FOCUS

Data Latency – Why Speed is of the Essence

by Sean O'Donnell, RealStream Product Director at Cognotec

The increased move by the Forex markets to a real-time trading environment means that there is a new urgency to address the risks associated with data latency amongst the bank participants. Fortunately, new technical solutions exist enabling the problem to be effectively addressed.

Data latency has become a critical issue for the banking industry. The evolution of electronic trading is seeing Executable Streaming Rates increasingly replace Request For Quote (RFQ) transaction models. Executable Streaming Rates (ESR) enable clients to deal with a single click on live rates which are updated in real-time, rather than requesting to deal on a rate which was made 'good' for a period of time and may be old.

The shift is from a trading environment in which rate data reflects an historical position (albeit one which is just one or two seconds old), to one where rate data reflects current positions.

For market-making banks serious about remaining at the forefront in the provision of forex trading services dealing in an ESR marketplace means it is imperative that they have the ability to provide true executable
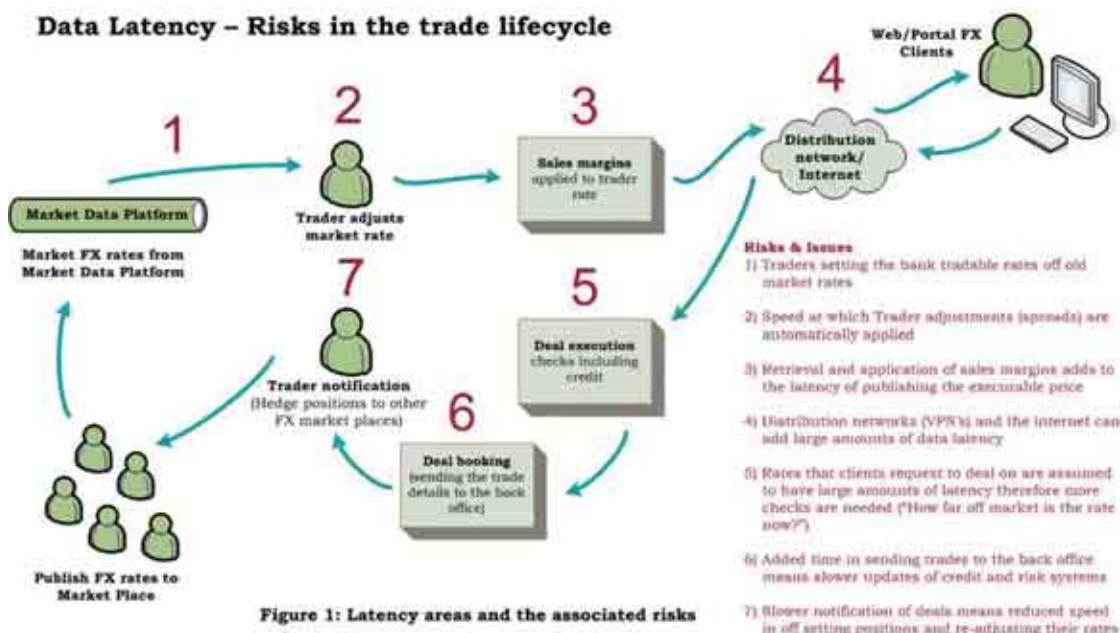
Data Latency – Risks in the trade lifecycle

Figure 1: Latency areas and the associated risks

streaming prices, i.e. ones designed around a push-pricing model, rather than using an RFQ designed pricing engine to simply replicate prices more quickly thereby simulating streaming prices which on execution may not be dealable.

Reducing latency between price publication and execution enables the market making bank to reflect an accurate picture of its trading position. As a result of this the trading desk is able to publish prices that more accurately reflect its market position, enabling it to maintain its profit margins. On the buy side, the client sees quality pricing and is able to deal on the rate that they see. In this more transparent market, where less defensive spreads are required to protect the bank against 'off market' prices, clients are more likely to get the best price for the volumes they want to deal on.

ESR trading

This new environment offers many positive opportunities for both banks and end-users, however it is extremely fast moving, and increased price transparency has lead to a narrowing of spreads as banks find they need to be increasingly competitive in order to get the deal. In addition, there is now a multiplicity of trading channels (proprietary online channels, multi-bank portals, ecn's and black box API's etc) through which the client can access their bank – and prices from competitors.

