An edited version of this article was published in the *London Review of Books* on 11 September 2014 (vol. 36, no. 17, pp. 27-30). http://www.lrb.co.uk

Be grateful for drizzle

Donald MacKenzie

Although you can't see their beams, which are infrared, lasers are starting to flash stock market data through the skies over New Jersey. If they work well there, lasers might also soon be flashing over London. They're the latest tool for 'high-frequency trading': the fast, entirely automated trading of large numbers of shares and other financial instruments.

Originally, the data needed for high-frequency trading travelled almost exclusively via fibre-optic cables, in which signals move at about two-thirds of the speed of light in a vacuum. It's a tried, trusted technology. Fibre has the bandwidth (the capacity) to transmit the huge volumes of data spewed out by today's financial markets. Although accidents do happen (farmers, for example, sometimes cut the buried cables when ploughing), fibre-optic links are very reliable.

But two-thirds of the speed of light isn't fast enough if other market participants are faster. So there's a race on to get as close as feasible to what the theory of relativity posits as the maximum speed physically possible, that of light in a vacuum. Unlike glass fibre, air slows light down only almost imperceptibly. Sending light or radio signals through the atmosphere is less than a tenth of a percent slower than the speed of light in a vacuum.

All the ways of sending market data through the air have limitations. Improved versions of an old technology — wireless microwave transmission, using 'dishes' (antennae) on a series of towers — gets you the necessary speed. Although

gales can blow the dishes out of alignment, and make it unsafe for workers to climb the towers to re-align them, those involved tell me that a good microwave link will work well over 99 percent of the time. Its bandwidth, however, is limited, meaning that streams of financial data usually need editing: 'you have to pare down the data', removing what isn't necessary, said a microwave user. Specialised computer hardware can do the necessary editing, compression and decompression within a couple of microseconds (millionths of a second), but delays measured even in nanoseconds (billionths of a second) now matter.[1]

Higher-frequency millimetre waves provide larger bandwidth than microwave, reducing the need for 'editing', but are more easily disrupted by bad weather. Three years ago, millimetre-wave links "would go down when someone sneezed," says Michael Persico, whose Chicago firm, Anova Technologies, deploys these links. Even today, they tend to stop working if it starts to rain heavily.

So Anova has begun to deploy lasers to supplement its millimetre waves. Lasers also have vulnerabilities (fog is a big problem), but Anova says its test results show that a combination of lasers and millimetre waves can achieve reliability comparable to that of fibre-optic cable, while getting close to the holy grail of the speed of light in a vacuum.

A decade or so ago, there was a vogue for asserting that globalisation had 'ended geography' and created a 'flat world'. When financial trading was a matter of human beings looking at screens, that had a certain plausibility, because the inevitable slowness of humans' eyes and brains can easily mask the brief delays that result from different geographic locations. But with computers, not humans, doing the trading,

[1] When I first wrote about high-frequency trading in the LRB of 19 May 2011, nanoseconds still seemed marketing hype. No longer.

geography matters exquisitely. With any of these technologies — fibre-optic cable, microwave, millimetre wave, laser transmission through the atmosphere — the exact route is crucial.

The shortest distance, and therefore the fastest route, on the surface of the earth between any two places is what's called the 'geodesic' or great-circle route. If you talk to high-frequency traders (something I've done a lot over the past four years), you quickly learn that the world's financially most crucial geodesic — the spinal cord of U.S. capitalism — runs from Aurora, a town in Illinois that's now basically an outer suburb of Chicago, to northern New Jersey. It's now close to impossible, I'm told, to put new microwave dishes along that crowded geodesic: all the space on the towers closest to the geodesic is already rented. The dishes can interfere with each other, so you need permission from the Federal Communications Commission to build new towers or install new dishes. As a result, no-one is easily going to beat McKay Brothers, which owns the fastest, geodesic-hugging Aurora-New Jersey microwave link.

Aurora matters to global finance because in 2012 the Merc, the Chicago Mercantile Exchange, relocated its electronic trading system to a new data centre there. The Merc trades futures: at first, futures on eggs, onions and other agricultural commodities, but since 1972 financial futures as well. Originally, Chicago futures trading was done face-to-face (by voice, or eye-contact and hand signal) in raucous, crowded trading pits. The Merc's pit traders fiercely resisted the coming of electronic trading: its leading advocate, Leo Melamed, received frequent death threats. By 2004, however, that resistance had crumbled, and now nearly all the Merc's trading is electronic.

