Raging Bulls: How Wall Street Got Addicted to Light-Speed Trading

By Jerry Adler

Link: http://www.wired.com/business/2012/08/ff_wallstreet_trading/
08.03.12 5:53 PM

Photo: Tim Flach/Getty

Wall Street used to bet on companies that build things. Now it just bets on technologies that make faster and faster trades.

Editor's note: *One of the most interesting things about the catastrophe at Knight Capital Group—the trading firm that lost $440 million this week—is the speed of the collapse. News reports describe the bulk of the bad trades happening in less than an hour, a computer-driven descent that has the financial community once again asking if their pursuit of profit has lead to software agents that are fast, dumb, and out of control. We're posting this story in advance of its publication in Wired's September issue because it examines how Wall St. got to the point where flash failures come with increasing frequency, and how much farther traders seem willing to go in pursuit of ever-greater speed.*

The 2012 New York Battle of the Quants, a two-day conference of algorithmic asset traders, took place in New York City at the end of March, just a few days after a group of researchers admitted they had made a mistake in an experiment that purported to overturn modern physics. The scientists had claimed to observe subatomic particles called neutrinos traveling faster than the speed of light. But they were wrong; about six months later, they retracted their findings. And while "special relativity upheld" is the world's most predictable headline, the news that neutrinos actually obey the laws of physics as currently understood marked the end of a brief and tantalizing dream for quants—the physicists, engineers, and mathematicians-turned-financiers who generate as much as 55 percent of all US stock trading. In the pursuit of market-beating returns, sending a signal at faster than light speed could provide the ultimate edge: a way to make trades in the past, the financial equivalent of betting on a horse race after it has been run.

"Between the time the first paper came out in September and last week, a guy in my shop had written two papers explaining how it could be true," a graying former physicist said ruefully, sipping coffee near an oversize Keith Haring canvas that dominated the room at Christie's auction house where the conference was held. "Of course, you'd need a particle accelerator to make it work."

If that were all it took, then by now someone would be building one. One of the major themes of this year's conference was "the race to the bottom," the cost-is-no-object competition for the absolute theoretical minimum trade time. This variable, called latency, is rapidly approaching the physical limits of the universe set by quantum mechanics and relativity. But perhaps not even Einstein fully appreciated the degree to which electromagnetic waves bend in the presence of money. Kevin McPartland of the Tabb Group, which compiles information on the financial industry, projected that companies would spend $2.2 billion in 2010 on trading infrastructure—the high-speed servers that process trades and the fiber-optic cables that link them in a globe-spanning network. And that was before projects were launched to connect New York and London by a new transatlantic cable and London and Tokyo by way of the Arctic Ocean, all just to cut a few hundredths of a second off the time it takes to receive data or send an order.

High-frequency traders are a subset of quants, investors who make money the newfangled way: a fraction of a cent at a time, multiplied by hundreds of shares, tens of thousands of times a day. These traders occupy an anomalous position on Wall Street, carrying themselves with a distinctive mixture of diffidence and arrogance that sets them apart from the pure, unmixed arrogance of investment bankers. A pioneering high-frequency trading firm, Tradeworx, has its relatively humble offices two flights up from an Urban Outfitters in a sleepy New Jersey suburb. Twenty people work there, about half of them on the trading floor, monitoring on triple screens the fractions of a penny as they mount up, second by second. Roughly 1.5 percent of the total volume of stocks traded on US exchanges on a given day will pass, however fleetingly, through the hushed, sunlit, brick-walled room.

On the first day of the New York conference, Aaron Brown, a legendary quant and former professional poker player, took the stage in rumpled chinos and a leather jacket to lecture the assembly on game theory. He began his talk by saying, "3.14159," and then pausing expectantly. From the back of the room came the response: "265358." Together they made up the first 12 digits of pi—a geek shibboleth. "You won't see a lot of masters of the universe here," said Charles Jones, a professor of finance and economics at Columbia Business School. "A lot of these guys, if they're wearing a tie, it might be the only one they own."

Faster and faster turn the wheels of finance, increasing the risk that they will spin out of control, that a perturbation somewhere in the system will scale up to a global crisis in a matter of seconds. “For the first time in financial history, machines can execute trades far faster than humans can intervene,” said Andrew Haldane, a regulatory official with the Bank of England, at another recent conference. “That gap is set to widen.”

