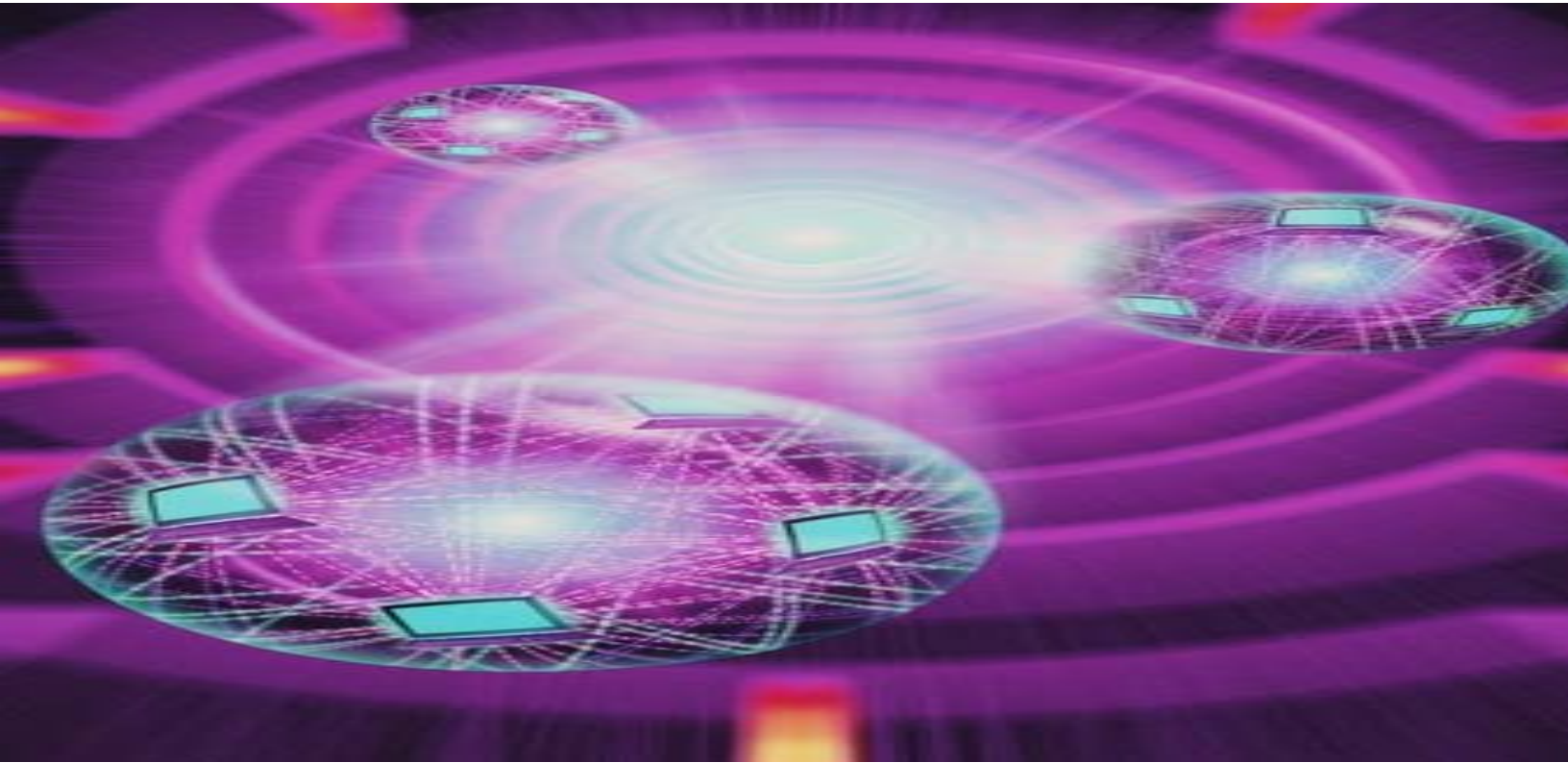

Faster Than a Speeding Bullet:

The New Low Latency Messaging



Vision

How fast can we go? Einstein told us the speed of light is the limit. So how fast can trading get? Assuming perfect conditions, if a server was one kilometer from the source at the exchange, the fastest an order message could reach the exchange would be 3.34 microseconds. If routing logic can be done in less than 660 nanoseconds, then an almost unbelievable 4-microsecond latency could be the end game. Throughout history, we've traveled farther and faster than anyone imagined years before. So as untouchable as they may seem, such speeds promise enough of an economic advantage that the struggle to reach them continues.

With the exchanges going electronic, the global race for alpha intensifying, and technology becoming ubiquitous, volumes and messaging are going through the roof. In 2007, the global equities and options markets produced an average of more than 7 billion messages a day. TABB Group estimates that global messages per day will increase 18-fold, to more than 128 billion by 2010. Fast, robust and reliable delivery is becoming even more critical to the success of buy-side and sell-side institutions.

In fact, reducing latencies is no longer a nice-to-have; it's a necessity. Financial services firms have invested millions in developing state-of-the-art, low-latency messaging to handle the rapid growth of black-box and algorithmic trading in the equities market. Now, we see similar trends occurring in a variety of asset classes and geographies. As more instruments trade electronically, latency will become a multi-asset class, global issue. Doing your low latency homework is the only way of keeping up with the Joneses, and only by earning "extra credit" can a select few surpass them.

