0への挑戦-フラッシュ・ボーイズ-

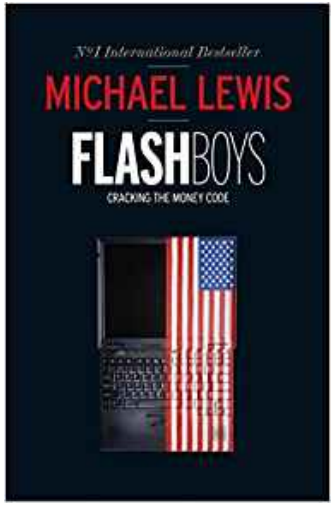
Shishio Tsuchiya
shtsuchi@arista.com

ARISTA

HFT

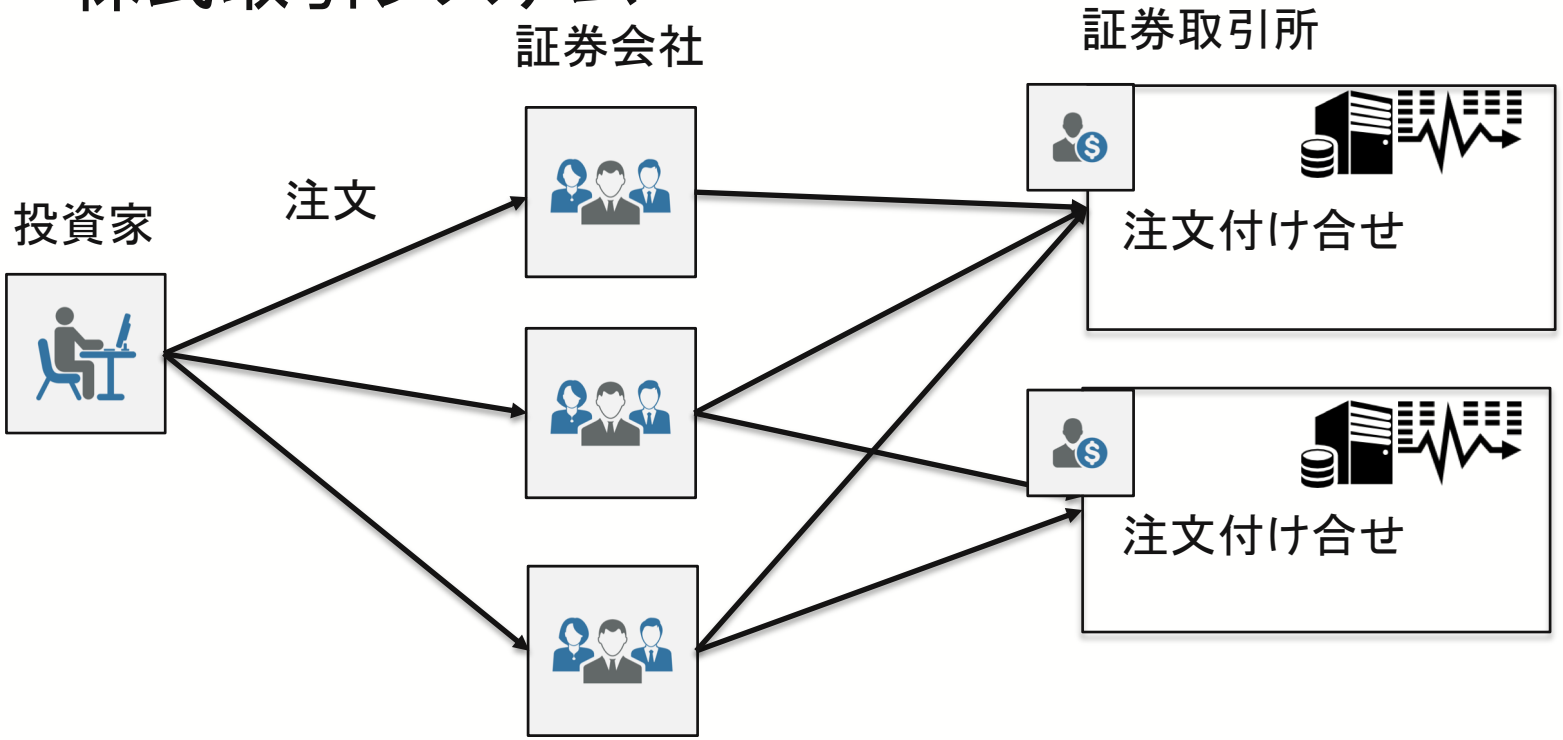
- 高頻度取引（high-frequency trading, HFT）とは、1秒に満たないミリ秒単位のような極めて短い時間の間に、コンピューターでの自動的な株のやり取り戦略を実施するシステムのこと。超高頻度取引、超高速取引とも呼ばれる。

フラッシュボーイズ



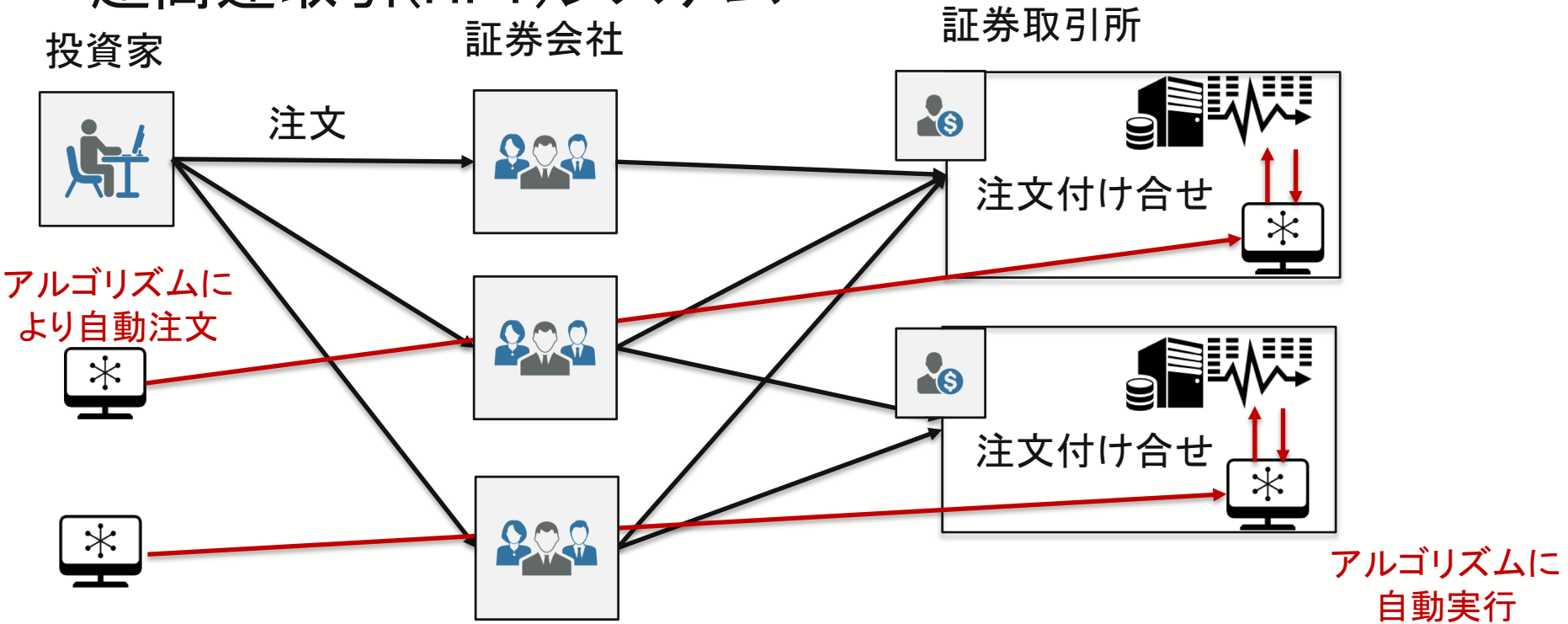
- 米国のノンフィクション作家マイケル・ルイス氏の高速高頻度取引(HFT)に焦点をあてた本
- あなたは1000円でA社の1万株を買おうとしたが次の瞬間1003円になってしまった
- それでも買いたいあなたは1003円で1万株購入した
- もし、あなたが1000円で買えていたらそれを1003円で売れたのかもしれない