This movement has been gaining momentum for more than a decade. Human beings who make investment decisions based on their assessment of the economy and on the prospects for individual companies are retreating. Computers—acting on computer-generated market trend data and even newsfeeds, communicating only with one another—have taken up the slack. Conventional economics views all this as an unalloyed good: It is axiomatic that all trades are a net benefit to the economy because they enhance “liquidity,” the ability of investors to buy or sell assets at the best price. Indeed, in 2007 the SEC instituted an ambitious new rule, the national market system, that opened the door to dozens of new venues for stock trading, but now that transaction times are measured in microseconds and prices are carried out to six decimal places, those opportunities have arguably gone past a point of diminishing returns.

So, barring any new breakthroughs in physics, we are in the final stages of a trend that began when the Rothschilds, by legend, used carrier pigeons to trade on the outcome of the Battle of Waterloo. For roughly a century leading up to 1970, the state of the art in financial communication was the telegraphic stock ticker (for receiving data) and the telephone (for transmitting orders). Now it is the high-speed server linked to a financial exchange by fiber-optic cable as short as physically possible, because each mile adds about eight microseconds of latency. There is so much money to be made that any expenditure on research and infrastructure to shave those microseconds is worth it.

That plays out in the very hardware of finance. The data center of NYSE Euronext, the international conglomerate that includes the New York Stock Exchange, is in a building in suburban Mahwah, New Jersey, 27 miles from Wall Street. Besides “matching engine” computers that process trades on the exchange, it also houses high-frequency trading servers, which receive data and spit out orders according to programs—algorithms. Traders pay to put their servers in the same building, and to make things fair, engineers scrupulously add extra lengths of cable to equalize the runs among all the servers. Yes, we are talking about a few feet plus or minus. At nearly the speed of light.

Now multiply that effort across the breadth of a continent or an ocean. Factor in the quantum and relativistic effects of machine-to-machine communication untouched by human hands, far faster than a human can react. A trend that began with pigeons ends with subatomic particles, carrying data that is outdated almost before it arrives at its destination.

Photo: Citywide Corporate

Not all trading takes place in New York. By historical accident, derivatives such as futures and options are mostly traded on the Chicago Mercantile Exchange, 720 miles away. So a few years ago, a company called Spread Networks began quietly buying up rights-of-way for a route that would lop about 140 miles off the shortest fiber-optic cable distance between the Chicago Merc and the communications hub of Carteret, New Jersey, the primary data center for Nasdaq. Existing networks tend to follow railroad lines and were designed to serve population centers, not to provide a point-to-point link for traders. Instead of dipping south toward Philadelphia, Spread's route heads northwest through central Pennsylvania and then due west to Cleveland. Latency is typically measured in round-trip times (i.e., an order and a confirmation); the shortest cable route before Spread lit up its network in 2010 clocked a round-trip time of 14.5 milliseconds, according to Spread executives, but capacity was inadequate, so most customers had to settle for 15.9 milliseconds. Spread cut that to as little as 13.1 milliseconds for its premium "dark fiber" service, a connection that doesn't have to be shared with other customers. Prices are a closely guarded secret in this world, although the consensus estimate among traders is "plenty."

Spread quickly began signing up customers, but by the spring of 2012, there was a faster competitor on the horizon. Because of some complicated physics, the speed of light through any medium is inversely proportionate to the medium's index of refraction—so signals travel about 200,000 kilometers per second through fiber-optic cable, compared with 300,000 through the atmosphere. The fastest communication between New York and Chicago would be line-of-sight through the air, which requires a chain of microwave relay towers. Tradeworx is building such a network, as is McKay Brothers, a California firm that hopes its system will be the fastest, with a round-trip latency of less than 9 milliseconds. Its route, cofounder Bob Meade boasts, uses the smallest possible number of towers, 20, and deviates from a perfect geodesic (more poetically known as a "great circle," the shortest line between two points along a planet's surface) by just 4 miles. It includes roughly another 2 miles of wiring in the towers themselves, connecting the dish antennas hundreds of feet in the air with amplifiers on the ground.

