

# How fast is state of the art HFT trading systems today?

Asked 11 years, 6 months ago   Modified 6 years, 5 months ago

Viewed 27k times

---



53

All the time you hear about high frequency trading (HFT) and how damn fast the algorithms are. But I'm wondering - what is fast these days?



## Update



I'm not thinking about the latency caused by the physical distance between an exchange and the server running a trading application, but the latency introduced by the program itself.

To be more specific: What is the time from events arriving on the wire in an application to that application outputs an order/price on the wire? I.e. *tick-to-trade* time.

Are we talking sub-millisecond? Or sub-microsecond?

How do people achieve these latencies? Coding in assembly? FPGAs? Good-old C++ code?

## Update

There's recently been published an interesting article on ACM, providing a lot of details into today's HFT

technology, which is an excellent read:

[Barbarians at the Gateways - High-frequency Trading and Exchange Technology](#)

low-latency

algorithmic-trading

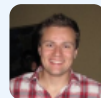
Share

edited Oct 18, 2013 at 10:03

Improve this question

Follow

asked Jun 22, 2013 at 22:47



Nicholas

2,195 ● 3 ● 24 ● 31

---

hmmm good question! Now i really want to know this!

– [Jorge Y. C. Rodriguez](#) Jun 22, 2013 at 22:50

---

- 2 "On the wire" is kind of a fuzzy boundary. It takes time for a complete data packet to arrive, and some of the processing may already have started before the entire message has been received. Everything is skewed through the different layers of the memory system and the kernel and the application, and people are paying close attention to that skew. – [sh1](#) Jul 1, 2013 at 12:49
- 

Thanks for the insight. Interesting! Can you provide any more details or examples? – [Nicholas](#) Jul 1, 2013 at 15:36

---

- 2 The bulk of the latency is usually for network I/O. A good network library should be able to process UDP/TCP with low single digit microsecond latencies. Here are [some benchmarks](#) you can check. – [rdalmeida](#) May 4, 2016 at 0:18



## 9 Answers

Sorted by:

Highest score (default)



56



I'm the CTO of a small company that makes and sells FPGA-based HFT systems. Building our systems on-top of the Solarflare Application Onload Engine (AOE) we have been consistently delivering latency from an "interesting" market event on the wire (10Gb/S UDP market data feed from ICE or CME) to the first byte of the resultant order message hitting the wire in the 750 to 800 nanosecond range (yes, sub-microsecond). We anticipate that our next version systems will be in the 704 to 710 nanosecond range. Some people have claimed slightly

less, but that's in a lab environment and not actually sitting at a COLO in Chicago and clearing the orders.

The comments about physics and "speed of light" are valid but not relevant. Everybody that is serious about HFT has their servers at a COLO in the room next to the exchange's server.

To get into this sub-microsecond domain you cannot do very much on the host CPU except feed strategy implementation commands to the FPGA, even with technologies like kernel bypass you have 1.5 microseconds of unavoidable overhead... so in this domain everything is playing with FPGAs.

One of the other answers is very honest in saying that in this highly secretive market very few people talk about the tools they use or their performance. Every one of our clients requires that we not even tell anybody that they use our tools nor disclose anything about how they use them. This not only makes marketing hard, but it really prevents the good flow of technical knowledge between peers.

Because of this need to get into exotic systems for the "wicked fast" part of the market you'll find that the Quants (the folks that come up with the algorithms that we make go fast) are dividing their algos into event-to-response time layers. At the very top of the technology heap are the sub-microsecond systems (like ours). The next layer are the custom C++ systems that make heavy use of kernel bypass and they're in the 3-5 microsecond range. The

next layer are the folks that cannot afford to be on a 10Gb/S wire only one router hop from the "exchange", they may be still at COLO's but because of a nasty game we call "port roulette" they're in the dozens to hundreds of microsecond domain. Once you get into milliseconds it's almost not HFT any more.

