Network based DAQ の最近と今後

五十嵐 洋一

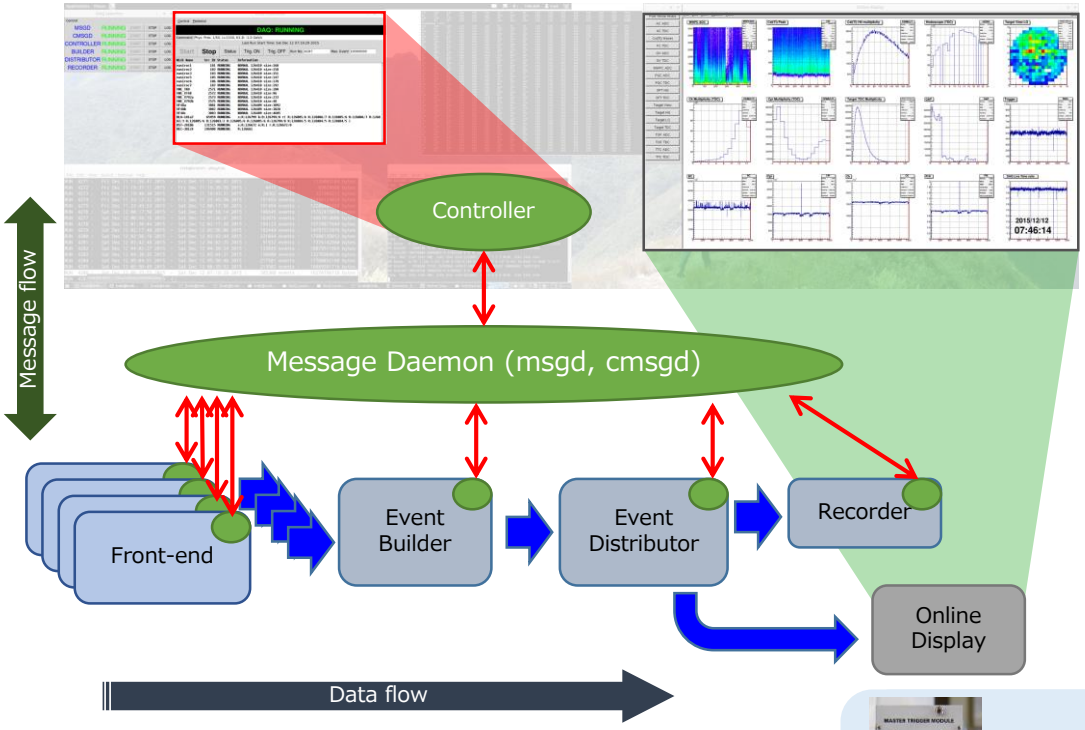
KEK

計測システム研究会 2020.11.26

Content

- Network based DAQ
  - その構成と弱点
  - 近年の機器を使った構成のパフォーマンス、COMET DAQ からの知見
    - 安価なネットワークスイッチ
    - スイッチレスネットワーク
    - RAID DISK
- より高いデータレートへ向けて
  - マルチエンドポイントの Network
  - ZeroMQ/FairMQ
  - 現状

Network based DAQ system

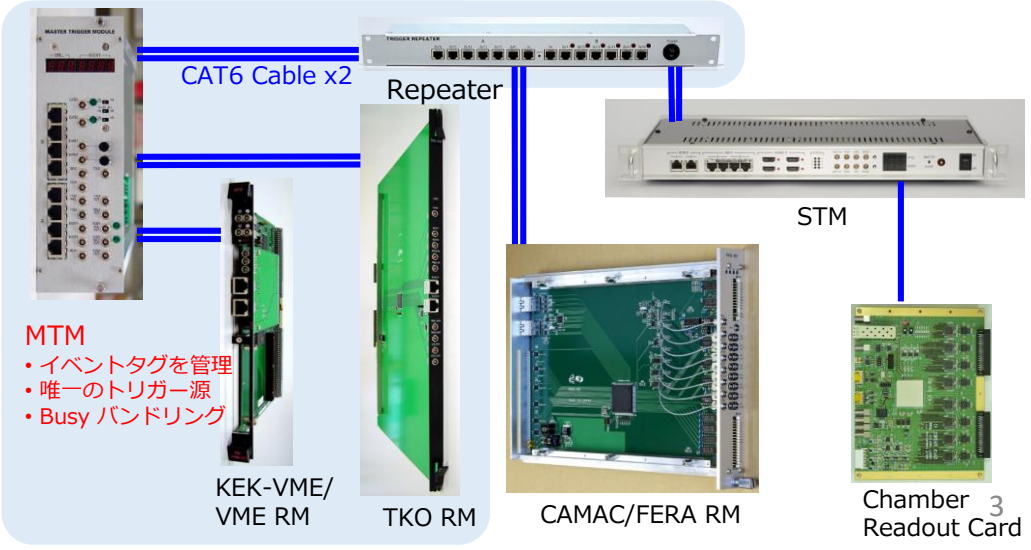


DAQ software (HDDAQ)

- 単機能の多数のプロセスによる協調動作
- Message Path(制御通信), Data Path(データ通信) の2種類の通信
- すべての通信は TCP/IP