株式取引システム



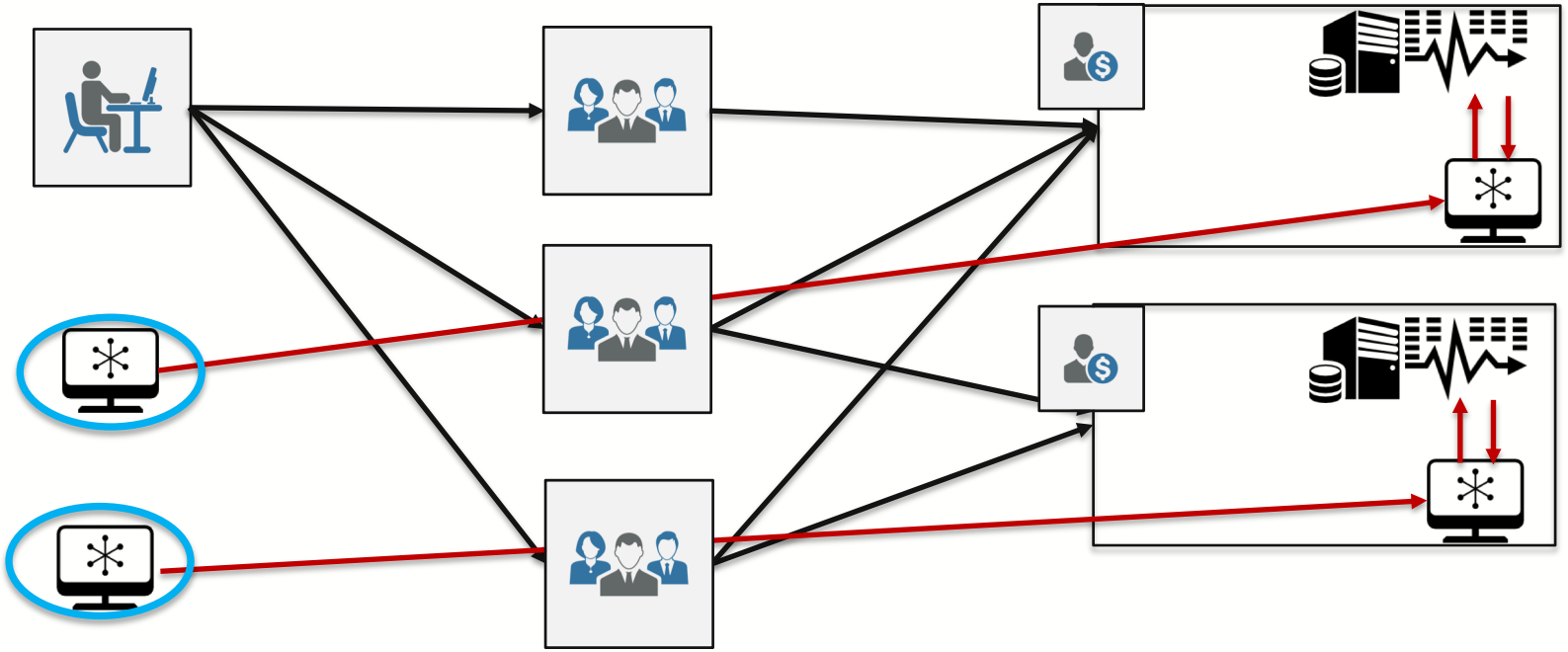
- 投資家は証券会社に売買を注文
- 証券取引所で注文付け合せを行う売買が成立

超高速取引(HFT)システム



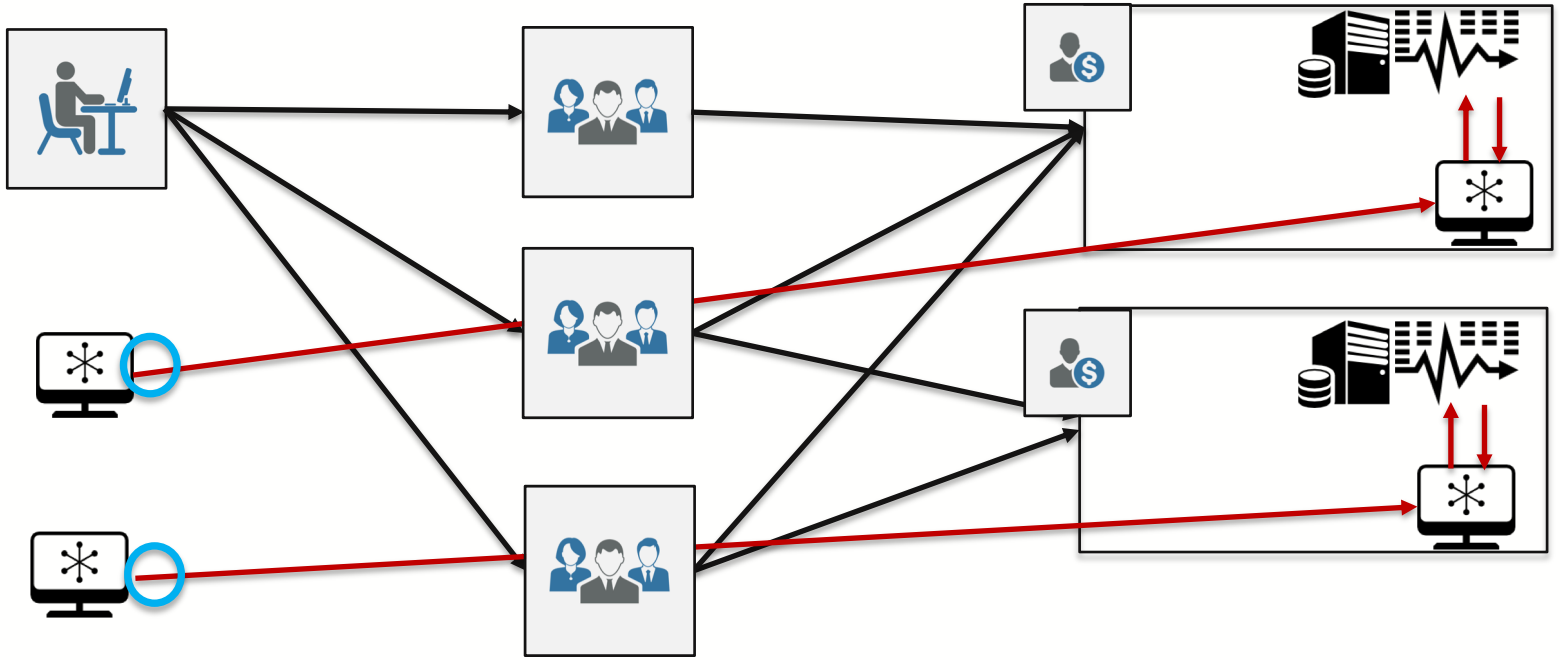
- 一定のアルゴリズムに基づき自動的に注文
- コロケーションエリアから自動実行
- トランザクションが高頻度で行われる/少しの遅れが大損害になる

