Upgrade of the CMS Event Builder

Andrea Petrucci *- CERN (PH/CMD) on behalf of the* ***CMS DAQ group***

19th International Conference on Computing in High Energy and Nuclear Physics
CHEP 2012, 21-25 May 2012, New York, United States

Outline

- Introduction to CMS upgrade
- Current DAQ system
- Architecture of the upgrade DAQ system
- Feasibility studies on advanced networking technologies for event builder
  - Network technologies
  - CMS online software
  - Test setups
  - Preliminary measurements
- Summary

CMS Upgrade

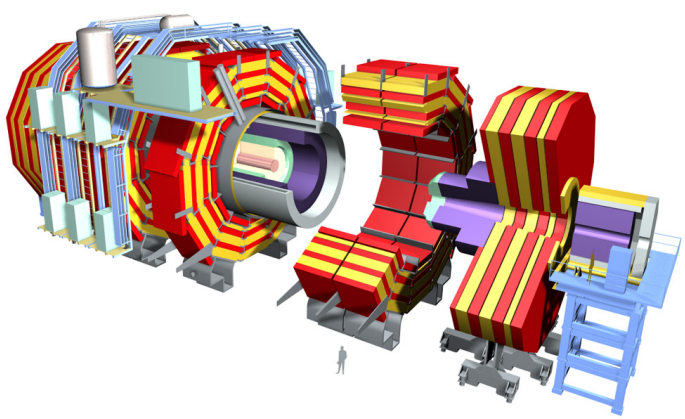
CMS Upgrade timeline

- Magnet Test at surface
- First beam in CMS September 10th
- First collisions in CMS November 23st
- First 7 TeV collisions in CMS March 30th
- 5fb $^{-1}$ collected
- **Higher pileup**
  (depending on LHC operation 2x 10^{34} at 25 ns or 50 ns bunch spacing)
  Pileup could reach 100 underlying interactions
  More data will be generated due to higher occupancy

2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021

7Tev 8Tev LS1 14Tev LS1.5 LS2

- Install Pixel 4th barrel layer and 3rd endcap layer



- Upgrade of detector:
  - HCAL introduces uTCA based readout system
  - Trigger system will be upgraded: the new system will be first operated in parallel to the existing system. Readout will be based in uTCA.

DAQ: new optical link to readout uTCA based systems

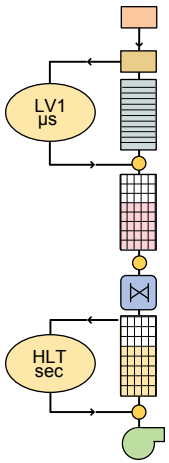
Motivations for upgrade of CMS DAQ

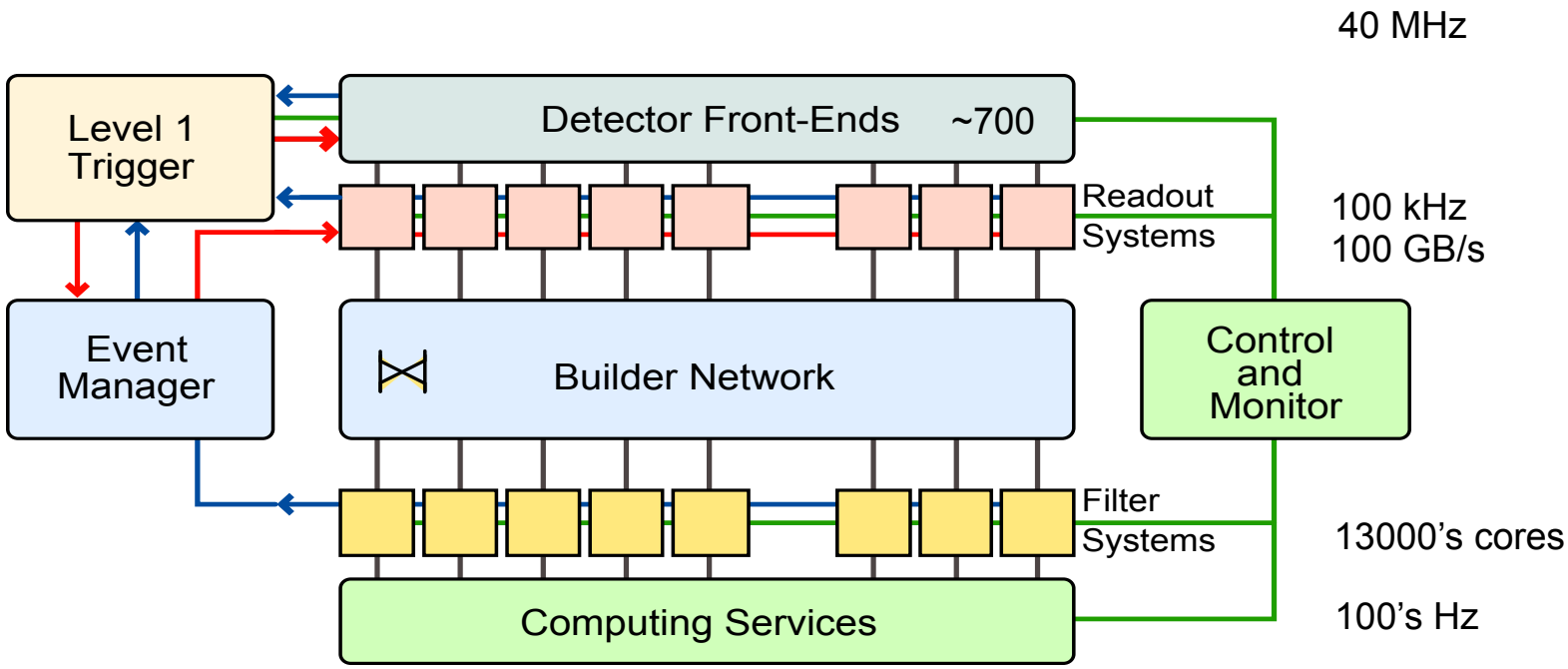
- Aging of existing hardware (PCs and Networks at least 5 years old)

- Accommodate sub-detectors with upgraded off-detector electronics
  - HCAL upgrade front-end electronics and readout system
  - New Pixel detector with 4 barrel layers and 3 endcap layers (LS1.5), estimated requirements: ~80 readout channels with 640MB/s
  - Trigger system will be upgraded: the new system will be first operated in parallel to the existing system.

Current CMS Data Acquisition System

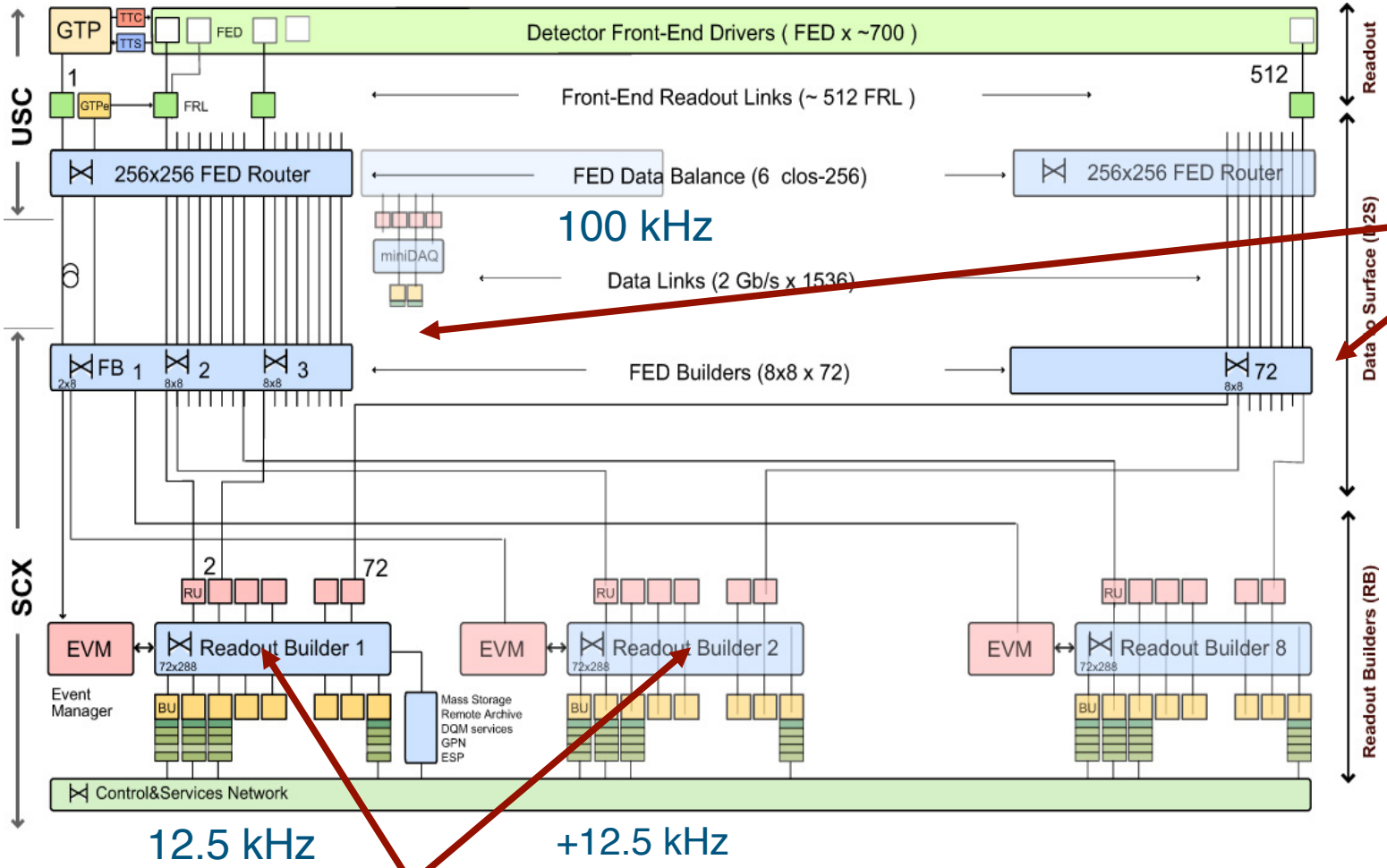
Current DAQ system





- Read out 700 detector front-ends (fragment size ~2 kB)
- Build complete events at 100 kHz (L1 trigger rate)
- Make them available to a filter farm of O(13000) cores
- Store 100's of Hz to disk (10's of TB/day)
- Scalable system employing commercial components wherever possible
  - Proprietary / Commercial: Front-Ends, VME, PCI, PC servers, networks, Protocols, OS

Current DAQ system (II)



72 FED builders: assemble FED (~2 kB) into super-fragment (~16 kB)

Eight RU builders (EVB DAQ slices):
Assemble super-fragments (~16 kB) into full events (~1 MB) at 12.5 kHz each

Current DAQ system (III)

- Full readout of all sub-detectors at L1 rate
- SLINK64 operates on 400 MB/s
- Fed builder operates on ~200-350 MB/s, depending on event size distribution
- 8 DAQ Slices, 100 GB/s event builder
- Event Filter
  - ~1300 dual-CPU (multi-core) PC nodes, 13 k-cores, 220 kHEPSpec06
  - running about ~20 k HLT processes, CPU budget is ~175 ms/evt @100 kHz.
- Storage Manager
  - Event data to disk
    - pp: ~200 MB/s, design for 600 MB/s
    - Heavy Ions: ~1,4 GB/s (up to 2.8 GB/s w/o transfer)

Upgrade of CMS Data Acquisition System for long shutdown 1

Upgrade of DAQ system

Requirements

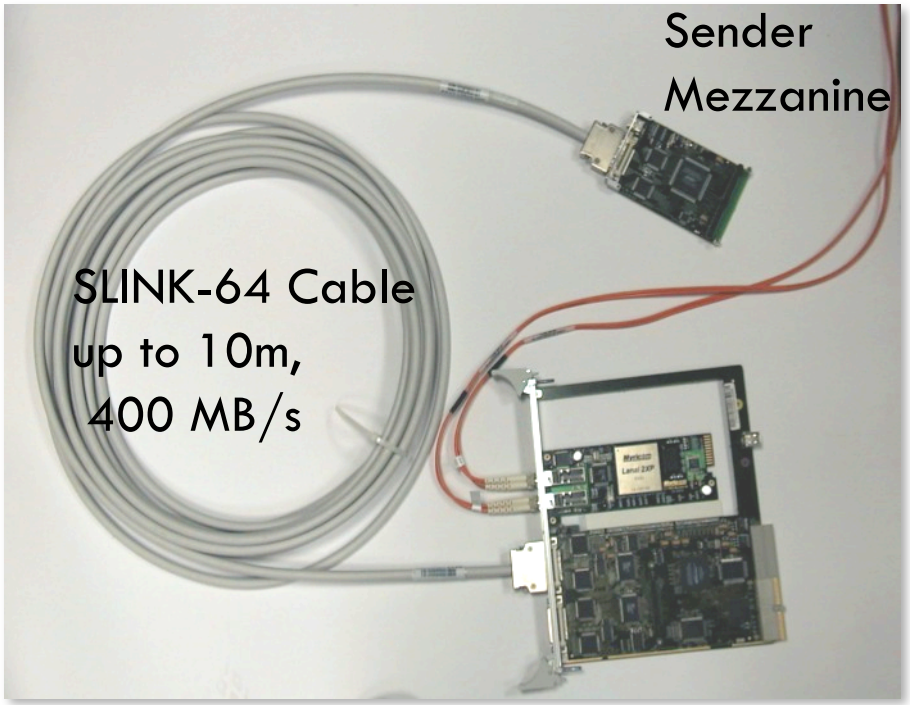
- L1 rate of 100 KHz
- Detector Front End Drivers (FED)
  - ~552 Legacy FEDs (fragment size increases from 2 kB to 4 kB due to pile-up)
  - 37 (TRG, HCAL, HF) + 40 (Pixel - 2 x 10 GbE links) new readout links from new FEDs (expected maximal fragment size 8kB)
- Frontend Readout Links (FRLs)
  - ~360 FRLs (Legacy FEDs, ~400 MB/s)
  - ~120 FRLs (new FEDs, ~640 MB/s)
- High availability (redundancy, load balancing, failover, etc.)