The Merc's first fully electronic product was the E-Mini financial future, launched in 1997. It tracks the S&P 500 index, made up of the 500 leading U.S. stocks. The buyer and the seller of an E-Mini each maintain a deposit known as 'margin' on account at the Merc's clearing house. Every night, the clearing house adjusts those deposits. If the S&P 500 index has risen by a single point, $50 is transferred from the seller's account to the buyer's; if it has fallen by ten points, say, $500 shifts from the buyer to the seller. If their deal is for a thousand E-Minis, the latter sum becomes $500,000. (The contract is called the 'Mini' because these transfers were five times larger for the contract that traders called the 'big', the corresponding pit-traded S&P 500 future.)

Traders tell me that new information relevant to the overall value of U.S. shares almost always shows up first in orders for and in the prices of the E-Mini, and only a fraction of a second later in the underlying shares; financial economists have also documented that pattern. The reason for it is probably that the E-Mini gives you greater 'leverage': a modest 'margin' deposit permits gains (and, of course, also losses) corresponding to buying or selling a large and expensive block of shares. So if traders think they or their automated trading systems have an information edge, it is to the E-Mini they will usually turn first. For example, the big crises of modern US stock markets have tended to show up first in the E-Mini (or, before 1997, in its predecessor, the S&P 500 pit-traded future) and only a little later in the stock market.

Changes in the electronic order book for the E-Mini are crucial information for automated share trading (this particular game is now too fast for human players, who wouldn't be able to react quickly enough to those changes). Suppose the price of the E-Mini has fallen, or even simply that the number of offers (sell orders) has risen sharply and the number of bids (buy orders) has fallen. Over the next fraction of a

second, falls in the prices of the underlying shares are more likely than increases. I'm told that a similar pattern generally holds for U.S. Treasury bonds, with the prices of the bond futures traded in Aurora leading changes in the prices of the underlying bonds (although the bonds do sometimes lead the bond futures). Foreign exchange, I'm informed, is different. Although Aurora is the world's main site of currency futures trading, the dominant foreign-exchange markets are still the 'spot' markets in which foreign exchange for immediate delivery is traded among banks, hedge funds, and so on.

US shares, US government bonds and much (although by no means all) of 'spot' foreign exchange are traded in data centres in northern New Jersey. Data centres resemble giant warehouses, and their size explains why this trading has shifted from its traditional sites in Manhattan to townships in New Jersey no tourist has ever heard of: real estate is much cheaper there. Data centres often have high-security features, such as a two-door entrance like a spaceship airlock. They're frequently windowless, and sometimes freezingly cold because of the need for fierce air conditioning to extract the heat generated by the tens of thousands of computer systems they contains. (The small numbers of maintenance workers who are needed can stay in warm rooms unless something goes wrong or new equipment needs installed.) Data centres are huge consumers of electricity, and while a single modern computer is close to silent, the combination of tens of thousands packed together and all the air conditioning makes for a lot of noise. Andrew Blum, author of a fine book on the physical reality of the internet, describes the experience of visiting a data

centre as 'like stepping into a machine, all rush and whir. … as loud as a rushing highway'.[2]

There are four main share-trading data centres in the US. The New York Stock Exchange owns its own data centre in Mahwah, New Jersey. NASDAQ's is in Carteret. The computer systems of their biggest rival, BATS (Better Alternative Trading System, set up in 2005 by a team from the Kansas City high-frequency trading firm Tradebot) are currently in NJ2 in Weehawken, owned by data-centre managers Savvis. The share-trading systems of the fourth main exchange in the US, Direct Edge (recently bought by BATS), are in NY4, a giant multi-user data centre owned by Equinix, the largest global provider of such centres. Despite its name, NY4 is actually in Secaucus, New Jersey, and much of the trading of bonds, foreign exchange and options also takes place there.

Each of these datacentres contains the exchange's or other trading venue's matching engines, which are the computer systems that maintain its electronic order books, search for a match (a bid and an offer at the same price for the same share, bond or currency) and, if they find one, consummate the trade and generate electronic 'confirms' to tell the parties to it that it has taken place. Surrounding the matching engines are the order gateways, further computer systems that receive the electronic messages containing traders' orders and their cancellations of orders, send those messages to the correct matching engine, and despatch 'confirms'. An exchange will also have one or more 'feed servers', which collate information on every change in its order book and send it out in a continuous stream of messages referred to as the 'raw' datafeed.