Stress-testing systems with unexpected and extreme market conditions for worst-case scenario response times is crucial to seeing the flaws of a particular technology. The recent market turbulence highlighted the fact that messaging is at the center of any low-latency solution. All market participants, including exchanges, market data providers, the buy side and sell side are now critically analyzing the benefits and disadvantages of the various solutions. Even the hares are looking over their shoulders for that sneaky tortoise.

Over the next few years, volumes will continue to explode. The costs of keeping up with competitors will grow while latency drops, driving firms to carefully choose how best to invest the dollars set aside for low latency. Vendors will form new partnerships, or even merge, as the best and most effective solutions are discovered. For the vendors in the low-latency game and their clients, the clock is ticking to find the most cost-effective solution to the low-latency challenge. Although there will continue to be uncertainty and change, one thing will always remain true...time is money.

Table of Contents

VISION	1
TABLE OF CONTENTS	1
INTRODUCTION	3
THE FASTEST PATH	4
HOW DID WE GET HERE?	5
BUILDING A BETTER MOUSE TRAP	6
SOFTWARE BASED MESSAGING MIDDLEWARE	7
HARDWARE-BASED MESSAGING MIDDLEWARE	8
HARDWARE/SOFTWARE COMBINATIONS – THE BEST OF BOTH WORLDS?	9
ON THE CLOCK	10
HOW BIG IS THIS?	11
GETTING THERE	14
THE TRADEOFFS	15
VENDOR SELECTION	15
CONCLUSION	17

Introduction

The creation and continued improvement of a low-latency infrastructure is near the top of Wall Street's priority list. August 2007 saw some of the highest volatility and volumes in recent memory, with NYSE coming within 20% of its peak quote volume. The seemingly bulletproof trading systems of major broker/dealers and hedge funds had a hard time dodging bullets under rapid fire message volumes. Some blame the sub-prime mortgage mess; some blame algorithms for erratic behavior of market prices near the close, and some still blame Alan "Irrational Exuberance" Greenspan. But whatever the actual cause, it is indisputable that the resurgence of volatility has stressed that access to a low-latency trading solution is now a necessity.

The rapidly changing structure of today's markets has placed a premium on speed and complex routing mechanisms. Early FIX messages were only a prelude to an electronic marketplace that views the traditional FIX message as "fat" and "slow." Now, throughput is viewed in millions of messages per second. The levels of data required to make the best trading decisions have far surpassed the capabilities of the hardware, software and networks of the '90s. What was considered fast just a few years ago can no longer keep up.

Institutions whose messaging infrastructures remain even a few microseconds behind their competition will consistently lose out on the best trading opportunities, as incremental technological improvements become available to everyone. The need for low-latency messaging infrastructures—and more specifically, low-latency messaging—is absolutely mandatory. Benchmarks held out as "super fast" just a year or two ago—10 milliseconds—are now too slow to hit the bid on most markets.

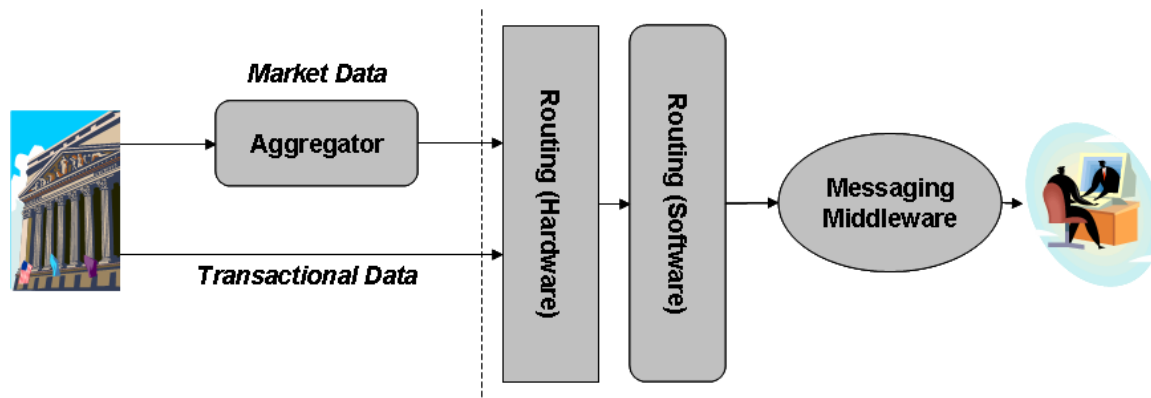
Despite the continuing improvements in messaging speed, the latency reduction from seconds to milliseconds to microseconds is no small feat. TABB Group estimates that nearly \$500 million will be spent in 2007 on solutions to increase overall messaging capacity and to manage that data in the form of middleware and network upgrades. How exactly will this money be used to lower latencies into the microseconds (or even nanoseconds)? By introducing the most advanced technologies to improve speeds at each sub-segment of the low-latency market-data infrastructure value chain, future generations of today's hardware- and software-based solutions will quickly bring us closer and closer to top speed.

The Fastest Path

It wasn't long ago that trade confirmations were printed and mailed to buy-side clients. The contents of those messages have changed little, but the delivery method has changed drastically. The "letter" is now broken into small groups of computer-readable text. That text is sent through networks and servers, rather than the post office, to reach its destination. Software translates those messages and displays them in easily readable formats for the proper recipients. The post office can send a letter overnight--these new financial messages arrive in milliseconds.

