LMAX Exchange: **TCA white paper V1.0** - May 2017

TCA and *fair* execution.

The metrics that the FX industry must use.

An analysis and comparison of common FX execution quality metrics between 'last look' vs firm liquidity *and* its financial consequences.

speed › price › transparency

LMAX EXCHANGE

blank page

Contents

Foreword

David Mercer
CEO, LMAX Exchange

Transparency has become a buzzword adopted by many in the FX industry but it is time now for the industry to lead with transparent action rather than just paying lip service to the buzzword.

Whilst there is broad consensus about the importance of transparency in a highly complex, fragmented and mostly OTC marketplace, we are some way off agreement on the means to achieve it. Specifically, the FX market does not have commonly agreed metrics to assess the true cost of trading, liquidity quality and certainty of execution.

If we are being honest, traders are often left groping in the dark, forced to navigate the inconsistent standards, market practices and evaluation methods of different liquidity providers. Clients face either complete opacity or the confusion of conflicting messages and benchmarks.

Transparency cannot merely mean deluging the client with an unworkable volume and variety of disclosure information. It should be about clear, concise and common metrics that consistently inform customers and allow them to take back control of their trading strategy. Reaching that point will require the creation of robust, commonly agreed Transaction Cost Analysis (TCA) metrics that compare and contrast the differences in firm and last look liquidity, and are applicable to all client segments. Cohesive, complete information will forge the path to informed choice for the trader and go a long way to sweeping away the distrust built up over the past few scandal ridden years.

The upcoming Global Code of Conduct and the principles stated therein, whilst not banning 'last look', will bring a sharper focus on how it is deployed in relation to fairness, transparency and the management of the conflicts of interest that it introduces. And although Spot FX is out of scope for MIFID II, MIFID II does introduce monitoring and reporting requirements for venues and participants to provide a better view of orders, when they are received, how they are executed, with more granular time stamps - providing regulators with comparative data from across the market. To the extent this becomes market practice for financial instruments, alongside the Global Code, it helps set the standards for the way orders are processed in financial markets, including FX.

Foreword *cont'd*

The purpose of this white paper is to propose a blueprint for FX TCA metrics that can equip clients with an effective evaluation of the cost of trading and quality of execution. In particular, it examines the differences between trading on firm vs last look liquidity; it highlights the failure of existing TCA metrics to capture the nuances and value of firm liquidity; and it demonstrates how the transparency of market dynamics can be used by traders to reduce trading costs and to regain control of their execution strategies.

We believe there are five metrics that need to be evaluated to conduct effective TCA in FX:

- **Fill ratio**: proportion of orders successfully filled
- **Price variation**: symmetry, or lack thereof, in price slippage or improvement
- **Hold time**: the cost of discretionary latency
- **Bid-offer spread**: the difference between buy and sell prices
- **Market impact**: market price reaction to a given set of trades

The analysis that follows examines in detail fill ratio, price variation and hold time and gives a respectful nod to market impact. It concludes by providing a comparative TCA example between firm and last look liquidity, which highlights the often unacknowledged execution advantages offered by firm liquidity venues. We intend to publish separately deeper analysis on bid-offer spread and market impact, but ultimately these are relatively simple calculations that require only efficient timestamps and consistent, precise market data.

Our intention with this white paper is to contribute to an industry-wide debate on how to conduct TCA in a way that benefits the customer, provides a fair comparison for liquidity providers, and creates genuine transparency: one that enables choice and aids quality decision making.

Ultimately, this is about helping the industry move towards a position where traders are in control of their trading costs and can set their liquidity strategy accordingly which, in turn, should give true market makers an advantage through transparent information. While there may be disagreements within the industry over the veracity of specific trading practices, we should all be able to agree that traders need better tools to understand the true cost of trading.

Only with these can they make informed decisions about their execution and liquidity strategies; and only then can we take the step the whole industry needs towards restoring trust and confidence in our market.

David Mercer

speed › price › transparency

LMAX EXCHANGE

blank page

Executive summary

Executive summary

The objective for this paper is to propose a blueprint for TCA metrics that will accurately assess and measure execution quality on both firm and last look liquidity. Compared to equity markets, FX TCA is still in its infancy and primarily captures execution costs on last look liquidity. With the growing prevalence of firm liquidity, TCA metrics need to evolve to reflect execution quality across all available styles of liquidity.

To achieve the stated objective, the white paper sets out the analysis from the buy side perspective with three questions in mind:

- **Do the commonly used TCA metrics accurately measure execution costs on both last look and firm liquidity?**
- **What are the metrics that measure all the underlying processes of trading on firm liquidity, and thus should be used to properly assess the cost of trading in the FX marketplace?**
- **How does the total cost of execution compare between last look and firm liquidity?**

The analysis that follows examines in detail, **fill ratio**, **price variation** and **hold time**, as well as touching on market impact and the resulting pre-trade information leakage. Bid-offer spread comparisons will be explored in future publications.

Our findings demonstrate that applying the 'standard' execution quality metrics which have been developed for last look liquidity does not provide the full picture of execution costs, and more importantly, misses quantifiable positives of trading on firm liquidity.

Unlike last look liquidity, fill ratio and price variation on firm liquidity venues don't measure any business processes - they measure market volatility. Comparing fill ratios between last look and firm venues is not a useful exercise in comparing like with like, and price variation must also be included for a complete picture.

The cost of hold time, non-existent for firm liquidity, is an important hidden cost for trading with last look; our analysis estimated this cost at $25/million for a rejected order after 100ms. Finally, market impact, not always measured, is crucial for understanding execution quality across both liquidity types.

As a result, strategies that work well on last look venues may not translate to firm liquidity. A simple application of the basic metrics of spread, fill ratio, slippage and hold time would not show the value firm liquidity brings through price improvement and more consistent execution latency:

- **Swapping price improvement for fill ratio is possible with firm liquidity - placing the trader directly in control of their execution costs. If a higher fill ratio is important, that can be chosen over increased price improvement;**

- The underlying processes of the LMAX Exchange anonymous central limit order book hold challenges and opportunities for customers used to trading with last look venues. There are higher market data update rates to contend with, but faster and more consistent execution. We have shown several ways to place a value on that fast execution by calculating the cost of last look hold time, regardless of whether hold times are unilaterally applied to all orders or selectively applied and visible as extended tail latencies;

- Last look optionality is used as a defence for the LP against stale prices in the market. That discussion normally focuses on fill ratio and spread. It does not explain the observation that the market processes that naturally lead to price improvement in market orders are conspicuously absent for limit orders sent to the same venues, or the asymmetry in slippage vs improvement seen with some venues. Only firm liquidity exposes the same underlying market dynamics for market and limit orders.

Finally, the comparative TCA calculation demonstrates that due to the absence of price improvement, combined with a net cost increase associated with trading after a discretionary hold time, trading costs on last look are higher by between $2.25/million and $48.86/million.

Furthermore, the transaction cost analysis shows the benefits of firm liquidity as the transparent, cost effective choice that places the trader in complete control of their execution quality, with no pre-trade information leakage.

blank page

About this white paper

This white paper has been prepared by LMAX Exchange.

The findings documented in this report solely reflect the data analysed by LMAX Exchange. LMAX Exchange has compiled the data received and interpreted the data for presentation purposes in order to produce this white paper.

About the authors

Dr. Andrew Phillips
Chief Technology Officer

Dr. Andrew Phillips F.R.A.S. has over 20 years of experience in leading innovative technology teams in the design, build and operation of reliable web-scale and low latency architectures.

Andrew Stewart
Director of Strategy and Research

Andrew Stewart has 25 years of software development experience and is responsible for directing our applications development to better understand, measure and improve the customer trading experience.

Dr. Sam Adams
Head of Software

Dr. Sam Adams has a background in Chemical Informatics research. For the last 5 years he has focused solely on building, delivering and optimising LMAX Exchange high performance architecture.

blank page

Introduction

Trading Foreign Exchange (FX) is frequently conducted on 'last look' venues. Last look is a trading practice where the liquidity provider (LP) provides a quote rather than a firm price into the trading system or execution venue. When a request to trade against the quoted price is received the LP may hold the request for some period of time, execute the trade ('fill') at the price quoted, offer an alternative price ('requote') or decline to trade ('reject'). This quote driven behaviour is commonly argued to be necessary to protect the liquidity providers in a fragmented and unregulated market place where there is no central exchange [1].

Firm liquidity venues such as LMAX Exchange are gaining traction in the market as an alternative to quote driven venues. Firm liquidity means that they operate using an open central limit order book where orders are matched based on published rules without optionality.

While they offer broadly similar trading features, there are very different execution models underlying a last look quote driven venue and a firm liquidity venue such as LMAX Exchange. The standard metrics used to assess execution quality have been designed to measure and address the negative impacts of last look on trading. To properly apply TCA to FX, traders must understand the underlying processes of both firm and last look liquidity in order to compare accurately their respective execution quality characteristics.

The paper is divided into three sections:

I **Part I** is an analysis of a third party database of trades executed on LMAX Exchange and both Bank and Non Bank last look venues. This database has been selected to showcase certain behaviours which highlight the differences between firm and last look liquidity. We apply the standard metrics common in FX to build a simple score card for each of the three classes of venue;

II **Part II** is a more detailed investigation of the results of the standard metric analysis illustrating the additional factors that are important to consider from an execution quality point of view. We present details of how we calculate trading costs and what the differences between the two execution models mean for a trader's control over their order execution;

III **Part III** is a side by side comparison of the relative costs of trading on the last look execution model compared to the firm liquidity model for a variety of real world trading scenarios.

Our objectives are to analyse firm and last look liquidity side by side using the standard metrics, to show how we approach TCA for FX and to lay out our approach to analysing execution quality. Our goal is to drive the debate and contribute to the formation of a common baseline for FX TCA metrics. The questions we aim to answer are:

- Do the commonly used TCA metrics accurately measure execution costs on both last look and firm liquidity?
- What are the metrics that measure all the underlying processes of trading on firm liquidity, and thus should be used to properly assess the cost of trading in the FX marketplace?
- How does the total cost of execution compare between last look and firm liquidity?

blank page

Part I

Applying standard metrics to a sample data set

Part I: Applying standard metrics to a sample data set

This section takes the buy side trader's perspective, using trade logging information from LMAX Exchange customers accessing a range of competing liquidity providers via an independent Third Party Aggregator (TPA). This data set provides us with an opportunity to perform a like for like comparison of trades executed on firm liquidity at LMAX Exchange alongside trades executed on a number of leading Bank and Non Bank last look liquidity providers.

The TPA trade database contains 7,137,576 orders sent to six last look (3 Bank, 3 Non Bank) LPs and the firm liquidity reference venue (LMAX Exchange) for a total of seven LPs in this study over the period from 1st January 2016 to 31st December 2016. Further detailed information about the data set can be found in Appendix A.

Where necessary to add extra background material, verify assumptions or provide sanity checks of the results from the TPA data, we have supplemented this database with metrics from the LMAX Exchange platform measuring the trade flow in the exchanges in London, Tokyo and New York. Where LMAX Exchange data is used to supplement the results from the TPA data is clearly indicated. Throughout this paper, we treat the data as if they were our own trades in order to illustrate what we currently consider to be best practice TCA for FX.

We propose that there are five key metrics that should be used to assess execution quality when trading FX:

- **Fill ratio** measures the number of successfully filled orders as a fraction of the total number of orders placed, normally stated as a percentage. Rejected or unfilled orders represent an opportunity cost - the trader must forego the opportunity to trade or attempt to trade again at a potentially worse price - therefore higher fill ratios are desirable;
- **Price variation** measures the difference between the price the trader expected and the price at which they were filled, arising from movements in the underlying market prices and often separately referred to as '*slippage*' or '*price improvement*' for adverse and favourable outcomes respectively. Traders normally focus on low slippage, or net neutral price variation, as it has traditionally been unusual for them to receive (and therefore measure) price improvement in isolation;
- **Hold time** measures the discretionary element of execution latency, which is the time observed by the trader between placing an order and receiving notification of the fill or rejection. Higher hold times and execution latencies not only exacerbate the opportunity cost associated with rejects but also represent an opportunity cost on filled orders because, while the trader is waiting for a response, they are committed to honouring the potential trade and may be unable to execute the remainder of their strategy;
- **Bid-offer spread** measures the difference between the buy and sell prices. The larger the difference between the prices the more the market will have to move to make a particular position profitable. When the spread is zero, this is referred to as a '*choice price*'. This is the simplest metric to compare between LPs;
- **Market impact** characterises the response of the market, typically in terms of price changes, to a given set of trades. The interpretation of market impact is highly subjective; one trader's strategy may rely on minimising impact while another's may actively benefit from a pronounced and consistent post-trade reaction, but also suffer if that reaction occurs before a set of related trades is completed.

Part I: Applying standard metrics to a sample data set

In the following sections, we will use the TPA data to compare the fill ratio, price variation and hold time characteristics of firm and last look liquidity. We will leave the comparison of bid-offer spread to a future study as it requires comparison of market data which was not present in this data set. The nature of the trading conducted through the TPA also means that it does not provide a useful source for comparing market impact when using different liquidity types, however we will return to this topic in Part II.

We will analyse our TPA data against each of these metrics and at the end of each section will present a three level qualitative score card for **Firm**, **Bank last look** and **Non Bank last look** liquidity, with the best performing class of liquidity receiving a score of 3 (best) and 1 (worst). Where all classes of liquidity have provided a similar service we have scored them all at 2.

We will be looking at data from a 12 month period - January to December 2016 - as a unit. This means that the data includes scheduled market events such as US Non Farm Payrolls, Fed rate decisions and Brexit, US Election announcements, as well as unexpected and disruptive events such as the October GBP 'flash crash'. There is also a broad market consensus that execution quality can vary by time of day, trade size and pair traded, and a comparative study of execution quality under different market conditions may provide the basis of a future paper. However in order to keep the scope of information presented manageable we have opted to treat the year as a whole and we believe that even with these caveats, the average of trading over a year will provide a useful insight into the differences between last look and firm liquidity. Furthermore the ability to apply TCA to trades during both smooth and turbulent periods has real world application for traders who are active in a range of market conditions.

Although the names of the venues are obscured due to commercial concerns, we would welcome a serious and critical collaboration by any interested or independent parties, and would value access to other trade databases with varying types of flow to the highly uncorrelated flow seen in the TPA data set we use here.

Box 1

About the analysis

The data used in this paper is from an independent Third Party Aggregator (TPA), which contains 7 million orders sent to seven LPs (both firm and last look) in 2016.

Three key metrics are analysed in-depth and are used to assess execution quality: **fill ratio**, **price variation** and **hold time**. The TPA data is analysed against each of these metrics with a three level qualitative score card presented. This methodology allows us to compare the value of each liquidity class on the same set of metrics.

Part I (i): Applying standard metrics to a sample data set

(i) Fill ratio/rejects

Fill ratio is usually expressed as the percentage of orders that have been filled as a fraction of the total number of those sent to a liquidity provider. This has the virtue of being easy to compute and understand. The higher the fill ratio - the better, as a reject will normally result either in a missed opportunity to trade or in a worse price if the order is later resubmitted to the same (or different) venue.

Although fill ratios are commonly used to compare execution in a commercial setting, the method by which the reject rate is converted into a cost per trade is often not clear. One method is to use a subsequent fill for a rejected order (either from the same or a different LP) to determine the opportunity cost of a reject [2]. This also requires an accurate measure of hold time. The strategy in use may also influence whether a higher rate of fast rejects with a lower opportunity cost per event is preferable to a lower rate of rejects with a longer hold time and potentially higher opportunity cost per reject.

For a quoted price stream from a single LP, the fill ratio should just be a measure of whether the deal was done at the agreed price or not - was the order filled or rejected (or requoted). Rejects (errors aside) are due to insufficient liquidity to match the trader's order or LP optionality.

We have calculated fill ratio as the number of orders receiving a fill, divided by the total number of orders (excluding errors). This has some shortcomings in that it does not discriminate between large and small orders nor does it adequately represent partial fills. We have compared the results using a more proportionate calculation of notional value traded divided by notional value ordered, but as this does not materially affect the findings we have opted for the simpler fill and order counts as this is the method we see used by the majority of LMAX Exchange customers.

Causes of rejects

Rejects often include a reason for the reject, and during this analysis we considered using this message field to understand the reason behind a reject with the aim of detecting the exercise of optionality as opposed to other causes such as lack of liquidity.

Order type	Errors	Non error rejects	Total
Market	99	10,480	10,579
Limit	709	36,479	37,188
Previously Quoted (PQ)	52	17,311	17,363
Total	860	64,270	65,130

Table 1: Rejects classified by reason

However, in practice this is not a reliable technique. Messages are not standardised across venues and some can be quite ambiguous in their meaning, covering a wide variety of potential causes. There was only one exception to this - client error messages.

Part I (i): Applying standard metrics to a sample data set

The clearly identifiable types of error messages seen in the data included: coding or message format errors with the FIX order stream, trading being attempted outside of market hours, the trader exceeding their position limits, having insufficient funds or failing some other pre-trade risk control or exposure tests. Error messages normally are very specific as it is in the interests of all LPs to clearly indicate a problem that requires correction by the trader.

These error rejects were discarded from the fill ratio analysis as they are caused by errors over which neither the venues nor LPs have control. There is one exception - for the analysis of execution latency they do have a particular use, which we cover in the 'Hold time and execution latency' section (p. 29). Error messages in the TPA data are dominated by rejects due to credit issues.