---

[2] *Tubes: Behind the Scenes at the Internet* (Penguin, 2012).

Nowadays, the majority of orders for US and European shares and futures (and increasing proportions of the orders for bonds and foreign exchange) are generated by computer systems, also often located within the data centre in question. High-frequency trading firms — and also, for example, the big banks that act as brokers for institutional investors — pay the exchanges or other data centre owners for space in which to place their computer servers. I'm told that a 'rack,' a cabinet-sized space that can accommodate around 30-40 computers, costs from around $1,500 to $15,000 a month. A big trading firm will need multiple racks in each datacentre, perhaps a whole 'cage', which can easily cost well over $1 million a year. (To protect against rivals' tampering, trading servers are often kept in locked cages.) It's possible for the owners of data centres to charge these high rents because it's a very concentrated world. Most of global financial trading probably takes place in no more than around 15 data centres: six in the US (the four big share-trading centres, the Merc's in Aurora, and a big multiuser centre closer to the Chicago Loop); another five in Europe; and a handful in the rest of the world.

The computer programs of a bank that is executing orders on behalf of an institutional investor will have a goal set by the latter, for example to buy 100,000 Apple shares. If those programs immediately and visibly placed the entirety of that order in the electronic order books of the exchanges on which Apple shares are traded, it would drive up their price sharply. So the bank's programs (its 'smart order router' and its 'execution algorithms') will split that 'parent' order into as many as a thousand 'child' orders, and execute them as surreptitiously as possible, often trying to make them look like large numbers of independent orders from small investors. The order router will usually at first send child orders to 'dark pools' (which are not exchanges, but trading venues in which participants, whether human or algorithmic,

cannot see the electronic order books). If the router then has to send child orders to exchanges such as BATS, Direct Edge, NASDAQ, the New York Stock Exchange – in which order books are visible – it will keep them small and therefore hopefully inconspicuous, buying or selling as few as 100 Apple shares (or even fewer) at a time. All this splitting and routing of a big order is complicated but eminently automatable. It's of course far cheaper to have computer systems do all of this than to pay human traders.

The goal of high-frequency trading (HFT) programs, running on computer servers inside data centres, is simply to trade profitably, ideally without accumulating too large, and therefore too risky, an inventory of the futures, shares, bonds or currencies being traded. The designers of HFT programs usually want them to end the trading day 'flat': with no inventory whatsoever.

Those involved describe two broad ways for HFT program to make a profit. The first is what's called 'market-making'. If a market-making program is trading Apple shares, for example, it will continually post, in the order book of the exchange or other trading venue in whose data centre it is running, competitively-priced bids to buy Apple shares and offers to sell them at a marginally higher price. The goal of market-making is to earn 'the spread', in other words the difference between those two prices — a few cents; in the case of many other shares, a single cent — together with the small payments (around 0.3 cents per share traded) known as 'rebates' that exchanges make to those who post orders that other traders execute against. The goal of these rebates is to encourage market-making, with the hope that competition amongst market-makers will encourage lots of keen pricing and thus attract other participants to the exchange.

The other way that a HFT program can seek to trade profitably is to be one of those other participants: it can ‘hit’ a bid or ‘lift’ an offer that is already in the order book. That’s often called trading ‘aggressively’. It’s more expensive than market-making: you have to pay the exchange a fee, rather than earning a rebate, and if prices don’t move your program can end up simply ‘paying the spread’ to market-making programs, because it will have to sell more cheaply than it buys. However, if a HFT program can identify a trading opportunity that’s larger than the ‘spread’ (a high probability that, for example, the price of the shares being traded is going to rise or fall by several cents), then it may well need to act immediately and aggressively, before other programs can act. There’s quite a bit of circulation of staff among HFT firms, and people in those firms tell me that in consequence the chief ways of identifying trading opportunities are common knowledge across multiple firms.

How can a computer program identify a profitable trading opportunity? Both macroeconomic and company-specific news is now often issued in machine-readable form, and a well-programmed machine can act in response to it at least a quarter of a second faster than a human being can. There’s also increasing interest in using automated analyses of social media as a source of trading advantage. However, mainstream high-frequency trading programs rely primarily on two other sources of information.