The downside is that microwaves in the 11-gigahertz band can be interrupted by rainstorms or certain atmospheric conditions that duct the signal away from the receiving dish. CEO David

Barksdale of Spread Networks, which claims “five nines”—99.999 percent reliability—for its fiber-optic link, says he’s not worried about microwave competition: “People have been talking about building low-latency wireless networks for years. We don’t believe long-haul microwave is a suitable technology for sophisticated trading applications, due to certain key limitations.” (He declined to specify what those were.) But Meade, a former Harvard physicist and quant, is convinced that speed, more than reliability, is the key. If your link is only 99 percent reliable, you don’t make money 1 percent of the time; if it’s slower than the competition, you don’t make money 100 percent of the time.

The other crucial routes are New York to London and London to Tokyo. (Trading hours in the US don’t overlap much with Asia, so there’s less demand for an ultrafast New York–Tokyo link.) At least three companies have announced plans for fiber-optic cables under the Arctic Ocean between Europe and Japan. One route skirts the Russian coast and comes ashore on the northern tip of Murmansk; the other traverses the Northwest Passage through the Canadian Arctic. When they go into operation around 2014, they will cut latency from about 230 milliseconds on routes through Asia to between 155 and 168 milliseconds.

Meanwhile manufacturers have begun making cable for a new New York–London link intended to shave 311 miles off the usual distance and cut the round-trip message time from 65 milliseconds to just under 60. It will do this by taking a great-circle route, traversing the shallow Grand Banks off Newfoundland. Most transatlantic cables head straight for deep water, to get away from sharks. In what some might consider a case of karmic justice, sharks threaten the financial industry by biting its cables, attracted by the electromagnetic fields generated by the wires that power the amplifiers at intervals along their length. Along the continental shelf, cables must be expensively armored against sharks and if possible buried to avoid damage from anchors and fishing trawls. The new cable will be armored for about 60 percent of its length, to take advantage of the shortest possible route. By summer 2013, two ships will begin laying cable, meeting mid-Atlantic in about three months, according to officials of Hibernia Atlantic, the company behind what it calls Project Express. Cost: around $300 million. Estimated useful life before obsolescence: hard to say. But what else can they do? Unlike the New York–Chicago route, the Atlantic Ocean is a highly unsuitable environment for erecting microwave towers. On the other hand, when I raised this point at the Battle of the Quants with Alexander Dziejma, chief architect at a high-frequency trading firm called Dymaxion Capital Management, he scoffed.

“They’re doing amazing things now with drones,” Dziejma said.

“Drones?”

Sure, he said. A fleet of unmanned, solar-powered drones carrying microwave relay stations could hover at intervals across the Atlantic. I started to come up with all the reasons that was a crackpot idea, then realized I’d heard 10 crazier things since 9 am that day. “Someone will do this eventually,” he said.

High-frequency traders make money in a vacuum, grabbing for pennies that appear and disappear like the virtual particles of quantum field theory. Their goal is to end each trading day “flat”—out of the market, their profits safely in the bank. Depending on their model, they can do well winning as little as 55 percent of their trades. They are continuously testing prices, looking for patterns and trends or the chance to buy something in one place for $1 and sell it somewhere else for $1.01, or $1.001. Sometimes they aren’t even looking to make money on the trade itself. Under the “maker-taker” model, some exchanges offer tiny incentive payments, or rebates, for

posting a quote (to buy or sell a stock) that results in a trade. The exchange charges the other side in the trade, the taker, a slightly higher fee and collects the difference. So an algo can buy a stock, earn a rebate, then sell the stock and earn a rebate for that too. All of this is governed by algorithms whose lifespans can be as short as a few weeks. Sometimes an algorithm does something as simple as look for a stock that ticks up in price several trades in a row. A "momentum" algo would buy the stock, expecting the rise to continue. A "mean-reversion" algo would sell, expecting a drop back to average price. They might both even be deployed by the same firm. Over the course of a minute, they might both be right.

One common algo strategy is to look for pairs of stocks whose prices are historically correlated. The canonical examples are the stock prices of oil companies, which rise with the price of crude, and those of airlines, which do the opposite. But they may not move all at the same time, so one strategy is to buy or sell the one that's trailing and wait for it to catch up. Similarly, "derivative" equities such as options and futures may get out of equilibrium with the underlying stocks. Some algorithms are "market makers" in a stock—they attempt to buy at a low bid price and quickly sell at a slightly higher asking price, pocketing the difference, or spread. The people who did this used to be called specialists, and it was a nice living when spreads were an eighth of a dollar. Since the New York Stock Exchange instituted "decimalization" in 2001, spreads have gone down to a penny or two, meaning you have to trade a lot more stock, a lot faster, to make the same amount of money. It's no place for a human being.