"clients are more likely to get the best price for the volumes they want to deal on"

Bank risk management in this environment must be more responsive to the needs of the trading desk. Latency, and the speed or slowness with which information is transported and updated, is key to many of these risks. For instance, arbitrageurs are more likely to be using algorithmic models which respond to price anomalies much more quickly than a human being can intervene to adjust prices.

In addition, narrower spreads increase the cost to the bank of prices being 'off market' – a risk which is exacerbated if the bank is dependent on an RFQ-based pricing model. Finally, instantaneous price updates can only aid transparency if they are accurate – faulty prices can engender a sense of false security on the trading desk, and the inherent risk of taking a particular position is increased as a result. Finally, there is the business risk that, within a multi-bank trading environment, deals can be lost to banks quoting more competitive prices more quickly.

The trading cycle

Linked to the above risks there are further outcomes associated with time lags which affect other areas of the trading cycle. Put simply, faster data delivery can only positively impact on shortening transaction times if other areas of the trading cycle are modernised to allow it.

For instance, credit checks can slow down deal execution if a check is to be made through to the bank's back office systems.

Solutions now exist which bring credit checks closer to the user-level which in turn speeds acceptance on deals and avoids post-trade errors. In addition by automatically reducing the deal volume for a time period post execution, selectively by currency pair, period, channel and client, the bank can have a greater level of control. Real-time notification and 'hit protection' help the trading desk to increase their deal flow in a controlled manner while maintaining a true picture of its position.

The solution

As stated above, a true 'push-pricing' engine in which prices are refreshed instantaneously rather than as a result of a client request is the foundation for an effective technical solution to the problem of data latency. With push-pricing, the possibility that published prices are dead and no longer tradable is vastly reduced. For this reason pricing engines that solely work on an RFQ basis are no longer adequate in the current trading environment.

Secondly, to reiterate the points about the impact of latency on the trading cycle made above, moving transaction processes such as credit checking and notification closer to the client level and out of the back office enables overall transaction times to be reduced and the whole trading cycle to become more efficient.

In addition, the number of end-users has effectively grown exponentially because existing clients may be accessing the trading desk via more than one trading channel. Therefore a solution which is easily scalable and can be quickly pushed out to hundreds or even thousands of end users is required.

"The forex trading environment is changing and banks need to invest in technical tools"

Further, the ability to tailor prices according to individual clients, groups of clients and how that client is trading (via proprietary channels or multi-bank portal for instance) ensures that end-users will receive the most competitive prices wherever and whenever they want to trade.

The forex trading environment is changing and banks need to invest in technical tools which are specifically designed to meet the demands of near-real-time trading. In order to build an effective business model around delivering executable streaming rates – and meet client demands for prices which display the tightest margins and are guaranteed to be tradable - banks have to be able to distribute real-time customisable prices across a multiplicity of trading channels. More importantly, in terms of their own risk management, banks have to be able to rely on these prices and their efficient delivery in order to retain a true picture of the bank's position in the market and eradicate the need for defensive pricing.

These are essential requirements to the competitiveness of the forex trading desk.

Business is already being lost by some banks to competitors who have made this investment, and as comparative prices are increasingly made available via the use of portals and multiple screens the pace of competition is likely to increase. However, efficiency of delivery and execution has to be matched by effective risk-management tools if banks are to compete profitably in a more transparent trading environment.

We go faster you go further

www.standardchartered.com

At Standard Chartered we believe in the spirit of achievement. As an established leader in FX, we actively trade in more than 100 currencies with particular strength in exotics. We also offer a range of risk management options to match your needs, from fixed income funds to interest rate derivatives. We set high goals for ourselves, and our ultimate goal is your financial success.

Standard Chartered

Currenex:
Tackling the Latency Issue

*e-Forex asks **Sean Gilman**, Chief Technology Officer at Currenex to tell us how they are tackling the problems of latency.*

Sean Gilman

Introduction

With a continuing investment in developing enhanced technology, Currenex – a leading multibank FX trading platform – has focused on what is one of the most vital aspects of today's online trading: SPEED. In migrating to a Linux open source environment they have established a competitive advantage in what they believe is a crucial facet of future success.