The first source of information is the contents of the order book of the particular exchange in whose data centre the HFT program is running. As I’ve said, the exchange’s feed server continuously broadcasts changes in the order book (new orders, cancellations of orders, consummated trades). In share trading, it’s not the only way of keeping track of the market — you can also subscribe to the official, multi-exchange ‘consolidated tape’ — but the raw datafeed direct from the feed

server is fastest, so any HFT program needs access to it and its owner must be prepared to pay the exchange for that access. I'm told, for example, that the Merc charges $10,000 per month for access, but also warned that that's cheap: other exchanges often charge more.

Prediction using the contents of its 'local' exchange's order book for the shares it is trading will involve a HFT program monitoring the raw datafeed to keep track of whether that order book contains more buy orders than sell orders, or vice versa, and whether the numbers of either are changing fast. Sometimes, too, a program may be able to detect in the order book a distinctive pattern in the arrival of new orders that might indicate that a big institutional investor is buying or selling a big block of shares.

In the early days of high-frequency trading, when banks' execution algorithms did dumb things like send in a child order exactly every 60 seconds, detecting such a pattern was often easy. Now, it's more of a 'dark art'. Who does it, with what success, and how are very difficult questions to answer. Even those who do it may not fully know whether or not they're doing it. In a field of complex electronic interactions, it may be hard to distinguish between a program that is successfully identifying and exploiting patterns of orders that result from the splitting of one big order, and a program whose success is based on less specific order-book patterns.

Unquestionably, though, the more basic techniques of making predictions based on the balance between bids and offers in the order book are well known to all HFT firms. The resultant need for speed is a major reason why HFT firms all have to pay for the fast, raw datafeed. They all must also rent space for their computer servers in an exchange's data centre, because they need to be able to receive that datafeed as

quickly as possible and have the orders they place in response to it — along with their cancellations of orders that have been rendered 'stale' (mispriced) by changes in the market — arrive at the matching engine with minimum delay. Some data centres, such as the Merc's, don't allow you a spatial advantage within the data centre: if your cage is physically closer to the order gateways, the cable that connects your servers to them is coiled so that it is the same length as others'. In other data centres, I'm told, you can get closer by paying a higher tier of rents. You'll also pay more if you want a '10 gig' — 10 gigabits, or ten billion binary digits, per second — connection to the order gateways, not just 1 gig.

The second source of information used by a HFT program is order books other than that of the exchange in whose data centre it is running. That's why HFT firms need the fastest possible communication links between data centres. If, for example, a HFT program is trading Apple shares in NASDAQ's data centre in Carteret, it will — for the reasons discussed above — most likely need access to a microwave link to Aurora. It will almost certainly also need access to the order books of the other exchanges on which Apple shares are traded. The firm that owns it doesn't need itself to organise that access: NASDAQ, for example, will sell you it to you. According to the most recent price list I've been able to find, the transmission time you can expect over the millimetre-wave link from BATS in Weehawken to NASDAQ in Carteret is around 105 microseconds, and that costs each firm using the link $7,500 a month. The link from Direct Edge takes 101 microseconds and also costs $7,500 a month; a new link from the New York Stock Exchange is expected to take 190 microseconds and cost $10,000 a month.

Those links are needed because, with the trading of shares in the US spread across multiple venues, a change in one venue's order book often presages changes in

others'. Here, one has to start looking at maps to follow what's going on. Michael Lewis's new book about high-frequency trading, *Flash Boys*, recounts what Lewis was told by his main informant, Brad Katsuyama, who worked in the Royal Bank of Canada's offices in New York.[3] Katsuyama would, for example, often buy a large block of shares from one of the bank's institutional-investor customers, and then need to sell those shares in smaller blocks. He would frequently try more or less simultaneously to hit multiple bids or lift multiple offers for those shares that were in the order books of several different exchanges. This would work for BATS – his orders there were typically executed in full – but by the time his orders reached other venues the bids he was trying to hit or offers he was trying to lift would have vanished from the order book. 'The market as it appeared on his screens was an illusion', says Lewis.