A more specific example: a simplified algorithm that Mani Mahjouri, chief investment officer of Tradeworx, presented at the Battle of the Quants. His hypothetical algorithm buys and sells shares of SPY—a security pegged to the level of the S&P 500 index, which is based on the value of 500 leading publicly traded companies. SPY trades on various exchanges, including Nasdaq in New York, but in Chicago there is a market in S&P 500 futures, contracts whose price reflects bets on the value of the index weeks or months ahead.

The prices of SPY and the futures contract generally move together but not in lockstep; as a rule, the futures contract tends to lead SPY by a few milliseconds. The reason is immaterial; all that matters is that the rule holds often enough for an algo to make money by using changes in S&P futures to predict the direction of SPY over the next few milliseconds. How much money depends on how quickly the algo can get data from Chicago to New York. As an experiment, Mahjouri applied this algorithm to the complete record of trades for an entire day and found that under optimum conditions—by which he meant transmitting data and orders at the speed of light—it would have made approximately 64,000 trades at an average profit of $0.0001 per share. The model specified a size of one share per trade; as a rule, high-frequency traders deal in relatively small volumes, a few hundred shares at a time, because large orders perturb the market. Strictly as an illustration, consider that on an average day about 150 million shares of SPY change hands. Tradeworx claims to account for around 4 percent of SPY trading, so by extrapolation, that would be 6 million shares of SPY, times $0.0001, equals ... hmm ... $600. By itself, this is a slow way to get rich. But multiply that figure by the number of such algorithms running at any given time—in the "high hundreds"—and it starts to get interesting. One reason high-frequency trading works at all is that it takes place much too fast for human beings to get in the way.

Here's something to try: Go to Yahoo Finance during trading hours and enter a ticker symbol for a quote, say, INTC for Intel. On the quote page, click on Order Book. You'll see, under "Bid," a list of four or five prices in descending order, and a similar list, arranged lowest to highest, under "Ask." These represent offers to buy or sell a specified quantity of stock. The prices will be

closely spaced, and the top numbers should converge on the current ticker price, reflecting the most recent trade.

So that is the capital markets at work, but it is also, of course, fiction. If you as a retail customer want to buy stock in Intel, your order will most likely be filled at or near the prevailing market price by your brokerage, out of in-house inventory. Or it may be handed off for a small fee to an "internalizer," who buys and sells in a "dark pool" where quotes aren't published. Perhaps you've heard of Bernard Madoff; in the legitimate part of his business, that's what he did. Bid and ask quotes are always to the penny, but the prices of executed trades may be carried out to three, four, or even six decimal places. No sane human trader would spend their time haggling over a ten-thousandth of a cent, but computers don't get bored.

The quotes in Yahoo's order book probably came from an algo, and you almost certainly can't trade at that price. Even if you had access to the exchange, which of course you do not, they would likely be gone long before you could jump in the market—either already executed or, much more likely, withdrawn before any shares changed hands. And that's where the really dangerous part begins to set in. It's not just that humans are less and less involved in trading; it's that they can't be involved. "By the time the ordinary investor sees a quote, it's like looking at a star that burned out 50,000 years ago," says Sal Arnuk, a partner in Themis Trading and coauthor of a book critical of high-frequency trading titled Broken Markets. By some estimates, 90 percent of quotes on the major exchanges are canceled before execution. Many of them were never meant to be executed; they are there to test the market, to confuse or subvert competing algorithms, or to slow trading in a stock by clogging the system—a practice known as quote stuffing. It may even be a different stock, but one whose trades are handled on the same server. On the Internet, this is called a denial-of-service attack, and it's a crime. Among quants, it's considered at most bad manners.

As one result, the consolidated quotation system that aggregates quotes from 12 American exchanges and transmits them to traders—the market's nervous system, in a way—is feeling the strain. "Each time they increase the capacity, the volume of messages gaps up to fill it," says Eric Scott Hunsader, CEO of Nanex, a firm that aggregates and analyzes market data. "You need a gigabit connection or better to stay on top of it—in 2000 I could keep track of the markets with a 56K modem. It's free to generate a quote, but we're all paying for it."

