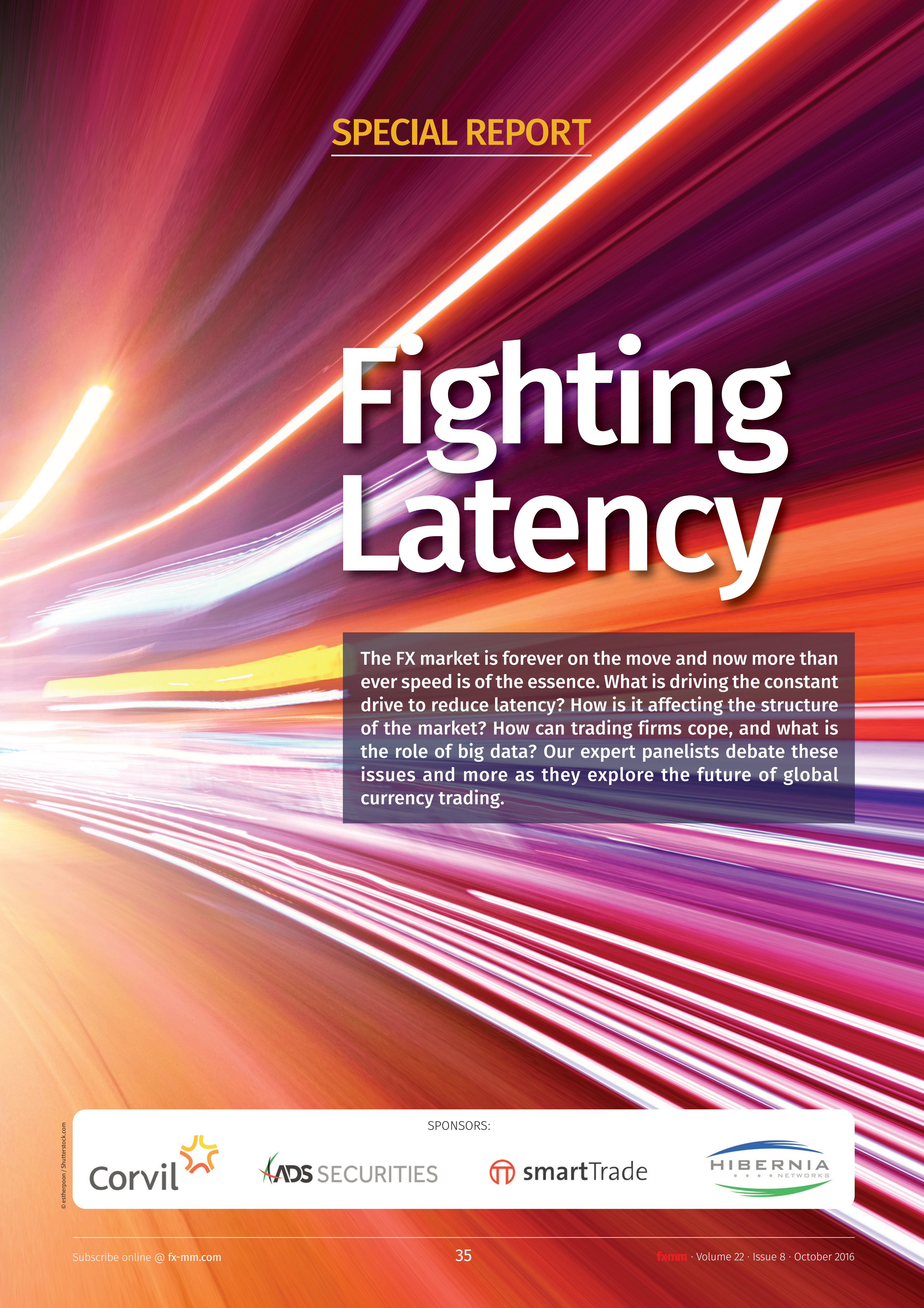


SPECIAL REPORT

Fighting Latency

The FX market is forever on the move and now more than ever speed is of the essence. What is driving the constant drive to reduce latency? How is it affecting the structure of the market? How can trading firms cope, and what is the role of big data? Our expert panelists debate these issues and more as they explore the future of global currency trading.

