UCL DEPARTMENT OF
COMPUTER SCIENCE

Research Note
RN/12/04

Colocation and Latency Optimization

May 2012

Ayub Hanif

Abstract

The proliferation of electronic trading has seen a race towards faster execution. This technological arms race has yet to be comprehensively and scientifically reviewed. We perform a systematic literature review into colocation and latency minimization with respect to high-performance algorithmic trading systems. We start by defining our protocol in which we identify research questions which aim to help us summarize research in the field of colocation and latency optimization, identify potential research propositions and facilitate production of novel research. Thereafter we execute the review and undertake summary analysis. Results of the analysis are synthesized to help answer the research questions.

1 Introduction

A systematic literature review was conducted to look into colocation and latency optimization with respect to rapid-order execution algorithmic trading systems in order to identify research activities in the hope of facilitating a wider discussion in the field rather than being prescriptive and looking at and advocating a single architectural style.

Advances in the trading environment both technical and regulatory have led to a number of distinct problems. Margins have been squeezed and profits eroded through a combination of decimalization and standardization coupled with disparate and fragmented liquidity. Such changes have led to greater and greater automation within the trading microstructure however there is a distinct lack of comprehensive and scientific review around latency optimization of algorithmic trading trading systems. Proximate and colocated solutions have long been purported though a rigorous and holistic inspection has yet to be conducted. We look into this field whilst keeping in mind the key requirements of speed, invariant execution and superior modeling, and the key challenges of market impact minimization, best execution requirements and platform economic viability.

2 Protocol

Our goal was decomposed into a number of research questions:

Why is latency an issue in algorithmic trading and what is its significance?

- What is it and why is it important?
- What is end-to-end, exchange and member latency?
- What are common metrics for these?

What are the current trends within rapid order execution?

- What are the current execution methods?
- How are they favored?

What are location based execution principles?

- What is colocated trading?
- What is proximity based trading?

What motivates a quantitative trading desk to use collocated execution strategies?

- Rack cost, information leakage, etc.
- TCA.

What are the common architectures in algorithmic systems?

- With respect to ECNs, DMA, internalization, crossing.

These questions were arrived at through a population, intervention and outcome decomposition as per the systematic literature review protocol template. RQ1-RQ4 are framing questions to aid in the analysis and answer of RQ5. The experimental design to be followed through in the review will be based on primary and secondary studies. The study shall avoid the use of surrogate measures and studies as they have been found to be misleading and conclusions from these are not robust.

3 Analysis

We shall now proceed to provide descriptions and analyses of primary studies. These shall be drawn upon as primary findings for answering the set of research questions.

Empirical Market Microstructure - Hasbrouck (2007)

Hasbrouck (2007) details trading mechanisms available for securities trading. Most of the trading arrangements we observe are continuous securities markets. The most important feature of such a market is the electronic limit order book, however there are several paths to execute a trade for a given security. There is a desire of participants (retail and institutional) and regulators for orders to execute at the national best-bid and offer (NBBO) being captured in the moniker best-execution however through fragmentation of liquidity this is proving an elusive aspiration.

Real-time feeds allow for continuous observation of the book and to condition/calibrate strategies accordingly. Hasbrouck notes two key limitations to the data emanating from a market. Firstly, the immense volume of data emanating from liquid markets presents computational challenges and makes all-inclusive modeling impossible. The most important limitation, however, is the unit of observation being the order and thus the inability of a participants to gauge market agents' knowledge and beliefs. Though there are some quite elaborate models which allow for this type of model building, Hasbrouck argues these models are not plausible. They do however allow us to build simple models of individual trading strategies.

Dark Pools, Fragmented Markets, and the Quality of Price Discovery - Schwartz (2010)

Price discovery is the process a participant goes through to find the fair value 'par' price of an asset in the market. The requirement of, particularly institutional, investors to handle large orders requires some discretion. Schwartz (2010) discusses discretionary trading mechanisms and market quality in. There are a couple of very important advantages to dark pools. Firstly, it is better to have an order submitted to a (electronic) trading venue than to have it submitted to a portfolio manager, broker or dealer. The trading system matches off orders received without gaming the investor. Secondly, as trades executed in a dark pool are reported, submitting large orders to a dark pool contributes to price discovery.

Investors have many venues to choose from when executing orders. These choices have come about from market designers and regulators advocating transparency and competition. This has led to liquidity fragmentation in the 'market-for-markets'. Whilst consolidating order flow weakens competition in the market-for-markets, it is necessitated to achieve trading goals and minimize execution costs. Discussions on consolidation have frequently focused on the spatial dimension, though as more and more institutional flow is being worked through retail flow, temporal fragmentation has become an increased threat to market quality. The temporal dimension needs to be included in discussions on market quality and structure.

Corrective steps should be taken to restore order to the markets. There should be clear demarcation between exchange and brokerage operations. Order flow should be consolidated in period call auction trading. Internalization should be studied closely for market manipulation, market quality and market integrity. Accurate price discovery should be a guiding regulatory principle, allowing for greater market quality and integrity. Steps should be taken to allow the markets to re-consolidate.

History Doesn't Repeat Itself, It Rhymes: The Coming Revolution in European Market Structure - Schack & Gawronski (2008)

Regulatory and market structure changes have been positively correlated in the U.S. markets, which have looked towards Europe for innovative and novel trading techniques where issues are beginning to arise after the introduction of MiFID. Schack & Gawronski (2008) present the structural changes we are observing within the markets in Europe and how investors should navigate them.