DAQ plans for upgrade

- Replace myrinet-based fedbuilder
- Replace event builder network
  - two options: conservative (10 GbE) and aggressive (40 GbE or Infiniband)
- New architecture between Event Builder and Filter Farm
- Replace of the Storage manager hardware

Read out link

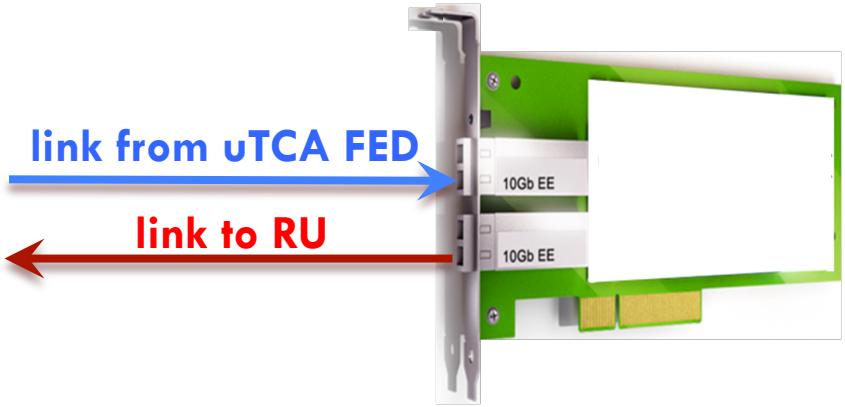
- New readout system must be compatible with old data sources (FEDs – SLINK) and with the newly designed FEDs.



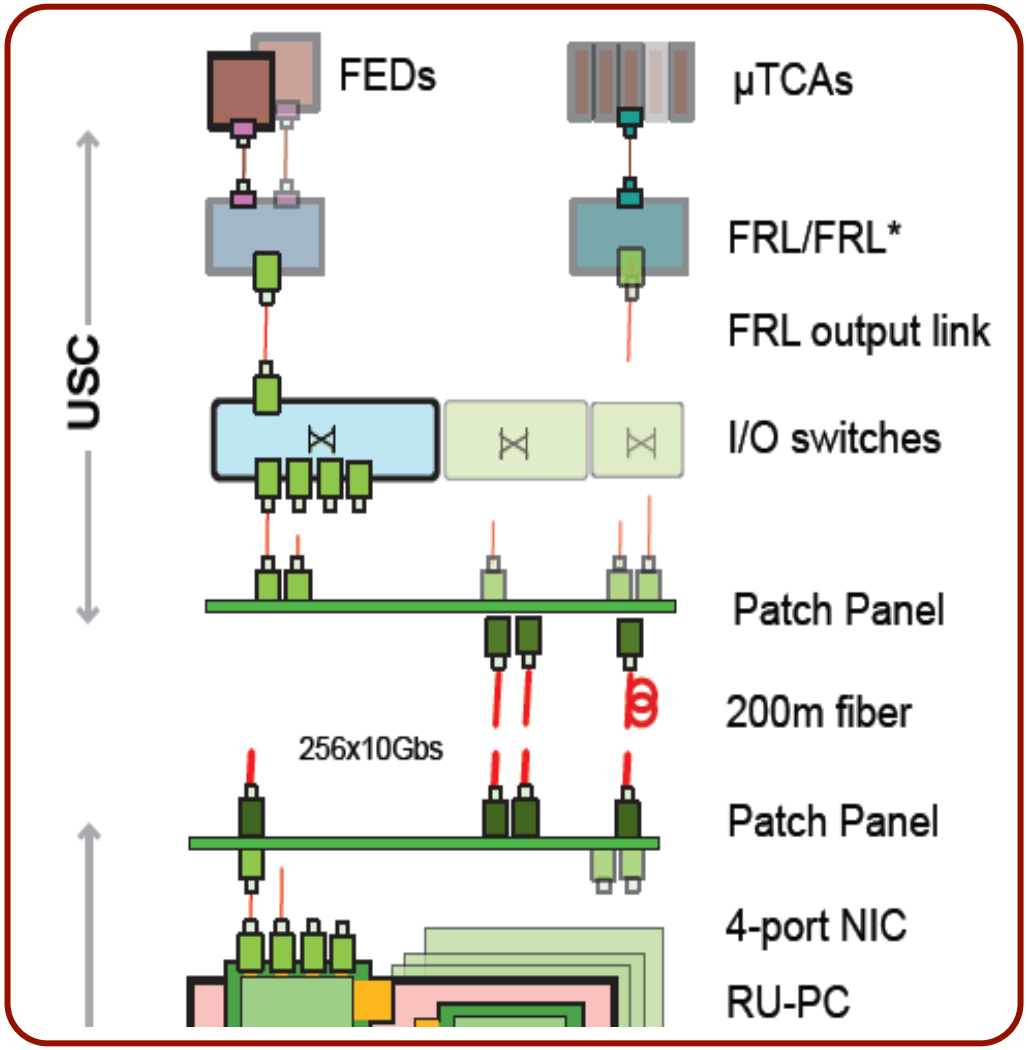
Frontend Readout Link
reads 1 or 2 links
Interfaces to Myrinet NIC

Build a new PCI card for the existing FRL

- Will implement new 10GB link to the RU
- Will be able to receive data via a fiber from newly built uTCA FEDs
- Pros: Small change wrt existing system. Rely on stable existing hardware
- Cons: bottle neck of internal PCI bus (520MB/s per FRL if reads SLINK)

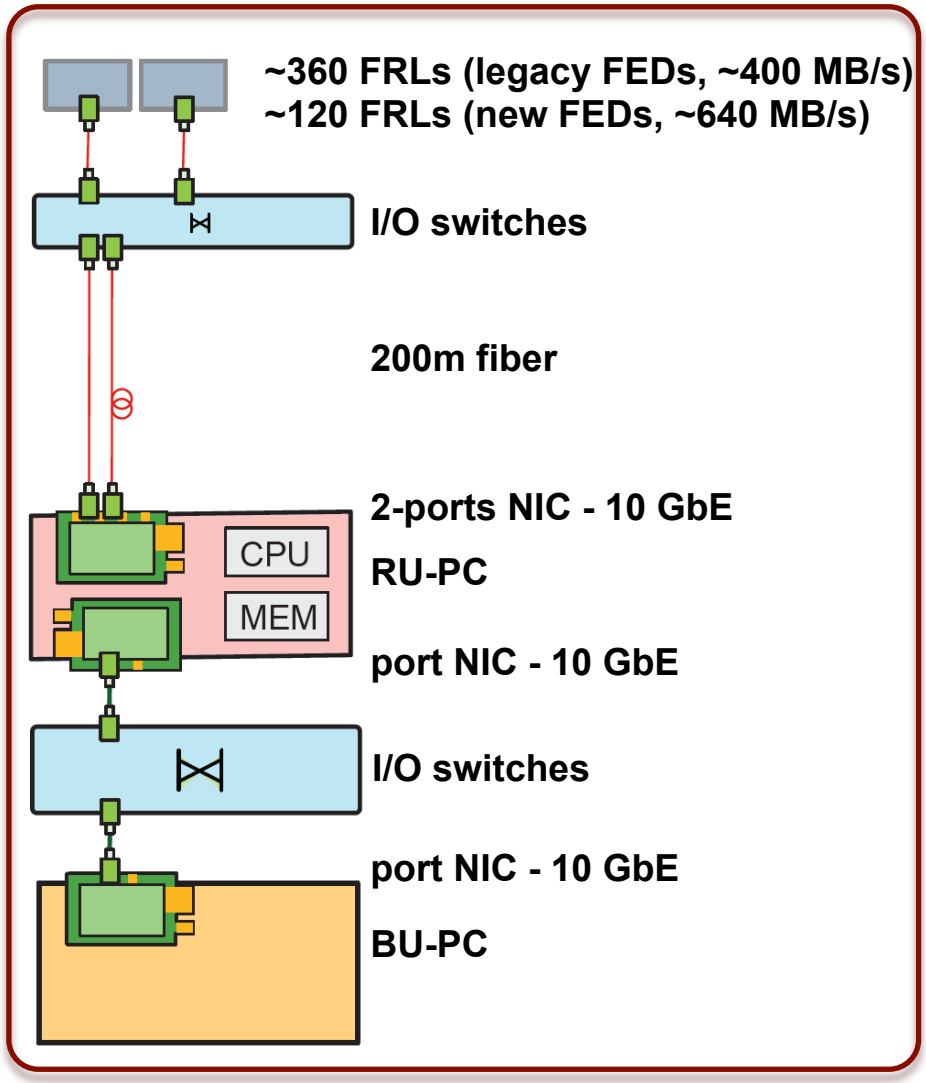


New Data to surface (D2S)



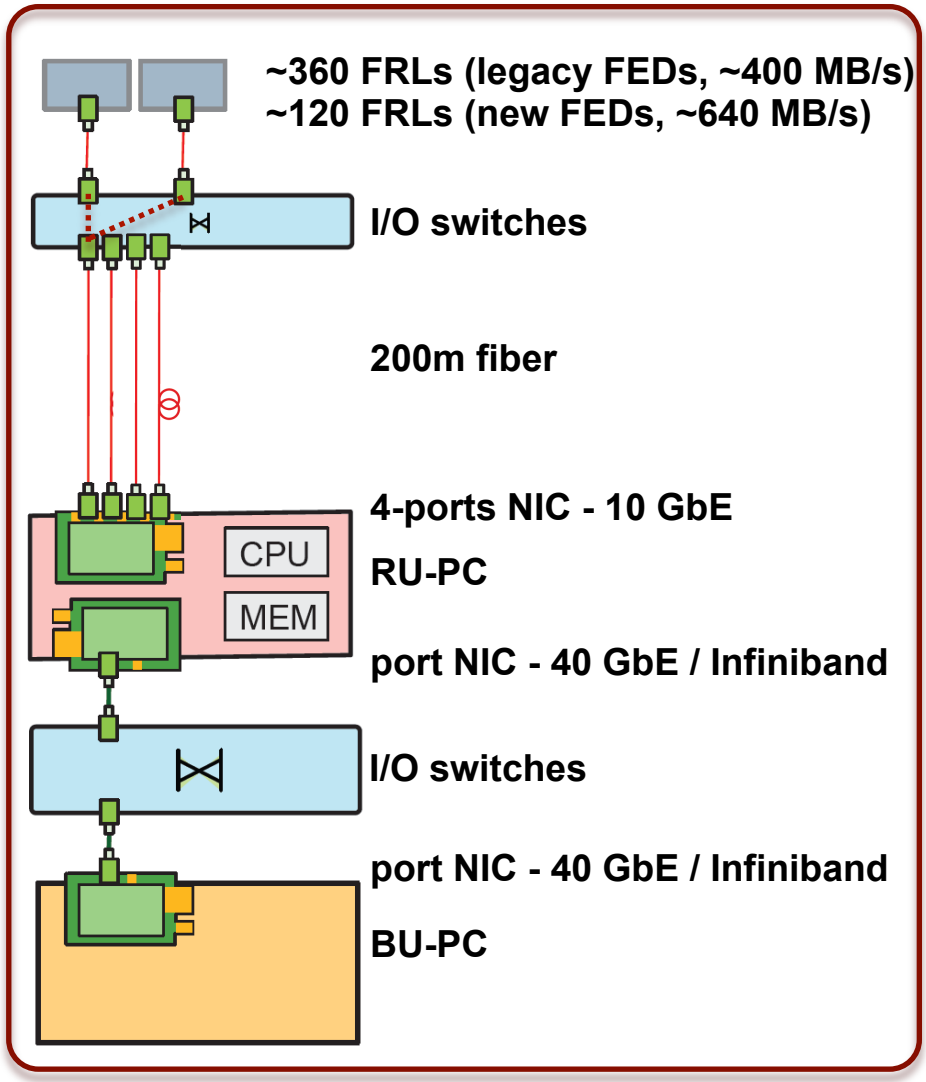
- Moving from myrinet 2x 2Gb/s to 10 Gb Ethernet
- Point to point protocol used between FRL and RU
- Concentrate FRLs sources in USC switch or/and RU

