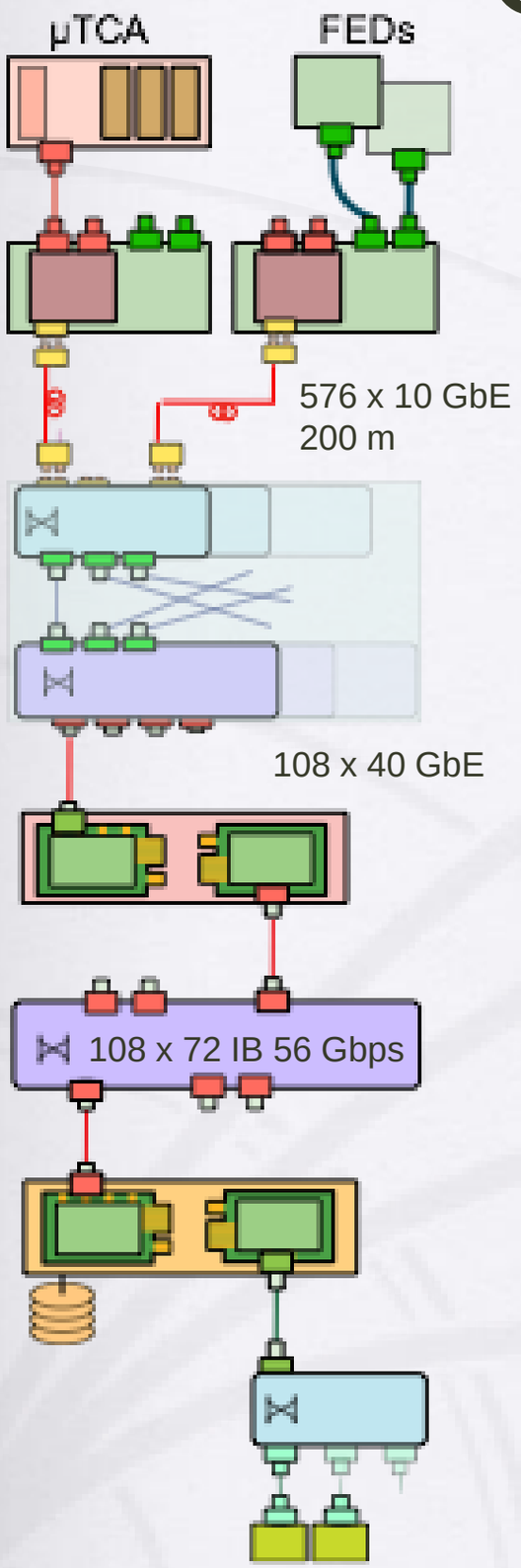
# CMS Run 2 Event Building

*Remigius K Mommsen*
*Fermilab*

# Overview

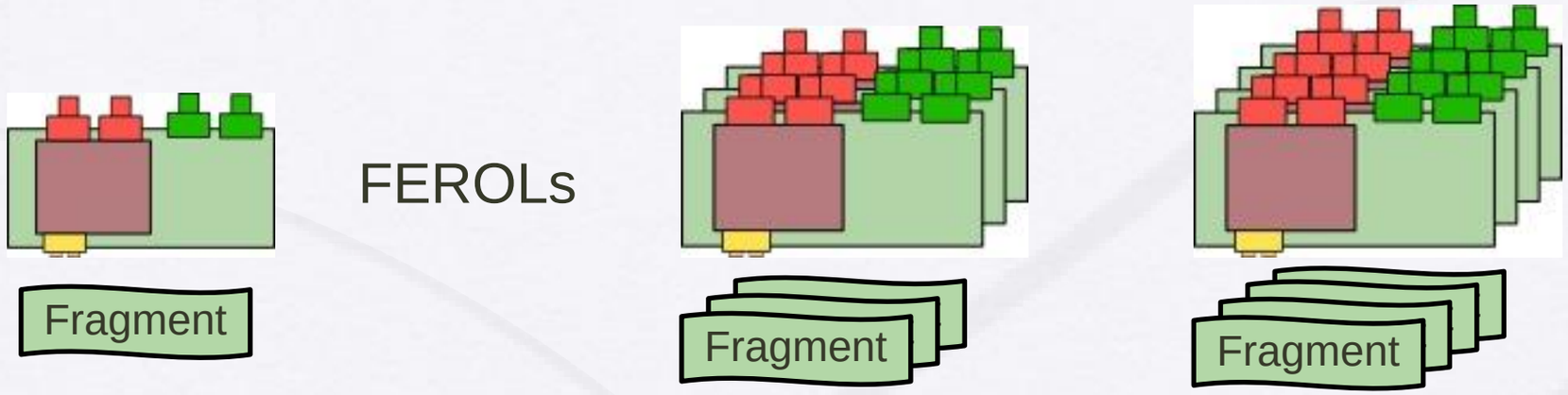
- Overview of CMS event builder for LHC run II
- Event-building protocol
- Measurements

# CMS Event Builder




- Detector front-end (custom electronics)
- Front-End Readout Optical Link (FEROL)
  - Optical 10 GbE TCP/IP
- Data Concentrator switches
  - Data to Surface
  - Aggregate into 40 GbE links
- Up to 108 Readout Units (RUs)
  - Combine FEROL fragments into super-fragment
- Event Builder switch
  - Infiniband FDR 56 Gbps CLOS network
- 72 Builder Units (BUs)
  - Event building
  - Temporary recording to RAM disk
- Filter Units (FUs)
  - Run HLT selection using files from RAM disk

# EvB Protocol



# Achieving Performance

Avoid high rate of small messages

- Request multiple events at the same time
- Pack data of multiple events into one message

Avoid copying data

- Operate on pointers to data in receiving buffers
- Copy data directly into RDMA buffers of Infiniband NICs
- Stay in kernel space when writing data

Parallelize the work

- Use multiple threads for data transmission and event handling
- Write events concurrently into multiple files

Bind everything to CPU cores and memory (NUMA)

- Each thread bound to a core
- Memory structures allocated on pre-defined CPU
- Interrupts from NICs restricted to certain cores
- Tune Linux TCP stack for maximum performance

# Computers

Readout Unit (RU)

- Dell PowerEdge R620
- Dual 8 core Xeon CPU E5-2670 0 @ 2.60GHz
- 32 GB of memory

Builder Unit (BU)

- Dell PowerEdge R720
- Dual 8 core Xeon CPU E5-2670 0 @ 2.60GHz
- 32+256GB of memory (240 GB for Ramdisk on CPU 1)

# Data Network

40/56Gb NICs (Infiniband or Ethernet)

