The Fed Robbery Revisited

Riadh Zaatour and Stéphane Tyč

Quincy Data

December 04, 2013

Abstract

We revisit the controversial release of FOMC news on September 18th 2013 using the time stamps generated by the exchanges and included in their raw feed. Our analysis is consistent with a simultaneous release of the information at the CME and at the Nasdaq colocation centres exactly at 2:00:00 pm.

Introduction

A heated debate followed the FOMC "no taper" information release of September 18th. This controversy was started in a paper by Nanex, a market data research company. In a study called "Einstein and The Great Fed Robbery"[1] Nanex asserts that markets in Chicago and New York reacted almost simultaneously to the Fed announcement. This is incompatible with the time needed to move information from Washington DC to either city at the speed of light. Nanex concluded that *There are 2 possibilities, and both aren't good news for Wall Street.*, alleging foul play. The analysis of Nanex was then criticised by Virtu Financial who pointed out inaccuracies in their method[2]. The FIA PTG[3] issued a statement saying that the rules governing the release of news were, presumably, not violated[4]. The FIA PTG further states that the method of releasing the information by the Federal Reserve had been changed from *lock up* to *embargo* in March of 2013 and that the *simultaneous release of the Fed announcement at 2:00 pm ET sharp in Chicago and New York* was to be expected.

We revisit this debate with our own data sources. We begin by presenting the data and especially the timestamps involved in the study, then we identify market activity right after 2:00 pm on September 18th. We then try to determine the release mode for the FOMC announcement of October 30, for which some changes in the release mechanism were reportedly made by the Fed[5]. We also open the discussion on the different methods of releasing news.

Sequence of events and available data

When the Fed's announcement is released, it is first made available to a select group of (presumably accredited) news providers. Then those providers compete to send it to their clients for use in each of the colocation centres they serve. Before the release time of 2:00 pm, the news is given in the form of a text which must be transformed into a machine readable format. The news release mechanism provides time for this transformation to occur in a locked up room. When the information has been transformed into bits, then those bits are sent through fibre or wireless and the rest of the story is purely handled by computers. This handover, from news agencies to their clients, can occur just at the gate of the lock up room at exactly 2:00 pm and this will subsequently be called the *lockup* news release. The news can also be transported by the news

[1] http://www.nanex.net/aqck2/4436.html

[2] http://www.virtu.com/news-22/study-of-federal-reserve.html

[3] Futures Industry Association Principal Traders Group

[4] http://www.futuresindustry.org/ptg/downloads/FIA-PTG-Statement-10-2-2013.pdf

[5] http://www.cnbc.com/id/101128781

agencies to various locations in the United States and then released at 2:00 pm. We'll call this *national embargo.* There can, likewise, be an *international embargo* where the news is released at 2:00 but simultaneously all over the world. We would like to know where the handover occurred and if it was consistent with rules then in effect.

The data that we have is the raw feed from the CME and the raw feed from Nasdaq. These feeds are time stamped by the exchanges in their internal matching processes. The CME feed is time stamped with a least significant digit of 1 millisecond and the Nasdaq feed is time stamped with a least significant digit of 1 nanosecond. In the following we assume that the Nasdaq has atomically synchronised time and that its internal time precision is of the order of 1 microsecond (μs) precision. This assumption is reasonable given the quality and criticality of the systems involved. Our analysis of the internal timestamps of the CME shows some variation between the timestamps of the internal CME messages and the time of their recording on our machines. This variation could be explained by a difference in publishing time for different CME channels or by a difference in the synchronisation of the internal CME clocks. However, all the conclusions presented here stand within this clock precision and, in particular, a $100\mu s$ jitter would not alter our conclusions. We collect this data respectively in Aurora and in Carteret, the colocation centres for both matching engines. Our timestamps are recorded to $1\mu s$ granularity and our servers are not atomically synchronised to this precision. Our servers are synchronised with a simple NTP mechanism and the accuracy is of the order or one millisecond. However, the clock of our servers does not drift significantly over a few seconds. We only rely on relative time in our records and the local clock drift is insignificant at the level of this analysis.