超高速取引(HFT)システムの遅延要求



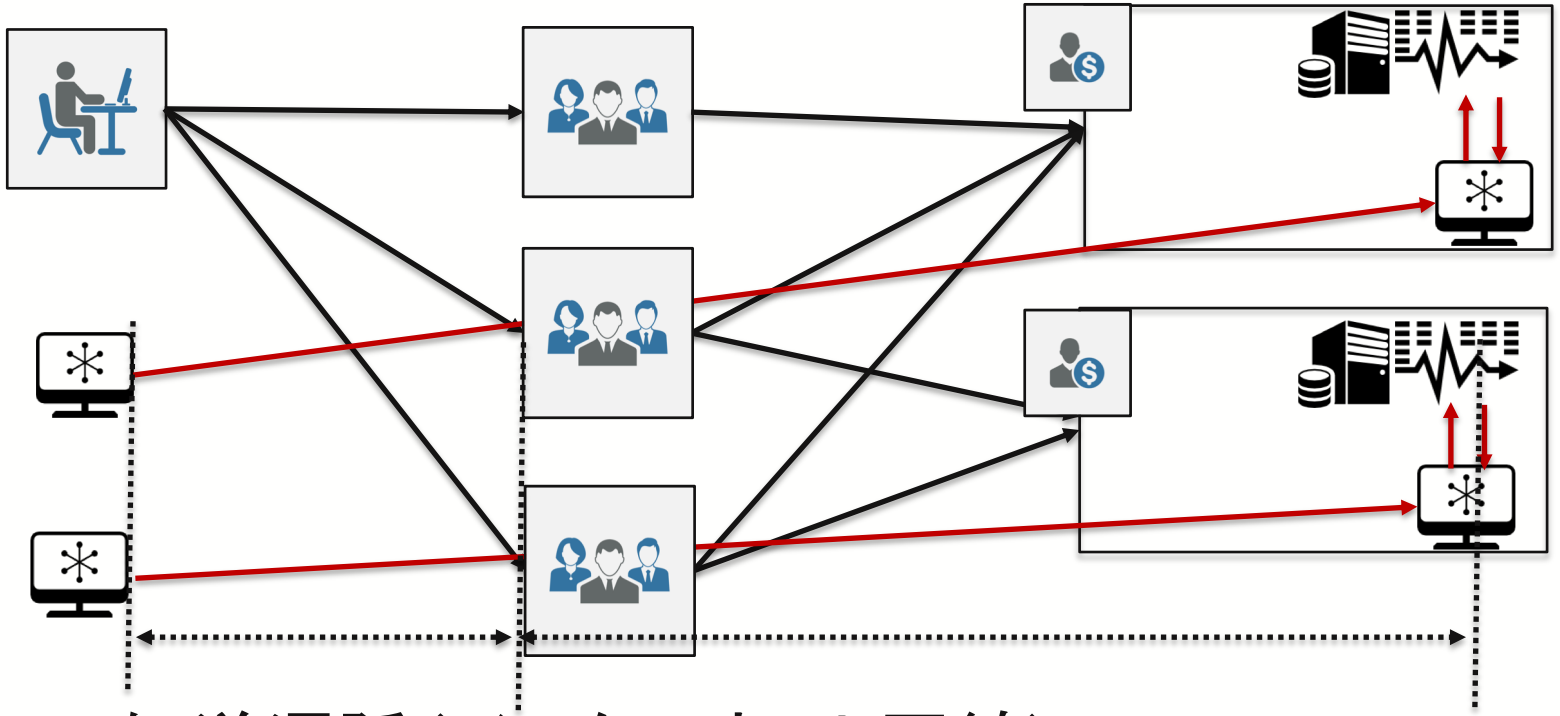
- 望む条件に応じて即座に注文

超高速取引(HFT)システムの遅延要求



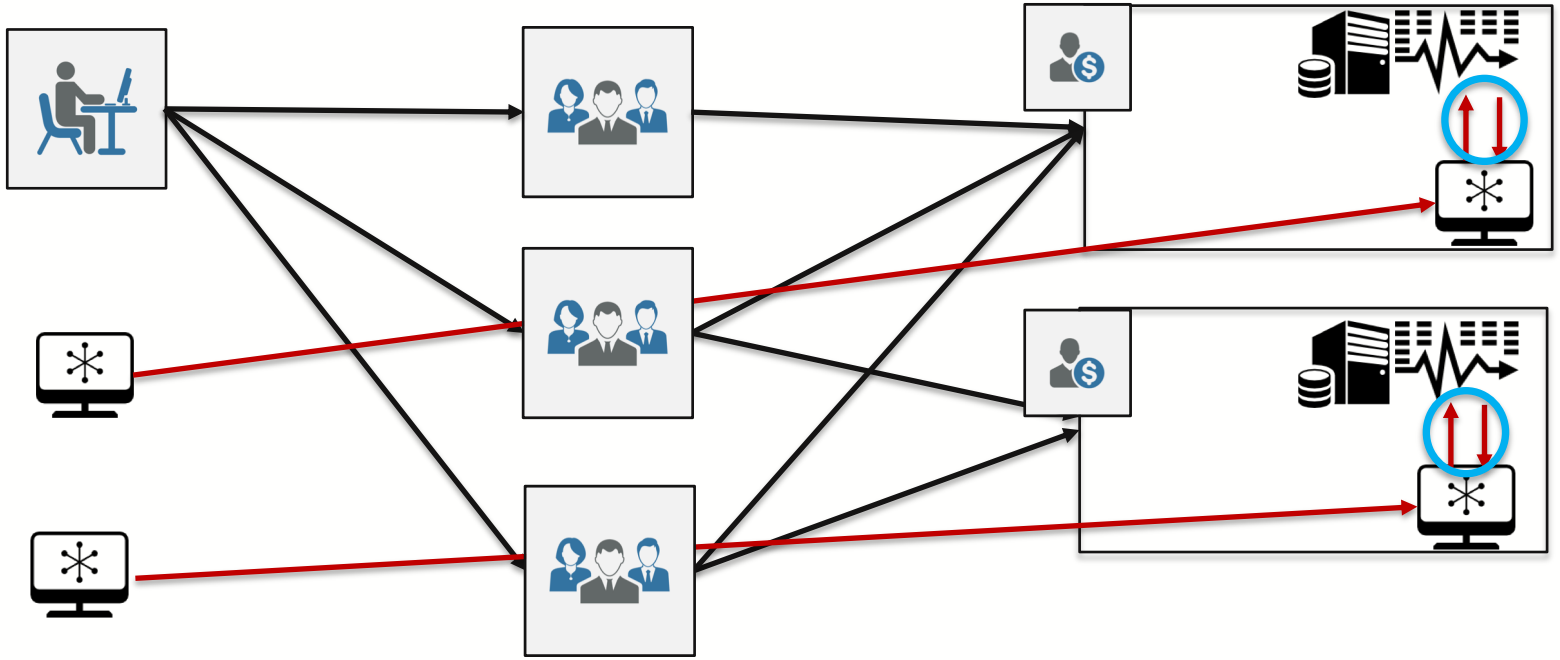
- アドレス変換/ルーティング

超高速取引(HFT)システムの遅延要求



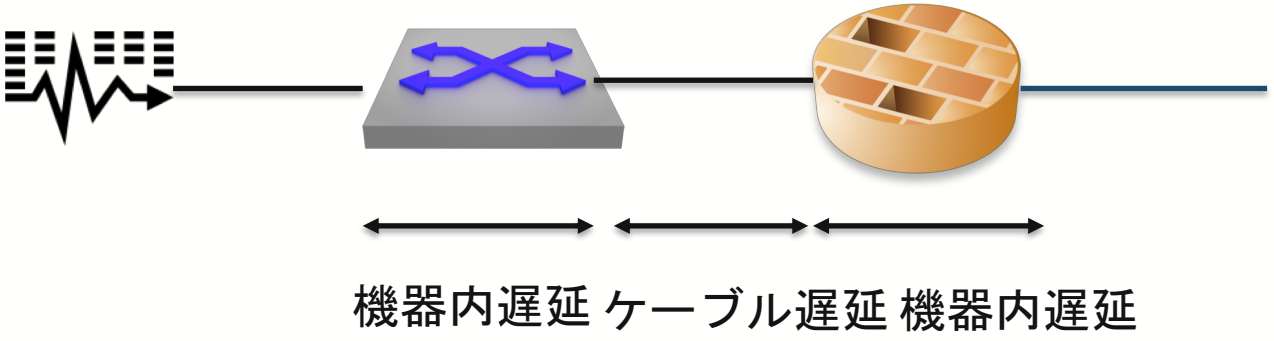
- 伝送遅延(インターネット回線)

超高速取引(HFT)システムの遅延要求



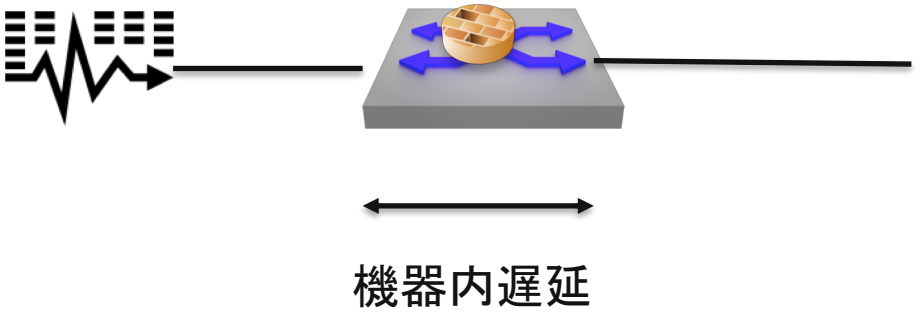
- 注文処理の実行遅延

とにかく高速で処理する機器が必要となる



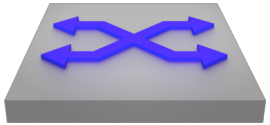
- NATデバイスは一般的に高遅延
- 処理ロジックはTCAMなどでプログラミング可能

