IBM Global Business Services

Financial Services

Tackling latency – the algorithmic arms race

How investment banks are applying technology innovations in order to gain competitive advantage in the move to algorithmic trading

Introduction

In any gun fight, it's not enough just to shoot fast or to shoot straight. Survival depends on being able to do both. And a single shot isn't always enough either – you also need to be quick to load and fire again. For gunfighters in the Wild West making use of the latest innovations, such as repeating revolvers, could mean the difference between life and death, and these innovations were rapidly adopted by all combatants as each sought every possible advantage.

In a similar way traders on the world's financial markets are also embarking on a massive arms race. The only difference is that the lone gun-slinger of the open-outcry trading floors is rapidly being replaced by ultra-fast, computerised trading systems which are more akin to robots with machine guns.

In the age of algorithmic trading systems – computer software that consumes realtime market data and trades automatically according to sets of rules, or algorithms – there are three factors define your competitive advantage:

1) ***Shoot straight*** *– the ability to define a trading strategy that adapts dynamically not only to changes in the market, but also to the impact of other firms' trading strategies. In addition, the ability to align your algorithmic trading system so that it effectively executes the defined trading strategy.*

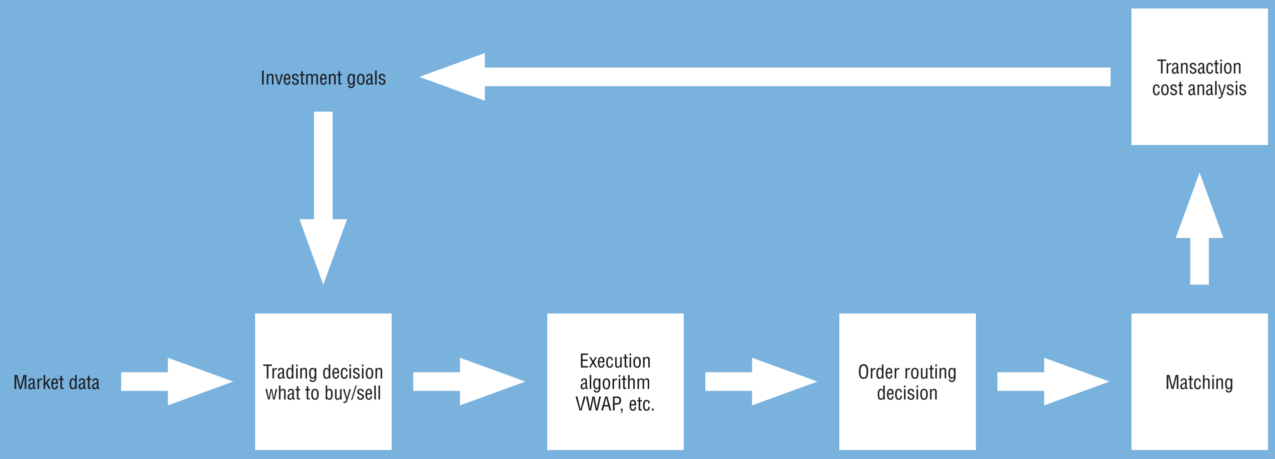
2) ***Shoot fast*** *– the ability to reduce latency (the time it takes to react to changes in the market and execute a trade) to an absolute minimum. Speed is an advantage not only because the first mover usually gets the best price in the market, but also because the competition between algorithmic systems increases the risk that late movers may not be able to complete a planned trade.*

   Indeed, with certain strategies, each trade informs the algorithm what it should do next. Consequently, the longer the delay, the greater the chance that the execution will fail (e.g. an arbitrage trade that seeks to take advantage of a short-term discrepancy between the cash market and the futures market is dependent on getting the fill done quickly enough so that the gain is preserved).

   Indeed the need for speed is now so great, that many are talking about latency arbitrage.

