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

TCA metric #3

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

(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.

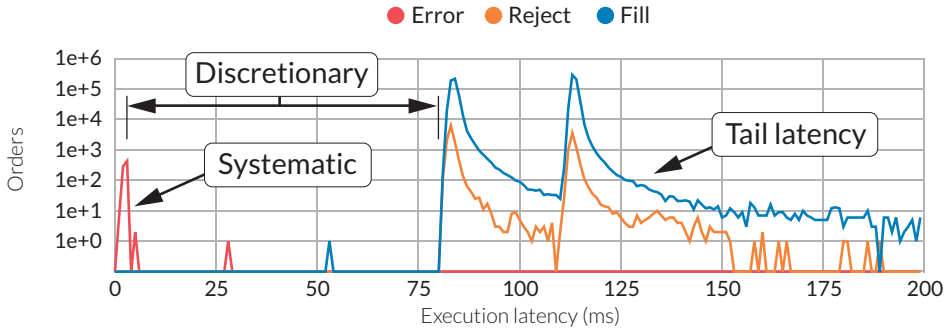

Chart 3: Detailed execution times for Bank 3

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

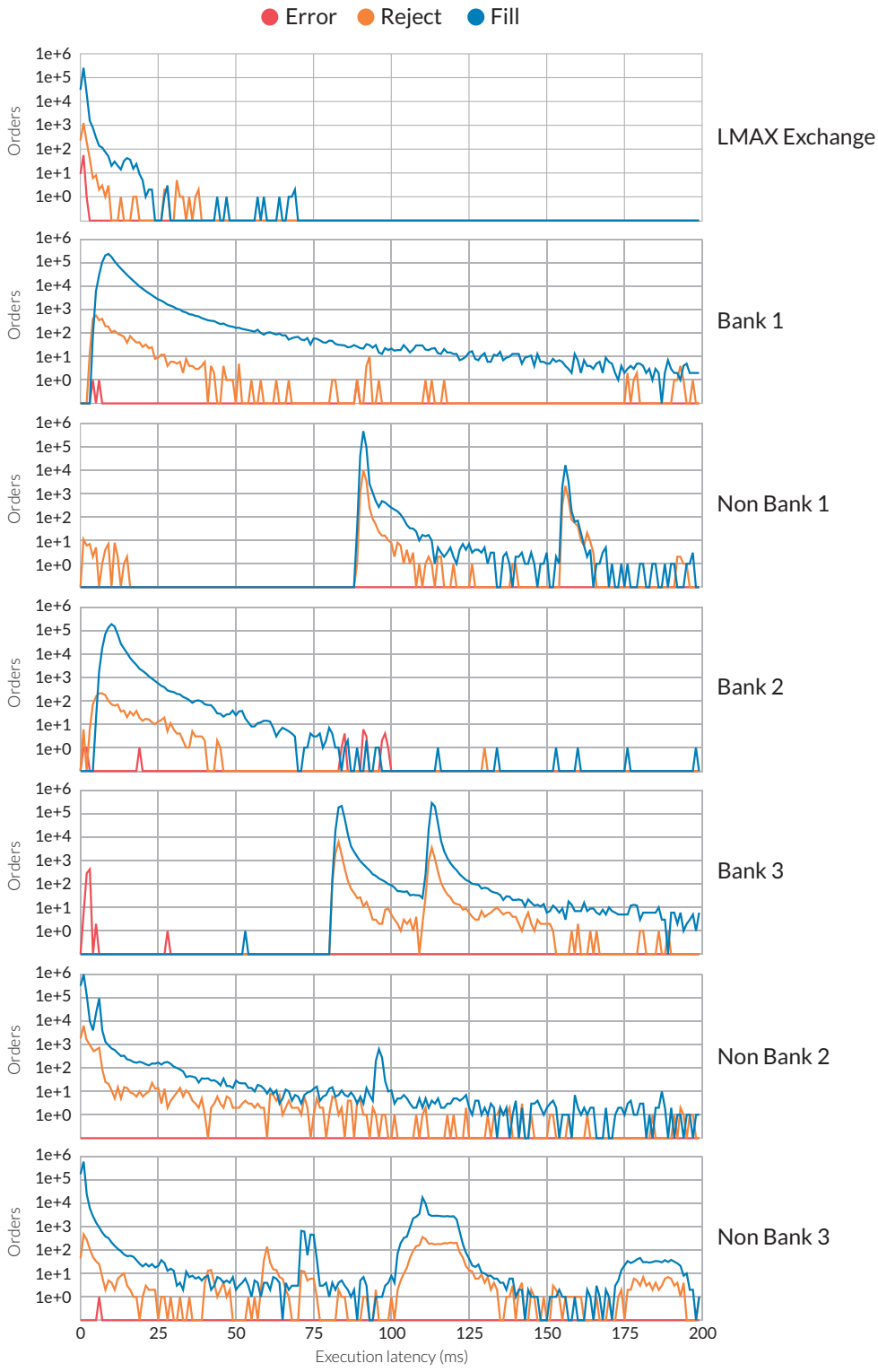

Chart 4: Execution times by venue for 2016

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.

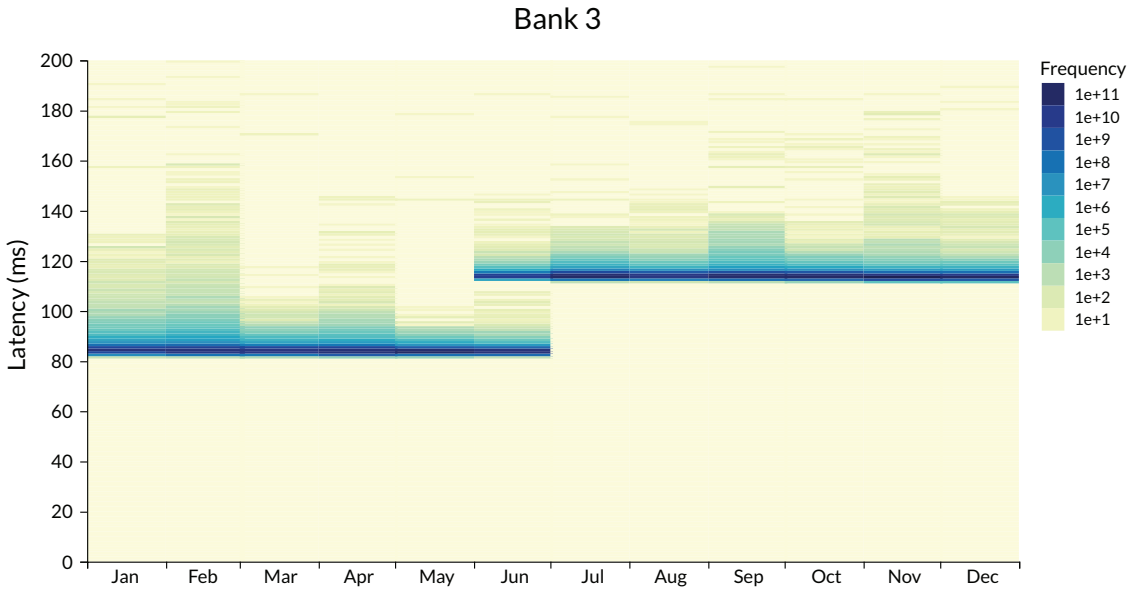

Chart 5: Execution latency by month for Bank 3

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.