とにかく高速で処理する機器が必要となる



- 高速なスイッチ上でNAT/BGP/マルチキャスト処理を行う

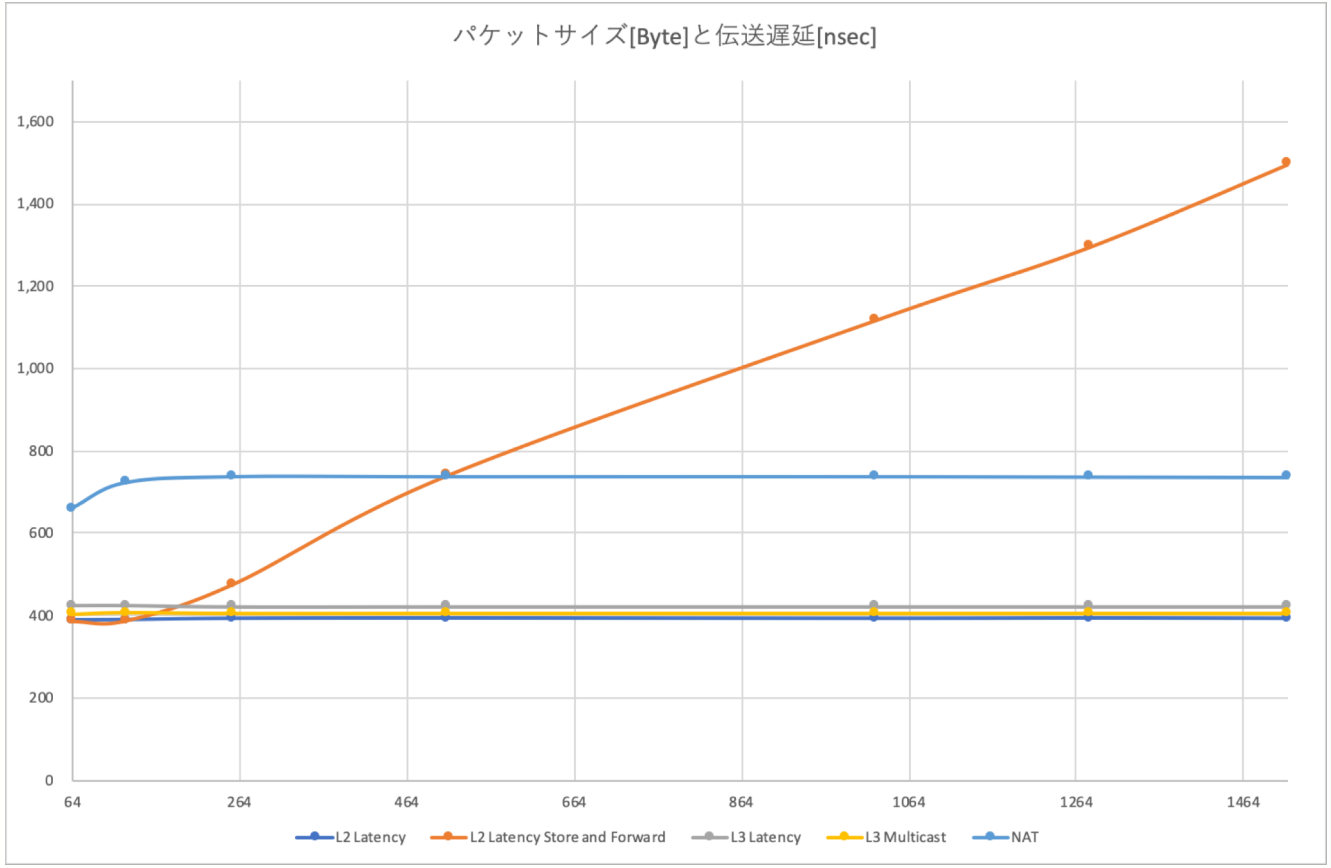
イーサネットフレームのどこを見るのか



プリアンブル (8バイト)	宛先MAC (6バイト)	送信MAC (6バイト)	イーサタイプ (2バイト)	データ	FCS (4バイト)

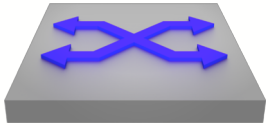
- **ストア&フォワード**:宛先MACをみて送信先を決定し、FCSでフレーム正常性を確認し、転送
- **カットスルー**:宛先MACを見て、即座に送信・イーサタイプや固定部分のヘッダーまで見るものもある
- フレーム長に関わらず伝送遅延は一定となる:約350ナノ秒(10^{-9}):**0.0000035**

パケットサイズと伝送遅延



- RFC2544サイズでの遅延時間を測定
- 同じ装置でも転送モードにより遅延時間の差分が明らか

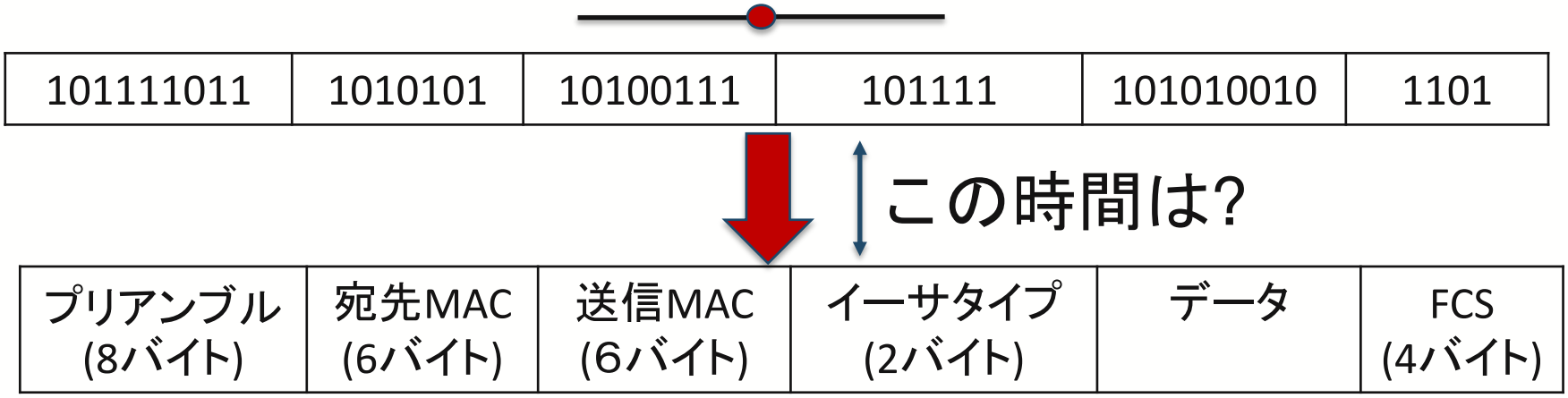
そもそもフレームとして見て良いのか?



プリアンブル (8バイト)	宛先MAC (6バイト)	送信MAC (6バイト)	イーサタイプ (2バイト)	データ	FCS (4バイト)
101111011	1010101	10100111	101111	101010010	1101

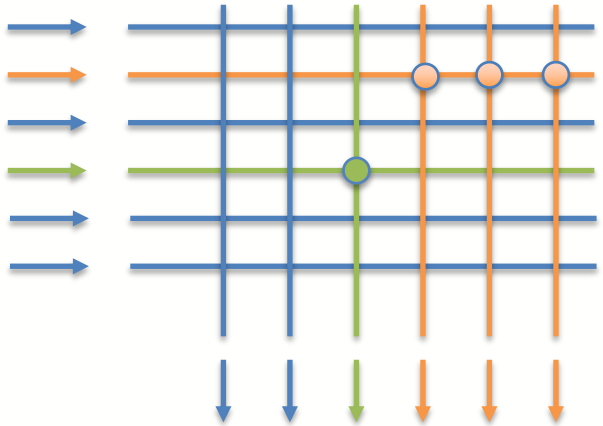
- もともとはただのビット列

そもそもフレームとして見て良いのか?