September 18th Data

The FOMC announcement on September 18th 2013 is particularly interesting because it was very different from the consensus. When announcements confirm what the consensus is predicting, the subsequent market move is small, as most of the information was already incorporated in the prices. The large move triggered by the FOMC news release is easily identified in the subsequent market activity and the precise time of its occurence can be read from the market data.

Nasdaq

Nasdaq data can be studied directly as the market timestamp has nanosecond granularity and very probably microsecond accuracy. Trades realised on the gold ETF, GLD after 2 pm are in the following table :

Table 1: Nasdaq GLD Activity

Timestamp	Volume	Price
14:00:00.000330613	100	126.83
14:00:00.000330613	100	126.83
14:00:00.000330613	100	126.83
14:00:00.000330613	3	126.84
14:00:00.000330613	100	127.00
14:00:00.000330613	100	127.09
14:00:00.000330613	100	127.42
14:00:00.000330613	100	127.91
14:00:00.000330613	500	127.91
14:00:00.000330613	1000	127.98
14:00:00.162620675	124	127.39
14:00:00.162620675	100	127.82
14:00:00.162620675	733	127.99
14:00:00.178121355	100	127.99

Note. Trades on GLD during the first 3 seconds after the no taper announcement

Notice that most of the trading activity takes place on the same nanosecond timestamp 14:00:00.000330613, with a large cumulated volume and an increase of 0.91% in the price. Subsequent trades take place only more than 160ms after, with a negligible cumulated impact on the price. All those trades are probably spawned by a small number of large aggressor trades with a highest limit price of 127.98. These orders took all the resting orders in the book. This is a very good example of the winner takes all game. There were probably many more trying to buy at almost the same time but only the first few executed their trades. After the first three matching events there is a long lull in activity lasting up to 3 seconds. There are no trades until a new equilibrium is reached.

As for SPY, 155 trades are timestamped 14:00:00.000390009, with a volume of 60726 and a price increase of 0.39%. Subsequent trades take place more than 10 ms after 2 pm.

Table 2: Nasdaq SPY Activity

Timestamp	Volume	Price
14:00:00.000390009	500	170.83
14:00:00.000390009	505	170.84
14:00:00.000390009	5704	170.85
14:00:00.000390009	2605	170.86
14:00:00.000390009	4005	170.87
14:00:00.000390009	4404	170.88
14:00:00.000390009	3703	170.89
14:00:00.000390009	3903	170.90
14:00:00.000390009	4003	170.91
14:00:00.000390009	2603	170.92
14:00:00.000390009	2602	170.93
14:00:00.000390009	26189	170.94 ↗ 171.49
14:00:00.010486893	1179	171.00
14:00:00.017643111	1	170.84
14:00:00.029170164	1179	171.00

Note. Summary of the trades on SPY during the 30 ms after the no taper announcement.

In both cases, regarding the volume and impact of the trades, these are clearly motivated by the no taper information. It is therefore clear that this information was received by the servers which submitted the orders before 2 pm + 330 microseconds. It takes some time for the servers to receive the data and submit an order and it takes some more time for the matching engine of the exchange to process the trades and to publish the results in the feed that we have captured. Therefore, a release of the information at 2:00 pm in Washington DC can be ruled out. It is possible that the information was released early in DC or it could have been released at 2:00 pm in New Jersey. In order to distinguish between those two possibilities let's turn to the study of the CME data. If the information was released in DC albeit early, then the arrival time in NJ and at the CME should be separated by the time it takes to move the information to the respective data centres.

The difference in time can be estimated simply by taking the difference in distance over the great circle divided by the speed of light. The existence of fully functional microwave networks between DC and Chicago and New Jersey is highly probable. The following picture of a subset of the FCC registered microwave paths is good circumstantial evidence of their existence.

