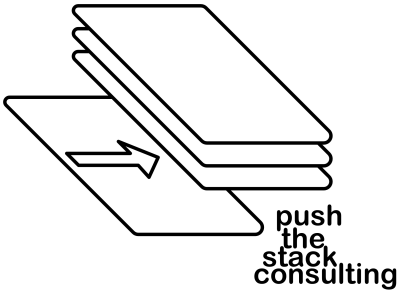
Security When Nanoseconds Count



a whitepaper on the security issues and
challenges with next generation finance and trading infrastructures
Blackhat USA Briefings 2011

Abstract

There's a brave new frontier for IT Security - a place where "best practices" do not contemplate the inclusion of a firewall in the network. This frontier is found in the most unlikely of places, where it is presumed that IT Security is a mature practice. Banks, Financial Institutions and Insurance Companies. High Speed Trading, High Frequency Trading, Low Latency Trading, Algorithmic Trading -- all words for electronic trades committed in microseconds without the intervention of humans. There are no firewalls, everything is custom and none of it is secure. It's SkyNet for Money and it's happening now.

Introduction

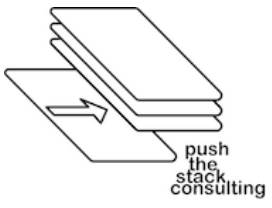
Throughout the course of modern financial times, technology has influenced the development and maturity of all markets, from the chalk boards and runners of the late 1700s to the current trend towards incredibly quick trades performed entirely within electronic systems without any human intervention.

The communications revolution of the 1800s brought about swift changes from the carrier pigeons used by Reuters in the early part of the century to the first telegraph based ticker systems of the 1860s. Advances through the early and mid 20th century lead to the introduction of computers as the trusted stores of data on trade pricing, volumes, opening and closing prices and more.

The electronic nature of stock markets became part of most people's general awareness with the opening of the NASDAQ exchange in the early 1970s and the move to the electronic small order execution system in the late 1980s.

High frequency trading likely started in concert with changes initiated by the US Securities and Exchange Commission in 1998 to permit electronic trading. The majority of the strategies in high frequency trading are related to time arbitrage – the ability to make (or lose) money based on minute differences in time between the time information is available and the time it is widely known.

The speed of trading, and therefore the available time in which to complete the calculations necessary for time-based arbitrage, has been shrinking at an accelerated pace over the last 50 years. With the advent of rapid communications, stock quotation systems and computer mediated order management, the time to complete a trade was brought from hours to minutes to seconds. The shift from seconds to 100s of microseconds has happened relatively quickly. The fact that a dollar value can be assigned to the length of time available for arbitrage is an indication that we're reaching the end-game on time-based arbitrage. Estimates on the value of a millisecond of unnecessary latency range wildly. Simply stated – never before in the course of human history has a microsecond been worth quite as many millions of dollars.



Low Latency Infrastructure Patterns

Understanding that the number of microseconds to complete the calculations necessary to take a position in the market with the intent to arbitrage in time is a fact, grasping what that means literally is still a very emotional and qualitative discussion.

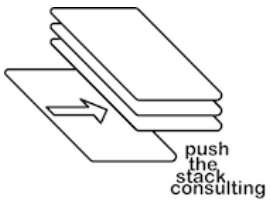
- ‡ If you are able to complete trades in seconds, you have no position.
- ‡ If you are able to complete trades in milliseconds, you lose nearly every time.
- ‡ If you are able to complete trades in 100s of microseconds, you're a bit player and missing a lot of action.
- ‡ If you are able to complete trades in 10s of microseconds, you're usually winning.

The predictability of the connection is nearly as important as the absolute latency value. If the connection is fast enough 80% of the time but too slow the other 20% of the time, you cannot ensure that your trades will result in the desired outcome – you have a 20% chance of failing on every single trade. This indeterminism is usually composed of jitter, packet-loss, and inefficient protocols (such as TCP.) Remember that a dropped packet is dropped cash.

This need for low latency and predictable performance drives a certain set of infrastructure design choices. These choices will include elements such as:

- ‡ Extreme systems (in 2011: 16+ core, 128GB, 10G-ether/Infiniband/PCIe interconnects)
- ‡ Custom hardware and software solutions which place the decision engine into FPGA processors
- ‡ Custom networking interfaces that bypass the kernel
- ‡ Extreme networks (cut through switching at 10GB/s)
- ‡ Proximity (the same data-centre as the exchange)

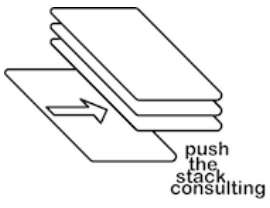
Many of these choices are made because interconnect technologies designed for wide area use have inherent inefficiencies and also because one of the increasingly popular interconnect technologies (PCIe) was originally intended for use within a single system and not as a data transport from system to system. The key issue for both latency and determinate data transport remains the speed of light.