Market data and transactional data make up the vast majority of financial messaging and are also responsible for the recent focus on speed. Market data consists of pricing and volume data, while transactional data consists of individual trades and executions. In both cases, that data must be first transmitted from the source to the recipient, and then across various layers within the recipient's network to reach the trading applications such as algorithms and transaction-cost analytics (See Exhibit 1). Acceptable latency for transaction data tends to be about twice that of market data, however, it is still in the range of microseconds.

Exhibit 1
Example Automated Trading Architecture



Source: TABB Group

A low-latency solution is critical for both parts of that journey. For example, receipt of market data from the Options Price Reporting Authority (OPRA) in microseconds is of little use if it then takes milliseconds or more to reach the algorithmic trading engine. The solutions for these external and internal communications, however, are not one in the same and require technology geared specifically to that need. Further, the state of the surrounding infrastructure along with the precise method used to integrate the intertwined applications can often be the difference between reaching total platform optimization and having components that under-perform their maximum speeds.

How did we get here?

The use of messaging middleware within financial services is nothing new. It came into being in the 1980s out of a need to connect old mainframes to newer applications. In the '90s, brokers remained competitive by making incremental improvements to traditional messaging software, a strategy that was coupled with "throwing hardware at the problem." What has changed is the importance of messaging to the trader's electronic execution infrastructure, which has highlighted the critical coupling of messaging and low latency.

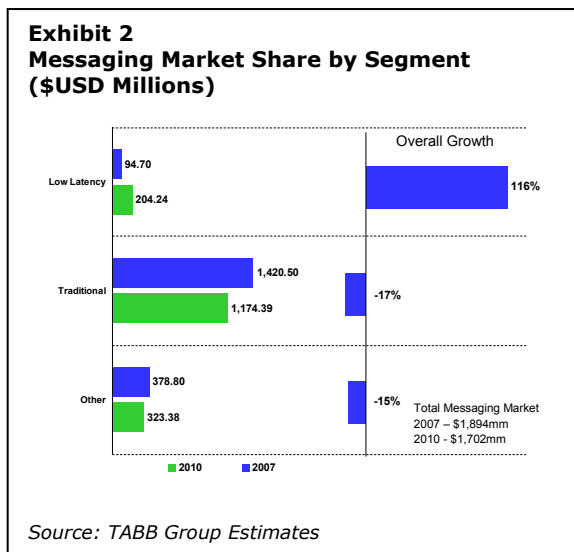
Broker/dealers first developed messaging solutions internally. As technology was less trusted, and with messaging standards still in their infancy, broker/dealers created messaging infrastructures focused on guaranteed delivery and ease of integration into existing applications. In parallel, vendor solutions were being created and brought to the market. Tibco's Rendezvous (RV) messaging product became well-entrenched in financial services as the first targeted messaging system, while IBM's MQ product (formerly MQSeries) has been a preferred solution since it came on the scene in 1992. Both were designed with a focus on guaranteed delivery of messages, which is in part why these products were well suited to the financial-services firms of the time.

During this time, messaging middleware was treated mostly as a transport mechanism, with a strong focus on reliability, using hardware to enhance speed. Increasing the number of CPUs allowed more messages to be processed in less time. However, the ability to improve performance through this kind of parallelization of hardware has proven to be limited over time, as it creates a hardware footprint that is so big and expensive that it becomes unscalable and cost-prohibitive. Moore's Law, if it's actually followed in the future, will not provide CPUs that are fast enough to keep up with the current market-data explosion. Hence, new solutions are needed to solve this dilemma and keep data moving at the speeds required, as legacy systems lag further behind.

Building a Better Mouse Trap

New entrants in the messaging market looking to solve these issues fall into three broad categories: software-based, hardware-based, and a software/hardware combination. All have strong merits and advantages above traditional solutions – the obvious one being low latency. In most cases, these offerings are relatively easy to integrate and highly customizable. One of the biggest benefits to a new messaging solution rests on each vendor having the benefit of building a system from scratch with today’s technology in today’s market. The pioneers of messaging middleware could not have predicted where we’ve arrived at today. Newer, innovative firms such as 29West introduced low-latency messaging solutions to the brokerage industry by building software applications based on a strong knowledge of today’s environment and geared specifically to today’s markets and technology. Their quick penetration has shown this approach to be highly effective against established and new firms alike. Other firms, such as RTI, are in the early stages of bringing solutions established in other industries to the financial-services table.

Although the traditional messaging middleware providers still maintain large market share in the overall messaging space, their market will not grow like that of new providers with a pure low-latency focus (see Exhibit 2). The newer entrants spend their time and energy on ultra-low latency solutions by focusing primarily on high-velocity trading applications. If the need for speed moves beyond front-end applications into downstream applications, the demand for low-latency messaging will increase even further.



Early messaging products focused on guaranteed delivery, rather than speed. Today’s low-latency products focus on speed, yet have kept guaranteed delivery in the picture. These new solutions persist messages in parallel with the send/receive process, so as not to interfere with message delivery and cause additional latency. Many of them allow the database to run anywhere within the infrastructure, removing further complexity from the existing equation.