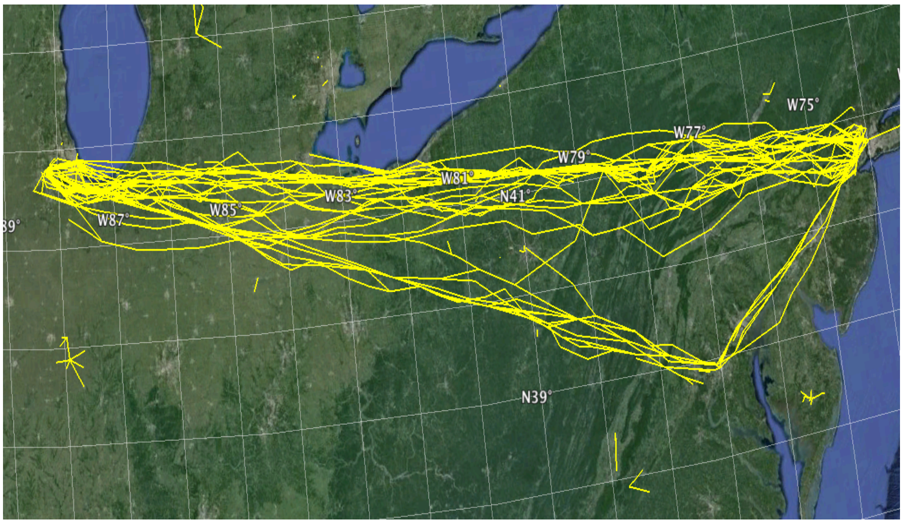

Figure 1: A subset of FCC frequency applications identified by McKay Brothers as probable low latency networks

The difference in distance divided by the speed of light yield a time delay of 2.34ms. To estimate this we took the distance as the crow flies between the K street data centre in Washington DC and respectively the CME data centre in Aurora (624 miles) and the Nasdaq data centre in Carteret (188 miles).

CME

In the CME case, we also consider futures on Gold and E-mini index. Liquid maturities are December 2013 for both, so that studied symbols are ESZ3 (E-mini) and GCZ3 (Gold).

The market publisher timestamps the packets it sends up to millisecond precision. In the table below, we put the number of trade messages by millisecond for the first ten milliseconds after 2 pm. Timestamps here are those embedded in the market messages. Let us stress that during the 100ms before 2 pm, there was no trade on either GCZ3 or ESZ3.

Table 3: CME Activity

CME timestamp	GCZ3	ESZ3
14:00:00.000	0	0
14:00:00.001	20	0
14:00:00.002	0	104
14:00:00.003	0	0
14:00:00.004	0	0
14:00:00.005	43	88
14:00:00.006	0	194
14:00:00.007	0	0
14:00:00.008	23	0
14:00:00.009	0	126
14:00:00.010	0	0

Note. Number of trades on GCZ3 and ESZ3 for the first ten milliseconds after 2 pm.

The 20 trades on Gold futures, timestamped 1ms after 2 pm cause a price increase from 1313.5 to 1316.8. The 104 trades on the E-mini future, timestamped 2ms after 2 pm cause the price to increase from 1695.75 to 1696.50.

The first trade in Gold futures could have occurred at 2 pm + 1.999 milliseconds. In order to be consistent with the first trades on GLD at Nasdaq, those trades should have occurred later than 2 pm + 0.330 ms + 2.34 ms = 2.67ms. So we can assert that if the CME timestamps are sufficiently accurate, then a lockup release mechanism was not possible. We will try to test this assumption by looking at the different CME channels because they correspond to different time stamping machines. We have also assumed the difference of propagation time was simply the difference in the distance divided by the speed of light. However, the microwave networks from DC to the CME and to NJ could be of different quality and this could introduce another 100 microsecond of uncertainty.

