Linux-RT in Financial Markets

Adrien Mahieux

Performance Engineer
Orness
gh: github.com/Saruspete
tw: @Saruspete
em: adrien.mahieux@gmail.com

BOFH - $(uname -a)

Adrien Mahieux

Orness
Performance Engineer aka Microsecond hunter

How I’m seen from:

- HW Vendor: Client with funny cases, that may want this new feature
- Kernel Dev: Client with weird requests that breaks his model
- App Dev: The one that refuses “quick deployment” in prod (“No-as-a-Service”)
- End-Users: Why is my Linux desktop not working?!

Agenda

- Linus Torvalds: Controlling a laser with Linux is crazy, but everyone in this room is crazy in his own way. So if you want to use Linux to control an industrial welding laser, I have no problem with your using PREEMPT_RT.

- Finance: Hold my beer

- What’s the Finance about ?
- Why would Finance actors need RealTime ?
- Real-time Challenges
- Monitoring and benchmarking

What’s the Finance about ?

Financial Markets 101

- Place where to exchange goods, currencies, insurances…
- Can trade **any type of product**:
  - **Physical**: commodities, currencies, goods…
  - **Virtual**: stock, index, insurance, futures, options..
  - ~~Rule 34:~~ As long as there's a **need** & a **supplier**, you can trade it
- Need to have a **Trading License**
- The market **takes a fee** for every transaction executed
- For an identical price, **First arrived, first served**
- You can **see the Order Book** (qties + price), but not "who"

Why would you trade

- Commodities (Fast-Food, beverage store, industrial...)
  - Need to buy every day physical goods (commodities) for your business to run
  - Need a delivery date and check product quality
  - Want to have a cover over steel in case of import tariff

- Investor (individual or corporate)
  - Stocks kept long-time (years...)
  - Plays in the tendencies of the market (people reactions)
  - May want an insurance in case of product fluctuations

Stock Market: from paper to electronic

Palais Brogniart 1973
Paris Stock Exchange

La Bourse, vue générale du marché, 1973.

Stock Market: from paper to electronic

Palais Brogniart 1997
Paris Stock Exchange

Stock Market: from paper to electronic

Palais Brogniart 2016
~~Paris Stock Exchange~~

Paris Stock Exchange is at Basildon, UK

Stock Market: from paper to electronic

Any Stock Exchange:

- Matching Engine
- Access Gateway
- Recorders

Stock Market: from paper to electronic

An history of technical optimizations

- Hire the **fastest runner** between “trading floor” (pit) to “trading offices”
- Went out of job with **Telex & Tickers** operated by secretaries
- 1983: Thomas Peterffy created the **first touch-screen tablet**: direct market access

- With reduced latencies, stock minimal price went from **25cts** to **1cts**.
- CoLocation: host your servers in the market’s DC.
- From Optical Fiber to Microwaves

Automation - New strategies created

Legal: taking advantage of technology

- **Arbitrage** : Play on different prices for the same product on different markets
- **Market Making**: Provide securities available elsewhere
- **Front Running**: Using faster infrastructure to buy first what others wants

Illegal : manipulating market / Generate fake events

- **Quote Stuffing**: send lots of quotes to slow down other people
- **Quote Stuffing(2)**: Place big sell orders (not 1st limit) to fake market pressure
- **Wash Trade**: Cross-trade: buy your own orders to generate fake volume
- **Smoking**: propose an interesting price, but cancel it immediately

HFT / High Frequency Trading

What is a “High Frequency” ?

Many definitions, no official one

Catch-All term for “**Automated**” or “**Algorithmic Trading**”
Current status: 80% of all trades made by bots

The problem: Bots reacts to events (really, any event…)
⇒ Risk of “**Flash Crash**” (amplification loop between bots)

Front Running

Use of a technical advantage to vampirise traders wanting to make a deal

Brad Katsuyama wanted to buy 100,000 AMD shares:

- 2006: Placed its order, got its 100,000 shares.
- 2007: Placed its order, got 80,000 shares
- 2008: Placed its order, got 70,000 shares
- 2009: Placed its order, got 45,000 shares