Currenex seeks to maintain a technological edge amongst FX trading platforms, knowing full well that such a determination could be the difference between feast and famine in the increasingly competitive FX arena. Currently, the growing adoption of electronic trading in FX is following the same course set by equities, derivatives and fixed income. However, the fact is much of what passes for "electronic trading" in FX is relatively unsophisticated and unsustainable. Other market experiences demonstrate that the quality of technology becomes the key determinant of which trading platforms flourish and which collapse during the period of inevitable consolidation. FX will be no different. With that in mind, Currenex's technology focus has been on reducing latency to an absolute minimum. An important aspect of Currenex's quest for limiting latency has been the migration to a faster Linux operating system.

The Importance of Speed

To understand the motivation behind the migration is to understand the overall importance of speed as it pertains to the challenges of execution in FX. RFQ (Request For Quote) and RFS (Request for Streams) workflow emulates trading over the phone, and there was less need to focus on latency when this was the dominant mode of trading FX. With the emergence of ESP (Executable Streaming Pricing), e-FX has evolved due to the fact that this mode of trading is decidedly faster than RFQ. The ESP model is extremely time sensitive.

For streaming rates, the two forces driving FX platform success are access to liquidity and speed (low latency). A customer will look for the destination where there is the best chance of getting executed; in other words the ECN with the deepest and tightest order book. However, speed is an equally important decision factor when choosing where to route an order. In particular, low latency technologies provide market participants with the most efficient executions (i.e. lower cost of execution, less slippage, etc.). In contrast, high latency systems cause poor execution, slippage, and allow for the "picking off" of market makers (i.e. banks). As a result, market participants – whether they are banks or buy side firms – will migrate to platforms with good initial liquidity and low latency. Put simply, volume will go to the fastest.

For example, a trader might wish to execute a fill or kill order. Due to latency, in a fast moving market with multiple ECNs available, the trader doesn't actually know the exact liquidity available at a given instant at each destination. In reality, the trader only has an approximation of the available liquidity based on market data. The accuracy of the approximation is inversely proportional to the time it takes to receive the market data plus the time it takes to submit an order to the ECN. Therefore the trader needs to make a strategic decision about which ECN should receive the order first. By reducing the "opportunity cost" (i.e. time) that a trader takes in routing an order, the destination with the lowest latency has an edge. It is this competitive advantage that Currenex has created.

Faster Boxes

In 2003 Currenex entered into a strategic relationship with Hewlett Packard to design and develop the optimal systems environment for low latency FX trading. As a result of this joint effort, Currenex has been able to reduce the execution latency to levels of less than 10ms, and this effort continues with the goal of reducing execution time to the microsecond (sub-millisecond) level.

Continuing the partnership with HP, Currenex has migrated its system to a Linux environment, running on Intel Xeon and AMD Opteron processors, to achieve greater speed. Before making the decision to move to Linux, Currenex built a test suite to measure the benefits. "We built tools to measure each step of the order execution process and have used that information to locate and streamline processing bottlenecks," says Currenex chief technology officer Sean Gilman. "When we were done, we saw a 75% reduction in processing latency and three fold increase in maximum throughput." Looking ahead, Currenex plans to perpetuate the benefits of this process as it continues to develop. "We chose Linux because of the performance/cost curve," adds Gilman. "We plan to refresh our hardware every six months and stay on the fastest boxes money can buy.

Practical Benefits of Speed

The focus on speed allows Currenex to overcome some of the obstacles inherent to online trading. In the process of connecting banks with their customers, the Currenex platform processes approximately five thousand spot orders per second. Additionally, each order can have its own options (partial fills, stop-loss, take profit, etc.) which makes the process more complex. Further, the sorting of the order book needs to be "atomic", meaning the work can't be split between multiple processors or threads because that could create inconsistencies. This is a particularly computation intensive part of the system that Currenex has focused on accelerating with the upgrade to the faster Linux processors (see Figure 1).