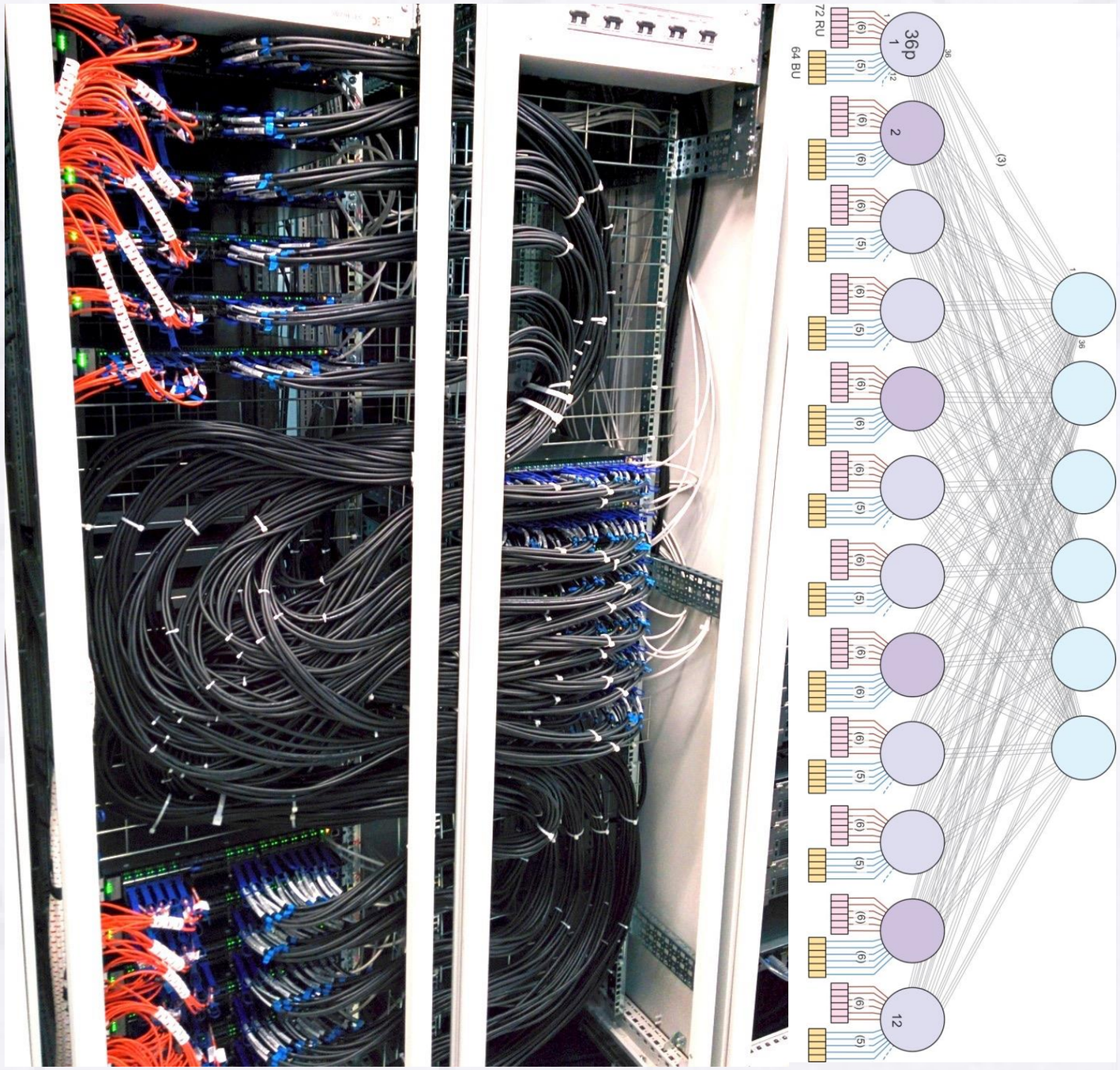
- Mellanox Technologies MT27500 Family [ConnectX-3]

10/40 GbE switches

- Mellanox SX1024 & SX1036

Infiniband switches

- Mellanox SX6036




Infiniband CLOS network

# Measurement Technique

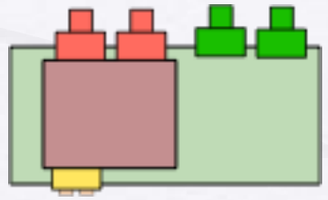
Semi-automatic scanning using python scripts (w/o run control)

- Generate fake data on the FEROL in free-running mode
  - EVM gets always 1024 Bytes fragments & is not counted as RU
- Full event building, but events not written to disk unless noted
- 20 measurements every 5s done at each point after waiting >60s

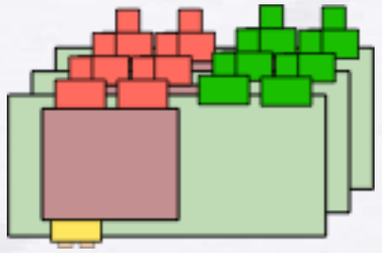
Measurement schemes

- Data-concentrator measurement
  - Vary number of FEDs (TCP streams) sent to 1 RU
- Canonical setup uses 8 FEDs (TCP streams) per RU
  - All streams use the same fragment size
  - Scan different fragment sizes
  - Measure various sizes of the event-building system
- Real FED builder setup with different number of FEDs
  - 1-18 FEDs per RU
  - Set fragment sizes to roughly expected size for each FED
  - Scale fragment sizes linearly for different event sizes

# Data Concentrator



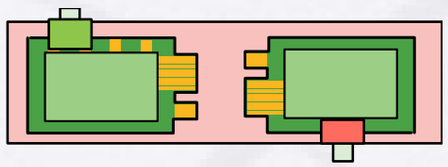
1 kB



256B - 16kB

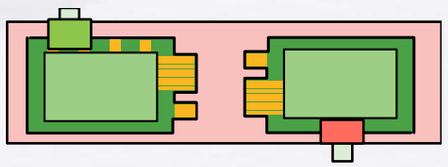
4 – 24 Streams
(1 stream / FEROL)

