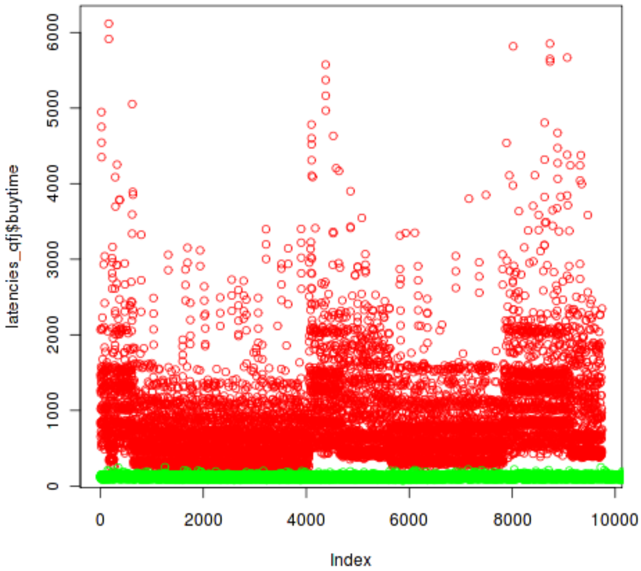
RAPID ADDITION LIMITED

Performance Whitepaper

Latency spotlight

Issue Date: 28 September 2010

Version: 1.0 Final



This document presents a discussion on latency measurement and importance to organizations wishing to trade in today's marketplace. Included are the results of controlled tests run by Rapid Addition to compare the performance of their leading FIX Engine, RA-Cheetah/J, against the open source FIX engine QuickFIX/J.

Table of Contents

Introduction

Latency is widely acknowledged to be extremely important in trading with many firms and exchanges investing significantly in reducing the latency profile of their systems. As latency has become more important so too does the effect of Jitter – the random trade by trade variation in latency.

This whitepaper outlines why latency and jitter matter, provides an approach for evaluating the value of latency for an organisation and then examines the latency and jitter performance of the Rapid Addition RA-Cheetah/J FIX engine (which uses the proprietary GenerationZero™ framework to reduce latency and achieve a low-jitter profile) and the QuickFIX/J FIX engine. The tests performed and the testing methodology utilised are also presented here.

About Rapid Addition

With over 70 clients worldwide, including stock exchanges, Rapid Addition is the leading provider of FIX and FAST related software solutions to the global financial community.

Rapid Addition is a technology partner of Microsoft and founding member of the Intel Low Latency Labs. Rapid Addition is the only FIX vendor to continuously test their products in the labs. RA-Cheetah, the flagship FIX engine, gives a consistent low-latency performance and is the only FIX engine to provide this level of consistent measurement. RA-Cheetah™ and GRHub™ (our order routing software) run on our advanced low-latency GenerationZero™ technology, which, inter-alia ensures no garbage collection.

For further information, please visit www.rapidaddition.com.

About the author

Company chairman Kevin Houstoun is the designer of the FIX Repository for FIX Protocol Limited (FPL). Mr Houstoun co-chairs the FPL Global Technical Committee and is an active member of the FPL Global Steering committee. He is also a member of the lead expert group for the UK Government's Foresight Committee on the future of computer-based trading.

Executive summary

The findings presented in this whitepaper show that the profitably that can be ascribed to any particular system can be adversely affected when latency characteristics are not given due consideration. The notion that systems that make money must be "fast enough" is shown to be naïve as such systems may actually represent substantial lost profit opportunities.

We also shows that whilst the QuickFIX/J FIX, open source, engine may be functionally complete it is unsuited for use as a component in a low-latency trading system as design limitations mean that even at low messaging rates it is not able to achieve sub-100 microseconds to process the following flow:

1. Take a FIX messages from the network card;
2. Create a new FIX message;
3. Send the message out to a counterparty destination.

At higher messaging rates, garbage collection introduces considerable variability to the performance of QuickFIX/J making it unsuited to strategies that rely being able to factor consistent latency figures into their calculations.

Contact Rapid Addition for an analysis of your trading environment.

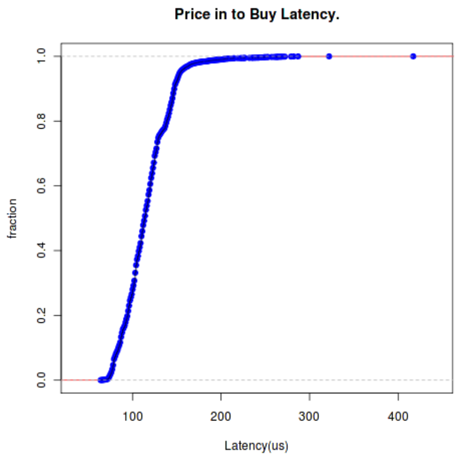
Why latency and jitter matters

There are two scenarios in all low latency trading activities where latency really matters:

1. System-A detects that a price is available at a venue that fits with its strategy. The system needs to create an order and send it to the venue for execution at the price the system originally saw.
2. The system detects that a price is being offered by your institution that is incorrect. A Cancel or Modify instruction needs to be sent to amend the price before someone else hits it.