Reject rates by venue

We will first look at reject rates for market orders only. Market orders have no price restriction and we expect a high fill ratio (theoretically 100%) from all venues once errors are excluded. Rejection reasons should relate purely to liquidity or optionality and by excluding limit orders, we can exclude cancels and rejects due to conditions on limit orders that are never met. We further exclude all fill or kill (FoK) market orders to ensure there are no rejects based on any size constraints.

Once all the variability from restrictions on matching is removed we should be left with the underlying best case fill ratios for each venue allowing a direct comparison of each venue's ability to fill trades. The only remaining causes for rejections should then be if there is zero useful liquidity on the book or if a reject happens due to last look optionality.

Venue	Filled	Non error rejects	Fill ratio
Non Bank 2	267,304	43	99.98%
LMAX Exchange	299,085	182	99.94%
Bank 1	207,157	130	99.94%
Bank 2	100,730	372	99.63%
Bank 3	173,571	3,047	98.27%
Non Bank 3	120,789	2,233	98.18%
Non Bank 1	115,823	3,648	96.95%

Table 2: Market order fill ratio by venue

The firm liquidity venue - LMAX Exchange - is near the top of the table, and there is a clear grouping with the top 3 having a better than 99.9% fill ratio and the rest clustered at 99.5% and below. Detailed investigation of the rejects for LMAX Exchange shows that they are all liquidity based rejects related to times when market conditions did not permit orderly execution. In common with some other venues LMAX Exchange includes a variety of protections against off-market trade execution arising from either errors during order submission ('fat finger') or discontinuities in the market price. Market discontinuities commonly occur immediately after market open and are analogous to the 'uncrossing period' seen on equities exchanges.

Part I (i): Applying standard metrics to a sample data set

This grouping implies a difference in execution model between the top three and the bottom four LPs. The presence of both Bank and Non Bank LPs in the bottom four (assuming the bottom two are not chronically short of liquidity) suggests that rejects due to last look optionality are common to both types of LP.

The next comparison is to look at orders with price restrictions. Some venues implement these as limit orders with a threshold price and some as the 'previously quoted' or PQ order type, where a specific price is referenced via a quote id. These price constrained orders are more representative of the majority of institutional trading.

Venue	Order type	Filled	Non error rejects	Fill ratio
Bank 2	PQ	621,250	1,222	99.80%
Bank 1	PQ	1,111,524	3,221	99.71%
Non Bank 3	Limit	768,467	4,190	99.46%
Non Bank 2	PQ	1,431,232	12,868	99.11%
Bank 3	Limit	964,857	14,391	98.53%
Non Bank 1	Limit	613,020	15,828	97.48%
LMAX Exchange	Limit	23,841	2,070	92.01%

Table 3: Limit order fill ratio by venue

This result shows a completely different picture. Looking at the firm liquidity LP, the LMAX Exchange fill ratio is now the lowest whereas for market orders it was near the top of the table. Another surprise is that for some of the bottom four in the market order fill ratio rankings the fill ratio for limits or PQ orders are higher than they are for market orders.

Detailed investigation of the LMAX Exchange rejects using FIX logs and internal tooling shows that 7% of the 2,070 rejects recorded for this set of trades were caused by market conditions that did not allow for orderly execution of risk for general trading (as described above), with the remaining 93% being order cancels. Of the cancels, 1.4% were due to insufficient quantity being available at the price point requested combined with a FoK strategy, and the remaining 98.6% were a limit price miss - i.e. the market had moved away from the limit price specified. **In other words, almost all the LMAX Exchange 'rejects' were driven by pricing behaviour on the venue**.

Part I (i): Applying standard metrics to a sample data set

As the number of limit order trades from this source is relatively small, we examined the limit order fill ratio for both the TPA and a variety of similar LMAX Exchange customers to see if this was an outlier.

Customer	Orders	Fill ratio
A	75,950	85.86%
B	136,872	91.24%
C	76,924	92.55%
TPA	24,910	92.58%
D	70,918	93.99%
E	4,378	95.34%

Table 4: Limit order fill ratio by client

The limit order fill ratio is consistent between the TPA data set and the LMAX Exchange internal view of the fill ratio, as expected. Furthermore the limit fill ratio for the TPA lies within a range of fill ratios for other similar customers and is not an outlier. It is notable that other customers using exactly the same order types (and to the best of our knowledge a comparable trading strategy) are able to achieve a higher fill ratio using the same market data.

Part I (i): Applying standard metrics to a sample data set

(i) Section summary: fill ratio/rejects

As expected, the order count fill ratio tells us that fill ratios on LMAX Exchange and 2 other LPs are close to 100% for market orders - barring those times when market discontinuities made it unsafe to trade. For price constrained order types - limit and PQ in this data set - the fill ratio for LMAX Exchange is much lower than the last look liquidity providers, with almost all the limit order 'rejects' being cancels due to a missed limit price. This implies that there is something specific about LMAX Exchange liquidity which makes it relatively hard to trade successfully with a high fill ratio using immediate execution limit orders. *(We will return to this in Part II).*

Metrics scorecard

- **Market order fill ratio.** LMAX Exchange is in the top three, but each of the Bank and Non Bank venues have members in the top, medium and low thirds of the table giving both of them medium scores.
- **Limit/PQ order fill ratio.** This is influenced by the two different types of order, with PQ having higher fill rates. Due to this, the Banks earn top marks, Non Banks come second, and LMAX Exchange trails in last.

Metric	Bank 'last look'	Non Bank 'last look'	LMAX Exchange
Market order fill ratio	2	2	3
Limit order fill ratio	3	2	1

Table 5: Fill ratio 'scorecard' points (higher is better)

Box 2
Analysis of fill ratios

The fill ratios on firm liquidity differ significantly for market and limit orders:

- Market order fill ratios are close to 100% on firm liquidity, as expected;
- Limit order fill ratios are much lower, with almost all the limit order 'rejects' being cancels due to a missed limit price.

The lower fill ratios for limit orders on LMAX Exchange imply that there is something specific about this type of liquidity, the trading strategy, or both (see Part II) which makes it relatively hard to achieve a high fill ratio.

Part I (ii): Applying standard metrics to a sample data set

(ii) Price variation - slippage and price improvement

Price variation is a trader's view of the difference between a desired or expected price and the actual execution price achieved by an order. While attention is often focused on slippage (i.e. execution at a worse than expected price) when using market orders we should expect to experience both slippage and improvement. Traders using price constrained orders (limit or PQ) may have been conditioned to expect neither; limit orders cannot slip and many traders do not even consider measuring price improvement.

Measurement of slippage or improvement requires information which may only be available in the trader's own logs. We cannot rely on orders to carry the price which prompted the decision to trade – market orders do not carry a price at all and the price on a limit order is not necessarily the same value as the decision price – making this metric potentially both opaque and highly subjective. However, the order placement behaviour of the TPA is far more predictable, allowing us to measure the impact of price variation consistently and objectively across LPs.

When the TPA receives a customer order, it waits until the market data it receives from the LPs indicates that the order can be filled, meeting all price or size criteria specified. Once suitable market conditions are identified the TPA selects one or more LPs, captures the current best price on the relevant side of the market and sends some or all of the order to the selected LPs for execution as a 'leg'. We have calculated slippage or price improvement per leg by looking at the difference between the logged market price at the time the decision to trade was made and the actual fill price received. This approach removes much of the individual variation from the data, treating the TPA as a single customer trading with each of the LPs and requesting the current price available for immediate execution.

We have excluded numbers from infrequently traded currency pairs (any instrument with less than 100,000 trades over the 12 month period of the data set). The remaining sample set consists of trades in EURUSD, GBPUSD, USDJPY, AUDUSD, GBPJPY, USDCAD, EURJPY, EURGBP, NZDUSD, USDCHF, EURCHF, EURAUD, AUDJPY and AUDCAD, which together represent 91% of all successful trades.

We have reported slippage and improvement using the FX conventions of 'pips', i.e. the 4th decimal place of the price other than for currency pairs priced in JPY where the 2nd decimal place is used. This introduces some comparability issues across currency pairs and over time, for example 1 pip GBPUSD is a smaller proportional slippage than 1 pip AUDUSD, and a 1st January GBPUSD pip is a smaller proportional slippage than a 1st November GBPUSD pip due to the depreciation of GBP over the year. However as all pip values fall within a range close to 0.01% of traded price (between 0.006% and 0.016% at the extremes) we have erred on the side of using familiar units over something abstract but more mathematically accurate such as basis points.

Part I (ii): Applying standard metrics to a sample data set

Market orders

Table 6 shows the proportion of market orders receiving fills where prices showed slippage, were as expected or showed improvement.

Venue	Slippage	As expected	Improvement	Ratio of slippage to improvement
Bank 3	0.00%	100.00%	0.00%	0
Non Bank 2	1.14%	98.18%	0.68%	1.68
Non Bank 1	19.40%	70.04%	10.56%	1.84
LMAX Exchange	4.36%	93.54%	2.10%	2.08
Non Bank 3	0.64%	99.15%	0.21%	3.05
Bank 2	3.47%	95.65%	0.88%	3.94
Bank 1	7.16%	92.15%	0.69%	10.38

Table 6: TPA market order price variation statistics

Chart 1 shows the percentage of orders that experienced slippage or improvement at 0.1 pip intervals. Negative numbers indicate slippage (a worse price than expected) while positive numbers indicate price improvement (a better price than expected).

In addition to the skew and shape of the distribution, it is important to note the scale is limited to +/- 5 pips for illustrative purposes. In many cases the maximum improvement observed is less than 5 pips (denoted by the green marker) whereas the maximum slippage observed is in most cases more than 5 pips away from the zero point (indicated by the red marker). In the TPA data, only LMAX Exchange exceeds 5 pips of price improvement.

The price variation of market orders falls into two distinct categories. There are those venues which show both slippage and improvement at an approximately 2:1 ratio and those for which the slippage is dominant with little or no price improvement.

As LMAX Exchange operates a firm central limit order book offering best execution in price-time priority, we might expect a more neutral result. The skew towards slippage suggests that behaviour in this data set is linked to market direction, demonstrating a propensity towards buying in a rising market and selling in a falling market. This leads to a natural bias towards slippage and away from improvement. If we take LMAX Exchange behaviour as an approximation of the pure market, then this ratio becomes an interesting metric for market order price variation. This allows us to distinguish between those venues which are passing the underlying market price behaviour straight through to the customer against those which show a higher bias towards slippage.

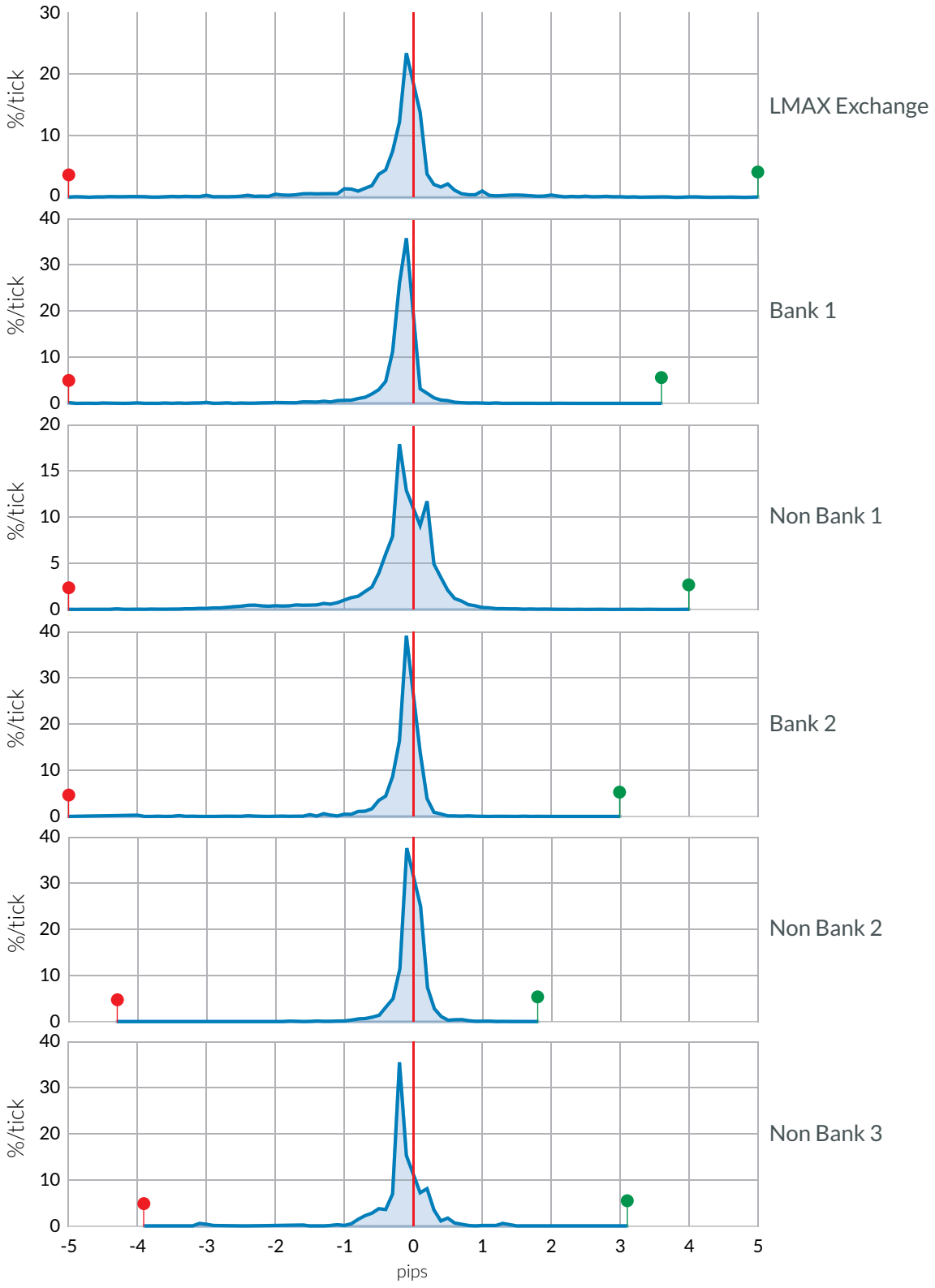

Chart 1: Market order slippage by venue

Limit/PQ orders

The situation for order types with price constraints is more interesting. These order types prohibit slippage, and the TPA sets its limit price to the same value it uses as a reference level to calculate slippage or improvement for market orders, so naively we might expect that the price variation for such orders would have a similar incidence and distribution to the price improvement side of the market order charts shown above.

With the exception of LMAX Exchange, this is not the case. Table 7 shows the proportion of limit or PQ orders receiving price improvement by venue, alongside the market order price improvement from the same venue for comparison.

Venue	Order type	Improvement	Market order improvement
LMAX Exchange	Limit	6.358%	2.10%
Bank 1	PQ	0.001%	0.69%
Non Bank 2	PQ	0.000%	0.68%
Non Bank 3	Limit	0.000%	0.21%
Non Bank 1	Limit	0.000%	10.56%
Bank 2	PQ	0.000%	0.88%
Bank 3	Limit	0.000%	0.00%

Table 7: TPA limit/PQ order price improvement statistics

Only LMAX Exchange exhibits a significant level of price improvement for limit orders. Improvement is either negligible or entirely absent for limit orders executed on all other venues.

As a further illustration of the mechanism driving limit order price improvement, chart 2 shows the distribution of the level of improvement received by both limit and market orders on LMAX Exchange, showing the percentage of orders that received improvement at 0.1 pip intervals.

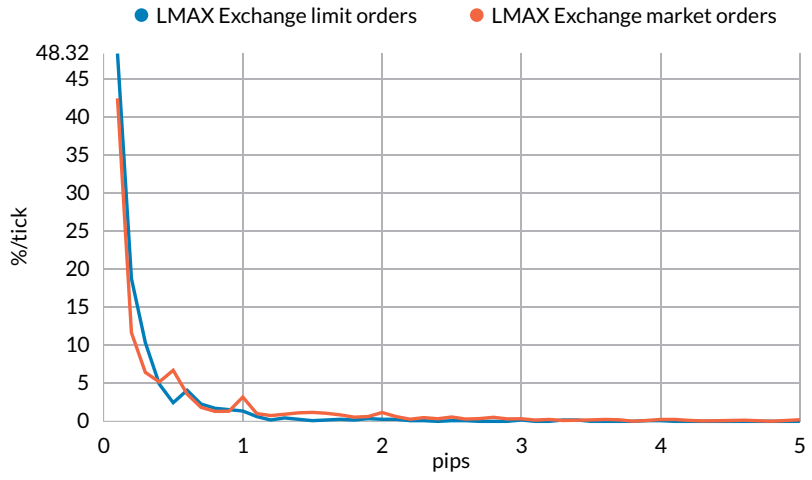

Chart 2: Price improvement for LMAX Exchange market and limit orders

Part I (ii): Applying standard metrics to a sample data set

The consistent distribution of price improvement observed for both order types is a key characteristic of firm liquidity. Limit prices only constrain the worst execution price for an order. When better prices are available, limit and market orders behave identically.

In contrast, the very different price improvement behaviour observed for market and limit orders on last look liquidity demonstrates a fundamentally different approach to filling limit orders in which LPs exercise their option to fill almost every order at its limit price, even though the evidence of fills on market orders indicates that a better price should be available for some proportion of the time.