This paradigm is used not only for capturing best execution-related data, but also for capturing messaging performance data that can be later used to fine-tune the implementation. The resulting database contains not only message contents, but also the length of time it took for that message to make its journey. The most tightly integrated metric-capture systems are also the most accurate and effective, so coupling this type of data collection directly with messaging middleware can provide huge benefits.

Close physical proximity to the data source is low latency's best friend. Rather than turning newly-minted suburban offices into market centers, co-location of servers has been a widely used and successful means of reducing latency in message delivery. Additionally, the use of direct feeds has become commonplace for bulge-bracket brokers, allowing them to gain speed by bypassing market-data aggregators for those markets deemed most crucial.

The FIX protocol, which is arguably the most widely used message standard in financial services, has also made steps towards lower latency by encouraging the use of industry-leading messaging middleware. Previous versions of FIX have coupled the message format and its means of transport, leaving only one standard way of moving those messages between applications or different firms. With the advent of FIX 5.0, however, the message format and the method of communicating that message have been decoupled. It is now possible to use standard FIX-style messages and transfer them using a low-latency messaging solution. A quick uptake of FIX 5.0 would help maintain what has proven a very useful standard since its inception in 1992, while allowing a move into the high-speed world.

These trends will continue, however, much of the millisecond savings have already been extracted from these methods, forcing advanced trading desks and their IT departments to look elsewhere. But with spending on messaging infrastructure still increasing, what else does the Street have to buy?

Software Based Messaging Middleware

The most widely used forms of messaging middleware over the past 15 years have been based on a publish/subscribe (pub/sub) or message-queue model. Pub/sub allows the message originator to send a message (the "pub") with no knowledge or concern of the potential receiver. The recipient of the message (the "sub") can then subscribe to receive only the message it needs. The queue model similarly does not require the sender and recipient to be aware of each other, and queues up sent messages until they are ready to be consumed.

The point of interest in both of these methods is the necessity for a software daemon, a program running in the background, through which all sent messages must pass. In contrast, new low-latency solutions remove this middleman and allow applications to speak directly with each other using custom application programming interfaces (APIs). 29West, for example, has had great success with its solution that operates on this paradigm. Others, such as Progress Sonic, provide some point-to-point messaging on top of the more traditional method discussed above.

Software messaging solutions not only bypass the pub/sub daemon, but also work to avoid additional hops through the network. The routers and switches on which hardware messaging solutions are based—we'll get to that later—are in many cases completely taken out of the picture, saving additional microseconds.

These solutions most often focus on moving messages within an organization. Examples include taking market data received by a ticker plant to an

algorithmic trading engine or moving an order to the connectivity engine that gets it to the exchange.

Hardware-Based Messaging Middleware

Hardware, consisting of routers, switches and cabling, is viewed by some to be the backbone of any infrastructure. For years it was accepted that adding more CPUs and memory was the only way to speed things up, and that Ethernet was the only viable choice for linking it all together. As new products and standards emerge to meet the growing demand for low latency, high-velocity trading desks are opening their eyes to what was once left to engineers in cold, dark server rooms.

Hardware solutions are installed primarily in trading institutions that seek a "content-aware" infrastructure. "Content-aware" networking equipment can reduce latency by placing business logic directly on the network router, rather than using software on an outside server. The bits and bytes that make up any software application must be transmitted to the CPU for execution. Removing this step is the main method through which hardware-based solutions reduce latency.

These devices are created with technology similar to that in the latest smart phones. Did you ever wonder why your cell phone can come online within a matter of seconds but your PC still takes minutes? In large part that can be attributed to the phone's software being encoded directly onto the chip, while the operating system for your PC resides on your hard drive. The same can be true of complex smart order-routing logic, and the faster those decisions are made, the faster the message gets where it needs to go.

Hardware solutions include the ability to translate FIX messages, create complex event processing (CEP) logic and enrich market data, all within microseconds. For the most part, upgrades to the core software must be done by the vendor, however, some vendors offer APIs to client developers. If you've had to configure your home or office's wireless router, you've seen the ways in which hardware logic can be modified with relative ease.

Currently, the primary users of these solutions remain high-frequency hedge funds and proprietary trading firms. Bulge-bracket investment banks are all aware of the technology and see its merits, but some still believe there are other more cost-effective methods to reducing latency...for now.

Another less-sexy, but equally important component to a low-latency hardware solution is the method of connecting servers, whether they are traditional or content-aware. Ethernet has long been the standard to connect both user PCs and servers. However, newer standards have emerged to support the high-throughput, low-latency needs of financial services and other industries.

Infiniband is a standard that is becoming more and more prevalent in the data centers of Wall Street. It connects servers to other servers and allows them to communicate several times faster than does traditional Ethernet. To improve the performance between servers with Ethernet, additional Ethernet cards are put in each machine to essentially create more pipes between the servers. The

downside, however, is that all of these pipes require the CPU to spend more time managing the data and less time running the actual application. Infiniband takes that processing from the CPU and performs it right on the network card, allowing the application to run optimally with all of the processing power it needs.

Hardware/Software Combinations – The best of both worlds?

It is interesting and refreshing that the majority of low-latency market providers are quite willing to discuss the benefits of solutions other than their own. By no means is that because they hope the others will take a nose dive; instead, it's because many of these solutions are made even better when paired with other technologies. Clients agree: the bulge-bracket brokers see the ideal solution as a best-of-breed approach, looking to various bits of technologies to solve problems across the low-latency path.