Event builder network: conservative option



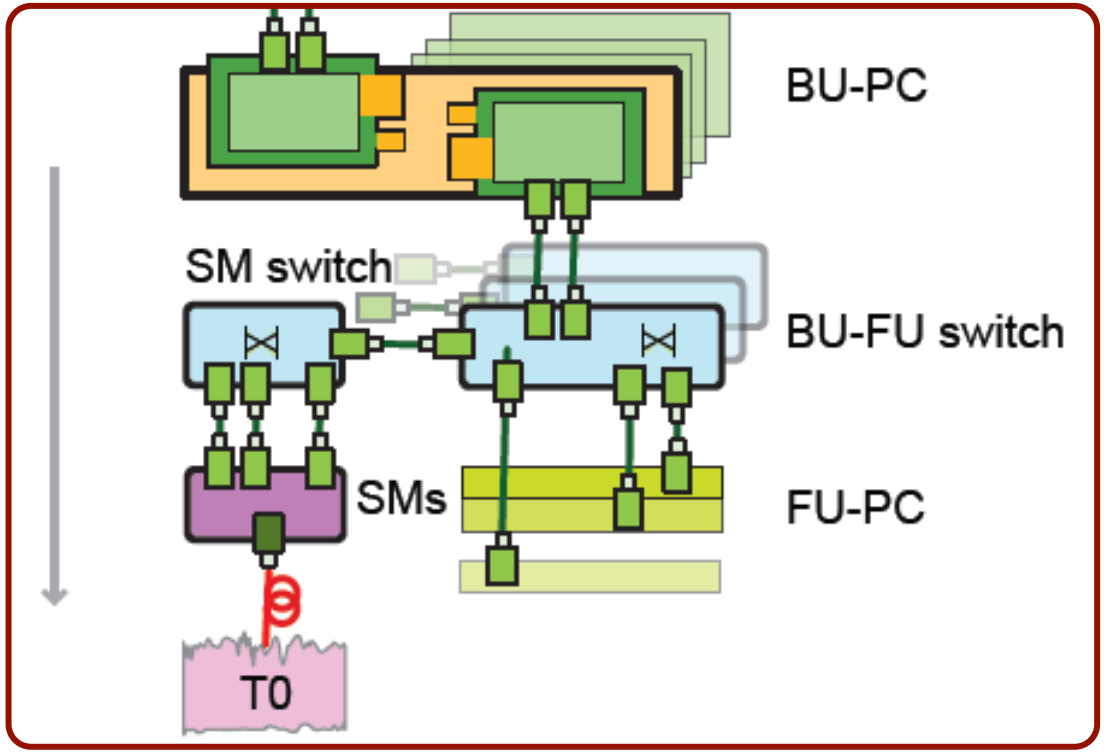
- Sources (in total 480)
  - At 100 kHz 2 to 4 KB fragments (legacy FEDs)
  - At 100 kHz 4 to 8 KB fragments (new FEDs)
- Input NIC 2 - 10 GbE ports
  - Required throughput per link100 kHz, 2 to 4 kB per fragment
- PC concentrates by factor 2 for legacy FEDs
  - 360 FRLs → 180 RUs
- Output NIC – 10 GbE port
  - Required throughput per link 100 kHz, 4 to 8 kB
- EVB configuration
  - 300 RU PCs
  - 300 x 300 switch 10 GbE

Event builder network: aggressive option



- Sources (in total 480)
  - At 100 kHz 2 to 4 KB fragments (legacy FEDs)
  - At 100 kHz 4 to 8 KB fragments (new FEDs)
- Front Switch
  - Concentrate by factor of 2 for legacy FEDs
- Input NIC 4 – 10 GbE ports
  - Required throughput per link 200 kHz, 2 to 4 kB
  - Required throughput per link 100 kHz, 4 to 8 kB
- PC concentrates by factor 4
- Output NIC – 40 GbE/ Infiniband port
  - Required throughput per link 100 kHz, 16 to 32 kB
- EVB configuration
  - 75 RU PCs
  - 75 x 75 switch 40 GbE / Infiniband

New Builder Units, Filter Unites and Storage Managers configuration



- Builder Units/Filter Units
  - Add 10 GbE switch to connect BUs to Fus
  - 2 links 10 GbE from BU and 1 link 10 GbE for each FU
- Storage Managers
  - 1 link 10 GbE from BU-FU switch to SM switch
  - 3 links 10 GbE for each Storage Manager

Feasibility studies on advanced networking technologies for event builder

Feasibility studies on advanced networking technologies

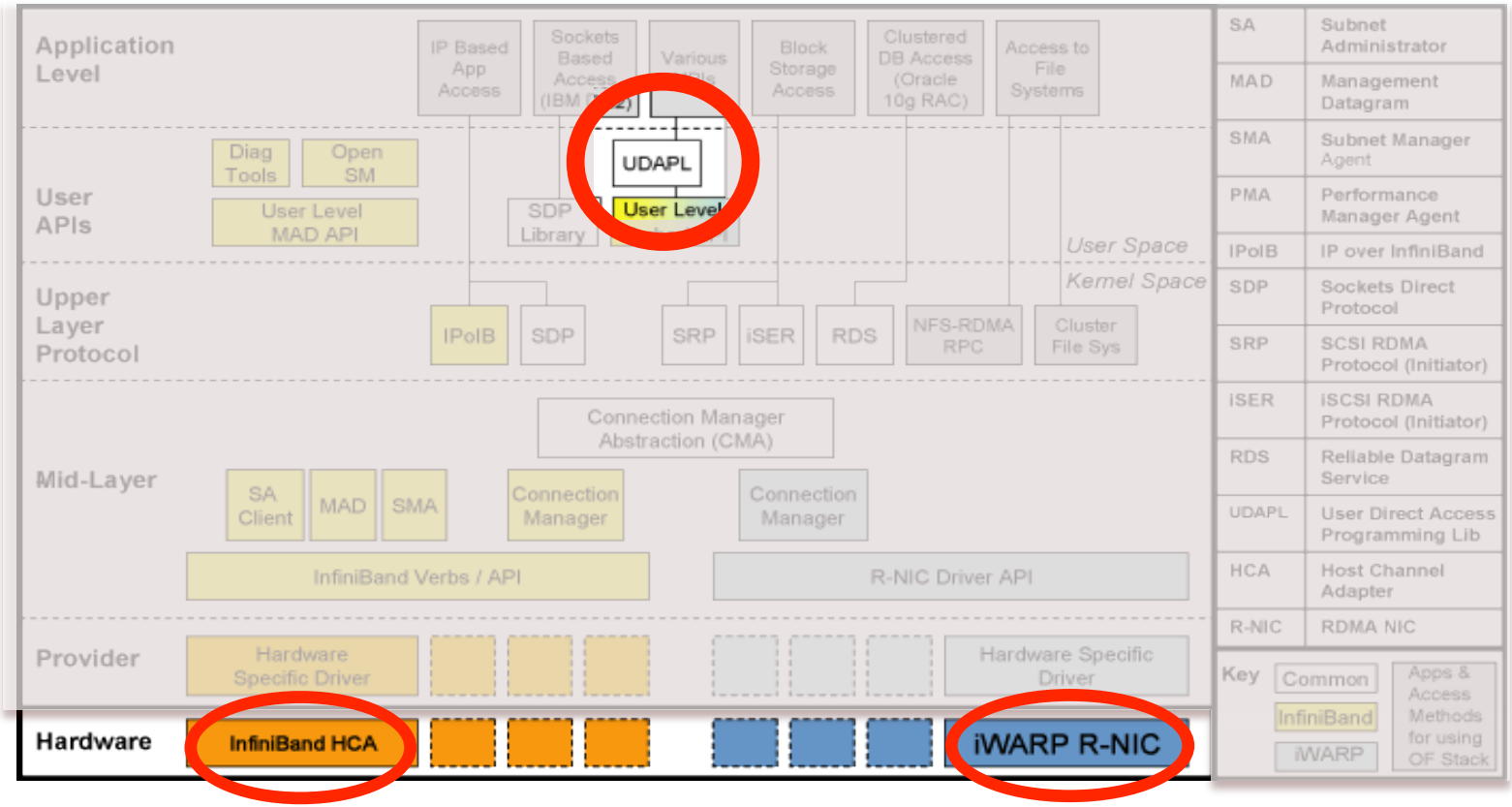
Our feasibility studies are focused in two network technologies:

- Ethernet
  - 10/40 Gigabit Ethernet (different vendors)
  - iWARP (RDMA) – TCP/IP full offload (Chelsio T4 Unified Wire Adapters)
  - performance measurements using TCP/IP and DAPL (Direct Access Programming Library- OpenFabrics)
- Infiniband
  - 4x quad data rate (QDR)
    - 40 Gb/s - 8B/10B encoding -32 Gb/s data rate
  - 4x fourteen data rate (FDR)
    - 56 Gb/s – 64B/66B encoding – 54.54 Gb/s data rate
  - performance measurements using DAPL (Direct Access Programming Library- OpenFabrics) and IPoIB (IP over InfiniBand)

The OFED Stack (source: OpenFabrics Alliance)

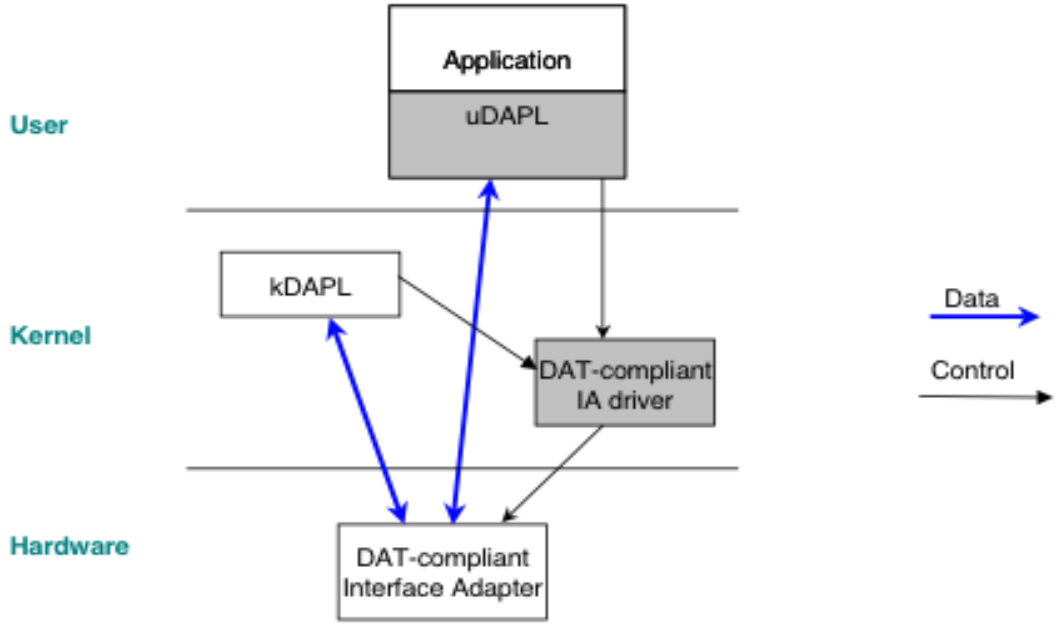
A unified, cross-platform, transport-independent software stack for RDMA and kernel bypass

- http://www.openfabrics.org/



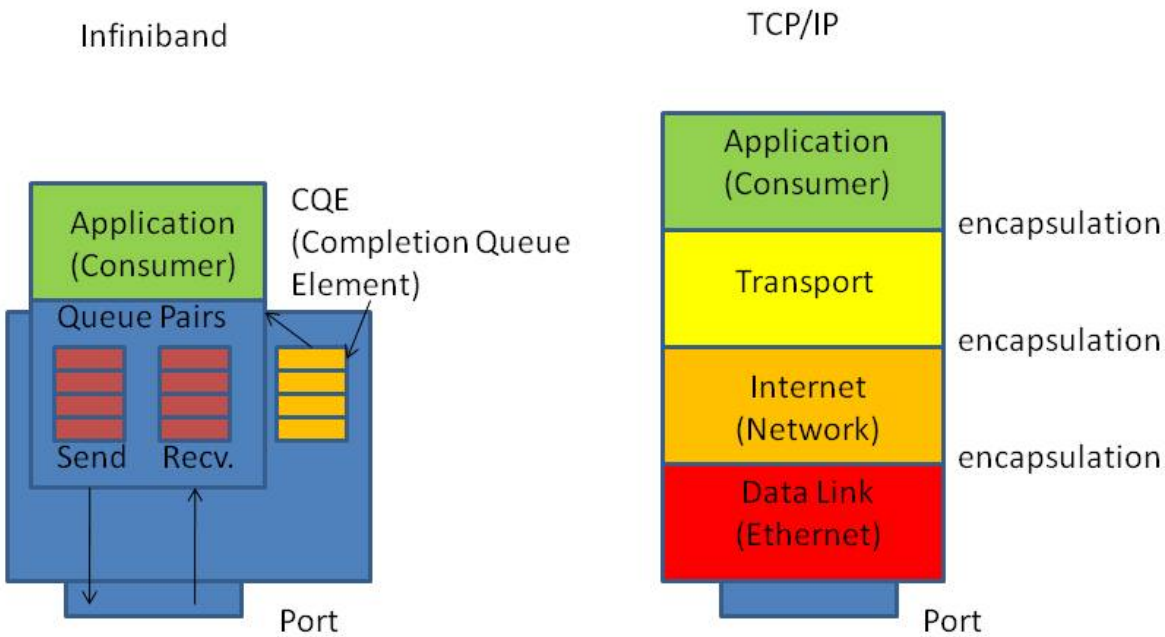
DAT Model (source DAT Collaborative)

- Developed by DAT collaborative
  - http://www.datcollaborative.org/
- Transport and platform (OS) independent
- Define user (uDAPL) and kernel (kDAPL) APIs
- DAT supports reliable connection
- Data Transfer Operations send, receive, rdma_read, rdma_write
- uDAPL Version 2.x, January, 2007



Comparison of the Stacks (Infiniband vs TCP/IP)

- The protocol is defined as a very thin set of zero copy functions when compared to thicker protocol implementations such as TCP/IP



CMS online software for data acquisition system

CMS online software

The CMS online application are based on XDAQ framework.

