IT Infrastructure for High-Frequency Trading

Bob Carasik
Wells Fargo Bank, bcarasik@wellsfargo.com

Follow this and additional works at: http://aisel.aisnet.org/whiceb2013

Recommended Citation
http://aisel.aisnet.org/whiceb2013/13
IT Infrastructure for High-Frequency Trading

Bob Carasik
Wells Fargo Bank, bcarasik@wellsfargo.com

Abstract: Algorithmic trading has become a huge proportion of the trades executed on securities exchanges around the world. Reduced transaction costs have facilitated a new generation of trading strategies which depend on computing and network speed. Algorithmic trading also poses a new set of risks and challenges regulators to keep up. This paper reviews the technology infrastructure supporting algorithmic and high-speed trading.

Keywords: HFT, algorithmic trading, analytics, FPGA

1 INTRODUCTION: WHAT IS ALGORITHMIC TRADING?

Most major financial institutions provide a wide range of services for its customers, and continually implements innovative technology to remain competitive. On the consumer side of the business, mobile technologies have been the most prominent area of innovation. In wholesale banking, support for algorithmic trading has been an area of major change. Investopedia defines algorithmic trading as “A trading system that utilizes very advanced mathematical models for making transaction decisions in the financial markets. The strict rules built into the model attempt to determine the optimal time for an order to be placed that will cause the least amount of impact on a stock’s price. Large blocks of shares are usually purchased by dividing the large share block into smaller lots and allowing the complex algorithms to decide when the smaller blocks are to be purchased.” “High frequency (HF) trading firms represent approximately 2% of the nearly 20,000 trading firms operating in the U.S. markets, but since 2009 they have accounted for over 70% of the volume in U.S. equity markets and by early 2012 were said to be fast approaching 50% of the volume in futures markets”. Algorithms typically control the timing, routing, quantity, and of orders. They dynamically monitor conditions across different securities and different exchanges. Firms attempt to limit the market impact of their orders by techniques such as randomly breaking large orders into different-sized smaller ones. Algorithms watch trades placed by other traders and alter their trading strategies accordingly, frequently placing and cancelling trades in order to mislead other traders and maximize their own advantages.

The rapid growth of automated markets and high speed trading merits the attention of IT staff at any financial institution. The volume of trades executed outside of the “brand name exchanges” has exploded in recent years. Traditionally, investment decisions are based on expectations about a company’s earnings and assets. By contrast, high-frequency trading (HFT) bases trading decisions on very small, very rapid changes in stock prices and on differences between prices on different exchanges, some public, some not. Improvements in technology have caused the average time to execute an order has dropped from 25 milliseconds(msec) in 2000 to 1 msec in 2010. Numerous press sources suggest that profits from HFT approach $21 billion per year, but there is only a guess since there are no clear definitions to follow. In the book “Dark Pools”, Scott Patterson reports that in the 1940's, stocks were held on average for four years. By 2000, stocks were still held for an average of 8 months. But by 2012, stocks are held, on average, for 22 seconds. A buy or sell order no longer involves a shouting group of traders on an exchange floor. Technology has made trading so swift, and so cheap, that 90% of orders that are placed by algorithms are later cancelled, within milliseconds of placement, just to game rival trading algorithms. The risks involved in algorithmic trading are poorly understood. These risks came to public attention in the “Flash Crash” of May 6, 2010, when the Dow Jones dropped nearly 1000 points
in a few minutes and then recovered almost as quickly. In August 2012, the "Knightmare" loss of $440 million almost bankrupted a Knight, leading trading company, when software that should have been un-installed somehow re-activated and was erroneously multiplying trade volumes by 1,000. At that time, many financial industry observers argued that new automated safeguards kept the erroneous errors placed by Knight’s algorithmic trading from snowballing into a much larger disaster. Some commentators suggest that the profitable days are over, as an HFT "arms race" has led to a stalemate situation where HFT traders face a zero-sum situation with profits declining from a peak in 2009. Further regulations will no doubt be forthcoming as algorithmic trading becomes better understood.