Essentially all low latency trading activities, be it arbitrage, smart order routing or market making are made up of these two scenarios. For example, a smart order router (SOR) will determine that a good price is available and try to route a child order to the appropriate venue before someone else. IN another example, "Risk free" arbitrage is where the system spots a profitable opportunity to execute a trade buying on one venue and selling it on another, requiring that both “legs” of the trade are executed before someone else causes the price to move.

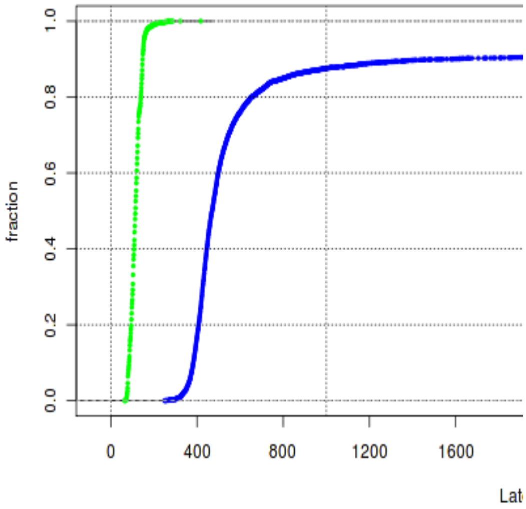
Each actor in the market has a trading system that will have a latency distribution that can be summarized in the table below. Here we show the percentage of messages written onto the wire, y-axis before a particular latency time, x-axis.



In the system described by the chart (left), we see that 25% of the orders are on the wire within 98 microseconds of the price being exposed to the network card. 50% of the orders are on the wire within 103 microseconds and 75% of the orders are on the wire within 114 microseconds. Such charts often do not show maximum latency well and numbers are frequently truncated. Here the maximum latency is 417 microseconds (and is shown on the chart).

The above chart is measures the time from a FIX Market Data Incremental Refresh (MsgType = X) message being presented to the network card on the server to the time that a corresponding FIX New Order - Single (MsgType = D) is presented to the wire on its outbound journey.

When we place the latencies from two competing systems on the same chart we can see how a slower system will occasionally still win the race to the price. Here we can see that fastest trades trade from the system represented by the blue line is faster than the slowest trades from the system represented by the green line. What percentage of the time the slower system will beat the faster system to the price will depend on any correlation between the latencies of the two systems. The important principle to understand here is: just because a system is occasionally making a



profitable trade, it does not follow that it is making as *many* profitable trades as it could be.

Effectively, in these scenarios each market participant is racing other market participants to be the first person to respond to a particular event or piece of information. Each has a system that will have a range of response times distributed around its average response time and so whether or not you get your fill, or modify your price in time, become a probabilistic function rather than a certainty. The higher the probability of receiving the correct fill or modifying a price, the higher the probability of profitability AND the higher, therefore, that profit is likely to be. By failing to achieve the highest probability the system will lose money even if it doesn't make a loss.

So, for any individual trade the benchmark latency number to beat is that of the closest competitor.

Obviously if all participants trying to exercise the same strategy, in aggregate. had this same distribution each participant would, over the long run, win an equal percentage of the opportunities and the profitability of the trading strategy would simply be determined by the number of participant which would rise over time until no one made more than a fair economic return.

Since competing funds are all operating software and hardware devices subject to the same laws of physics, the competition can effectively be represented by one of these charts. However, practical difficulties in measuring the performance of competitor systems mean that for most trading strategies firms are in the suboptimal position of only having a "feel" for the benchmark number.

Often firms do not try to estimate the spread of competitors' latency. This can be determined by careful analysis of slippage in a firm's trading; however the real time telemetry to measure this is expensive and may not even be available (see the FIX Inter-party latency working group for details of an initiative on how to address the problems of measuring latency consistently: http://www.fixprotocol.org/working_groups/ipl). For the purposes of this document we make the simplifying assumption that competitive market funds can be summarized by a line: below the line a firm can trade successfully; and above it strategies will fail.

So, what characteristics should one look for in a trading system's latency profile? The slope of the line should be as steep as possible and as far to the left as possible to minimises slippage and opportunity costs. Sometimes it may be necessary to choose between two competing solutions which have different jitter and latency characteristics, both capable of getting some trades through fast enough to trade successfully. Here you should chose the system that offers the highest expected return. This normally means a latency weighted average trade return.

Latency-weighted average trade return

To calculate a latency weighted average trade return we need to know a number of things but the two most important are the upside of a successful trade and the downside of an unsuccessful trade.

Slippage is simply the difference between the price you expected to get based on the price you see at the time of placing the order, and the price you actually got when your order reached the market[1].

[1] Note, some people argue that slippage can only exist on limit orders but that is a separate debate not within the scope of this document.