RESEARCH ARTICLE OPEN ACCESS

# Electronically aided High frequency trading

Dr. S. S. Limaye Principal,
Jhulelal Institute of technology, Nagpur 441111
shyam_limaye@hotmail.com

**Abstract—**
High Frequency Trading (HFT) is a kind of algorithmic trading on electronic stock exchanges where trading decisions are taken in a very short time-only a few milliseconds- in response to some stimulus in the system. The algorithms are different from those used for traditional buy and hold strategies or pattern recognition in technical analysis. The success basically depends on quick response to stimulus and hence, HFT firms use high speed communication links to stock exchange and high performance computing. There is increasing use of multiple processors of Graphic Processing Unit or massively parallel special hardware implemented on FPGAs. This paper offers a review of the technological challenges and market risks.

***Keywords—***algorithmic trading, HFT, FIX, FPGA, GPU

## I. INTRODUCTION

In old days, stock market floor was populated by hundreds of jobbers (called market makers in USA) and trading took place in open outcry mode. To overcome the din and chaos, they received visual clues from brokers in a strange sign language. Market price was determined by jobbers depending on the relative buy and sell pressures. Nowadays markets are made electronically with bid and ask queues maintained in a computer for each scrip. Earlier buy and sell decisions were made manually through fundamental analysis and technical analysis. Fundamental analysis is based on profitability projections based on anticipated economic, technological and political factors and it is valid for long range price prediction. When a fundamental factor for a share changes, it starts ripples in the price of that share and many other shares too. Technical analysis is based on study of these ripples. It detects patterns in the price movements and generates buy-sell advices.

People have been using computers since 70's for arriving at buy-sell advices based on the market data. It got a boost in 1992, when NASDAQ in New York introduced FIX protocol for trade communication. It was then adopted by several other stock exchanges and now it has become the de facto messaging standard in the global equity markets. With this protocol, it is possible to see the order book, get research data and execute the order automatically, without any manual intervention. This process is called 'Algorithmic trading'**.** Today about 70% of trade in USA is algorithmic.

High-frequency trading (HFT) is a specialized form of Algorithmic trading, where the execution of computerized trading strategies is characterized by extremely short position-holding periods – just a few seconds or even down to milliseconds. The success of an HFT algorithm depends on its ability to react to a situation faster than others. This has given birth to another variant of HFT called Ultra High Frequency Trading (UHFT). Here, the execution of trades happens in sub-millisecond times. The technology used by UHFT traders is co-location of servers with exchange, direct market access, using parallel processing on GPUs and using special hardware like FPGAs.[1].

## II. Overview of HFT components

### A. **Information infrastructure**.

Consolidated Tape Association(CTA) oversees the collection, processing and dissemination of consolidated quote and trade data at NYSE. Securities Information Processor(SIP), is the technology that enables collecting quote and trade data from the exchanges, consolidating it, and sending it out as a continuous stream of best bids and offers (quotes) and last sales (trades). SIP has to work at enormous speed. On average, NYSE handles average 2 lakh quotes per second out of which 28000 per second get converted into trades[2]. The traders talk to the exchanges using FIX protocol. FIX stands for Financial Information eXchange. The standard is managed by a nonprofit organization called FIX Trading Community. The message consists of ASCII characters and the format is an extension of XML, called FIXML[3]. Recently Citibank has announced that it will provide FIX functionality to NSE in India.

### B. **High speed data links**.

The links between traders and stock exchange have high throughput, typically 100Gbps. Also they have low latency. People are stretching the limits of physics to reduce latency. In 2010, traders in Chicago switched to a new fiber optic link built at the cost of 300 million dollars. The old line went along railway lines in a zigzag way and the signal took 16 miliseconds for a round trip. The new line was laid straight and it reduced the round trip time from 16 to 13 miliseconds[3]. People are trying to use fibers with low refractive index so that it will have faster speed of light. The new trend is to use free to air channel using lasers or microwaves. Such speed races are taking place everywhere. A microwave link between London to Frankfurt has reduced the travel time by 40% compared to optical cable[4].

C. **HPC hardware**. There is increasing use of High Performance Computing platforms like GPU multiprocessing and FPGA.

D. **HFT algorithms**. They are fast and parallelizable. They are specifically designed to make money by exploiting tiny, lightning-fast price changes in shares.

## III. HFT STRATEGIES

Traditional trading strategies are known as fundamental, technical and quant. HFT strategies are different.