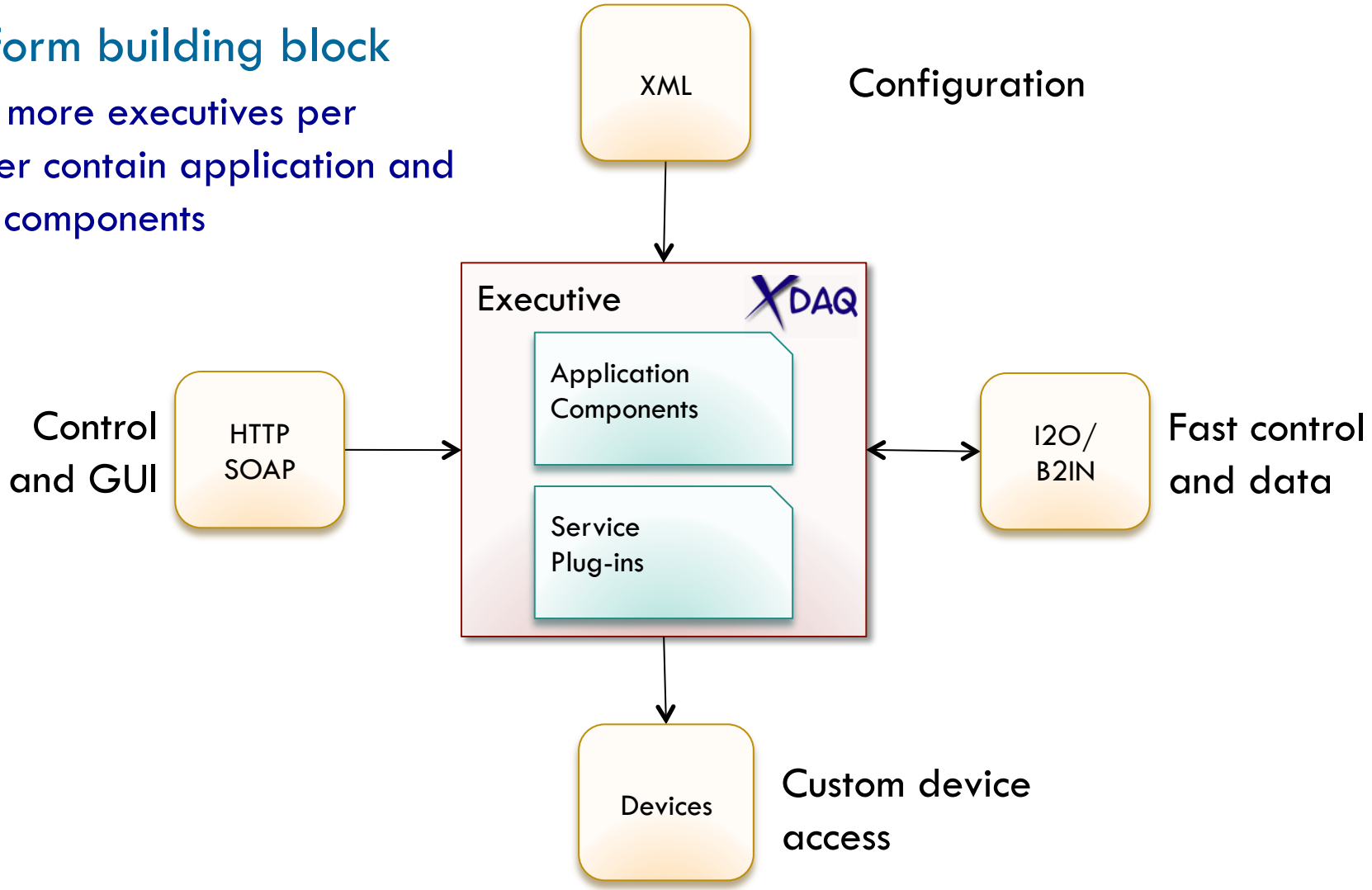
What is the XDAQ framework?

- XDAQ is a software platform designed specifically for the development of distributed data acquisition systems.
- XDAQ is the definitive across the board middleware that eases the tasks of designing, programming and managing data acquisition applications by providing a simple, consistent and integrated distributed programming environment.
- The framework builds upon industrial standards, open protocols and libraries.

Architecture Foundation

- Uniform building block

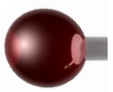
One or more executives per computer contain application and service components



Architecture Foundation (II)

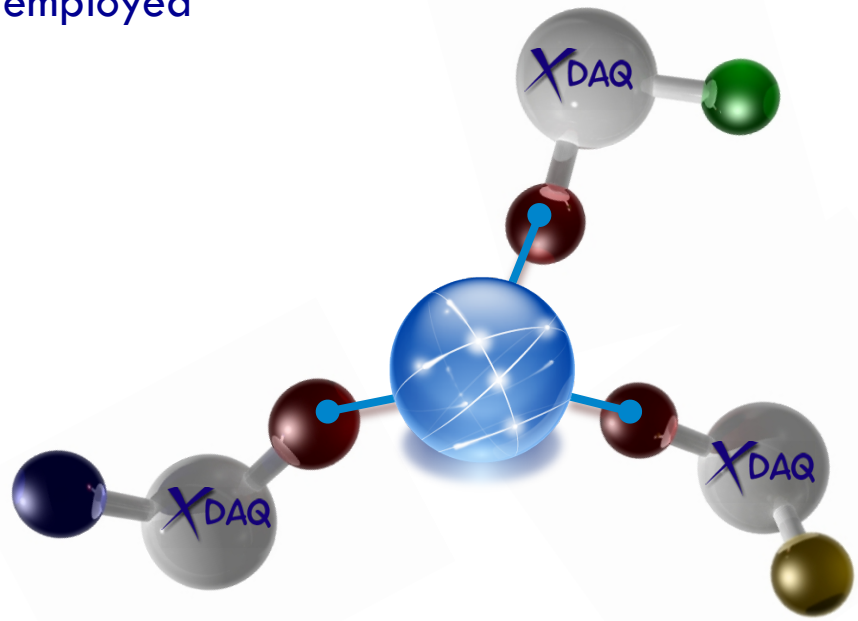
Peer-Transport a pluggable layer of various networking medium

- The XDAQ framework provides peers with a mechanism for determine a route to an endpoint, allowing the peer to send data to the remote endpoint
- Peer transports are the entity responsible for conducting the actual exchange of information over a network
- Peer transport encapsulate a set of network interfaces, allowing a peer to send and receiver data independently of the type of network being employed



pluggable Peer-Transport:

- SOAP/HTML: HTTP
- I_2O: TCP, Myrinet, FIFO, etc.



Application code does not change when moving from Ethernet to Myrinet or Infiniband

Pluggable Peer-Transport for DAT library

- pt*u*DAPL is a new peer-transport for DAT library (Infiniband and iWarp) using I_2O messaging
  - Use of smart memory pool based on uDAPL memory region allocator (random access to memory with no intermediate management by using cookies)
  - Profiting for inherent non blocking and queuing of uDAPL API for minimizing latency
  - All I/O operations centered on dedicated uDAPL memory pool (full zero-copy between XDAQ applications and DAPL driver)
  - Based on DAT Spec 2.x

Test setups

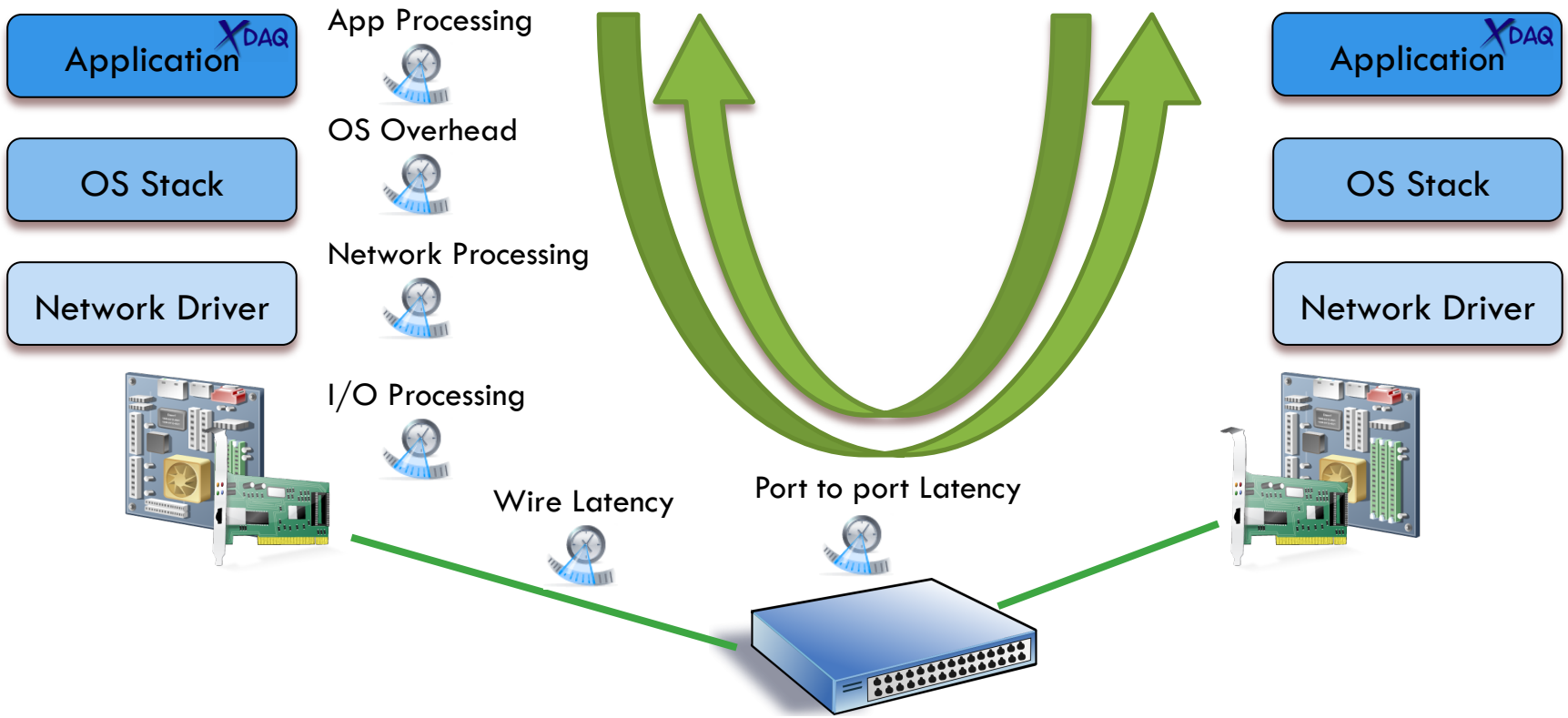
Test setups

	Setup 1 (LHCb)		Setup 2 (daqval 2)	Setup 3 (daqval 2)	
Nodes	8		16	32	
Type	DELL R710		DELL R310	DELL C6220	
CPU	Xeon E5530 2x 4-core at 2.27 GHz		Xeon X3323 2x 4-core at 2.50 GHz	Xeon E5-2670 2x 8-core at 2.6 GHz	
Memory	3 GB		4 GB	32 GB	
Network	Ethernet	Infiniband	Ethernet	Ethernet	Infiniband
Adapter	Chelsio T420-CR 10GBASE-SFP RNIC (iWarp)	Qlogic HCA, qle7340 4x QDR PCIe	Silicom PE210G2SPI9-SR	Mellanox - ConnectX-3 VPI (single FDR or 40GbE) - MCX353A-FCBT	
Switch	Voltaire Vantage 6048, 48 ports, 10 GbE	Qlogic 12300-BS01, 36 ports, 4x QDR	Brocade MLXE16, 16 ports, 10 GbE	Mellanox 36 - port 40 GbE Switch - MSX1036B-1SFR	Mellanox - IB switch (36 ports FDR QFSP) - MSX6025F-1