During the 1990's, in the U.S. competition for order flow and thus liquidity was fierce and the ECNs

introduced pricing structures to lure order flow away from the major exchanges. The most effective of this was the market-taker fee structure which paid users for posting liquidity to the book (i.e. limit orders) and charged slightly higher fees to users who removed liquidity. As liquidity was becoming more and more scarce, brokerages created their own dark pools, primarily for internalization. Internalization is the process of netting customer flow within your organization before looking to external liquidity sources. Issues such as the lack of a central counterparty and securities depository, fragmented back-office infrastructures and nationalist tendencies of clearing houses, need to be addressed but shall not hamper changes in market structure. Increasing trade volume shall dictate which exchanges do better than others, and which venue acquires which venue.

Divergent Expectations - Davis, et al. (2008)

Homogeneous expectations is a critical assumption underpinning most capital pricing and utility valuation models we observe in the market. It states investors with the same information set form a consensus (identical) price. The rationale behind homogeneous expectations does not translate well into the real-world, with it's implausibility evident by simple observation of the two-sided nature of the markets. Davis et al. (2008) explore heterogeneous (divergent) expectations, where investors who have the same information arrive at different valuations and prices.

Regulation and market design have been relatively expectation agnostic however where expectations do matter, price and quantity discovery, they have favored homogeneous expectations through lack of inquiry into these key facets. Under divergent expectations, trades are not made simply because of the differences between the trichotomy of traders (informed, uninformed and noise), but rather because they all disagree. Along with the ability to idiosyncratically change their expectations, participants are influenced by what they observe in the market a term referred to as adaptive valuation (Paroush, et al. 2007).

Microstructure application and discussion should be based on a divergent expectation and adaptive valuation environment. These two issues address why we trade and why market structure matters. Trading is not driven simply by the trichotomy of traders, but also by participants who adaptively value securities and idiosyncratically change their expectations.

Anonymity, Frontrunning and Market Integrity - Comerton-Forde & Tang (2007)

The orderly operations of markets and protection of clients and investors is primary to financial regulation and informs market structure. Comerton-Forde & Tang (2007) look into fraudulent, misleading and manipulative practices in securities markets. Post-electronic proliferation, pre-trade anonymity has been the focus of increased attention from market participants and regulators alike. Surveys into anonymity and market quality have typically looked into information asymmetry however, note Comerton-Forde & Tang, current research does not explore the impact of anonymity on market integrity. This lack of exploration is quite surprising given the positive correlation between manipulative practices and anonymity (Garfinkel & Nimalendran 2009). The authors assert market integrity is compromised further in an anonymous environment due to the lack of transparency as clients cannot observe brokers trading in front of them.

Comerton-Forde & Tang ran some models to test for frontrunning activity within various scenarios using data provided by the Toronto Stock Exchange. They found amongst a small number of brokers evidence of systematic trading ahead of clients anomalous to an active proprietary trading desk. They noted that a client who is perceived as being informed is more likely to be front run than an uninformed or noise (momentum) trader.

Convergence of the U.S. Options and Equities Markets: Is the Party Over, or Just Getting Started? - Schack & Gawronski (2009)

Options markets in the U.S. are maturing and are converging on the cash equities market model. The most important source of this convergence is the preponderance of large institutional investors using options

actively in trading strategies. Schack & Gawronski (2009) explore this convergence and look at the future of securities markets. Some of the convergence aspects of mature markets infiltrating the options markets include: quoting spreads in smaller penny increments, algorithmic trading, market-taker fee structures, price-time priority, new exchanges, high-frequency trading and crossing networks.

The main outcome from this convergence is options trading shall become highly competitive with much lower margins. Intermediaries shall make less from trades however will recoup profits from volume business as market-taker fee structures induce high-frequency traders. Exchanges such as NYSE Euronext, Nasdaq OMX and ArcaEx which operate market-taker fee structures, price-time priority matching and advanced dealer exchanges are likeliest to best benefit from this convergence, though due to innate differences between the fundamentals of equities and options this convergence may run asymptotically for some time.

Execution Quality at the New, Fast NYSE - Abrokwah & Sofianos (2008)

As part of the electronic drive post-Reg NMS, the NYSE rolled out fast-execution features as part of its Hybrid Markets suite. Abrokwah & Sofianos (2008) assessed the execution quality of the new execution features. Analysis was carried out on published execution quality statistics.

Execution time is measured from when an order first touches the execution venues systems till execution and trading costs are measured according to the execution shortfall measure, which for buy orders is the execution price less the mid-quote at order arrival. A sharp drop is noted for the transformation in the middle of the roll-out for all bucket sizes and overall we observe a comprehensive decrease in execution times. Inspecting the execution shortfall for executing market orders less than 10,000 shares, we observe little change in execution shortfall when contrasted with the dramatic drop in execution speed. There is contention with these results being skewed towards retail flow, however with the rise in algorithmic trading and DMA institutional flow and large orders are less likely to find there way onto the NYSE and as such these results are valid. In summary, the NYSE dramatically decreased order execution times with limited changes in true trading costs.

Exchange Mergers - Oberhaus & Ezrati (2006)

With fragmentation of liquidity and multiple execution venues, primary exchanges are looking at mergers and expansion nationally and internationally. Oberhaus & Ezrati (2006) explore the reasons behind such activity and the associated risks. Economic viability of an exchange depends on volume, which in turn requires a venue to tap into the world trading volume. This objective can only be accomplished through strategic mergers and acquisitions, on the national and global stage. The need to increase volume, is coupled by the desire of investors to tap into previously low-traded asset classes such as derivatives.