In order to get a more precise picture, we need to have timestamps at microsecond precision. For that purpose, notice that we have two inputs : timestamps at millisecond precision embedded in market messages, and timestamps at microsecond precision (*pcap timestamps*) made by our own machines when they receive market messages. Whereas our machines are clearly not microsecond-synchronized with the market ones, it remains plausible to suppose that the clock difference may be considered constant during a short time interval of, say, ten milliseconds. As is explained in the appendix, this enable us to recover microsecond timestamps in the market clocks from our two mentioned inputs.

We obtain the following results, where we show the corrections made to the first millisecond where we find trades after no taper announcement:

Table 4: ESZ3 timestamps milliseconds after 2:00 pm

market timestamp	pcap timestamp	corrected timestamp	Number of messages
2	1.369	2.070	18
2	1.437	2.138	17
2	1.526	2.227	17
2	1.603	2.304	18
2	1.730	2.431	17
2	1.777	2.478	15
2	1.796	2.497	2

Table 5: GCZ3 timestamps milliseconds after 2:00 pm

market timestamp	pcap timestamp	corrected timestamp	Number of messages
1	1.286	1.621	12
1	1.387	1.722	7
1	1.664	1.999	1

As explained in the Appendix, the offset needed to align the times for GC is probably due, in part, to an inaccuracy in the synchronisation time on the CME side. However, the ES symbol was timestamped only $80\mu s$ later. So, the time of the first publication of a trade was 2ms after 2:00 pm with less than $200\mu s$ accuracy.

Now, with these refined timestamps for the ES symbol, the inconsistence in time that must be explained is 2.67 - 2.070 = 0.600 ms. This is too large to be explained the uncertainty in our assumptions or by measurement errors. We can safely conclude that the handover of the FOMC announcement did not occur in DC, it was not a *lockup* release.

Discussion

What happened?

The data is not consistent with a release in the lockup facility in Washington DC, even at a time slightly before 2:00 pm. The data can be explained by an embargoed release simultaneous at the data centres in Carteret and Aurora. However, it is interesting to see that the first trades are published by Nasdaq much sooner than by the CME. The whole matching to publishing process is probably faster at Nasdaq. The data could also have been released a little late in Aurora. Maybe a release time of one millisecond after 2:00 pm could explain the first matching of Gold futures at 1.621 milliseconds. However, we consider this hypothesis unlikely. It is very simple to have a synchronised time within 10 microseconds of the atomic time and, with many news providers competing to release the FOMC news, we believe this was released very sharply at 2:00 pm. The probable cause of delay is that the avalanche of orders slows some component of the CME system. It could be the gateways, the matching engine, the publisher or some other component. The first resulting publication of matching occurs 1.6 milliseconds after the submission of the first order to the market gateway.

Our conclusion is that the most likely scenario is that the FOMC data was released under embargo at 2:00 pm exactly both at Aurora and at Carteret.

Why is it so difficult to understand what happened?

Was the embargo release mechanism in conformance with the current rules? Nanex stated that it was not, FIA PTG said it was, what do the rule makers say? Eamon Jeavers from CNBC has a great statement about this :

> A second organisation also declined to say whether or not it transmitted data out of the lockup room before the deadline. Instead, Market News International, which is owned by the Deutsche Boerse Group, sent CNBC this statement: "MNI follows the rules set by the Fed as we do with all data releases."
>
> So what exactly were those Federal Reserve lockup rules? Were organisations allowed to transmit information out of the room before 2 p.m. or not? The Federal Reserve won't say?a Fed spokesman declined to answer that question from CNBC.[6]

The Fed declined to say anything to CNBC. So we do not know if this was actually allowed or not. It is difficult to understand what happened because the rules do not seem to be very

[6] http://www.cnbc.com/id/101062081

clear and are not publicly available. It is hard to understand why the Fed would not answer the questions by Eamon Jeavers.

What is the right mechanism to disseminate information?