For full disclosure, it is worth noting that liquidity providers other than LMAX Exchange offer price improvement. Unfortunately the TPA did not route a sufficient number of orders with any such provider for us to be able to make statistically valid comparisons. As noted above, we would welcome collaboration under NDA with traders with trade databases that include LMAX Exchange, ECNs, Bank and Non Bank venues to further our understanding of execution quality. Returning to the data at hand, the obvious question to ask is where has the price improvement on limit orders gone for all other venues?

(ii) Section summary: limit/PQ orders

The only venue that offers the same improvement on both limit orders and market orders is LMAX Exchange. This arises from the use of firm execution against a central limit order book, and is driven by exactly the same market behaviour that gives rise to both slippage and price improvement in market orders. None of the other venues provide significant limit order price improvement, even though a subset of them clearly expose similar underlying price volatility on market orders.

Metrics scorecard

- **Market order slippage rate.** Looking purely at slippage percentages, LMAX Exchange is in the middle of the table while each of the Bank and Non Bank venues have members in the top, medium and low thirds of the table. As a result we will award all the benefit of the doubt with a medium place.
- **Limit price improvement.** LMAX Exchange is the only venue to offer significant price improvement in this data set.

Metric	Bank 'last look'	Non Bank 'last look'	LMAX Exchange
Market order fill ratio	2	2	2
Limit order fill ratio	1	1	3

Table 8: Price variation score card points (higher is better)

Box 3

Price variation analysis

Only firm liquidity venues offer consistent price improvement on both market and limit orders:

- The analysis of market orders across the LPs shows that price variation can be either symmetrical (both price slippage or price improvement are passed to the customer without restriction) or asymmetrical (where the price improvement passed to the customer is limited but price slippage isn't);
- Only LMAX Exchange demonstrates symmetrical price variation on both market and limit orders.

The observed price improvement behaviour on last look liquidity demonstrates a fundamentally different approach to filling limit orders in which LPs exercise their option to fill almost every order at its limit price, even though the evidence of fills on market orders indicates that a better price should be available for some proportion of the time.

The obvious question that comes out from this analysis is 'where has the price improvement on limit orders gone for all other venues'?

(iii) Hold time and execution latency

Execution latency is the time taken between an order being transmitted from the trader's system and the receipt of a response. Hold time is the commonly used name for discretionary latency where the execution of an inbound order from a trader is deliberately delayed pending a decision to fill or reject by the liquidity provider's systems. This period of time is also referred to as the last look window.

Hold time/discretionary latency is just one component of execution latency, so we must first look at other causes of latency before we can assign hold times to each venue in order to compare this aspect of the execution quality of last look and firm liquidity.

We will divide execution latency into the following components:

- **Systematic.** The time required to complete the necessary operations to execute the trade, including network round trip time, transit through any pre-trade risk control system, matching engine cycle time and any other systematic delay applied across all customers of the LP;
- **Tail.** Each cause of systematic latency will also have a characteristic jitter with causes at network, operating system or application level. In addition, platform capacity constraints ranging from microbursts to sustained higher traffic rates during market announcements can lead to queueing and congestion giving a familiar long tail latency distribution;
- **Discretionary.** Any time added where the order is held prior to executing a trade. LPs may apply or vary hold time based on their assessment of a customer's market impact, the current market conditions or their own appetite to trade in a given direction.

Each of these components is subject to variation over time. Systematic latencies may be affected by hardware or software upgrades which may change the LP's latency profile. Tail latencies may likewise be affected by capacity upgrades or constraints. Lastly hold time may be adjusted by LPs in response to a change in market conditions, strategy, policy or simply based on developing insight into a customer's trading behaviour.

While we are primarily concerned with discretionary latency in the direct comparison of firm and last look liquidity, information regarding the non-discretionary causes of latency is also valuable in its own right, as this can be used to make order routing decisions as well as for TCA purposes. For example, if the latency of a particular LP degrades badly during busy times, this information may be used to augment best price or volume criteria in selecting an execution venue.

Part I (iii): Applying standard metrics to a sample data set

Chart 3 shows the execution times for rejects and fills for a particularly interesting last look LP in the TPA data set providing a clear example of each of these different types of latency. The execution time is recorded to the nearest millisecond and the frequency of occurrence is shown on a logarithmic scale. The chart spans the whole year of 2016.

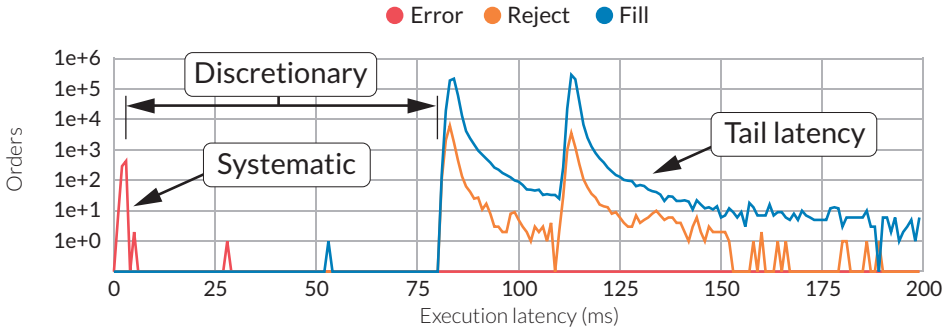

Chart 3: Detailed execution times for Bank 3

An analysis of this kind would normally use supplemental information gathered by the trading infrastructure to determine some parts of the systematic latency. For example the base network latency can be estimated by using session level FIX messages – heartbeats or test requests – which are typically processed at the edge of the LP's trading platform. Unfortunately that level of data was not available to us in the TPA data set, and we were then forced to determine the systematic latency from the execution time profiles available. Fortunately there are some markers in the data that can help us.

For the LP in chart 3, there is an interesting pattern in that fills and non-error rejects indicate that the minimum response time is around 81-82ms. However, when we looked at rejects due to errors – as defined earlier – a response time of 2-3ms is evident. 99.7% of these errors were caused by a reject at the pre-trade risk control level, rather than a programming or FIX session level error. This is then an error from within the platform – not an immediate reject at the edge. [i]

With the moderate assumption that the next logical step within the platform would be matching the order against available liquidity, we can then assign a systematic latency of at least 2-3ms. The discretionary latency or hold time would then be 80ms for this LP. It is unlikely that an order would transit the network and pre-trade risk control systems within 3ms and then take a further 80ms to be placed unless there was a hold time in play.

The execution latencies for all of 2016 for each LP in our set are shown in chart 4 (p. 32), which plots the millisecond latencies for fills, errors and rejects against the number of orders experiencing that level of latency. There are several features which stand out and bear further investigation:

- Histograms for the same class of event (e.g. fills) which display multiple peaks in the latency histogram;
- LPs where the peaks for fills, errors and rejects occur at different modal latencies;
- Long tails to execution latency distributions.

[i] ***Warning:*** *using a combination of network pings, FIX session level messages and deliberately generating different error conditions as a basic probe of the systematic latencies within an LP's platform, is a practice we do not advocate or condone and may be in breach of your terms of connection or other agreements with the LP.*

Part I (iii): Applying standard metrics to a sample data set

Our first task is to investigate each of the features above so that we can determine a characteristic systematic latency and hold time for each LP. We will investigate the first 200 ms of latency in detail. In some cases the latency distributions extend beyond this, however, latencies much beyond 200ms are usually a very small proportion of trades and our goal here is to derive the base characteristics of hold time for each LP.

Defining the systematic latency as being the mode of the first peak in the execution time histogram (whether from fills, rejects or errors) and the hold times as being the difference between the systematic latency and the mode of the second peak, we can produce the following table of systematic latencies and hold times. Rejects and fills are examined separately as their latency histograms may differ as in the example above.

Venue	Systematic (ms)	Fill hold time (ms)	Reject hold time (ms)
LMAX Exchange	1	0	0
Bank 1	4	5	1
Non Bank 1	1	90	90
Bank 2	1	9	5
Bank 3	4	80	79
Non Bank 2	1	0	0
Non Bank 3	1	0	0

Table 9: First glance modal hold times by LP

A quick comparison of table 9, which attributes very similar latency profiles to Non Bank 2, Non Bank 3 and LMAX Exchange, and chart 4 (p. 32), which shows a very different visual signature for each, indicates that our initial scorecard is not telling the whole story.

Box 4

Execution latency

In a direct comparison of execution latencies on firm and last look liquidity, the primary concern is discretionary latency or hold time.

A more detailed investigation of execution latency can reveal systemic latency and the tail of the latency distribution, providing further insight into what is really driving an LP's latency characteristics.

LPs may apply or vary hold time based on their assessment of a customer's market impact, the current market conditions or their own appetite to trade in a given direction.

Part I (iii): Applying standard metrics to a sample data set

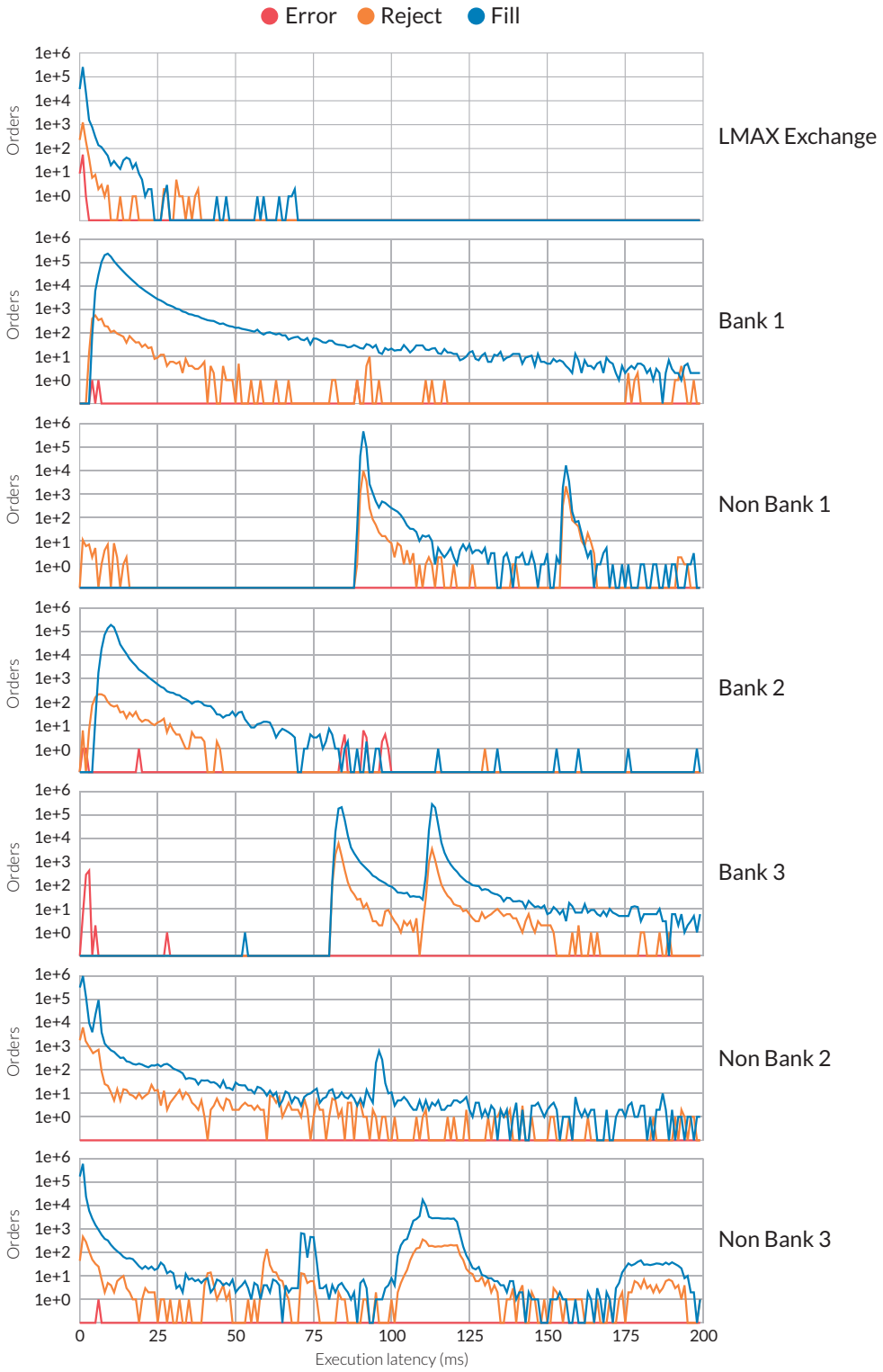

Chart 4: Execution times by venue for 2016

The multiple peaks and wide variation in the tail distribution of the latency histograms displayed by last look LPs require further investigation, and are suggestive of arbitrary changes to discretionary latency which, by definition, do not occur on firm liquidity.

Part I (iii): Applying standard metrics to a sample data set

Tail latency

The single hold time number needs to be supplemented by an evaluation of tail latency to form a complete view of the impact of hold times on execution quality. On paper, tables 9 and 12 (p. 31 & 38) place LMAX Exchange, Non Bank 2 and Non Bank 3 in the same category. However, chart 4 shows qualitatively very different profiles, ranging from the very well constrained to very long tails - sometimes including one or more peaks at higher latencies (for example Non Bank 2 and Non Bank 3). These peaks match the criteria outlined above for hold times – the only difference is that they are applied selectively to parts of the flow.
The distributions are not Gaussian, Poissonian or power law, so to quantify them we have to look at the percentiles of the distribution rather than statistical measures like standard deviation or mean.

As the curves in chart 4 are averages over all of 2016 it is useful to include the dimension of time to help us separate the cases where multiple peaks are due to changes over time as opposed to those where hold times may be selectively applied to parts of the flow on short timescales. To do this, we will use heatmaps.

Chart 5 shows a log scale heat map of execution latency for Bank 3 by month.

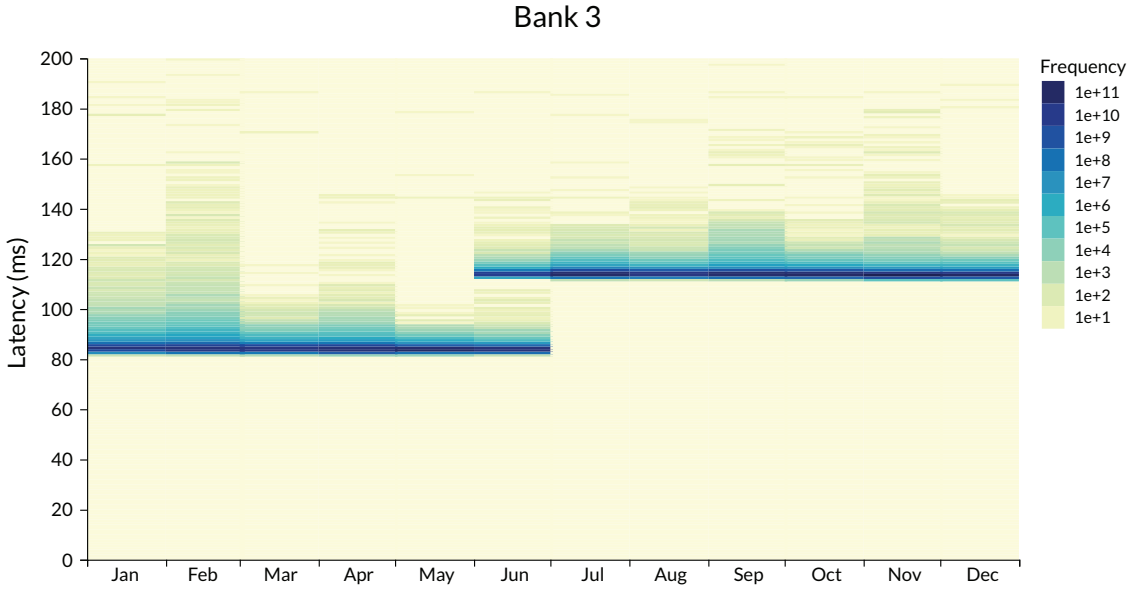

Chart 5: Execution latency by month for Bank 3

The x-axis shows calendar month with execution latency on the y-axis and the number of orders experiencing this latency indicated by the intensity of the coloured area on the chart – the modal latency (or latencies) showing up as darker. We can now see that the two peaks in the execution latency charts 3 (p. 30) and 4 are in fact a result of a distinct change in modal latency in June 2016.

Part I (iii): Applying standard metrics to a sample data set

Similar heatmaps for the remaining LPs are displayed in chart 6. Bank 1 and Non Bank 1 were not pricing during the first couple of months of the year, so their charts start in March and April respectively.

When LPs alter their base hold time the whole latency distribution is moved. This is particularly evident in chart 5 (p. 33). To remove the effect of the LPs varying their base hold times over the year, we calculated the delta between the 50% and the 99% and 99.9% levels on a month by month basis to arrive at a measure of the long tail mostly independent of value of the base hold time.

Table 10 below shows the results of this method and some sample characteristics of each LP's tail latency distribution displayed as the difference between the 50% and the higher percentiles.

Venue	50% (ms)	99% - 50% (ms)	99.9% - 50% (ms)
LMAX Exchange	1.0	1.4	6.8
Bank 1	9.7	22.3	70.7
Non Bank 1	98.2	1.9	9.0
Bank 2	10.1	15.7	32.6
Bank 3	98.7	6.8	21.7
Non Bank 2	1.0	2.7	32.0
Non Bank 3	10.3	23.3	103.2

Table 10: Tail latency percentiles

There are some objections to this rather simplistic approach – although for most LPs their tail latency is unchanged by moves in their base hold time (their whole distribution moves as a unit) this is not always the case – for example, in December 2016 Non Bank 3's tail latency improves dramatically (their 99.9%ile – 50% delta drops from 110ms to 25ms) once its base hold time has moved to 110ms. That is a strong indicator that part of their pre-December tail latency may be due to a more fine grained selective application of hold times to different parts of the flow and not purely the result of stochastic variation to the systematic latency, whereas from December rather than part of the flow being subjected to a hold time of 110ms, all of it is.