**Strategy 1-Market microstructure**

Traders can place two types of orders for the securities listed on the exchange - limit orders or market orders. Limit orders are characterized by a triplet - action (buy or sell), quantity and price. e.g. buy 400 shares of Infosys at Rs 3350 or less is called a ‘bid’ order and sell 500 shares of Infosys at 3360 or more is called a ‘ask’ order. There is a gap of Rs 10 between the expectations of bidder and asker. This gap is known as ‘bid-ask spread’. All bidders for 3350 level are entered in a FIFO queue. Similarly all askers at 3360 level are also entered in a FIFO queue. Suppose these are the best bid and best ask offers. Now there are other bid and ask offers which are not as good as best offers. They are also queued up. For example, there may be a bid for 1000 shares at Rs 3345 and an ask for 2000 shares at Rs 3362. These are also entered in the queue. The collection of bid and ask offers is called the limit order book and it appears as shown in figure 1.

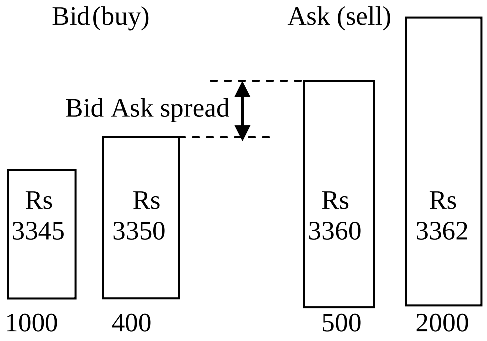

Figure 1
Limit order book

Mid point of best bid and best ask may be considered as the prevailing market price. The probability of upward movement pUP depends on the bid ask imbalance and is given by

$$pUP = \frac{bs_0}{as_0 + bs_0} = \frac{400}{400 + 500} * 100 = 44\%$$

Where $bs_0$ is best bid size and $as_0$ is best ask size. Some people take a weighted average of the entire limit order book instead of the best bid and best buy sizes. Similarly the probability of downward movement pDN is given by

$$pDN = \frac{as_0}{as_0 + bs_0} = \frac{500}{400 + 500} * 100 = 56\%$$

A trader who is waiting in the queue can either hope for the better offers to be wiped off and opposite offer match his expectation or jump the queue by offering a better price. If he places a market order, then he will be immediately allotted the best offer from the other side.

A trader can anticipate a bull run, based on the brief history of pUP and pDN. He can then place a market order to buy, hold it till the price rises and then sell. This is called ‘squaring up’ a transaction. He can make profit if the price rise is sufficient to cover the bid ask spread twice and also take care of transaction costs. The position is held for a very short time, sometimes as short as a few milliseconds. HFT traders reverse their positions hundreds of times in a second.

**Strategy 2- Correlated securities**.

There are many related pairs like gold and gold ETF or mutual fund and its chief constituent scrip. The relationship can be modeled by linear regression analysis. When one of the securities is found to be underpriced according to the regression relation, it can be bought before others make the discovery. Arbitrage is a special case of this strategy where prices of same security are compared in two markets. It is bought from the market offering lower price and immediately sold in the market offering higher price. Eric Budish[5] has demonstrated that correlation holds good when prices are averaged over human response timeframes i.e. days or hours. At minute level it starts breaking down and at 250 ms level, it completely breaks down. HFT traders take advantage of the momentary loss of correlation by acting fast.

**Strategy 3 – Responding to events**
Certain announcements are routinely made by various government agencies and business associations. These could be corporate results announcements, labor employment and wage statistics or car sales figures from automobile manufacturers' association. Figure 2 shows the response of share price to a good news.

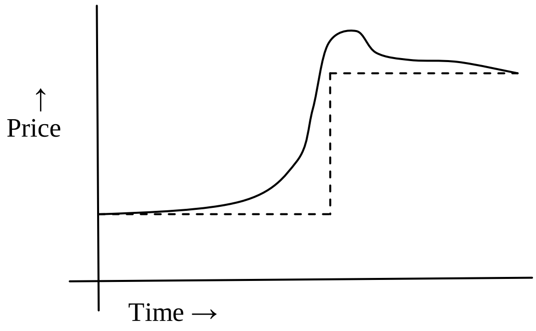

Figure 2
Market response to a good news

The dotted line shows the theoretical response for anticipated change in EPS of the share as a result of the good news. As the news trickles down to the investors, the price starts rising, as shown by the solid line. It overshoots the correct price and then stabilizes. HFT traders electronically decipher the news feed and act before the others.