The goal must be to have clear and enforceable rules that guarantee a level playing field in the sense that anyone can have access to the same information in the same way. It is also a worthy goal to reduce the price at which trading participants can have access to the information. Let's examine the different mechanisms.

Lockup

The lockup is the simplest to understand. The information is provided in advance to news organisations. This advance release is necessary to ensure that the information is well understood before it is disseminated. It is necessary to be able to ask questions and get clarifications if needed. Then all the journalists can prepare their reporting and their machine readable news. The connection with the outside world is opened at exactly 2:00 pm and the race to the markets begins. Let's think what it takes to really isolate a room from connectivity with the outside world.

- A Faraday cage.
- Anechoic sound absorption, and, of course only chemical toilets to prevent tapping on the plumbing.
- No windows.
- No physical network connection.

Then at 2:00 pm exactly the network connection is plugged or, even better, an optical mirror is used to toggle the fibre in a connectivity mode. Of course the rate at which the servers of the reporters can send messages (presumably UDP) to the aggregation switch must be throttled to avoid an embarrassing colapse.

All of this sounds feasible, actually quite easy. There would be a single point to police in addition to all the people in the Fed who know the news before hand. One drawback is that the race to move the information from DC to the various trading centres is quite onerous. To win this race, you have to subscribe to all the machine readable news feeds in DC and then compete to carry the information to the data centres. Another drawback is how should the television reporting be treated? If anyone is allowed out of the room before 2:00 pm to prepare fro the broadcast then it becomes difficult to physically ensure that there are not leaks.

National Embargo

In order to reduce the economics cost imposed on participants by the expensive race to move the information, one could decide that the information can be sent under embargo to any place in the United States. This sounds very easy, all news organisations can transport the information to any place in the USA. Then a set of technical guidelines is published and those organisations must obey the guidelines and keep evidence of their compliance.

The difficulty becomes in enforcing the rule in foreign countries. Maybe this is not a concern of the Fed. However, if a news were released early in Europe, it could find its way back in the US before the 2:00 pm time.

Market pause

All the regulated markets could have a mandatory trading halt for a short period around the announcement time. This would presumably protect the less careful and less savvy investor. It would impose an administrative burden on the markets but this would presumably be surmountable. It would also have the advantage of being easy to police and of reducing the cost of the race.

Why is anyone providing liquidity at this time?

There seems to be nothing to gain by providing liquidity at a time when prices are likely to move very rapidly. This is very similar to selling an option for a zero premium. If you are providing liquidity on the bid, either the market moves up and you do not get executed, or the market moves down and you buy at price which is too high. So why would anyone be providing markets at this time? If everyone was a rational, well informed economic agent and if the losses incurred were large enough for people to care, then nobody should be providing tight markets at this time. We can speculate as to why the bid/ask is not very large at the time of announcements.

- Trading is complicated and not everybody pays attention to all the details
- Some agents have an obligation to make markets at all time
- Some agents execute trades on behalf of others who are not sophisticated enough to give the proper instructions
- Economics is called the dismal science precisely because its assumptions of rationality are extremely unrealistic

The fact remains that if everybody was a rational agent we would not have to discuss this topic because there would be no money to be made by being fast.

What happened on the FOMC release of October 30, 2013?

It would be interesting to see if the Fed did actually change its rules for the release of October 30. However, the announcement was not a surprise and the markets did not move. So we have to wait for the next surprise to be able to detect in a reliable way the news release mechanism.

Conclusion

The FOMC decision not to taper was not released in a lockup mode according to our internal data and to the timestamps of the exchanges. It was most probably released in an embargoed mode simultaneously in many US data centres as suggested by the FIA PTG in its statement[7].

