



Contents

Speed makes the difference between winners and losers _____	2
The growth of automated trading systems _____	2
Automation _____	2
Algorithmic trading fuels order flow _____	3
Turbo-charging the Systems _____	4
A yoctosecond? _____	4
What are firms doing about speed? _____	4
Co-location _____	5
Proximity Hosting _____	6
The Issue with Latency _____	7
Future _____	8
COLT's Proximity Services Portfolio _____	8

Speed makes the difference between winners and losers

For most of this decade, the global investment markets have been turbo-charging their systems. This need is being driven by a huge increase in automated trading, with some predicting that most trading will be electronic by the end of the decade. These automated systems are connecting brokers, dealers, traders, investors, exchanges, locally and globally. What are the drivers behind these changes and how did they come around? This report examines the history in depth, the current state and future of the world's investment markets with a clear conclusion: speed makes the difference between winners and losers.

The growth of automated trading systems

Only a decade ago, a great deal of trading was performed through open outcry, where traders would go into 'the pit' and shout over each other to try to win business. Back then, the major differentiation was having bodies on the floor who could gesticulate and shout the loudest. Today, it is very different. Today, the major differentiation is having systems connected to each other that can trade the most in the fastest time possible. What happened to create such radical change so fast?



Automation

The origins of automated trading systems date back to the first Electronic Communication Network (ECN), Instinet established in the USA in 1968, and stock market, NASDAQ created in 1971. These changes to trading, alongside the introduction of Request for Quotes (RFQ), SWIFT standards and other trade processing services, started the automation of the investment markets.

In Europe however, not a great deal changed until the 1980's, when the 'Big Bang' occurred in London. By the 1980's, the City of London and the London Stock Exchange (LSE) had become tired and uncompetitive. It was based upon high commission cartels, long lunches, brokers and jobbers. The cosy City Club was ready to be broken, as it was losing ground to the American style of trading, particularly the sophisticated networking the USA markets were using known as Electronic Communication Networks (ECN), and so new regulations were introduced. Out went the long lunches and in came the American banks and electronic trading.

This revolution in trading has changed the face of investment banking, with Direct Market Access (DMA) and program trading displacing the old style pit trading. This point cannot be seen more clearly than with the closure of the LSE's 1970's office, a 26-floor tower of trading at Threadneedle Street. This tower is now offices and a shopping centre, as the LSE moved to a much smaller building in 2004. This move is just one example of the impact of electronic trading.

With the rise of more sophisticated networking, also came the rise of crossing networks. Crossing network services, known as Alternative Trading Systems (ATS) in the USA, meant that traders could place an electronic offer to buy or sell on the network. When another order file crosses the offer and matches, the deal was filled. ECN and ATS services became critical to creating higher returns for investment firms as it meant that no human was involved in the process. Orders could be traded just by leaving them sitting on the network. For this reason, order flows have exploded in recent times, with the volume of orders nearly doubling each year whilst the value of these orders is halving.

This is all due to automation.

This explosion is further enhanced by the development of other services, particularly algorithmic trading.



Algorithmic trading fuels order flow

Algorithmic trading has really come into its own in the 21st century, thanks to developments in technology. All of these technological evolutions are fuelling the ability to create very complex trading strategies.

By way of example, the original form of program trading, which appeared in the mid-1990's, operated with rules that were sophisticated for the time. The rule might be, "if Microsoft shares increase in value by more than 0.5% in the day, and IBM shares reduce by more than 0.5% at the same time, then buy IBM (IBM) and sell Microsoft (MSFT)".

This program trading required complex systems for the time too, as Straight Through Processing (STP) systems were appearing for the first time. These STP systems focused upon bringing together various pieces of the trade lifecycle from Order Management System (OMS), Direct Market Access (DMA), fast pre-trade messaging using open standards such as the FIX Protocol, and efficient post-trade clearing and settlement.

Today, through algorithmic systems, a broker-dealer can enter a trading rule such as:

- > if Microsoft shares move by more than 0.5% based upon their Volume-Weighted Average Price (VWAP, which is the ratio of the value traded to total volume of shares traded) in a fifteen-minute period; whilst
- > IBM VWAP falls by more than 0.5% in the same fifteen-minute period; and
- > the S&P500 and FTSE250 are both rising; then
- > buy IBM (IBM) and sell Microsoft (MSFT).

In other words, the complexity of trading instructions is almost unlimited by time, volume, value, exchange and indices. In fact, using the latest algorithmics, trading systems can be fine-tuned to change strategies based upon news alerts from Reuters and Bloomberg. You can even use market data to simulate trading strategies. You could take the example above and try that strategy through the course of yesterday's or last week's data to see what would have happened. If the results are good, then why not run this in real-time.