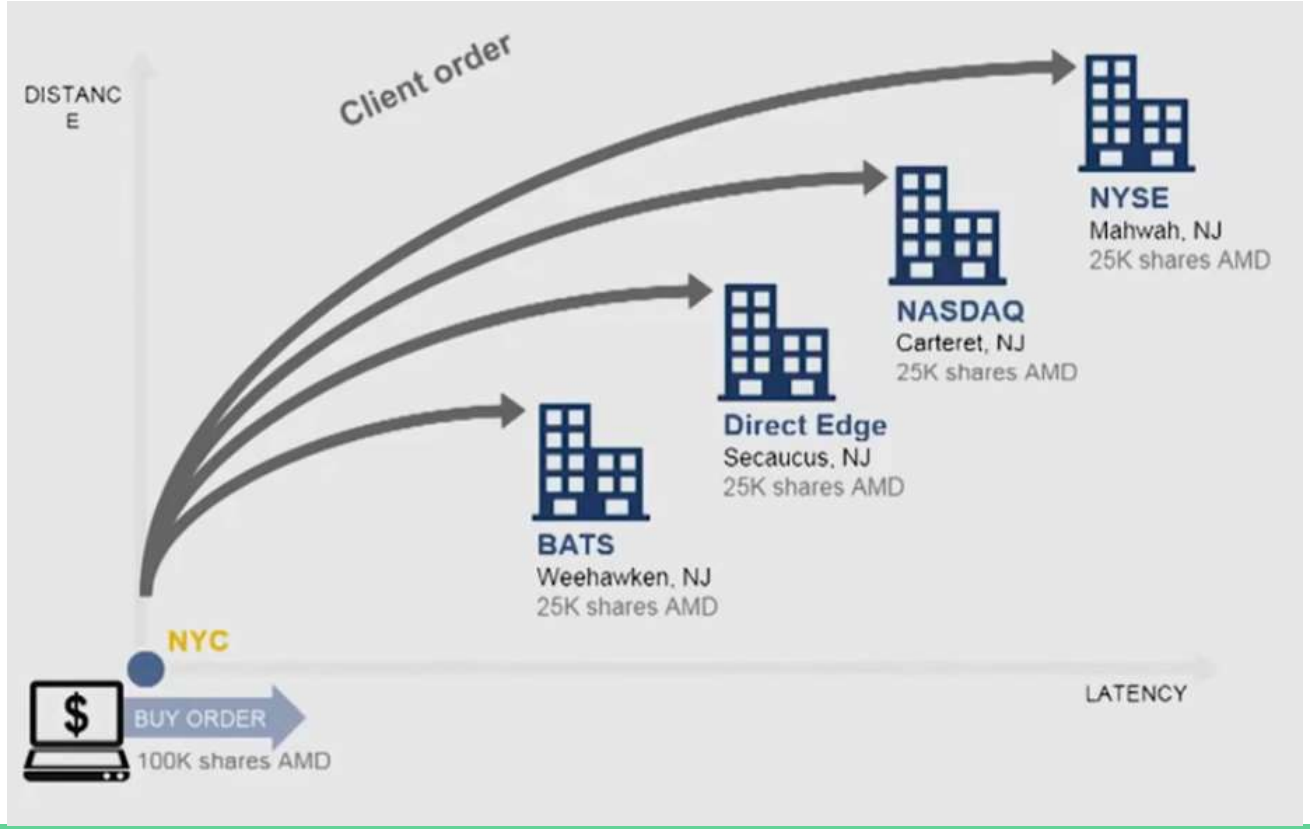
Maybe more people want to buy AMD at the same time?

Front Running

There’s about 13 stocks exchanges in US

Split 100K order in 4 * 25K

Brad latencies were **2ms** from first (BATS) to last (NYSE)

