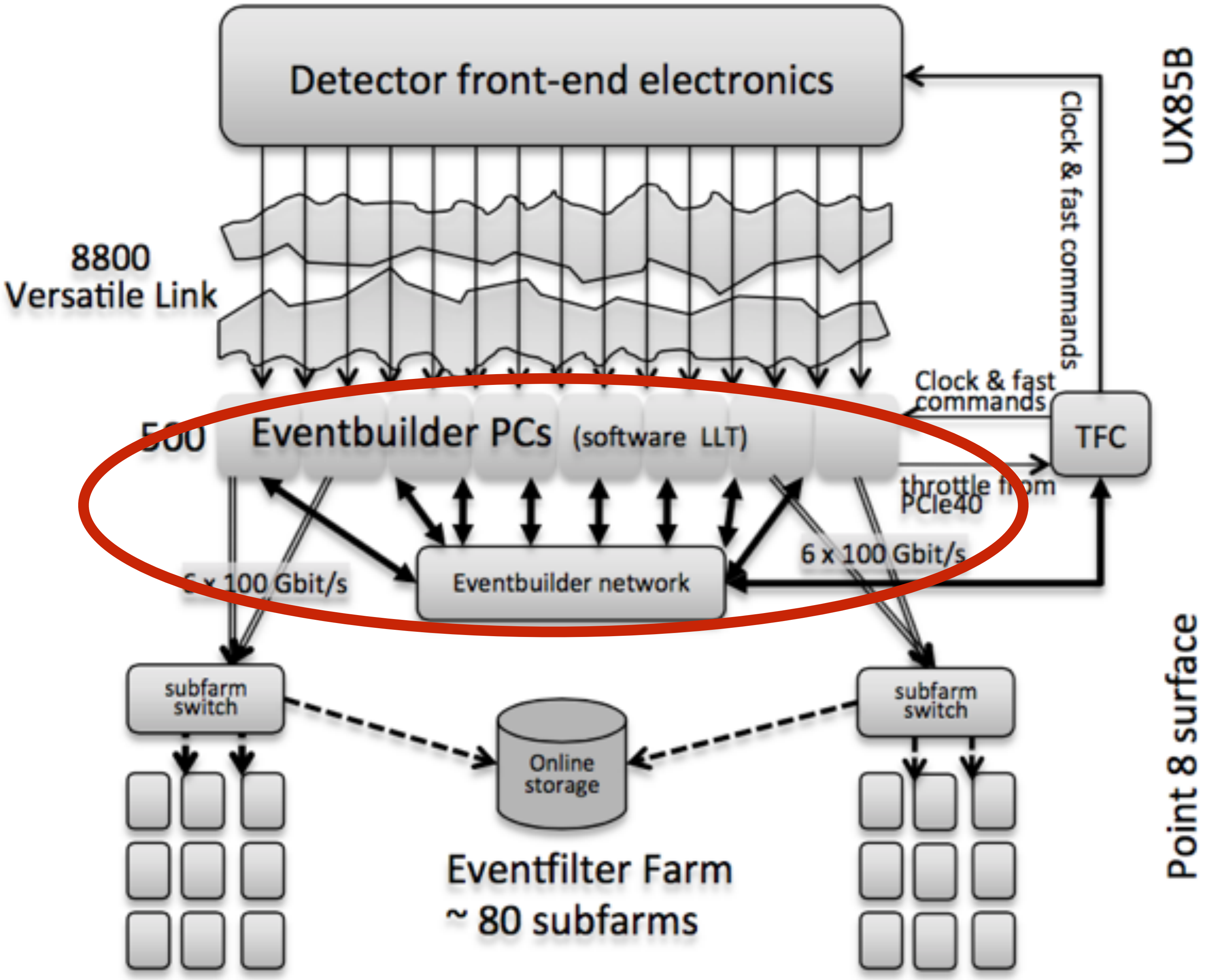
Status of online upgrade

Matteo Manzali [2,3], Antonio Falabella [2], Domenico Galli [1], Francesco Giacomini [2], Umberto Marconi [1]