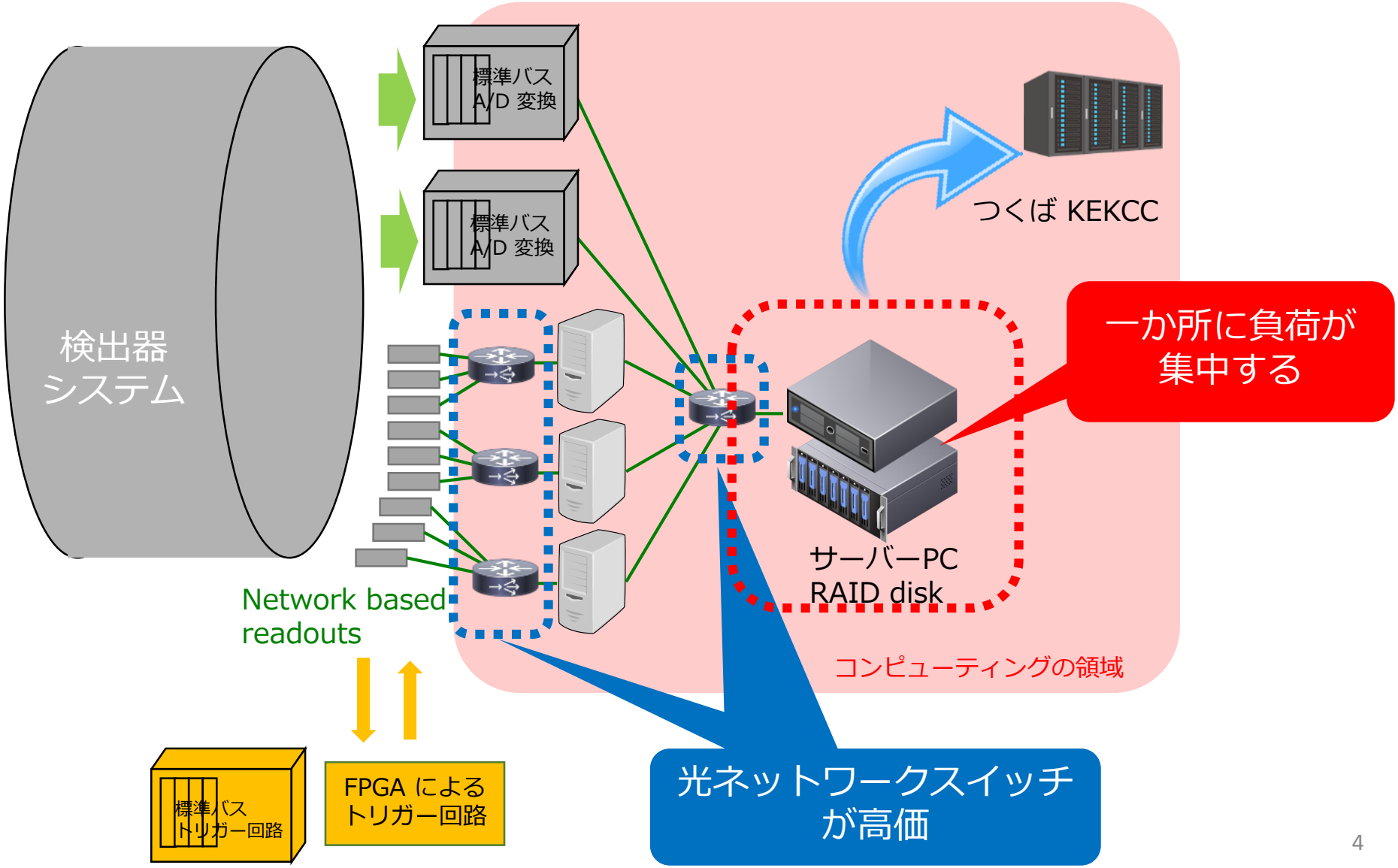
Event synchronization

- MTM/RM トリガー配布システム
  - マルチイベントバッファを持ったサブシステム間でイベントを同定する。
  - Trigger/Busy ハンドシェイク
  - Trigger 毎に唯一のイベントタグを各サブシステムに送る。



Network event building

- 現状の Network based DAQ



安価な Network switch で Event build は可能か？

- On-detector readout → 読み出し機器が磁場中 → メタル Ethernet のパルストランスが機能しない。
- 最近安い光ネットワークスイッチが登場してきた。
  - FS.COM, QNAP, Mikrotik, …
- Requirements
  - SiTCP 1G を受けるために たくさんの 1G SFP 光ポート
  - 後段にデータを送るための一つ以上の 10 Gb ポート

FS.COM (Fiberstore) S3900 24F4S

- 20× 1 GbE SFP
- 4× 10 GbE SFP+
- 4× 1 GbE Metal
- Layer2+

30倍の価格差?

Cisco MS410-32

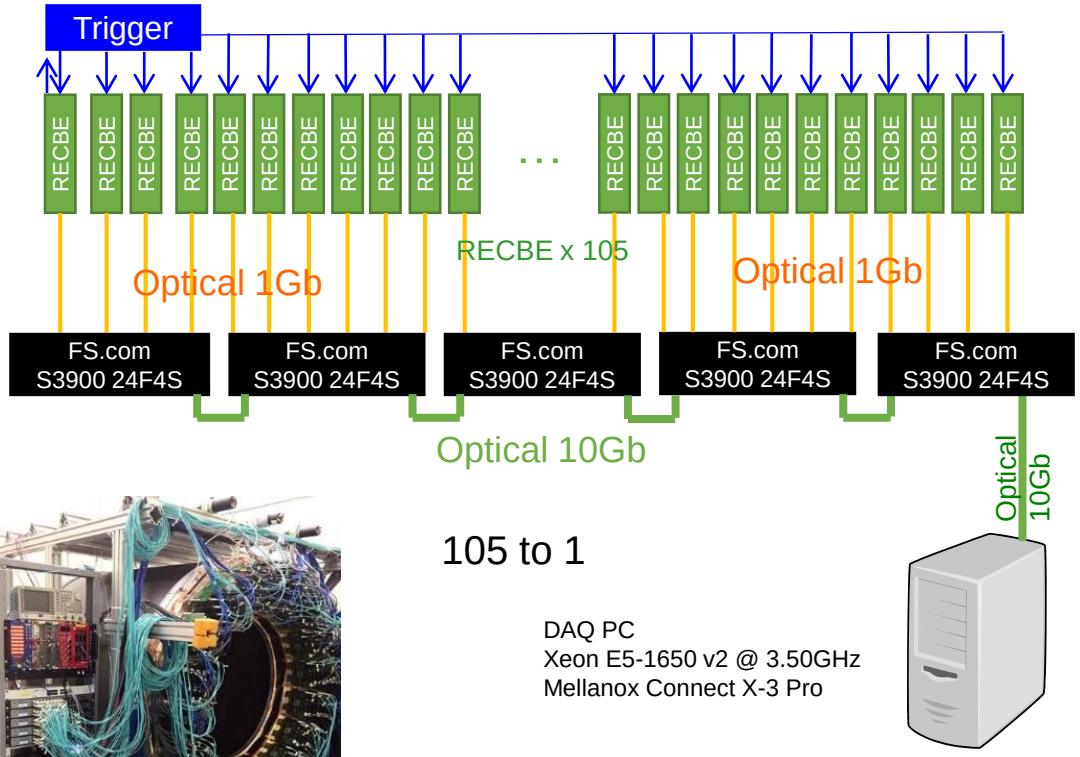
- 32 × 1 GbE SFP
- 4 × 10 GbE SFP+
- Layer3 capable