2 ANALYSIS OF TECHNOLOGIES

While the business rules (algorithms) for algorithmic trading are proprietary, the infrastructure technologies needed to support it are not, and there is lots of public information about them. The computing resources required to support high-speed trading are optimized to maximize computing performance and to minimize latency.

![Hardware acceleration applied to a trading infrastructure](image)

**Figure 1: Block diagram of trading infrastructure, from the Tabb Group, showing possible benefits of FPGA, GPU, and Multicore optimization technologies.**

The lowest bound on latency is of course set by the speed of light, so algorithms need to be run as close to the exchanges as possible. Light travels approximately 100 miles per millisecond, so when trades are made in a few milliseconds, a few miles begin to matter. Co-location, or “co-lo” facilities are built close to national exchanges, and sometimes even in their data centers, to support algorithmic equities trading.

Networks themselves can be highly optimized, with optical fiber, and even dedicated microwave links being pressed into service in order to bypass the public Internet and minimize the total number of hops. Network connectivity is an important locus of optimization. Merely using low latency cards can speed up access by a factor of four. Field-programmable gate arrays (FPGA) and application-specific integrated circuits
(ASIC) can execute complex logic in hardware. Either technology provides an order of magnitude increase in execution speed. FPGA technology is newer and costs far less to develop and program than ASIC, and has therefore become popular for many roles in computing infrastructure.

<table>
<thead>
<tr>
<th>Latency</th>
<th>Standard 10GE Network card</th>
<th>Low Latency 10GE Network Card</th>
<th>FPGA</th>
<th>ASIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20 micros + application processing</td>
<td>5 micros + application processing</td>
<td>3-5 micros</td>
<td>Sub-micro</td>
</tr>
<tr>
<td>Ease of Deployment</td>
<td>Trivial</td>
<td>Kernel driver installation</td>
<td>Retraining of programmers</td>
<td>Specialist</td>
</tr>
<tr>
<td>Man Years Effort to Develop</td>
<td>Week</td>
<td>Weeks</td>
<td>2-3 man years</td>
<td>2-3 man years</td>
</tr>
<tr>
<td>Elapsed Time</td>
<td>Week</td>
<td>Weeks</td>
<td>6 months +year</td>
<td>Year +</td>
</tr>
<tr>
<td>Costs</td>
<td>$50 - $200</td>
<td>$500+</td>
<td>$100 - $20,000</td>
<td>$1million+</td>
</tr>
</tbody>
</table>

**Figure 2:** costs and benefits of specialized network hardware, from HFT Review

Algorithmic trading orders need to be generated in real time based on trading data, so processing needs to be optimized as well. In-memory databases, extensive caching, and kernel bypassing help squeeze maximum performance out of processing resources.

In-memory databases can speed up calculations by orders of magnitude. Assuming that RAM access requires 400 nanoseconds (ns) while disk access takes 10 milliseconds (ms), placing data in RAM provides at 25-fold performance improvement.

Processor architecture can be optimized in many other ways, for example graphics processing chips can be pressed into service in a new role: data analytics. GPU units can support parallel execution streams to greatly accelerate data analysis tasks.

Kernel bypassing is used to speed data from network interfaces into processor main memory, and thus directly to applications that evaluate incoming information and execute trades. With kernel bypass, packets avoid the NIC driver, TCP stack, and socket layer and can be consumed by applications. This allows the system to "Reduce copying and synchronization costs incurred by traditional kernel-based receive operations and optimized for user applications tightly coupled with data and achieve throughput under 5 microseconds with jitter (variability) of a microsecond or less."

Large banks typically run order management systems (OMS) and execution management systems (EMS), provided by vendors such as Bloomberg, Charles River, and FlexTrade, among others. OMS support the workflow involved prior to a trade, while EMS's, as the name suggest, actually execute the trade. Both EMS and OMS both benefit from low-latency computing since they depend on rapid access to information, and either system may support algorithmic trading functionality. OMS and EMS are supported by data feeds such as those from Bloomberg, BrokerTech, and TradeWeb.