EVM



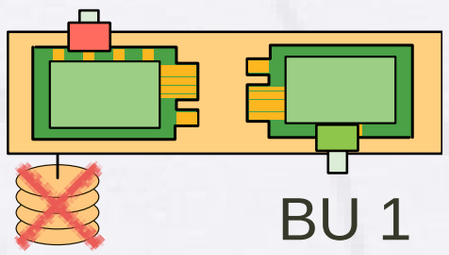
1 kB

RU



1 – 256 kB

1 RU

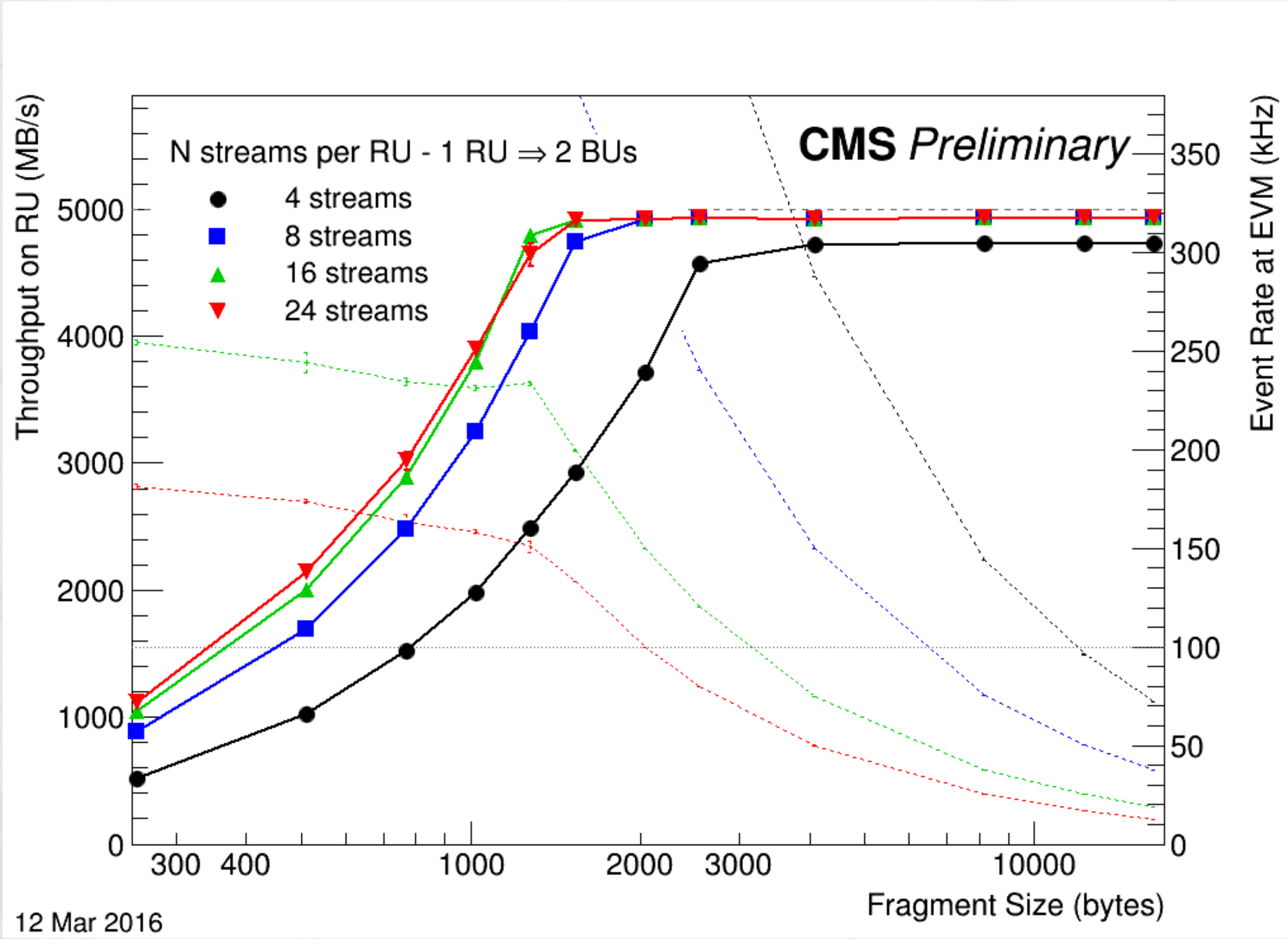


BU 1

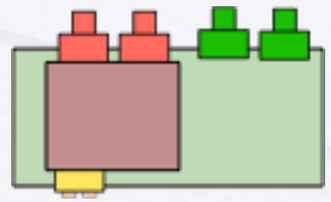
BU 2

2 BUs

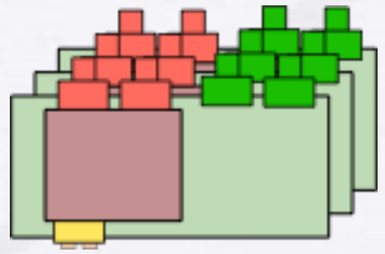
# Data Concentrator



# Scalability of the Event Builder



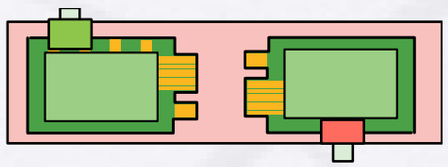
1 kB



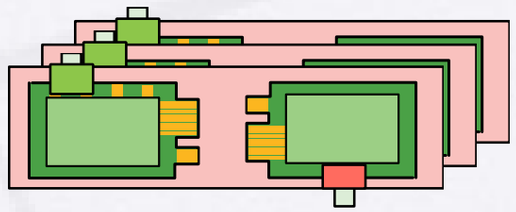
8 Streams per RU
(1 or 2 streams / FEROL)