3) ***Shoot often*** *– the ability to process massive volumes of trades. While humans have a limited trading capacity and get tired quickly, algorithmic trading systems have a massive capacity and can operate continually without ever getting tired. Firms are capitalising on this increased capacity by breaking large block trades into a number of timed smaller trades that will have less market impact and will help disguise the trading strategy.*

Algorithmic trading


Investment goals
Transaction cost analysis
Market data
Trading decision what to buy/sell
Execution algorithm VWAP, etc.
Order routing decision
Matching

Automated models for entering orders:

- From simple 'order slicing' of large orders to complex quantitive models
- VWAP (Volume Weighted Average Price)
- TWAP (Time Weighted Average Price)
- TVOL (Target Volume Average Price)
- Arrival price
- Correlation: pairs, baskets, waves
- Sensitivity
- Arbitrage (index, stat, risk, etc.)
- Implementation shortfall
- Pegging.

Constant testing optimisation

- Pre-trade and post-trade analytics
- Achieve trading goals through transaction cost analysis as well as market data.

Market impact

- 'Fuzz factors' to disguise models from detection and reverse-engineering (e.g. buy back some product that has already been sold…)
- Stealth algorithms: ITG COBRA – places orders randomly based on historic trading trends (trades are undetectable)
- Adaptive algortithms: BoFA Ambush – different order placement strategies (aggressive, neutral, passive).

The trading environment

While the world's equity markets may rise and fall, its stock exchanges are firing on all cylinders. Recently, the London Stock Exchange (LSE) said sales for the second quarter 2006 rose by 25% year-on-year to a new record.[1] This is largely courtesy of SETS, the automated system that accounts for about two-thirds of trading in London. SETS volume rose by an incredible 69%,[2] a rate that makes the LSE's target to double activity by 2008, set as part of its takeover defence, look attainable.

The fact that London's volume growth is running at about double that of its peers is partly explained by the upgrade to the SETS technology last autumn that cut latency – or the time delay - in communicating with market users. An improvement from 30 milliseconds (ms) to 2 ms may sound like a pedantic boast, but is material for the algorithmic trading programmes that are driving SETS volumes.

It is estimated that around 40% of the trades made on the LSE now originate from algorithmic trading systems.[3] These systems thrive on instant information. Ironically, the LSE is now making more money from market data than it is from actual trading, because while algorithmic trading has been fuelling the growth in the volume of quotes, the ratio of quotes to completed trades has actually fallen dramatically – the number of OPRA quotes per transaction has increased from 300-400 in 2001 to over 3,000-3,500 today.[4]

Indeed, unless the algorithmic trading system is quick enough to complete the end-to-end process – receiving the data, to analysing it, placing an order and executing the trade – then it simply incurs the overhead of processing the information, without the benefit of profiting from the end transaction. Applying the gun-fighting analogy, this means that a great deal of ammunition is being wasted, with very few shots hitting their target.

[1] London Stock Exchange.
[2] London Stock Exchange.
[3] Sunday Telegraph – 27 August 2007.
[4] OPRA.

Algorithmic trading is also being applied far beyond just the equity exchanges, and now incorporates not only many other asset classes, but also complex trades that offset a number of different classes. In this environment where a trade may have components in several different asset classes, the failure of any one of the components of the trade directly impacts the overall trading position. This means that there is immense pressure to have systems that are fast enough to complete the end-to-end process across all components of a complex trade.

Goldman Sachs predicts that within 12 months, 60% of the deals struck on the London market will be generated from black box systems.[5]

The death of the trader has been predicted for 20 years or more. But with many trades now being executed by computers (based on information fed from other computers), it seems that some of these fears are beginning to be justified. Indeed, the exploitation of short-term arbitrage is just the sort of quick-fire, low-value trading that is ideally handled by algorithmic systems.

However, in areas like proprietary trading where taking a strategic risk position as well as a tactical trading position is essential in order to beat the market, algorithmic systems have a smaller role to play. It is therefore the low-value end of the trading spectrum where traders are under greatest threat, while the star traders at the high-value are safe for now – as long as they continue to beat the market.

This low-touch versus high-touch split is further evidenced by the recent IBM Institute for Business Value study entitled, 'The trader is dead, long live the trader'.[6] This study found that for every 40 traders that are active today for a given product group, there will be only four left standing by 2015 as more of the high-touch turns to low-touch due to electronification of markets. The four traders will be the stars that assume risk, achieve true client insight and, of course, consistently beat the market.