U.S. exchanges are looking for overseas expansion for two primary reasons. Firstly, Sarbanes-Oxley, a costly auditing process, has discouraged listings on American exchanges. Secondly, the trade-through rule of Reg-NMS has effectively increased execution costs of the exchanges, thus driving down their profit margins. There are risks, primarily monopoly risk, associated with change though effective, consolidated, global regulation shall be able to control such risks. To conclude, the benefits of such mergers for investors vastly outweigh the monopoly concerns of regulator and politicians.

If Best Execution Is a Process, What Does That Process Look Like? - Wagner & Edwards (2007)

Best execution is an investment goal set by regulators in both the U.S. and Europe to ensure the investor is getting the best available price from the available liquidity in a market. The vagueness of process modeling will necessarily lead to conflicting models, however Wagner & Edwards (2007) have tackled this in a pragmatic, objective manner. They describe a six step process of implementing investment ideas and process:

1. *Establish goals*: trading is a key tenet of investment management, where we try to capture as much return as possible for the given resources.

2. *Define the process*: investment managers' routines do not have simple causal implications as one would find in other similarly complex physical tasks.
3. *Analyze the data*: this is a complicated but automated task. We clean the data and store it in a well-defined and suitable database management system.
4. *Identify solutions*: this is the hardest step of the process. The TCA reports outputted from the system need to be handled carefully by senior management. Proposing solutions requires some diligence and matureness to be executed effectively. Problems need to be prioritized, with problems with the easiest solutions usually the best place to start.
5. *Implement the solutions*: key step in the process. This requires organizational effort, and is solved top-down not bottom-up through rules and regulations.
6. *Review and revise*: need to repeat the process.

Best execution process modeling and TCA allow us to reduce slippage and control performance. However, there are various factors outside the control of the investment manager, though through organizational best efforts, factors within their control are controlled.

Systematic Internalizers - The New Trading Animals in Europe - Gomber & Wittner (2006)

The advent of MiFID saw the reclassification of internalization engines as trading venues in their own right. Gomber & Wittner (2006) explore the MiFID classification of internalization, present empirical results of German investment firms and discuss this addition to the European execution landscape. There are several motivational factors for internalization. The clearest of these is earning the spread. Furthermore internalizers are exempt from trading, clearing and settlement costs, these savings instead being passed on to customers. There has been debate whether internalization contributes to market quality and competitiveness or whether it detrimentally affects the price discovery process through enhanced fragmentation. MiFID addresses contentious issues such as these and focuses on market center competition and venue transparency.

Transparency regulations could deter entrants, which Gomber & Wittner believe will force some firms to try and circumvent the rules to avoid classification as a systematic internalizer (SI). Significant order flow is required to be successful at internalization and given the already deeply fragmented marketplace could see delayed entrance as well. MiFID is quite detailed and descriptive in it's treatment of SIs, aiming to reach a balance between market quality and competition. However, the onerous nature of this directive equates to substantial regulatory costs for firms. Pan-European operations are seen as prime candidates for classification and the effectiveness of these venues shall dictate the success of this legislation.

The OMS as an Algorithmic Trading Platform: Five Critical Business and Technical Considerations - Decker (2009)

With the proliferation of algorithmic trading, algorithmic trader's saw an execution management system (EMS) as essential to meet fundamental investment needs. The traditional order management system (OMS) was used primarily for trade blotting, portfolio reporting and compliance. Decker (2009) provides a thorough review of EMS and OMS technology, and sheds light on some key considerations. The consolidation and integration of the two platforms into one is a considerable challenge, however a clear opportunity exists in todays market for an integrated OMS which has the benefits of traditional OMS tasks and is coupled with algorithmic and fast-execution features of an EMS. For the purposes of algorithmic trading there are five key business and technical aspects which need careful consideration:

- *Unification and integration*: the ideal OMS should unify front, middle and back-office functions.
- *Adaptability*: the OMS should support a suite of algorithms from multiple vendors, in a consolidated platform.
- *Speed*: a consolidated OMS speeds up the investment workflow, allowing for optimal translation of investment ideas into realized trades.

- *Scale*: broker neutrality allows the firm to trade and connect with whomever they wish to do business with.
- *Architecture*: an advanced OMS has a flexible architecture, facilitating proactive strategic and tactical decision making.

The utility of a consolidated platform is firm specific and as such standalone architectures should not be dismissed as outdated. A firm should undertake self-assessment of their OMS and EMS with respect to their workflow, taking the above considerations into account of whether they should keep with the split or unified setup. EMSs definitely have a use in todays fragmented marketplace, though there is value in a unified OMS.

Toward "Best Settlement": Thinking Beyond Execution - Bird & Payne (2010)

Fragmentation of liquidity has driven regulation forcing venues, brokers and the like to provide best execution facilities to their clients. Bird & Payne (2010) feel there has been significant and deserved effort in best execution requirements, though there has not been enough progress to tackle settlement processes which are a larger bottleneck in the trading process. They propose three areas which are ready for improvement, the addressing of which shall allow for true straight-through-processing (STP) from investment idea generation through to settlement:

- *Broker notification*: OMS developers recognized broker-neutrality was a key feature required for a unified trading platform.
- *Custodial and prime-brokerage notification*: trading desks need to notify their custodians and prime-brokers of trade details, which is still often done manually.
- *Back-office processing*: external communications with counterparties, brokers, venues, etc. all require collaborated effort.

Market signs indicate awareness of the issues addressed above. Settlement quality is becoming an important feature for broker evaluation. As there is more momentum behind such initiatives, Bird & Payne feel we shall see standardization and consolidation in this space. Client satisfaction is a key factor for profitability and as such brokers, investment and portfolio managers, and traders should take a holistic view on the investment workflow and realize post-trade optimization is key to future growth.