Cheers

Share Improve this answer

answered Feb 27, 2014 at 23:00

Follow



NVG Associates Inc

681 ● 1 ● 5 ● 3

---

Brian, you might want to know, that your Linked-In URL fortunately fails to work. – [user3666197](#) Aug 14, 2014 at 22:28

---

Mercury-Minerva: what does you mean by "port roulette" here? – [rahul.deshmukhpatil](#) Aug 17, 2015 at 10:55

- 
- 2 "port roulette" refers to the assignment of ports to customers by the exchange, in an environment where not all ports have the same hop-distance from the source server(s) and/or the same network hardware. When you request a new port, it's a gamble (the "roulette" part) whether you get a port with good latency or not. – [David Arnold](#) Nov 19, 2015 at 21:51

---

@Mercury-Minerva, you mention 'latency from an "interesting" market event on the wire' and FPGAs. [I was wondering](#) whether there are byte-oriented Ethernet NICs to avoid the 1-frame latency forced by typical frame-at-a-time NICs (which is at least 67 ns in 10GbE). Would you know? – [hmijail](#) Jan 27, 2017 at 14:58 ✎

---

If the cable connecting your computer to the exchange's computer is 1 meter long, it's going to take the signal at least

3.3 nanoseconds to reach the other side. It's not meaningful when you're talking about 700 nanoseconds, it is meaningful when you talk about 10. – [zmbq](#) Jun 13, 2017 at 8:56

---



25



You've received very good answers. There's one problem, though - most algo trading is secret. You simply don't know how fast it is. This goes both ways - some may not tell you how fast they work, because they don't want to. Others may, let's say "exaggerate", for many reasons (attracting investors or clients, for one).

Rumors about picoseconds, for example, are rather outrageous. 10 nanoseconds and 0.1 nanoseconds are exactly the same thing, because the time it takes for the order to reach the trading server is so much more than that.

And, most importantly, although not what you've asked, if you go about trying to trade algorithmically, don't try to be faster, try to be smarter. I've seen very good algorithms that can handle whole seconds of latency and make *a lot* of money.

Share Improve this answer

answered Jul 2, 2013 at 6:09

Follow



[zmbq](#)


39k ● 15 ● 105 ● 183

---

I will accept your answer because I think it has key insights but award the bounty to sll as he delivered the best answer regarding the bounty description. Thanks. – [Nicholas](#) Jul 2, 2013 at 8:06

---

For 2014 - check [goo.gl/3fxqQU](http://goo.gl/3fxqQU) for a description of a setup that is smaller. – [TomTom](#) Sep 2, 2014 at 12:25

10 nanoseconds and 0.1 nanoseconds are exactly the same thing, because the time it takes for the order to reach the trading server is so much more than that. This is a 6 years old comment and I reckon things have changed since. After all, what matters is the relative order of the traders' instructions compared to their competitors. For anyone curious, you can see some stats on reaction times (2018,2019) [here](#). – [Vasilios Mavroudis](#) Nov 6, 2019 at 14:56 



7



+100



Good article which describes what is the state of HFT (in 2011) and gives some samples of hardware solutions which makes nanoseconds achievable: [Wall Streets Need For Trading Speed: The Nanosecond Age](#)

With the race for the lowest “latency” continuing, some market participants are even talking about **picoseconds–trillionths of a second**.

EDIT: As [Nicholas](#) kindly mentioned:

The link mentions a company, Fixnetix, which can "prepare a trade" in 740ns (i.e. the time from an input event occurs to an order being sent).

Share Improve this answer

Follow

edited May 23, 2017 at 12:18



Community Bot

answered Jul 1, 2013 at 9:03



sll

62.4k ● 22 ● 108 ● 157



5

"sub-40 microseconds" if you want to keep up with Nasdaq. This figure is published here

<http://www.nasdaqomx.com/technology/>



Share Improve this answer

answered Jul 1, 2013 at 8:32

Follow



bbaassssiiee

6,734 ● 2 ● 47 ● 58



2 Can you provide a more specific link? – Flavio Jul 1, 2013 at 10:01



3

For what its worth, [TIBCO's FTL messaging product](#) is sub-500 ns for within a machine (shared memory) and a few micro seconds using RDMA (Remote Direct Memory Access) inside a data center. After that, physics becomes the main part of the equation.



So that is the speed at which data can get from the feed to the app that makes decisions.



At least one system has claimed ~30ns interthread messaging, which is probably a tweaked up benchmark,



so anyone talking about lower numbers is using some kind of magic CPU.

Once you are in the app, it is just a question of how fast the program can make decisions.

Share Improve this answer

Follow

edited Jul 2, 2013 at 7:24



Nicholas

2,195 ● 3 ● 24 ● 31

answered Jul 2, 2013 at 6:17



sasbury

301 ● 1 ● 4



3



Every single answer here is at least four years old and I thought I would share some perspective and experience from someone in the HFT / algorithmic trading field in 2018.