The liquidity at a given ECN destination is constantly moving, causing the prices to change. If prices are updating every 50

milliseconds, any latency above 50 milliseconds can result in sub-optimal execution. If a trader sends an order to a slow ECN, the market can have already moved before the trader even knows if the trade has been executed. Even 50 milliseconds of latency can cost a customer or market maker a couple of pips in execution. "Our customers have an advantage with us because they have faster access to liquidity than their competitors," says Gilman. "Instant access to these volatile markets makes a difference as to whether a trader makes money or chases the market, therefore a key component in our superior execution performance is, and will be, limiting latency."

Better Trading

Reduced latency also means more effective blackbox model trading. With faster execution, black box traders will experience less slippage in their models. Unlike their competitors, Currenex does not throttle customer orders, meaning there is virtually no limit to the amount of orders that can be entered per second over the platform. During typical traffic the Currenex matching engine handles about 5,000 messages per second, and has experienced peaks of 11,000 per second. The system has been tested in a controlled environment at 25,000 messages per second. Because the system is so fast, order-throttling techniques employed by other platforms are not necessary on Currenex.

Throttles drive liquidity away for the simple reason that platforms employing throttles will not keep up with a volatile market. Rejected orders will clearly mean less liquidity and lost opportunity. Furthermore, Currenex provides market updates in real-time along with displays of market depth. By contrast, the leading interdealer system's market data is only updated every _ second, which is, at best, merely indicative of the market and not a reflection of prices that can be transacted.

Market makers garner benefits from reduced latency as well, as speed will also contribute to a more balanced exchange. A platform like Currenex can ensure fairness to both sides since speed of execution can help banks avoid potentially harmful arbitrage situations.

Legacy of Innovation

Currenex builds, owns, and operates its own technology and is focused on maintaining its advantage by continuing to invest in technology and foreign exchange expertise. Their association with HP and Linux has been another step in this process. Thanks to a technology advantage, Currenex has positioned itself as the best platform for model trading and the best for market making, which subsequently results in unbeatable liquidity. Sustained profitability in FX – for the buy or sell side – requires a high-performance trading platform that can capitalize on speed. Currenex has made this idea a reality.

Figure 1

Is it time to raise your game?

Andy Webb is a freelance
journalist who regularly
contributes to e-Forex.

*Considering the size and liquidity of
the market, and the number of
quantitative methods applied to it, it
is perhaps ironic that FX has in some
respects been rather imperfectly
served in terms of raw data.*

Just as with any other market, foreign
exchange is populated by individuals and
organisations looking for an edge.
Whether they are trying to predict the
direction of a currency pair, arbitrage its
inefficiencies, or capture business for
their FX e-commerce platform, all these
entities are to some extent dependent
upon market data.

This data appetite has further expanded in
recent years as interest in areas such as
high frequency trading has grown.