Latency in Electronic Securities Trading: A Proposal for Systematic Measurement - Budimir & Schweickert (2009)

It is critical for execution services providers to meet the high performance requirements of algorithmic traders, though implicitly speed is crucial to all types of traders and participants. Issues surrounding this space are captured under the term latency, though this term has varying meaning for each participant. Budimir & Schweickert (2009) discuss definitions of latency and propose a standard methodology for latency measurement. They propose a comprehensive approach to analyzing latency in securities trading which will be used to propose a systematic measurement process. Their motivation lies in a pertinent observation: latency is *the* factor affecting profitability of algorithmic traders and provides them with a competitive advantage in the market.

The requirements elicited using the defined approach are: *i)* our focus on latency is from the bi-directional perspective; *ii)* sound latency measurement should reflect the order-action latency from customer initiation to the host and response back to them customer's point of initiation; and *iii)* sound latency measurement is comprehensive and takes measurement across trading days. Using these requirements they proceed to define latency as a single order level measure of delay for order-action round trip from customer access endpoint to market center host and back again.

Budimir & Schweickert's empirical analysis contained a dataset of Xetra order-actions (entry, modification, deletion), covering a 10-day (Sample 1:- Mar 12-23) and 5-day (Sample 2: Oct 29 - Nov 2) period in 2007. All order-actions which reached the central execution engine were recorded along with three observed latency measurements. Key findings from their analyses include latency in the investigated infrastructure is inelastic to increasing activity, the latency peak is attributable to release of economic data and as hypothesized, latency increases with distance.

Budimir & Schweickert have effectively proposed a new definition for latency and proposed a comprehensive latency measurement methodology. They have illustrated both through application to a complex dataset exposing the three main driver's of latency: *i)* trading activity - increasing activity disproportionately increases latency; *ii)* time of day - latency fluctuates throughout the course of the day; and *iii)* geographic distance - latency is directly correlated to physical distance between the customer and host advocating a dislocate between customer access and execution engine.

Trade Shredding: SRO-Sponsored Payment for Order Flow - Kugele & Wood (2007)

Trade shredding is the controversial practice of breaking orders into multiple trades purely to earn market data revenue. Kugele & Wood (2007) explore trade shredding and try to understand the driving market dynamics, observability and regulatory elimination of trade shredding. Payment for order flow arrangements where securities markets pay brokers to route orders to them for execution, are widespread within the market. There have been many reported cases of market abuse of this inherently distortive behavior. Market data revenue, which is received and allocated by the Consolidated Tape Association (CTA), has experienced substantial growth as electronic trading has proliferated. Revenues have nearly doubled from $223 million in 1994 to $434 million in 2004. Revenues are allocated amongst constituent self-regulatory organizations (SRO) on percentage of trades, and percentage of trades coupled with percentage of share volume. Payment for order flow, the increasing market data revenue pool and low profit margins on executions all increased competition amongst market centers for the tape revenue pool and became the driving market dynamic for trade shredding.

Empirical trade shredding isolation is a non-trivial task. Initial shredding activity shred trades into 100-share lots, however the same feat can be achieved with other bucket sizes without arousing suspicion. To compound issues further, there are valid reasons for shredding trades completely unrelated to increased tape revenue i.e. minimizing market impact of institutional orders. Kugele & Wood undertook a study into evidence of trade shredding and found as hypothesized, trade shredding activity migrating from less favorable to more incentivized assets. There have been attempts by the SEC to eliminate gaming in this area, though experts argue the problem lies in the size of the revenue pool not in it's allocation, however given the complexity of current allocation regulation it is a matter of time before a sophisticated participant exploits tape revenues again. Kugele & Wood summarize by positing the impossibility of automatic trade shredding surveillance and regulation. They propose a solution which addresses the revenues themselves, where opening up market data in a competitive market would remove distortive practices such as trade shredding.

Algorithmic Trading: A Primer - Palmer (2009)

The proliferation of automated computer-controlled trading techniques have come into question after accentuated short-term volatility affects were felt in the equity space during the recent crisis. Some buy-side traders have come to question the role and future of algorithmic trading. Palmer (2009) addresses this question whilst also, more importantly, focusing on the construction, underlying assumptions and operations of trading algorithms.

Abstract analysis of trading algorithms shows a common framework. The key to an algorithm is its performance criterion, commonly referred to as its benchmark. Algorithm selection is based on selecting one which has a benchmark which meets your trading objective. Benchmark statistics along with other inputs feed the algorithmic engine. Here inputs are processed and the course of action is decided. Key to the al-

gorithmic engine is its decision framework which in turn works with execution logic to realize the strategy. Algorithms run in data-driven mode, reacting in a rule-based manner to inputs most of which are market data feeds. Complementary to this feature, some algorithms allow manual tuning of algorithms i.e. altering parameters.

When considering use of an algorithm, traders should take the time to find out the assumptions and models in the algorithm. Whilst these may not be readily available, they have direct implications for performance and as such may not be applicable to specific trading strategies. Buy-side desks should not purely choose algorithms based on underlying assumptions and models, however these assumptions and models should be checked for applicability to the desk and it's trading objectives. Algorithms are designed to increase the productivity and performance of traders and as traders get experience using these tools, optimal algorithm selection follows.

Algorithmic Trading in Turbulent Markets - Flatley (2008)

Negative market sentiment is evident in the fluctuating prices and volumes of major companies and stock indices. Trading patterns are generally unfamiliar adding to the general complexity and pressure on the market. Flatley (2008) discusses new trading practices which need to be implemented in the face of the parallel rise of automation, uncertainty and volatility.