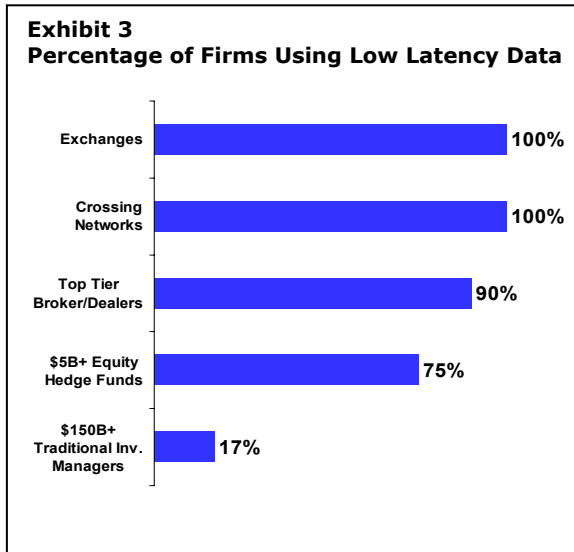
Hardware- and software-based ticker plant solutions can both utilize low-latency messaging solutions. If the information coming in is slow, then a solution that delivers that information to the desktop with low latency loses value. Conversely, it is of little value if market data arrives at the server in microseconds if the algorithm doesn't see it until hundreds of milliseconds later.

Whereas the vendors in this space continue to make friends with each other, technology groups within their potential client base often remain quite separate. Software teams often have very little interaction with the architecture teams. Further, even within the architecture team, tight integration between the networking group and the server team is not common. Since seamless integration of all components and resources remains a critical aspect of maintaining a system with the lowest latency possible, the firms that can create technology teams by using knowledge that stretches from the routing switch to the user interface will be the ones who take the lead in low latency.

On the Clock

More than 200 years ago, Benjamin Franklin said “time is money.” No wonder his face is on the hundred-dollar bill. In Ben’s day, the Pony Express provided a fast way to get your message across the country. The financial markets are still in a horse race today, but the rules have changed. As recently as the 1990s, being in the right place at the right time on the NYSE floor was all the speed you could ask for. In those days, low-latency messaging was the person with the fastest handwriting, and volumes were limited by manual processes.

Low-latency messaging is in use across a number of industries. National security, medicine and entertainment—after all, you’ve got to get last week’s “The Office” on your iPhone—to name a few. Financial services often acts as an early adopter of technology that can provide an edge over the competition. Low-latency trading is not for everyone, however. Only 17% of investment managers currently use low-latency data feeds, primarily because the complicated infrastructure required to deliver the lowest latency is too costly to implement and maintain (see Exhibit 3).



To justify the cost of a low-latency infrastructure, the strategies and volumes transacted through the system must produce consistently improving returns. If five milliseconds saved will cost \$200,000 or more a year, will the incremental improvement in market access time pay for itself? In many cases, such as a quantitative trading strategy, five milliseconds is the difference between multi-million dollar trading profits and a consistent run of losses. Five milliseconds were exciting five years ago, but now, reducing latency to less than 100 microseconds should be the goal. It is certain that as the speed bar is set higher, the cost per millisecond will also rise. Investment managers trading long-only funds are less likely to reap the rewards of such high-speed trading and thus will be far more challenged to justify this growing expense.

Although the focus of low-latency technology has been primarily on the front office, we are starting to see newer and more innovative applications in other areas. Global trading strategies require information to be available simultaneously to local trading desks all over the world. Low-latency messaging technology can be used to keep databases in sync in near real-time, as opposed to the more widely used replication technology of today. As we see this technology work its way down the trading lifecycle, we may even see it contribute to the realization of T+1 settlement for equities. But as with the

front office, the costs must be justified through returns or through cost-cutting before the investment makes sense. This will probably leave the back office in batch-process land for the foreseeable future.

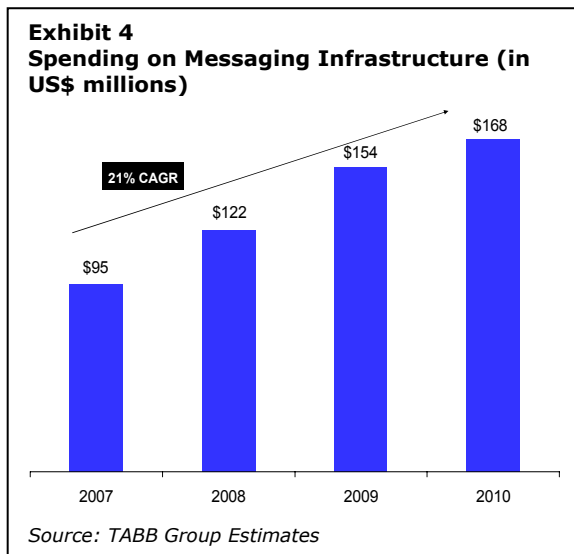
As traditional messaging frameworks change, the way they are created, managed and supported will change as well. The growing use of electronic trading tools has already caused technologists and traders to coalesce into a single entity: the electronic trader. More recently however, the need for low latency has moved the focus of these teams from trading front ends to messaging middleware. Additionally, the hardware infrastructure over which the messaging runs is often managed by a centralized, firm-wide infrastructure group. Formation of tightly integrated teams encompassing market, software and hardware specialists will enable the more efficient creation of the most optimal low-latency infrastructure.