INFN Bologna - INFN CNAF - University of Ferrara

The readout-system



The Event Building

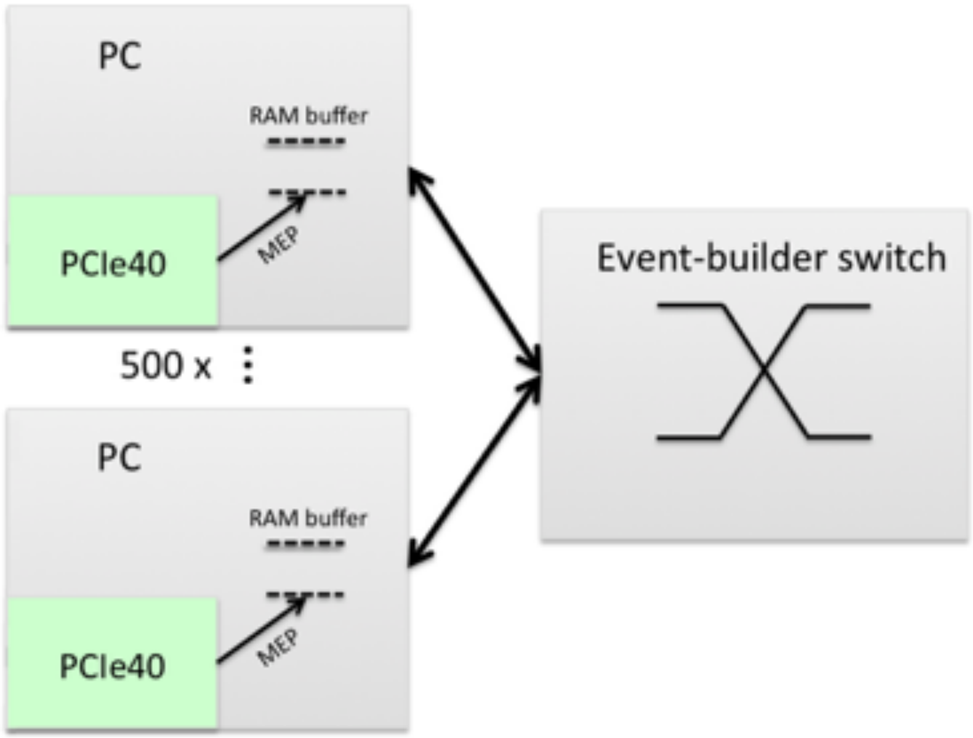
- It’s composed of ~500 nodes interconnected together with a dedicated network (the Event-builder switch)

- PCIe40s write directly in RAM memory the incoming data from the detector

- Each node sends fragments of a specific event to an elected node (~100Gb/s)

- Each elected node gathers fragments together to generate a full event and and sends it to the filter farm (~100Gb/s)

- Each node sends and receives data with an aggregate bandwidth of ~200Gb/s



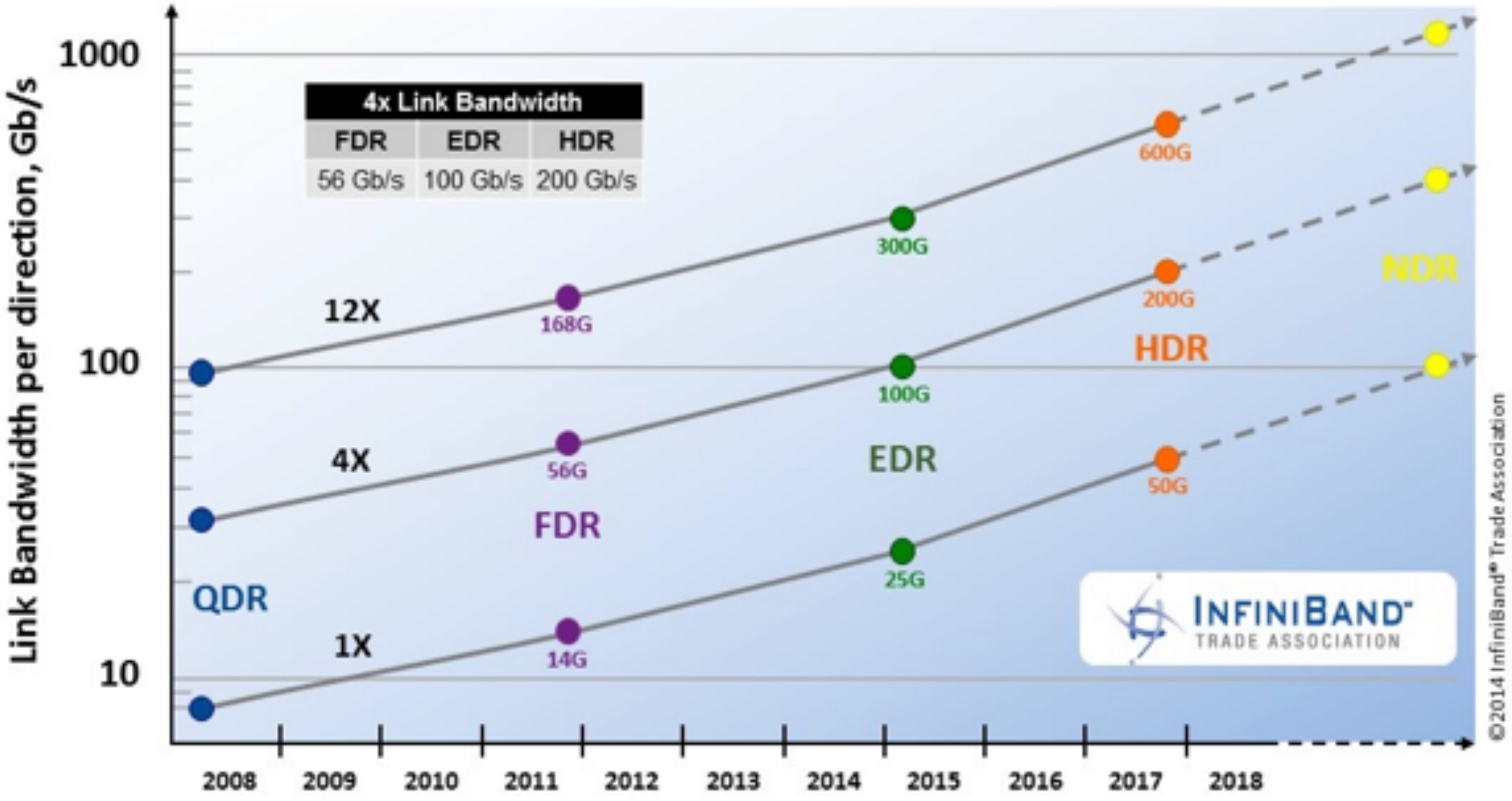
- In order to reach this performances a high-speed network is required