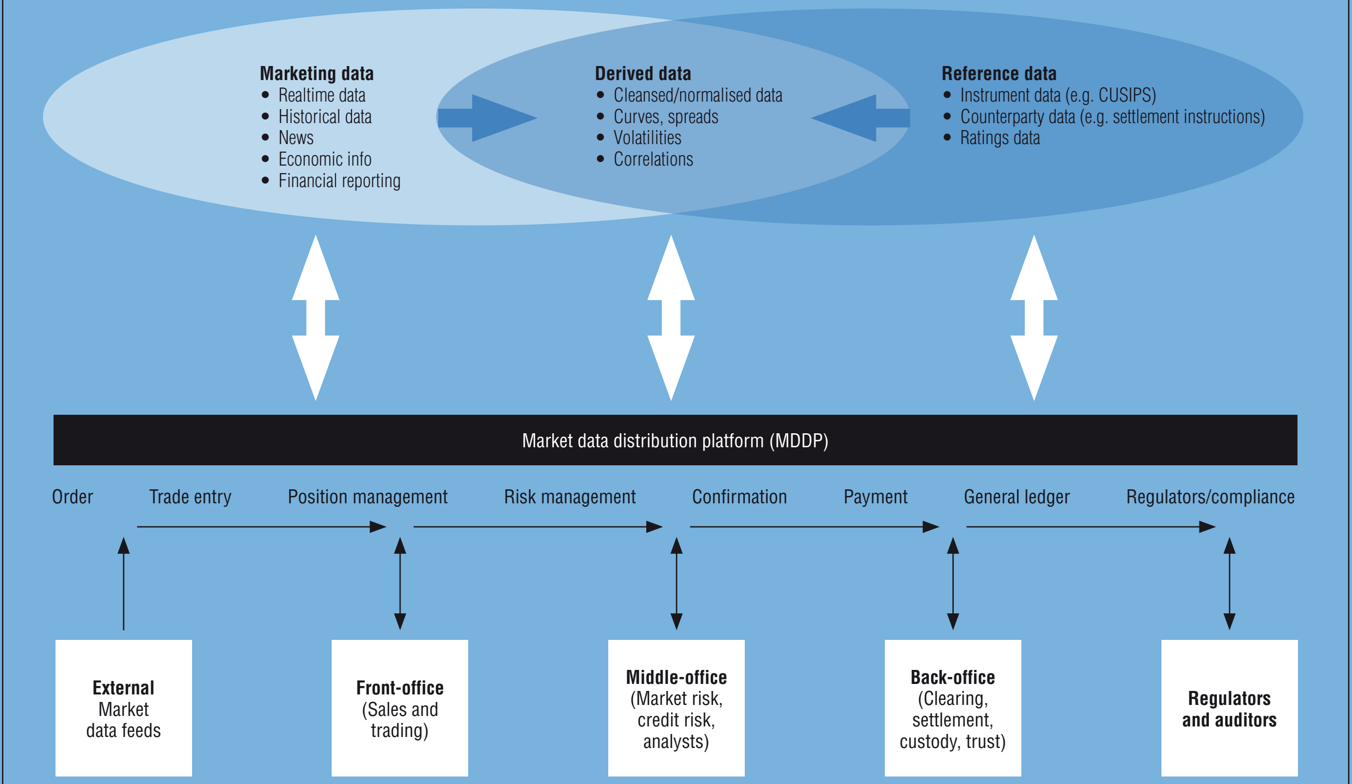
According to research by TABB Group, only 31% of institutional US equity order flow is currently communicated via phone, while 69% is communicated electronically.[7]

These numbers are also rapidly changing. Last year buy-side firms only communicated 52% of their order flow electronically. However, by 2007 firms project their electronically routed orders will increase to an incredible 80% of total order flow. This 54% increase in electronic orders over a three-year period will have a tremendous impact on firms' infrastructures, as it will cause the number of electronically traded shares to triple.