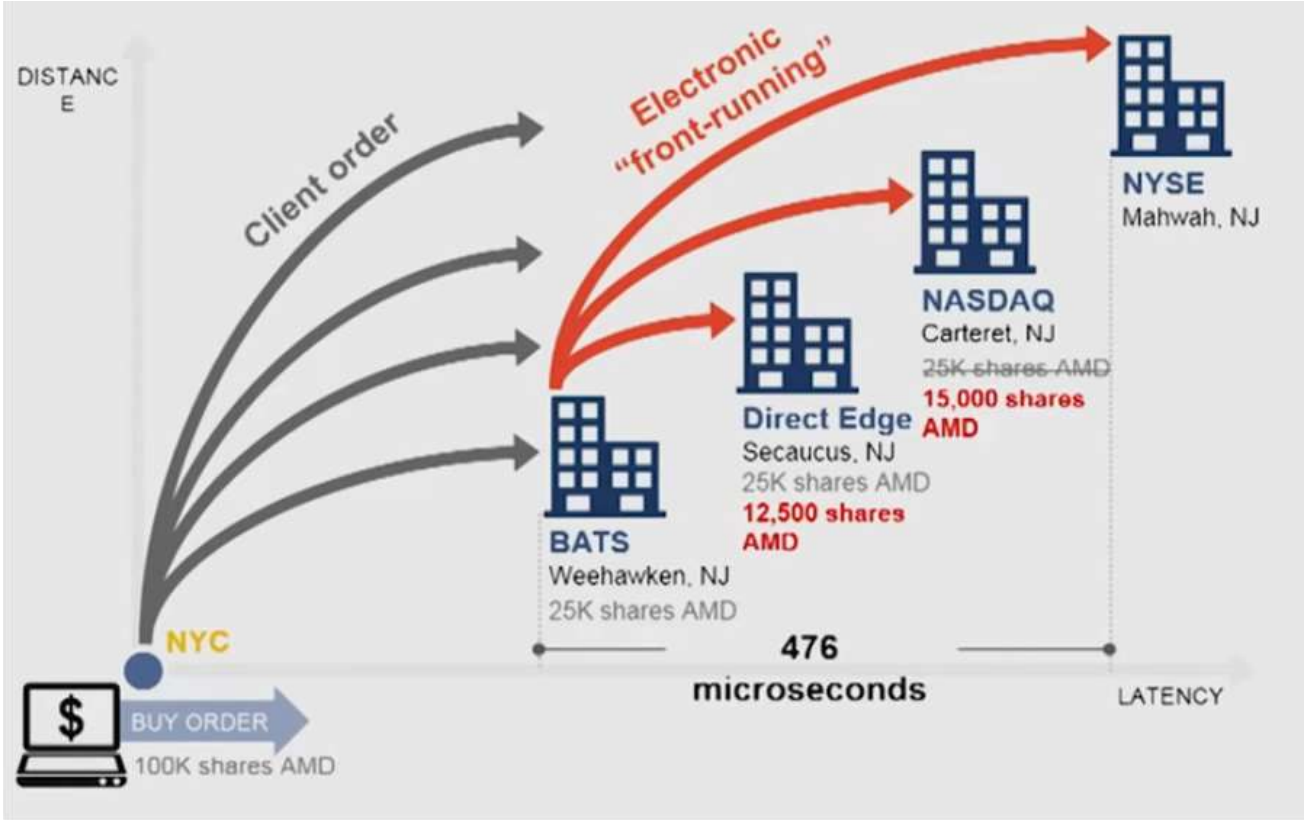


Front Running

Front-Runner with optimized infra is **0.5ms**

FR cancel their Sell order on the different markets

FR buy shares ahead of Brad and sell them back at a higher price (+1 cts)



Why would Finance need RealTime ?