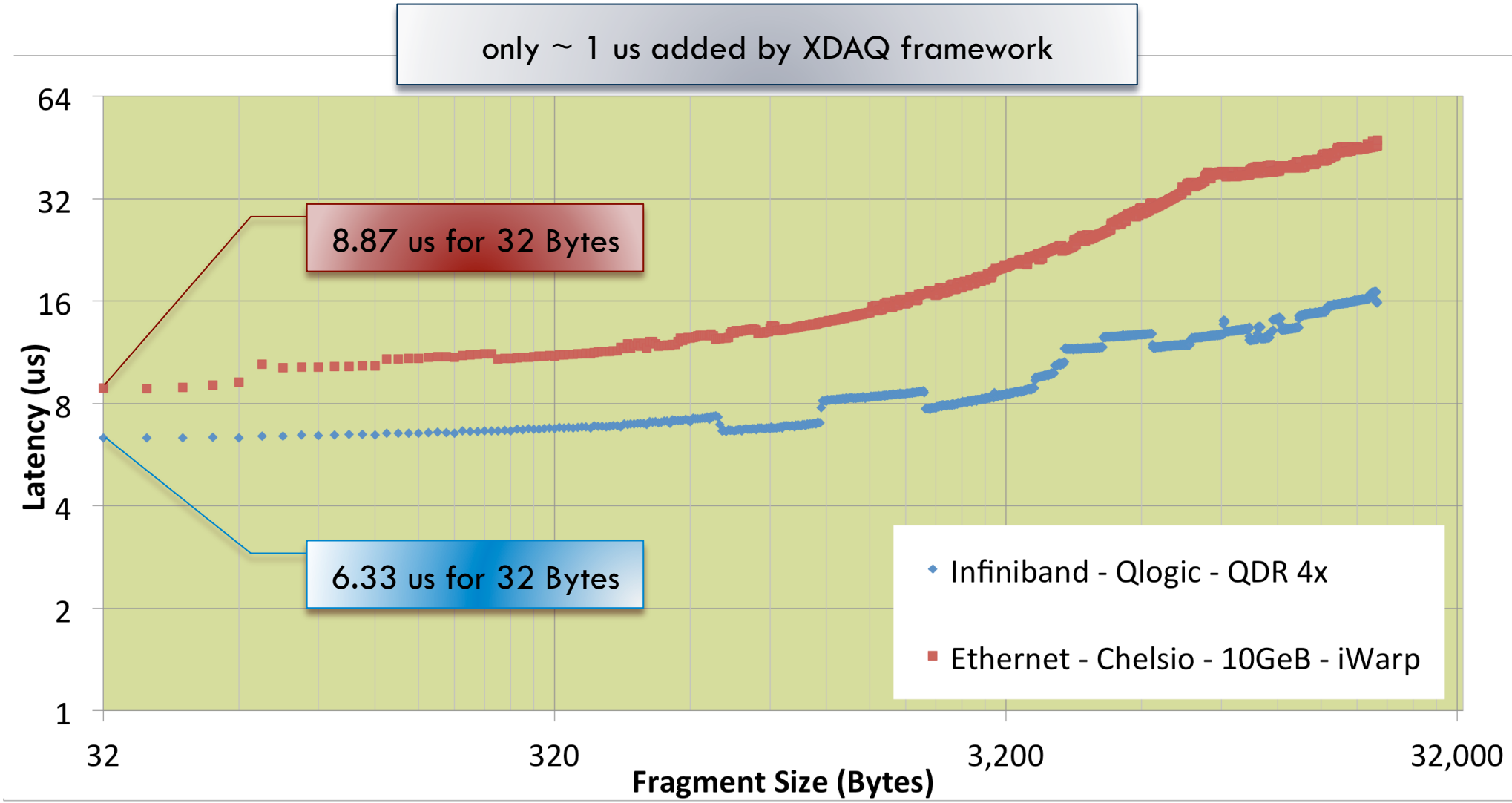
Preliminary Measurements

Roundtrip

- Simple XDAQ application to compute the One-way delay
- Time packet to travel from a specific source to a specific destination and back again
- One-way latency is measured by timing a round-trip message and dividing the obtained result by two



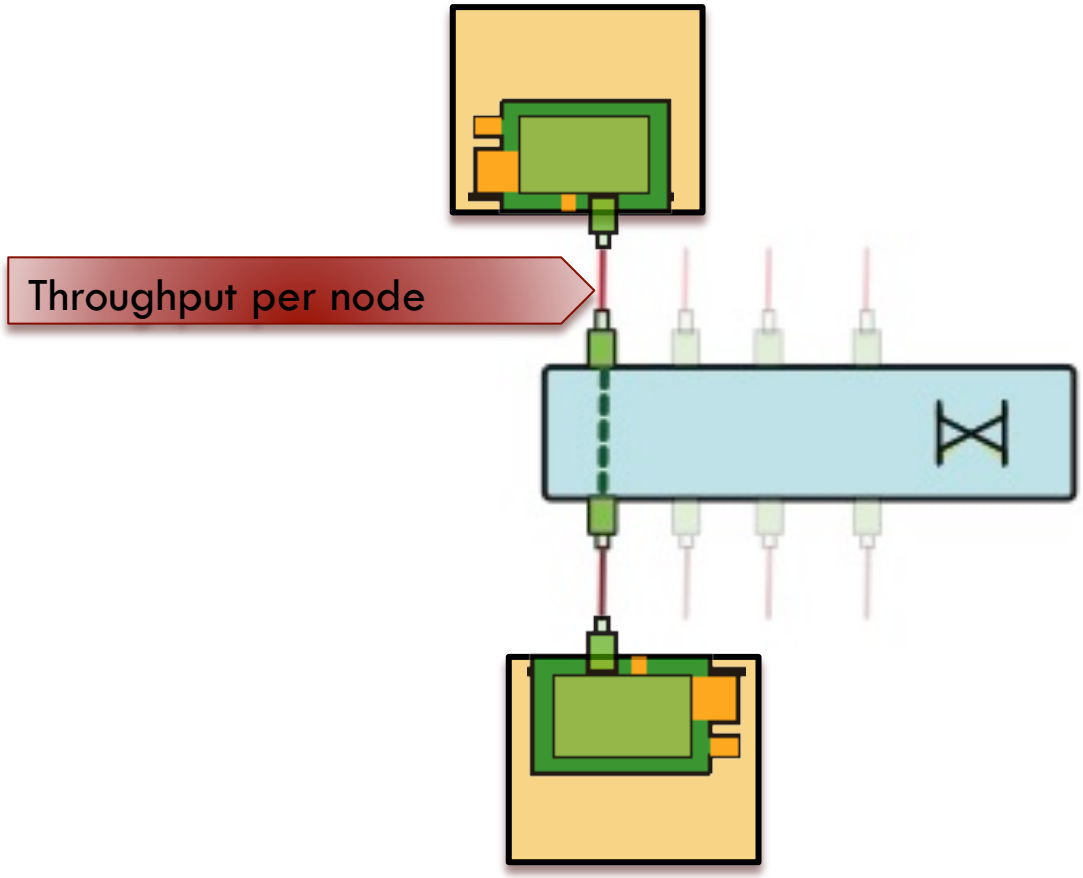
Latency measurements for ptuDAPL with Infiniband and iWarp



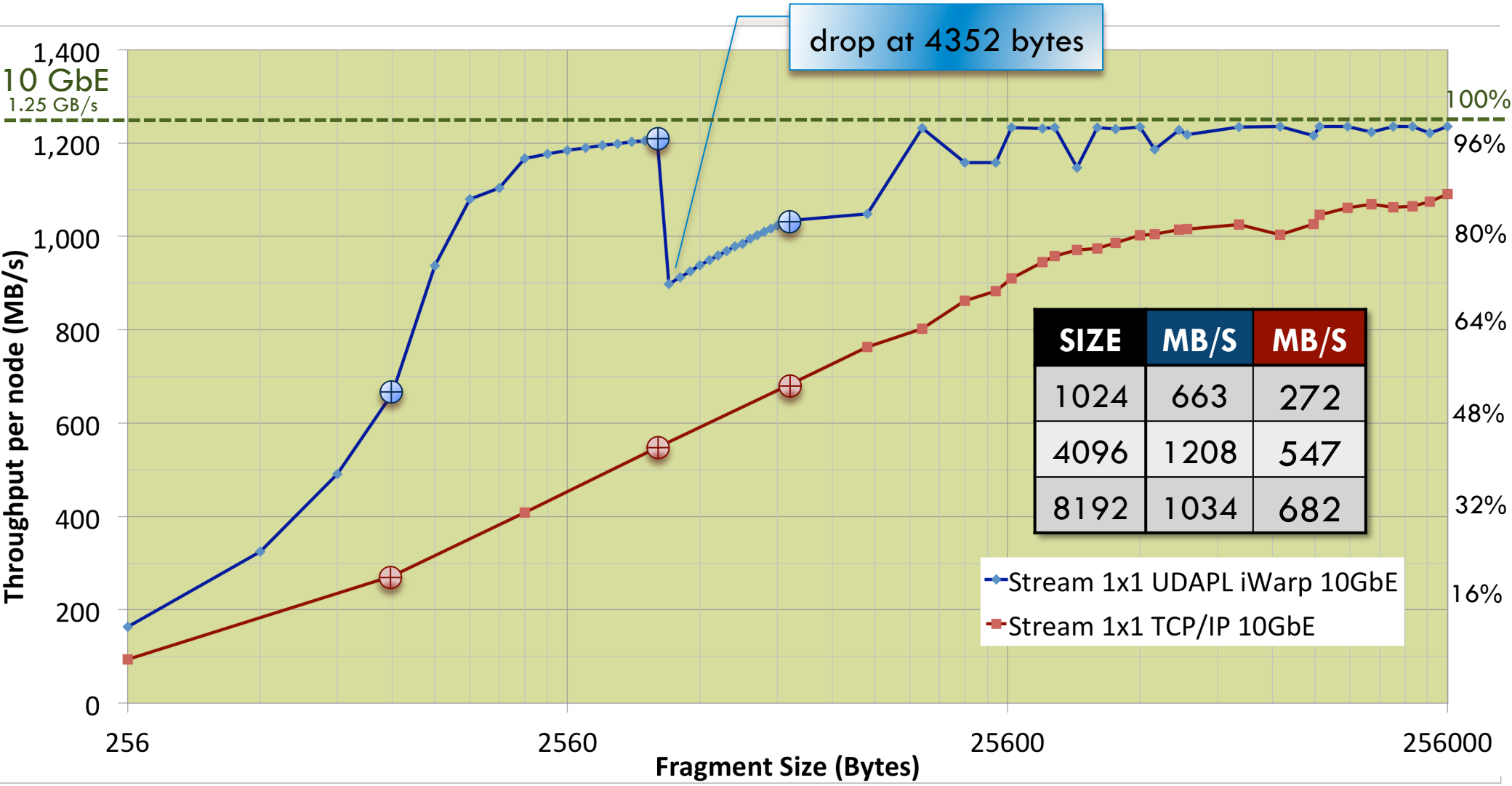
Setup 1

Stream I/O

- Unidirectional throughput is measured using a unidirectional send of N messages. Time sampling is done at the receiver side and starts with the first incoming message



