Safety in numbers

The approach taken to address low-latency in equity markets is changing – the historical drivers of the pursuit of low-latency no longer hold in equity markets as the business imperative for low-latency has expired, and banks are beginning to take a more holistic view of latency management. By Saoirse Kennedy, GreySpark Partners.

Since pre-historic human communication and barter systems, trading has suffered the confines of latency. The pursuit of low-latency trading is evidenced in horses racing to bring information on arriving ships from the harbour to the market; signal lanterns in the 1840s that sent messages from New York to the Philadelphia stock exchange in under 30 minutes; telegrams and telephones, and fibre optic cables that transport information in milliseconds and now nanoseconds.

In equity markets, interest in technology to achieve low-latency and ultra-low-latency peaked just after the financial crisis in 2009. Before 2009, banks engaged in a technology arms race to reduce latency and develop superior quantitative algorithms to take advantage of market opportunities. Electronic trading, including algorithmic trading (AT) and high-frequency trading (HFT), is the source of the capital markets industry's interest in low-latency trading technologies – the highly electronic equity markets are an opportune asset class for the pursuit of low-latency strategies. Prior to the financial crisis, a technology arms race was decisive for equity market players pursuing AT and HFT in achieving the lowest possible latency, fuelling the race-to-zero latency that is now over.

Fig 1: Business outcomes of latency in equities

Why is the race to zero-latency over?

A 2012 analysis by GreySpark showed that a bank must halve its trading systems' latency every three years to keep abreast of the pace of change. This means that the end-to-end

processing time of electronic trades had to decline by 90% in the 10 years prior to this study to stay competitive. Low-latency technologies in equity markets degraded latency almost to the level of latency experienced in flow FX streaming and, as such, the race to zero-latency ended (see Figure 1).

The business outcomes of latency are determined by three latency zones (see Figure 2). Between each zone, for each asset class and business line, there is a latency threshold that varies by market and which evolves over time. These thresholds delineate the three zones of latency. At present there is no longer a business case in equity markets to continue pursuing low-latency, and it is understood across the capital markets industry that it is acceptable to remain in the safe zone.

New systems and algorithms continue to have low-latency as one of their core, non-functional requirements. However, the cost-benefit of moving to ultra-low-latency platforms is no longer compelling. Tier 1 banks are comfortable when on par with one another, with no single bank getting too far ahead of the pack – there is little advantage in being faster than the pack, it is sufficient to be fast enough. This is reinforced by the availability of information that was once not widely disclosed, but which is now easily accessible, and widely understood and put into practice.

HFT is no longer a driver in the low-latency arms race. HFT, which emerged in the late 1990s, experienced peak trading volumes in 2009. By this time, HFT accounted for almost 75% of all US equity trading volume[1]. As HFT strategies became more widespread, the cost of maintaining a competitive advantage increased. As a result of the proliferation of HFT strategies, the larger investment required for those firms to maintain a competitive advantage amid increasing regulatory pressures caused HFT profitability to decline from a peak of around USD 7.5bn globally in 2009 to an estimated USD 2bn in 2013[2]. In 2014, we see that HFT's share of equity trading volumes in the US stabilised at about 50%, while there is still room for HFT growth in Asia and slight growth is also expected in Europe[3] (see Figure 3).

Regional market structures and conditions, including auto-execution, maker-taker pricing structure, small lot sizes, availability of liquidity, fragmentation and trade-through protection explain differences in levels of HFT activity across various jurisdictions. These conditions are all present in

Fig 2: Latency zones and thresholds

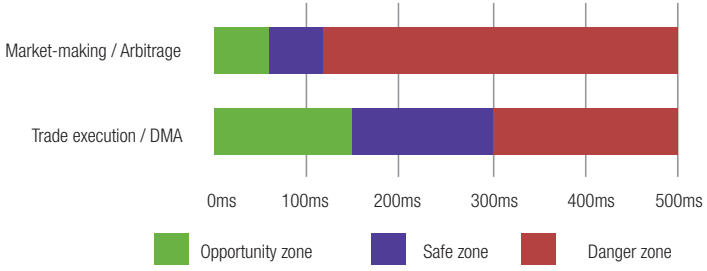
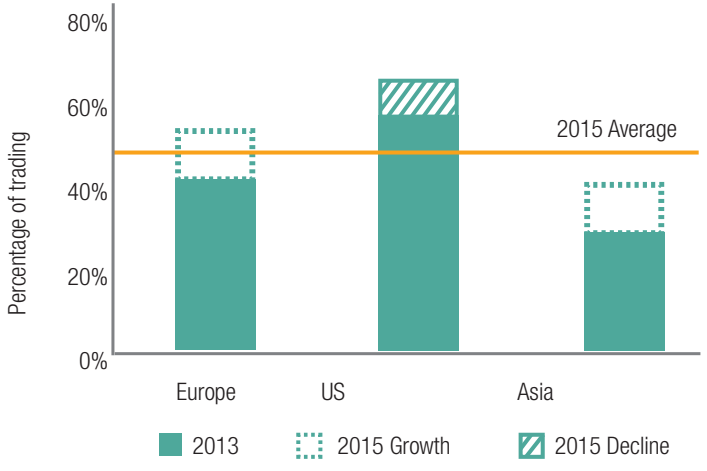