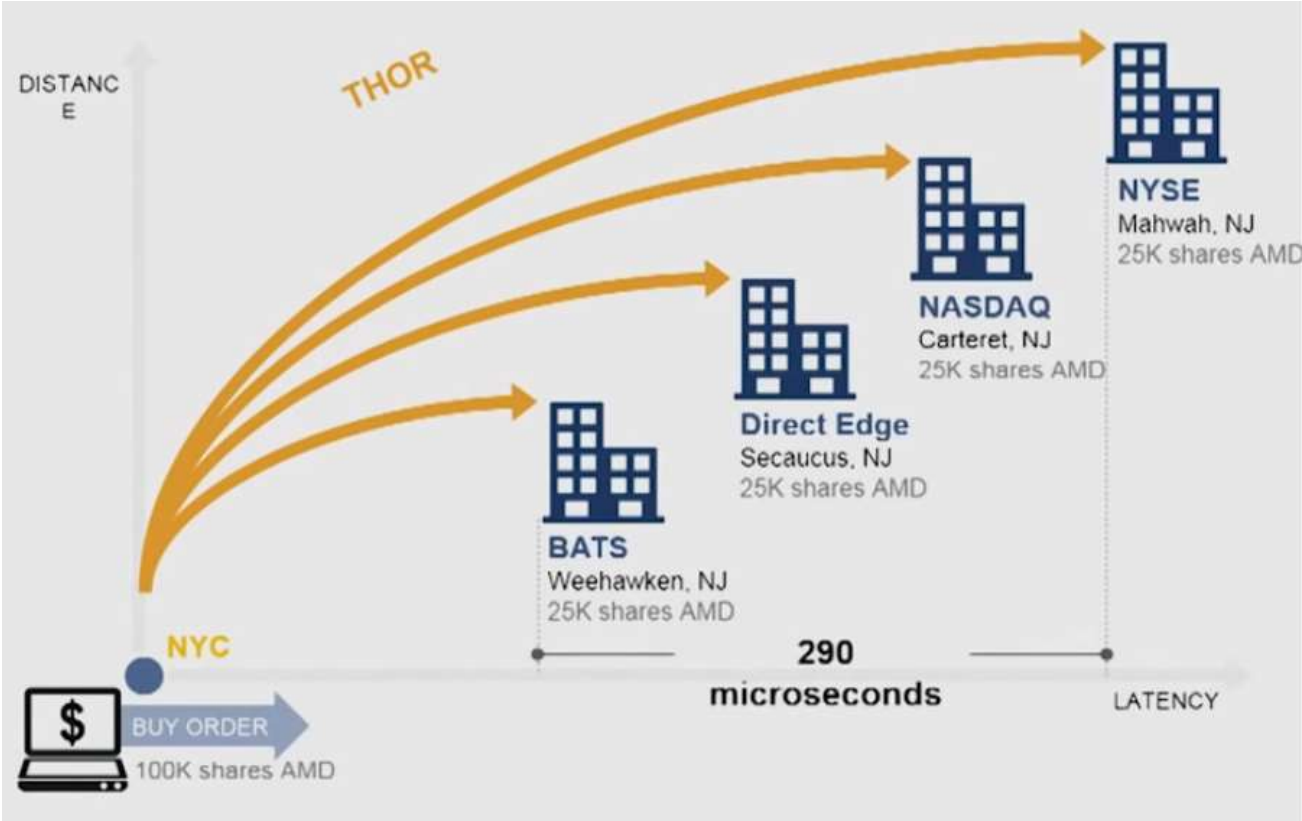
Front Running - Protect against it

Don’t send all orders at the same time: **slow down**

Delay sending based on the time needed to reach them

Closest locations will be sent after farther ones.

Goal: they all touch their target market at the same time



Regulations for Market Members

MIFID II requires Timestamp orders accuracy of **100µs**

Everyone wanting to acquire stocks must go through a “Market Member” (usually a Broker or a Bank) which has **extensive regulations** and risk analysis.

Market rules requires actors “to place an order with willingness to be executed”

Need accurate statistics for strategy backtest

Exchanges need fairness

A Stock Exchange is also a regulated entity.

Should ensure “**Fairness for all**” actors (not yet enforced)

- Co-Location hosting
- Same length of cable / Optical Fiber: 1m = 5ns
- Distributed Gateways to handle workload

Technical Challenges

x86 - What could go wrong ?

Tend to be Throughput optimized rather than Latency optimized

Interrupts: SMI, LOC, IPI, NMI

Frequency: Turbo Boost makes real freq vary greatly

Assembly: Transcribed to µops

Out Of Order Execution: need memory barriers

Compiler Optimization: Some neat tricks depending on options

NUMA: control placement of CPU, memory and cache snoop

x86 - CPU Isolation to reduce jitter

Goal: dedicate cores for specific processing and don’t disturb them ever.
Most frequent combination:

- **IRQ Isolation**: Move IRQs away from processing cores
- **isolcpus=** Change the default affinity of all userland process
- **nohz_full=** Don’t fire the scheduler interruption (thanks Frederic)
- **rcu_nocbs= & rcu_nocb_poll** Move away kthreads for RCU mechanism
- **intel_idle.max_cstate=1** disable CPU c-states
- **idle=poll** when really angry
- **noibrs noibpb nopti nospectre_v2 nospectre_v1 l1tf=off nospec_store_bypass_disable no_stf_barrier mds=off mitigations=off**

x86 - Application optimization

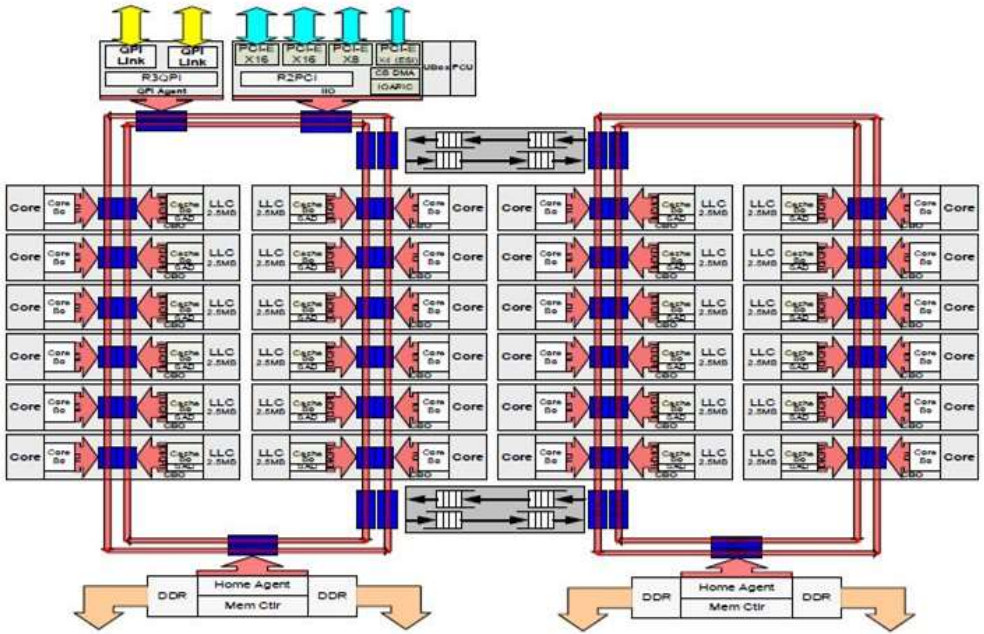
Application also has its share of work:

- **No memory allocation** in the critical path
- **Cache isolation** to avoid noisy neighbor (Intel CMT-CAT)
- **Keep cache hot** for critical path with fake events
- **Network Polling** by application (spinning 100% CPU)
- **Memory locking** to avoid swapping (mlock* sysctls in mm/mlock.c)
- **Hardware Timestamping** when available

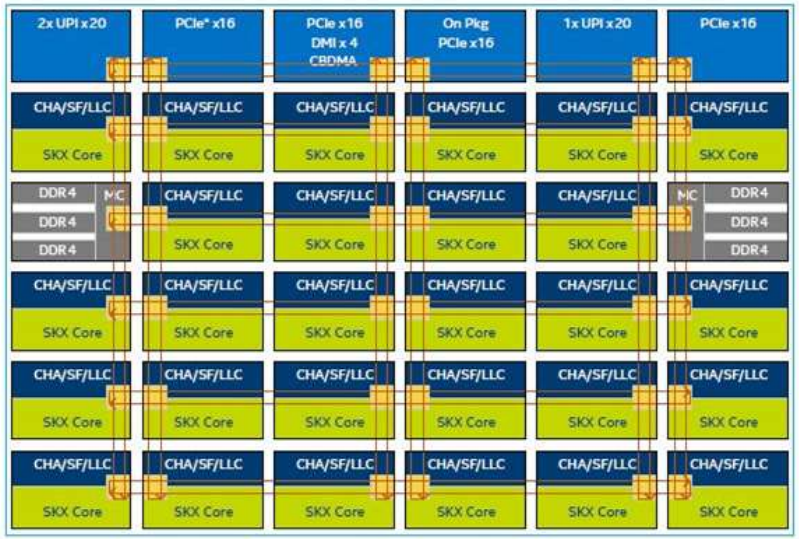
x86 - Hardware modifications

Core placement & Pinning is hard : ACPI Table, Firmware update, Model changes...