SPONSORS:

Corvil ADS SECURITIES smartTrade HIBERNIA NETWORKS

© estherpoon / Shutterstock.com

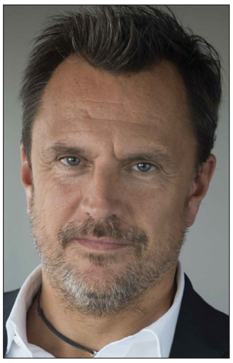
© estherpoon / Shutterstock.com

ROUNDTABLE

With contributions from:

Andrew Rossiter
Managing Director – Global Head, Information Technology, ADS Securities

Donal Byrne
CEO, Corvil



David Vincent
Co-Founder and CEO, smartTrade Technologies

Omar Altaji
Chief Commerical Officer, Hibernia Networks

Peter Garnham
Editor, FX-MM

Peter Garnham: Why is latency such a critical issue in FX compared to other asset classes?

David Vincent: Latency is not necessarily more critical in FX than it is in equities but as the fragmentation of the FX market has increased a lot over the last decade it has become more important. Today, there is a multitude of FX Liquidity Providers (LPs) and in order to improve their trading execution, to obtain better pricing and to comply with regulations, buy-side and sell-side firms need to connect to a dozen of them. Of course the better your latency is, the more accurate your view of the market will be and consequently the better your trading execution will be.

Donal Byrne: Unlike other asset classes, with FX there is no central market. Instead participants are essentially market makers trading across a complex and globally distributed exchange. With the electronification and inter-connectedness of the FX market, rapidly updating rates are available from many sources leading to arbitrage opportunity that favours those with the lowest latency.

Andrew Rossiter: You could almost say it should not be critical, because FX is an over-the-counter (OTC) market and is not like an exchange, where the first one to the exchange matching engine wins. But it is interesting that it has become an issue in what is an OTC market. You could also ask a wider question: does it indicate there is something wrong with FX, in that an OTC market has become latency dependent? It is OTC, it shouldn't be latency dependent.

One of the reasons this has happened in my view is that the market is so much more fragmented than before. And I think the fragmentation that everyone talks about over the past 12 to 18 months is leading towards latency being an issue in an OTC market. Deeper within that fragmentation on the latency side is the question: is the liquidity really there, or is it just a price, someone fishing for some kind of market

© hxdyl / Shutterstock.com

indicator. And also if that liquidity really is there, is it being warehoused, or is it being passed on? In this case there is market impact.

And market impact now, because of fragmentation and lower latency, is higher. The bigger banks that would have warehoused this risk have stepped back from the market, and have allowed new entrants to come in. but many of these new entrants aren't that well capitalised. They can't warehouse the risk, so they are passing that risk on as a price update. I think they are all feeding each other in this loop. So the newer entrants who are market making are passing their risk on, which in the exchange world would be like another print on the exchange. They are having market impact, so everybody is looking at everyone else and they are all moving their prices. They are tied together in my mind for those reasons. Latency should not be an issue, but it has become an issue because of fragmentation.

Now the market is chasing itself into this latency battle. The quote 'time to live' – how long that quote is good for – obviously depends on a lot of factors, but it keeps coming down. Whereas a quote would be good for 500 milliseconds a few years ago, now some quotes are only good for 10 milliseconds. Those people that are going for higher fill rates are doing it by either one of two methods, either widened spreads so you are protected, or quote more often. I think a lot of people are going down the quote more often route. I think that is going to continue, the quote frequency is going to go up and the quote 'time to live' is going to come down.

Peter Garnham: Is there much more latency to be squeezed out of communication systems? Or are we reaching the limits?

Omar Altaji: The short answer is yes. Faster speeds and latency performance continue to play critical roles in financial trading platforms. With machine-time becoming more prevalent as a driver of financial trading activity, it is only a matter of time before latency will be measured in nanoseconds. Latency performance is dependent on the pure physics of connecting two endpoints, the characteristics of the underlying transmission medium whether fibre or wireless, and lastly the optimisation of algorithmic software efficiencies.