The InfiniBand standard (1)

- A computer-networking communications standard that feature very high bandwidth and low latency

- Low CPU utilization with RDMA (Remote Direct Memory Access) - Unlike standard Ethernet, communication bypass the OS and the CPUs

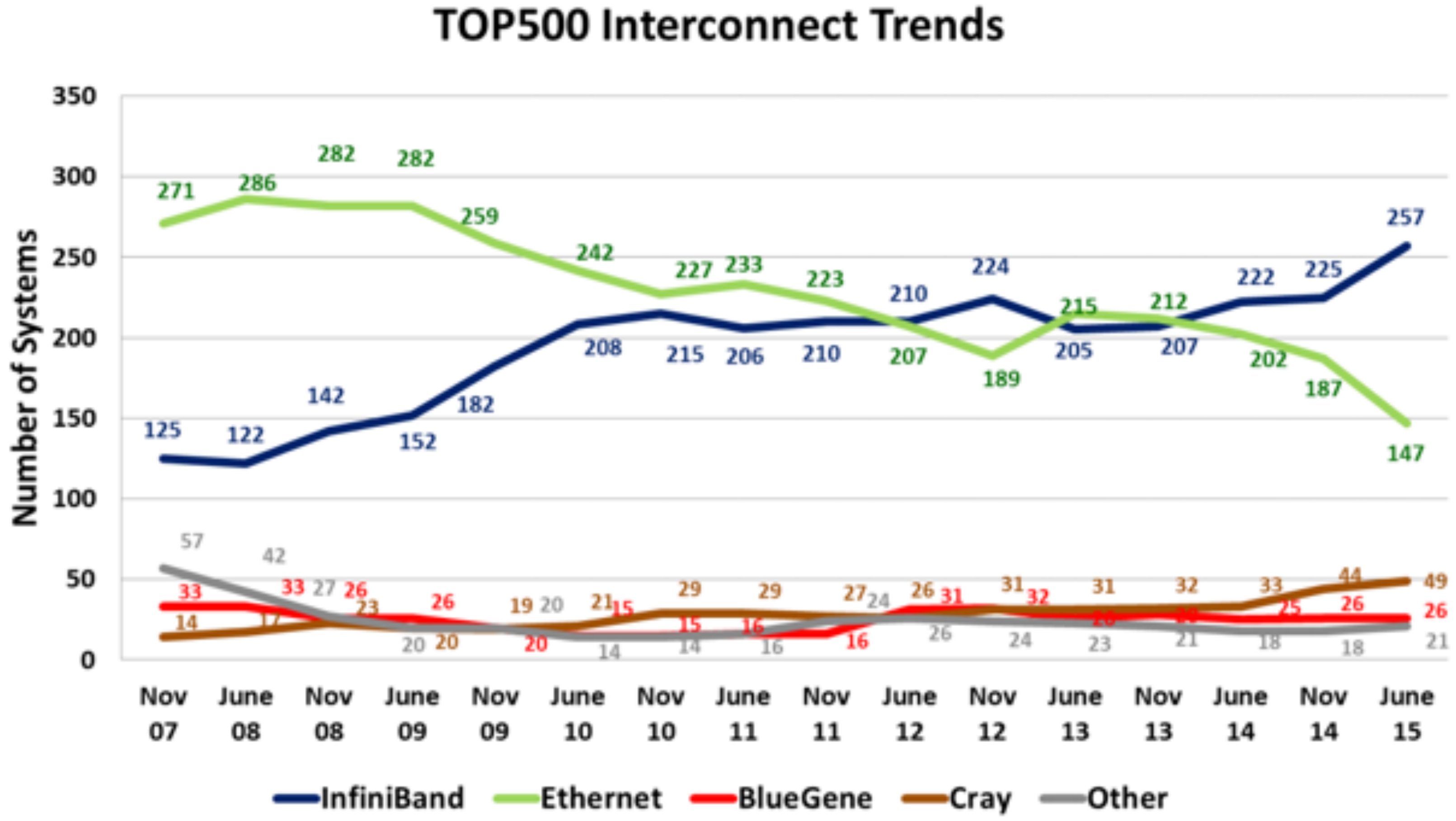
- Constant speed evolution and cost-effective technology