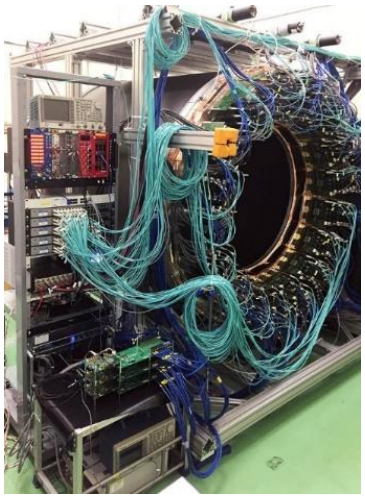
機器	価格
S3900 24F4S	￥43,284
SFP (1000BASE-SR)	￥660
SFP+ (1GBASE-SX)	￥1,800

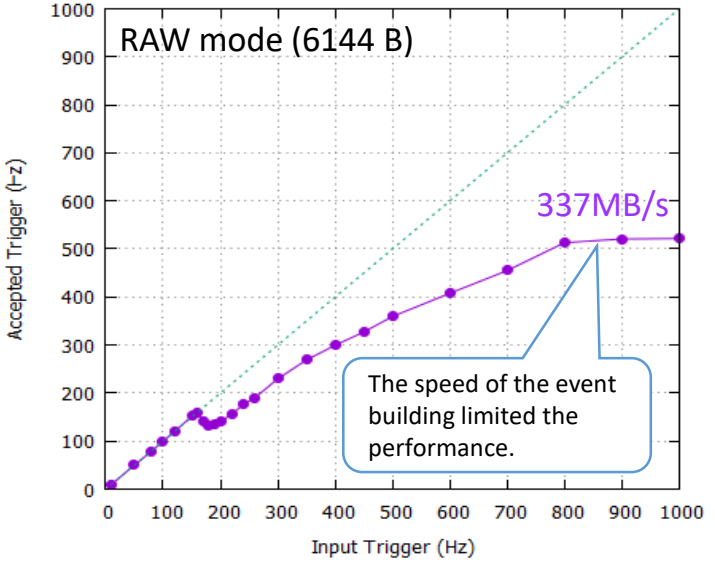
現実的な評価をしてみた。

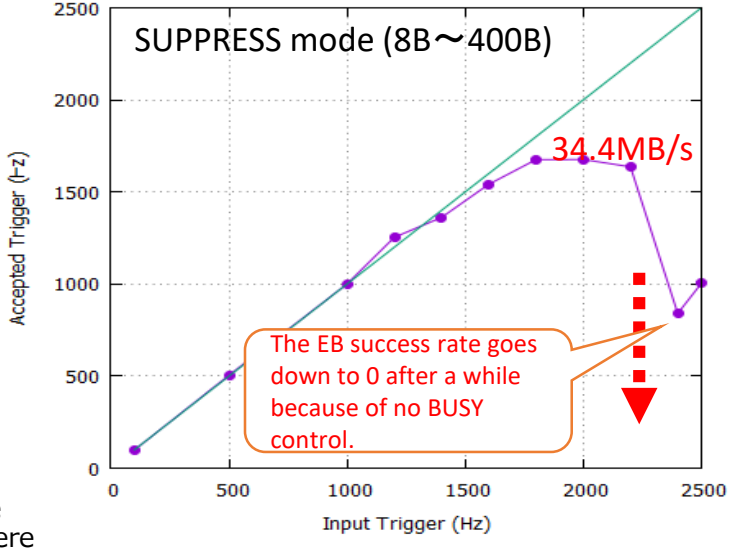
COMET CDC を使って試験

- CDC の宇宙線による試験が進行中
  - このセットアップで安価な network switch S3900 を使用してみた。









CDC cosmic ray test setup

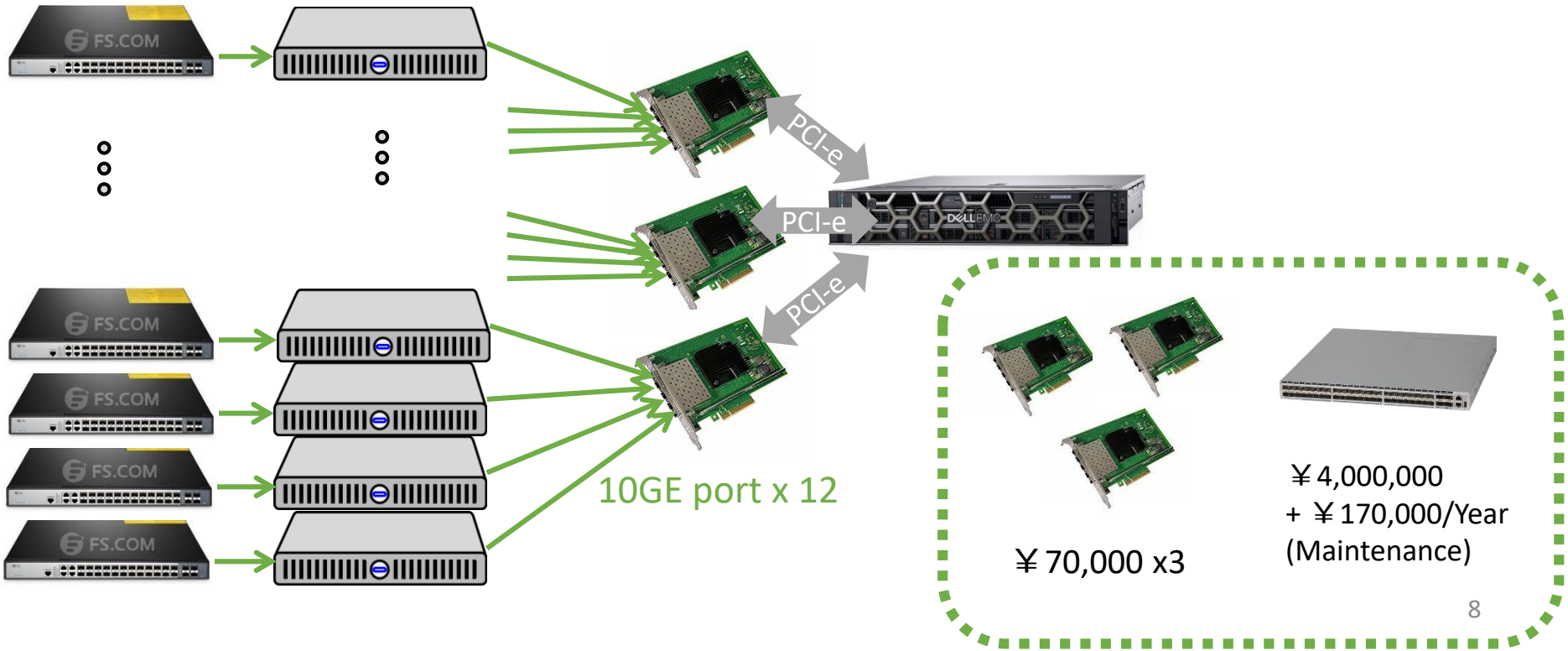
One hundred five readout cards (RECBE) and five low price optical network switches worked on the setup. The data were taken with the full event building and without the BUSY control.