How big is this?

There certainly doesn't seem to be a lack of budget for creating and improving low latency architectures. TABB Group estimates the market for messaging middleware to be \$95 million in 2007, growing to \$168 million in 2010 (see Exhibit 4). Many investment banks

plan to double or even triple their budgets, driven by the need to capture every possible profit opportunity in the hyper-competitive trading markets. This vote of confidence from major sell-side institutions will likely lead to smaller shops following suit in an effort to provide a significant value-added service for buy-side clients. While many innovative vendor solutions successfully shave microseconds off of message delivery, a high percentage of these resources will likely be spent on internal development projects, enhancing current applications, creating new and better algorithms and improving trading tools. Including integration and customization of both internal and external applications, TABB Group estimates the overall spend on low-latency infrastructures—including not only message buses, feed handlers, ticker plants, complex-event processors, physical transport and data storage—to be closer to \$300 million.

With only about a third of low-latency budgets currently spent on vendor products, huge opportunities exist for third-party solutions. Furthermore, faster messaging allows trading strategies that surely will utilize all available bandwidth. Therefore, as messaging accelerates to keep up with volumes,



volumes will increase as a result of faster messaging. Just as liquidity begets liquidity, speed begets speed.

And speed is not cheap. TABB Group estimates show that the cost per microsecond of improvement in today's market place is about \$250 (see Exhibit 5). That might not seem like much at first glance, but if you consider we're reducing the latency from a standard of 5-6 milliseconds to 65 microseconds or less—a nearly 6,000-microsecond drop—the costs grow quickly. The cost per microsecond is not linear either. Each second shaved off below 65 microseconds can increase cost by 5-10 times per microsecond.

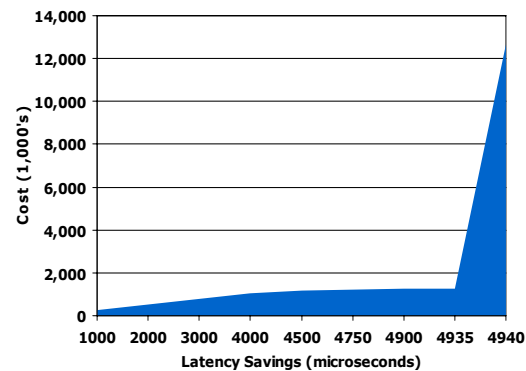
So how will these huge budgets be applied to create the needed speed? Several top-tier banks intend to buy "crazy amounts" of hardware in the coming year. That hardware, however, is mostly in the form of servers and blades, not highly specialized low-latency routers. Most believe there is still enough improvement provided by software solutions and traditional hardware, so the push for accelerated

hardware is still in its infancy. TABB Group estimates show that by 2010, 34% of servers within financial services firms will be dedicated to high-performance computing. The big question is, exactly what type of servers and how smart those servers will be?

As the overall market for messaging middleware grows, use of traditional messaging products will likely remain steady or decrease slightly with new market entrants taking big chunks of the pie. It is probable, however, that providers of traditional messaging products will soon step into the low-latency space. Buying a provider with already-established low-latency technology would provide the quickest way to gain traction in this space. Taking a "build" rather than "buy" approach, IBM has recently released its new low-latency product, dubbed WebSphere MQ Low-Latency Messaging. It's not yet determined how well the technology stacks up against the other well-established low-latency players, however, IBM's existing penetration in financial services firms based on their legacy messaging products may give them a leg up on the competition.

As low-latency software messaging continues to come into its own, low-latency messaging providers are less often competing with each other or hardware-based solutions, and more often with internally developed technologies at top-tier banks. Although there will always be buy vs. build competition, as products become more mature, banks traditionally look outside to fill those needs. We've seen this happen to a large degree in the OMS market, and we will likely

Exhibit 5
Costs for Reduced Latency



Source: TABB Group Estimates

see this in 5-7 years in the low-latency market. For unlike the OMS market, which is virtually unlimited, there is a finite ceiling for speed.

Getting There

For institutions that are heavily reliant on trading, selection of the proper solution for each component is paramount. First, the institution will select the most appropriate protocol, such as FIX or FAST. The next challenge is choosing the most effective method of getting those messages out and around, whether it is through a traditional Ethernet solution or a newer technology such as Infiniband. The combinations and choices are nearly infinite. Although feed handlers, ticker plants, and complex event processing (CEP) engines are all key to a successful low-latency infrastructure, without the proper messaging solution, the environment will be no better than running a paper ticket down Broad Street.

When it comes to low-latency messaging, one size definitely does not fit all. When buying a suit, some times you get lucky and pick one up off the rack at Saks and it just fits. More often than not, however, the legs need to be lengthened and the jacket sleeves shortened. Occasionally you just need to have that suit custom-made elsewhere, because even though a poorly fit suit is still a suit, it will never feel just right.

The new suppliers of low-latency infrastructure components have begun offering unique and incremental improvements for different segments of the value chain. When chosen in the right combination and implemented properly, the best solutions can provide the boost that traders need to stay competitive. However, the sheer number of choices and combinations and the challenges inherent in molding together their sometimes disparate underlying technologies has created more questions and headaches for today's electronic trading desks and their technologists.