©2014 InfiniBand® Trade Association

The InfiniBand standard (2)

- InfiniBand is widely used in High Performance Computing:



- However Ethernet is also a potential candidate (100Gb/s is already available)
  - Possibility to perform RDMA operations (with RoCE or iWARP)

What are the “verbs”?

- Verbs are a low level description for RDMA programming and provide best performance
- Any other level of abstraction over verbs may harm the performance
- Same API for all RDMA-enabled transport protocols:
  - InfiniBand
  - RDMA Over Converged Ethernet (RoCE)
  - Internet Wide Area RDMA Protocol (iWARP)

RoCE™

Chelsio
Communications
Accelerate

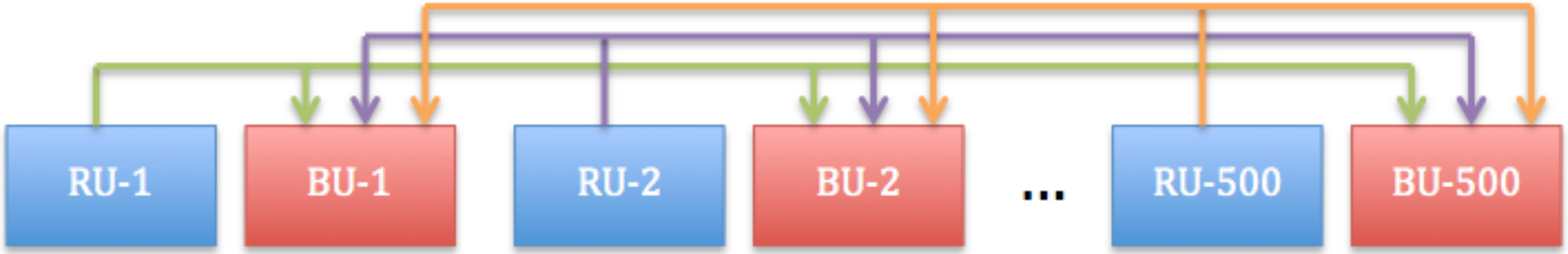
The Large Scale Event Builder

- The Large Scale Event Builder (LSEB) simulates the event building design in a realistic way in order to investigate the possibility to use an InfiniBand based network to perform the event building.

- It is composed of two distinct logical components, the Readout Unit (RU) and the Builder Unit (BU):

    - Each RU receives data from a generator that simulate the detector, creates the event fragments and ship them to receiving BU in a many-to-one pattern.

    - Each BU gathers event fragments together to generate full events.



The Galileo cluster

- Galileo is a new cluster of the CINECA consortium devoted to scientific computing and special HPC oriented projects:

    - Nodes: 516
    - Processors: 2 8-cores Intel Haswell 2.40 GHz per node
    - Cores: 16 cores/node, 8256 cores in total
    - RAM: 128 GB/node, 8 GB/core
    - Internal Network: Infiniband with 4x QDR switches

- Due to the policy of the cluster is not possible to tune the InfiniBand drivers and switch off the CPU power management in order to run always with the max CPU frequency.