Perhaps the most vehement complaint against high-frequency trading is just that: that it leads to a vanishing market, a market that disappears as soon as you start to try to trade. You can begin to understand Katsuyama's eventual explanation of it if you take a taxi from Manhattan to Newark Airport. Quite likely, you'll be driven through the Lincoln Tunnel underneath the Hudson River. When you emerge into New Jersey, the first place you come to is Weehawken. The main fibre-optic cables from Manhattan to New Jersey follow the same route as you, through the Lincoln Tunnel.

The office in which Katsuyama worked was in Manhattan, like those of most of the big banks that operate in the US as stock brokers. So Katsuyama's orders were reaching NJ2 in Weehawken, where the BATS matching engines are located, before they reached the data centres of the other exchanges. Somehow, HFT systems trading

[3] John Lanchester reviewed *Flash Boys* in the LRB of 23 May.

in these other data centres were learning of his other orders before they arrived, and buying, selling or adjusting their price quotes, thus causing the market Katsuyama could still see on his screen to have disappeared in reality.

I'd heard such complaints myself, and I admit I'd been sceptical, because they seemed to violate the laws of physics. A HFT system in the NASDAQ data centre in Carteret, for example, can't learn by magic about a child order that has arrived at BATS in Weehawken. There are essentially two ways it can find out: from the BATS raw datafeed once that arrives in Carteret, or from a message from one of the same firm's servers in Weehawken, which has learned that one of its bids or offers has been executed against. Even with a millimetre wave or laser link, all of that takes time: more time than it would take for the child order sent to Carteret to get there if it was travelling on fastest fibre-optic route.

I was put right by a quiet-spoken physicist I'd met, who knows a lot about speeding up communications. Mistakenly, I'd been assuming that banks would know how to find the fastest cables, and would ensure that their own orders and orders from their institutional-investor customers travelled down them. I now think I was wrong.

Michael Lewis's *Flash Boys* has been much debated, but no-one seems to have focussed on its crucial sentence, which comes on page 49. The sentence reports the difference in time taken between an order sent from the Royal Bank of Canada's office in Manhattan to BATS in Weehawken and one sent to NASDAQ in Carteret: around two milliseconds (two thousandths of a second). That sounds tiny, and indeed no human being could perceive a time interval that short, but in the world of New Jersey's high-frequency trading it's an eternity. Two milliseconds is ample time for news of the arrival of the order in Weehawken to reach HFT systems in Carteret. It's

around ten times longer than you should be able to achieve in a fibre-optic cable that is optimised for speed and follows the most direct route.

Lewis's *Flash Boys* is widely being read as a morality play, a story of evil-doing high-frequency traders. But it can equally be read as an account of banks that either wouldn't, or didn't know how to, take best care of their own or their institutional-investor customers' orders. To their credit, the Royal Bank of Canada team took action once they realised the disadvantage they were under, a process that is the main theme of Lewis's book. The extent to which other banks have done so is, however, unclear. Behind orders from banks' institutional-investor customers are people's savings and pension funds. It's right to care about what happens to them, but we need to be careful in allocating blame. (I suspect that the worst problems aren't high-frequency trading or even incompetent handling of orders by banks, but the high fees charged by those who manage pension funds and other savings, and the excessive trading they often engage in.)

Is there an equivalent to the Lincoln Tunnel in Europe: a geographical quirk that creates exploitable predictability in the arrival of orders at different data centres? I don't yet know, but both regulators and those responsible for handling customers' orders need to pay careful attention to Europe's geography of automated trading, which is an issue I've never seen discussed in public.

As I've said, there are five main trading data centres in Europe. Two are in London: the London Stock Exchange's centre and the Reuters foreign-exchange trading centre, which is in Docklands. (I'm told that, unusually, trading firms originally couldn't place their servers in the building containing the Reuters matching engines, which made locations close to that building very valuable real estate.)

Two other data centres are just outside of London. Equinix's LD4, in Slough, hosts BATS Europe and one of the three global sets of matching engines of EBS, the main rival of Reuters as a foreign-exchange trading venue (EBS's other two sets are in NY4 and in Tokyo). A data centre in Basildon, now owned by the US-based IntercontinentalExchange, contains the matching engines for LIFFE (the London International Financial Futures Exchange) and for much of the trading of continental European shares. Europe's fifth main financial data centre is Equinix's FR2 in Frankfurt, which hosts the matching engines of Eurex (Europe's leading futures exchange) and of the Deutsche Börse.