Turbulent markets call for 'order agility' from a trader to automate trades successfully, protecting and enhancing alpha. Agility is achieved through automatic re-calibration of orders, responding to changing market conditions during the lifetime of the order. Algorithms must include current market conditions into scheduling decisions and need to make use of dynamic market access to navigate fragmented liquidity.

Are You Playing in a Toxic Dark Pool? A Guide to Preventing Information Leakage - Mittal (2008)

Dark pools have attracted substantial order flow with the lure of liquidity and low market impact. The assumption amongst traders is that as matches in dark pools are *dark* there is no information leakage. This, however, is a widely known fallacy amongst market designers. Mittal (2008) attempts to demystify dark pools, their classification and discusses various issues around information leakage and toxicity levels. The taxonomy of dark pools contains five general categorizations: public crossing networks, internalization engines, ping destinations, exchange-based pools and consortium-based pools (observed pools may fall into multiple categories).

There are two common misconceptions amongst traders regarding dark pools. Firstly, trader's believe dark pools are *dark* and hence do not leak residual order information, either into the lit or dark markets. However, residual order size information is leaked by dark pools through gaming and information sharing. Secondly, trader's believe that as dark pools operate derivative pricing mechanisms, trading in dark pools does not impact price. In the taxonomy of dark pools, some pools interact with lit flow, with these interactions causing these pools to support prices in favor of the dark order. These information leaks have three general affects; information leakage can impact the price of your order, information leakage can lead to manipulation of prices in opposition to you and information leakage can result in adverse selection (where the price moves in the traders favor ex-post) for the trader.

Dark pools provide an excellent facility for order execution whilst reducing market impact and improving price, though trivial and uneducated usage can have dire performance repercussions. Fill rate should not be *the* evaluation metric when assessing dark pool quality, as fills can come at the expense of quality. Owing to their unique nature, traders and customers interacting and using dark pools, should understand the liquidity characteristics of the pools they use.

Articulating True Willingness to Trade: A Note on Marketplace Software Design - Mårtensson (2007)

Market centers attempt to draw liquidity and facilitate orderly price discovery, both of which are achieved through eliciting the true willingness to trade of participants (Harris 2003). Mårtensson (2007) describes how to enable articulation, rather than persuasion, of participants' true willingness to trade.

Electronic venues need to enable articulation through careful design considerations during development, which is currently through what can be achieved on a technical interface. Soft factors which are found in floor-markets cannot be communicated through a machine. In cash markets, users have views which are expressed through market valuation of securities, whilst in derivatives markets users tend to be price driven. Similarly, simple trading strategies are fairly easy to be expressed, however complex strategies display enhanced non-trivial expressiveness as complexity increases. Current approaches such as these are costly and increase operational risk, the major drawback of these approaches being their insistence on expressing willingness to trade in monetary terms. When inspecting at this level of abstraction, one will draw incorrect conclusions.

Marketplace software should allow users to express themselves in two additional dimensions. The pricing dimension allows users to express willingness in non-monetary terms. The price packaging dimension is concerned with at which actual level prices are expressed. This allows complex strategies to be expressed as a whole without pricing constituent orders. Such an approach allows a user to express true willingness to trade at an 'optimal' level of abstraction, creating a competitive advantage for a market center, thereby attracting liquidity.

Transaction Costs Analysis and Liquidity Discovery with Equity Options - Larison (2008)

There has been substantial volume growth in U.S. options, with tighter spreads and a conversion from quote-driven to order-driven markets. This harmonization is similar to changes observed in the equities space. Larison (2008) proposes a novel option execution comparison measure, thereby allowing institutional clients to evaluate performance. Institutional client concerns lie solely on the existence of the volume price point. They require liquidity to complete their trades and are not concerned with price improvements, especially when executing large trades. When executing in the lit markets, benchmarks such as VWAP are used to measure execution quality of institutional orders. There are many such execution quality measures for equities but none for options. The convexity measure of an option (gamma) can be used to construct a benchmark in a similar manner to which VWAP is applied to the equity market. The GWAP (gamma-weighted average price) measure allows the VWAP of the underlying to be calculated, which can be used to accurately estimate sensitivity between volume and the underlying.

Institutional demand for block options necessitates the development of an accurate, tradable benchmark. The proposed GWAP measure will see block trades move back onto exchanges away from more expensive OTC markets. The benchmark provides a number of advantages including tradability, measure and rate performance and an increase in depth for asset and portfolio managers. The GWAP benchmark provides a simple elegant, analytical solution to the TCA needs of institutional clients.

Applying Event Processing to Electronic Trading - DeLoach & Wootton (2009)

As the complexity of trading is increasing, complex event processing (CEP) technology is being harnessed across the trade lifecycle. There are many benefits of a CEP platform utilized in a high-performance, real-time architecture including rapid implementation and deployment, scalability and flexibility. DeLoach & Wootton (2009) provide some examples of CEP applications in electronic trading:

- Latency minimization in market data.
- Market makers need to be able to react to changing conditions using low-latency auto-quoting.
- Perform complex validation tasks and workflows in parallel.

- Implementation and deployment of flexible, real-time trading strategies.
- CEP is the base for most execution algorithms in the market.
- SOR implementation which can easily adapt to changing market conditions and regulations.
- Position, profit and loss, risk management can leverage the real-time capabilities of the event processor.

The event processor has frequently been shown to deliver low-latency results for trading desks and has now become an essential piece of algorithmic trading platforms. CEP architecture is geared towards the real-time requirements of trading, where the need to process, aggregate, analyze and react in real-time is critical to profitability and survival.