Photo: Tim Flach/Getty

Last March an upstart exchange called BATS, based in Kansas City, organized an initial public offering of its own stock. Within a few seconds after trading opened, something went wrong: A software bug froze trades in BATS stock on the BATS exchange, and in the process took down a server that handled all the ticker symbols at the top of the alphabet. So the shutdown also affected a certain Cupertino-based company whose name starts with A, erroneously reporting a price drop of almost 10 percent on BATS and causing a brief suspension in all trading of the world's most closely watched stock. Meanwhile, the only exchange quoting BATS stock was Nasdaq, and something strange was happening there as well. In 900 milliseconds, too fast for anyone to react, the stock plummeted from its $15.25 opening price to $0.28, reaching a fraction of a cent before trading was halted.

BATS executives apologized, took responsibility, withdrew the IPO, and canceled the trades. But maybe there was more going on than a glitch. Analyzing the transactions, a Nanex engineer named Jeffrey Donovan saw the fingerprints of an algorithm designed to feed stock into the market at successively lower prices. "You can see it waiting a few milliseconds after each trade for the bid side to lower its price, and then the cycle repeats until the stock goes to zero," he says. Whoever may have done this presumably wasn't in it to make money; Nanex CEO Hunsader thinks it was an attempt to obliterate BATS, which in just a few years has captured some 10 percent of US trading volume from older competitors.

BATS officials wouldn't comment on this theory, which was kicking around stock message boards last spring. (The SEC is looking into it.) Other market observers were skeptical. In general, says Tradeworx CEO Manoj Narang, "computers don't engage in market manipulation. Humans do. Any HFT who encodes manipulative logic into an algorithm that anyone could subpoena is probably too stupid to make money."

Most of Wall Street looked on the bright side. Because the BATS debacle didn't precipitate a widespread meltdown across other exchanges, some market observers think that regulations meant to prevent a repetition of the Flash Crash of May 6, 2010—when the Dow Jones Industrial Average fell 600 points in five minutes—did their job. High-frequency traders believe that event has misled the public into blaming them for every financial misadventure since the Great Crash of 1929. "People literally shake when they learn what I do," says Irene Aldridge, a leading algo trader and a panelist at the Battle of the Quants. In a 100-page report that avoids using the words fault or blame, regulators found that at the onset of the 2010 crash, "HFTs began to quickly buy and then resell contracts to each other—generating a 'hot potato' volume effect as the same positions were rapidly passed back and forth." To someone keeping an eye on their retirement account, that doesn't sound good; investors have withdrawn more than $300 billion from long-term mutual funds in the two-plus years since then. "When someone can set off a panic with a single bad trade, it creates an environment that's unfriendly to investors," Themis Trading's Arnuk says, "which is why they've been fleeing ever since."

High-frequency trading raises an existential question for capitalism, one that most traders try to avoid confronting: Why do we have stock markets? To promote business investment, is the textbook answer, by assuring investors that they can always sell their shares at a published price—the guarantee of liquidity. From 1792 until 2006, the New York Stock Exchange was a nonprofit quasi utility owned by its members, the brokers who traded there. Today it is an arm of NYSE Euronext, whose own profits and stock price depend on getting high-frequency traders in the door. Trading increasingly is an end in itself, operating at a remove from the goods-and-services-producing part of the economy and taking a growing share of GDP—twice what it did a century ago, when Wall Street was financing the enormous industrial expansion of the economy. "This is counterintuitive, to say the least," wrote New York University economist Thomas

Philippon in an article for the Russell Sage Foundation. “How is it possible for today’s finance industry not to be significantly more efficient than the finance industry of John Pierpont Morgan?”

At a press briefing a few months ago, Mary Schapiro, chair of the Securities and Exchange Commission, said she was concerned about the volume of trading “unrelated to the fundamentals of the company that’s being traded.” Proposals have been floated for ways to rein it all in, such as penalizing traders for canceling too many orders. Most of them are stranded in the proposal stage. “Circuit breakers” arrest a spiraling market, but right now high-frequency trading is essentially unregulated. (The SEC now says it will at last audit microsecond-scale trading—by buying access to Tradeworx’s data.)

The defense of automated trading is that it lowers the cost of trading and improves liquidity by lowering the spread between bid and ask prices. “At the end of each trading day, equities wind up in the hands of people who want to hold them for appreciation or for dividends,” says Bernard Donefer of the Subotnick Financial Services Center at CUNY’s Baruch College. “If that’s you, you don’t have to pay attention to what goes on between 9:30 and 4:00 and can still get the benefits of lower costs and faster executions.” Fair enough, but as Arnuk and his partner Joseph Saluzzi point out in Broken Markets, you also bear the risk of heart-stopping price swings when the algos shut down their programs to avoid losses for their own reasons. They aren’t in business to make money for you.