Europe's data centres are connected not just by fibre-optic cables but also by microwave links, and those in and around London by millimetre wave links too. London's rain is different from New Jersey's — I'm told the average droplet size is smaller — making life in London somewhat easier for millimetre waves and, in the words of one of my contacts, 'much worse for lasers'. So if you're a Londoner, and are spooked by the idea of lasers flashing stock market data overhead, be grateful for drizzle: it might keep the lasers at bay.

Go beyond the Chicago-New Jersey and London-Frankfurt clusters and you are soon back in the slower world of fibre-optic cable. Oceans are a major barrier to microwave, millimetre wave and laser transmission through the atmosphere. Because of the curvature of the Earth and attenuation of the signals, all three require a series of 'line of sight' repeater stations. I've been warned, however, not to imagine that this barrier can't be overcome. There's serious thinking going on about ideas such as suspending microwave repeater stations from a series of balloons.

For now, though, transatlantic and other global financial trading links run through undersea cables. The routes followed by existing cables are often at some distance from the relevant geodesics, for example to minimise the extent to which they had to be laid on the continental shelves. In shallow waters, cables need to be buried in the ocean floor, which is expensive, because a cable that isn't buried is vulnerable to trawlers and ships' anchors, and also to attacks by sharks (strangely, sharks are attracted by the electromagnetic radiation from fibre-optic cables).

So an obvious thing to do is to lay new cables closer to the geodesic. It's brutally expensive: laying a transatlantic cable can cost you more than $300 million, and to my knowledge nobody has done it since the dot.com boom ended. A firm called Hibernia Networks, however, has had plans to do just that, with the intention of shaving around 2.6 milliseconds off the one-way transmission time on the fastest existing cable, Global Crossing's AC1. However, amongst Hibernia's partners was the Chinese equipment maker Huawei, and the project seems to have hit trouble because of US cybersecurity concerns.

But you can do quite a lot without a new cable. You can add microwave links between the landing stations of existing undersea cables and the world's main financial data centres. Doing that has, for example, reduced the total one-way transmission time from Aurora, Illinois to LD4 in Slough or FR2 in Frankfurt to 35-38 milliseconds. That gives high-frequency trading systems in those data centres usefully up-to-date information, for example about the order book for the E-Mini (which is helpful data not just for those trading US shares, but also for those trading European shares and futures).

As well as needing links of this kind, HFT firms have to speed up their computing as much as possible. Even the fastest conventional computer hardware, running conventionally, is now not fast enough. During the lunch break of one of the HFT conferences I attended, I noticed the staff of a firm that was exhibiting fanning out across the lunch room, one of them going to each table. The technology they were promoting involved submerging computer systems in liquid nitrogen, which permits 'superclocking': getting a computer system to run faster than it would be safe to run it at room temperature.

As far as I can tell, however, that approach hasn't been adopted much by high-frequency traders: if nothing else, installing a liquid-nitrogen system in your 'cage' in a data centre is likely greatly to increase the rent you have to pay. Much more widely used are specialised forms of computer hardware known as FPGAs, or Field-Programmable Gate Arrays. The basic idea is to shift as much computing as possible off the microprocessor chips that make up the heart of a computer system, and to do that computing not by software but in fast FPGA hardware.

I was in Chicago in April, and was told that there was a lot of 'buzz' amongst the city's high-frequency traders for the FPGA technology of a firm called Solarflare. It enables you to circumvent the kernel – the programs that manage the computer's central processing unit, memory and input/output devices – and send bits (binary digits) direct from the raw datafeed to designated locations in a computer's memory. 'They [Solarflare] promise you 1.2 microseconds one-way trip from the network to your user memory', a programmer told me.

But you can't do everything using FPGAs: sometimes, high-frequency trading does require use of a computer server's central processing unit. Sitting on a bench in

Chicago's Millennium Park, the programmer tried to explain to me the software style that was necessary to ensure speed. 'There are rules you need to follow to write fast code', he told me: 'Don't touch the kernel. Don't touch main memory. …Don't branch.' (That third commandment means don't fill your program with "if" statements – with lines of program with the generic form "**if** A **then** do B **else** do C" – because those get in the way of a modern microprocessor's capacity to do several things in parallel.) But he also warned me that 'we don't know ahead of time what those rules are because every piece of code [program] comes with a different rule book. … I call them rules, but they're more guidelines … That's why it's hard to teach someone: they either get it or they don't. Those that get it, awesome.'