Back-end は?

- データは一つの計算機に集中する。
  - PC はどの程度のデータをさばけるか?
- ちゃんとした 10G 光スイッチはやっぱり高価
  - 安い 10G 光スイッチも出てきている。
  - 10G NIC はほどほどの価格、たくさん PC に挿せる。
- COMET DAQ の構築で割と高速な Server PC や RAID HDD が使えたので、評価したみた。

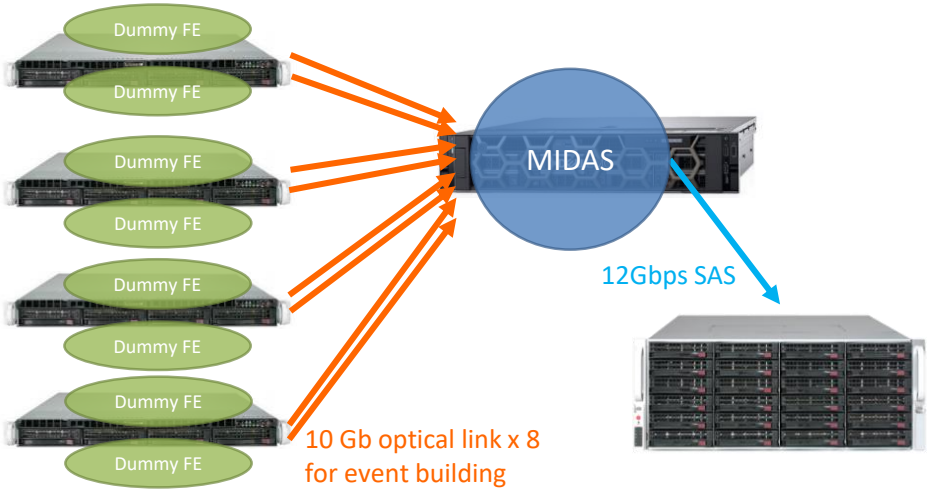
Switch-less network event building

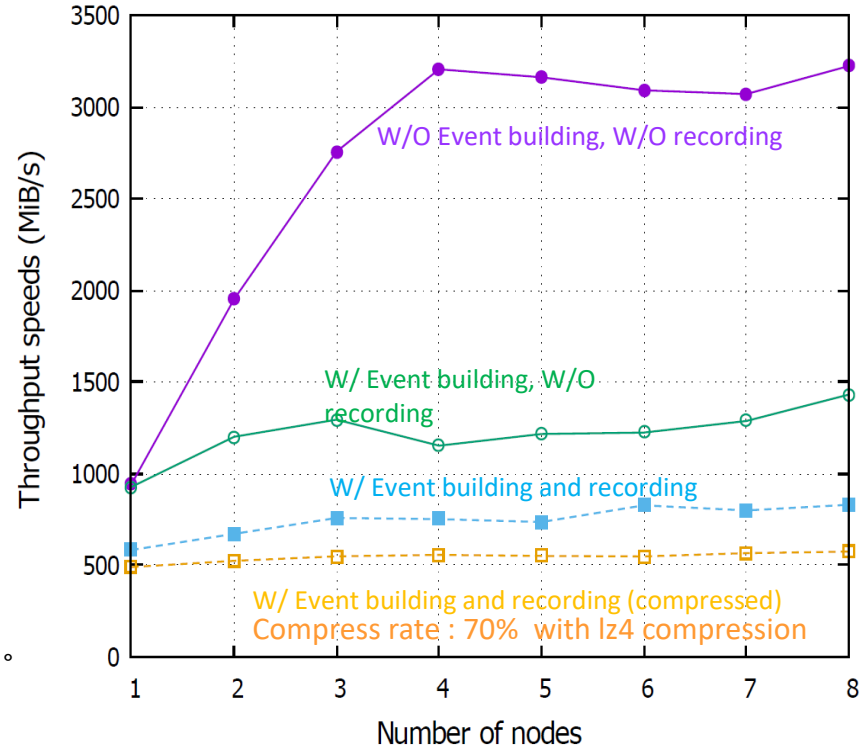
- PCIe は共有バスではなく、それぞれのスロットが直接プロセッサ内部のブリッジの繋がっている。
- 2U Sever PC は 4つ以上 8 lane 以上の PCIe スロットを持っている。
- PCIe Gen3 8 lane の速度は 7.877 GB/s
- たくさん 4 port NIC を挿せばたくさんの FE につながる。