TABB Group estimates that electronically routed buy-side orders will increase from approximately 1.2 billion shares a day in 2004 to more than 3.1 billion shares per day in 2007.[8]

[5] Goldman Sachs.
[6] IBM Institute for Business Value: 'The trader is dead, long live the trader! A financial markets renaissance'.
[7] The TABB Group – October 2005.
[8] The TABB Group – October 2005.

FM data management data can be grouped into three overlapping segments…


Marketing data
• Realtime data
• Historical data
• News
• Economic info
• Financial reporting
Derived data
• Cleansed/normalised data
• Curves, spreads
• Volatilities
• Correlations
Reference data
• Instrument data (e.g. CUSIPS)
• Counterparty data (e.g. settlement instructions)
• Ratings data
Market data distribution platform (MDDP)
Order
Trade entry
Position management
Risk management
Confirmation
Payment
General ledger
Regulators/compliance
External Market data feeds
Front-office (Sales and trading)
Middle-office (Market risk, credit risk, analysts)
Back-office (Clearing, settlement, custody, trust)
Regulators and auditors

…supporting the end-to-end financial institution value chain

Making sure you shoot straight – innovations to improve trading strategies

The algorithmic trading systems incorporate pre-trade and post-trade analytics that allow them not only to respond dynamically to multiple criteria in making a trade decision (from VWAP and TVOL, to correlation and sensitivity – each across a number of asset classes), but also allow them to intelligently route the trade – even tracking both the transaction's cost, and its impact on the market and on further trading decisions.

However, shooting straight is a great deal more difficult when seeking to hit a moving target. Increasingly, firms are not only seeking to second-guess the trading strategies of their competitors, but are seeking to disguise their own. In order to prevent their competitors detecting or reverse engineering their algorithmic trading strategies, firms often buy back some product that has already been sold, or they use additional algorithms to place extra orders randomly, or base trades on historical trading trends so that trading patterns cannot be detected. Some firms are even developing adaptive algorithms that vary their placement strategies in certain areas or at certain times in order to adopt aggressive, neutral or passive positions.

We're also seeing algo-busting trades at the end of the trading day – a time when the algorithmic trading typically reaches its peak. Algo-busting trades are used by traders who believe that they've spotted a competitor's algorithmic trading pattern. They use this information to make algo-busting trades that push an algorithmic trading system to trade in a certain direction.

Making sure you shoot fast – innovations to improve trading speed

For some, latency isn't just part of the challenge – it IS the challenge. These days, the trading world measures throughput to liquidity pools in thousandths of a second. Ultimately, trading is about responding to information and transferring risk. It may sound obvious, but whoever accurately analyses and responds the fastest, and transfers risk most efficiently, has an edge that can mean significant profits. And while shaving 10 or 20 ms (i.e. one or two hundredths of a second) may not sound like much, it can be the difference between transacting or not transacting, or getting order flow from a hedge fund or seeing that flow go elsewhere.

The most obvious ways to reduce latency are to obtain direct access to market feeds (fast, low-cost access to market data), to optimise event stream processing (ESP) and then to obtain direct market access (DMA – defined here as fast, low-cost access to execution centres). Each of these has its own requirements and its own implications:

- ***Direct access to market feeds – requirements:*** *in gaining fast access to market data, firms already need to be able to streamline their support for a number of different feeds, including NASDAQ Totalview, NASDAQ UDQF, NASDAQ UTDF, NASDAQ NIDS, SIAC CTS, SIAC CQS, Archipelago ARCA, Instinet ITCH, LSE, Euronext, SWX, Eurex, BrokerTec and TradeWeb. There is the need to be able to handle various different data structures, application programming interfaces (APIs) and message sets, all of which conform to formats that aren't static but are continually evolving. Time series and XML data conforming to different ontologies and taxonomies from different data sources needs to be managed in realtime.*

  In addition to the current lack of common standards, access to market data is being further complicated by the potential proliferation of data sources being brought about by regulatory changes. The EU Markets in Financial Instruments Directive (MiFID) and the North American RegNMS are going to require complete pre-trade and post-trade transparency which will lead to an explosion of data sources. Not only will this be a challenge for the analytic capacity of the trading systems, but it will also impact everything from data compatibility and integration to data storage and retrieval.