## IV ONE PASS ALGORITHMS

All HFT algorithms have to solve three basic problems before generating trade orders. They are: moving average, variance and regression. Given the high rate of input data, "One pass algorithms" are used [6]. Unlike standard algorithms, these algorithms use a data point only once. This makes them fast and memory efficient. The entire program and data can reside in L1 cache. It also makes FPGA implementation easy.

1. Moving average. We need to calculate it for various parameters. For example we need a running measure of liquidity to determine the size of an order that is likely to execute successfully. In technical analysis, moving averages with a long window is often compared with moving average with a small window to predict future trends. An N-element moving average system requires N memory locations for storing past N values. Memory can be saved by using a first order recursive filter. For estimation of running mean m(n) of a time series x(n) we use the following equation
   $m(n) = (1-\alpha)\, m(n-1) + \alpha\, x(n)$
   it is an average of infinite number of past values with exponentially reducing weights. That is why it is also called Exponential smoothening.
2. We need to get a running estimate of volatility to quantify the short term risk of a position. The difference between the mean m(n) calculated above and current sample is used to estimate variance v(n), which is the measure of volatility.
   $v(n) = (1-\alpha)\, (x(n) – m(n))^2 + \alpha\, v(n-1)$
3. We need to fit a straight line between prices of closely related securities by regression analysis. We use a recursive version of least square error method. Suppose N pairs of the independent variable x(i) and the dependant variable y(i) are available. The relation between x and y can be estimated as
   $y = \beta 0 + \beta 1\, x + \varepsilon$
   Where β0 is offset, β1 is slope and ε is random error.
   In matrix form, we can write:

$$\begin{pmatrix} y(0) \\ y(1) \\ --- \\ --- \\ y(N-1) \end{pmatrix} = \begin{pmatrix} 1 & x(0) \\ 1 & x(1) \\ --- & \\ --- & \\ 1 & x(N-1) \end{pmatrix} \begin{pmatrix} \beta 0 \\ \beta 1 \end{pmatrix} + \begin{pmatrix} \varepsilon(0) \\ \varepsilon(1) \\ --- \\ --- \\ \varepsilon(N-1) \end{pmatrix}$$

i.e. $Y = X\, \beta + \varepsilon$

The regression parameters β0 and β1 are estimated as
$\beta = (X^T X)^{-1} X^T Y$
We can construct a one pass version of the algorithm as proposed in [6]. We need to construct a 2X2 matrix M and a 2X1 matrix V. These matrices need to be updated with each data point $X_t = \{1\ x_t\}$ and $Y_t = \{y_t\}$ as follows.

$$\begin{pmatrix} m_{00}(n) & m_{01}(n) \\ m_{10}(n) & m_{11}(n) \end{pmatrix} = \alpha \begin{pmatrix} m_{00}(n-1) & m_{01}(n-1) \\ m_{10}(n-1) & m_{11}(n-1) \end{pmatrix} + \begin{pmatrix} 1 & x_t \\ x_t & x_t^2 \end{pmatrix}$$

This can be written in a compact form as
$M(n) = \alpha\, M(n-1) + X_t^T X_t$
The vector V is updated as:

$$\begin{pmatrix} v_0(n) \\ v_1(n) \end{pmatrix} = \alpha \begin{pmatrix} v_0(n-1) \\ v_1(n-1) \end{pmatrix} + \begin{pmatrix} y_t \\ x_t y_t \end{pmatrix}$$

In matrix notation,

$V(n) = \alpha\ V(n\text{-}1) + X_t^T\ Y_t$

We can observe that $m_{01}$ and $m_{10}$ are an estimates of $\Sigma x$ , $m_{22}$ is an estimate of $\Sigma x^2$, $v_0$ is an estimate of $\Sigma y$ and $v_1$ is an estimate of $\Sigma xy$. The regression coefficients $\beta$ can be calculated as

$$\beta_t = M_t^{-1} V_t$$

## V HARDWARE

The hardware must provide massively parallel operation, must be flexible and will have limited production volume.

FPGA is a good solution to achieve this. The FPGA boards can be interfaced to PC through PCI slots. Nallatech company provides PCI compatible boards with ALTERA chips. Public domain platforms like OPENCL can be used for integration and software development. The web site OPENCL.org provides useful guidelines for this purpose. Another alternative is to use a Graphics Processing Unit (GPU). GPU was originally developed to handle 3D animation in real time. It has hundreds of processors and they can be used for general computing if we are not using them for animation. GPU chips from NVIDIA company have become very popular. NVIDIA has also developed an open platform called CUDA for software development.