[7] http://www.futuresindustry.org/ptg/downloads/FIA-PTG-Statement-10-2-2013.pdf

Appendix

Algorithm to align the CME internal clocks and our clocks

We describe here the algorithm allowing to retrieve microsecond precision timestamps out of our inputs. The idea is simple, if you have a continuous stream of market messages with millisecond timestamps, you can detect very exactly the transition from on millisecond to another. Using this information and the microsecond granularity of our records its is possible to recover a microsecond granularity for the market timestamps. This is akin to using two rulers, one for the atomic precision the other one for microsecond granularity.
Recall that we deal with two time stamps :

- The time of receipt of each market data packet at microsecond granularity, timestamped by our clock.
- Timestamps embedded in the market data packet at millisecond granularity, timestamped by the market clock.

Our clock and the market one are not synchronised. Our goal here is to mix these two information in order to come with microsecond precision timestamps synchronised with market timestamps. A naive way of doing this is to calculate the best offset and keep it constant over 20 milliseconds. However, a simple offset is not sufficient to explain all the correspondence between the two clocks. The time it takes a market data packet to reach our servers from the time it was timestamped in the first place is not a constant. So, two trades with exactly the same millisecond timestamps from the CME could be received at more than one millisecond of interval on our servers. In order to accommodate for this and to still ensure that the function transforming the matching engine time into our time is simple, continuous and monotonous we take a piecewise affine model. We could make other reasonable assumptions and we would produce the same results. The actual transform is very close to a simple offset.

Let's define notations and note: t_i the market time stamps in millisecond, so that $t_{i+1} - t_i = 1ms$ for every integer i.

Then, for a fixed millisecond bin t_i, let $\tau_{i,j}$ be the packet time stamp (in microseconds) of any pcap embedding a market timestamp t_i. So the sets $\tau_{i,j}$'s for a fixed i can be empty if the concerned millisecond t_i does not contain any packet with a market time stamp inside it.

Our goal is then to come with a function f mapping microsecond pcap time stamps into microsecond market timestamps. The function is then ***increasing*** and such that:

$$t_i \leq f(\tau_{i,j}) < t_{i+1}.$$

We further suppose this function is ***continuous*** and ***piecewise affine***.
Therefore, knowing $T_i = f^{-1}(t_i)$ completely defines the function.

We look then for T_i's such that

$$\sum_{i=1}^{n} ((T_{i+1} - T_i) - (t_{i+1} - t_i))^2$$

is minimum, subject to the constraints

$$T_i \leq \tau_{i,j} < T_{i+1}$$

We make the optimisation on the first 20 ms following 2 pm. We have then 21 parameters. We have also performed the optimisation on the first 10ms to verify the robustness of the procedure and it yields very similar results.

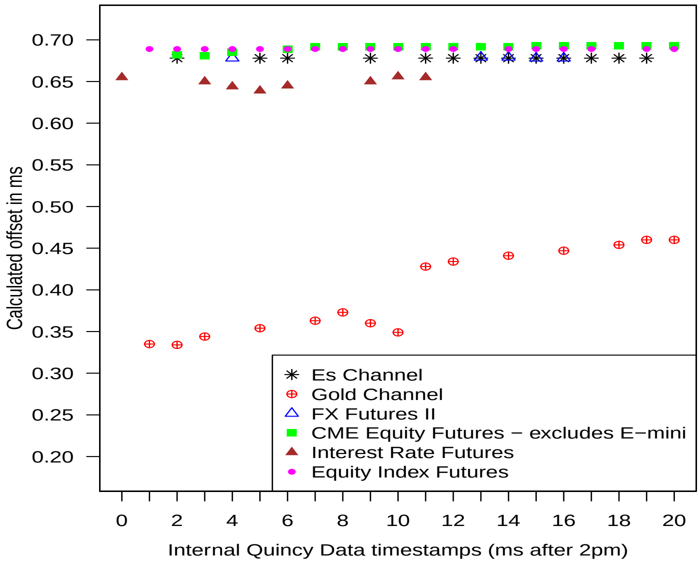

Figure 2: The offset between the CME publishing timestamp and the internal Quincy time stamps for several channels

The offsets have two components. The first one is the difference in clock synchronisation. It is constant over the time interval of 20ms. The second is the processing and propagation time between the timestamps of the CME and the reception on the Quincy Data servers. This second time can vary with the load of the CME system. The Gold channel seems to suffer from some kind of load and is also probably not synchronised with the other channels to a very high precision. We have relied on the results from the other channels in our discussion.