256B - 16kB

EVM



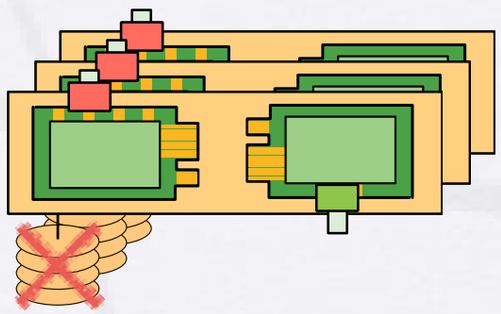
1 kB



1 – 101 RUs

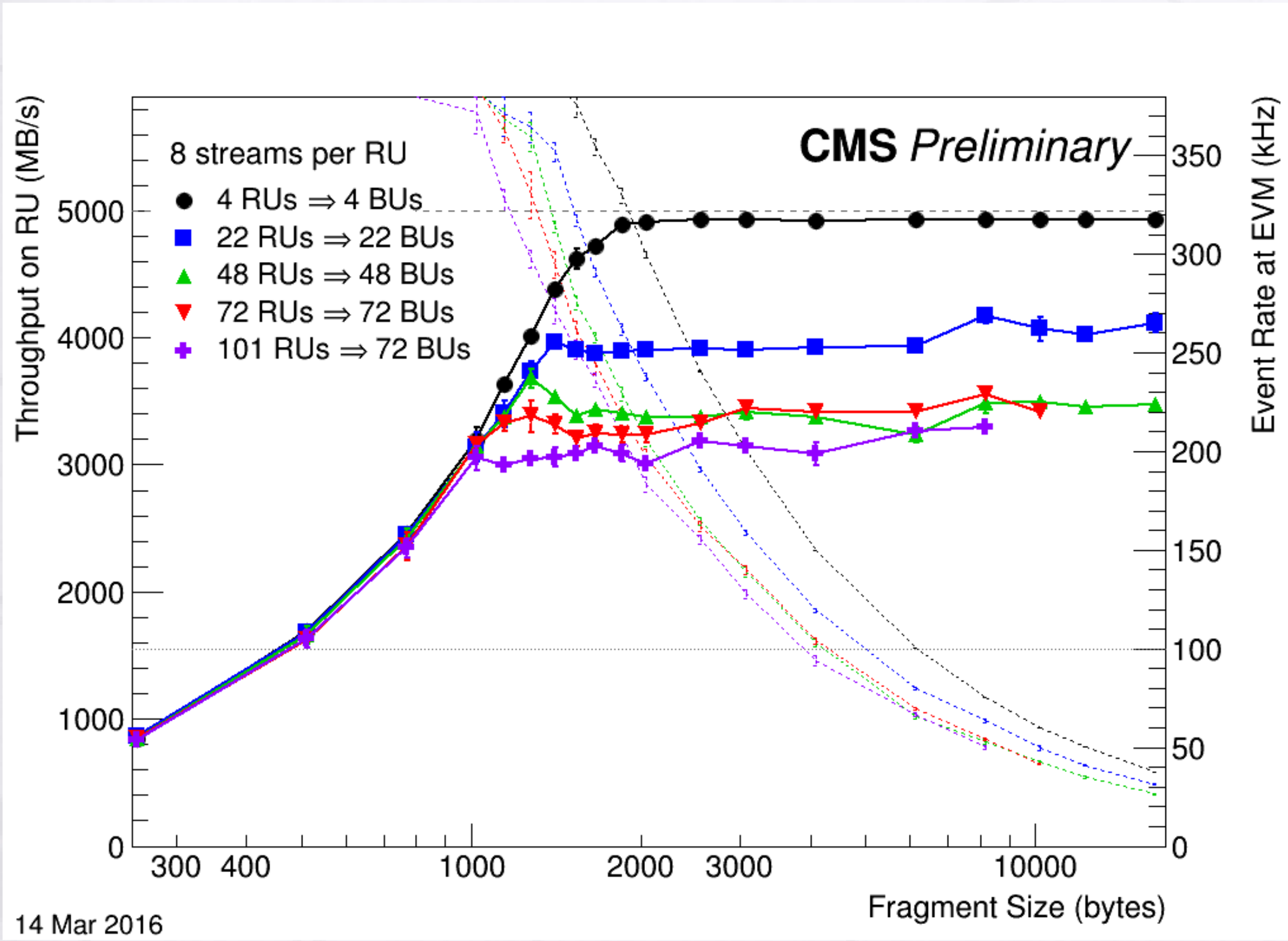
2 - 128 kB

3 kB – 8 MB

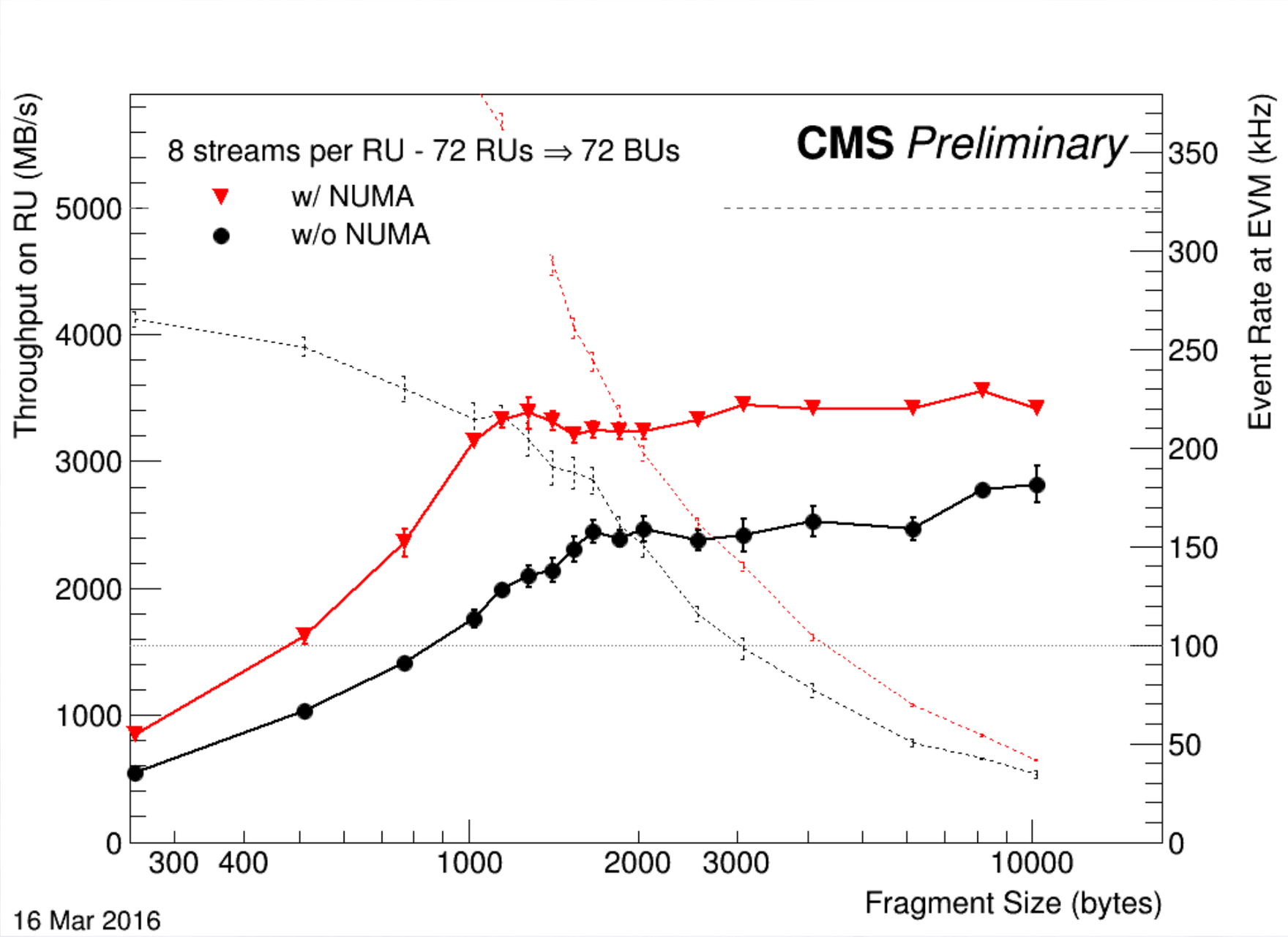


1 – 72 BUs

# Scalability of the Event Builder

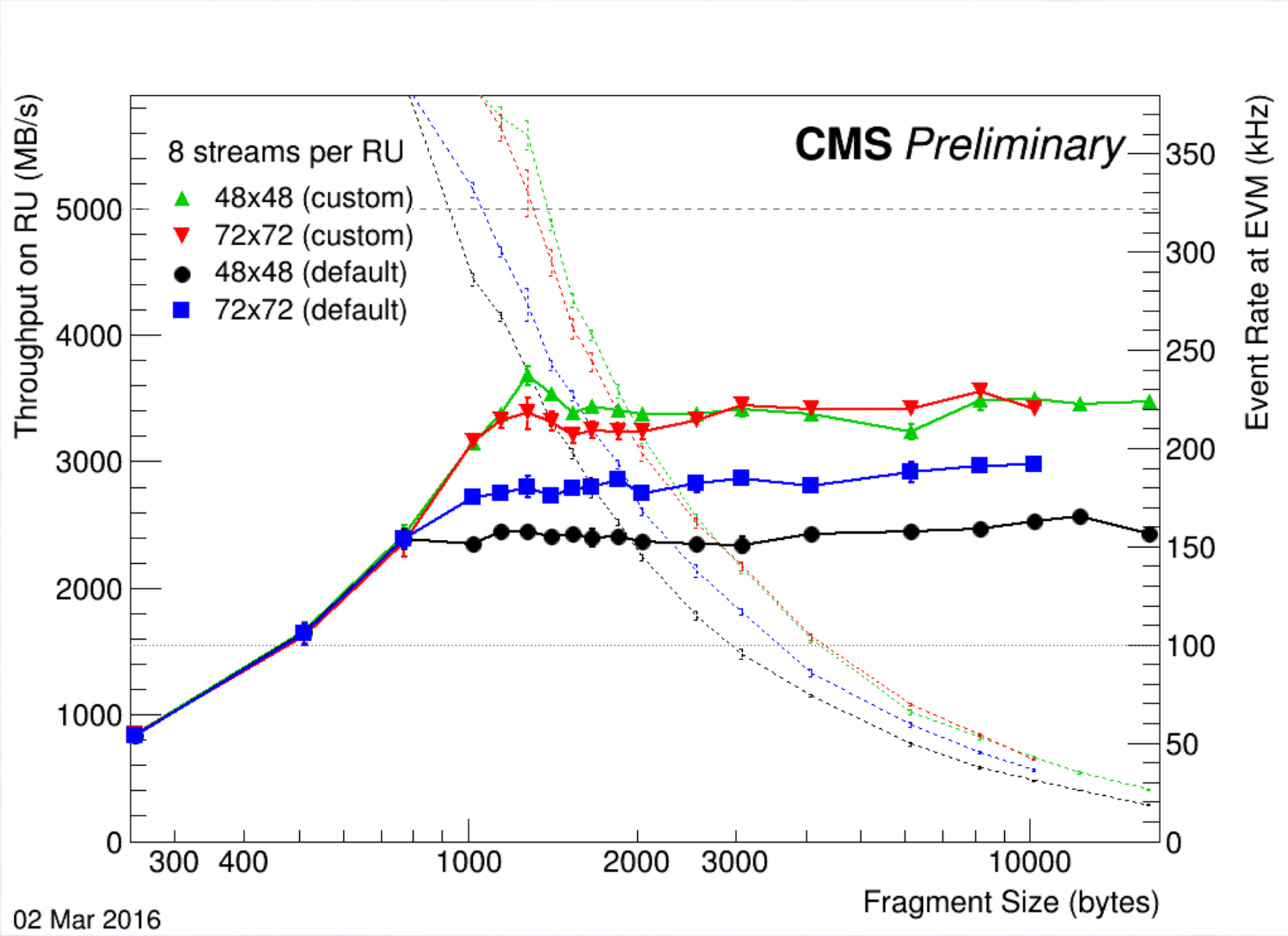


# Effect of NUMA Settings



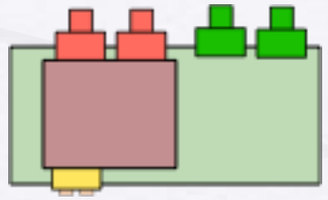

- '/usr/bin/numactl —physcpubind=10,12,14,26,28,30 —membind=1' used to start executives
- Threads and memory structures are bound to cores/memory using XDAQ policies