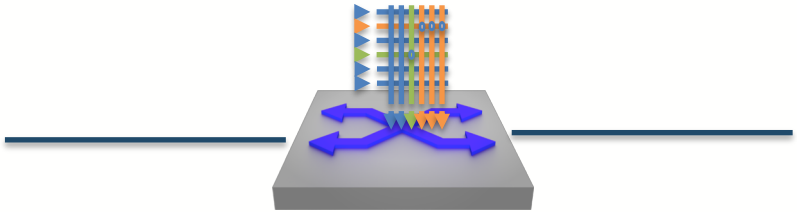
- 複雑な処理が必要であればフレームとして認識してよいが、処理が決まってるのであれば電気信号として見ても良いかもしれない

レイヤー1+スイッチという考え方



- あるポートで受信したものは別なポートに送信する
- ビットレベルで転送
- あるポートで受信したものを複数のポートに送信する事も可能(マルチキャスト的な送信)

レイヤー1+スイッチ

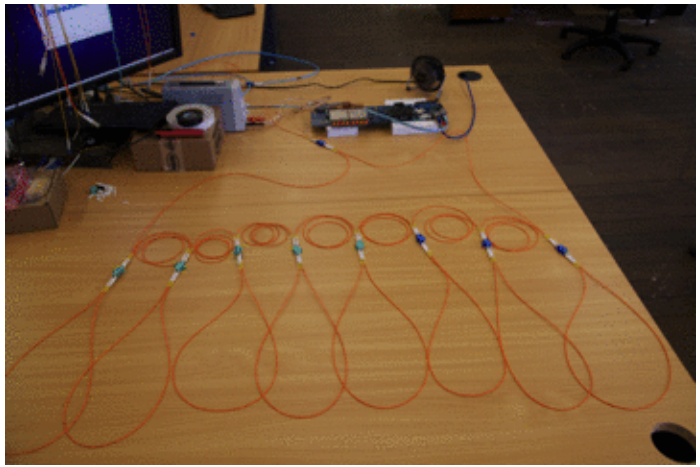


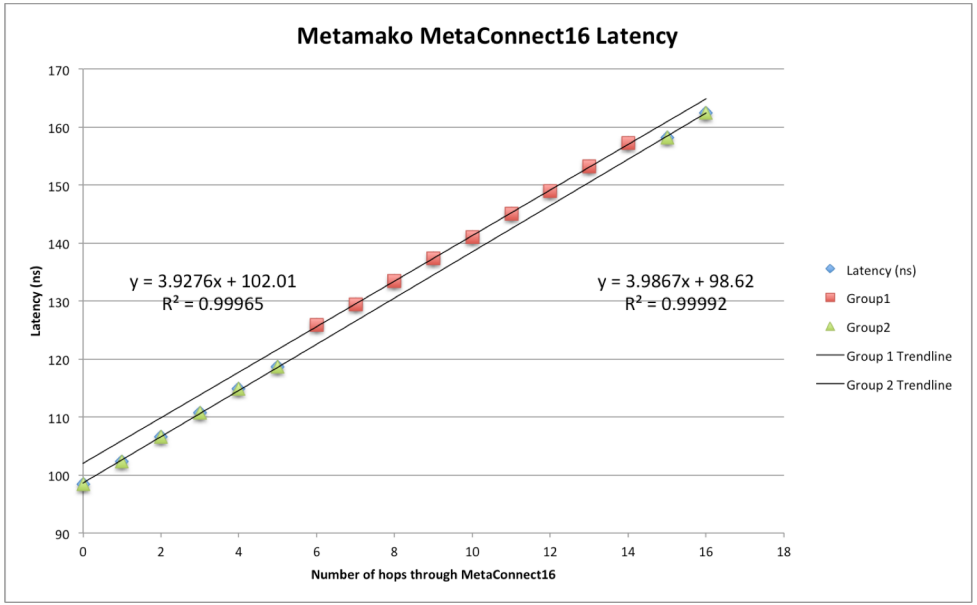
101111011	1010101	10100111	101111	101010010	1101

- 光減衰しない様にオプティクスは使用
- ビットをそのままポート間転送:5ナノ秒(10^{-9}):**0.00000005**
- この遅延時間は1メートルファイバーの伝送遅延とほぼ同等

そもそもどうやって5nsecの遅延を測定するか

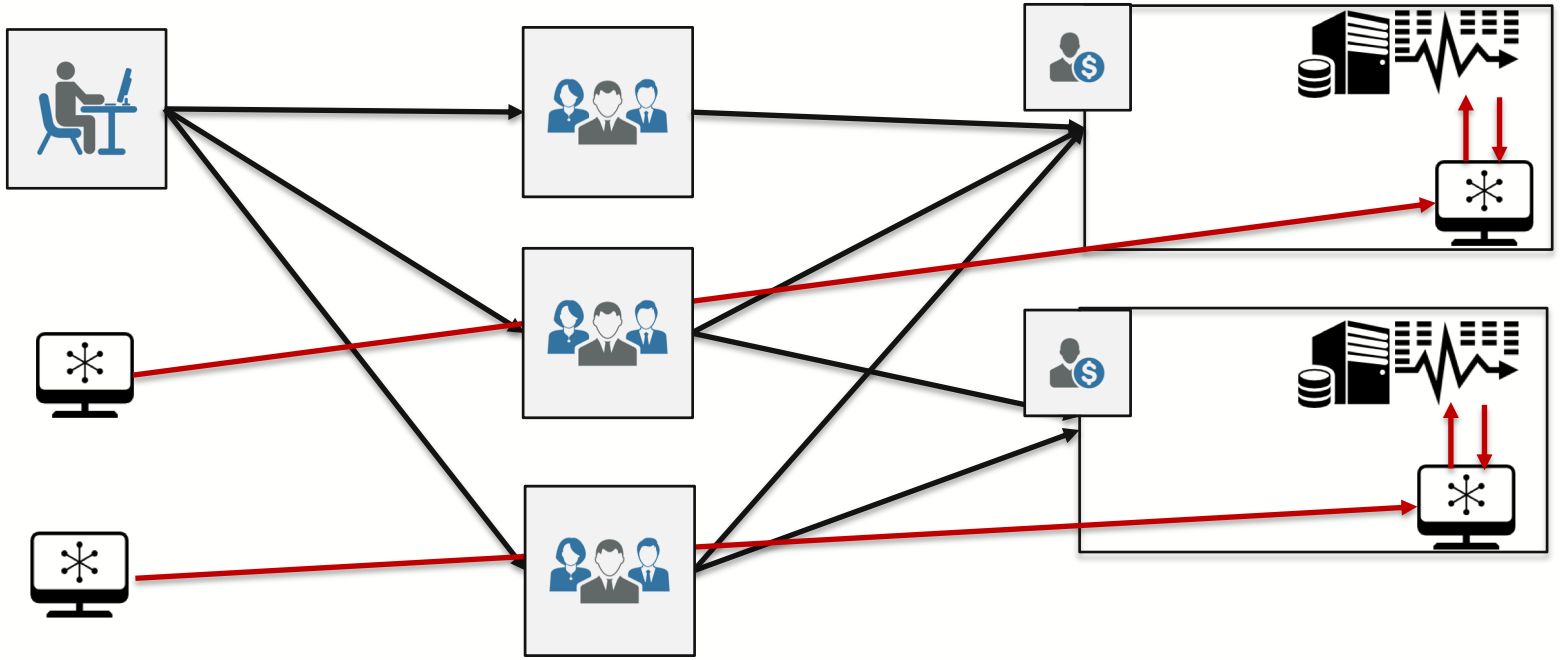
https://blog.metamako.com/how-to-measure-the-latency-of-a-4ns-switch





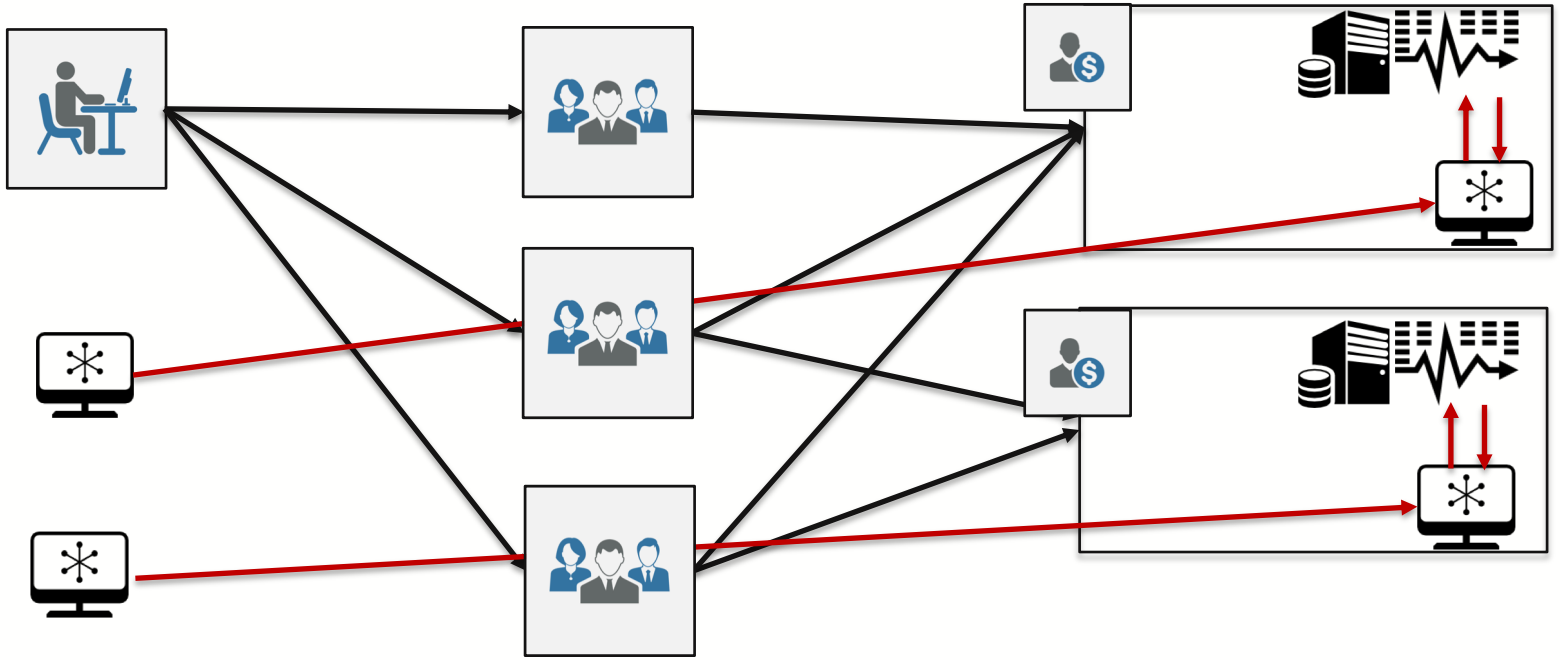
- ファイバーとカプラを接続
- 実際のパケットにハードウェアタイムスタンプをつけてLatencyを測定
- 少しづつカプラーを外し、L1+スイッチへ接続する