Stream I/O (TCP/IP vs uDAPL) – Chelsio T4 – 10 GbE

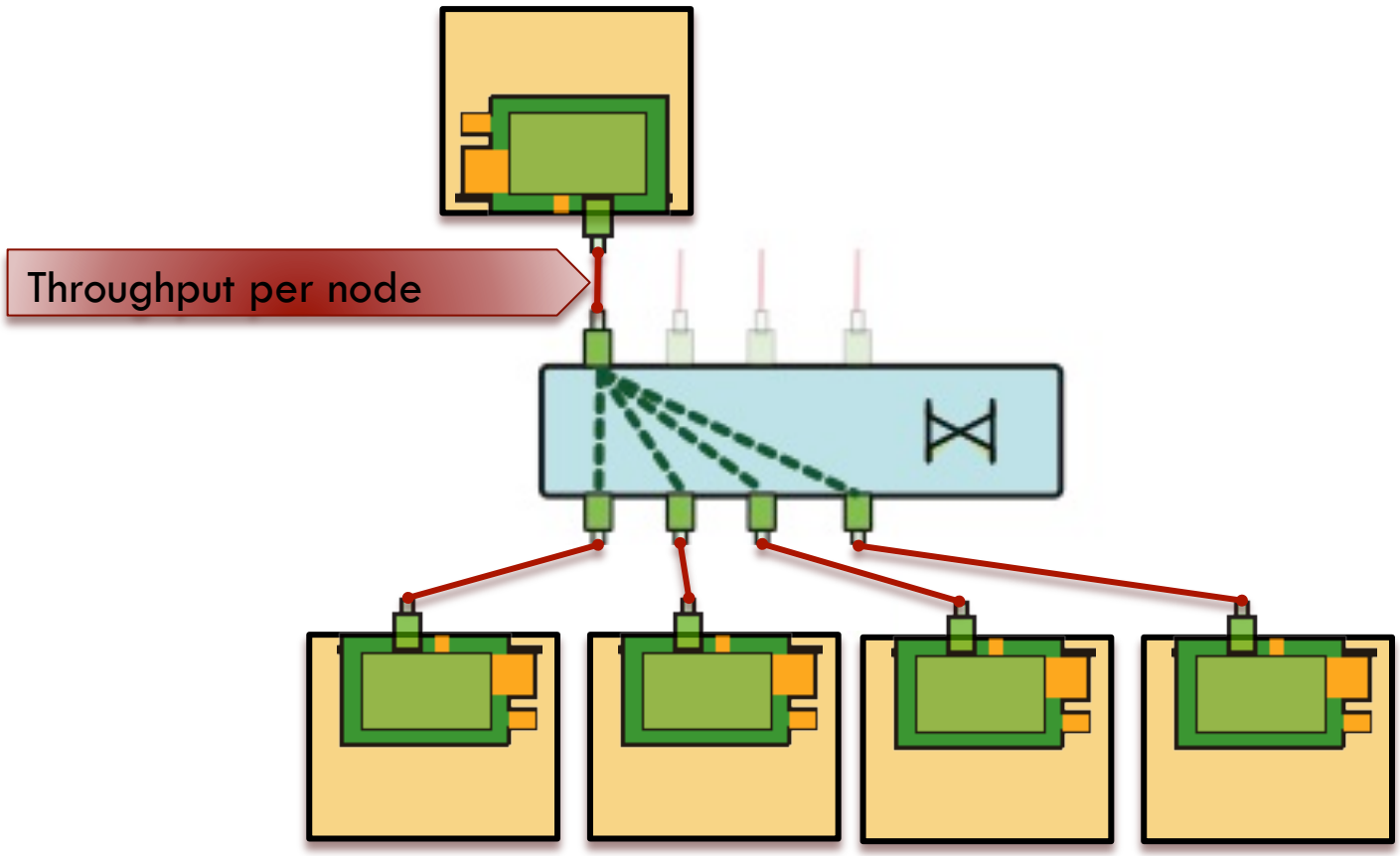


SIZE	MB/S	MB/S
1024	663	272
4096	1208	547
8192	1034	682

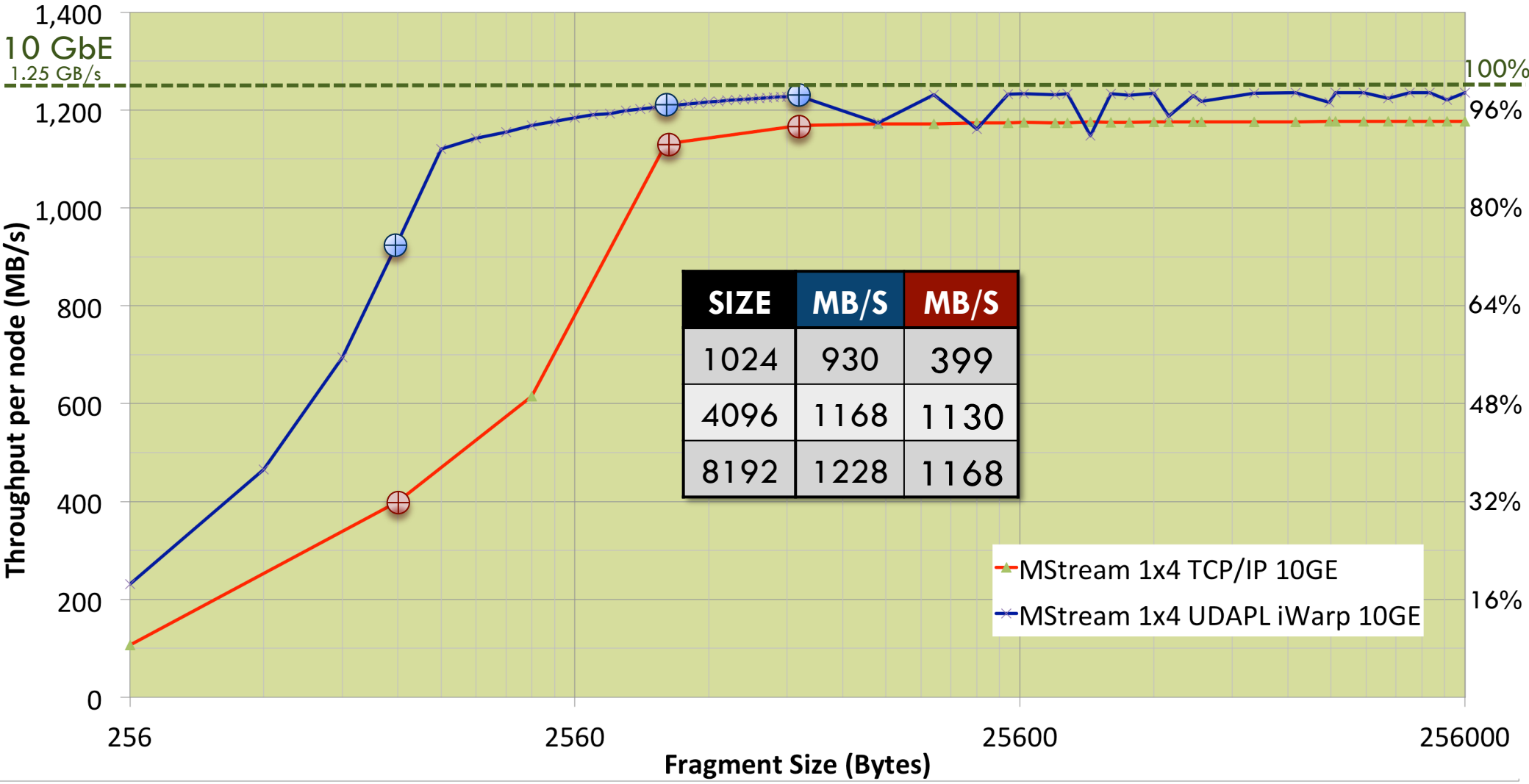
Setup 1

Multi-Stream I/O

- Unidirectional throughput (bandwidth) is measured using a unidirectional send of N messages to N receivers. Time sampling is done at the receivers side and starts with the first incoming message.



Multi-stream I/O 1 to 4 (TCP/IP vs uDAPL) – Chelsio T4 – 10 GbE

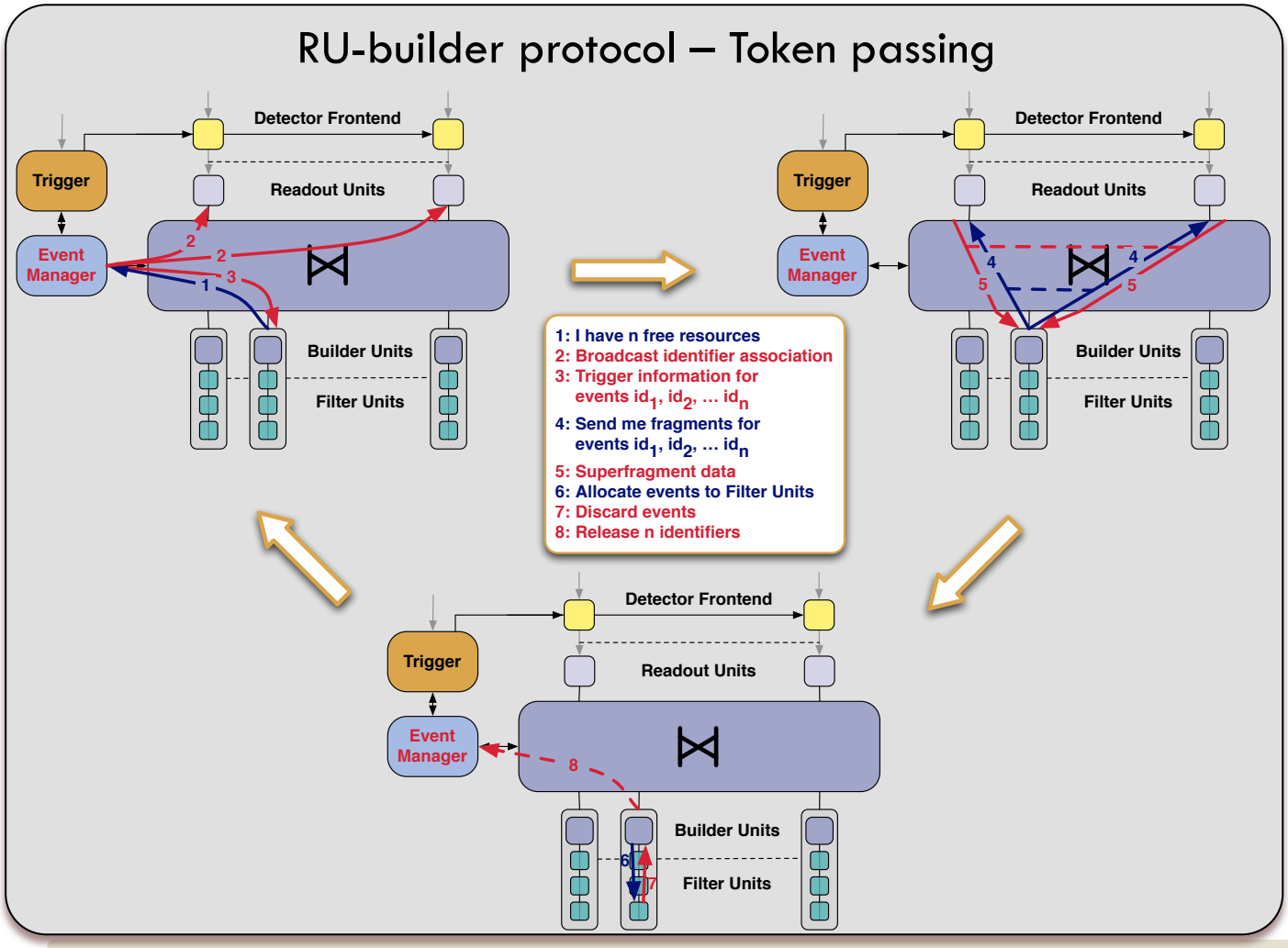


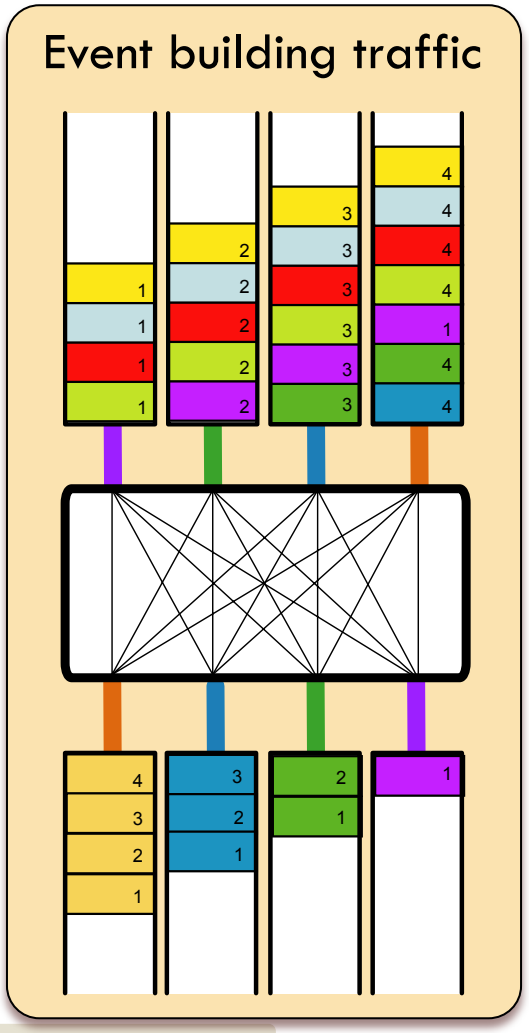
SIZE	MB/S	MB/S
1024	930	399
4096	1168	1130
8192	1228	1168