# Custom IB Routing

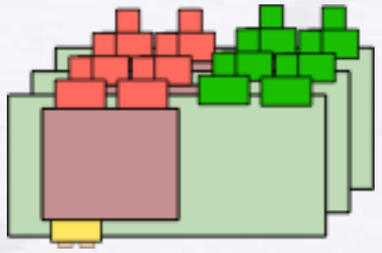



- Optimize the routing to prefer traffic from RU to BU
- See next talk from Andre about IB routing

# BU Scalability



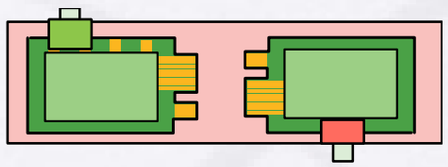
1 kB



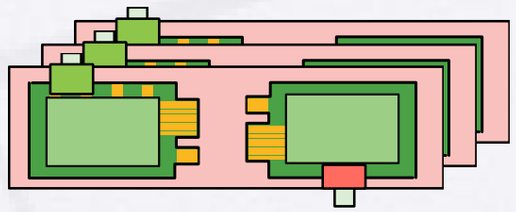
8 Streams per RU
(1 or 2 streams / FEROL)

256B - 16kB

EVM

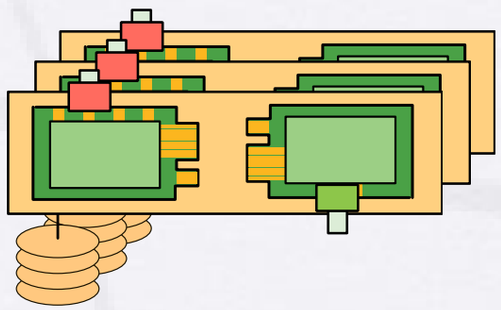


1 kB



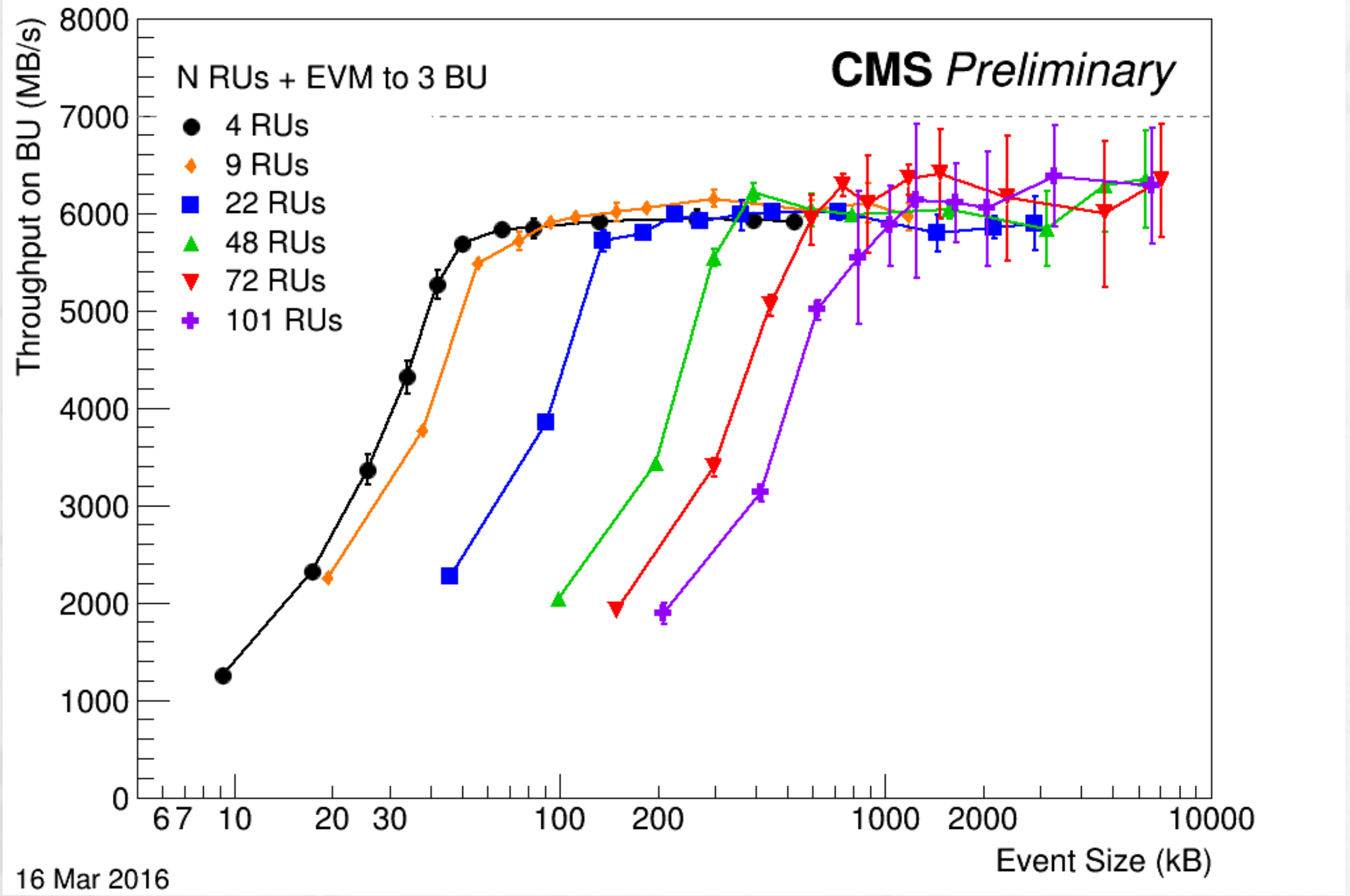
1 – 101 RUs

2 - 128 kB

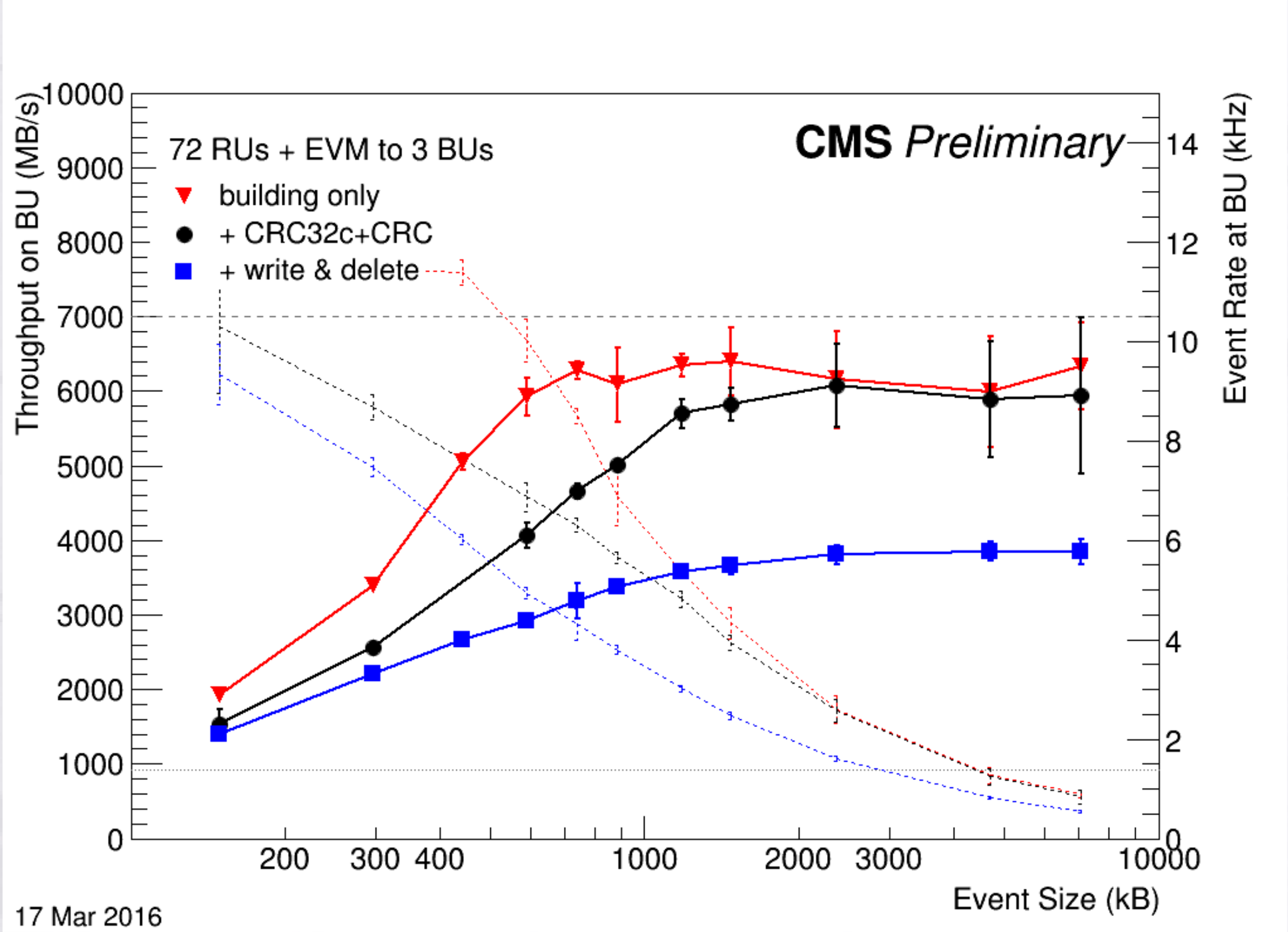


3 BUs

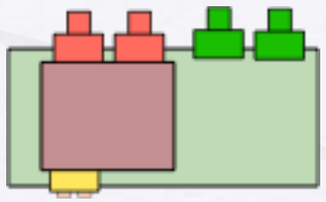
# BU Scaling vs Number of RUs



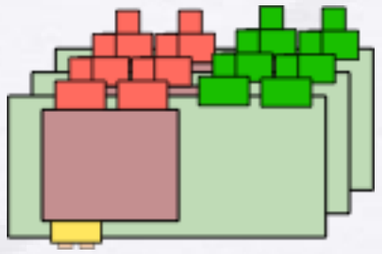
# BU Performance



# Production FED Builders



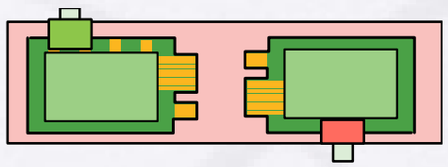
1 kB



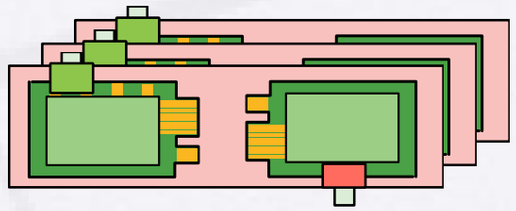
2-18 Streams per RU
(1 or 2 streams / FEROL)