Switch-less event building の評価試験

- 2 port の 10GBASE-SX を持った Front-end PC 4 台を Event-building PC に接続
- 8 本のコネクションまで測定
- DAQ software としては MIDAS を用いた。





Logger process limited by the recording speeds.
Total throughput : 830 MiB/s

それぞれで動いている process

- FE PC：Dummy FE
  - 16 kiB のランダム値を詰めたイベントフラグメントを EB に送る。
- EB PC：いくつかの MIDAS プロセス
  - FE：ネットワークからデータを読む。 8 FE プロセスが動作
  - EB：MIDAS FE が読んだイベントフラグメントビルドする
  - Logger：ファイルにビルドされたデータを書き込む

	Processor	Memory	10 Gbps NIC	SAS/RAID system	HDD
Event building PC	Xeon Gold 6126 @ 2.6GHz	64 GB	Intel Network Adapter X710-DA4	Broadcom / LSI MegaRAID SAS-3 3108	SEAGATE 10 TB 7200RPM
Front-end PC	Xeon E-2134 @ 3.5GHz	32 GB	Mellanox Connect-X 3 Pro		

RAID disk performance

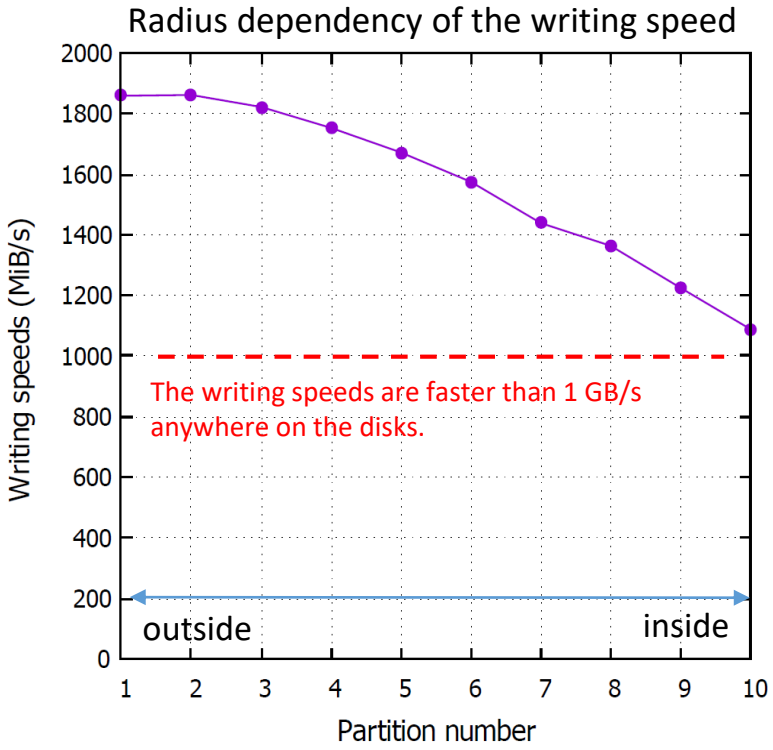
- RAID disk system
  - 12 Gbps SAS (Serial Attached SCSI) で EB PC に接続
  - 現在は 10 台の HDD がインストールされている。フォーマット後データ容量 74 TB
  - 最大 44 HDD がインストール可能
  - RAID5+0 で構成。
  - RAID5＋0：Striping RAID5, 冗長 bit と Dual データアクセス