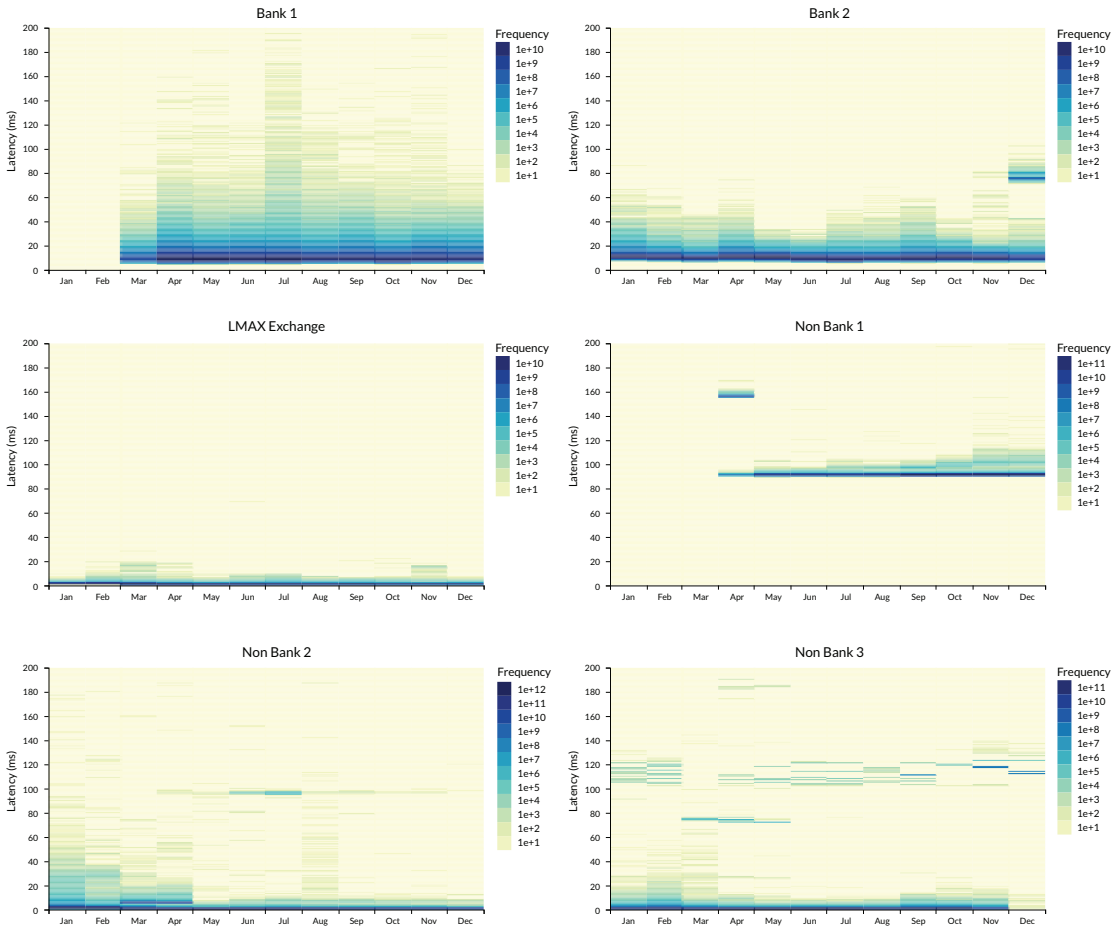

Chart 6: Execution latency by month heatmap

Ranking the 'quality' of tail latencies can be somewhat subjective – what may be acceptable for one trading strategy may not be for another. However, we can say that the last look LPs offer a choice of either a low hold time but a long tail latency (probably composed of selective application of hold times) or a high hold time and a reasonable tail latency (e.g. Non Bank 1 post April 2016). Only LMAX Exchange in this data set consistently provides both a low latency and a small tail latency.

Latency variation with time

The variation in execution latency can be continuous or discrete - both behaviours are shown in the heatmaps above. Tail latencies are a manifestation of continuous variation - particularly queueing or congestion under load as noted above. In this section we will examine the modal latency and discrete changes to it, as this indicates a change affecting all orders and may be a signature of a change to last look hold times.

At a high level we considered two possible causes of discrete changes in the latency profile:

- A step change in systematic latency due to upgrades or infrastructure changes;
- A step change in discretionary latency/hold time.

Part I (iii): Applying standard metrics to a sample data set

To determine if the change was due to a change in systematic latency or due to a change in the discretionary hold time, we looked at the timing of the change relative to the working week on the assumption that it is extremely unlikely that any infrastructure or software upgrade work would be performed within market hours unless a severe loss of service occurred.

A second consideration is the direction of the change. Infrastructure and software changes act mostly to improve the modal latency and reduce tail latencies. While it is possible that such a change could introduce an unintentional deterioration in latency characteristics we would expect to see those changes quickly rolled back or corrected. Our assumption therefore is that changes which increase the modal latency are more likely to reflect a conscious business decision. The last consideration is the size of the change. Humans are drawn to multiples of 5 and 10 [3]. Upgrades or infrastructure changes rarely show such preferences.

In chart 5 (p. 33) above there are two distinct populations with similar tail latencies. Between January and Wednesday June 22nd 2016 the modal latency was 84ms. After Wednesday June 22nd the modal latency changed to 114ms. Subtracting the systematic latency derived above gives us hold times of 80 and 110 ms. As this occurred mid week, the numbers are nice and rounded. We can then be confident that in this case a commercial decision was made to increase hold time around the UK EU referendum and never reset.

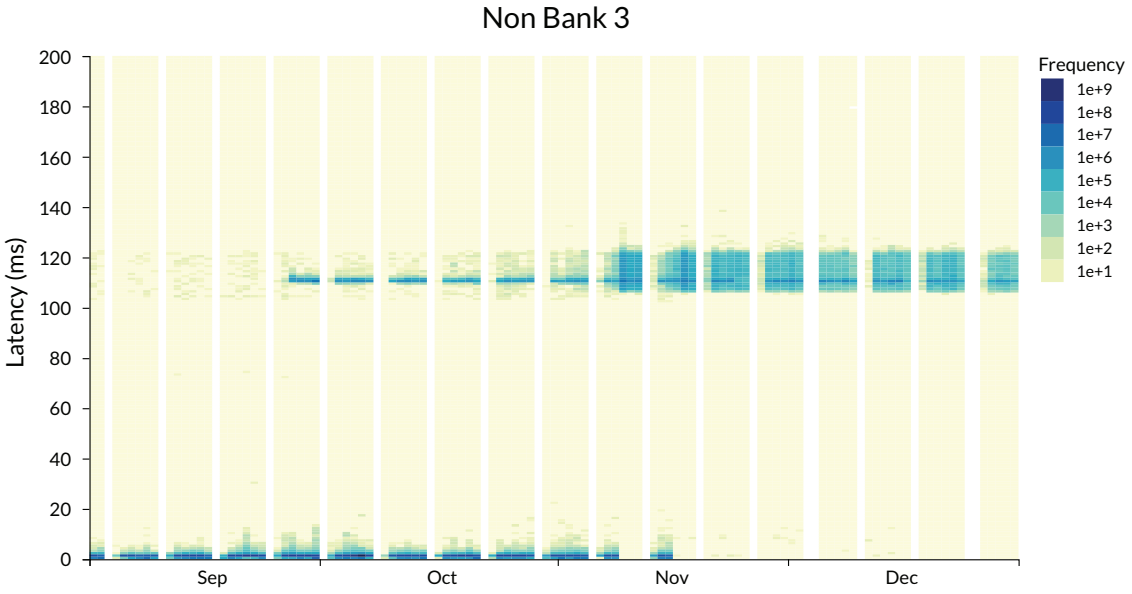

Chart 7: Execution latency by day for Non Bank 3

A final consideration is any evident selectivity in the change. Unplanned increases to systematic latency are more likely to affect all orders or increase the length of the distribution tail.
A change which introduces multiple concurrent modal latencies indicates that the higher latency is incurred selectively and is more readily explained by longer hold times being applied only to specific customers, orders or only at certain times.

Part I (iii): Applying standard metrics to a sample data set

An example of a set of selectively applied changes is shown in chart 7:

- Between January and September this LP consistently exhibited modal execution latency of 1ms, and at the beginning of September 95% or more orders are executed within 5ms. Other less pronounced latency peaks (containing less than 1% of orders) could be observed around 70, 110 and 180ms;
- On Tuesday 27th September the proportion of orders in the 110ms peak (distributed between 105 and 125ms) rises to 15%;
- Between Tuesday 27th September and Tuesday 8th November, the proportion of orders in the higher latency peak varies daily between 1.5% and 22.6%;
- On Wednesday 9th November the proportion of orders executed between 105 and 125ms rises to 99.7%;
- With the exception of a temporary re-appearance of low latency execution between Sunday 13th November and Tuesday 15th November, the higher latency peak represents more than 99.5% of orders for the remainder of the year.

The selectivity, timing and abruptness of these changes, with the lower latency execution being 'turned off' twice in the middle of the trading week, are all strong indicators of discretionary latency being applied with the higher range of hold times being run in parallel with the previous low latency settings before the increase is applied across the board.

In practice all of the LPs with the exception of LMAX Exchange and Bank 1 appear to experiment with different hold times during the year, for some or all of the trade flow, as summarised in table 11.

Venue	Time period	Day of change	Modal hold time(s) (ms)
LMAX Exchange	Jan 1st – Dec 31st		0
Bank 1	Jan 1st – Dec 31st		5
Non Bank 1	Jan 1st – Apr 28th	Thu	155
	Apr 28th – Dec 31st	Thu	90
Bank 2	Jan 1st – Dec 19th	Mon	9
	Dec 19th – Dec 22nd	Mon-Thu	70
	Dec 22nd – Dec 31st	Thu	9
Bank 3	Jan 1st – Jun 22nd	Wed	80
	Jun 22nd – Dec 31st	Wed	110
Non Bank 2	Jan 1st – Dec 31st		0
	Mar 16th – Apr 13th	Wed-Wed	5
	Jun 20th – Jul 21st	Mon-Thu	95
Non Bank 3	Jan 1st – Nov 16th	Weds	0
	Nov 16th – Dec 31st	Weds	110
	Mar 15th – Jun 17th	Tue-Fri	70
	Mar 15th – Jun 17th	Tue-Fri	180

Table 11: Timing of changes to hold times

Part I (iii): Applying standard metrics to a sample data set

Timing of rejects vs fills

Although not directly related to our TCA calculations, we also looked at the timing between rejects and fills. In conversation with traders it is often taken as a given that rejects will take longer to process than fills as part of a last look process. The usual rationale for this is that LPs will know quickly when they want to fill, but may take some time to review market conditions before deciding on a reject.

This turns out not to be true for the TPA data set. Rejects happen at the same time as fills or before, which is the opposite of the conventional wisdom. It is possible that this is an artefact of this particular data set and the LP's assessment of this particular trade stream's market impact.

Nonetheless, visual inspection of the profiles in chart 4 (p. 32) which are summarised in table 12 below, show that there is a clear division between Bank LPs and the rest. Traditional Bank platforms typically reject faster than they fill. Non Bank platforms and LMAX Exchange reject at the same time as they fill. The deltas are also not as large as we had expected and seem to be constant by LP irrespective of whatever the hold time is currently set to.

Venue (times in ms)	Fill modal hold time	Reject modal hold time	Fill - reject delta
LMAX Exchange	0	0	0
Bank 1	5	1	4
Non Bank 1	90	90	0
Bank 2	9	5	4
Bank 3	80	79	1
Non Bank 2	0	0	0
Non Bank 3	0	0	0

Table 12: Relative timing of fills vs rejects

Box 5
Hold time and execution latency analysis

The analysis of different components of execution latency shows that in this data set:

- Last look LPs adjust hold time depending on the market conditions throughout the year;
- Last look LPs with a shorter hold time exhibit less consistent execution latency, which may be indicative of longer hold times being applied selectively, while those with a higher hold time show more consistent latency;
- The hold times preferred by the last look LPs cluster around 100ms (2016 data).

The absence of discretionary latency means that only LMAX Exchange provides consistently low execution latency with no arbitrary variation over the year.

(iii) Section summary: hold time and execution latency

We have explored the different components of execution latency in this section. The hold times preferred by the last look LPs tend to cluster around 100ms, and we will use this figure as the basis for calculating the cost of hold time in the next section. Table 13 summarises the main findings. There is only one LP (Non Bank 2) with a comparable latency profile to LMAX Exchange, and even so there are horizontal lines on chart 6 (p. 35), most prominently in June/July but present throughout most of the year that indicate selective hold times at the 95ms level.

Venue (times in ms)	Fill modal hold time	Reject modal hold time
LMAX Exchange	0	6.8
Bank 1	5	70.7
Non Bank 1	90 & 155	9.0
Bank 2	9 & 70	32.6
Bank 3	80 & 110	21.7
Non Bank 2	0	32.0
Non Bank 3	0 & 110	103.2

Table 13: Summary of execution latency characteristics by venue

Two out of three Non Bank LPs favour selective application of hold times, whereas two of three Bank LPs favour a simple base hold time and a moderate latency tail.

Only LMAX Exchange provides low and consistent execution latency with no hold times or variation over the year.

Metrics scorecard

- **Execution time 50%ile**. LMAX Exchange is the clear winner here, with two Non Bank LPs being competitive and one with an obvious hold time. Only one of the Bank LPs is competitive, leaving them last.
- **Hold times/long tail latency**. Again LMAX Exchange is the clear winner. The second and third places swap due to the presence of multiple hold times for the Non Bank LPs.

Metric	Bank 'last look'	Non Bank 'last look'	LMAX Exchange
Execution time 50%ile	1	2	3
Hold times/long tail latency	2	1	3

Table 14: Hold time score card points (higher is better)

Part I: Summary of findings:

Full metrics scorecard for each class of LP on our sample data set:

Metric	Bank 'last look'	Non Bank 'last look'	LMAX Exchange
Market order fill ratio	2	2	3
Limit order fill ratio	3	2	1
Market order slippage rate	2	2	2
Limit order price improvement	1	1	3
Execution time 50%ile	1	2	3
Hold times/long tail latency	2	1	3
Totals	11	10	15

Table 15: Combined scorecard by class of LP

For simplicity we have restricted the scorecard to two metrics per category. We are aware of the "well they would say that" observation, irrespective of the lengths we have gone to internally to conduct an unbiased and fair assessment. The TPA data set was chosen to compare the behaviour of a single common trading strategy on both styles of liquidity. In common with many FX traders today that strategy has been developed for last look and poses apparent problems for firm liquidity. We discuss the limit order fill ratio findings and explore the causes behind it in Part II.

There are several clear points that emerge for the TPA sample data set:

- The market order fill ratio is close to 100% on firm liquidity, LMAX Exchange;
- The limit order fill ratio for LMAX Exchange appears to be comparatively very low;
- Limit order fill ratio on LMAX Exchange is related to pricing behaviour on the venue;
- LMAX Exchange is the only provider offering significant limit order price improvement;
- LMAX Exchange exhibits consistently lower execution latency and shorter tail latency.

In addition, the following extra metrics or aspects have emerged from the analysis:

- Expected slippage/improvement ratio is 2:1; higher values relate to non market factors;
- The degree of price improvement for limit orders should mirror that for market orders. If not, that is an indication limit orders are not being filled at the underlying market price;
- Latency thresholds which change during the week are hold times set for commercial reasons;
- Hold times may be selectively present on parts of the flow, and manifest as long tail latency.

Box 6

Summary of the key metrics analysis

Fill ratio, price variation and hold time execution metrics suggest the following differences when trading FX on firm liquidity:

- The market order fill ratio is close to 100%; whereas the limit order fill ratio is lower when compared to last look venues, suggesting pricing behaviour differences on firm vs last look liquidity;
- Only firm liquidity offers significant limit order price improvement;
- Execution on firm liquidity is not subject to discretionary latency;
- Firm liquidity exhibits lower and more consistent execution latency than other venues.

blank page

Part II

Execution quality metrics and firm liquidity

Part II: (i) Execution quality metrics and firm liquidity

In this part we will investigate the underlying mechanisms behind the apparently anomalous limit order fill ratio seen on the sample TPA data set and explain in detail our approach to quantifying the financial value of price improvement and cost of higher execution latency or hold time preparing the ground for the relative TCA calculations in Part III.

During the second half of 2016, while researching what would become this paper, we extended the market and execution quality statistics collected by LMAX Exchange venues. The data in this section is based on insights from statistics collected on our London, Tokyo and New York exchanges supplemented by other third party data sets where noted.

In our sample TPA trade database the overwhelming majority (over 90%) of limit order rejects on LMAX Exchange are cancels caused by price fluctuation/volatility. This observation and the limit order fill ratio result leads to the questions we will address first:

- Why do the flow characteristics in the TPA data set allow for higher fill rates from last look providers?
- Can we observe higher price volatility on LMAX Exchange which may lead to a lower fill ratio for limit orders?
- Are there any order placement strategies that will improve limit order fill rates?