4B – 12kB

EVM



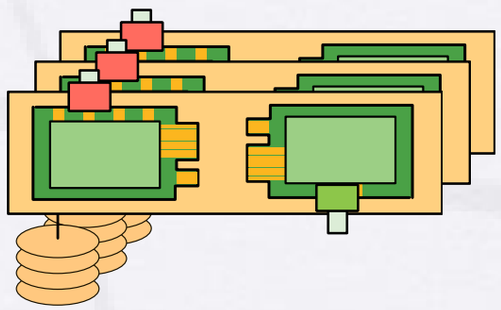
1 kB



60 or 82 RUs

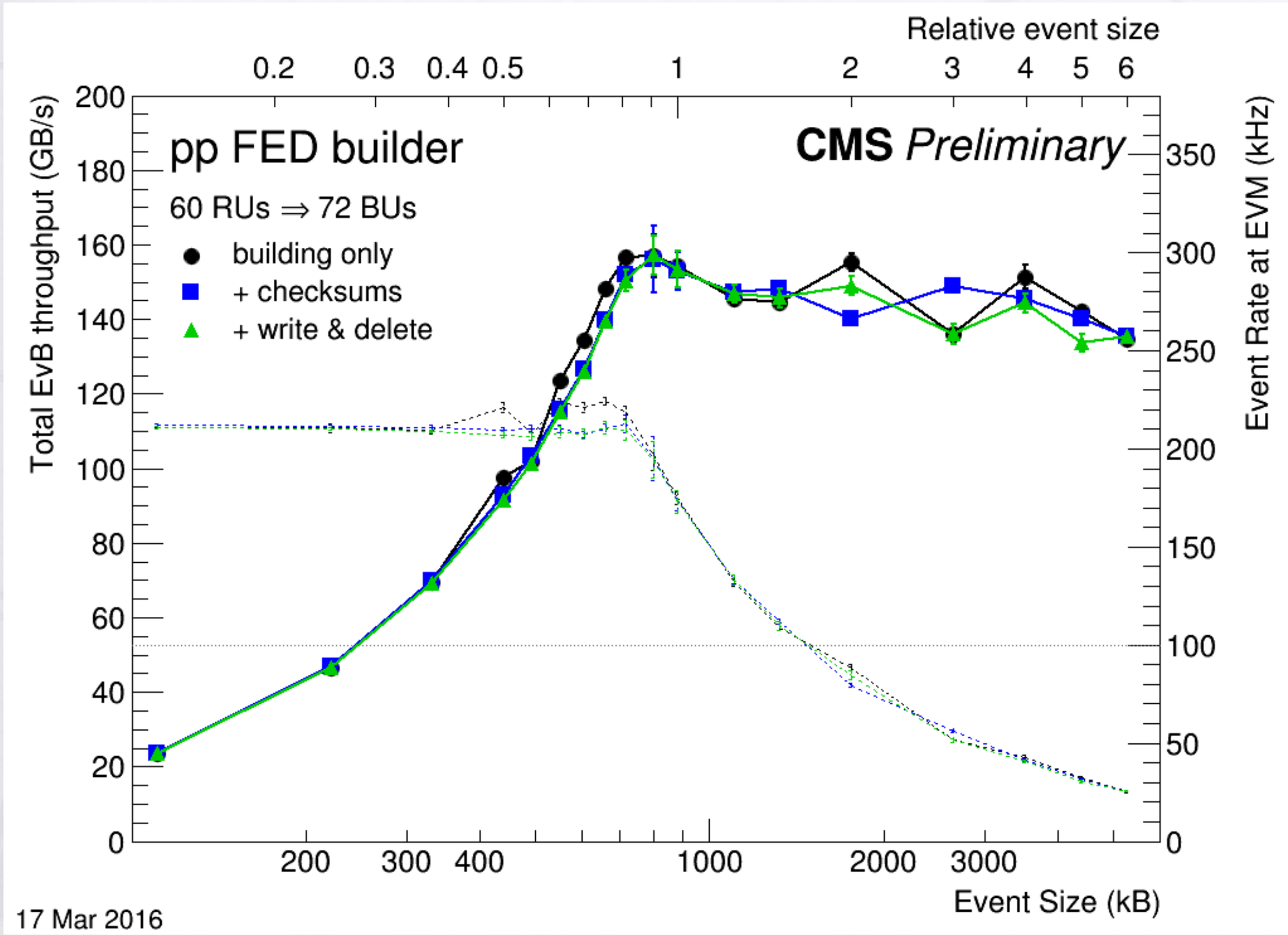
1 - 118kB

100kB – 5MB

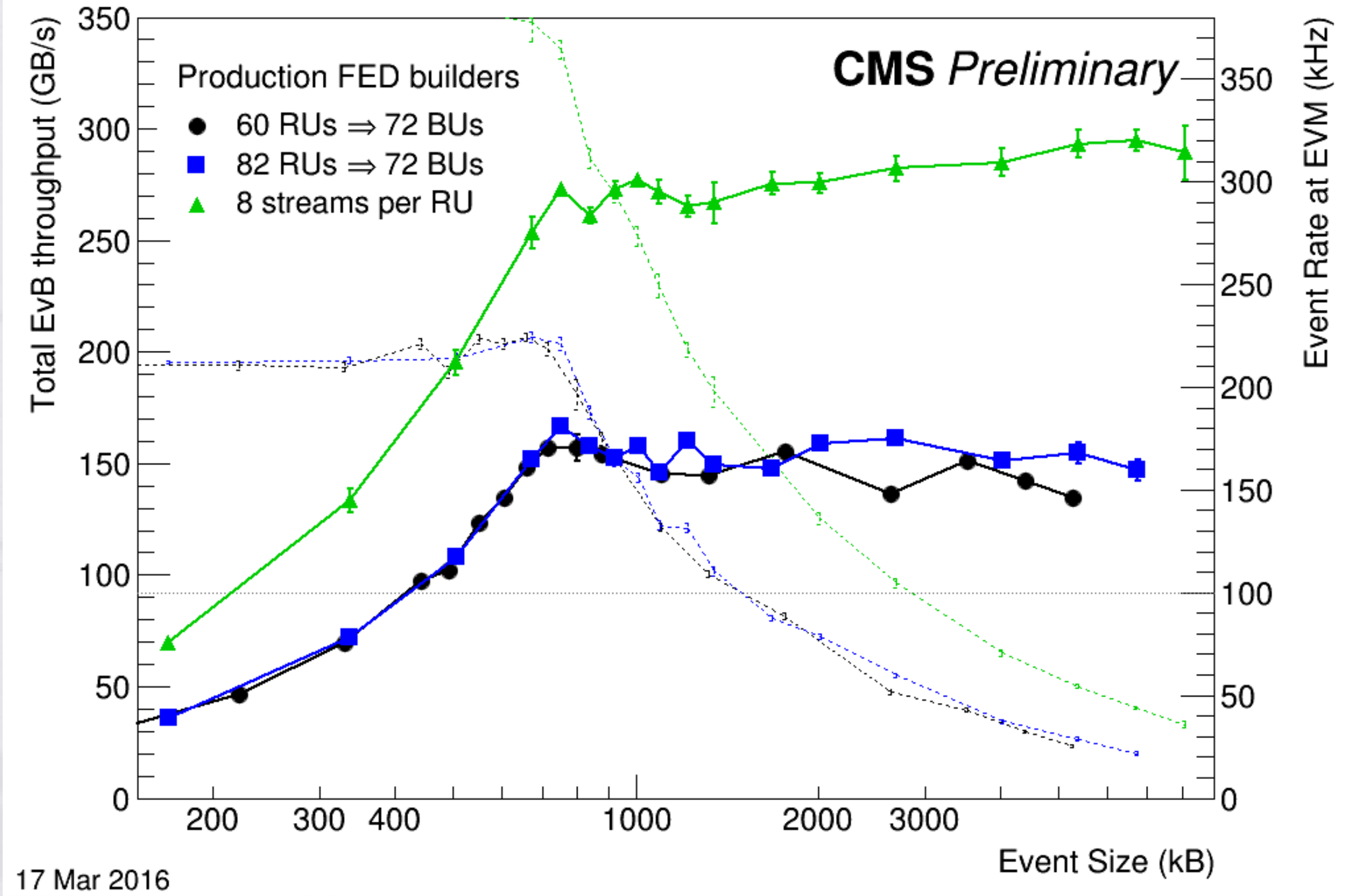


72 BUs

# Standard pp FED Builder



# Prod. FED Builder vs Canonical



# Summary

CMS has a complete new event-builing system for LHC run 2

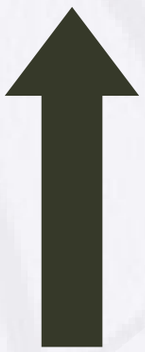
- State-of-the-art technology
- Order of magnitude smaller than run-1 DAQ system

Optimal use of high-end hardware

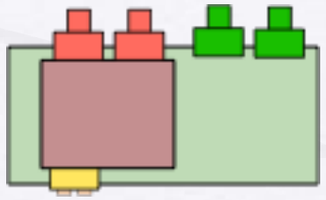
- New event-building protocol
- New software to exploit hardware capabilities
- A lot of fine-tuning to get full performance

Infiniband performance sensible to traffic pattern

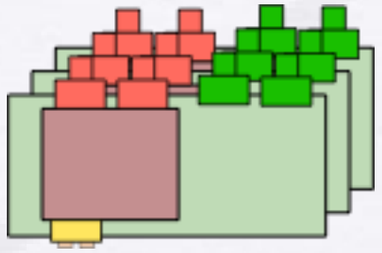
- Requires custom routing
- Performance diminishes the more uneven the system becomes