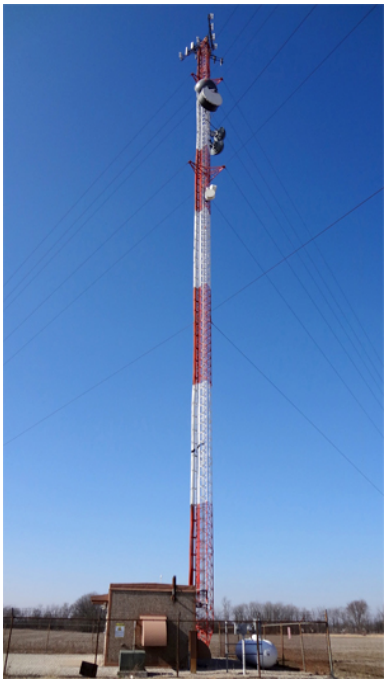

Photo: Steve Ikam

One of the effects of the explosion of algo trading is a boom in the market for financial data. Some of the most valuable is historical; quants require it by the terabyte to back-test their trading models. Even free information, such as corporate earnings reports and government statistics, can be sold if you format it to be machine-readable. That did not escape the notice of Dow Jones, which realized that all the money it had invested in getting the news out first was

wasted on humans, who took entire seconds to digest it. The result was the creation in 2007 of an "elementized" direct newsfeed to customers' trading platforms—all the news fit to quantize.

Since then Dow Jones has expanded the feed from numeric data to cover "unstructured" news, anything deemed likely to affect profits and stock prices. That move spawned a whole industry of "news analytics" dominated by companies like RavenPack, which turns 100,000 news articles a day into trade-ready data. "An asset manager says, 'I want to model a strategy based on central bank events,' and so we keep track of every article that mentions both the head of the European Central Bank and the word vigilance," says Rob Passarella, vice president and managing director of Dow Jones Institutional Markets. One of the rules of analytics, disconcertingly enough, is that other things being equal, companies that don't get written about in the media tend to outperform those that are widely covered.

And it's not just the words of central bankers that matter in this world. Almost any kind of data that in any way bears on economic activity, no matter at what remove, is being aggregated and tested for its potential impact on stock prices. GPS data, showing concentrations of cell phone users in malls or office buildings, has been used to get a real-time read on economic activity. Even the most ephemeral shards in the digital scrap heap, old Twitter posts, are proving of value, if you can get your hands on enough of them. For their starkly titled paper "Twitter Mood Predicts the Stock Market," Johan Bollen, an informatics researcher at Indiana University, and two colleagues collected almost 10 million tweets from 2008, aggregating phrases that indicate emotional state and analyzing them along dimensions of feeling such as "calm," "alert," "sure," "vital" and "happy." They then looked for correlations with stock prices and discovered that a surge in "calm" sentiment reliably predicted an increase in the Dow Jones Industrial Average two to six days later. No one, including Bollen, knows what this means or why it should be so, but if quants had a coat of arms, it would say, "If it works, trade on it."

Furthermore, trade on it now, this microsecond. It is only a matter of time, perhaps a few decades, says Alexander Wissner-Gross, a Harvard physicist, before some hedge fund decides it needs a particle accelerator to generate neutrinos, and then everyone will want one. Yes, they travel slower than light, but they indisputably can tunnel through the earth, cutting thousands of miles off an intercontinental message. And just a few days before the Battle of the Quants, right before the bad news about faster-than-light neutrinos, researchers announced they had sent a message by neutrino from the Fermilab accelerator in Chicago to a detector a kilometer away. According to Dan Stancil of North Carolina State University, the signal traveled at "very close to" the speed of light. Unfortunately, the data rate was only about 0.1 bits per second, meaning it would be useless for much more than sending a yes/no signal. "With the right modulation scheme, this could be increased by at least one or two orders of magnitude," Stancil said, adding "I don't know of a compelling commercial application." But we've all heard the (apocryphal) story that Thomas J. Watson of IBM predicted "a world market for maybe five computers." Stancil knows physics, but he seems never to have worked in finance. Any quant would know what to do next.

Buy neutrinos.