  A further overhead is the variation in data quality and the need to handle exceptions in an efficient and effective manner.

- ***Direct access to market feeds – implications:*** *flexibility will be required in order to rapidly develop the capability to build and implement additional adapters in order to incorporate additional feeds as and when they are required. And as composite feeds are implemented, there will be a need not only to ensure that they respect the permissions and commercial agreements with each source, but also to ensure that the aggregation of the feeds does not impose any additional latency. Complex algorithmic trading strategies that incorporate multiple asset classes will compound the complexity here. All this is leading some market data vendors to split their market data service to provide one data feed for electronic trading applications and separate feeds for screen-based applications that are intended for human eyeballs.*

  In addition, not only will everyone be seeking to aggregate information from the growing number of sources, but they will become market data providers themselves. This will in turn lead to changes to the market for market data with new charging models and tariff structures evolving.

  As the current limitations for speed and reliability are reached, players will start considering not only their proximity to key market data sources with the possible construction of co-location centres, but also the construction of 'military-grade' networks for assured network services.

- ***Optimise ESP – requirements:*** *essentially the main requirements here are not only to streamline both the pre-trade analytics and the process for making trading decisions, but also to increase the processing speed. Previously, databases were used to store, index and query trading data. The game has changed dramatically with the introduction of automatic trading systems, electronic trading platforms, algorithmic trading systems and direct market feeds. The focus is now on the refinement of highly efficient applications and algorithms, and the application of massive amounts of processing power. The competitive threshold has changed dramatically with new products like IBM WebSphere* DataPower appliances offering wire-speed data translation of XML and other data formats as well as built-in security – with wire-speed decryption, adding of digital signatures and encryption for example. We're also seeing the evolution of the first ultra-efficient, feed-agnostic integrated platform for the acquisition and delivery of market data (see the ultimate solution section on page 10).*

- ***Optimise ESP – implications:*** *spending on ESP is set to rise exponentially, according to a report by Tower Group. The Boston-based research house is predicting outlays on ESP third-party solutions will be US$67 million in 2006.*[9] *Tower expects that spending to explode to US$600 million in 2010. In terms of processing power, there will continue to be an ongoing arms race between competing firms, with each keen to adopt the latest innovations in order to enhance their performance. Ever-faster systems, including Infineon and Cell processors, are being implemented in ever-greater server arrays as players seek step-changes in processing power to stay ahead of competitors.*

  The new processing capacity and streamlined systems are allowing firms to exploit arbitrage between different exchanges – with the wide area network latency for a typical international carrier remaining at about 80 ms across the Atlantic, 16 ms between European locations, about 250 ms between Europe and Asia, and 205 ms between Singapore and the US.

  Additional trends are for greater off-market trading as tier-one banks exploit their coverage and their internal liquidity pool, and for ever-more sophisticated scenario generation – with players already supplementing their 'Monte Carlo' systems that rely on random number scenario generation to 'Darwinian Flows', which model evolving, adaptive and selective scenarios. Players will also use their new processing capacity to analyse petabytes of historic market data and news to simulate ever-more complex scenarios – techniques that are akin to war games or chess gambits – as they prepare a set of gambits for any possible scenario. In simulation terms, chess has always been thought to have the right combination of human flair alongside serious number-crunching to simulate real-world requirements. IBM cracked the chess problem in May 1997 when Deep Blue beat the chess champion Gary Kasparov, but the scenarios being simulated by trading systems today need to consider a complex combination of factors across multiple indexes and asset classes, requiring far great processing power.