Broadwell EX 24-core die



Skylake-SP 28-core die



CHA – Caching and Home Agent ; SF – Snoop Filter; LLC – Last Level Cache ;
SKX Core – Skylake Server Core ; UPI – Intel® UltraPath Interconnect

Time Synchronisation

Needed for legal records, traceability and monitoring

NTP: sub-ms, very network dependant (from client to local server)

- Relies on server clock source (Often TSC)

PTP: sub µs, close to client thanks to boundary clocks:

- Smart switchs can offset their propagated values
- SmartNICs of final servers have an embedded oscillator

WhiteRabbit: sub ns, implemented by CERN and Eurex (among others)

Heavy network pps load

Awful kind of network workload:

- Packets < 256 bytes
- Spike only lasts a few milliseconds
- No cache miss allowed
- 10 Gbps at Pkt Size 1500 ~ 800.000 pps
- 10 Gbps at Pkt Size 128 ~ 10.000.000 pps

Better to just **bypass the Kernel** (DPDK, Onload, VMA)

Optical Fiber is slow : Go Microwave !

Optical Fiber:

- **⅔ of speed of light** (wire is not empty and beam bounces)
- **not straight** : has to follow ground
- Compact and secure channel

Microwave:

- **Full speed of light** in vacuum
- **Direct line of sight** (only need repeaters for earth curvature)
- But small packet size: 64 (raw sockets)
- Subject to weather, Marine Traffic and Solar flares

Optical Fiber is slow : Go Microwave !

Houtem tower in Belgium:
Auction started at 255K€
sold for 5M€

Direct line of sight from Houtem to England

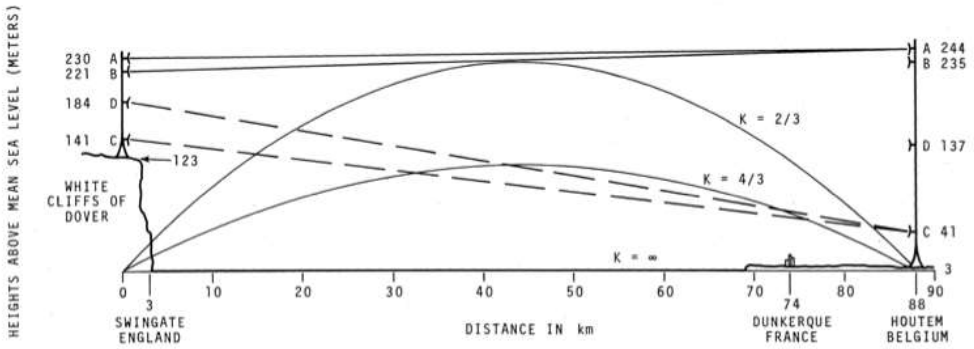


Figure 2. Terrain profile for 5 GHz microwave link.

Time requirements

IEX : Speed bump to protect from Front Runners

Best tech is **no tech**

A very long Optical Fiber adds 350µs latency

Monitoring & Profiling

Monitoring - Netdata

github.com/netdata/netdata

- 1s granularity
- Standalone
- Integrated alerting
- Lowest overhead monitoring
- 1000’s of metrics per second
- Self-explained configuration
- No conf needed to use

Profiling & Measurements

Measurement will have an impact

many TSC Flags: rdtscp, constant_tsc, nonstop_tsc, tsc_known_freq, tsc_adjust

CPU Perf Counters readable through MSR (Intel PCM: github.com/opcm/pcm)