Data analytics is at the heart of the algorithmic trading universe. Firms need trading data, not just about equities, but about futures, options, currencies, and commodities, in order to create predictions about market behavior. Not just traders, but their regulators rely on sophisticated data analysis. It took five months for the SEC to identify high-frequency trading as a cause of the "Flash Crash" in May 2010, when U.S. markets briefly lost $862 billion in shareholder value. Starting in late 2012, the S.E.C will be using an analysis system called Midas, and in the future, a consolidated audit trail, in order to identify anti-competitive behaviors and develop new regulations to match new financial products.
Figure 3: Data Analytics in Support of Financial Regulation

A typical trading operation at a large brokerage firm could include several hundred trading desks, each would include monitors for order entry and risk analysis. A typical trading workstation consists of a blade server with 160 GB disks, with 8 or 16 processors and 4 or 8 GB of RAM, driving several 24-inch screens for each trader. Portals link the computing hardware to the traders’ desks. The supporting systems for these securities and commodities traders can include grid computing resources from Data Synapse and Calypso.

3 BUSINESS OPPORTUNITIES

Investopedia estimates that in 2009, “more than 50% of exchange volume comes from high-frequency trading orders.” Many estimates are even higher, since a growing proportion of trades take place in so-called “Dark Pools”, outside of the regulated exchanges. A U.S. Securities and Exchange Commission SEC proposal to create a consolidated audit trail, was announced in July, 2012: " The Securities and Exchange Commission today voted to require the national securities exchanges and the Financial Industry Regulatory Authority (FINRA) to establish a market-wide consolidated audit trail that will significantly enhance regulators’ ability to monitor and analyze trading activity ".

The speed at which new technologies and algorithms are created threaten to make any regulatory proposal obsolete by the time it could be implemented. Any beefed-up regulatory regime that could provide more transparency in the world of high-speed trading will require even more bandwidth and less latency in order to distribute a near-real-time audit trail. It is possible that the early excesses of algorithmic trading may soon be behind us. Instead of being a secretive technology that enabled a minority of firms to make unfair profits, it may become a nearly universal part of the securities markets, and perhaps even deliver on its promise of providing greater liquidity to the trading public.

4 RECOMMENDATIONS

The systems development community should be aware of the opportunities presented by processing and networking innovations that are driven by high speed trading strategies. Just as military and aerospace drove IT innovation in the past, we can expect further applications of some of these new technologies in areas far from the financial industry, including other real-time Big Data applications.
5 REFERENCES


Bodek, Haim, 2013. 'The Problem of HFT - Collected Writings on High Frequency Trading & Stock Market Structure Reform'


6 ENDNOTES

1 Wikipedia article: http://en.wikipedia.org/wiki/High-frequency_trading
3 Dark Pools: High-Speed Traders, A.I. Bandits and the Threat to the Global Financial System, by Scott Patterson, July 2012
4 Nanosecond Trading Could Make Markets Go Haywire http://www.wired.com/wiredscience/2012/02/high-speed-trading/
5 http://www.economatters.com/2012/08/440-million-knightmare-wall-streets.html
11 http://www.hfreview.com/pg/blog/mike/read/6868/fpga-the-next-wave-of-hft-technology
12 http://www.hfreview.com/pg/blog/mike/read/6868/fpga-the-next-wave-of-hft-technology
16 Emulex white paper: High Performance Networking Made Easy with Emulex OneConnect Network Xceleration Solutions
18 Ravi Kalakota, "Practical Analytics" blog, http://practicalanalytics.files.wordpress.com/2013/01/regulatoryregimes.jpg
19 http://www.investopedia.com/terms/h/high-frequency-trading.asp
20 http://sec.gov/comments/s7-02-10/70210-129.pdf