回線遅延 注文システム～証券取引所



- そもそもコロケーションをしてもらう必要がある
- とにかく近く

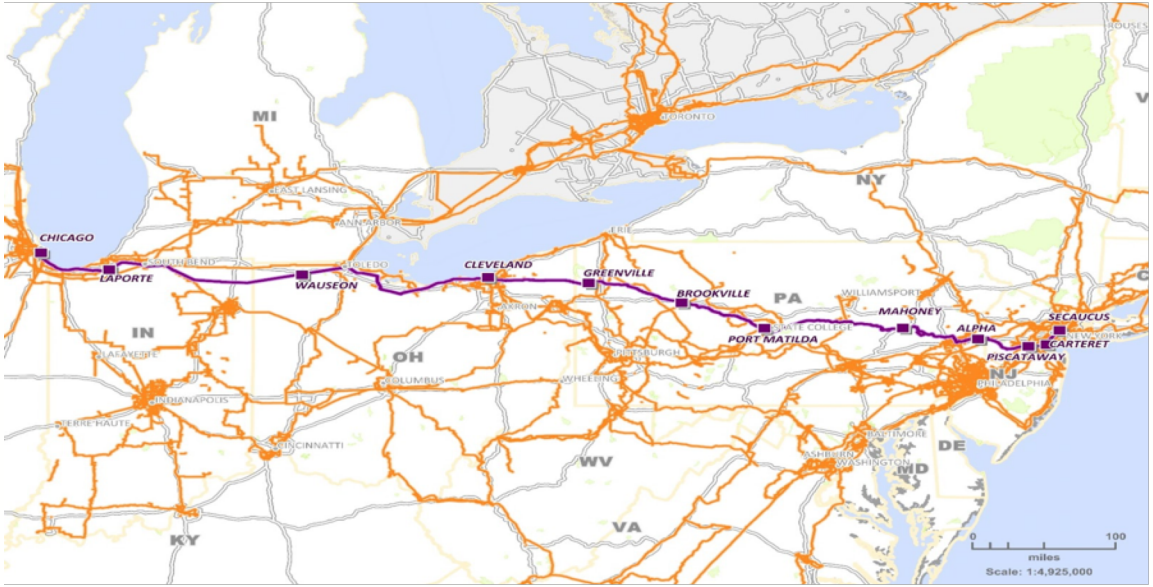
回線遅延 証券取引所間



- 動きを察知した後の関連株のオーダなどを他の証券取引所などにもすばやく行わければならない

SPREAD NETWORKS BY ZAYO

https://www.zayo.com/services/spread-networks/

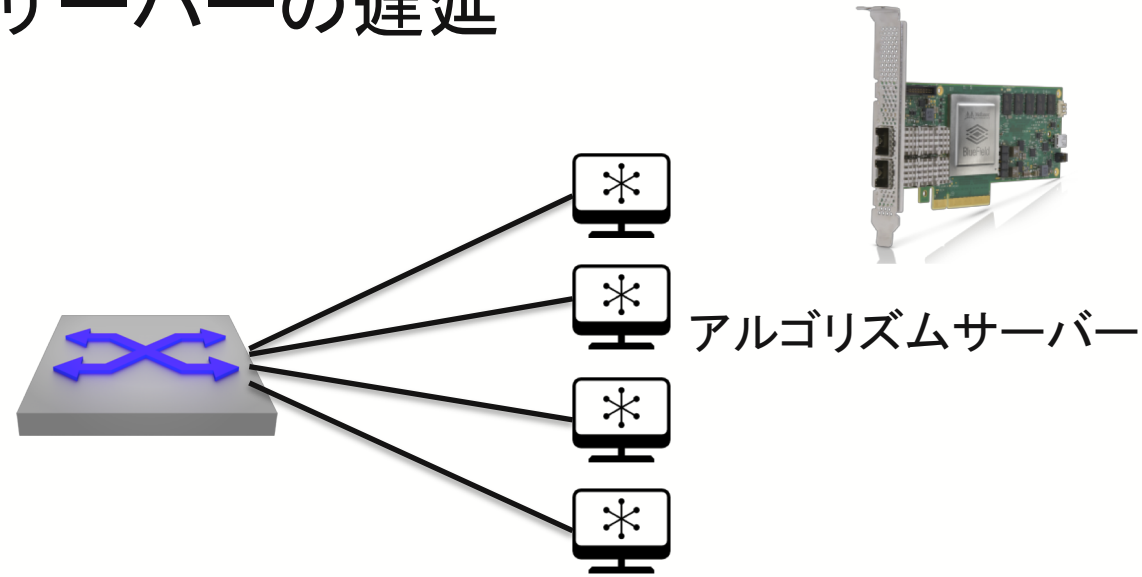


- シカゴとNASDAQのデータセンターNJ 1331kmをダークファイバーで接続
- ラウンドトリップタイムが約12.98msec

マイクロウェーブの使用

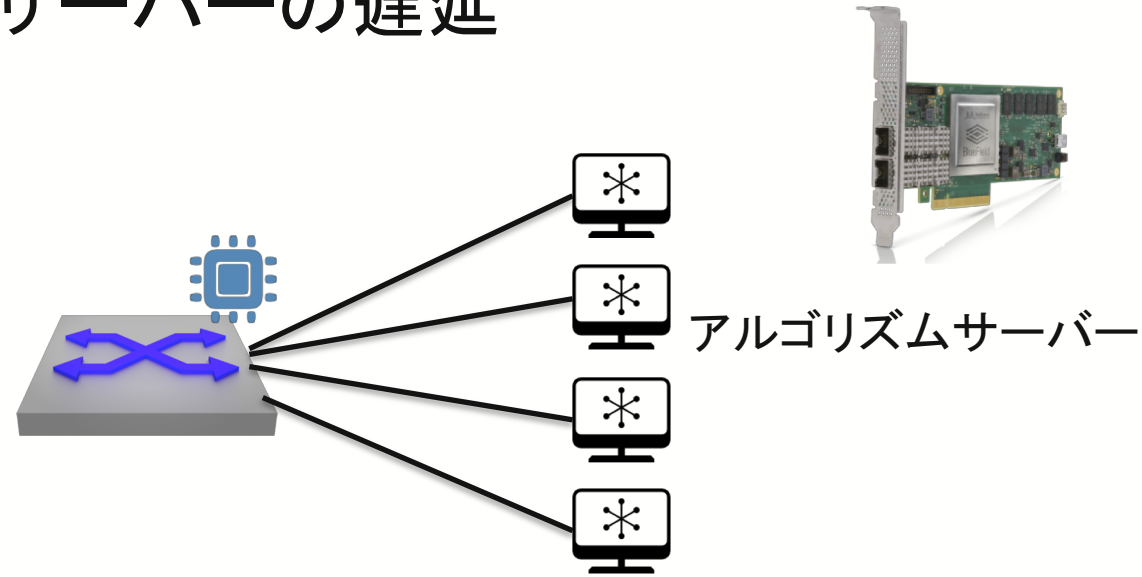
- マイクロウェーブ使用して空気中に伝送を行うとファイバーを通すよりも50%ほど遅延は良くなり、理論的には7.5-8msecにする事が出来る
- 1215095 - The Flash Boys Mystery Solved
  - http://blog.themistrading.com/2014/04/1215095-the-flash-boys-mystery-solved/
- High-frequency traders use 50-year-old wireless tech
  - https://www.itworld.com/article/2827482/mobile/high-frequency-traders-use-50-year-old-wireless-tech.html