Colocation and The Art of Rapid-Execution Trading - Barr (2009)

Colocation has come about through the desire for low-latency amongst financial markets participants. Optimizing organizational infrastructure can result in efficiencies, however if a competitor is colocated, they have access to liquidity before you. Barr (2009) examines the colocation opportunity explaining why latency is an issue and proposes a strategy to tackle it. Minimizing latency is an organizational objective to be the first to market in light of changing market conditions. Colocation is a strategy whereby two parties operate from the same physical building, whilst proximity indicates nearness but not in the same building. These definitions are open to debate and are, unfortunately, frequently abused by marketing and sales departments.

Investing in colocation and similar low-latency initiatives necessitates a clear understanding and optimization of latency. The first step to monitoring latency is measurement, which can vary from dedicated hardware to software analytics. Summary analysis is either carried out in real-time or post-fact. Traditionally, such analysis was carried out by trading firms but is increasingly becoming a concern for financial service providers. This has led to inter-party latency monitoring, providing transparency in the marketplace and promoting venues. Such analysis highlights a key point; latency is not important, *relative* latency determines success. The spectrum of colocation facilities on offer varies from neutral providers on one end, with a managed service in the middle and exchange owned facilities on the other end. The trade-off is between flexibility and integration. Empirically, firms which employ low-latency, colocated strategies trade on multiple venues which crucially implies the criticality of colocation, proximity and connectivity when deciding where to locate an execution engine.

The Competitive Landscape for Global Exchanges: What Exchanges Must Do to Meet User Expectations - Robin & Green (2008)

The competitive environment for exchanges is being influenced by a number of factors including regulation, trading uncertainty and technology. To understand the market-for-markets the Cisco Internet Business Solutions Group (IBSG) surveyed senior industry executives to understand what exchanges must do to meet user expectations (Robin & Green 2008). From the survey, unsurprisingly, high-performance and low-latency are both very important exchange capabilities, however the most critical capability for an exchange is liquidity, without which it is impossible to trade. It is noted, liquidity follows from succeeding at all other capabilities. Another key point from the survey is that it is clear that there is no clear, universal definition of latency.

From an exchange perspective, the key latency metrics are price dissemination and order execution whilst from the member's perspective the opposite holds; how quick can data received from an exchange be processed and an order executed. Exchanges are meeting user expectations, however to continue this they must invest strategically in important capabilities which are controllable and linked to increased revenues. Some recommendations are reductions in tariffs, improved latency performance and increased peak transaction capabilities.

Algorithmic trading: Can you meet the need for speed? - Detica (2008)

The challenge of latency measurement and optimization is addressed by Detica (2008), forming the base for a discussion on low-latency architectures. Reaction times to changing market conditions is the key differentiator for success in automated trading. An optimized infrastructure with reliable low-latency will enable traders to benefit from market opportunities. The quantification of value-impact of latency is frequently inadequately addressed at firms owing to calculation complexities but, a targeted infrastructure which takes a holistic viewpoint on trading infrastructure is crucial to strategic and tactical investment. Latency minimization techniques are outlined including:

- Network and application optimization should be addressed, with particular focus on applications and components which influence trade paths.
- Application and infrastructure architectures should be conscious of agility and upgrades.
- Latency-sensitivity analysis should be carried out on components and split into dedicated groups.
- Peak performance is a key differentiator and consistent low-latency increases trader confidence and the firms profitability.

High-Performance Automated Trading Network Architectures - Cisco Systems (2010)

Cisco Systems (2010) discuss low-latency networks and their implementation for algorithmic trading, touching on some key factors and demystifying network platform suitability. As algorithmic trading has grown, firms have focused on end-to-end trading flow latency trying to reduce this to gain a competitive advantage. This focus is not in respect of absolute latency exclusively, but focuses on providing predictable and reliable latency in face of market conditions. Network platform latency analysis highlights five latency contribution factors, which are presented in order of increasing importance: serialization delay, propagation delay, nominal switch latency, queuing latency and retransmission delay.

The only way to improve latency is by measurement and subsequent analysis. Firms are interested in end-to-end latency metrics which may be difficult to gather as a trade flows through a number of functional modules, but can be facilitated through measurement platforms. Network traffic which causes short-lived congestion on the network is known as a microburst. Such periods are often attributable to heightened activity in the market. It is very important to benchmark microburst periods as this is when automated firms make most of their profit, though unfortunately these periods are relatively poorly understood and are not considered in platform evaluations. Low-latency infrastructure increases profitability of a firm, though when evaluating the underlying network platforms, excessive focus is placed on the nominal latency of the switch which has been shown to be overwhelmed by queuing and retransmission delays owing to microbursts. Such periods must be included in platform evaluations.

Design Best Practices for Latency Optimization - Cisco Systems (2007)

Cisco Systems (2007) discuss mitigation options for key latency factors. Comparing these factors we find the application and middleware factors contribute most to latency in trade flow, and offer the greatest potential for latency reduction. Mitigation options include:

- Networks need to be re-engineered to handle microbursts thus to avoid queuing delays.
- Propagation delay can be minimized through reducing the distance data travels.
- Processing and serialization delays can be reduced by using features which support hardware assistance.
- Latency is inversely proportional to network utilization so you must ensure you are using the smallest packets possible.
- The trading application must be scalable without adding to latency.

Latency minimization efforts should be holistic, taking the whole trade flow into account end-to-end. Latency reduction and mitigation techniques should be undertaken at each level in the design whilst ensuring organizational cohesion on agility and profitability.