ファイルシステムによる速度比較
"dd" コマンドに "oflag=direct" を付けて測定

File system	Speed	
Xfs	1.84 GiB/s	adopted
Ext4	1.66 GiB/s	

DAQ PCs installed on the 3rd floor of COMET experimental building

- 2 EB PC and 4 FE PC (for the main detector readout).
- RAID disk system

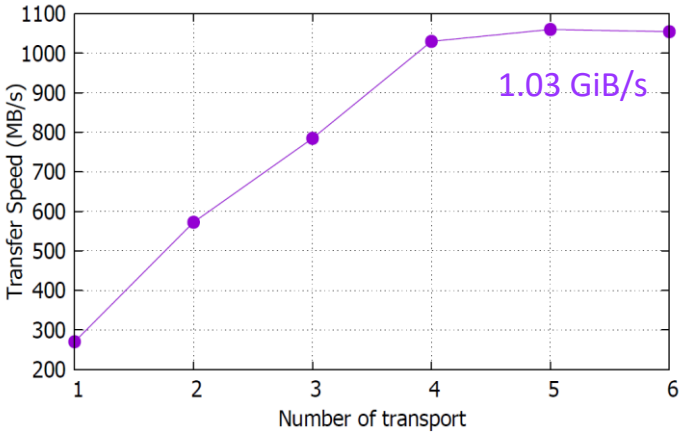


東海つくば間データ転送

実験データは最終的にはつくばの KEKCC のデータストレージに保存したい。

“scp”によるデータ転送試験

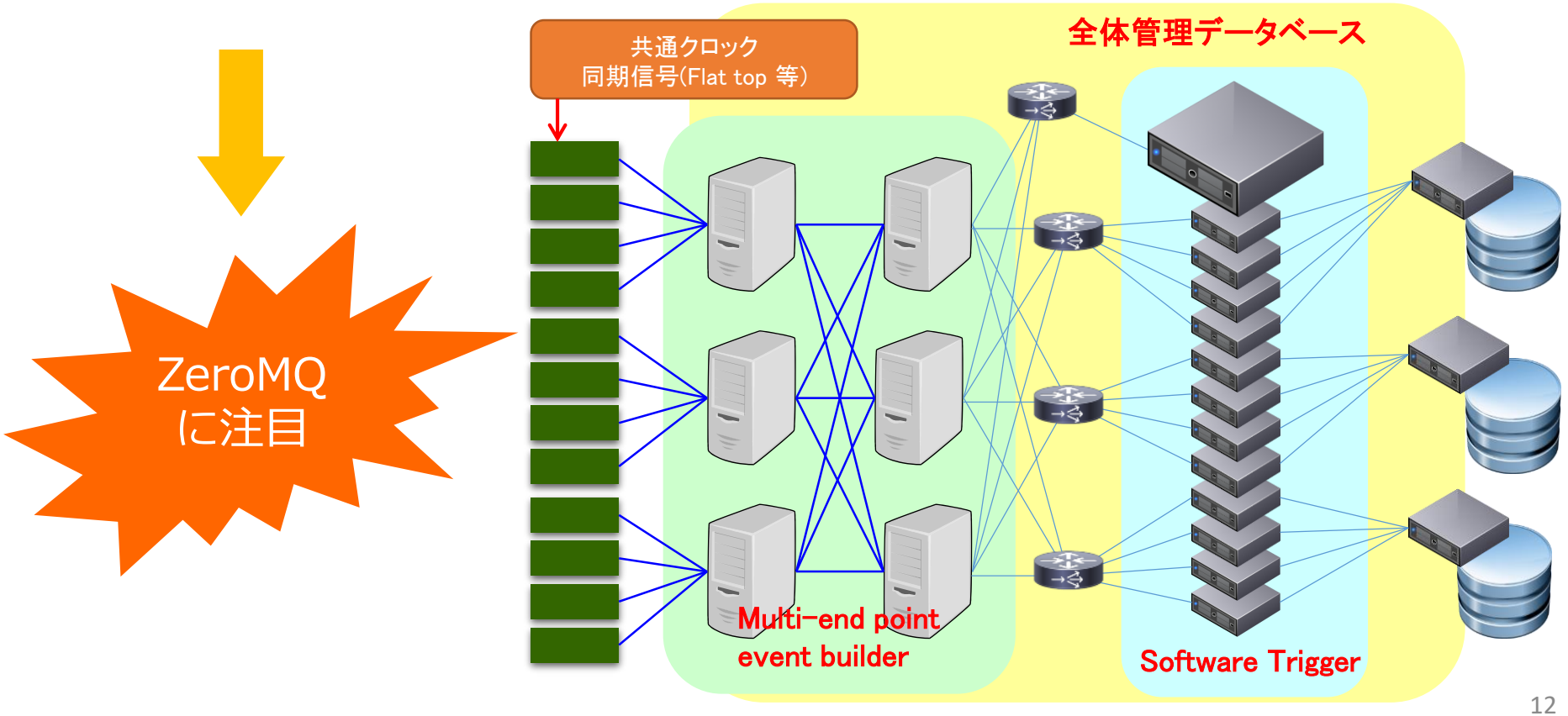
- つくば－東海間は 10Gbps の転送幅を SINET から借りている。
- 一つの “scp” あたり 280 MB/s 程度の転送速度



これらを超えるデータレートに向けて

- 1GB/s までは樹状ネットワークで行ける。それ以上のデータレートの時は?

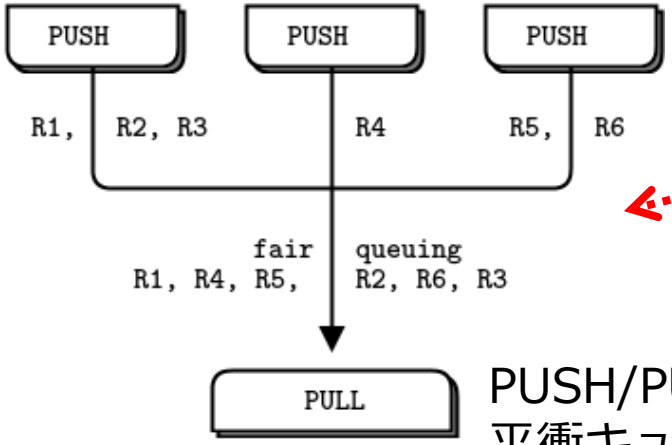
- データを一か所に集中しないように
  - マルチエンドのネットワークデータ収集
- トリガーレス/ストリーム型
  - オンラインイベントフィルタリングは必要?