アルゴリズムサーバーの遅延



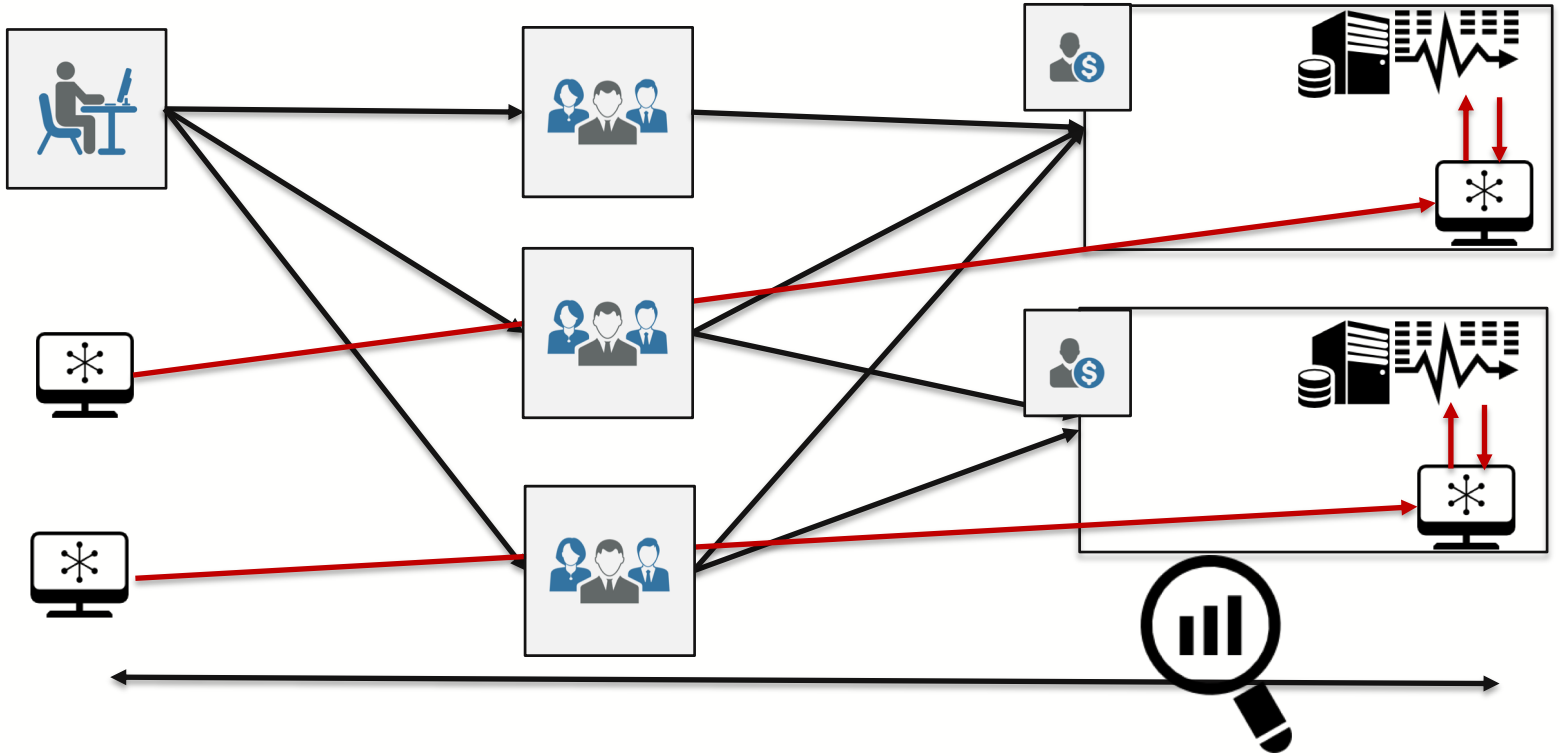
- ネットワークの遅延のみならず取引を成立させるサーバーでも遅延は重要となる
- NICに載せたFPGAでアプリケーションロジックを載せてパースし処理時間を短縮

アルゴリズムサーバーの遅延



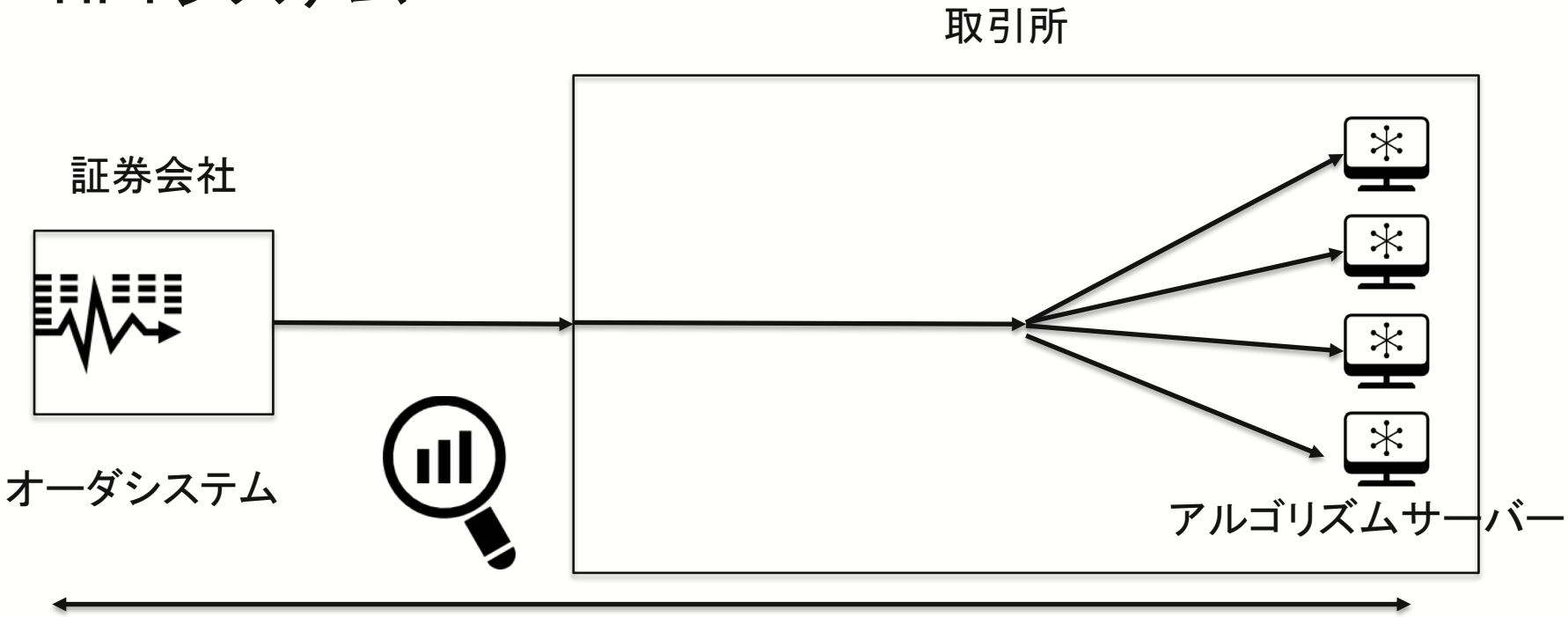
- アプリケーションロジックを搭載したFPGAをスイッチへ搭載させる、これによりさらなる高速処理が可能に