4 Discussion

Why is latency an issue in algorithmic trading and what is its significance?

Implicitly speed is an issue for all market participants however it is particularly pertinent to algorithmic traders, providing a competitive advantage in navigating the fragmented marketplace. Success in an automated environment depends on your platform's reaction times to changing market conditions. Speed is thus critical to automated execution, being the key to sustainability and profitability. Hence, traders are concerned with bidirectional communication speeds, between themselves and execution venues. This concern allows one to measure *latency* as the round-trip delay for an order-action from the trader to exchange host and back (Budimir & Schweickert 2009). Minimizing latency requires organizational effort, aiming to be the first to market in light of changing market conditions, benefiting from market opportunities which are ever-increasingly short-lived and gaining and sustaining a competitive advantage (Detica 2008).

Low-latency has been highlighted as a key capability of an exchange in the survey carried out by Cisco IBSG (Robin & Green 2008). Key metrics from an exchange perspective are price dissemination and order execution latencies, whilst from a trader's perspective inbound market data, processing and execution latencies are key. Robin & Green (2008) detail three types of latency. Exchange latency is the delay incurred through the processing and traversal of an exchange's proprietary systems. Similarly, member latency is the delay incurred in the processing and traversal of the trader's systems. End-to-end latency, the critical measure Budimir & Schweickert refer to, is the delay in a round-trip of the overall investment flow, from trader to exchange and back. We can see latency incurred by a trader is part-controllable by the trader and part not, advocating the necessity of inter-party latency management. Such endeavors provide transparency in the marketplace and promote venues. Analyses of end-to-end latency coupled with inter-party latency management indicate the importance of participants' relative latency, being the key determinant of success (Barr 2009).

What are the current trends within rapid order execution?

Traditional order execution functionality lay within the order management function of the algorithmic trading platform. Increased market complexity and the proliferation of algorithmic trading has seen this function extracted as a separate real-time execution-oriented engine. EMS appeal lay traditionally with sophisticated active investors though we have seen this appeal broaden, leading to an integration back into the order management function (Decker 2009).

Propagation delays between the EMS and OMS increase slippage. Hardware advances provide an opportunity for fully-featured OMSs to be developed, maintaining the core features of order management and benefiting from fast-execution features of an EMS. OMS consolidation speeds up the investment workflow by avoiding synchronization pitfalls of the old split setup. The integration debate does not have a clear answer for a firm, requiring careful introspection to decide between a split or unified setup. Consolidated architecture utility is firm specific, necessitating careful consideration (Decker 2009).

What are location based execution principles?

Fragmented liquidity and the associated disparate execution venues requires a firm to have agile execution processes. Propagation delay overwhelms architecture placement questions and identification of an optimal location is crucial at overcoming geographical lag in the underlying network topology. Latency has empirically shown to be reduced when utilizing distance minimization techniques such as colocation (Budimir & Schweickert 2009). Budimir & Schweickert's analysis of end-to-end latency in the network platform shows a direct correlation between the physical distance between the trader and execution venue.

This advocates a dislocate of trader access and execution engine. Implementing a high-performance, low-latency platform requires geographical optimization to realize the true potential of the system, as otherwise a colocated competitor will have access to liquidity before you (Barr 2009).

Specifically, colocation is a strategy where two parties operate from the same physical building, whilst proximity indicates nearness but not in the same physical building. POPs provide proximity solutions for clients, facilitating dedicated connections between the client and the exchange matching engine. Both strategies have a plethora of offerings and are sold on a per rack, cabinet or cage basis. Inter-party latency highlights the shortcomings of benefits of colocation when compared to a well connected POP, again stressing the importance of relative latency. Physical latency minimization techniques must include considerations of colocation, proximity and connectivity (Barr 2009).

What motivates a quantitative trading desk to use colocated execution strategies?

Primarily, highlighted by Budimir & Schweickert, latency is *the* factor affecting profitability of algorithmic and high-frequency traders, providing them a competitive advantage in the marketplace. With latency being directly correlated to the physical distance between the trader and the execution venue, colocated strategies address this through physical distance minimization (Budimir & Schweickert 2009).

Minimizing information leakage and providing best execution are further drivers for location-based execution strategies i.e colocated trading and proximity trading. Information leakage leads to manipulative practices such as gaming to impact the price of your order, manipulation of prices in opposition to you and to systematic adverse selection. Trivial navigation of fragmented liquidity can lead to severe performance decreases (Mittal 2008). Execution shortfall and implementation shortfall allow for TCA of filled trades. These are directly affected by any market impact of your trade. Propagation delay leads to increased execution and implementation shortfall (Budimir & Schweickert 2009). This can be directly controlled through a colocated execution strategy; organization infrastructure optimization can provide benefits up to a limit thereafter the laws of physics dictate costs, with colocated execution strategies accessing liquidity faster (Barr 2009). 'Invariant execution' allows traders to pursue more aggressive strategies (Robin & Green 2008). Microburst performance and deterministic latency increase trader confidence and firm profitability (Cisco Systems 2010). Short-lived arbitrage opportunities can be exploited in a reliable colocated architecture, which would lead to detrimental performance otherwise (DeLoach & Wootton 2009).

What are the common architectures in algorithmic systems?

There are five major components of an algorithmic trading platform as abstracted out by Detica (2008). The front-office component handles order and execution management along with algorithmic strategies. Market access handles execution venue connectivity, implementing the FIX and venue-specific protocols. The data acquisition component handles data feeds, connectivity, quality management and tickerplant. Support functions include position, risk, compliance, reporting and settlement management. These are all tied together by a high-speed messaging backbone, which provides a consistent, reliable low-latency messaging implementation.