Most importantly we will show why it is worth going to this trouble for traders who are currently satisfied with the nominal fill rates offered by their last look LP.

Box 7

Approach for quantifying the value of execution on firm liquidity

To quantify the value of firm liquidity, the previous analysis of fill ratio, price variation and hold time metrics needs to be developed by:

- Exploring the relationship between higher price volatility and lower limit order fill ratios observed with firm liquidity;
- Understanding order placement strategies that improve limit order fill rates;
- Calculating the value of price improvement;
- Quantifying the cost of higher execution latency or hold time on last look venues.

(i) Market impact

Over the last few months we have discussed early drafts of our results with a variety of industry experts. The anecdotal evidence from comparison with their own experience is that the fill ratios observed for limit orders for all of the last look venues in the TPA data are unusually high, leading to a conjecture that the TPA trade flow was very benign from the perspective of the liquidity providers. In order to prove this theory, we have applied a similar methodology to the majority of FX market makers (see [4] for an example from a leading Non Bank LP) and started the collection of market impact statistics on the LMAX Exchanges.

Our approach to measuring market impact is to calculate the profit and loss of each trade relative to the market mid price at various time points after execution (from milliseconds to

minutes). In normal conditions each trade will instantaneously appear as profitable for the LP (and therefore loss making for the trader) due to the bid-offer spread. The LP's profit will then rise or fall as prices change. Flow is considered to be 'benign' (also described as 'soft' or 'uncorrelated') if it takes a long time before becoming unprofitable from the LP's perspective or remains profitable over the entire measurement horizon. The other extreme is 'toxic' (also described as 'sharp' or 'highly correlated') flow which becomes unprofitable from the LP's perspective over very short time horizons (milliseconds to seconds). Each LP will have their own view as to what time horizons demarcate the boundaries between correlated, average and uncorrelated flow.

We can analyse the net impact of a large number of trades by aggregating the profit or loss at the same post trade measurement times. To allow for trades of different sizes and on different instruments we can express the profit as a fraction of the notional value traded to create a composite market impact. An LP would normally expect profits to decline from the value at trade execution given the directional bias in trading activity (i.e. given a bias towards buying on rising and selling on falling prices, most trades will show a short term deterioration for the LP). The extent and gradient of this decline, and the recovery (if any) of profits over longer time scales characterise the difficulty of servicing a particular trader's flow from the perspective of the LP.

An aggregate trading pattern which remains consistently profitable for the LP, even if it shows some short term decline, provides the LP with the option to hold positions and internalise this trading against other clients or hedge with limited impact on the underlying market, and would usually be considered benign and uncorrelated.

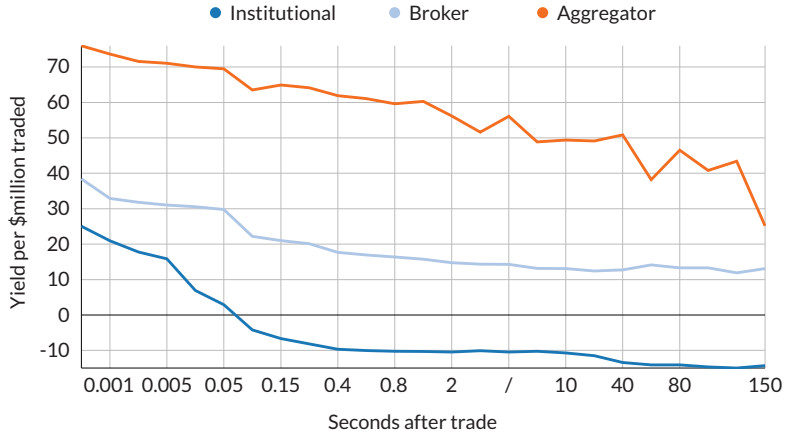

Chart 8: Market impact by profile on LMAX Exchange

Chart 8 shows the market impact for a set of different customer profiles with different flow types on LMAX Exchange London. The three different profiles are;

- **Broker:** Customers of LMAX Exchange Broker
- **Institutional:** Customers of LMAX Exchange MTF
- **Aggregator:** Customers with a similar profile to the TPA

The criteria for each profile are covered in more detail in the section 'Quantifying the value of price improvement' (p. 51).

Part II: (i) Execution quality metrics and firm liquidity

The Aggregator profile flow is very benign with a net positive P&L over all time horizons shown. For this type of flow, last look providers can provide customised very tight pricing. In contrast the LMAX Exchange firm liquidity model of an open anonymous central limit order book which caters for a wide variety of flow types is placed at a natural disadvantage. As a second consequence of the uncorrelated nature of the flow, we would also expect that the reject rates offered by the last look LPs should be very low when compared to other more correlated flow types, as observed previously.

The Aggregator profile flow is, however, not typical of all trading, or indeed all flow through aggregators. More typical average impact profiles for all Broker and Institutional customers on the London exchange are also shown in chart 8 (p. 45), with an expected crossing of the zero point for the Institutional customers at a very familiar number - 100ms.

This answers our first question – for this specific customer profile, one of the LMAX Exchange USPs of the 'same price for everyone' places firm liquidity at a disadvantage to last look where tighter prices can be offered based on the assessment of market impact.

However, even for this profile there are other considerations of the firm liquidity trading model which can offset this head start for last look, and we will now look at the impact of price stability on the particular execution style favoured by this subset of aggregators.

Box 8

Assessment of market impact and differentiated pricing by last look providers

- The analysis of different trading profiles shows that the trade flow from the TPA had minimal market impact;
- Trading strategies with more sensitivity to market impact risk information leakage, potentially resulting in disadvantageous price changes by last look LPs ahead of full execution;
- These strategies benefit from trading on firm liquidity, which does not suffer from pre-execution information leakage.

Part II: (ii) Execution quality metrics and firm liquidity

(ii) Price volatility in firm liquidity

The commonly cited benefits of last look in FX are tighter prices for traders and the optionality protection it affords market makers. In the event that the LP's pricing is wrong with respect to market conditions then last look gives them the option to reject trades.

The corollary of that is that for a firm liquidity venue where optionality is not present, the time to cancel a stale (or even unnecessarily competitive) price and update with a newer price becomes key to an LP's profitability, leading to a very rapid turnover of liquidity at the top of the order book. This is similar to the latency 'arms race' seen in other asset classes – for example equities – where competition amongst cross connected or co-located LPs to position themselves within the book on scales of micro to milliseconds is well documented [5]. As a result, a much more dynamic market pricing structure is to be expected on firm liquidity venues. This manifests as far more price variation or 'jitter' on short time scales.

The TPA data set includes the number of top of book market data updates per second recorded for each venue. Unfortunately the data is recorded at minute level averages, which eliminates all ability for us to compare burst rates or other microstructures in the data. Nonetheless the number of updates can stand as a proxy for the general volatility of the book.

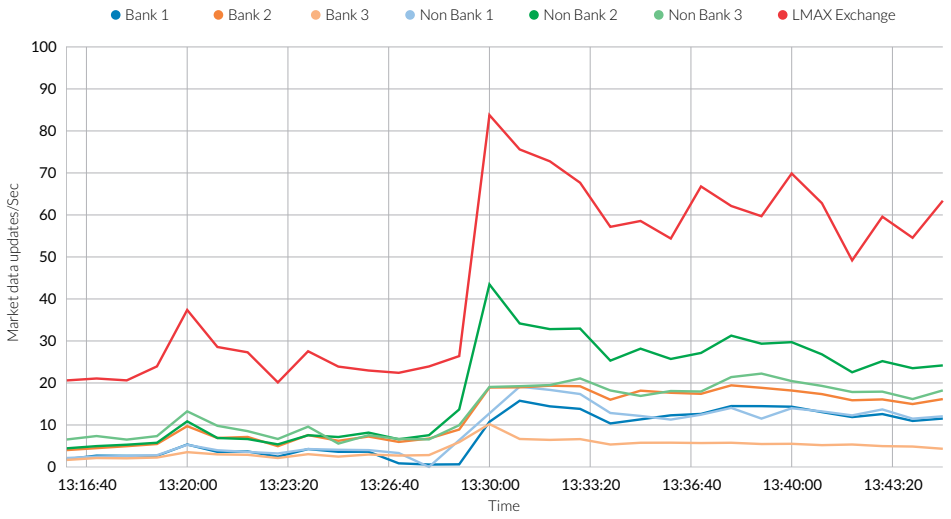

Chart 9: market data update rates for venues over Non Farm Payroll

Chart 9 plots the top of book data rates across a US Non Farm Payroll (NFP) announcement. As expected from the cancel time argument above LMAX Exchange has a much higher update rate than the last look LPs, even with the high degree of smoothing or 'averaging down' in per-minute aggregates. The approximate ratio of LMAX Exchange market data rates to the next most active provider (a factor of 2-3) is maintained over normal trading as well as over market events which cause spikes and step changes in data rates.

The sample data above however can understate the real situation in a way that is particularly relevant to the execution style of the TPA. The actual rate at which prices change in the matching engine, which will determine whether or not a limit order can be filled will be higher than the rates most clients observe in their market data for a variety of factors (see Appendix B – factors affecting Internet delivery of market data).

Part II: (ii) Execution quality metrics and firm liquidity

We can measure price jitter by examining the distribution of lifetimes of local price peaks in the raw tick by tick market data. That is events when an increase in the best bid is immediately followed by a decrease (or a decrease in the best offer is immediately followed by an increase) with a skew towards shorter lifetimes representing higher price jitter. While short lived peak prices may be caused by both liquidity churn and trading, from the trader's perspective the effect is the same – the best price in the recent time window appeared and disappeared before there was time to execute on it.

Short lived peaks are not the only form of problematic price change - a price which is consistently but rapidly rising or falling may be just as difficult for specific trading strategies – however they do characterise a behaviour which is difficult to trade with immediate execution limit orders regardless of strategy or sentiment.

Analysing internal LMAX Exchange market data directly from the matching engine for the 30 minute window surrounding the NFP market event illustrated above, we find that:

- 69% of price peaks last less than 10ms;
- 90% last less than 40ms.

Even at less volatile times the rate of price updates remains challenging. The full day and two week period around the same event showed:

- 62-63% of price peaks last less than 10ms;
- 90% of peaks last less than 110ms.

The comparison with the 'fundamental' clock of primary venues used for price discovery in FX with market data update times of 100ms or 250ms is quite striking. Even with the recent launch of 20ms and 5ms interval 'ultra' updates [6] the fact remains that at present the tick by tick price movements from a firm liquidity venue are happening at rates of a least a factor of 5 - 10 higher than even a 5ms tick would allow.

The TPA's 'price-sniping' execution strategy, which is nominally designed to secure the end customer the best price (and all other things being equal, lowest transaction cost) by holding their order in the aggregator awaiting 'the right price' is in fact counter-productive when firm liquidity is included in the mix. The orders placed have close to top of book limit prices and if they have a 'tick-to-trade' latency of more than a millisecond or so they are almost guaranteed to fail more frequently due to misses caused by price volatility and jitter.

The simple application of a reject ratio in this case misses the point that these are the expected market dynamics in an active competitive firm liquidity venue.

Part II: (ii) Execution quality metrics and firm liquidity

To recap, in order to be successful with its price-sniping strategy, with limit prices targeting a potentially jittery top of book, the aggregator must:

- **First be able to ingest a higher rate of market data than typically received from other LPs;**
- **Have access to a high bandwidth, high market data update rate connection in the same data centre;**
- **Process market data to identify the desired trading conditions and place orders with a sufficiently low tick-to trade latency to ensure the desired price is still available.**

Viewed from this perspective the relatively low limit order fill rates of the TPA on LMAX Exchange liquidity are not surprising when compared to the far more sedate price volatility behaviour on last look venues, quoting prices in response to a much slower 5ms, 20ms or 100ms price tick.

This appears to create something of a perfect storm for LMAX Exchange limit fill ratios for this particular execution style. If the true value of execution on firm liquidity, as shown in Part III: 'Comparative transaction cost analysis' (p. 68) is to be realised then this particular last look optimised strategy needs to be adapted to work with firm market price dynamics and microstructure.

However, this is a time limited strategy even for customers with no interest in firm liquidity. The price-sniping approach works well when the latency for the tick-to-trade is less than the latency of market data updates. The increase in market data update rates across all venues, as noted above, is likely to continue to at least 1ms update rates, and possibly real time streaming prices such as those already found on LMAX Exchange.

This is likely to drive customers reliant on this method of execution to consolidate their FX trading even closer to venues.

Last look LPs will face trade-offs of either maintaining hold times or shortening quote lifetimes and falling back on a higher reject rate or – if they're already pricing firm liquidity venues - simply applying the same fast cancel approach already employed to avoid rejecting orders hitting a stale price. Either way, it is likely that quote rates will increase and some of the factors listed above as negatives for LMAX Exchange will start to come into play across more venues. The market impact of the customer will become the main factor driving the decisions to accept trades based on 'stale prices' that are more than a few ms old.

As we will see below having transparent market dynamics means that a trader can redesign their execution strategy on firm liquidity to take account of this new world of increasingly high pricing or quoting rates in a way that is very hard to match on last look liquidity.

Box 9

Understanding implications of higher price volatility on firm liquidity

Market makers respond to firm liquidity by updating prices more frequently, resulting in more price variation or 'jitter' over short time scales.

A comparison of the number of price updates received by a sample customer shows LMAX Exchange market data rates at 2-3 times those of the most demanding last look venues, while an examination of the internal matching engine statistics shows the majority of new top-of-book prices lasting less than 10ms.

If the true value of execution on firm liquidity is to be realised, then it needs to be adapted to work with firm market price dynamics and microstructure:

- Customers with high market impact profiles, for whom last look LPs face trade-offs of either maintaining hold times or shortening quote lifetimes, are likely to derive substantial value from transparent execution on firm liquidity;
- Having transparent market dynamics means that a trader can evolve their execution strategy on firm liquidity to take account of this new world of increasingly high price or quote rates in a way that is very hard to match on last look liquidity.

Part II: (iii) Execution quality metrics and firm liquidity

(iii) Quantifying the value of price improvement

For traders using limit orders, price improvement represents a profit relative to the trade as filled at limit price. If we view the TPA flow as a single 'customer', then using the LMAX Exchange execution quality reports we observe a price improvement over the entire period equivalent to $3.17 per $1M notional traded with monthly averages ranging from $0.96/million to $8.48/million. The notional volume traded is based on all limit order trades, not just those receiving price improvement, hence the per-million equivalent should be interpreted as an aggregate reduction in trading costs across all trading activity.

Month	Improvement (USD)	Notional (USD)	$/million	Trades	
				Total	Price improved
Jan	224.10	122,659,512	1.83	1,304	63
Feb	200.62	151,925,601	1.32	8,155	212
Mar	71.37	52,094,610	1.37	1,676	39
Apr	297.74	59,143,226	5.03	597	88
May	381.64	81,025,408	4.71	1,053	126
Jun	626.11	109,260,141	5.73	2,552	270
Jul	227.22	97,732,787	2.32	2,546	137
Aug	167.54	80,191,782	2.09	916	60
Sep	661.04	77,956,809	8.48	1,778	103
Oct	231.47	78,953,242	2.93	1,488	107
Nov	151.52	82,229,754	1.84	1,372	112
Dec	42.12	43,830,233	0.96	822	43
Q4	**425.11**	**205,013,229**	**2.07**	**3,682**	**262**
Annual[ii]	**3,282.49**	**1,037,003,105**	**3.17**	**24,259**	**1,360**

Table 16: TPA price improvement by month

The price improvement shows a variable, but significant, potential for reducing the overall cost of trading. As LMAX Exchange is the only LP offering discernible price improvement on limit orders, we cannot benchmark this behaviour against other sources of liquidity.

To provide some context for the price improvement achieved by the TPA we will look at the relative performance of other LMAX Exchange customers matching the three customer profiles introduced earlier. As this relies on the execution quality statistics developed during the course of 2016, data covering all customer profiles is only available for Q4 2016.

The aggregator profile consists of a small group of customers with known trading patterns similar to those of the TPA. While they may not trade via an aggregator as such, they are co-located, low latency and price sensitive. They target prices in the same way as the TPA, resulting a similar profile with relatively low price improvement and relatively high fill ratios.

[ii] *The total here represents a relatively small segment of the total TPA data set of $85bn notional value, of which LMAX Exchange received $11bn, $10bn in market orders and $1bn in limit orders*

Part II: (iii) Execution quality metrics and firm liquidity

The Broker and Institutional profiles gather a wider set of customers with more diverse behaviour, divided into clients of LMAX Exchange Broker (typically other brokers, individual traders and small funds) and direct Institutional clients of LMAX Exchange MTF (typically banks, larger funds and HFTs). While the main structural distinction is that Institutional clients have their own prime broker and do not require the services (and associated risk controls) of LMAX Exchange Broker, the group are also characterised by lower transit latencies, often achieved by co-location, and higher overall investment in trading infrastructure. Compared to the TPA and the aggregator profile, customers in the Broker and Institutional profiles exhibit much more varied approaches to setting limit prices. While some customers place limit orders at or close to the price at which they expect a fill, others are willing to pay a premium to get into the market during periods of high price volatility. In some cases prices are even further off market, but the customer actually expects to trade at a better price, so what is superficially traded as a limit order really represents a market order with the limit price setting a slippage threshold.