- ***DMA – requirements:*** *DMA is used in a different context to describe two very different things. The first is direct connectivity to execution centres or exchanges, and the second is the associated disintermediation of the sell-side as the buy-side assumes ever-more control over the sales process. With regard to the first of these, proximity is again a very real issue and in order to reduce latency as much as possible, firms would want to locate their systems as close as possible to major execution centres, within the execution centres or even ideally alongside the matching engine within the exchanges themselves. In order to ensure parity, if exchanges offer such co-location to some firms then they may have to offer it to all, which in itself would provide further challenges.*

[9] Traders Magazine – 1 July 2006.

Direct market access and algorithmic trading

What is direct market access (DMA)?

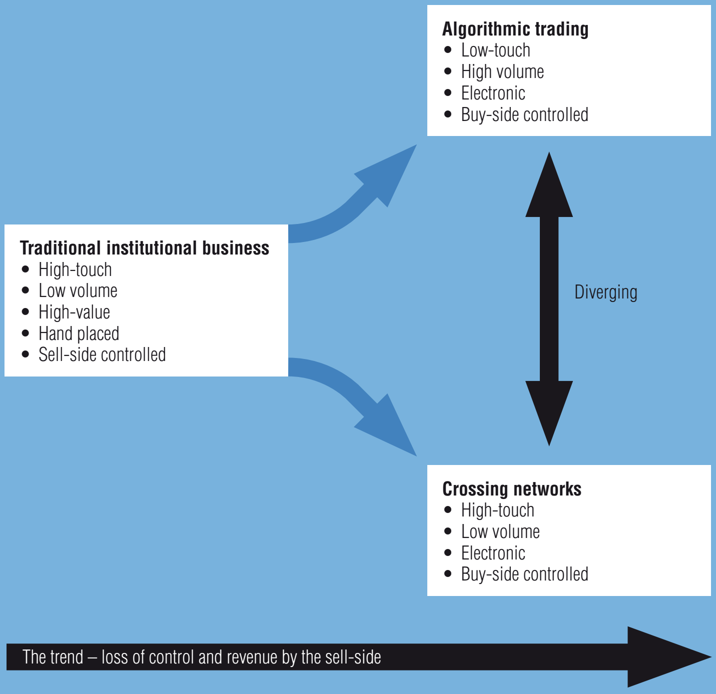
- Fast, low-cost access to execution centers for the buy-side (low-touch, no-touch)
- Self-directed trading
- Applies to equities as well as futures and options, and now FX.

Price aggregation and order routing

- Common access to multiple execution points
- Smart order routing
- For competitive markets (US equities and options markets)
- Lava (Citigroup), Sonic (BNY Brokerage), Direct Access Financial (Bank of America).

Fast pipes through sell-side to execution centre

- Sell-side offers connectivity and low-latency access to markets
- 'Pure DMA' – risk management is done asynchronously and in parallel
- Sell-side infrastructure must be optimised.



Traditional institutional business

- High-touch
- Low volume
- High-value
- Hand placed
- Sell-side controlled

Algorithmic trading

- Low-touch
- High volume
- Electronic
- Buy-side controlled

Diverging

Crossing networks

- High-touch
- Low volume
- Electronic
- Buy-side controlled

The trend – loss of control and revenue by the sell-side

Interaction between the two markets

- Algorithms are now interacting with crossing markets: Liquidnet H2O, CSFB CrossFinder, GS Sigma X.