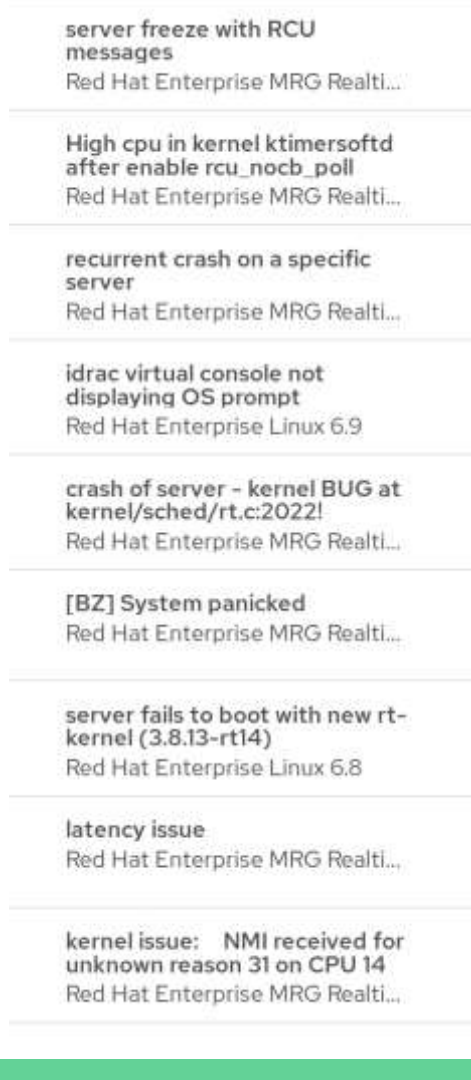
Traces when it goes bad

ftrace ring buffer on production + crash plugin (thanks Steven)

Had multiple issues involving NFS, RCU, RT_Mutex...

Most of the time, it was an application issue with 3 factors:

- Spinning
- Realtime (FF) Scheduling
- Bad affinity



Network - FPGA to the rescue

The best code is the one not run

FPGA are perfect for Latency optimized workloads
⇒ High end FPGA are expensive

Still needs monitoring software on the host
⇒ Relies on communication channels and drivers

Network - Wire 2 Wire

Metamako Layer 1 Switch

Fast active tap: 4ns

Profiling and Dev Tools

Opengrok - https://github.com/opengrok/opengrok
Fast code browser with indexing and ctags support

Compiler Explorer - https://godbolt.org
Show compiled ASM and matching source lines

Agner Tables - https://www.agner.org/optimize/instruction_tables.pdf
List of x86 instruction latencies per CPU (Intel, AMD, Via)

QuickBench - http://quick-bench.com

Benchmarks

Dynticks-Testing

LinuxRT Cyclictest

Solarflare Sysjitter

Bitmover lm_bench

The Best benchmark...

Is still to test in production !

Bibliography

Trading - Market Operations

Michael Lewis - Flash Boys - ISBN: 978-0393244663

Alexandre Laumonier “4” (ISBN: 978-2930601106) & “5|6” (ISBN: 978-2930601106)

The Wall Street Code - youtube.com/watch?v=kFQJNeQDDHA

Flash Crash - youtube.com/watch?v=aq1Ln1UCoEU

Les Nouveaux Loups de Wall Street - youtube.com/watch?v=0KNwcJgKMbo

Trading - Technical reads

www.algo-logic.com/sites/default/files/Algo_Logic_How_To_Build_An_Exchange.pdf
Timing White Rabbit at Eurex: https://www.eurexchange.com/exchange-en/resources/initiatives/technical-changes/high-precision-time-white-rabbit-pilot

Meanderful - meanderful.blogspot.com/
Sniper In Mahwah - sniperinmahwah.wordpress.com/
Microwave KMZ: https://sniperinmahwah.wordpress.com/2015/04/15/hft-in-my-backyard-the-map/

NYSE vs IEX https://www.businessinsider.fr/us/iex-vs-nyse-on-speed-bump-2017-3
Eurex Connection Gateways details: https://www.eurexchange.com/exchange-en/resources/initiatives/technical-changes/connection-gateways

Trading - Official Business Websites

IEX about NYSE FIX GW: https://iextrading.com/about/press/op-ed/

Cboe (BATS/CHIX) markets.cboe.com/europe/equities/market_share/index/cboe

Questions ?

Bonus: Finance is always trending

Trading Fashion - High Heels

First of all, you have to stand out from the crowd. That's why you see all those Technicolor jackets.

A lot of guys wore shoes with three-inch soles so they'd look taller. There was a guy in the Loop who made his living resoling shoes for traders.

Bibliography

Trading Fashion

Business Insider

Bloomberg

Twitter - SniperInMahwah