Yet, despite this demand, definitive real-
time and historical data sources for FX
have not always been readily available.
For example, whilst very low latency real
time information has obviously been
visible through the trading interfaces of
primary FX markets, until comparatively
recently (with the launch of feeds such as
EBS Live) it has not been readily
accessible in a format that can be used for
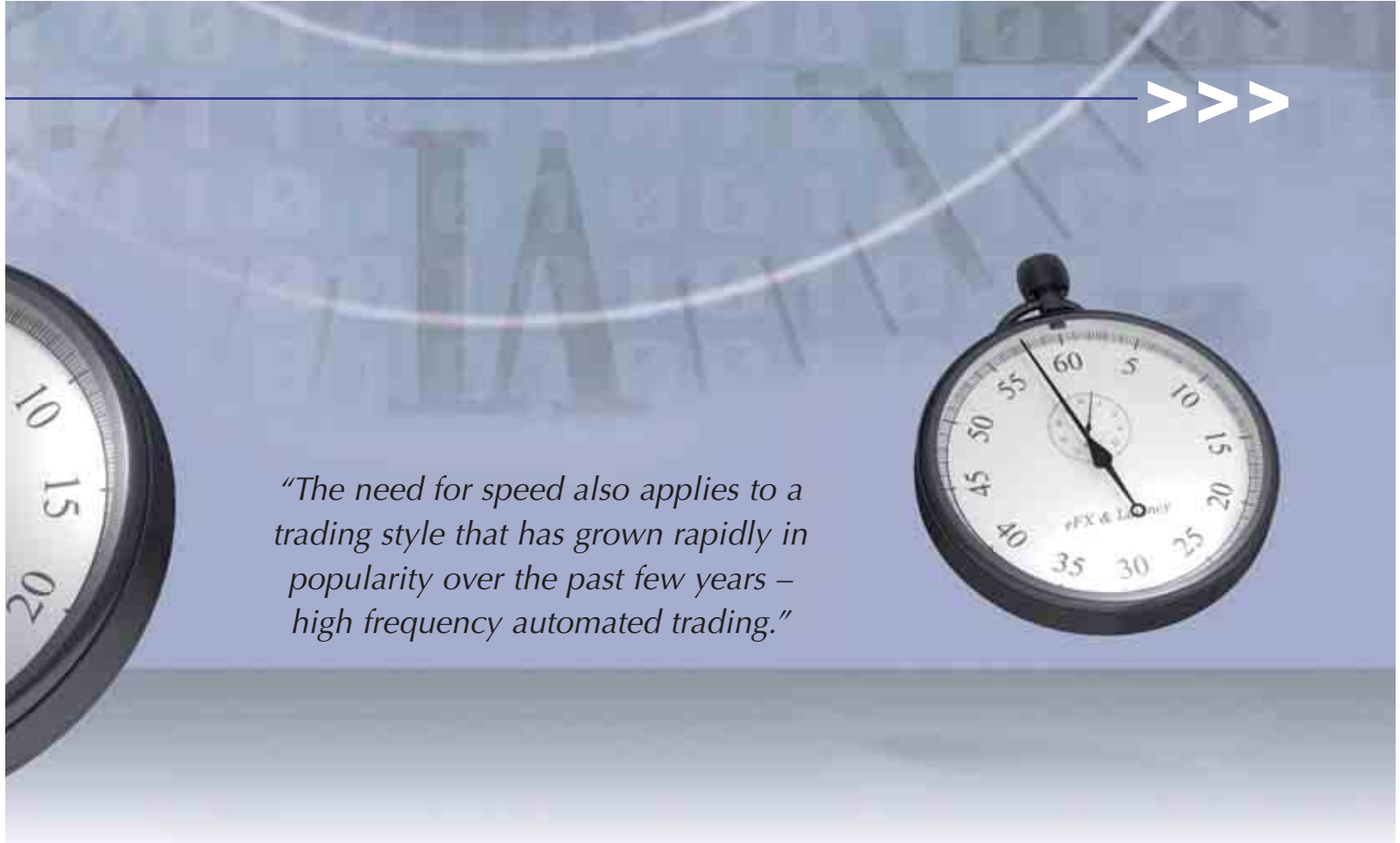other purposes, such as powering e-
commerce engines etc.

Real time speed, real time reward

A very low latency data feed provides a
significant edge in a number of highly
competitive areas, with FX e-commerce an
obvious case in point. In the future, banks
that try to skimp on the speed and
integrity of the feed powering their FX e-
commerce servers will find themselves
exposed. Furthermore, as the use of low
latency feeds and primary market access
spreads beyond the market making
community, these banks may even find
themselves increasingly vulnerable to the
buyside as well as their direct competitors.

A similar situation applies to FX arbitrage.
This has become an increasingly popular
activity among banks over the past few
years, as the risk profile of the arbitrage
business (which typically grinds out
consistent returns) looks particularly
attractive - especially when compared with
conventional proprietary trading activity.
However, FX arbitrage opportunities (even
outside the major pairs) tend to be short
lived, so those automated FX arbitrage
operations that adopt the lowest latency
price feeds will undoubtedly have the edge.

The need for speed also applies to a
trading style that has grown rapidly in
popularity over the past few years – high
frequency automated trading. Rather than
capturing arbitrage opportunities, this
attempts to capitalise on very short-term
directional moves lasting only a few
minutes, or perhaps even seconds.

Several thousand trades a day may be
involved, and in comparison with
traditional longer-term trend following
methods, the average profit per trade is
small. This makes it imperative that the
data feeding the automated trading
model has the minimum possible latency,

where any delay is measured in milliseconds rather than seconds. If not, the trading model will be delayed in making its calculations and therefore late in entering its trade orders. As a result, a small average profit per trade can be quickly and disproportionately eroded by slippage of as little as one pip per order.