# RU Scalability



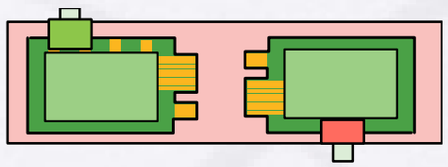
1 kB



256B - 16kB

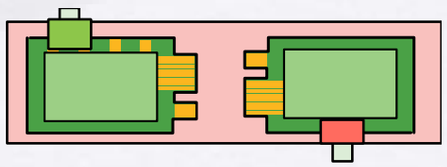
8 FEROLs per RU

EVM



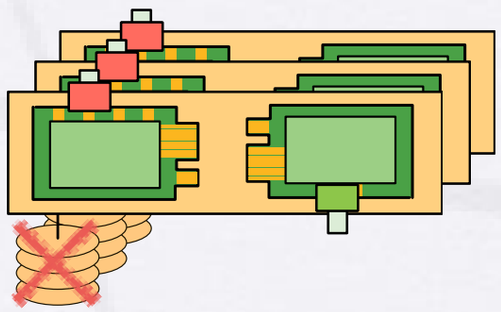
1 kB

RU



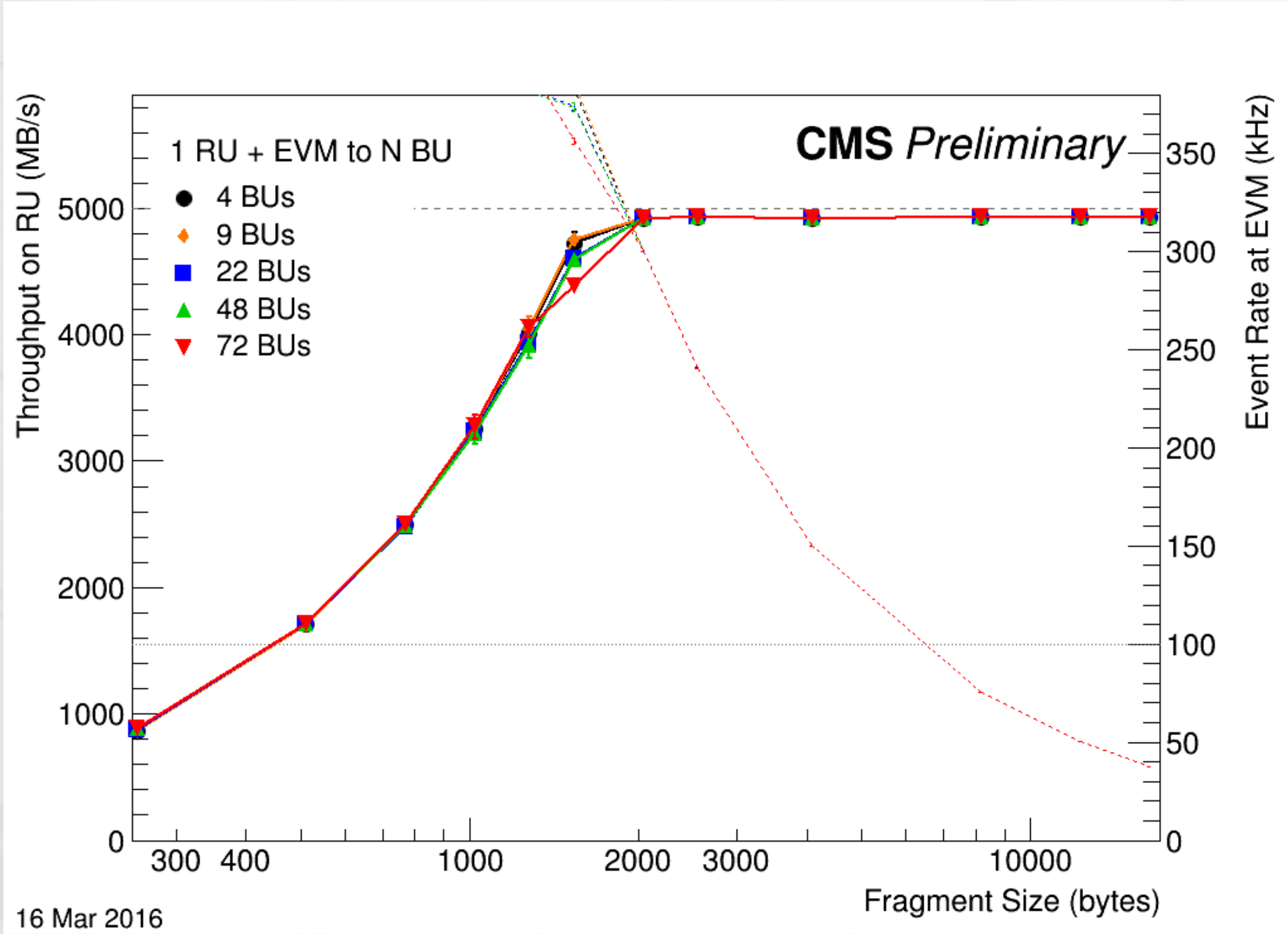
1 – 256 kB

1 RU

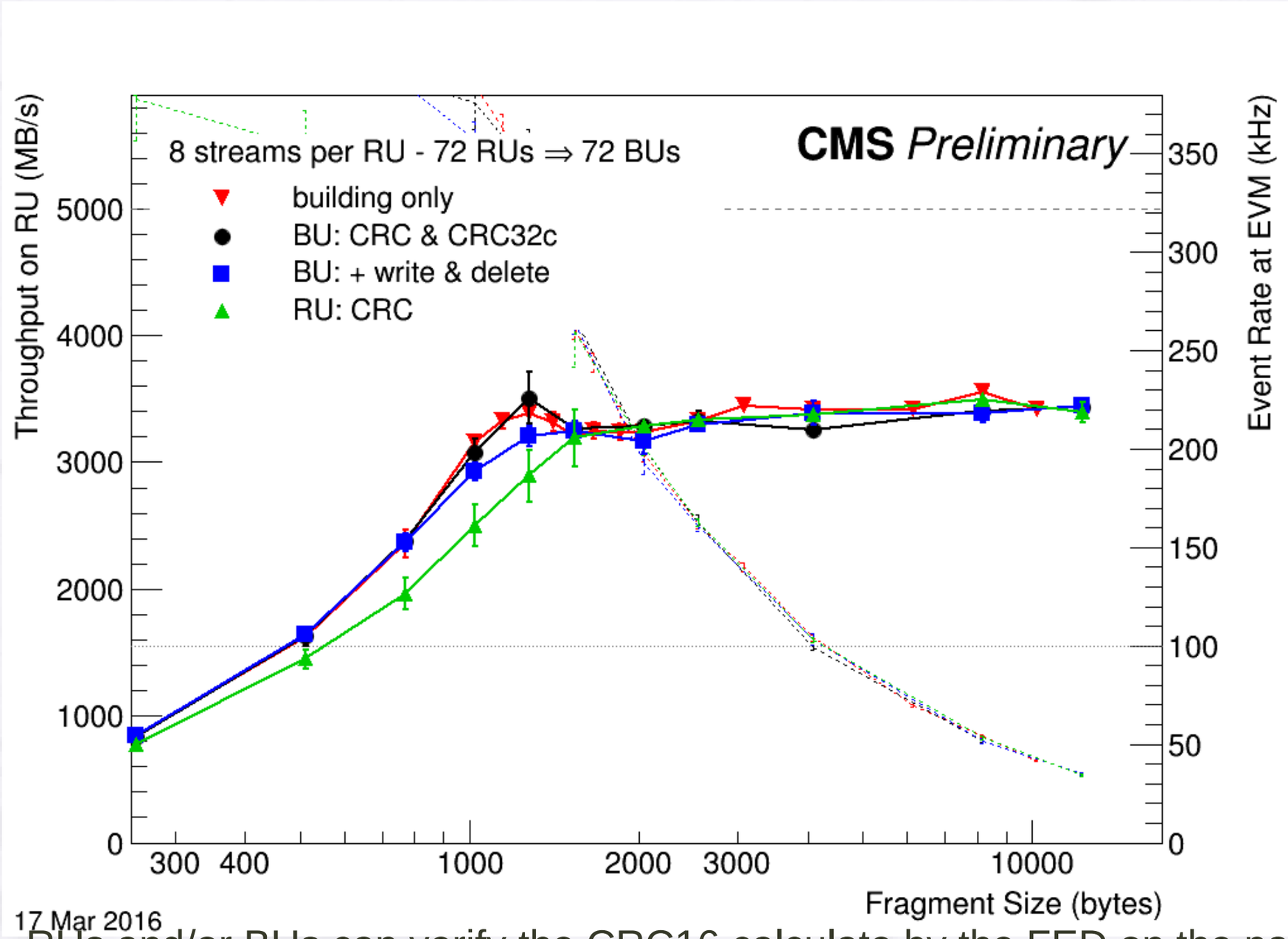


1 – 72 BUs

# RU Scaling vs Number of BUs



# Checksums




- RUs and/or BUs can verify the CRC16 calculate by the FED on the payload
- BUs calculate a CRC32c on the complete event which is rechecked by the HLT

# Number of Builder Threads