(This is not to say that any of these answers are poor as they most definitely are not however I believe it is necessary to provide insight regarding the topic that is more up to date).

To directly answer the first question: **We are talking approximately 300 billionths of a second (300 nanoseconds)**. Recall this is latency introduced *by the program* itself.

There is always going to be some variance firm by firm regarding the latency of systems, however the numbers I

am going to provide are the common values for internal HFT engine latency.

1. On average, one third of this time (300 nanoseconds) is attributed to latency *introduced by the program* as you stated in your question.
2. The remaining of the time is latency that exists due to co-location and other variables relating to the exchange, the matching engines, fibre optics, etc.

The question is about *how fast* high frequency trading systems are, and what the infrastructure looks like in terms of the hardware involved. The technology has advanced since 2014, however contrary to what a great deal of what the literature discusses in the field, *FPGAs are not necessarily the go-to choice for the big players in the HFT space*. Large companies such as [Intel](#) and [Nvidia](#) will cater to these firms with their specialized hardware to ensure they get everything they need from the trading system. With Intel obviously the system is going to be built more around CPUs and the kinds of computations best performed by CPUs, and with Nvidia the system will be more GPU oriented.

For systems on field programmable gate arrays (FPGAs), languages such as Verilog and VHDL are commonly used. However not everything is in assembly even for FPGA systems, most of it is highly optimized C++ with embedded inline assembly, this is where the speed often comes from. Note that this is the case for firms using all

sorts of hardware (FPGAs, specialized Intel systems, etc.)

It is unfourtunate however that the [top answer here](#) states something completely false:

10 nanoseconds and 0.1 nanoseconds are exactly the same thing, because the time it takes for the order to reach the trading server is so much more than that.

This is completely false as the co-location aspect of high frequency trading has become **completely standardized**. *Everyone* is just as close to the matching engine as you are thus the internal latency of the system is of great importance.

Share Improve this answer

Follow

answered Jul 19, 2018 at 4:52



ofey73

155 ● 3 ● 12



2



These days single digit tick-to-trade in microseconds is the bar for competitive HFT firms. You should be able to do high single digits using only software. Then <5 usec with additional hardware.

Share Improve this answer

Follow

answered Jan 29, 2014 at 16:02



Nathan Doromal

3,537 ● 2 ● 26 ● 28



1



According to the [Wikipedia page on High-frequency trading](#) the delay is microseconds:

High-frequency trading has taken place at least since 1999, after the U.S. Securities and Exchange Commission (SEC) authorized electronic exchanges in 1998. At the turn of the 21st century, HFT trades had an execution time of several seconds, whereas by 2010 this had decreased to milli- and even microseconds.

Share Improve this answer

answered Jun 22, 2013 at 22:53

Follow



[Gabriella Gonzalez](#)

35.1k ● 3 ● 79 ● 137

2 2010 - that's three years ago! That's eons in this space :-) We could be in the nano-second range by now. Also, it would be great if someone "from the inside" could reveal some figures. Even if it's just old ones from 2012 :) – [Nicholas](#) Jun 22, 2013 at 22:54 ✎



0



it will never be under a few microseconds, because of the em-w/light speed limit, and only a lucky few, that must be in under a kilometer away, can even dream to get close to that.

Also, there is no coding, to achieve that speed you must go physical.. (the guy with the article with the 300ns switch; that is only the added latency of that switch!;



equal to 90m of travel thru optical and a bit less in copper)

Share Improve this answer

edited Jul 1, 2013 at 9:28

Follow

answered Jul 1, 2013 at 9:11



[user2485149](#)

91 ● 7

- 
- 4 Right, many companies put hardware in 10meters from stocks not kilometers – [sll](#) Jul 1, 2013 at 10:03

---

well, if we're playing that game, why not just pay to install an application on the stocks main hubs.. or even add a patch to all the kernels responsible for real time updates.. i think it is a bit of overkill already – [user2485149](#) Jul 1, 2013 at 11:39

---

There's no doubt about overkill being in effect here. But there's real money to earn and I believe some go to extreme to achieve the lowest latency. – [Nicholas](#) Jul 1, 2013 at 12:22

---

You can't have a lot of hardware 10 meters from the stock exchange servers, there's not a lot of room there... – [zmbq](#) Jul 2, 2013 at 6:11

- 
- 2 Co-location is the new normal for years. And FPGA's run the algo traders, in hardware. – [bbaassssiiee](#) Feb 27, 2015 at 6:05
-