> *The better your latency is, the more accurate your view of the market will be and consequently the better your trading execution will be*
> *David Vincent*

Concurrent with financial trading platform acceleration, Hibernia Networks is committed to improving latency performance for customers on a continual basis by optimising network architecture and deploying measures that drive transmission speeds even faster to achieve lower latencies. In the end-to-end trade flow cycle, the most significant latency improvements remain in the physical routing of the service – a geodesic route wins the latency race, as demonstrated by Hibernia Express connecting the key financial centers in North America and Europe.

Peter Garnham: How can trading firms ensure the speed and reliability of their data transfer? How fast is fast enough?

Donal Byrne: Infrastructure must be appropriately engineered and scaled to maximise speed. In FX, relying upon one second averages to size network links and server performance no longer cuts it because transient peaks occur within millisecond and even microsecond timeframes. Visibility and analysis at the timescales that machines can execute a trade are critical. This is machine-time analytics. It provides transparency and understanding between infrastructure and trading performance. But once bottlenecks have been removed, further investment has a quickly diminishing return. With the right machine-time analytics, traders can correlate the relationship between trade outcome and latency. This is the key to understanding how fast is fast enough.

Andrew Rossiter: It is the same in any asset class: you have got to have good, well written software. You need the basics – as lower latency network as you can afford. That is a given, but one of the things we look at closely is something called jitter. If 99.9% of your quotes are relatively quick but you have got 0.1% that are spikes in latency, those can cause real issues. That is called jitter.

So it is not only about low latency in your system, but whether it is consistent. As latency becomes more and more important, then these outliers become more and more important, and you need to spend more and more time getting rid of this jitter, because these spikes in latency become more and more expensive. In order to do that, you have got to capture a lot of data. As your messages go through your system, you need to capture how long they are spending in each part of the system, right down to how long they are spending on the switch, firewalls, network cards, on the wire, in the software.

© amesteohart / Shutterstock.com

Omar Altaji: Looking at the transmission platform, pure end-to-end optical service is not the only determinant of latency performance. RF (microwave or millimetre wave) has firmly established itself for those seeking the absolute lowest latency. However, RF technology has its limitations.

The economics of RF limit distance, due to cost of tower space as well as other operational constraints. Distance also impacts both latency and bandwidth. Increasing the distance to reduce tower count on a route can impact line error rates. Increasing it too high may provide the fastest theoretical path, but may render it unusable for periods due to environmental factors causing poor reliability. Striking the right balance is important, which is driven by the trading strategy. Predictability is extremely important to any strategy. An industry has emerged around measurement and monitoring with a focus on accurate time-stamping for surveillance purposes.

Hibernia Networks realises the value of a future-proofed solution to protecting our customers' competitive edge. To that end, we are continuously upgrading capacity, expanding route diversity and leveraging innovative technologies to lower latency between the major financial centres. Our priority is to give customers the peace-of-mind that their network performance is unparalleled, enabling them to focus on their core business objectives.

> *It is not only about low latency in your system, but whether it is consistent*
> *Andrew Rossiter*

David Vincent: Latency can occur in different points of your trading process: your connections to your LPs, your network, your hardware set-up, your trading platform location and your platform performance. This is the reason why outsourcing to a single provider, such as smartTrade technologies, has become a more and more popular solution with FX trading firms. Your technology provider is in charge of optimising all the technological processes of your trading system and is held accountable for the performance, speed and reliability of your data transfer. Our teams are constantly monitoring our clients' systems' performance but are also in charge of benchmarking our software in different scenarios so that latency efficiency is insured even in exceptional events, such as the SNB crisis or Brexit, where systems had to handle very high volumes of data. Our R&D team constantly tests new hardware and software techniques and solutions to further optimise our trading platforms performance. Transparency is also key, this is why we provide regular performance reports to our clients.

Additionally your need for speed is different according to the kind of trading you do. For example, determining where to host your Smart Order Routing (SOR) according to where you trade might be a complex exercise if you trade on multi-venues and even more difficult if you do multi-asset trading. One effective way to deal with these issues is by outsourcing to a specialist.