Including orders with limit prices significantly off market would skew results towards higher price improvement figures. In the interests of having a fair and realistic assessment of the value delivered to the customer by price improvement, when reviewing the Broker and Institutional profiles we also need to restrict ourselves to orders where we believe the customer was trading with some expectation of being filled at their stated limit price. We will identify these as orders within a given price range – or 'volatility band' – of the market price at the time we received the order.

We examined several approaches to measure volatility in the literature [7] in order to define characteristic market ranges. These included Average True Range (ATR), realised volatility of returns and simpler statistical measures of variance. The most useful turned out to be the traditional Bollinger Band calculation [8], which as well as being familiar and easy to compute, has several immediately useful properties. As opposed to the more dimensionless volatility metrics we can directly use it to establish a pricing range and it accounts for moves in market direction which the simpler statistical methods do not. Additionally the market ranges given by this method also passed a 'real world' sanity check as to what constituted 'close to market' from colleagues with trading backgrounds.

The volatility band was computed as a daily mean for each instrument using minute level tick aggregates from LMAX Exchange's historic market data service. The parameters used to compute the volatility band was the standard B(20,2), i.e. 2 standard deviations measured over 20 observations. If B(t, N, m) is the Bollinger Band for observation at time t for m standard deviations measured over N observations, and P_t is the mid price at time t (computed from the high and low of an OHLC bar), then over n samples in a day (where typically n=1435 due to a 5 minute closure at end of day) our volatility band is calculated as:

$$B(t, N, m) = m \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} \left(P_{(t-i)} - \left(\frac{1}{N} \sum_{s=0}^{N-1} P_{(t-s)}\right)\right)^2},$$

$$Volatility\ band = \frac{1}{n-19} \sum_{j=20}^{n} B(j, 20, 2)$$

Figure 1: Daily 'close to market' volatility band

Part II: (iii) Execution quality metrics and firm liquidity

We take the daily volatility band and apply it to each immediate execution limit order for that day discarding any which have a limit price further than the range from top of book. Days with higher sustained levels of price change will have wider volatility bands while relatively stable days, even those punctuated by a small number of market events, will have narrower bands. Some sample dates and volatility band values for EURUSD, which has an annual mean volatility band of 4.5 pips, are shown in table 17.

Date	Calculated Daily Range (Pips)	Notes
2016-12-02	4.5	December NFP
2016-12-23	2.9	Pre-Christmas 'quiet' day
2016-06-23	7.0	Day of EU referendum
2016-06-24	23.3	Day after EU referendum
2016-12-05	8.8	Volatile day

Table 17: Sample volatility band values

Orders outside the volatility band are excluded from our calculation of price improvement. For example, any EURUSD limit order placed on 2nd December with a limit price more than 4.5 pips away from the market price, or on 23rd December with a limit price more than 2.9 pips away from the market price, was assumed to be a customer deliberately setting an off-market limit price to secure a fill.

Appendix C lists summary volatility band details for each currency pair. Tables 18, 19 and 20 (p. 54) show the price improvement for Broker and Institutional profile customer orders within the volatility band on the LMAX Exchange London venue, and Aggregator profile customer orders across multiple LMAX Exchange venues.

Broker profile

Month	Improvement (USD)	Notional (million USD)	$/million	Trades	
				Total	Price improved
Oct	963,487	27,049	35.62	200,972	116,940
Nov	1,271,394	37,314	34.07	264,502	144,490
Dec	2,522,175	35,349	71.35	246,180	167,009
Q4	**4,757,056**	**99,712**	**47.71**	**711,654**	**428,439**

Table 18: Broker profile price improvement, orders within volatility band (2σ)

Institutional profile

Month	Improvement (USD)	Notional (million USD)	$/million	Trades	
				Total	Price improved
Oct	60,071	29,846	2.01	95,752	10,096
Nov	252,307	63,511	3.97	212,542	31,416
Dec	158,474	43,380	3.65	180,770	20,530
Q4	**470,852**	**136,737**	**3.44**	**489,064**	**62,042**

Table 19: Institutional profile price improvement, orders within volatility band (2σ)

Part II: (iii) Execution quality metrics and firm liquidity

Aggregator profile

Month	Improvement (USD)	Notional (million USD)	$/million	Trades	
				Total	Price improved
Oct	14,792	6,484	2.28	18,930	2,832
Nov	48,585	19,536	2.49	49,589	6,968
Dec	36,856	18,595	1.98	45,510	4,444
Q4	**100,233**	**44,615**	**2.25**	**114,029**	**14,244**

Table 20: Aggregator profile price improvement, all orders

Comparing the Q4 number of $2.07/million from table 16 (p. 51) for the TPA with the wider aggregator profile above gives us a similar price improvement figure. The institutional profile fares slightly better at $3.44/million, however it is the broker profile that really stands out at $47.71/million, table 18 (p. 53).

In case these numbers for the broker profile appear to be unduly high, we also calculated the same price improvement numbers following the same method, but assuming a much tighter volatility band of 1σ or B(20,1). These are shown in table 21.

Month	Improvement (USD)	Notional (million USD)	$/million	Trades	
				Total	Price improved
Oct	382,454	25,005	15.30	174,867	90,835
Nov	880,920	35,942	24.51	249,964	129,953
Dec	1,523,378	31,602	48.21	215,872	136,701
Q4	**2,786,752**	**92,549**	**30.11**	**640,703**	**357,489**

Table 21: Broker profile price improvement orders within tighter volatility band (1σ)

Even using this much tighter restriction we see that the average price improvement seen by broker customers on LMAX Exchange London was $30.11/million over Q4.

To address the concern of cherry picking our quarter, we would point out that reviewing the data by quarter in table 16 (p. 51) indicates that Q4 is certainly not the best quarter for us to have picked. As mentioned above, this is just the data available to us at the time of writing.

The causes of the difference between the broker profile and the aggregator or institutional profiles will be further explored in the section 'Optimising trading on firm liquidity' (p. 56). For now the main observation is that both the aggregator and institutional profile customers are often co-located in the same data centre as the exchange and have dedicated high performance hardware.

This is however, not typical of the average broker profile customer and so market dynamics and latency are once more factors. Price improvement is a function of many variables. The customer's order placement strategy and market volatility all play a part in determining the actual price improvement seen. Nonetheless for the majority of the LMAX Exchange customer base who match the broker profile, we conservatively estimate price improvement on their limit orders is worth between 3 and 4 ticks on the spread.

Box 10

Quantifying the value of price improvement

For traders using limit orders, price improvement represents a profit relative to the trade as filled at limit price. Price improvement, only achieved at any significant level on firm liquidity, shows a variable but considerable potential for reducing the overall cost of trading:

- The customer's order placement strategy and market volatility play a part in determining the actual price improvement seen;
- The average price improvement for the broker profile is $47.71/million in Q4'16;
- More conservatively, the majority of the LMAX Exchange customer base who match the broker profile, achieved price improvement on their limit orders worth between 3 and 4 ticks on the spread.

Part II: (iv) Execution quality metrics and firm liquidity

(iv) Optimising trading on firm liquidity

So far, we have considered price improvement and fill ratio independently. Chart 10 demonstrates the relationship between these two metrics for our sample data set. Each data point represents a day of trading which includes at least 100 orders and plots the average fill ratio and price improvement ratio (i.e. the proportion of orders filled at better than limit price) observed on that day. The size of the circle represents the relative number of orders placed on each day.

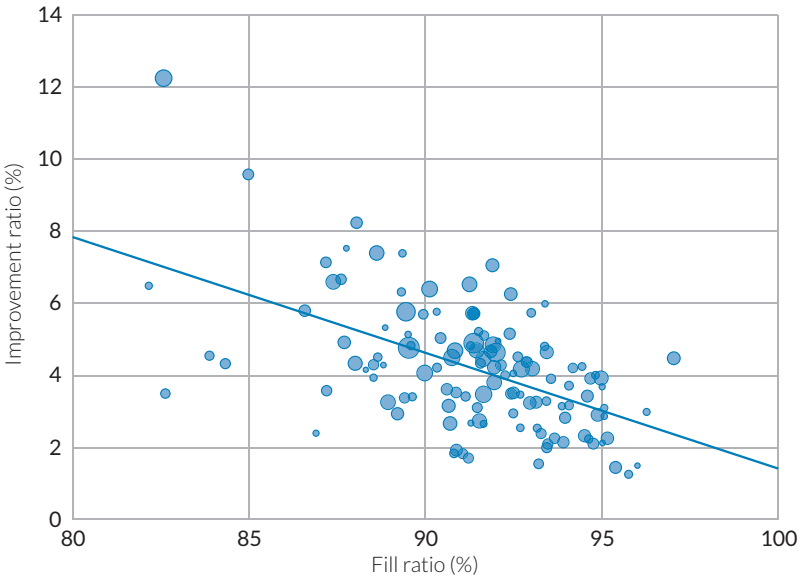

Chart 10: Correlation of fill ratio and improvement ratio

Higher levels of price improvement generally correspond to lower fill ratios. This outcome is entirely intuitive; if we assume a reasonably agnostic sentiment across the underlying trading decisions then the market conditions likely to produce larger favourable moves for some orders are also likely to produce a higher level of rejects for others.

This effect is amplified as the time for a trader to receive and respond to the market data associated with a price change increases. Traders may adopt a variety of approaches to address this gap and improve fill ratios:

- Reducing market data and trading latencies by tuning software and infrastructure or by geographical co-location;
- Throttling or rate-limiting market data to eliminate latency introduced by peaks in message rates causing queues of unconsumed messages to accumulate in the infrastructure between the LP and the trader (see Appendix B – factors affecting Internet delivery of market data);
- Implementing their own hold time by placing and cancelling Good-for-Day or Good-for-Time limit orders if venues supported them. These will result in a fill if the desired price becomes available again within an acceptable time frame (which brings with it the same opportunity cost issues of venue hold time, albeit under the trader's own control).

Part II: (iv) Execution quality metrics and firm liquidity

A less intuitive step for traders who have optimised their strategies for last look is that there is an option to improve fill ratios on firm liquidity by introducing a small level of slippage. This works by trading the price improvement received on filled orders for a smaller cost increase on orders that would otherwise be rejected.

To quantify this opportunity, we have recorded a virtual execution for every order placed on the LMAX Exchanges since September 2016. The virtual execution is performed against liquidity identical to that available both at the time the order was executed and at multiple time points up to 100ms later, but removes any limit price restriction. This allows us to model, order by order, the amount of slippage that would be required to secure a complete fill.

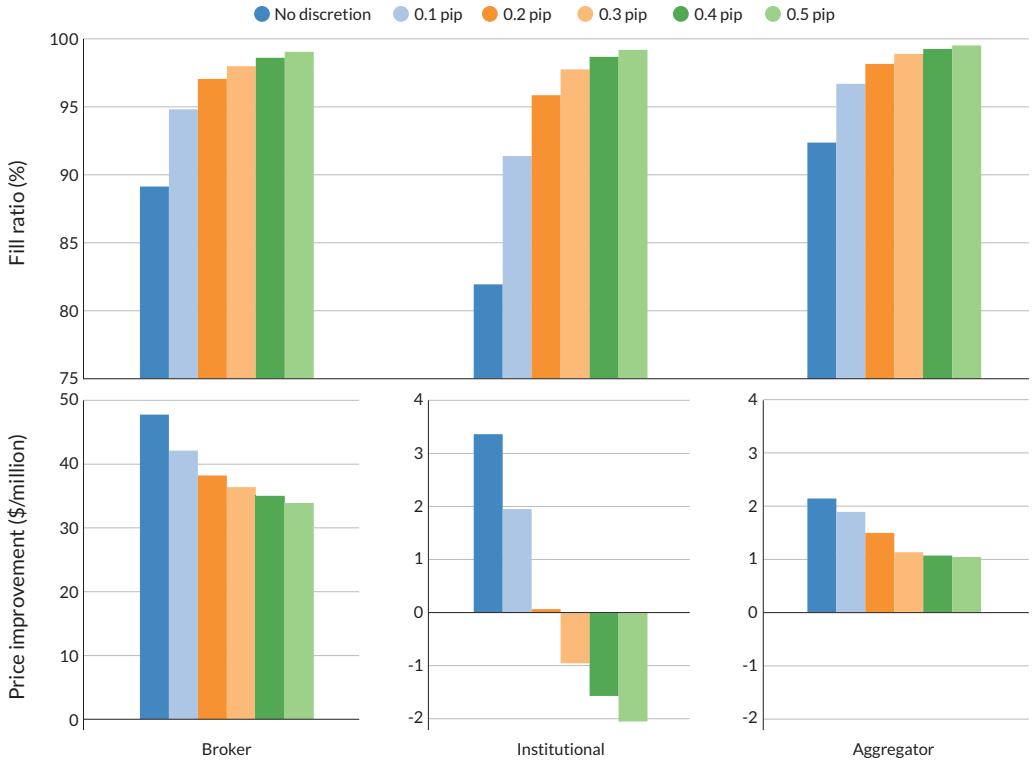

Chart 11: Trading price improvement for fill ratio

Chart 11 and table 22 (p. 58) show the results of this analysis for each of our three customer profiles for all trading in Q4 2016.

For each profile we have calculated the number of additional orders that would have been filled by applying price discretion by increasing levels of slippage in 0.1 pip intervals, represented as an increased fill ratio. Price discretion giving slippage at levels above 0.5 pips is possible but, for this data set, there is a clear case of diminishing returns as discretion levels above 0.2 or 0.3 pips are used.

This is expected behaviour given that the underlying driver for both price improvement and limit fill ratio on firm liquidity is price volatility and microstructure, with the diminishing returns displaying the familiar signature of a Chebyshev inequality or price distribution with the majority of missed prices very close to the requested limit price. Traders can make use of

knowledge about the shape of the volatility distribution to determine their ideal fill ratio/price improvement mix.

It must be stressed however that the above charts are averages for each different customer profile over a quarter, and should not be used as a guide or expectation for any individual trading session or strategy. For example applying a price discretion of 0.3 pips improves the mean fill ratio for the aggregator profile from 92.37% to 98.91% while still delivering price improvement and therefore offering scope for further fill ratio improvement before reaching break-even. By comparison at the same level of discretion the institutional profile shows net slippage with a 95.87% fill ratio. Individual customers will see quite different numbers depending on their latency and trading strategy.

Profile	Maximum slippage	Fill ratio	Net $/million
Aggregator	None	92.37%	2.25
	0.1 pip	96.69%	1.81
	0.2 pip	98.17%	1.55
	0.3 pip	98.91%	1.36
	0.4 pip	99.27%	1.24
	0.5 pip	99.52%	1.14

Profile	Maximum slippage	Fill ratio	Net $/million
Institutional	None	81.95%	3.44
	0.1 pip	91.40%	1.83
	0.2 pip	95.87%	0.19
	0.3 pip	97.77%	-0.90
	0.4 pip	98.69%	-1.60
	0.5 pip	99.20%	-2.09

Profile	Maximum slippage	Fill ratio	Net $/million
Broker	None	89.15%	47.71
	0.1 pip	94.82%	42.19
	0.2 pip	97.06%	38.30
	0.3 pip	98.00%	36.48
	0.4 pip	98.62%	35.11
	0.5 pip	99.06%	34.01

Table 22: Trading price improvement for fill ratio

In summary, the optimal level of price discretion and allowable slippage will depend on a customer's risk appetite and the relative value they place on securing a fill and getting the best possible price. Feedback from customers, as well as, the negligible limit order price improvement from last look LPs observed in the TPA data indicates that the option to exercise this discretion and level of control over trading is, at least currently, unique to firm liquidity venues.

Box 11

Optimising trading on firm liquidity

- Firm liquidity enables traders, who have optimised their strategies for last look, to improve fill ratios by permitting a small level of slippage, which is offset by price improvement on other trades;
- The optimal level of price discretion and allowable slippage will depend on a customer's risk appetite and whether they place more value on securing a fill immediately, or filling at the requested price even at the risk of missing a move in the market.

Part II: (v) Execution quality metrics and firm liquidity

(v) Quantifying the cost of hold time

Hold time represents one of the most significant hidden costs of trading. This is most apparent in the case of rejected or unfilled orders. A trader makes the decision to place an order and must wait an unknown period of time before discovering that the order cannot be filled as requested, by which time the market may have moved significantly. LPs with published methodologies for last look typically cite an adverse change in underlying prices as the principal reason for rejecting an order, and we can see from our analysis of the limit order fill ratio on LMAX Exchange liquidity that the vast majority of unfilled orders are caused by unfavourable price moves. As a result we can expect the distribution of price changes to be dominated by those which increase transaction costs relative to the limit price of a rejected order.

The impact of hold time for successful orders will depend on the individual trader's trading strategy. Those with a high frequency strategy will clearly be restricted as the required interval between trades approaches the range of hold times experienced. Similarly traders who rely on working large orders into the market in a series of smaller orders will be constrained on how quickly they can execute such strategies.

Using the virtual execution data collected for every order placed on LMAX Exchange venues allows us to measure the effect of market volatility for a range of hold times. Because the virtual executions are triggered by actual customer orders rather than being arbitrary samples, this provides results which are more representative of times when there is interest in trading. Since the calculation of an actual cost is somewhat subjective and sensitive to the interaction with the different trading strategies, we will present a range of values using conservative assumptions.