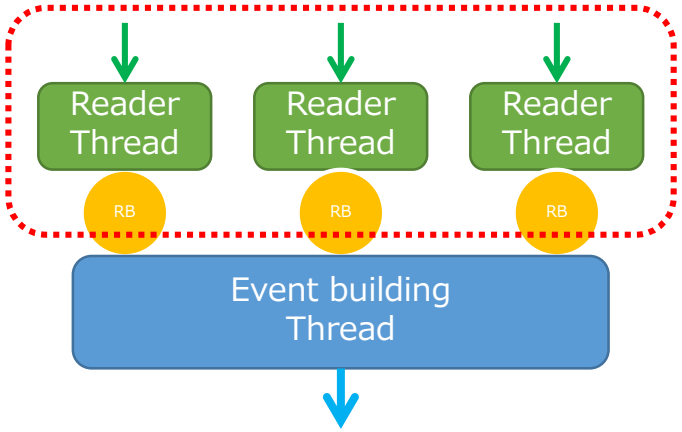
ZeroMQ

- なぜ ZeroMQ か?
  - 非同期メッセージレイヤーの一つ
    - 通信を TCP/IP, UDS の上に実現している。
    - コネクションの接続・切断を気にしなくてよい。
    - 長さとコンテンツだけの単純なフレーム
    - バッファリングを考えなくて良い。(といいなあ)
      - Memory allocation / free をよく管理する必要がある。
      - キューがいっぱいになるとブロックする。
        - Non-blocking だと思っていると止まることがある。
  - 多くの Linux distribution に含まれ、Windows を含めいろいろな OS で動いている。
  - 有用な通信モデルが構築されている。
  - スケーラビリティが高い。
    - たくさん接続しても破綻しない。(はず?)

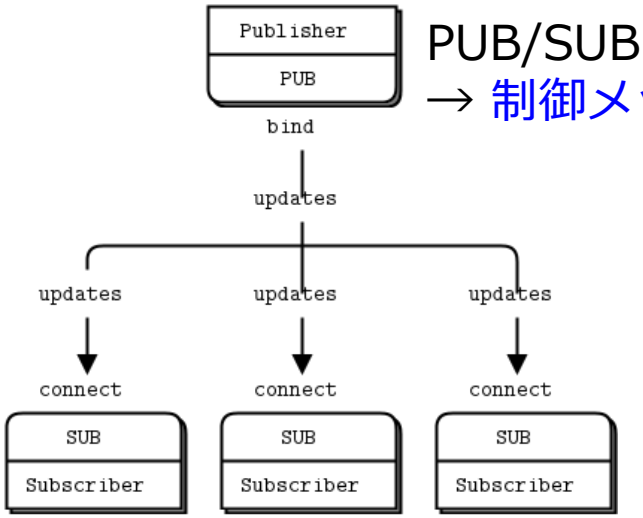
ZeroMQ 通信モデルの一部



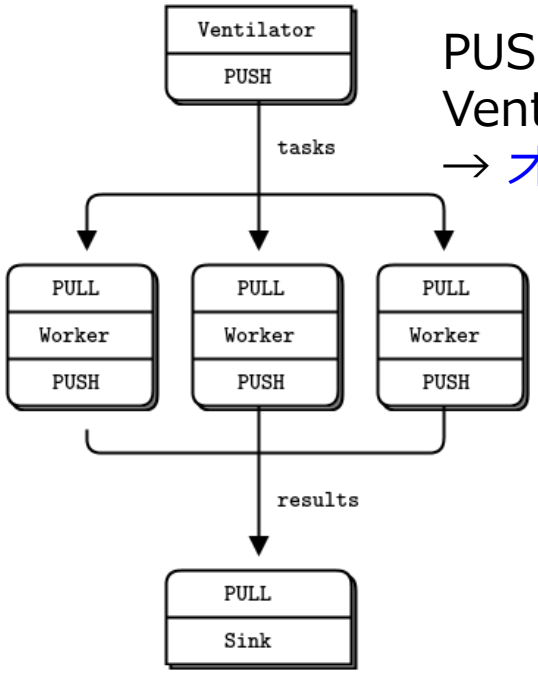
PUSH/PULL による
平衡キューイング



HDDAQ の Eventbuilder



PUB/SUB
→ 制御メッセージ



PUSH/PULL による
Ventilator/Worker/Sink
→ オンラインフィルタ

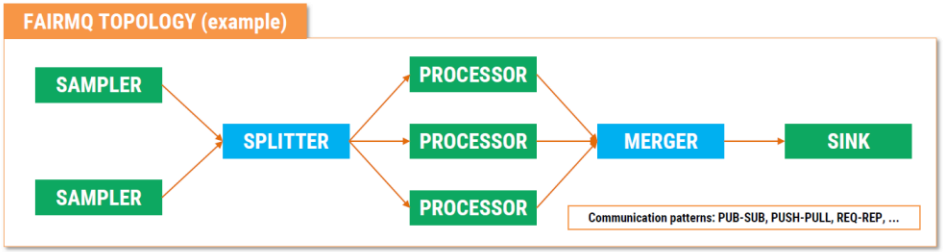
FairMQ

- GSI/FAIR のために開発されている DAQ フレームワーク
- ZeroMQ + State machine + 制御 plug-in + たくさんの周辺の統合

What is FairMQ?

Organize processing tasks in topologies, consisting of independent processes (Devices), that communicate via *asynchronous message queues* over network or inter-process.

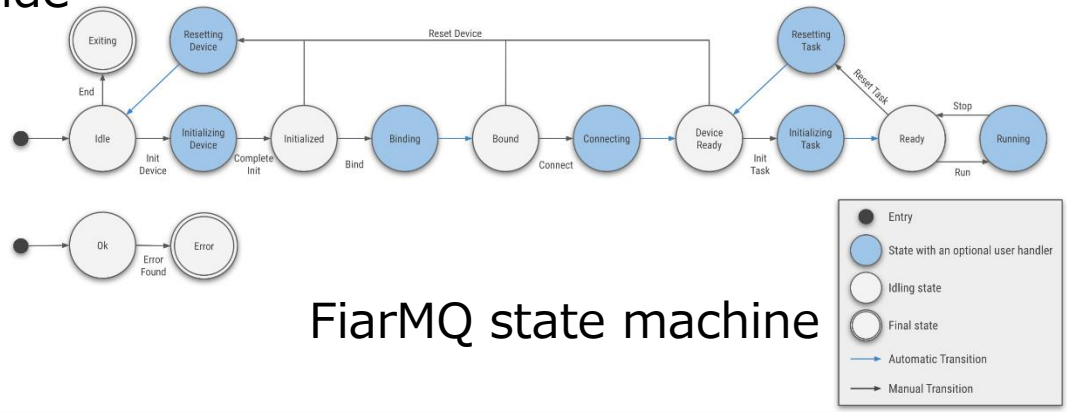
Ethernet, InfiniBand (IP-over-IB)



Ready to use devices are provided for typical scenarios.
User-defined devices can be implemented by inheriting from `FairMQDevice`.