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

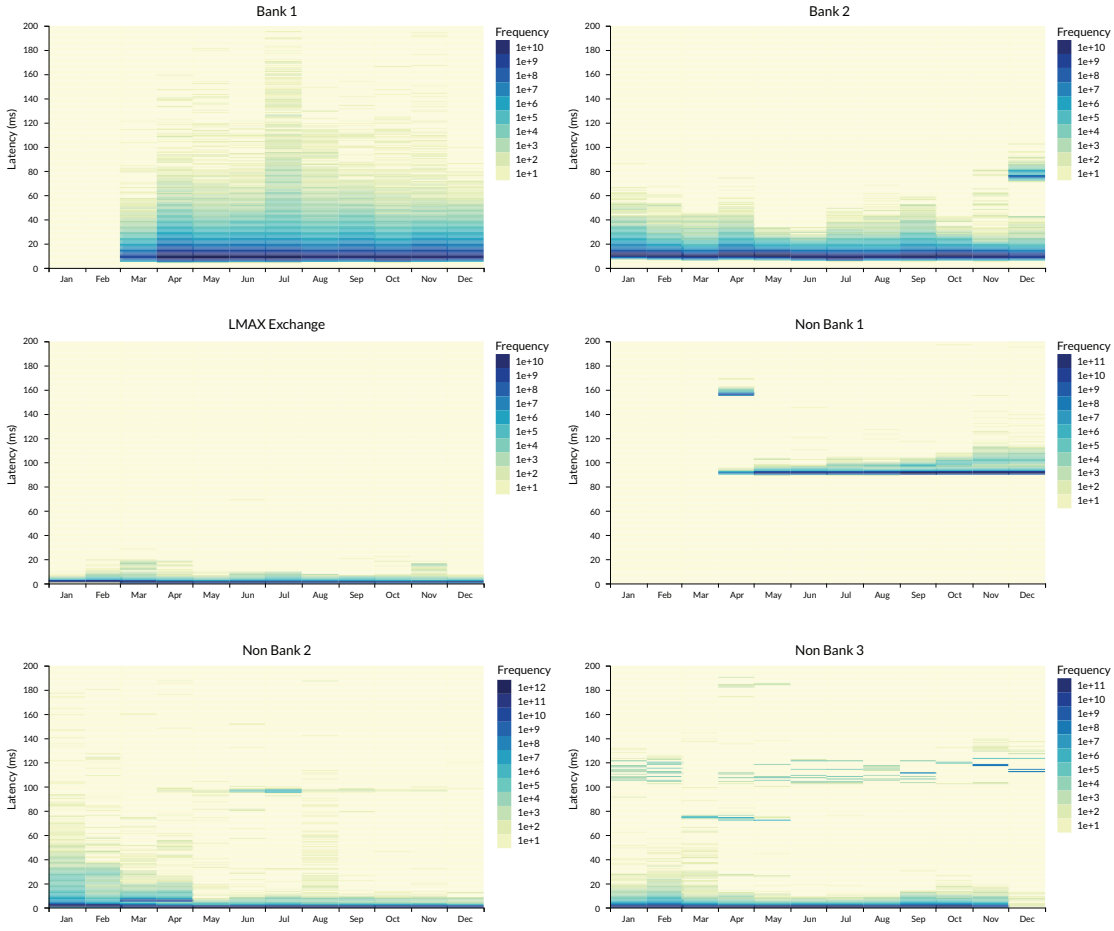

Chart 6: Execution latency by month heatmap

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.

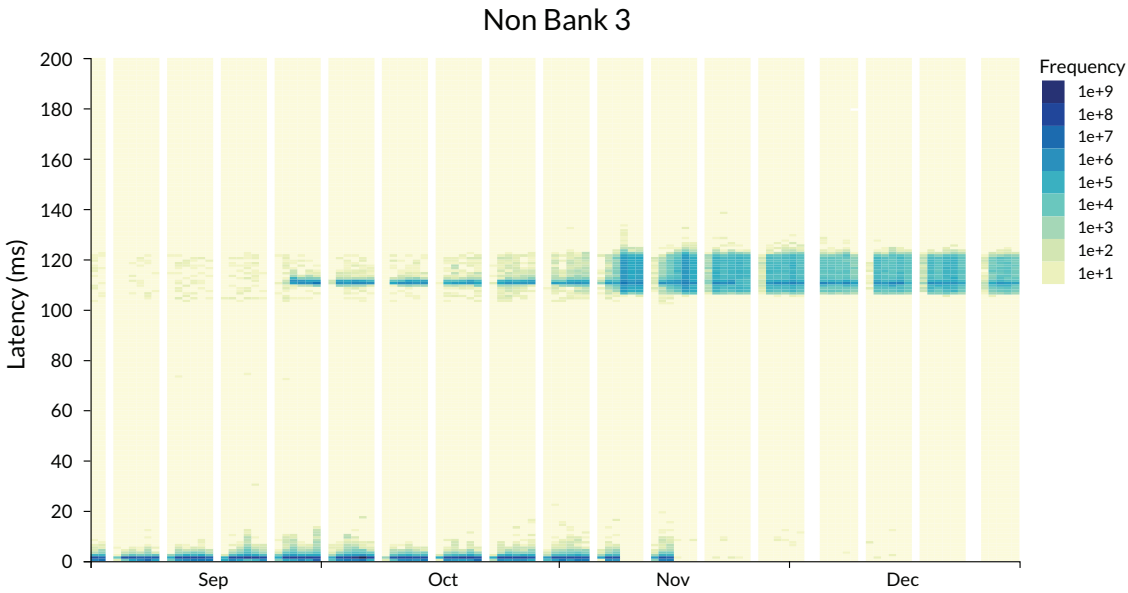

Chart 7: Execution latency by day for Non Bank 3

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)

Contact

speed › price › transparency

LMAX™ EXCHANGE

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