Latency minimization should be a holistic, end-to-end effort. This should be undertaken at each level of design and should be underpinned by organizational cohesion on agility and profitability (Cisco Systems 2007). The five main network platform latency contribution factors in order of increasing importance are serialization delay, propagation delay, nominal switch latency, queuing latency and retransmission delay (Cisco Systems 2010). Value-impact analysis has been found by Detica to frequently inadequately address these factors, with excessive focus placed on nominal switch latency which has been shown by Cisco Systems to be overwhelmed by delays attributed to microbursts. It is argued a sound infrastructure, which takes a holistic viewpoint, is key to efficient strategic and tactical investment (Detica 2008).

5 Summary

We have presented the systematic literature review process followed. Five research questions were identified looking into colocation and latency optimization with respect to rapid-order execution algorithmic trading. We understand that speed in execution, in decision making and in order-action provides a competitive advantage to algorithmic traders. Execution can be completed with the assistance of a dedicated execution engine (the EMS), or through an integrated order and execution management OMS.

Propagation delay is a key component of latency in the underling network topology to overcome which location based execution principles such as colocation and proximity based trading are engaged in. Algorithmic trading architectures are generally homogeneous, comprising of a core set of components, each with it's own latency profile. Addressing latency needs to be a holistic effort, end-to-end effort requiring organizational cohesion to ensure agility to market conditions.

References

K. Abrokwah & G. Sofianos (2008). 'Execution quality at the new, fast nyse'. *The Journal of Trading* **3**(1):29–34.

J. Barr (2009). 'Colocation and the Art of Rapid-Execution Trading'. *Infrastructure Computing for the Enterprise* .

J. Bird & C. Payne (2010). 'Toward Best Settlement: Thinking Beyond Execution'. *The Journal of Trading* **5**(1):63–65.

M. Budimir & U. Schweickert (2009). 'Latency in Electronic Securities Trading: A Proposal for Systematic Measurement'. *The Journal of Trading* **4**(3):47–55.

I. Cisco Systems (2007). 'Design Best Practices for Latency Optimization'. *Financial Services Technical Decision Maker White Paper* .

I. Cisco Systems (2010). 'High-Performance Automated Trading Network Architectures'. *White Paper* .

C. Comerton-Forde & K. Tang (2007). 'Anonymity, frontrunning and market integrity'. *The Journal of Trading* **2**(4):101–118.

P. Davis, et al. (2008). 'Divergent Expectations'. *The Journal of Trading* **3**(1):56–66.

T. Decker (2009). 'The OMS as an Algorithmic Trading Platform: Five Critical Business and Technical Considerations'. *The Journal of Trading* **4**(3):36–39.

D. DeLoach & J. Wootton (2009). 'Applying Event Processing to Electronic Trading'. *The Journal of Trading* **4**(3):56–58.

Detica (2008). 'Algorithmic trading: Can you meet the need for speed?'. *Detica in Financial Markets - White Paper* .

R. Flatley (2008). 'Algorithmic Trading in Turbulent Markets'. *The Journal of Trading* **3**(4):7–13.

J. Garfinkel & M. Nimalendran (2009). 'Market structure and trader anonymity: an analysis of insider trading'. *Journal of Financial and Quantitative Analysis* **38**(03):591–610.

P. Gomber & R. Wittner (2006). 'Systematic Internalisers-The New Trading Animals in Europe'. *The Journal of Trading* **1**(4):104–110.

L. Harris (2003). *Trading and exchanges: Market microstructure for practitioners*. Oxford University Press, USA.

J. Hasbrouck (2007). *Empirical market microstructure*. Oxford University Press.

L. Kugele & R. Wood (2007). 'Trade Shredding: SRO-Sponsored Payment for Order Flow'. *The Journal of Trading* **2**(1):9–29.

S. Larison (2008). 'Transaction Cost Analysis and Liquidity Discovery with Equity Options'. *Trading* **2008**(1):39–43.

A. Mårtensson (2007). 'Articulating True Willingness to Trade: A Note on Marketplace Software Design'. *The Journal of Trading* **2**(1):73–78.

H. Mittal (2008). 'Are You Playing in a Toxic Dark Pool? A Guide to Preventing Information Leakage'. *The Journal of Trading* **3**(3):20–33.

T. Oberhaus & M. Ezrati (2006). 'Exchange Mergers'. *The Journal of Trading* **1**(3):17–19.

M. Palmer (2009). 'Algorithmic Trading: A Primer'. *The Journal of Trading* **4**(3):30–35.

J. Paroush, et al. (2007). 'Stock Price Volatility, Price Discovery, and the Endogeneity of Fundamental Value' .

P. Robin & J. Green (2008). 'The Competitive Landscape for Global Exchanges: What Exchanges Must Do to Meet User Expectations'. *Cisco Internet Business Solutions Group - Point of View* .

J. Schack & J. Gawronski (2008). 'History Doesn't Repeat Itself, It Rhymes: The Coming Revolution in European Market Structure'. *The Journal of Trading* **3**(4):71–81.

J. Schack & J. Gawronski (2009). 'Convergence of the US Options and Equities Markets: Is the Party Over, or Just Getting Started?'. *The Journal of Trading* **4**(1):56–67.

R. Schwartz (2010). 'Dark Pools, Fragmented Markets, and the Quality of Price Discovery'. *The Journal of Trading* **5**(2):17–22.

W. Wagner & M. Edwards (2007). 'If Best Execution Is a Process, What Does That Process Look Like?'. *The Journal of Trading* **2**(3):32–36.