STP appears to be more like Slow Through Processing rather than Straight Through today. Today, technologies based upon massively parallel processing, grid computing, data centre and server farms, virtualisation, cloud computing and more have changed the landscape of trading and allowed the original form of programming a trade to become much more complex.

And this is the most critical point: it is all in real-time.

This is the reason why speed is critical.

Latency and Proximity Services



Exceed Together



Turbo-charging the Systems

Speed is critical because there has been a major increase in order volume over the past decade, due to these automated systems algorithmically trading. This can be illustrated through a review of the volume of trading through the New York Stock Exchange (NYSE). In June 2006, the NYSE averaged a daily volume of 3.841 billion shares. Two years later, in June 2008, that figure had increased to 4.973 billion shares per day, a third more.

With now 5 billion shares traded per day on the world's largest exchange you can see why firms need to be conscious of speed as any delays make the difference between winning and losing, especially as all trading is in real-time.

The earlier example of IBM stocks to buy and Microsoft stocks to sell, all based upon a 15-minute VWAP, means that if the trader misses the 1 yoctosecond opportunity where the window was there to complete that trade then the whole trading strategy is worthless.

A yoctosecond?

A yoctosecond is a trillion trillionths of a second. The fact that we talk about milliseconds, a thousandth of a second, and microseconds, a millionth of a second, these days is due to the focus in this area. However, sub-yoctoseconds are the target for latency, the delay in communications between systems. In between, you have nanoseconds, picoseconds, femtoseconds, attoseconds and more:

Fractions of a second	Metric name
0,000 000 000 000 000 000 000 001	yoctosecond (ys)
0,000 000 000 000 000 000 001	zeptosecond (zs)
0,000 000 000 000 000 001	attosecond (as)
0,000 000 000 000 001	femtosecond (fs)
0,000 000 000 001 (a trillionth)	picosecond (ps)
0,000 000 001 (a billionth)	nanosecond (ns)
0,000 001 (a millionth)	microsecond (µs)
0,001 (a thousandth)	millisecond (ms)
0.01 (a hundredth)	centisecond (cs)

The fact is that each layer of technology in the system adds a yoctosecond or worse, seconds, into the network and every yoctosecond counts when you have traders and market makers sending around 10,000 order/answer pairs a second. This is why firms are trying to

minimise speed between their systems, to ensure that orders are filled at the right price and at the right time, as prices change continuously. The more yoctoseconds or zeptoseconds you can take out of the network, the better, as this makes you more competitive.

What are firms doing about speed?

The fact that orders are being filled and completed in real-time, continuously and non-stop during the trading day is the reason why order flows have become so important. Within the order flow however are a variety of hand-offs and connections between the fund managers, the traders, the broker-dealers, the banks, the exchanges and, of course, the post-trade clearing and settlement operators. But it is the pre-trade to trade execution order flow that is critical for speed.

This is why, during the late 2000's, many broker-dealers acquired providers of Execution Management Systems (EMS) in order to link these systems into their client's OMS to ensure faster dealing with smart order routing, where the systems manage orders in real-time to deliver best results. Examples include Citigroup's Lava Trading, Goldman Sachs' REDIPlus, Instinet's Newport, ITG's Radical and Triton, Morgan Stanley's Passport and more. All of these systems aim to minimise the hand-off between providers and systems.

The result is that the time it takes for a client to send an order through the broker to the exchange and receive an acknowledgement back, a round trip, is generally less than half

a second. For example, Merrill Lynch is one of the highest volume traders in the world, illustrated by the fact that in July 2008 their smart order router executed over 100 million shares on Chi-X in a single day. According to Merrill Lynch, the benchmark round trip time from order entry to receipt of an exchange's acknowledgement has come down from 850 milliseconds to less than 60 milliseconds over

the last three years. Within this, the broker's contribution has come down from over 500 milliseconds to under ten, and the fastest exchanges response times have come down from 450 milliseconds to 30. This is despite the fact that order volumes have massively increased, and demonstrates significant improvements in both performance and capacity for all firms.

How this is being achieved is through various developments in technology, particularly through two services:

- > **co-location**, where brokers place their algorithmic trading servers inside the stock exchange next to their systems, and
- > **proximity hosting**, where brokers and their clients place their systems and servers together in the same physical space.

Co-location

Co-location physically places the broker's hardware inside an exchange's data centre, in order to ensure that the physical distance for trading is zero. The exchanges have been facilitating such capabilities by purpose building co-location centres with Deutsche Boerse the first European exchange to offer such services on Xetra and Eurex from the middle of 2006. Deutsche Boerse state that for every 100 kilometres of distance, an extra millisecond is added to execution times. As a result, round-trip order cycle times to Eurex averaged less than 10 milliseconds, compared to 29 milliseconds for the fastest connection from London and 128 milliseconds from Chicago.