Setup 1

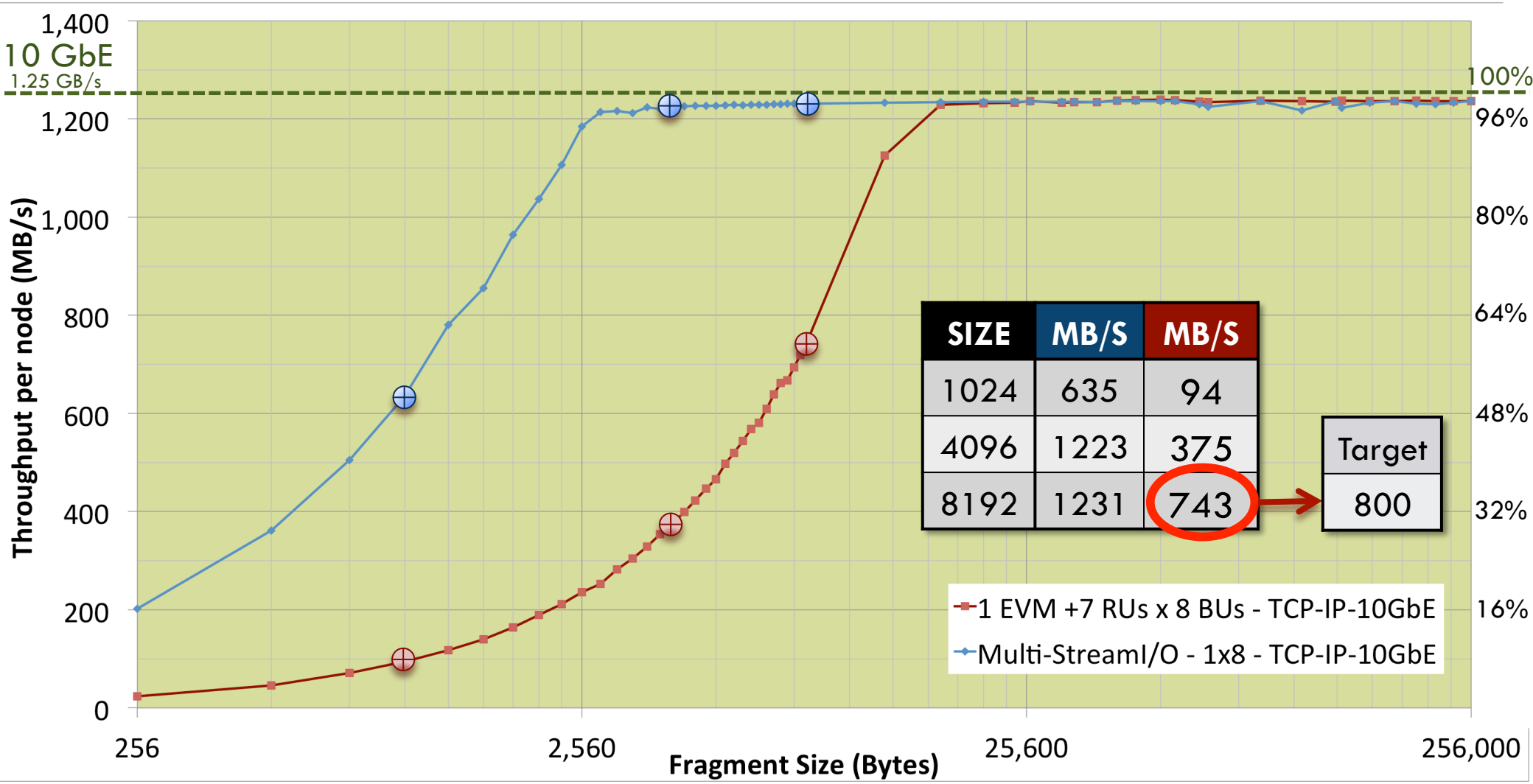
Event Builder

- CMS RU-builder
  - Currently used in CMS DAQ for data taking





Event builder 1+7x8 vs Multi-stream 1x8 (TCP/IP - 10GbE)

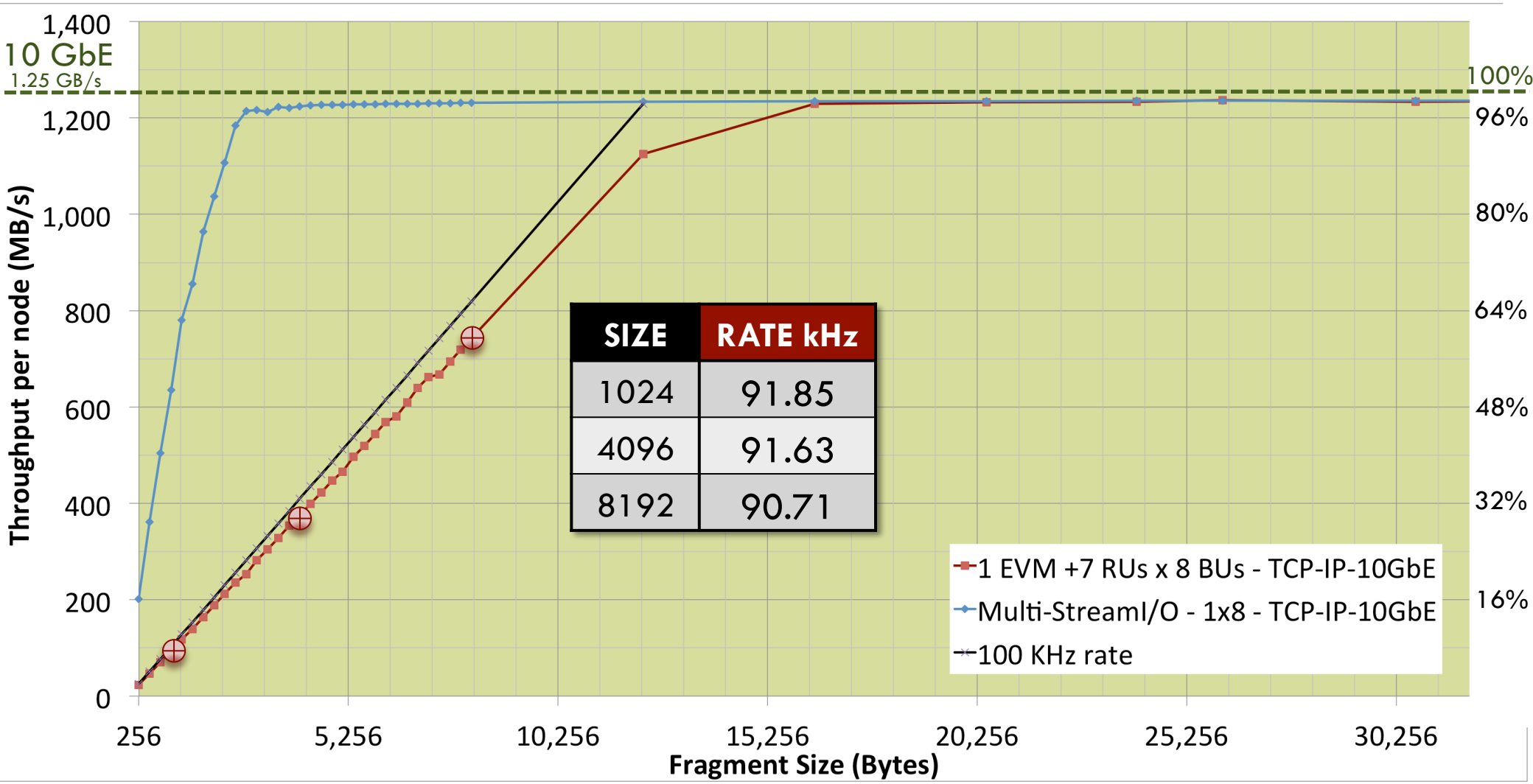


SIZE	MB/S	MB/S
1024	635	94
4096	1223	375
8192	1231	743

Target
800

Setup 2

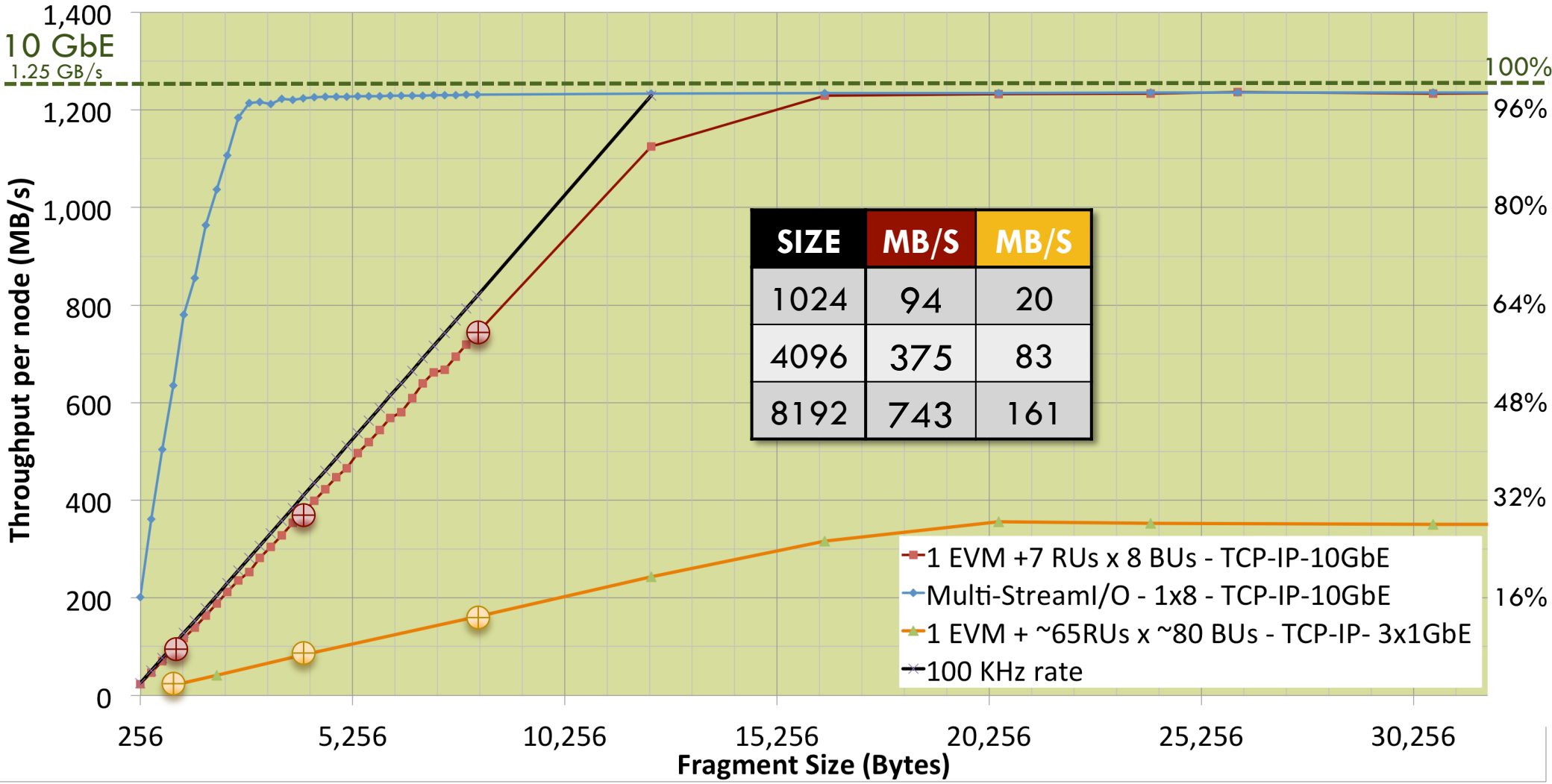
Event builder 1x7x8 vs 100 KHz (TCP/IP - 10GbE)



SIZE	RATE kHz
1024	91.85
4096	91.63
8192	90.71

Setup 2

Event builder 1x7x8 vs current event builder (TCP/IP - 10GbE)