We analysed 10,842,292 immediate execution market and limit orders placed by customers on the LMAX Exchange London venue between the beginning of September and end of December 2016. The cost of hold time was calculated by comparing the difference in notional value between the virtual execution at actual execution time and the virtual execution at a set of hold times as shown in figure 2.

$$Cost\ Delta(D) = \sum_{i=0}^{N} QB_i \left(P_{t=D} - P_{t=0} \right) + \sum_{i=0}^{N} QS_i \left(P_{t=0} - P_{t=D} \right)$$

Figure 2: Cost of hold time calculation

Where QS are sell orders, QB are buy orders, D is the delay time, P is the mean fill price at time t for an order of size QB or QS. As the limit price is not a part of the calculation we do not need to correct for skew due to off-market limit prices as we did for price improvement earlier.

Chart 12 shows the daily aggregate outcome of this analysis comparing execution cost at 0ms to execution cost at 100ms. A cost increase indicates that the 100ms execution price was worse from the trader's perspective, while a reduction indicates a better execution price. We can see that the outcome is heavily skewed towards cost increase on each day (this is consistent with the skew we see towards market order slippage over improvement, in that there is a bias in behaviour towards buying on a price rise and selling on a drop).

Given the skew towards an overall cost increase, we will use the terms Net and Gross Increase

from here on. Net increase is a more conservative measure including both cost increase and reduction in the calculation. Gross ignores any reduction, taking the view that a 100ms cost reduction observed on firm liquidity would have resulted in a fill at the quoted price for an equivalent order placed on a last look venue with a hold time of 100ms or greater.

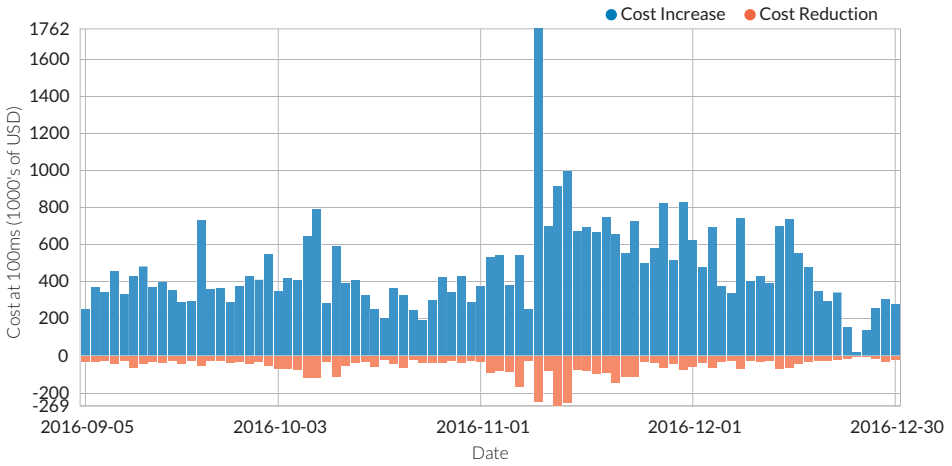

Chart 12: Cost increase vs reduction at 100ms for Q4 2016

A possible shortfall of our approach is that the price used to revalue each order may have been artificially influenced by concurrent trading activity transiently depleting liquidity resulting in a lower bid and/or higher ask and therefore exaggerating the cost increase of the execution. To gauge the potential impact of this effect we have identified orders in our sample set which are isolated by at least 100ms from any event which consumes liquidity, hence the subsequent price changes are purely in response to these orders and changes in the underlying market (8,805,137 out of the original sample of 10,842,292 orders). By removing the orders placed during times of higher concurrent activity we are implicitly selecting less active and therefore typically less price-volatile time periods, so the calculation of cost of hold time using isolated orders should be viewed as a lower bound on the range of probable values.

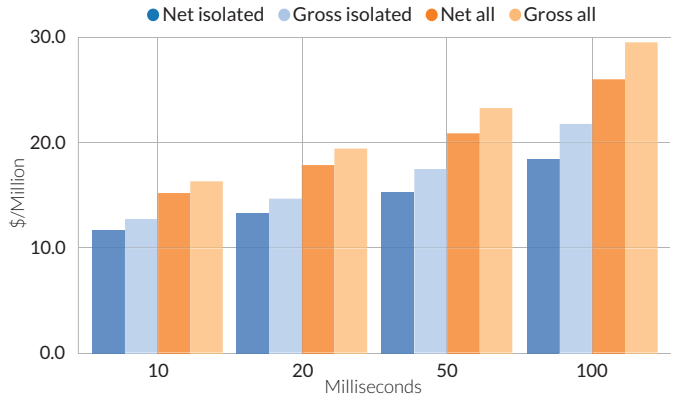

Chart 13: $/million cost of hold time - alternative methodologies

In order to create comparable results for the various approaches to measuring hold time costs (net vs gross, isolated vs concurrent) we have converted the cost into a per-million notional traded value for each of the possible combinations at 10ms, 20ms, 50ms and 100ms hold times. The results are shown in chart 13.

Part II: (v) Execution quality metrics and firm liquidity

As anticipated, when looking only at isolated orders we see a lower cost increase at any hold time. All approaches show a cost increase, net or gross, that grows steadily with increasing hold time, and there is no pronounced short term (10/20ms) 'reaction and correction' that might be expected if the order being measured was artificially altering prices by depleting liquidity which was replenished before the later (50/100ms) calculations were performed.

The true costs of hold time can only be estimated given a detailed understanding of a trader's order placement behaviour and how it is affected by these delays. The most intuitive impact of hold time is the opportunity cost of a reject. Taking the 100ms hold time, where estimates of cost range from $17.80-$28.60/million and applying a simple average we might assume that a trader receiving a 5% reject rate and 100ms hold time is experiencing a cost between $0.89 and $1.43/million in aggregate over their total trading volume. The real experience is likely to be significantly worse. The costs identified above show the outcome of a calculation for every order placed, but we expect rejections to be higher during times of higher market volatility. A superficially satisfactory fill ratio may mask a small number of extremely expensive rejections.

We can attempt to gain some insight into how much worse than the average these events might be by restricting our data set to unfilled immediate execution limit orders. In this case, we are no longer interested in the cost increase after execution time as at this point the trader has already missed their desired price level. Instead we will compare the order limit price to the price available when the order was received.

Using only the unfilled orders and cost relative to the order limit price, our prior measurement of $17.80-$28.60/million increases to a cost of hold time at 100ms in the range of $297-$336/million. We should approach this value with a healthy degree of scepticism as it relies on the order limit prices being realistic, however the value is within the range of the extremes of price swings observed over a 5ms time period during a major market event. Although the sample size is much smaller we can also use unfilled orders for the trading account used by the TPA as another data point – this returns values between $83-$86/million.

A final consideration is that calculating the cost of hold time per order only tells part of the story and that rejections may not be single events. Traders receiving rejections will often place multiple orders in order to secure a trade, hence the headline per-order hold time figures may well mask a much more extended opportunity cost from the trader's perspective. A proportion of the orders resubmitted after rejection may themselves be rejected creating a 'reject' chain. In the TPA data 11.2% of all rejected orders which eventually fill are rejected more than once. If those reject chains were clustered around significant changes in the market then the effects of hold time would be amplified beyond the conservative averages shown above. Unfortunately time precluded a detailed investigation of these effects, but this is an interesting avenue for future work.

We have picked 100ms as a representative hold time based on the TPA data, and we are open to the criticism that the market has moved on since 2016. That said, cost of hold time is not linear. **Approximately 60% of the 100ms hold time costs are incurred in the first 10ms, and for customers sensitive to market impact any hold time which means the market reacts before their trade executes may represent a significant cost**.

Box 12

Quantifying cost of hold time

Hold time represents one of the most significant hidden costs of trading:

- The impact of hold time for successful orders will depend on the individual trader's trading strategy;
- Approximately 60% of the 100ms hold time costs are incurred in the first 10ms; $15/million is the estimated net increase in cost of execution at 10ms;
- For customers sensitive to market impact any hold time during their order execution represents a significant cost;
- The true costs of hold time can only be estimated given a detailed understanding of a trader's order placement behaviour and how it is affected by these delays;
- The most intuitive impact of hold time is the opportunity cost of a reject and depending on the methodology used can range from $17.80-$28.60/million to $297-$336/million at 100ms. Given such a broad range, we conservatively settled on $25/million, as an estimated cost of last look at 100ms;
- The effects of hold time can be further amplified if the reject chains are clustered around significant market events, thus creating information leakage about the customer's trading strategy.

Part II: Execution quality metrics and firm liquidity

Part II: Summary of findings:

In this part we set out with some leading questions about flow characteristics, volatility and order placement strategies as a way to explore the unexpectedly low limit order fill ratio for firm liquidity calculated from the basic metric analysis of our TPA trade database.

We have identified the following key considerations when using limit orders:

- On firm liquidity price improvement and fill ratio are linked and are both related to price volatility;
- The level of the price volatility or jitter on firm venues is higher than on last look venues. Critically, price lifetimes can be much shorter than network latencies or the intervals between a customer receiving market data updates;
- Price improvement on firm liquidity can be traded for increased fill ratio at the trader's discretion as the two are linked. Understanding the market dynamics on firm venues can help determine ideal price discretion to give the best balance between fill ratio and improvement.

For firm liquidity, price variation and fill ratio are driven purely by market dynamics. For last look liquidity, this is not usually the case - the lack of price improvement indicates that there are additional commercial factors at play.

Slippage and fill ratio remain useful separate metrics for last look venues, but on firm liquidity venues these two metrics don't measure any business process - they measure market volatility. Comparing fill ratios between last look and firm is therefore not a useful exercise in comparing like with like. Price variation/improvement must also be included for a complete picture.

We have investigated multiple options for calculating the cost of hold time, and we have briefly touched on market impact to address the question of why the fill ratios observed were so much higher with last look. By examining the execution strategy and the associated market impact of the TPA we can explain many of the observed features or anomalies of the TPA data set.

Last look fill ratios appear competitive for the low market impact trading style of the TPA, making it an ideal data set to assess the suitability and completeness of the commonly used TCA metrics when applied to both last look and firm liquidity. However trading with greater market impact may not be treated so kindly by last look venues, and predicting the execution quality in the context of an arbitrary maze of commercially set hold times and reject rates makes the open deterministic nature of a firm liquidity central limit order book a far more predictable and appealing proposition for the trader:

- Fill ratio is not directly comparable between last look and firm liquidity as different processes are measured;
- Price improvement is important and should be included in the base metrics set along side slippage as combined price variation.

Part II: Summary of findings cont'd:

- The average price improvement achieved on LMAX Exchange in Q4 2016 by the TPA and our three customer profiles was:
  - TPA: $3.17/million
  - Aggregator Profile: $2.25/million
  - Broker Profile: $47.71/million
  - Institutional Profile: $3.44/million
- The cost of last look hold time for a rejected order was estimated to be $25/million at 100ms (across all orders, netting both cost increases and decreases from delayed execution), with 60% of costs incurred in the first 10ms.
- Market impact is not usually measured, but is crucial to consider when interpreting execution quality.

Box 13

Understanding firm liquidity

The analysis of flow characteristics, volatility and order placement strategies demonstrates that:

- For firm liquidity, both fill ratio and price variation on limit orders are linked and related to price volatility, thus purely driven by market dynamics;
- Last look liquidity is not driven by market dynamics, and the lack of price improvement indicates that there are additional commercial factors in play;
- Fill ratio and slippage remain useful separate metrics for last look venues, but on firm liquidity venues these two metrics don't measure any business process - they measure market volatility;
- Comparing fill ratios between last look and firm venues is not a useful exercise in comparing like with like - price variation must also be included for a complete picture;
- Market impact, not usually measured, is a crucial for understanding execution quality across both liquidity types.

blank page

Part III

Comparative transaction cost analysis

Part III: Comparative transaction cost analysis

In Parts I & II of this paper we have characterised the two main liquidity options available to the FX trader:

- Last look liquidity has been shown to be more straightforward to trade when using limit orders, but places control firmly in the hands of the liquidity provider. This restricts the choices available to the trader when executing in challenging market conditions;
- Firm liquidity, represented by the LMAX Exchange central limit order book, exposes the underlying market directly to the trader, giving them the challenge of dealing with more dynamic pricing but offering transparent execution and price improvement without pre-trade information leakage.

In this section, we will present a theoretical comparison of the costs of using trading styles optimised for last look and firm liquidity to achieve the same target fill rate. We will show results for a number of customer profiles using the execution quality data captured for clients trading on LMAX Exchange liquidity in Q4 2016. We have used the aggregate results of actual customer trading as the baseline for each customer profile.

Starting from this baseline we have calculated:

- The unfilled orders which could have been filled at the original time of execution by adding varying levels of price discretion at 0.1 pip intervals from 0 to 0.5 pips;
- The cost of filling these orders by adding price discretion on LMAX Exchange;
- The cost of filling the same orders by refiring the order (or unfilled portion thereof) on a theoretical last look liquidity source 100ms after the original execution time. The last look comparison assumes that no price improvement is given, in keeping with our observations detailed earlier, and is calculated based on actual market prices available 100ms after the order was originally accepted.

The different customer profiles are composed as detailed earlier:

- **Institutional**: All Institutional clients placing limit orders within the daily volatility band
- **Aggregator**: Cross connected clients with trading behaviour similar to a 'price-sniping' aggregator, exhibiting no price discretion, high fill ratios and low price improvement
- **Broker**: All Broker clients placing limit orders within the daily volatility band, see section 'Quantifying the value of price improvement' (p. 51)

The comparative results are calculated as a $/million impact on trading costs, where a positive value indicates a net profit (or reduction in costs) and a negative value indicates a net loss (or increase in costs) for the client. These values have been calculated using the Aggregator profile at 0.1 pip discretion as an example:

- The baseline is a traded volume of $44.62bn and a fill ratio of 92.37%;
- On LMAX Exchange, these clients achieved price improvement of $100,233, equivalent to $2.25/million;
- By applying 0.1 pip discretion the same clients would have traded an additional $1.94bn, increasing the volume traded to $46.56bn and fill ratio to 96.69%;
- The cost to the client of adding 0.1 pip discretion on these additional orders is -$15,944;

- The price improvement, net of the cost of 0.1pip discretion is $100,233 - $15,944 = $84,289, equivalent to $1.81/million on the total volume of $46.56bn;
- The cost of filling the same orders using a retry at 100ms was -$33,378, equivalent to -$0.72/million on the total volume of $46.56bn.

Therefore for this scenario we estimate that customers can achieve a fill ratio of 96.69% with a net $1.81/million impact (i.e. a reduction in costs compared to trading at limit price) using discretion on firm liquidity vs a -$0.72/million impact (net increase in costs) using retries at 100ms. Expressed as a comparison, the customers are better off by $2.53/million using firm liquidity.

The fill ratios and $/million values for all customer profiles and trading scenarios are summarised in table 23.

Profile	Scenario	Fill ratio	$/million impact on trading costs		
			LMAX Exchange	Last look retry	Comparison
Aggregator	Baseline	92.37%	2.25	0.00	-2.25
	0.1 pip	96.69%	1.81	-0.72	-2.53
	0.2 pip	98.17%	1.55	-1.09	-2.64
	0.3 pip	98.91%	1.36	-1.34	-2.70
	0.4 pip	99.27%	1.24	-1.46	-2.70
	0.5 pip	99.52%	1.14	-1.58	-2.72
Institutional	Baseline	81.95%	3.44	0.00	-3.44
	0.1 pip	91.40%	1.83	-9.84	-11.67
	0.2 pip	95.87%	0.19	-14.34	-14.53
	0.3 pip	97.77%	-0.90	-16.49	-15.59
	0.4 pip	98.69%	-1.60	-17.71	-16.11
	0.5 pip	99.20%	-2.09	-18.50	-16.41
Broker	Baseline	89.15%	47.71	0.00	-47.71
	0.1 pip	94.82%	42.19	-6.67	-48.86
	0.2 pip	97.06%	38.30	-9.74	-48.04
	0.3 pip	98.00%	36.48	-11.15	-47.63
	0.4 pip	98.62%	35.11	-12.06	-47.17
	0.5 pip	99.06%	34.01	-12.78	-46.79

Table 23: Comparative cost of price discretion and retries

In every scenario, the absence of price improvement, combined with a net cost increase associated with trading 100ms later, means trading costs on last look are higher by between $2.25/million and $48.86/million.

The assumptions behind this comparison and the exact cohorts selected for each customer profile may be open to dispute, however, the fact remains that by withholding transparent price improvement, current last look practices restrict choice. If a trader cannot apply price discretion with the knowledge that that discretion has the possibility to be refunded – in the form of price improvement – when not required to secure a fill, then the only strategy available is to wait for a rejection and try again. All of our analysis, both the TPA data set and the broader LMAX Exchange customer base, indicates that at the times customers wish to trade, re-executing even at hold times as low as 10ms will lead to a net increase in costs.

Box 14

Retaining control over execution quality with firm liquidity

Trading on firm liquidity enables traders to:

- Protect their valuable trade information pre-execution;
- Optimise trading strategies through the ability to adjust price discretion to control fill ratios;
- Gain value from price improvement, which transparently and fairly is passed on to the customer;
- Avoid the cost of hold time;
- Firmly retain control of their own execution metrics.

Box 15

Key findings from the comparative transaction cost analysis example

The cost comparison of using trading styles optimised for last look and firm liquidity to achieve the same target fill rate demonstrate the considerable value to the trader of executing on a central limit order book. Our analysis demonstrates:

- Trading costs on last look are higher by between $2.25/million and $48.86/million due to the absence of price improvement combined with a net cost increase associated with trading after a discretionary hold time;
- By withholding transparent price improvement, last look practices restrict traders from applying price discretion as part of their trading strategy, leaving them with no choice but to wait for a rejection and try again.