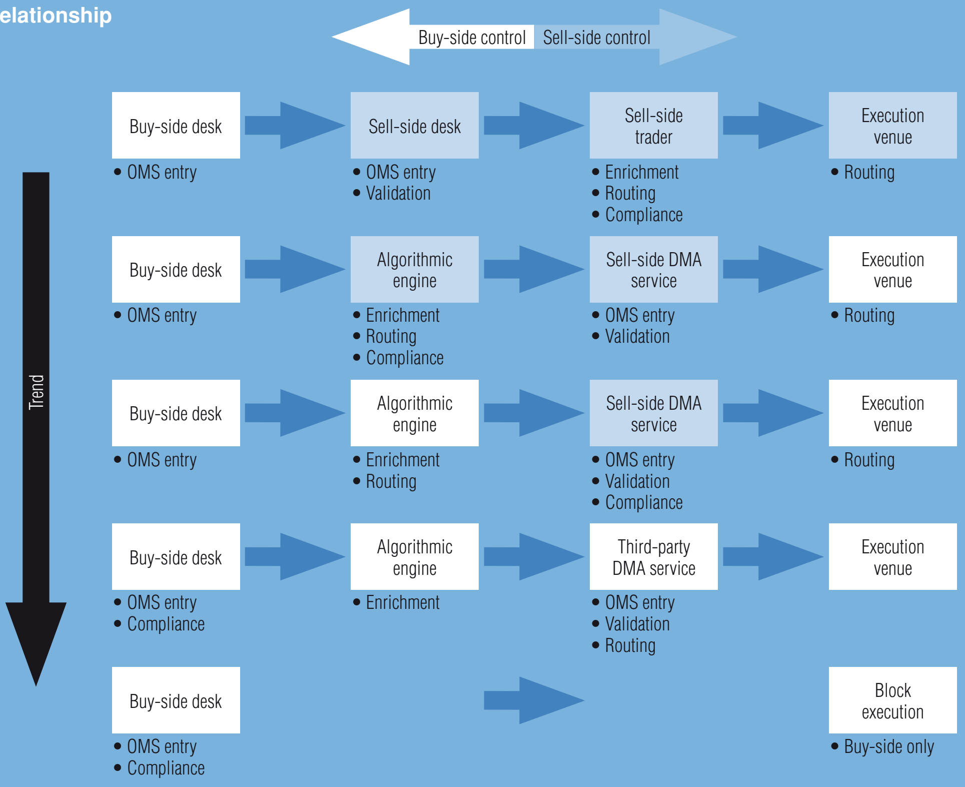
- ***DMA – implications:*** *Hitherto, buy-side firms have effectively outsourced execution risk to the sell-side firms. This is coming to an end and there is a clear trend towards buy-side control of more and more of the process. This has resulted in the sell-side firms losing the revenue and the valuable information that comes from controlling order flow. In addition, the emergence of block crossing networks that completely exclude the sell-side has meant that the sell-side firms now often compete against their buy-side counterparts. This in turn is leading to innovation on the sell-side firms as they respond by seeking to develop new sources of value generation.*

Indeed, a recent paper by IBM's Institute for Business Value[10] *predicted that today's terminology may soon start to lose relevance as firms we currently classify as 'buy-side', 'sell-side' or 'hedge funds', are in future simply classified as either 'advisers' or 'principals', or as 'risk assumers' or 'risk mitigators'.*

[10] IBM Institute for Business Value: 'The trader is dead, long live the trader! A financial markets renaissance'.

Changing buy-side/sell-side relationship

- Tradition of buy-side firms outsourcing execution risk to the sell-side is ending
- Clear trend towards buy-side control of more and more of the process
- Sell-side is losing revenue and valuable information that comes from controlling order glow
- Final stage is the movement of the block market to block crossing networks that completely exclude the sell-side
- Sell-side are now often competitors to buy-side
- But… sell-side is not giving up



Making sure you shoot often – innovations to improve trading capacity

While speed, capacity and price are all important factors, speed and price are of paramount importance on every single individual trade, whereas capacity only becomes important when volumes increase. Trades are not evenly distributed over time. Peaks occur typically at the beginning and end of any trading session or in response to news events, and it is the ability of a firm's systems to cope with the ever-increasing peak volumes that put the greatest strains on their capacity. Firms such as BNP Paribas have already concluded that it is less efficient to host such capacity themselves, opting instead for Deep Computing Capacity on Demand (DCCoD) services from trusted service providers.

Designing the ultimate machine gun-toting robot

In this algorithmic arms race, if all you had to consider was algorithmic trading, then the ultimate system would need a number of very straight-forward attributes, including accuracy, speed and capacity.