The CME events per channel

In order to illustrate differently the synchronisation of our clocks with the internal clocks of the CME we present the messages on the different channels as a function of their recording time on our servers. The horizontal axis differs from the absolute time by a simple offset.

This graph shows that, indeed, the gold future GL was the first contract which was published as traded. It does not mean that the orders were sent first to the matching engine, it only means that the CME published those first. Then the E-mini trades were published about $80\mu s$ later. One interesting fact to notice is that the same offset exists on the Nasdaq side. The gold ETF was traded about $60\mu s$ earlier than the SPY ETF. This could be explained by a faster algorithm on Gold compared to the S&P index.

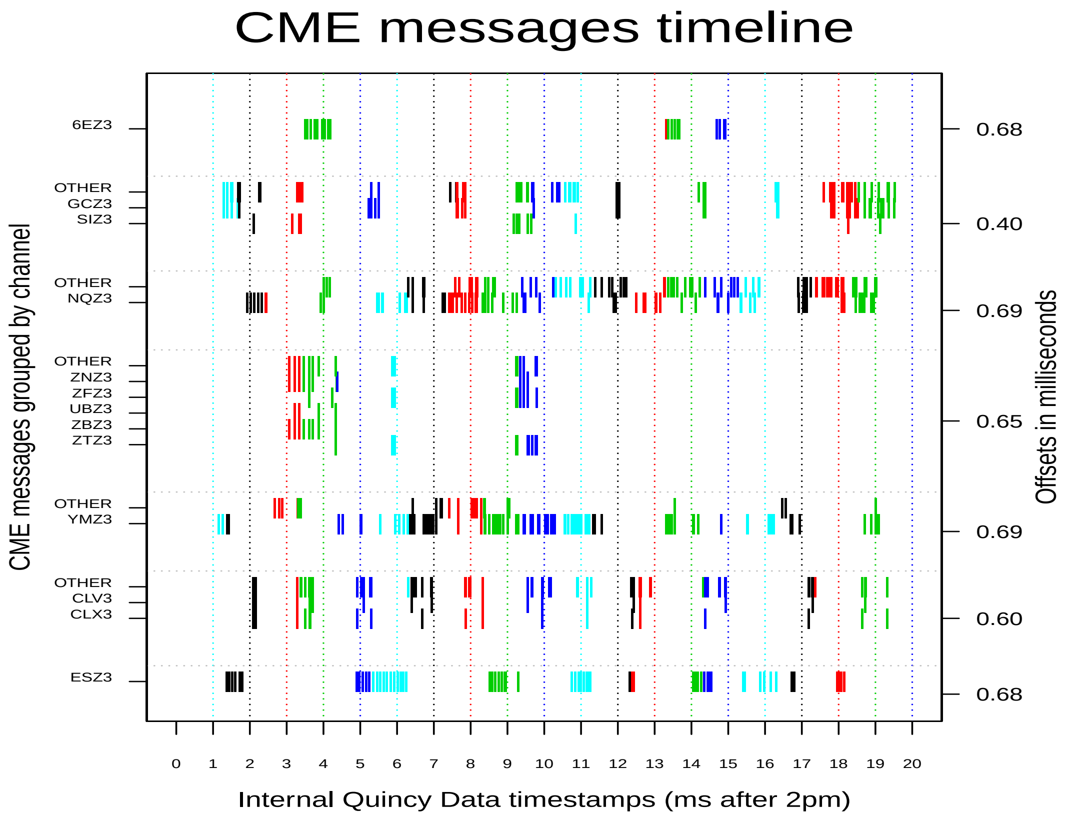

Figure 3: Time line of the recording of events on several CME channels.