Consider that light travels:

~300km (~186 miles) in 1 millisecond (1/1000th of a second)
~300m (~328 yards) in 1 microsecond (1/1000th of a millisecond)
~30cm (~1 foot) in 1 nanosecond (1/1000th of a microsecond)

These values start to place absolute limits on the distance between processor nodes (the trading engine and the exchange) in order to stay within the limits necessary to complete a time-arbitrage trade. This is not the first time that the speed of light has been an issue in data-centre design, but it may be the first time that there was a financial case for neat and tidy interconnect cabling.

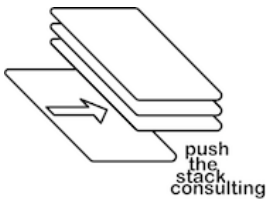


Variations from Common Practices

In order to meet the requirements dictated by the low latency / deterministic performance model identified above, a number of critical variations from common practice must be made.

For nearly all installations, the usual perimeter defensive mechanisms will be completely absent. You won't find a firewall, you won't see routers with ACLs, you won't see IDS and frankly, anything that you'd recognize as a security tool.

The essential reason that security devices are largely (if not wholly) absent from most implementations is that the best the IT Security industry can offer falls short. Most commercial firewalls process data and add a few milliseconds of additional latency. In the vast majority of interconnection scenarios, a few milliseconds isn't that much of a problem. In the case of low latency trading, it's about 100,000 times too slow. In addition to products which simply do not support this mode of operation, there's a skills gap in the practitioner space, the majority of IT Security workers in very large organizations which are utilizing low latency trading don't have the necessary background to implement some very old-fashioned and very basic network security while at the same time determine how to properly secure a host with custom *everything*.



Threat Models

Training staff with the necessary skills to do network security like it's 1999 and also the insight necessary to find, understand and communicate flaws in custom FPGA based network interface hardware is not a trivial exercise. And for this reason, the vast majority of installations simply skip security and rely on market-data providers and exchanges' commitments around the security of the network itself without real comprehension of the potential threats.

Developing an appropriate trust model should be trivial – we already know that we're missing our entire set of controls – but how to describe the real issue and how do we determine the most appropriate response?

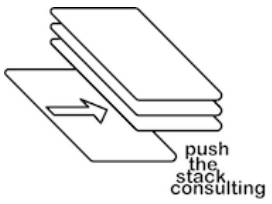
Start by managing the three largest threats:

The Developers – In most algo-trading, the developer isn't a traditional developer with all of the usual SDLC controls. The developer is probably a trader or trader's underling who has live access to the production algo engine and can make on the fly changes.

The Insider – This is not the "financial insider", but rather a trader or an administrator of a set of low-latency systems who is utilizing access to market data networks or exchange networks to cause negative effects of the other participants.

The Market Itself – This is an odd kind of technical threat – but as the other party in a communication, could the market cause issues with your systems? What about malformed messages? What about other participants with compromised systems?

As you build the picture of what your threat model should encompass, ensure that you consider that even very odd cases might actually be the common case. At the speed of transaction flow, can you really prevent things from occurring or should you build your threat model around post-fact detective controls?



Beginning the Solution

‡ *The journey of a thousand miles begins with a single step. ~ Lao-tzu*

From where most low-latency or algo trading systems are currently in terms of security, any change would be a positive change. Due to the need for speed, some of the techniques utilized in the late 90s are completely appropriate – bastion hosts, router acls, layer 3 and 4 firewall rules – assuming you are using sufficiently fast equipment – and top-of-rack switches are now available that offer sub-microsecond performance for cut through layer 4 switching.

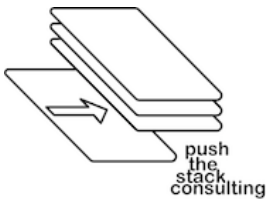
Even if you cannot implement any changes, it would be an improvement in security posture to have a complete architectural understanding of the systems as implemented. Situational awareness is important.

Product vendors – it's time to step up and give this market some attention. There's money to be had by individuals who want more than checkbox protection.

Risk / Process / Policy / GRC – work with the business, they understand risk – probably better than you do – but have a different set of tolerances. Use their knowledge to help make good decisions rather than blindly following dogmatic statements.

IT Compliance – meet the financial compliance people – I'm sure you'll find things to talk about. Finally.

Practitioners in the Trenches – research everything. Be prepared to operate at all levels simultaneously with reaction times that match your low-latency business partners. Work on proof of concept to see where you can and cannot actually help. And most of all, be prepared for the continued downward pressure on transaction times.



About James Arlen, CISA

James Arlen, CISA, is Principal at Push The Stack Consulting providing security consulting services to the utility and financial verticals. He has been involved with implementing a practical level of information security in Fortune 500, TSE 100, and major public-sector corporations for more than 15 years. James is also a contributing analyst with Securosis and has a recurring column on Liquidmatrix Security Digest. Best described as: "Infosec geek, hacker, social activist, author, speaker, and parent." His areas of interest include organizational change, social engineering, blinky lights and shiny things.