Alexey Rybalchenko | ALICE Offline Week | 01.07.2015 3/21

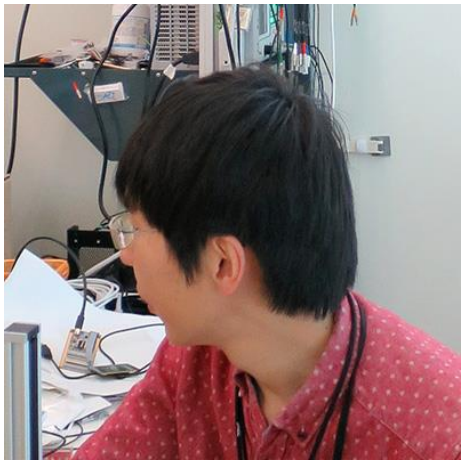
from Alexey Rybalchenko(GSI)'s slide



FiarMQ state machine

J-PARC HD での高速DAQ

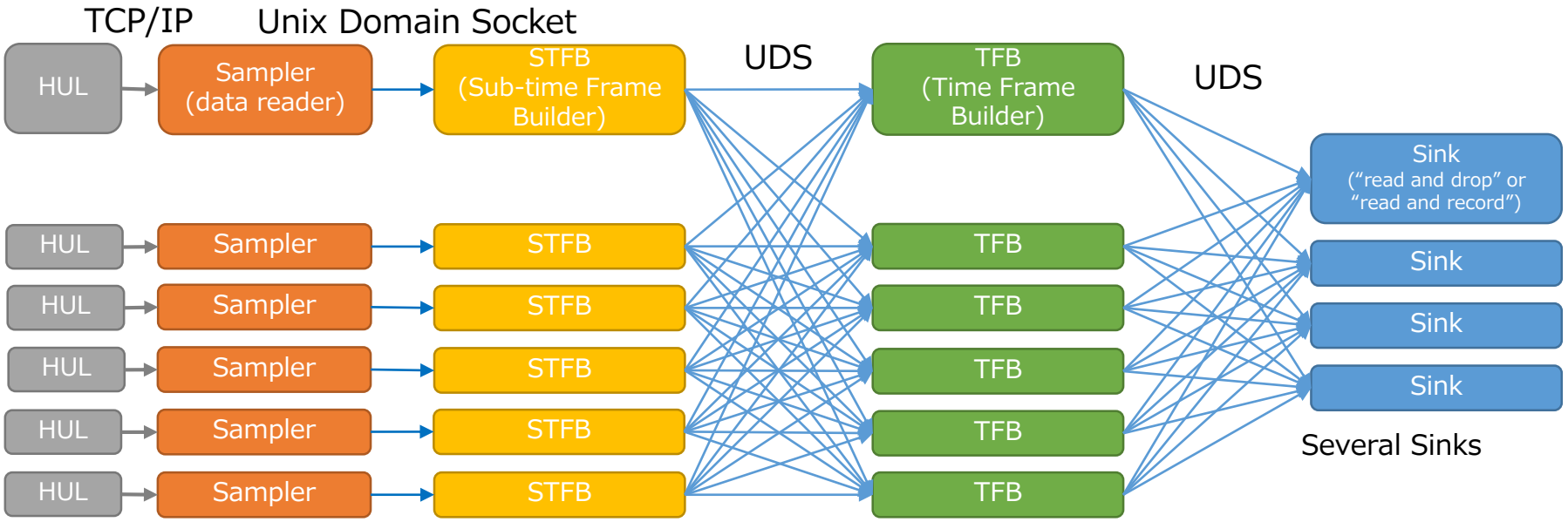
- E16/E50 をモデルケースにして開発が進んでいる。
  - Pipeline front-end readout (HUL/AMANEQ + …)
  - FairMQ のコア部分を利用
  - 全体管理、制御には Redis を検討
    - NoSQL データベース/キーバリュー型
    - メモリ指向/高速
    - キー・スペース通知 → 制御に使える
    - Redis API を FairMQ 制御 Pulg-in に組み込んで Redis からデータを読んだり、制御されたり。



高橋智則氏（理研）が強力に推進中

検出器 Beam 試験における試行

- E50 Fiber tracker と Cherenckov counter の試験を ELPH で 2018 年に行った。
- FairMQ のコア部分を使用
  - HUL TDC の連続読み出し
  - 時間分割による Time frame building
  - ソフトウェアによるコインシデンストリガー
  - 独自制御 plug-in
- “PUSH/PULL” 通信を使用
  - 一応動作
  - うまく動いているときはメッセージがなくなることはなかった。



Full mesh connection and round robin network

まとめ

- 近年の高速機器で Network based DAQ を行うとそれぞれのコンポーネントのパフォーマンスがだいたい 1GB/s でバランスする。
  - Eventbuild：～1 GB/s
  - Data recording：～1 GB/s
  - Data transport：～1 GB/s
- 安価なネットワークスイッチでも DAQ は出来た。
- ネットワークスイッチなしでもたくさんNICを入れればかなり読み出せる。

- トリガーレス DAQ を指向した FairMQ ベースの DAQ の開発が進行中
  - 次の機会の研究会での高橋智則氏の報告に期待しよう。