Fig 3: HFT is Stabilising Globally



US equity markets, which have the highest degree of HFT. In Asian equity markets, the ratio of HFT to total traded volume is smaller and will remain so for as long as the market structure is unchanged. The projected growth of HFT in Asia and Europe is fragile and will depend on whether regulators impose further controls on HFT activity. HFT requires both algorithmic trading flows and voice-trading flows to remain viable because HFT only exists where electronic trading prevails sufficiently for the returns to justify the technology investment. With the reduced profitability of HFT, a period of *détente* has begun with fewer players participating in the race-to-zero latency.

Hardware-accelerated trading tools are also impacted

Demand for hardware-accelerated trading tools, such as field-programmable gate arrays (FPGAs), has stopped growing, this acts as further evidence of the end of the arms race. FPGA solutions, in the form of market data handlers and line handlers, and for pre-trade risk checks, deliver the most latency-efficient solutions among hardware acceleration tools. They were marketed extensively during the peak of the latency race, but their up-take was low, primarily due to their costly nature.

The latency benefits delivered by FPGA solutions are costly – they are expensive because of the initial development effort required to implement their usage and because of their long-term maintenance cost. Although standardised FPGA development languages such as VHDL helped, a typical FPGA development cycle still requires 20-to-40 times the development effort of traditional software. Trading venues tend to update or enhance their protocols at least once per year, which requires significant FPGA redevelopment.

The business case for maintaining FPGA solutions is limited to a group of ultra-low-latency traders that utilise FPGAs for pre-trade risk feeds and feedhandlers. A 2014 GreySpark survey of equity market participants and third-party technology vendors shows that banks not pursuing ultra-low-latency trading strategies are happy to stick with software-optimised solutions and that technology vendors are not generally making further investments in developing FPGA technology. Banks that continue to pursue ultra-low-latency strategies using FPGA solutions are looking to reduce the total cost of ownership by migrating from in-house solutions to hosted, end-to-end solution providers that benefit from economies of scale.

The importance of low-latency is refocusing

Low-latency remains important, but it is no longer wholly concerned with the last millisecond of latency. As it was important in the past to achieve ever lower-latency, it is important now to improve the distribution of latency by reducing the likelihood of jitters.

Latency must be approached from a monitoring, consistency and reliability perspective. Holistic performance monitoring of infrastructure latencies, order-to-fill latency, performance latency, firewall latency and external latency, for example, must be performed passively so as not to add to overall latency. Additionally, onboarding the most suitable middleware solutions based on individual use cases and on the basis of a holistic view of an organisation's infrastructure allows an effective latency-reduction strategy or latency-management strategy to develop. Adopting this holistic approach to latency will prevent equity market participants from falling into the latency danger zone, and help them maintain the pace of the pack. ■

Footnotes:

1. *Popper, N., 2012. High-Speed Trading No Longer Hurtling Forward. NYTIMES.com. [online] 14 October. Available at: <http://www.nytimes.com/2012/10/15/business/with-profits-dropping-high-speed-trading-cools-down.html?ref=highfrequencyalgorithmictrading&_r=0>.*
2. *Ibid.*
3. *Ibid.*

For further information on the subjects of equities, low-latency and HFT, please see the following GreySpark research reports, available at research.greyspark.com:

- *Trends in Equities Trading 2014*
- *Low-latency Messaging Middleware: Pursuing Nanosecond Trading*
- *Low-latency Faster than Light*
- *Low-latency in Asia-Pacific: An Infrastructure View*
- *HFT I: Defining HFT Activity and its Regulatory Landscape*
- *HFT II: How Appropriate Risk Management Practices Can Offset HFT Risks*

Additional contributors: Asif Abdullah, Jon Batty, Anna Pajor, Frederic Ponzo