ICE Futures also offered co-location services around the same time, with the intention of processing 99.9% of order messages in less than 50 milliseconds.

For these reasons, there has been a rush towards providing co-location services, with most European exchanges offering such facilities today, including the LSE, NYSE Euronext, NASDAQ, the Australian Stock Exchange (ASE) and more.

In Europe, this rush was hastened by the newly competitive environment that the MiFID, the Markets in Financial Instruments Directive, introduced. As a result of MiFID, NASDAQ OMX, Chi-X, BATS, Turquoise, Equiduct and more launched trade execution services with low latency. These new services are competing head-to-head with the traditional exchanges, with most offering co-location services¹. Using such services, BATS claim that they can bring down order processing times to around 400 microseconds, whilst NASDAQ OMX Europe claim even faster cycles of 250 microseconds.

There is a downside to co-location though, in that it adds an overhead for those brokers who need to trade on multiple order books, as they will need to co-locate with multiple exchanges. This adds significant complexity and cost. For example, by placing servers in the LSE, the dealer cannot gain seek order placement at the same speeds with NASDAQ OMX, BATS, Chi-X or other markets without co-locating servers at every exchange's office. Even if the dealer does this, it adds incredible complexity when you may have multiple clients placing orders through smart order routing across multiple order management systems, and the broker's system has to stream these to every co-location system in tandem. That gets difficult.

Therefore, co-location is mainly popular with traders who are happy to deal with a single connection to a single exchange in a low latency environment. For these reasons, it is less popular than proximity hosting.



¹ See Colt Telecom's report: "The State of Europe's Equities Markets One Year after MiFID", September 2008

Latency and Proximity Services



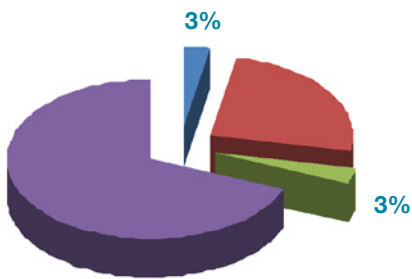
Exceed Together

Proximity Hosting

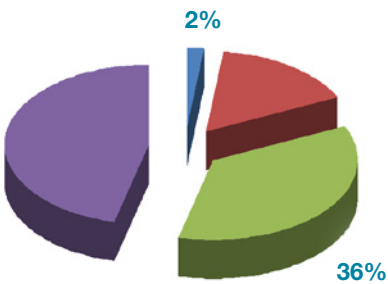
Proximity Hosting differs to co-location as it draws together broker and client systems in near range of the key exchanges. The aim is to ensure that the movement of instructions for client indication of interest and request for quote to acknowledgement is communicated as speedily as possible from the client's OMS through the broker's EMS to the exchange.

Merrill Lynch shows the impact this can have on order cycle as follows:

Round trip without proximity hosting averages 54 milliseconds



Round trip with proximity hosting averages 36 milliseconds²



- Client Network Connection
- Broker Contribution
- Exchange Network Connection
- Exchange Contribution

Proximity hosting is often discussed in the same breath as co-location, even though the two areas differ. For example, some brokers and fund managers may place business in a data centre near St Paul's for proximity to the LSE. Alternatively, others might co-locate servers inside the firewall of the LSE in Paternoster Square to lose an extra microsecond.

In some instances, may provide a combination of the two, with a proximity hub for clients to co-locate with their servers whilst, feeding from the hub, the broker deploys a range of exchange-based co-located servers to provide the lowest latency possible.

Proximity Hosting is offered by a range of providers with COLT Telecom partnering with the Deutsche Boerse, BT Radianz working with NYSE Euronext and SAVVIS the LSE.



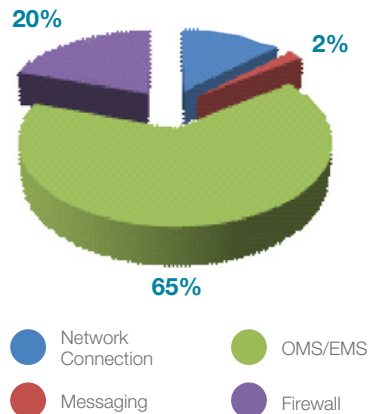
² Source: Merrill Lynch article "The Need for Speed", published in Automated Trader <http://www.automatedtrader.net/automated-trader-sponsored-articles-970.xhtm>

The Issue with Latency

Whether dealing with low latency feeds, co-location, proximity or any other aspect of market cycle times, most of the delay today is between applications rather than servers. Of course, it is true that distance makes a difference, with an extra millisecond added for every 100 kilometres of distance it makes sense to consider where to locate servers, routers and smart order routing and processing. If it takes 240 milliseconds to route an order from London to Tokyo via New York versus 180 milliseconds via Moscow, then it may well make sense to relocate the server farm for Asia connectivity to Russia. However, if you save sixty milliseconds by doing so and then leave the applications with the same processing platform and coding, it may make little difference to have that extra sixty millisecond window. After all, as discussed, a millisecond is a millisecond, potentially making the difference between an order filled and an order lost.