The hardware can be divided into three major blocks.

**1 Input parser**. It parses the data feed and passes the data to the concerned strategy engine. Most stock exchanges use a protocol named Fix Adapted to Streaming, i.e. FAST. FAST packets are sent through UDP protocol. Heiner Litz [7] has described an FPGA based engine that receives Ethernet packets. It decodes the Ethernet layer, IP layer, UDP layer and FAST layer and presents the data in binary form. It has a latency of only 2.4 µs.

**2 Strategy engine.** This forms the core strategy as described in section 3. It may be implemented either in FPGA or in GPU. Mohammad Sadoghi [8] has presented an FPGA based event processor.

**3 Order processing engine.** It receives buy or sell signals from the strategy engine and sends the orders to stock exchange in FIX format.

## VI PROBLEMS OF HFT

HFT strategies purely based on market microstructure can amplify small blips to dangerous proportions.

Gordon Baxter[9] has noted that a great dip called Flash Crash happened in the USA on May 6th, 2010. The US's Dow Jones Industrial Average Index was down by 300 points on the day, but then it fell a further 600 points in five minutes between 14:42 and 14:47, effectively wiping $1 trillion from the value of the market. By comparison, The total cost of the September 11 attack on the World Trade Center—comprising earnings losses, property damage, and the cleanup and restoration of the site—is estimated to be between $33 billion and $36 billion. After investigations it turned out that the crash was triggered by a large (75000) sale order of E-mini S&P contracts which exhausted the number of available buyers. This was followed by HFTs aggressively selling.

Baxter further notes that Knight Capital was a successful stock investing company. On 1st August 2012, it started live trading using their new Retail Liquidity Provider (RLP) market making software on the NYSE. Immediately they started losing millions of dollars a minute. It was forty-five minutes before the software was stopped, by which point Knight had lost a total of $440 million.

There are social problems associated with HFT. Richard Finger in his blog at Forbes.com has commented that HFT represents a dark force against ordinary investors. In March 2013, Getco, one of the world's largest automated trading firms, became the first Western trading firm to gain approval to trade in India, the world's fastest growing derivatives market. It is likely that it will have an unfair advantage over local firms, at least initially. But in the long run it is likely that Indian techies will outsmart everyone else. Eric Budish [5] describes the clamor for acquiring HFT technology as arms race.

## VII REMEDIES

After the flash crash, NYSE introduced a mechanism of 'circuit breaker'. It halts the trade of a stock if the price changes by more than 10% in 5 minutes. During this period, market cools off and reaches equilibrium. All stock exchanges in the world now use similar circuit breakers.

Eric Budish [5] proposes uniform-price sealed-bid double auctions conducted at discrete time intervals, e.g., every 1 second as a market design alternative to continuous limit order books.

## VIII CONCLUSION

High frequency trading is new but growing phenomenon. It provides interesting research opportunities in Financial management, market dynamics, FPGA hardware, multicomputing on platforms like CUDA and legal and ethical aspects.

**References**

[1] Philip Treleaven, Michal Galas, and Vidhi Lalchand, "Algorithmic Trading Review" Communications of the acm | november 2013 | vol. 56 | no. 11 pp.76-85

[2] Clark, C. "Improving speed and transparency of market data", https://exchanges nyx.com/cclark/improving-speed-and-transparencymarket-data. 19 Jan 2011

[3] http://en.wikipedia.org/wiki/FIX_protocol

[4] http://www.cnbc.com/id/100695563, 1 May 2013

[5] Eric Budish, Peter Cramton, and John Him , "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response. http://home.uchicago.edu/~shim/Papers/HFT-FrequentBatchAuctions.pdf, 17 Sept 2013

[6] Jacob Loveless, Sasha Stoikov, and Rolf Waeber "Online Algorithms in High-frequency Trading" Communications of the ACM. vol. 56 , no. 10, | OCTOBER 2013. Pp 50-56.

[7] Heiner Litz et al, "DSL Programmable Engine for High Frequency Trading Acceleration", Proceeedings of ACM Workshop on High Performance Computing WHPC 11 Nov 2011, pp 30-38

[8] Mohammad Sadoghi "Efficient Event Processing through Reconfigurable Hardware for Algorithmic Trading", Proceedings of the VLDB Endowment, Vol. 3, No. 2 pp 1525-1528

[9] Gordon Baxter et al. "Flying by the seat of their pants",Third International Conference on Application and Theory of Automation in Command and Control Systems (ATACCS13) May 2013 pp 64-73.

.