Small quantitatively-driven hedge funds have needs that are much different from that of a sell-side trading desk. The strategies employed by these funds tend to span products, markets and currencies and necessitate the acquisition of data for each of these areas to accurately make trading decisions. As if maintaining consistent low latency for equity market data weren't tough enough, throwing in options pricing and FX rates raises the complexity to a new level. Although sell-side trading desks do rely heavily on low-latency data to support their algorithms, they often focus on single-stock equities, creating a more straightforward problem set.

Solutions to handle high-speed market data can vary greatly from those used to manage the transport of transactional data such as execution reports. Communicating between applications internally is a different animal than dealing with market-data providers or trading customers. Although throughput and latency figures are no doubt important when making a vendor selection, truly understanding your needs and environment is a crucial first step in making the right choice.

The Tradeoffs

In all cases some integration of the new messaging solution is to be expected. Nearly all suppliers will work closely with their clients to ensure the integration is as seamless and painless as possible. Software solutions often require changes to existing applications, as they need to use the new API to properly utilize the new high-speed messaging. In many cases, industry-standard libraries, such as JMS (Java Messaging Service), are used so that changes to the applications are minimal, as the existing library calls simply point to a new set of procedures.

Hardware solutions can often be installed with no changes to front-end applications, however, data-center changes provide their own level of complexity. Content-aware hardware must have the appropriate business logic coded and optimized to the client's infrastructure. It is also likely the new hardware will run in parallel with the old for some time until all are comfortable with its stability.

In most cases, the more standard the hardware and software in a particular organization is, the more uniform any upgrade will be. For example, if a firm has applications in C++, Java and COBOL (gasp!), integrating new API calls at this firm will be three times more difficult – separate integration for each language - than an organization standardized on Java. Our conversations with some major brokers have shown a move toward a Linux environment; however, the programming platform is still quite diverse, with C++, .Net and Java all still prevalent.

Unfortunately, there is no perfect combination of components that will suit all of Wall Street's needs. There are providers offering technologically superior products backed by the right experience and expertise, but each firm must still ensure they are not trying to put a square peg in a round hole. The major investment banks are still focusing more energy on software messaging solutions, and their choices are generally a good barometer for the direction smaller firms should and will take. Hardware and software messaging can and should coexist, however, so only by garnering a deep understanding of a firm's low latency needs can the best selections be made.

Vendor Selection

When doing any vendor selection, several qualities must always be examined. Does the product you're evaluating represent the core business of that supplier or only its secondary focus? Will the company be around in five years to provide continued support? Can it be easily integrated into your firm's architecture? Will the benefit outweigh the cost? Can we build something better internally?

When discussing low-latency solutions however, only one word seems to come out when talking with traders, technologists and suppliers: Consistency. Throughput rates mean very little if they cannot be maintained in any market environment. The more volatile the day gets, the more crucial it becomes for traders and their algorithms to have fast and accurate data. The messaging

solutions that will have the most success are the ones that can maintain their high levels of throughput on even the most high-volume days.

Most bulge-bracket brokers are using a combination of internally developed and vendor-supplied messaging technology. In some cases, the brokers began years ago with a vendor solution, but over the past several years have highly customized the product to a point where it has become their own. As these big shops will always maintain core bits of their architecture in-house, their shopping habits are similar to that of someone out on 5th Avenue looking only for a few high-quality items, likely buying each somewhere different.

Smaller and medium-sized firms, however, enjoy a bit more of a one-stop-shopping environment. One trip to a department store will get you all of the high quality pieces you need. Just like a department store, however, one-stop shopping provides you a final solution made up of products from several different suppliers. This approach has proven very successful for many vendors in the low-latency space. Joint ventures allow each vendor to remain focused on their core competency while still providing an integrated solution to the client.

Many in the industry believe that legacy messaging solutions cannot keep up with today's low-latency needs. Often, benefits that can be gained by tweaking existing software and installing additional hardware have already been recognized. Maintaining a leading-edge status in today's markets require messaging solutions not only built for today, but scalable enough to keep pace with increasing volumes and lower latency in the next five to ten years.

Ultimately, the only way to really decide what best meets your low-latency messaging needs is side-by-side testing. To see (or not see) speed increases based on a particular solution is the only way to truly understand the benefit. Using real-life data from your most hectic trading day can really show the benefits the solution may or may not provide. The proof is in the pudding.

Conclusion

Messaging middleware has allowed disparate applications to communicate efficiently for years. It brings market data to trading screens; it brings orders to exchanges; it allows the quick integration of vendor products with home-grown applications. Unfortunately, many of these solutions were built in a world where sub-second was fast. Now, sub-second is sub-par.

While technology enabled volumes to increase, it simultaneously allowed the trading desk to manage the growth in data. As flat-panel monitors gained prevalence just after the turn of the 21st century, the explosion of information was managed with six or more monitors at each trading station. In today's market, algorithms and other computer-based models have the ability to consume even more data in only fractions of a second.