Candidate algorithmic trading requirements:

- *Co-location to both exchanges and data sources for faster communication*
- *Application-specific hardware integration for optimised wire-speed throughput*
- *In-built processes for everything from security to adaptive trading strategies*
- *Massive storage capacity for access to both current and historical market data*
- *Capability for wire-speed translations, data enrichment and exceptions management*
- *Significant inherent processing capacity, with additional capacity available on demand to handle peak volumes.*

However, such an ultimate solution would never sit in isolation and a broader consideration of the business context is required. While algorithmic trading will account for an increasingly large share of low-touch trades, there will remain a significant market for high-touch trades where traders assume risk, apply insight and seek to beat the market. The algorithmic systems will also need to be efficiently integrated with all the firm's other applications, not least of which are the firms regulatory compliance systems that authorise access, monitor trading and provide a full contextual audit trail including data such as time stamps. Ideally, while the core trading system needs to be streamlined (with speed being of paramount importance), the peripheral application interfaces need to be flexible (with standards-based service-orientated architecture (SOA) interfaces being essential to maximise adaptability). In reality, the ultimate solution needs not only to provide a streamlined platform for algorithmic trading, but also a versatile and efficient platform for all other requirements. This makes the design of the ultimate system a far more daunting challenge. The additional attributes that it would require would be:

- ***Streamlined:***
  - *An efficient, integrated platform that can optimise the acquisition, processing and delivery of market data (see previous requirements)*
- ***Adaptable:***
  - *An adaptable platform that provides very high speed transmission of market data and transaction messages to other applications and users*
  - *An open, vendor-agnostic platform, that is able to accept and distribute data from any market data vendor or alternative source*
  - *A platform that includes pre-integrated security, metering, and monitoring for both compliance and cost-effective operations*
- ***Reliable:***
  - *A platform that enables superior service levels and continuous delivery of market data*
  - *A platform that is based on robust and proven technology, and that is able to support the needs of the front-office*
- ***Open:***
  - *Open standards promote interoperability by using open published specifications for APIs, protocols, and data and file formats*
  - *Open architectures enable companies to build loosely coupled, flexible, reconfigurable solutions.*

Conclusion

The impact of electronification will be significant across all trading areas and asset classes, but will have a particularly significant impact on the low-touch environment of algorithmic trading. This will lead inevitably to an algorithmic arms race as firms compete predominantly for speed in the immediate term. But as systems across the industry improve and latency arbitrage becomes less important, firms will start to compete more on adaptability (allowing new feeds, instruments and services to be integrated quickly and efficiently) and strategic sophistication (allowing the development of ever-more sophisticated scenarios and gambits). Indeed, the algorithmic arms race may mimic the military arms race over recent decades, with the short-term focus on weaponry (as was seen in the Cold War), being replaced by a longer-term focus on combat flexibility and war-gaming strategy.

In addition, we are going to see the replacement of bespoke trading platforms with packaged trading platforms as the cost of maintaining ever-more complex applications with ever-more interfaces becomes increasingly prohibitive. Aligned to this will be a move from proprietary environments towards more open industry and technology standards. The decoupling of market data feeds from market data screens is just the first step in this direction. MiFID and RegNMS will lead to the emergence of some new frameworks and standards, simply to handle the subsequent proliferation of data sources, but further interoperability and flexibility will require the use of innovations such as SOA.

Firms need to be focused on exactly how they will compete both in the immediate and longer term as decisions they make in their overall strategy and value proposition, as well as their investment in trading systems, will impact their competitiveness for years to come.

IBM authors

Keith Bear
Partner and Solutions Executive,
Global Financial Markets
E-mail: keith_bear@uk.ibm.com

Zohar Hod
Associate Partner in FM Practice,
Lead in Front & Middle Office Trade
Process Transformation Practice
in America
E-mail: zoharhod@us.ibm.com

Phil Enness
Global Solutions Manager,
FM Data Management
E-mail: philip_enness@uk.ibm.com

Andrew Graham
Client IT Architect in Financial Markets
E-mail: Andrew_Graham@uk.ibm.com

IBM United Kingdom Limited
76 Upper Ground
South Bank
London
SE1 9PZ

The IBM home page can be found on the Internet at **ibm.com**

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. Copying or downloading the images contained in this document is expressly prohibited without the written consent of IBM. All products and/or services are subject to availability.

This publication is for general guidance only.