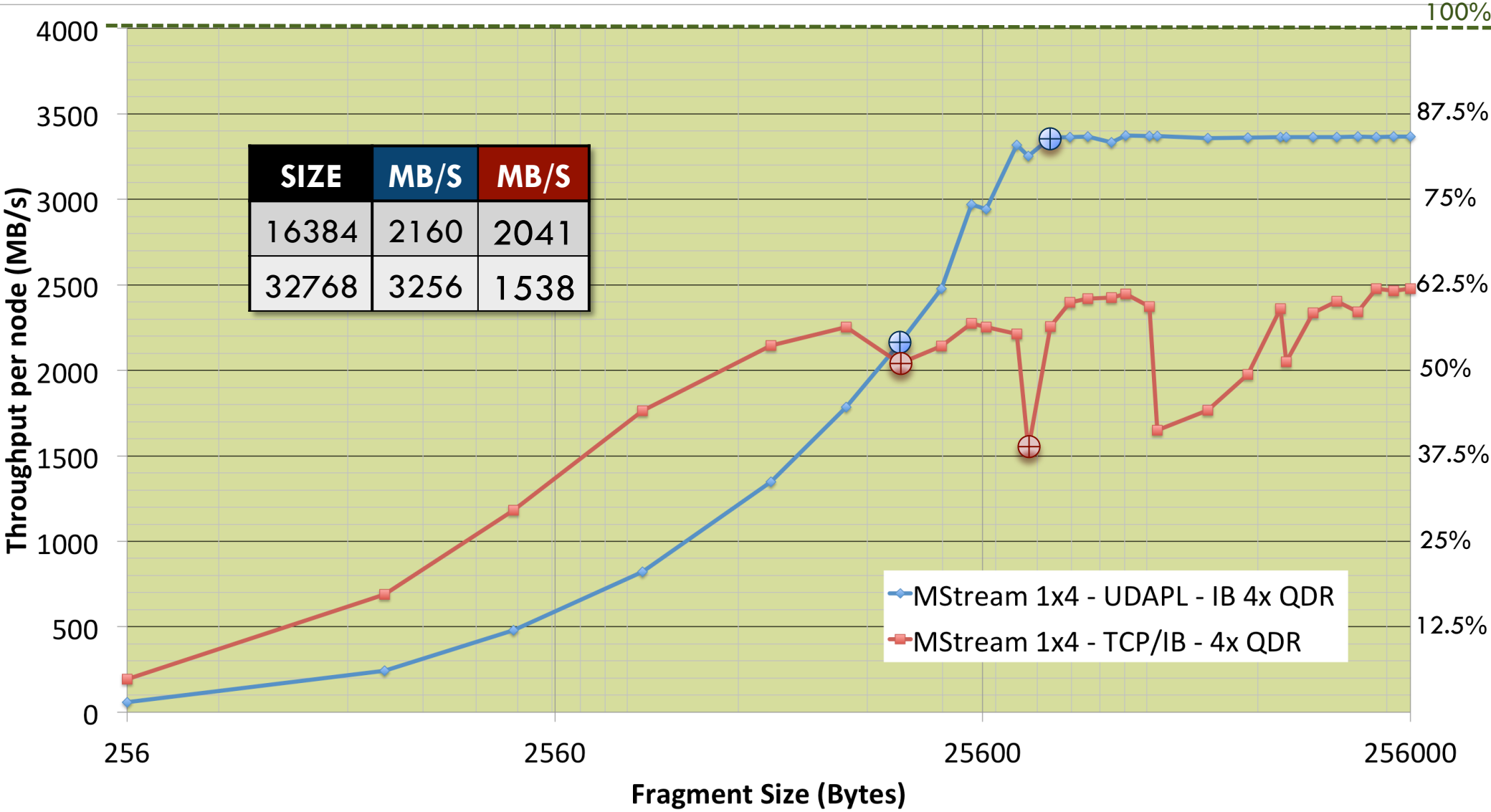


SIZE	MB/S	MB/S
1024	94	20
4096	375	83
8192	743	161

Setup 2

Infiniband Measurements

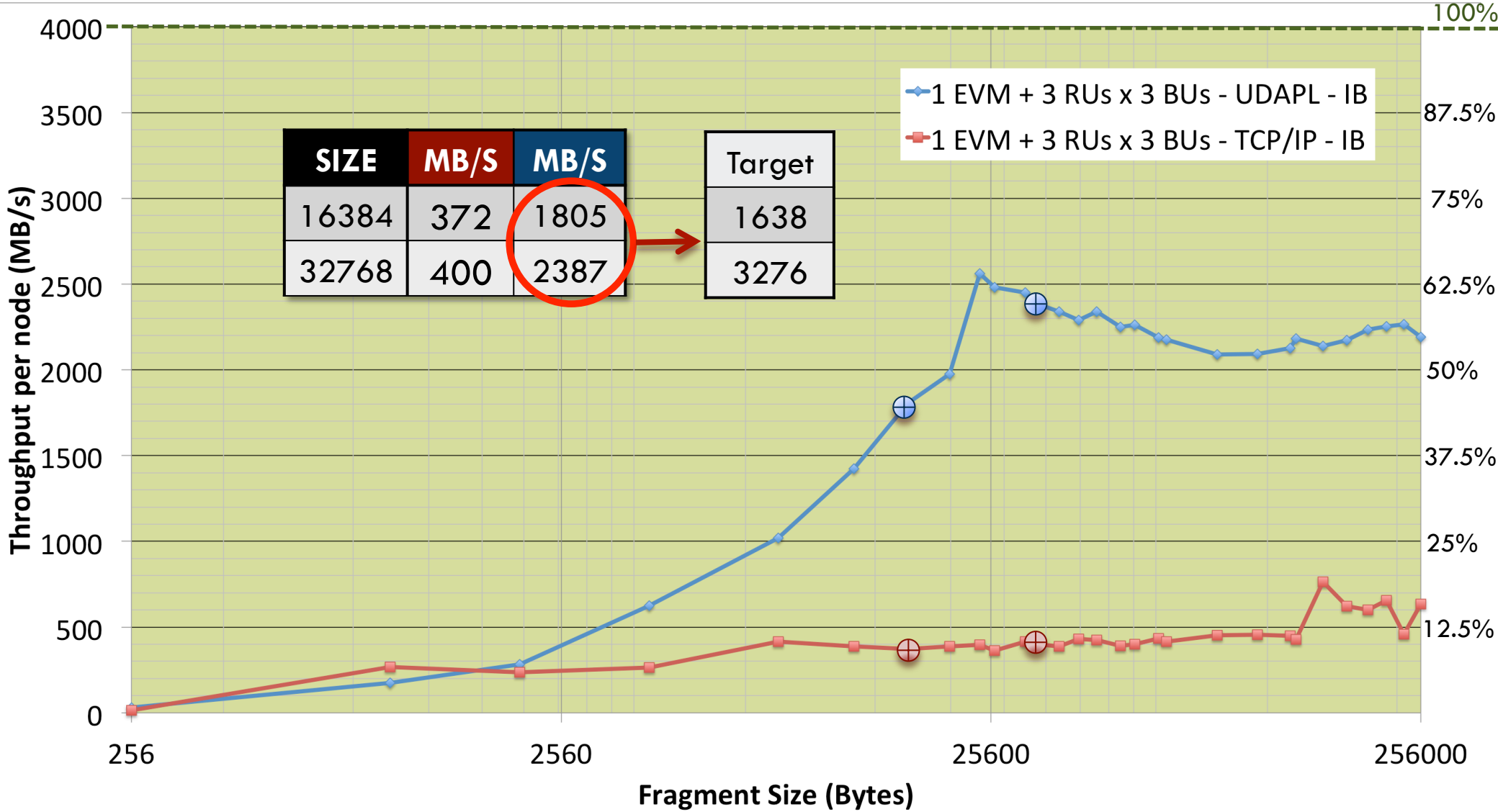
MStream 1x4 uDAPL - IB vs TCP/IP - IPoIB



SIZE	MB/S	MB/S
16384	2160	2041
32768	3256	1538

Setup 1

Event builder 1+3x3 uDAPL - IB vs TCP/IP - IPoIB



SIZE	MB/S	MB/S
16384	372	1805
32768	400	2387

Target
1638
3276

Setup 1

Summary

- New readout system must be compatible with old data sources and new data sources
- New DAQ column architecture replaces aging existing hardware and improves readout performance
- Feasibility studies on advanced networking technologies for event builder
  - Different network technologies under investigation
  - Possibly Quality of Service will give high throughput and scalability

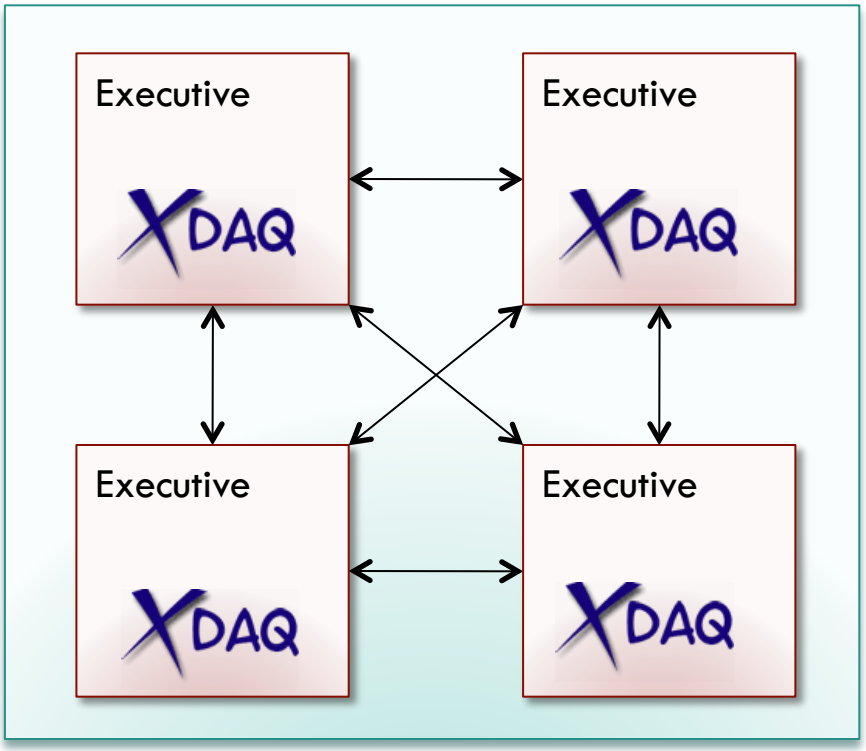
Thank you for you attention!

Backup

Architecture Foundation (II)

- Replicated building blocks

Scalable cluster system architecture



Software Distribution

coretools

Core framework

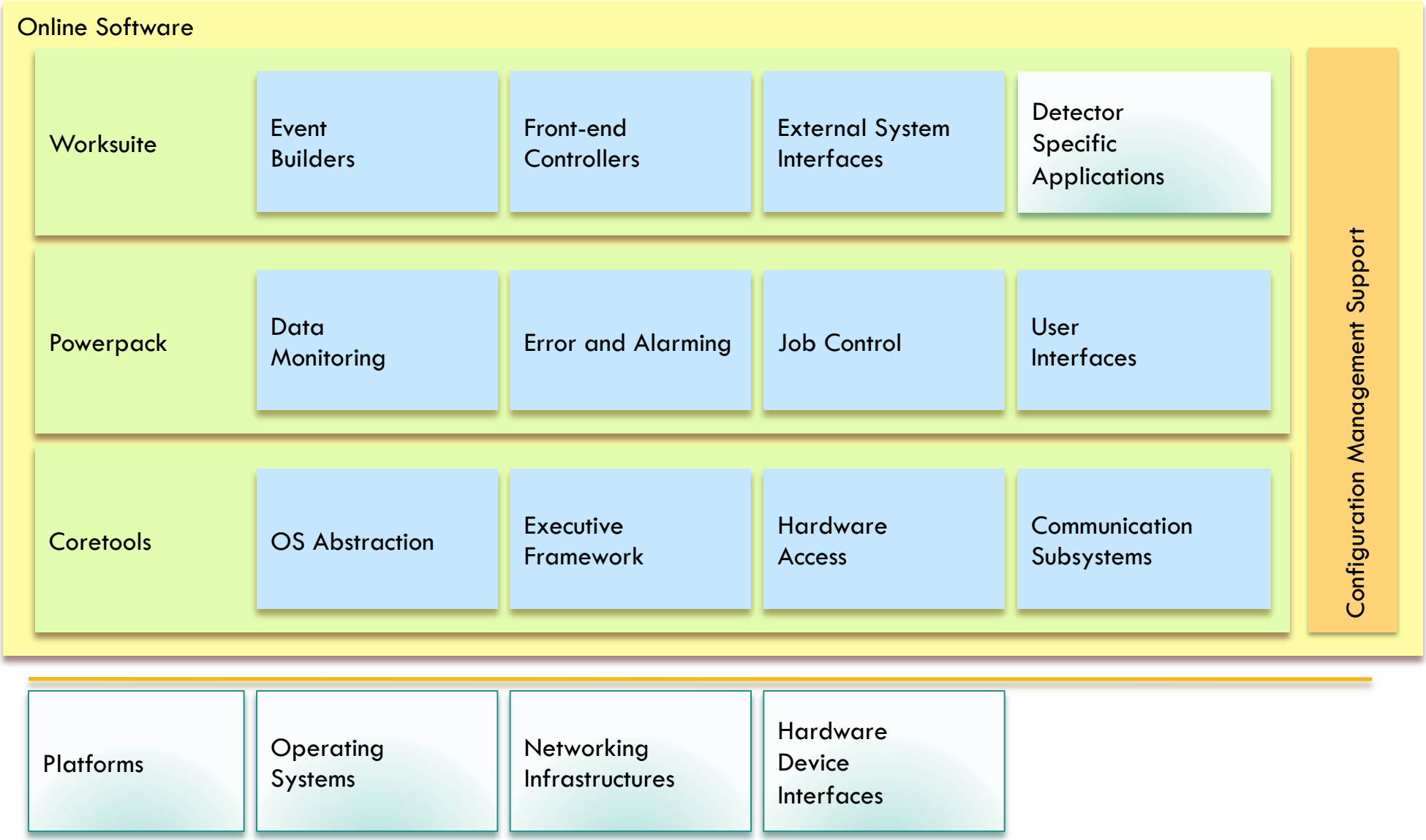
powerpack

Reusable applications

worksuite

CMS specific applications

Layered View



Timeline

Well consolidated after twelve years of development and use

2012 — Successefully **used in the first tree years of LHC**

2008 — **Commissioning and first beam event** successfully achieved

XMAS – monitoring, orthogonal to applications

2006

2004 — XDAQ 3 – **experiment wide adoption**

2002

2000 — XDAQ 2 – Web enters DAQ (SOAP)

First version of XDAQ - I_2O communication kernel

Outlook

- Integrated web technologies
- Monitoring
- Errors and Alarms
- Reusable Event Builder