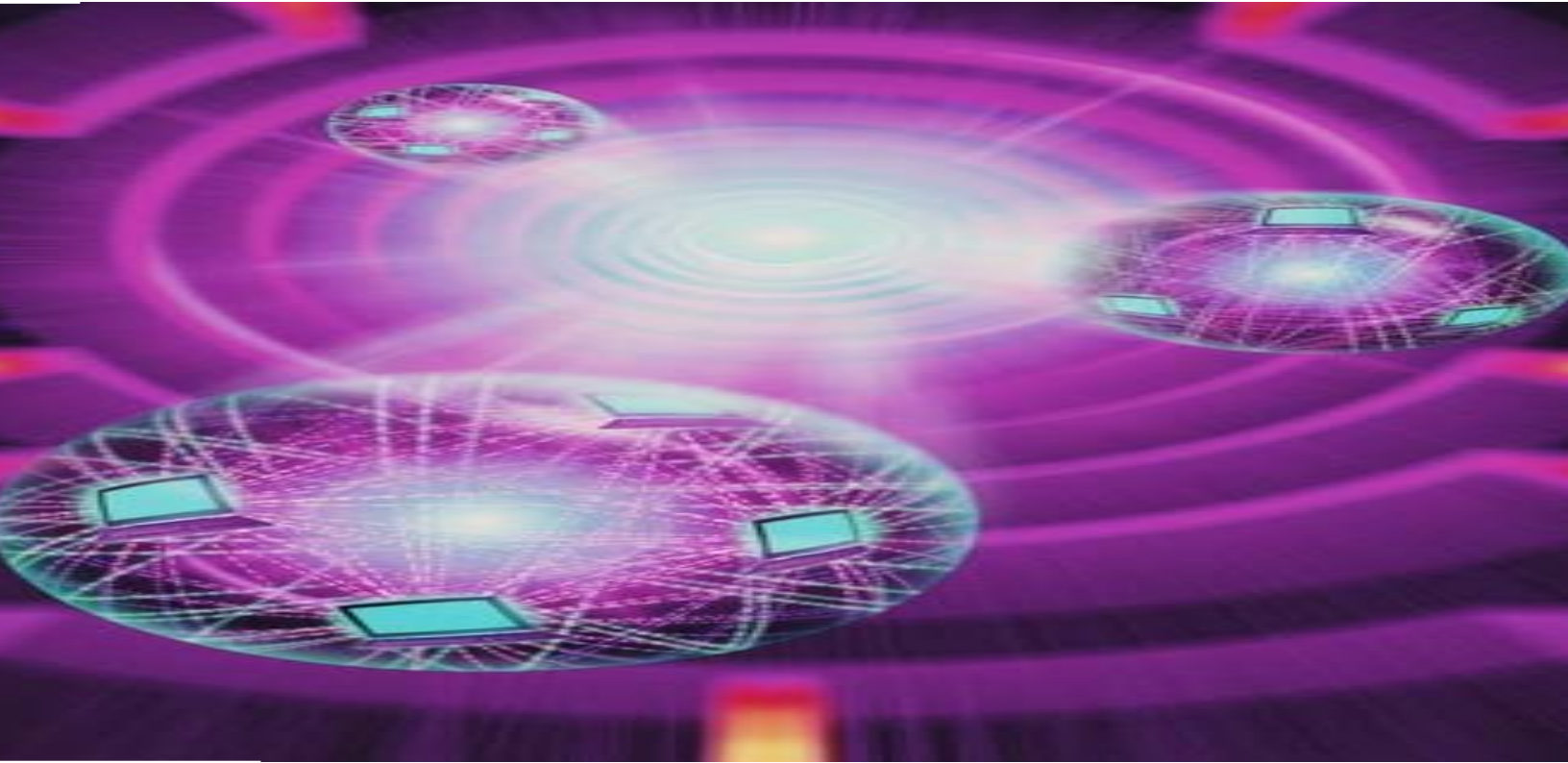
Now we have arrived in a place where low-latency messaging is required to support the messaging volumes, and the messaging volumes continue to grow in large part due to the increase in low-latency messaging. This chicken-and-egg scenario is an age-old tale, but in this case, one that must not be ignored.

Innovations in the messaging space have hit the Street full-force in the last few years. Software providers have removed "hops," allowing applications to speak directly with each other, hardware providers have built logic directly into silicon, removing software from the equation, and others have melded the two ideas in an attempt to leverage the benefits of each paradigm. Traditional daemon-based messaging still has its place, but achieving a low-latency label in today's market requires new technology built with only the latest and greatest.

The need for, and approach to low latency is different for each market participant. The benefits of sub-millisecond transactions must be well understood before seeking out the best solution. Building low-latency middleware is only viable for the biggest firms, and even then the benefits of homegrown solutions versus today's vendor-provided solutions are arguable. Finding bottlenecks in your infrastructure and running side-by-side comparisons with new solutions will provide much-needed answers for every situation.

Low-latency architectures will continue to be a key part of every sell-side trading offering, with quant funds adopting the technology and asset managers seeking access through broker-provided EMS and DMA platforms. Legacy messaging middleware can no longer keep pace, making this move to newer solutions a necessity. Gaining such low-latency messaging is no easy – or cheap – task, however. Only through truly understanding the available technology and gaining a deep knowledge of the business needs can all involved stay in the game. Paper tickets may have been easier to understand, but they are "so last-century."





www.tabbgroup.com

Westborough, MA
+1.508.836.2031
New York
+1.646.722.7800