TCA white paper conclusions

TCA white paper conclusions

In this paper we have analysed a third party trade database, which included both firm and last look liquidity, from the buy side perspective with three questions in mind:

- Do the commonly used TCA metrics accurately measure execution costs on both last look and firm liquidity?
- What are the metrics that measure all the underlying processes of trading firm liquidity, and thus should be used to properly assess the cost of trading in the FX marketplace?
- How does the total cost of execution compare between last look and firm liquidity?

Focusing mainly on three common TCA metrics (fill ratio, price variation and hold time), our findings demonstrate that naively applying the 'standard' execution quality metrics which have been developed for last look liquidity does not provide the full picture of execution costs. More importantly, it misses quantifiable positives of trading on firm liquidity.

In particular, the analysis demonstrates that for firm liquidity, fill ratio and price variation on limit orders are linked and both are related to price volatility, thus purely driven by market dynamics. For last look liquidity, fill ratios are subject to LP discretion and the asymmetrical application of price improvement indicates that there are additional commercial factors in play. Though fill ratio and price variation remain useful separate metrics for last look venues, on firm liquidity venues these two metrics don't measure any business process - they measure market volatility.

Comparing fill ratios between last look and firm venues is not a useful exercise in comparing like with like, and price variation must also be included for a complete picture. The cost of hold time, non-existent for firm liquidity, is an important hidden cost of trading with last look; our analysis estimated this cost at $25/million for a rejected order. Finally, market impact, not usually measured, is crucial for understanding execution quality across both liquidity types.

As a result, strategies that work well on last look venues may not translate to firm liquidity. They translate less when you mix the two. In the example of the limit order price-sniping strategy employed by the TPA, a simplistic application of the basic metrics of fill ratio, slippage and hold time would not show the value firm liquidity brings through price improvement and more consistent execution latency:

- Swapping price improvement for fill ratio is possible with firm liquidity - placing the trader directly in control of their execution costs. If a higher fill ratio is important that can be chosen over increased price improvement;
- The underlying processes of the LMAX Exchange anonymous central limit order book hold challenges and opportunities for customers used to trading with last look venues. There are higher market data update rates to contend with but faster and more consistent execution. We have shown several ways to place a value on that fast execution by calculating the cost of last look hold time, regardless of whether hold times are unilaterally applied to all orders or selectively applied and visible as extended tail latencies;

TCA white paper conclusions

- **Last look optionality is used as a defence for the LP against stale prices in the market. That discussion normally focuses on fill ratio and spread. It does not explain the observation that the market processes that naturally lead to price improvement in market orders are conspicuously absent for limit orders sent to the same venues, or the skew towards slippage and away from improvement seen with some venues. Only firm liquidity venues expose the same underlying market dynamics for market and limit orders.**

We have also introduced some new metrics to TCA, particularly market impact, or assessment of how correlated the flow is. This allowed us to understand the abnormally high fill ratios from the last look LPs seen in the TPA data in context with what institutional customers have shared with us as their view of 'normal fill ratios'.

Put simply, the trade flow of the TPA and its execution strategies create some of the worst possible conditions to showcase the value of LMAX Exchange against last look liquidity providers. Even so, the comparative TCA calculation (described in Part III) demonstrates that for each scenario, due to the absence of price improvement combined with a net cost increase associated with trading after a discretionary hold time, trading costs on last look are higher by between $2.25/million and $48.86/million. Furthermore, the transaction cost analysis showed the benefits of firm liquidity as the transparent, cost effective choice that places the trader in complete control of their execution quality, with no pre-trade information leakage.

It is hoped that this research will spark discussion across the industry on how TCA can be standardised in a way that creates clarity and promotes choice for traders. Our goal was to assist in driving the debate about TCA and contributing to the formation of a common baseline on what is important to measure, how to measure it and how to assess the cost.

Ultimately, once there is agreement on core metrics, the arithmetic that underpins TCA is straightforward. The challenge for the industry is to reach consensus on which factors should be considered and how the different metrics should be calculated. In doing so, it must properly reflect the distinctive nature of firm liquidity and correct the current imbalance of favouring strategies optimised for last look trading venues.

There are undoubtedly limitations to the research presented here. The data sets need to be widened and more detailed analysis of the five core metrics outlined (fill ratio, price variation, hold time, market impact and bid-offer spread) is required, including a deeper focus on variable trading and market scenarios, from order size to time of day, different currency pairs and the impact of major news events. LMAX Exchange will continue to publish further research in this area, drawing where possible on new data.

We encourage industry stakeholders to view this analysis as a starting point towards development of a commonly-agreed TCA framework, one which will need further research and development to refine. We welcome all feedback and critique on what is set out here.

We would like to thank our colleagues and the many customers of LMAX Exchange, forward looking liquidity providers, market makers and TCA experts who have helped formulate this analysis over the last 12 months.

blank page

Appendices and references

Appendix A: about the data

The third party aggregator (TPA) used as the principal data source in this study is co-located directly in the major financial centre for trading FX in London and cross connected to several major platforms and liquidity providers (LPs) including LMAX Exchange. Network round trip latencies between the TPA and LPs are usually less than 1ms. Given the geographical location, we are confident that when hold or execution times are discussed in this paper, the systematic network latency of the TPA to the execution venues is not a significant factor.

The TPA receives market and limit orders from clients. Both market and limit orders can specify Immediate or Cancel (IoC) or Fill or Kill (FoK) execution. Deferred execution (Good for Day or Good until Cancelled) is not supported for limit orders through the TPA.

The TPA performs some limited order management, holding the client order for a preconfigured time to live between 1 and 5000 ms while it waits for the available liquidity to meet size (i.e. fill or kill) or limit price constraints. At this point it will submit some or all of the order to the LP(s) offering the best qualifying prices for immediate execution. From the perspective of this paper, the order management function has the useful side effect of capturing the most recent market price (best bid for orders to sell, best offer for orders to buy) received from each LP whenever an order is submitted for execution. This provides us with an 'expected' price as the basis for analysing price variation, with the expectation being set by the TPA itself rather than individual clients who might display more varied latency profiles.

The TPA issues three types of order to LPs:

- Market orders
- Limit orders
- Previously quoted (PQ) orders

Previously quoted orders are particular to FX venues that are quote based – in other words last look liquidity providers. The market data provided by the LP consists of a stream of prices with attached 'Quote Ids'. These quote ids are then referenced in the order when it is sent to the LP.

Order type	Filled	Partially filled	Rejected	Total
Limit	2,370,185	32	37,188	2,407,405
Market	1,538,220	3	10,579	1,548,802
PQ	3,164,006	0	17,363	3,181,369
Total	**7,072,411**	**35**	**65,130**	**7,137,576**

Table 24: TPA order counts by type and outcome

We have eliminated all LPs receiving less than 250,000 orders over the time period of the study to ensure a statistically significant data set when calculating percentage fill ratios to 2 decimal places ('Four Nines').

In this data set only LMAX Exchange returns partial fills – usually where an order is matched against several counterparties. As the number of partial fills is very low, and as they are only returned when explicitly permitted by the customer, we have treated these as 'fills' rather than 'rejects'.

Market orders are accepted by all LPs. Orders with execution price restriction are submitted either as PQ (3 of 7 LPs) or limit (4 of 7 LPs) orders.

Currency pairs other than those priced in JPY are quoted to 5 decimal places. Under normal trading conditions prices do not vary by more than the last two decimal places and this paper follows the standard FX shorthand of 'pips' and 'ticks' – one 'pip' being the 4th decimal place and one 'tick' being the 5th decimal place. Currency pairs priced in JPY are quoted to 3 decimal places, with a pip corresponding to the second decimal place and a tick to the 3rd decimal place.

Appendix B: factors affecting internet delivery of market data

Delivering high rates of market data over the internet presents a challenge to the LMAX Exchange technology team. TCP/IP and FIX are both ill suited as protocols for the delivery of large quantities of market data over distance. The internet is an unreliable (un)co-operative network which offers no guarantees as to latency, packet loss or even being there at all [9]. Transit problems often manifest themselves in ways that are hard to diagnose and defy identification of a root cause.

Market data, in common with streaming video or voice, is a 'better never than late' data stream, where it is preferable to drop frames or reduce resolution and maintain a consistent low latency over halting the stream and attempting delivery of all data at high resolution with delays. TCP/IP however provides the latter behaviour and will attempt to deliver all data transmitted or die trying. This is the opposite of the behaviour we desire and in our case can lead to customers getting latent market data, assuming that their FIX engines are sufficiently well resourced to consume market data at the rate requested. This is not always the case – for example customers using non-dedicated public cloud systems may encounter contention for CPU or network resources from other virtual machines.

Going back to the network: TCP is designed with congestion control algorithms which aim to make it a 'good citizen' of the network and it will reduce or back off from transmission if congestion is encountered. In practice this limits the equilibrium network throughput [10], often to a fraction of the 'sticker speed' of a customer's network connection. With round trip (RTT) times of more than 100ms and the packet loss rates often encountered over longer distances we frequently see TCP unable to transport much more than top of book at a few 10s or 100s of updates per second. This can be hard to explain if the audience is not familiar with networking or versed in the finer points of the equations in the Mathis paper referenced above.

Fortunately we are able to take some steps on our side to address this.

The first step is measurement. Through our 'Streamstats' dashboard we are able to estimate a customer's maximum sustainable bandwidth. We can then match the average rate and depth of market data sent to reduce the occurrence of latent market data due to network congestion.

Within our internet facing FIX market data servers, we implement a 'coalescing ring buffer'. In the event of transient network congestion, this allows us to send the most recent market data once the network buffers have drained, discarding stale updates and resuming the stream with the most current data. We also use specialist peering and traffic engineering techniques like WAN optimisation. This uses features such as local cache tokenisation and local acknowledgements to 'trick' TCP into thinking the customer and the venue are actually much closer than they really are allowing a higher equilibrium bandwidth than it would normally settle on.

The above mitigate the impacts of transiting the internet by manipulating TCP/IP and reducing the bandwidth requirements on average and dynamically to provide a responsive market data stream. However even with all of the workarounds our team can devise, the Internet can still have a bad day.

The best solutions remain the simple ones of either avoiding the Internet completely via private transit or through co-location within the same data centre or metro area. We realise that this may not suit all customers however. It is also true that even when co-located, FIX over TCP/IP also has throughput problems when faced with ultra low latency interconnections where the RTT is much lower than the internal TCP timers. LMAX Exchange, in common with many other venues, offers a UDP based ITCH protocol which allows for true low latency streaming prices without the overhead of the binary to ASCII and back again translations of FIX and free from interruption by a complex protocol like TCP.

TCP/IP (and FIX) are really the wrong choices for delivering market data over the Internet. This is not 'where we would start'. However it is what we have available and while we remain interested in newer developments – for example Google's suggestions for 'short fat networks' [11] we have to be conscious of the need to maintain compatibility with a wide range of network stacks deployed by our customers and the requirement for us to serve any customer, no matter their location, technical ability or the varied 'problems' the Internet will continue to throw at us along the way.

Appendix C: volatility bands used for price improvement

Volatility band (2**σ**) ranges by FX instrument for 2016 from section 'Quantifying the value of price improvement' in Part II (p. 51). The daily volatility bands referred to in that section are available upon request.

Symbol	Volatility band		
	min	mean	max
AUDCAD	0.000238	0.000469	0.001514
AUDCHF	0.000227	0.000456	0.001626
AUDJPY	0.024353	0.061606	0.297316
AUDNZD	0.000229	0.000468	0.001168
AUDUSD	0.000207	0.000435	0.001529
CADCHF	0.000202	0.000407	0.001428
CADJPY	0.023881	0.059334	0.259619
CHFJPY	0.027308	0.060352	0.359139
EURAUD	0.000415	0.000884	0.003038
EURCAD	0.000380	0.000776	0.002784
EURCHF	0.000161	0.000298	0.001928
EURDKK	0.000149	0.000151	0.000154
EURGBP	0.000227	0.000428	0.002943
EURHKD	0.001655	0.003558	0.018066
EURHUF	0.054194	0.100152	0.162726
EURJPY	0.029353	0.064938	0.452184
EURMXN	0.006906	0.016609	0.122569
EURNOK	0.002204	0.004379	0.015539
EURNZD	0.000449	0.000990	0.002871
EURPLN	0.000895	0.001893	0.009932
EURSEK	0.001838	0.003314	0.014203
EURSGD	0.000352	0.000653	0.002549
EURTRY	0.000993	0.002140	0.008985
EURUSD	0.000197	0.000451	0.002330
EURZAR	0.007722	0.016577	0.064136
GBPAUD	0.000574	0.001170	0.007194
GBPCAD	0.000535	0.001033	0.007457
GBPCHF	0.000376	0.000756	0.006054
GBPCZK	0.009771	0.019957	0.134765
GBPDKK	0.002463	0.004812	0.033750
GBPHKD	0.002727	0.005667	0.048932
GBPHUF	0.136669	0.237968	1.526575

Symbol	Volatility band		
	min	mean	max
GBPJPY	0.048939	0.111364	0.915977
GBPMXN	0.010347	0.020702	0.130111
GBPNOK	0.004010	0.007211	0.044407
GBPNZD	0.000619	0.001334	0.007290
GBPPLN	0.002050	0.003421	0.021961
GBPSEK	0.004255	0.006907	0.045215
GBPSGD	0.000633	0.001024	0.007633
GBPTRY	0.001550	0.002866	0.016612
GBPUSD	0.000338	0.000726	0.006354
GBPZAR	0.011141	0.020826	0.099886
NOKSEK	0.000309	0.000491	0.001775
NZDCAD	0.000281	0.000515	0.001504
NZDCHF	0.000244	0.000445	0.001534
NZDJPY	0.025621	0.057843	0.267923
NZDSGD	0.000370	0.000561	0.001583
NZDUSD	0.000216	0.000437	0.001389
USDCAD	0.000301	0.000615	0.001725
USDCHF	0.000194	0.000409	0.001707
USDCNH	0.000560	0.001243	0.003621
USDCZK	0.005575	0.010601	0.053750
USDDKK	0.001332	0.002771	0.014045
USDHKD	0.000238	0.001144	0.001681
USDHUF	0.075297	0.146655	0.758817
USDJPY	0.026343	0.059517	0.247971
USDMXN	0.006529	0.014539	0.097505
USDNOK	0.002923	0.004904	0.023934
USDPLN	0.001027	0.002160	0.012728
USDSEK	0.002565	0.004327	0.024124
USDSGD	0.000222	0.000447	0.001318
USDTRY	0.000893	0.001889	0.006703
USDZAR	0.009374	0.016013	0.060790

References

[1] Foreign exchange markets with last look. Cartea et al. 2015
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2630662
< back-link

[2] The true cost of rejects in $PNL
https://www.linkedin.com/pulse/true-cost-rejects-pnl-matt-clarke?trk=prof-post
< back-link

[3] Do consumers prefer round prices? Lynn et al. Jnl of Economic Psychology 36 p96 2013.
http://www.sciencedirect.com/science/article/pii/S016748701300024X
< back-link

[4] Trading firm XTX challenges 'last look' in FX
https://www.fnlondon.com/articles/trading-firm-xtx-challenges-last-look-in-fx-20170131
< back-link

[5] Market liquidity: Theory, evidence and policy. Foucault, Pagano, Roell 2013 p 38
https://global.oup.com/academic/product/market-liquidity-9780199936243
< back-link

[6] EBS Live Ultra to provide data at 5ms intervals
http://www.ebs.com/news-and-events/2017/20170201_ebs_live_ultra_5ms_feed.aspx
< back-link

[7] Algorithmic and high frequency trading. Cartea, Jaimungal and Penalva 2015 4.3.2 p 76
http://www.cambridge.org/gb/academic/subjects/mathematics/mathematical-finance/algorithmic-and-high-frequency-trading
< back-link

[8] Technical analysis of the financial markets. Murphy. 1999 p. 209
http://www.penguinrandomhouse.com/books/350647/technical-analysis-of-the-financial-markets-by-john-j-murphy/9780735200661/
< back-link

[9] ISP Level 3 goes TITSUP after giganto traffic routing blunder
https://www.theregister.co.uk/2015/06/12/level_3_down_after_routing_through_malaysia_like_idiots/
< back-link

[10] The macroscopic behaviour of the TCP congestion control algorithm
http://ccr.sigcomm.org/archive/1997/jul97/ccr-9707-mathis.pdf
< back-link

[11] TCP options for low latency: Maximum ACK delay and microsecond timestamps
https://www.ietf.org/proceedings/97/slides/slides-97-tcpm-tcp-options-for-low-latency-00.pdf
< back-link

blank page

Notes

Contact

A unique vision for global FX

For specific feedback directly addressed to the authors, please email: TCAfeedback@lmax.com

For more information on LMAX Exchange:

Institutional clients
Telephone:
+44 20 3192 2682
Email:
institutionalsales@lmax.com

24-hour helpdesk
Telephone:
+44 20 3192 2555
Sun 22.00 - Fri 22.00 UK time

General enquiries
Telephone:
+44 20 3192 2500
Email:
info@lmax.com
Fax:
+44 20 3192 2572

blank page

LMAX Exchange: **TCA white paper V1.0** - May 2017

TCA and *fair* execution. The metrics that the FX industry must use.

An analysis and comparison of common FX execution quality metrics between 'last look' vs firm liquidity *and* its financial consequences.

speed › price › transparency

LMAX EXCHANGE

+44 20 3192 2555 | info@lmax.com | www.LMAX.com