Finally, colocation is of course a 'must have' in order to reduce the latency of your applications.

Peter Garnham: What challenges do trading firms face as they seek connections to more trading venues and counterparties? How can these be overcome?

Andrew Rossiter: The fragmentation of the market means that there are a lot more places to get your prices from. Some of those places are quoting very quickly – with thousands of updates per second across a host of currency pairs. The challenge we have is not only consuming that amount of data into our trading platform, but also into our big data platform. As we keep adding more and more venues and storing more and more data, we have to keep updating that system in order to be able to cope.

We monitor the market more or less in real-time, so we have a blotter dashboard that for every order that comes in, it gives us within a couple of seconds of that order being done, a complete breakdown of the market data that was involved in the system at that time: how long it took from the LP all the way through our system out to the client, how quickly the client responded,then how the order went through the system, how the LP responded , how the market responded to that fill being done. As we go to more and more venues, it becomes a bigger and bigger challenge.

For us it is not the traditional challenge that we have got a lot of cross connects and we have to manage the networks, because there are people that can help you with that – you can outsource that. I think that is becoming more commoditised. The challenge that we are having is to really understand how our system behaves, because we are using that to understand how an LP behaves, how the market behaves. There is a dedicated team here at ADS just focused on big data capturing.

David Vincent: I think the most important challenge for trading firms is to make sure their time to market is as short as possible. In these uncertain and very changeable market conditions, if you are not able to rapidly implement a solution, you are losing money and trading opportunities. We have over 70 established connections to LPs which allows our clients to have access to additional markets in just a few days. And to develop a new connection from scratch, our savoir-faire as a technology company enables us to develop it in less than a few weeks.

Finally obtaining a list of prices from a variety of venues is important but implementing a powerful aggregation system, to insure your traders have the most effective way to visualise the market, is crucial too.

Donal Byrne: As connectivity increases, so does complexity and cost. Plus, trading with more counterparties increases exposure and risk. Obtaining an accurate and consolidated view of all activity across all venues and counterparties becomes a necessity in managing the environment. Many are adopting a model that comprises of distributed, real-time data collection feeding a centralised analysis, reporting and alerting function

Omar Altaji: Close to 90% of financial trading activity is executed across 14 key financial centres along the East-West axis. This trading dynamic lends itself to a single global network platform for consistency and optimised latency performance. Trading firms seek to strike a balance between proximity to the key venues and reach, especially for the more complex algorithmic strategies.

As new hubs of financial trading activities emerge, trading firms and service providers will need to move rapidly to ensure access to those new exchanges. Hibernia Express has transcended its physical route to become an integral part of the low latency financial trading ecosystem across the globe given its strategic positioning and access to Chicago, New York, London and Frankfurt.

Peter Garnham: How have providers sought to overcome the latency challenge of the increased use of high-frequency and algorithmic trading techniques? Is there a role for big data analytics in reducing latency?

David Vincent: Latency can occur anywhere in your trading cycle, in order to avoid it and to be as efficient as a high frequency firm you need to work with a provider who will help you optimise your trading workflow. On top of an effective connectivity to your LPs, the trading platform needs to be enhanced. Our most sophisticated clients use our Order Management System (OMS), our Smart Order Routing (SOR), and our Distribution system which allows them to build a more performant trading workflow, to improve their fill ratios and reduce their slippage. Furthermore to better fine-tune a trading platform, leveraging on big data analytics has become strategic. With our smartAnalytics solution our clients can store, analyse and visualise all their trading data, they can conduct pre-trade and post-trade analyses which allows them to choose the most adapted algos and the most suited LPs according to the type of strategies they are trading. Additionally, this solution enables clients to store and monitor their latency metrics across their entire trading infrastructure.

> *As connectivity increases, so does complexity and cost. Plus, trading with more counterparties increases exposure and risk*
> *Donal Byrne*

Donal Byrne: High frequency and algorithmic trading are terms synonymous with equities trading. As these activities crossover into FX, so do the supporting technologies that address the challenge of latency. A prime example is the adoption of co-location for FX, essentially to provide a high speed access to multiple Liquidity Providers under one roof. Regarding big data analytics, it may prove useful over time in spotting trends that could optimise trading strategies. Analysis of streaming machine-time data is critical for transparency and regulatory compliance. Certain big data tools will be helpful for this purpose.