超高速取引(HFT)システムの遅延要求



- フラッシュボーイズにとってネットワークの遅延情報は超重要情報

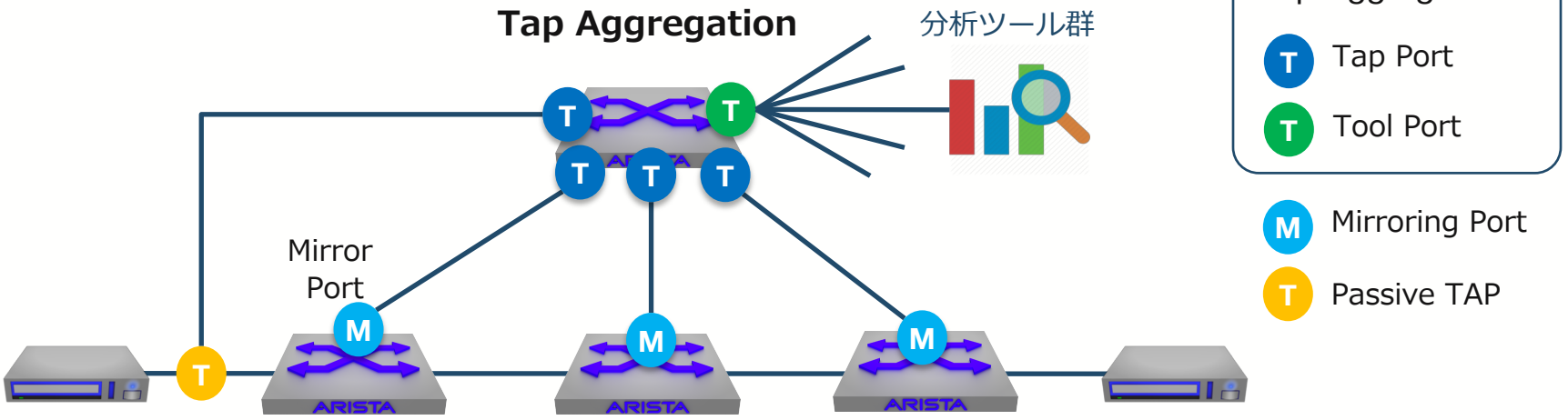
HFTシステム



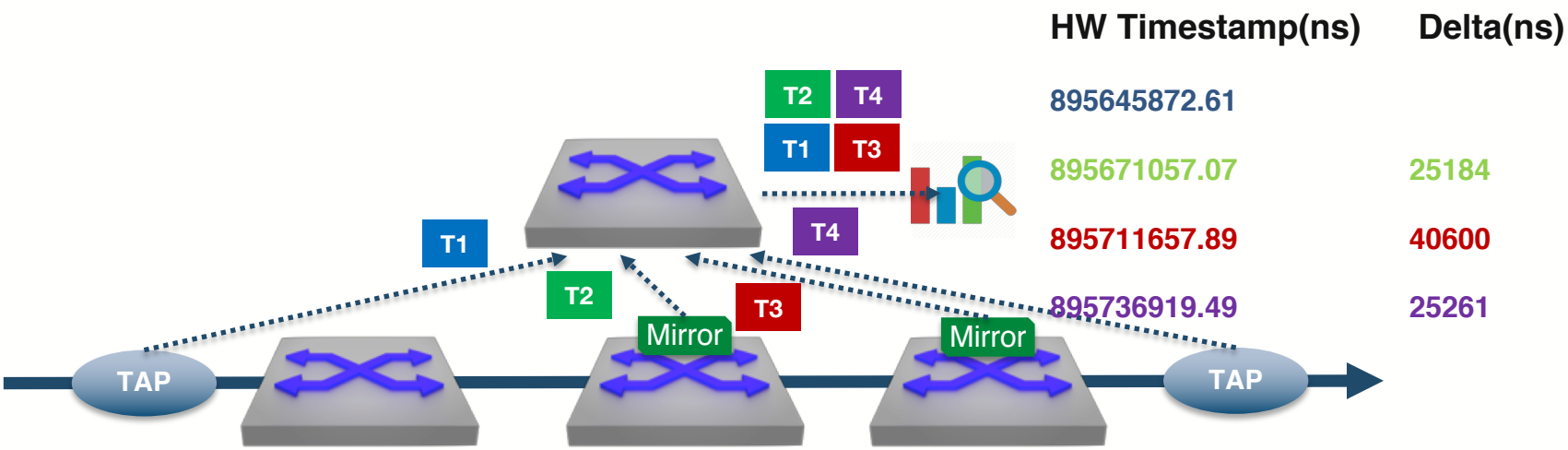
- フラッシュボーイズにとってネットワークの遅延情報は超重要情報

Tap Aggregation

- タップまたはミラーリングされたデータの集約と分析ツールへのデータ転送
- 分析ツールへの投資を最小限に



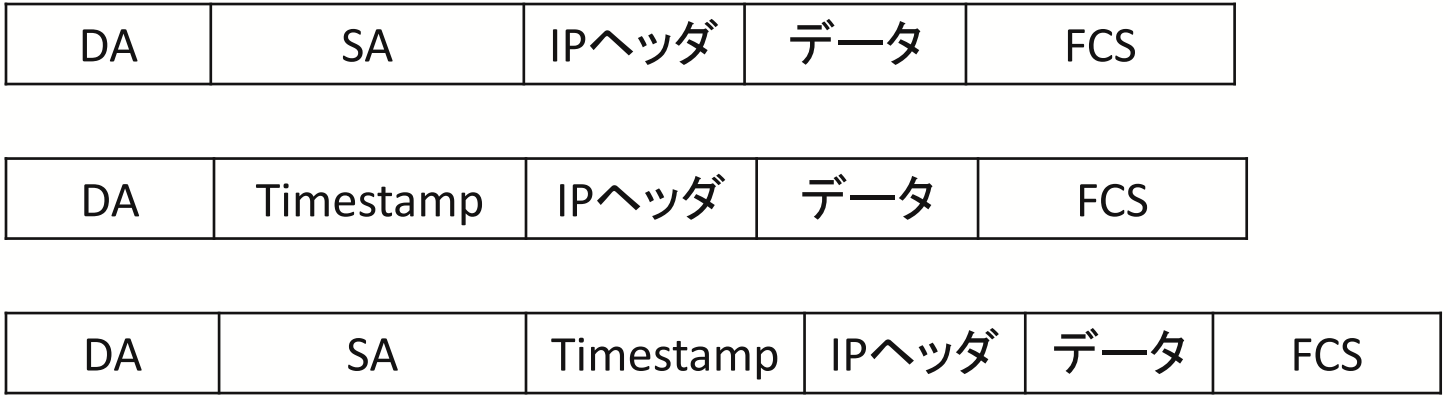
ネットワークの遅延情報のモニタ



- 各ポートでタイムスタンプ情報を付加し、TAPアグリゲーションでモニタリングする

ARISTA

フレームタイムスタンプ



- イーサネットフレームにtimestampを付加する様々な方法
- 送信元MACを置き換える(48ビット)
- タイムスタンプヘッダーを付与する(48ビット/64ビット)

アリスタフレームタイムスタンプ

```
▶ Frame 41: 91 bytes on wire (728 bits), 91 bytes captured (728 bits) on interface 0
▼ Ethernet II, Src: AristaNe_a5:1d:0b (44:4c:a8:a5:1d:0b), Dst: AristaNe_2f:60:b8 (44:4c:a8:2f:60:b8)
    ▶ Destination: AristaNe_2f:60:b8 (44:4c:a8:2f:60:b8)
    ▶ Source: AristaNe_a5:1d:0b (44:4c:a8:a5:1d:0b)
      Type: 802.1Q Virtual LAN (0x8100)
▶ 802.1Q Virtual LAN, PRI: 0, DEI: 0, ID: 47
▶ Internet Protocol Version 4, Src: 10.255.255.1, Dst: 10.10.1.2
▶ User Datagram Protocol, Src Port: 51097, Dst Port: 9711
▶ Data (41 bytes)
```

```
▶ Frame 50: 105 bytes on wire (840 bits), 105 bytes captured (840 bits) on interface 0
▼ Ethernet II, Src: AristaNe_a5:1d:0b (44:4c:a8:a5:1d:0b), Dst: AristaNe_2f:60:b8 (44:4c:a8:2f:60:b8)
    ▶ Destination: AristaNe_2f:60:b8 (44:4c:a8:2f:60:b8)
    ▶ Source: AristaNe_a5:1d:0b (44:4c:a8:a5:1d:0b)
      Type: Arista Timestamp (0xd28b)
▼ Arista Networks
    ▼ TapAgg Header Timestamp
        Version: 0x0010
        Timestamp: Dec 18, 2018 14:00:12.921219118 JST
▶ 802.1Q Virtual LAN, PRI: 0, DEI: 0, ID: 48
▶ Internet Protocol Version 4, Src: 10.255.255.1, Dst: 10.10.1.2
▶ User Datagram Protocol, Src Port: 51008, Dst Port: 9325
▶ Data (41 bytes)
```

まとめ

- HFTで超高速化するネットワーク要件と進化をまとめみた
- 常識と考えてきたものが通用しない事もある
- 0秒に近づくネットワーク:その要素技術はどこかで使えるかもしれません