These considerations have not been lost on the providers of such real time data, who have responded by reducing the latency and improving the granularity of their feeds the EBS Live feed (which is sourced from the activity on the EBS Spot system) exhibits latency of less than 200 milliseconds and is delivered in 500 milliseconds time slices.

However, speed is not the only crucial data requirement for highly automated FX systems – when it comes to real time data, cleanliness is also next to godliness. While growth in electronic trading has seen a commensurate improvement in data quality, it still remains a mission critical consideration. An undetected data discrepancy that is fed to an automated system (be it FX e-commerce, arbitrage engine, or directional trading model) can have catastrophic results.

Raising the bar internally

Although the advent of high quality, minimal latency data feeds, such as EBS Live, address the vital requirements outlined above, they may also require additional effort and investment on the part of the user. Such feeds certainly have the potential to confer a competitive edge – but only if the user's own internal networks and systems are up to scratch.

Buying a super fast data feed, but then plugging it into an internal network clogged with non-critical traffic, is an exercise in futility. By the same token, if the system consuming the data is a sub-optimal performer the real time data investment will be wasted.

In some cases, banks and other financial organisations have already fine-tuned their automated systems to the nth degree and have dedicated networks and data platforms for mission critical data. However, this isn't always the case, with smaller organisations in particular often accommodating data and generic business traffic (printing, file sharing etc) on the same network.

Finally, there is the additional problem of data creep. Like work, data appears to follow Parkinson's Law by expanding to consume all available capacity. Last year's high-speed network can therefore quickly become this year's treacle. Preventing this so that an investment in a minimal latency feed is not wasted requires vigilance in prioritising traffic, a willingness to invest in new networks/ upgraded systems, or both.

Historical and real time dislocation

The explosion in electronic trading and data capture across financial markets has drastically reduced the problem of bad ticks, or corrupt data points. (In an open outcry environment this was often due to keystroke error by the hard-pressed clerk entering traded prices from a booth on the exchange floor). As a result, there has been an increasing demand for historical data for trading model development and testing.

However, unlike many other markets, foreign exchange hasn't really seen the full benefit of this improvement. Part of the problem has been that high quality primary market FX historical transaction data has not been readily available.

As a result, development and testing of FX trading models has often been conducted on indicative or composite historical data. These index FX data sources are typically created by sampling a number of bank contributed quotes and then adjusting them with some form of algorithm in order to allow for outliers etc. Apart from the risk of outright errors, this has often resulted in products that are disproportionately expensive – especially when one considers that they do not represent a true record of actual traded prices from a primary FX market.

The problem with this approach is that the trading model often ends up being developed and tested on a different data set from the one upon which it will actually be traded. Depending upon the time frame and trading frequency, this disparity can be highly significant.

For a long term trading model that uses daily data as an input, the impact of this mismatch may not be hugely significant, as the number of bars and trades may be relatively small. However, it can be far more problematic with shorter time frames and high trade frequencies, where just one pip discrepancies on (say) five trade constant volume bars[1] can cause a large number of spurious entries or non-entries. In short, the model has been designed and proven to capture particular (and numerous) statistical anomalies in the historical data set that may be very differently distributed (or perhaps even non-existent) in the real time data of the live trading environment.

Market depth

The normal convention when designing and testing quantitative trading models with historical price data is to use a record of prices that actually traded. There is now significant and growing interest in developing models that also use historical market depth data, such as the bids and offers sitting either side of the market when a trade took place, in their calculations.

Depending on the amount of market depth information available, this effectively allows the market order book at any moment in time to be recreated. The depth of the order book either side of the market can be a vital model input for determining market sentiment – especially for short timeframe, high frequency, trading. Some form of weighting of this data is advisable (since bids/offers can be withdrawn), so the further bids/offers are from the traded price, the less weight they would typically be given in a trading model.

Historical depth of market data is also invaluable in improving the accuracy of historical simulations when larger order sizes and/or illiquid markets are involved.

A trading model might flag an order for (say) $5m at x price, but only part of that order might have been executable at that price. Obviously, this can have a dramatic impact on the accuracy of any historical simulation. If no allowance is made for this, the simulated performance of a trading model can be substantially and artificially inflated.