I've heard the requisite style of programming described as 'bit fucking'. I confess I'm a tad uncomfortable with the term, but it conveys the need for an intimate understanding of precisely the best, subtlest way to handle the flow of binary digits from the raw data feed, through the computer system and the additional specialised hardware, and then back (in the form of orders) into the network connection to the relevant order gateway.

A more polite term for bit fucking is 'close-to-the-metal programming'. The underlying issue is rather like the question of geography. Just as discussions of globalisation tended to assume that geography and local phenomena such as London's rain had become unimportant, so it is all too easy to think that digital financial markets are abstract and virtual. But, just like bankers who never look at maps, programmers who think like that will do a poor job in high-frequency trading: they will be amongst those who never 'get it'. If you're going to write really fast code, you have to understand the computer that you are programming not — as you might have been able to think about it in academic computer science — as an abstract machine,

but as a specific physical device through which electrical signals pass. Only then can you work out the most efficient way of channelling those signals.

If you're a certain kind of person (it's tempting to write 'a certain kind of man', but that would be simplistic), there's pleasure to be had in much of this. I don't mean simply in bit fucking, although pleasure is of course one of the term's connotations. Pleasure can also be found in other forms of skilled engineering, and in working for a firm with at most a few dozen employees that can outwit banks that are big, rich, but also often bureaucratic and sometimes simply stupid. (Nearly all HFT firms started small, and only a few have grown to be more than medium-sized.) I would even confess that some of the pleasure rubs off on me. I'm enjoying studying a domain in which economic life is so blatantly physical – a world of wind and rain and fog, of tunnels and oceans and sharks, of the geography of unfashionable places such as Aurora, Weehawken and Slough.

There's actually less money to be made from the pleasures of high-frequency trading than you might think. For example, if your program is market-making, you might hope that it would make a couple of cents for every share it buys and sells. Even a medium-size HFT firm can be trading a billion shares a month, and so the cents would really add up. However, in this domain prediction is almost always probabilistic. For example, a HFT program can seldom really know that a particular little order is a child of a bigger parent: it can only guess. More generally, most programs' predictions will be wrong almost as often as they are right. So, in reality, a market-making program is doing pretty well if it turns a profit of a tenth of a cent per share traded. True, there are — I'm told — opportunities that are quite a bit more profitable than that (and it is those that 'aggressive' programs live off), but those are less common than the more routine stuff.

A good high-frequency trading firm can almost always earn money in its trading, but costs of the kind I've listed weigh heavily, for example because you have to pay at least some of them at each of the many venues on which you are trading (the U.S. has 12 share-trading exchanges and dozens of dark pools). The fact that a lot of these venues are in the same handful of data centres — I'm told NJ2 in Weehawken is a favourite of dark-pool operators — creates some economies of scale, but each extra venue to which you feel compelled to connect adds costs. I get the impression that if a HFT firm fails it is most usually because it slowly drowns, submerged in a sea of costs. To a visitor like me, a sign of the risk of this fate is an office with lots of empty desks. One trader's offices had a fine view, but there seemed to be few staff to enjoy it. 'You have to love it [high-frequency trading]', he said.

Many pleasures, of course, are risky, and so it is with automated trading. Occasionally, an automated trading firm blows up spectacularly, rather than quietly drowning. The most dramatic case was the US market-making firm and stock broker, Knight Capital, which lost $440 million in 45 ghastly minutes on the morning of 1 August 2012. An old, long-unused trading program mistakenly left on one of its servers suddenly sprang to life, and unfortunately in the interim the piece of program that kept track of the execution of its orders had been moved and no longer worked. So the trading program kept on pumping more and more orders into the market. Unaware that the program was still there, Knight staff wrongly guessed that the fault was in newly installed trading software, so they lost valuable time uninstalling the software from Knight's servers. By the time the old code was found and switched off, the firm was on the brink of bankruptcy.