Omar Altaji: We are not seeing an impact on latency performance on our network based on the type or volume of traffic that traverses it. Nonetheless, we are always seeking opportunities to improve latency performance, resiliency and predictability. As it relates to big data and cloud analytics, clearly there's a role to potentially adjust network routing based on communities of interest between financial centres that drive trading activity.

© jamesteohart / Shutterstock.com

> *Trading firms seek to strike a balance between proximity to the key venues and reach, especially for the more complex algorithmic strategies*
>
> *Omar Altaji*

Andrew Rossiter: You have to have consistency. You don't really want a single quote, or piece of market data that is slow in the system because that can have knock-on effects – as I said, you have to eliminate jitter.

Big data analytics can reduce latency because you can spot parts of your trading system that are slow. You really need this forensic analysis to be able to look at the data.

Peter Garnham: Is trading speed limited by geography, or is physical location no longer an issue as regards latency?

Donal Byrne: Once servers, applications and algorithms are optimally tuned, to lower latency further attention is generally focused upon the communication between those machines, or "how quickly can we get data from one system into another". Locations, distance and the speed of light through optical fibres are all now being considered. So yes, geography and physical location is becoming increasingly relevant as these microsecond tolerances make a real difference when machines are making the trading decisions. However, the FX market is a global market and highly distributed and fragmented. Co-location can be a good point solution for key market spot but not a universal solution. This is why information about speed in the FX market will be as valuable as speed itself. Understanding precisely the parameters of latency and time to all FX venues allows traders to maximise trade outcome and minimise risk of trading on stale quotes.

Omar Altaji: We view geography as a dictating factor of latency performance, especially in the financial ecosystem. Financial trading is globally-based on city pairing. Latency performance between the financial exchanges is deterministic. The physical location of the trader is less relevant.

Hibernia Networks connects the world's top 13 world financial exchanges. We continuously innovate on route design to improve latency performance between these exchanges as well as extend the reach of our network connectivity platform to new exchanges.

Andrew Rossiter: This depends on the type of client. For a corporate client that wants to trade $50 million, latency is not an issue as long as you give a good price and a full fill.

If your client is someone who is trying to hedge, or you are receiving their hedging flow, then yes latency is an issue and physical location will be an issue. You are going to want your matching engine as close to the liquidity as possible. I think this is why you are seeing some firms pulling back from the market because the cost of chasing this type of client is huge.

David Vincent: It is not an issue as we have very efficient solutions to manage this challenge, such as proximity hosting and colocation. New York, London and Tokyo are still the big FX hubs where most of the FX players are located. Having your applications hosted as close as possible to price sources, we give the fastest time-to-market to our clients. Because the FX market is tradable 24 hours a day, we also advise our clients to have a follow-the-sun infrastructure with applications located in each of the FX hubs.

Peter Garnham: How can cloud solutions assist trading firms to overcome the latency challenge?

Andrew Rossiter: Again, it depends on the type of client. If you have a mobile platform, then clearly latency isn't such an issue as when you are dealing with a high frequency trader trying to hedge.

For a mobile platform, however, a cloud solution can help. For instance, you can have a very fast pipe to, say, Asia, and then you can use cloud providers so clients in Malaysia or the Philippines can connect to their local providers. You can then use a direct connection from that cloud provider to your trading engine. That can definitely help to lower overall latency, and it can reduce your costs, especially when dealing with retail clients.

Donal Byrne: Addressing the latency challenge is all about speed to liquidity. Co-location has proven to be a popular approach in Equities for many years and we're now seeing it cross over into FX. With some FX liquidity moving into the cloud, we are likely to see cloud data centres playing a role equivalent to traditional co-location. In practice, this is likely to increase fragmentation and complicate the latency landscape.

David Vincent: The cloud can actually add some latency if you do not set it up properly. We at smartTrade only use private cloud solutions because we are able to control, set-up and manage their environment.

For further information: **www.fx-mm.com**