By contrast, if historical depth of market data is available, far more accurate simulation of order fills becomes possible. Even if the entire order could not be filled at the price indicated by the model, reasonable assumptions about partial fills at adjacent prices can be made. This is particularly important when an automated trading model also incorporates a mechanism for finessing orders (algorithmic trading). Armed with an indication of the depth of the market, the process of developing the finessing algorithm becomes a far more straightforward and precise process.

Inevitably, the possibilities offered by historical market depth data have not been lost on the trading community and,

while it has been available for some time for equity markets, data providers for other markets are now also beginning to respond to this interest.

The flow jigsaw

Banks have been providing their clients with FX flow information for some time. This typically consists of aggregate data on currency transactions segmented in various ways, such as type of participant (speculative or commercial) or instrument (futures, forwards etc). While this is undoubtedly useful in terms of the macroeconomic picture, the downside is that no one bank can provide the complete FX flow picture.

The logical sources for this complete picture are the primary FX market platforms. They witness the primary interbank and professional flows 24-hours-a-day and are therefore ideally placed to provide a comprehensive picture.

In addition, while many bank flow sources can provide only periodic (often daily) updates, the primary markets might some day provide this data in real time, thus allowing traders and trading models to track flows by region and currency throughout the day.

Good prospects

It is apparent that although FX market data have been historically sparse in some areas, the situation appears to be improving rapidly. The primary FX markets have now started to make new data sets available (both historical and real time) that address a number of existing limitations, such as latency. Furthermore, some of these new data sets, such as historic market depth, should open up entirely new possibilities in quantitative financial modelling.

---

[1]Constant volume bars (CVBs) are based upon the number of trades, not the amount of time elapsed. For example, a five trade CVB will show the first, highest, lowest and last traded prices of the last five trades. By contrast a five-minute price bar would show the first, highest, lowest and last traded prices of the last five minutes.

Built for Speed

*e-Forex asks **Mike Mistretta**, Senior Director, Infrastructure at FXall to tell us about their Latency Monitoring System.*

Mike Mistretta

Speed of execution is key to maximizing foreign exchange performance and productivity. For banks streaming prices to a multibank portal, speed is central to winning and retaining client business, while for model-driven traders, it means that trades can be executed with minimal slippage.

Research from Greenwich Associates demonstrates that speed of execution is regarded as the single most important benefit of online trading. 68% of top-tier financial institutions – including banks, asset managers, and hedge funds – cited it as one of the most important advantages realized by trading foreign exchange online. Active traders, such as hedge funds and CTAs, place particular emphasis on lightning-fast pricing and execution – our clients in this sector often tell us that speed was a primary motivation for moving volumes online to FXall.

Accelerating the trading proces

Speed of execution depends on two things – maximizing speed and minimizing latency. To increase the speed of trading, we have delivered:

Deep markets for one-click execution – FXall offers deep liquidity from more providers than any other portal.

Our diverse customer base delivers non-correlated flow, giving banks an incentive to put up more liquidity. This depth of liquidity means orders can get filled in one execution – far more efficient than managing risk over five or six transactions. What's more, our one-click execution capabilities ensure slippage is kept to a minimum.

Flexible connectivity options – Clients can interface to FXall seamlessly through FIX, API (including Java* and Microsoft COM*) or our STP solution QuickConnect, enhancing efficiency and eliminating the need for manual re-keying.

Powerful infrastructure – FXall is built for speed. Last year, we embarked on an ambitious project to design and deploy a new generation of software for FX trading.

Our new infrastructure is based around a grid technique of processing engines, which ensure the shortest code-path for maximum speed. It is highly scalable for increased throughput, meaning that as volumes rise it can be easily extended by adding individual software engines and hardware servers.

Minimizing latency

In the quest to deliver speed of execution, investing to make FXall the fastest, most effective trading platform is only the first step. Equally important is making sure it stays that way.

By monitoring latency both internally and externally, and by working with banks and clients to trouble-shoot issues as soon as they occur, our company-wide Latency Monitoring System (LMS) ensures that we consistently achieve the sub-second transaction time targets we set when we built FXall. 99.99% of trades are executed in less than one second.

*Java is the registered trademark of Sun Microsystems; COM is a trademark of Microsoft Corp.