A research study by consultancy NET2S in July 2008 found that the network added 13% to the overall latency of the trading process, and messaging only 2%. The biggest sources of latency, the study found, are banks' applications, such as order and execution management systems, which contribute 65% to overall latency, and firewalls, which account for 20%.

Additional latency added by the components of the network



Therefore, although firms do need to focus upon reduction in connectivity, it is far more important to study end-to-end connectivity, not just network connectivity, to ensure low latency trading. This will mean re-architecting more of the systems to be more tightly coupled between buy-side, sell-side and execution venues in the longer-term, as well as further improvements in interoperability through FIX connectivity standards.



Latency and Proximity Services



Exceed Together

Future

The future will take trading systems further to reshape the world of investments from an open outcry trading pit to a proximity hosted systems pit. The pit moves man to machine.



These co-located and proximity services will be come even more complex and integrated, fine-tuned to perform at peak all the time with non-stop, high availability. The new world of co-located systems in close proximity to each other will continue to push the boundaries of trading, to try to achieve that yoctosecond faster capability for arbitrage and alpha.

The future focus will be a mixture of human traders creating complex trading strategies across asset classes and geographies. These traders will have direct market access, with buy- and sell- side becoming blurred as traders use systems for directly executing trades as well as dealers doing the same. Any human overhead or involvement in the direct execution and processing of orders will be purely for the purposes of advice and research, rather than processing and execution.

As a result, the idea that all human traders and broker-dealers disappear is a fallacy. You just have a reconstructed trading cycle where the humans must be very capable of driving automated systems strategies, combined with systems engineers who can truly support the trade process.

The integrated man-machine operation will be one where real-time no longer allows latency, but will be true real-time, and the winners will be those who can most clearly leverage real-time trading strategies alpha returns.

COLT's Proximity Services Portfolio

COLT recognises that network latency is crucial for financial organisations. Our network was built to provide the resilience and high availability that financial institutions require to support these services.

COLT has 18 purpose built data centres designed to provide the optimum environment for locating IT and telecoms infrastructure. These facilities are directly connected to our network backbone so are inherently suited to latency sensitive applications such as algorithmic trading. As COLT data centres are located in the key financial city centres across Europe this enables financial organisations to host their servers in a secure environment but close to the trading infrastructure for high speed availability of data.

COLT's Proximity Services offering provides customers the highest possible quality of hosting, managed services and network access for financial transaction management. By bringing together COLT's wholly owned data centres, fibre network and operational environment, customers are able to get the fastest possible access to data feeds and European financial exchanges. COLT Proximity Services are available as a standard offering at all 18 COLT data centres in Europe.



Latency and Proximity Services



Exceed Together

COLT offers three primary models for Proximity Services:

- > Customer trading platform colocation hosting with Financial Exchange within a single Data Centre
- > Customer trading platform colocation hosting with fibre connectivity to Financial Exchanges of customer choice
- > COLT managed services delivering monitoring and management of equipment delivering customer trading platform, with either of the connectivity options above.

Customers can choose to co-locate their existing trading infrastructure in the same physical building as their chosen market, or they can choose to centrally locate their platform, utilising COLT's fibre network to give them access to multiple exchanges across Europe from one trading system.

Customers can opt to continue to manage their own systems, or COLT can take on some of this load through its managed service portfolio, offering 24x7 proactive management and monitoring.

Where customers are additionally concerned with Disaster Recovery or Business Continuity Planning, they can engage with COLT's Professional Services team to design solutions that will allow for the management of disaster situations. These can include voice services,

high availability network architectures as well as high availability hosting and managed services infrastructures.

COLT was established in the City of London, with Fidelity as the prime shareholder, principally to serve financial institutions. Today COLT's network underpins 20 Stock Exchanges enabling financial customers to access their data rapidly and securely throughout Europe. COLT host 7 Exchanges and provide connectivity to the world's largest 25 financial services companies and 67% of the top 100 European banks. COLT is also the largest accredited SWIFT network provider for 750 customers.

For more information on COLT's Proximity Services offering please contact Terry.Quigley@colt.net.