By no means all such events become public. In a coffee house in New York's Greenwich Village, a former high-frequency trader told me in a matter-of-fact way

that one of his colleagues had once made the simplest of slip-ups in his program: what mathematicians call a ‘sign error’, interchanging a plus and a minus. In consequence, when the program started to run it behaved rather like the Knight program, building a whole set of bigger and bigger trading positions, in its case at an exponential rate: doubling them, then redoubling them, and so on. ‘It took him 52 seconds to realise what was happening, [that] something was terribly wrong, and he pressed the red button’, stopping the program. ‘By then we had lost $3 million.’ The trader’s manager calculated ‘that in another 20 seconds at the rate of the geometric progression’, the trading firm would have been bankrupt, ‘and in another 50 or so seconds, our clearing broker [a major Wall Street investment bank] would have been bankrupt, because of course if we’re bankrupt our clearing broker is responsible for our debts … it wouldn’t have been too many seconds after that the whole market would have gone.’

What is most telling about that story is that not long previously it couldn’t have happened. High-frequency firms like my informant’s are sharply aware of the risks of bugs in programs, and the firm in question had had an automated check that would have stopped the errant program accumulating trading positions well before its human user realised something was wrong. However, the firm had been losing out in the speed race to other firms, and so had launched what my informant called ‘a war on latency’, trying to remove all detectable sources of delay. Unfortunately, the risk check had been one of those sources.

Pleasure and risk are important, but we do need to come back to money. The right question to ask about high-frequency trading is not just whether high-frequency traders are good or bad, or whether they add liquidity to the markets or increase volatility in them, but whether the entire financial system of which they are part is doing what we want it to do. Of course, we want it to do several things, but I’d posit

that high on the list should be whether the system is taking people's savings and putting them to the socially most productive uses, without too much of those savings sticking to the hands of financial intermediaries on the way.

As I've said, less money sticks to the hands of high-frequency traders than you might think, but that doesn't mean that we've got an optimal financial system. Speaking a couple of years ago to Bloomberg *Businessweek* about new, faster cables such as that planned by Hibernia, Manoj Narang (founder of the Red Bank, New Jersey, HFT firm Tradeworx) commented: 'Nobody's making extra money because of them: they're a net expense on the industry. … All they've done is impose a gigantic tax on the industry and catalyse a new arms race.'

The chief economic characteristic of most arms races is that all participants have to spend more money, and none of them ends up any better off because of it. Although it may not be the most important source of waste within the financial system, an arms race within it does mean that money is being wasted. Sitting on the bench in Chicago, the programmer I was speaking to told me that his firm had 'to spend god awful amounts of cash on IT'. The competition for speed was 'a fun ride but bloody let us off!'

He had a simple proposal for how this could happen. The Securities and Exchange Commission (SEC), which regulates share trading in the US, could simply rule that matching engines should not search for matches all the time, but can do so only 'every hundred millis' – every hundred milliseconds. That would turn what is in effect the continuous auction conducted by a matching engine into a series of what economists call 'batch auctions'.

I'd actually heard the proposal to turn continuous into batch auctions before, most recently from the University of Chicago economist Eric Budish, who (along with his colleagues Peter Cramton and John Shim) has developed an elegant formal model of the arms-race aspect of modern financial trading. As both they and my programmer suggest, batch auctions could end the arms race by eliminating the advantage a participant gets from an increase in the speed of its trading.

Variants of the batch-auction proposal are being experimented with in foreign exchange, but that's a very different context. The big banks retain market power in foreign exchange that they have largely lost in share trading, and the difficulties they have with fast trading give them reasons of their own for wanting to slow it down. The programmer I was speaking to wasn't optimistic about the prospects of batch auctions in share trading. 'It's a dream, but it's never going to happen': too many people benefit from the current set-up.

One would also need to be very careful about how batch auctions were introduced. The feed servers would have to fall silent during the period between them, because otherwise the arms race would simply concentrate in its final millisecond. More generally, changes in a fast, interactive, highly complex system of the kind I've described can have unexpected side effects. But a regulator such as the SEC could introduce a change such as this on an experimental basis, for only some stocks, and evaluate whether it had benefits.

Although I don't think my programmer knew it, his proposal for auctions every 100 milliseconds stands in an intriguing genealogy. 100 milliseconds – a tenth of a second – is approximately the threshold of human perception of time, and so has played an important role across a variety of both human and physical sciences as well

as in fields such as cinematography, a role explored in Jimena Canales's, *A Tenth of a Second: A History*.[4] It's also a marker of how fast finance has become that being able to trade only every tenth of a second would now count as slow trading.

[4] Chicago University Press, 2009.