- If power management is active the performances of low latency network technologies usually are affected.

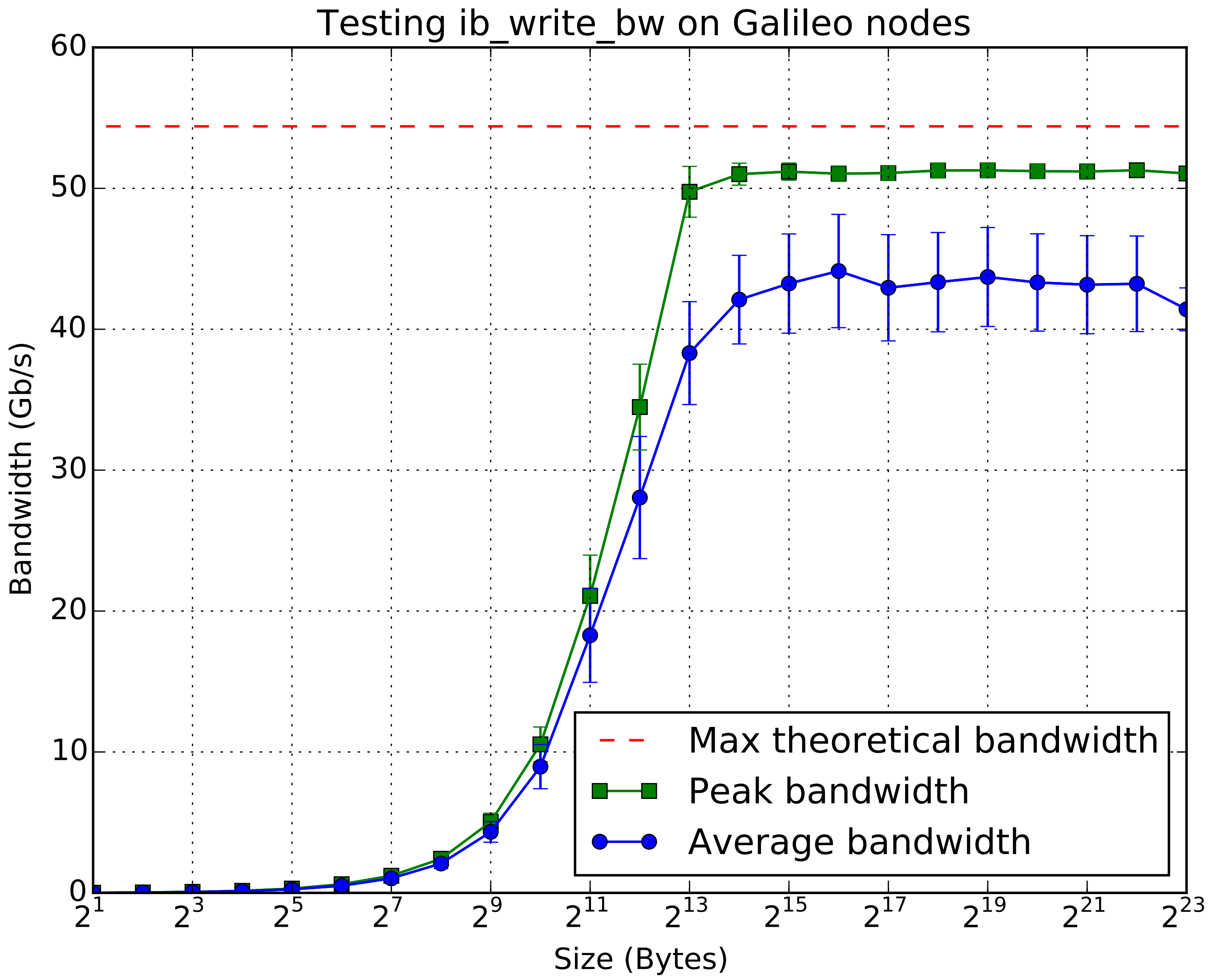
Raw bandwidth on Galileo

- This plot shows the gap between peak and average bandwidths reached running the ib_write_bw tool on two nodes of the Galileo cluster (these results are the average of 10 tests).

- The green line represents the highest bw (on 5K iterations) reached for each size.

- The blu line represents the average bw (on 5K iterations) for each size.

- As can be seen it's not easy to reach and keep the max bw, also the average bw is really different from test to test.



LSEB on Galileo

- We ran tests with 4, 8, 16, 32, 64 and 128 nodes

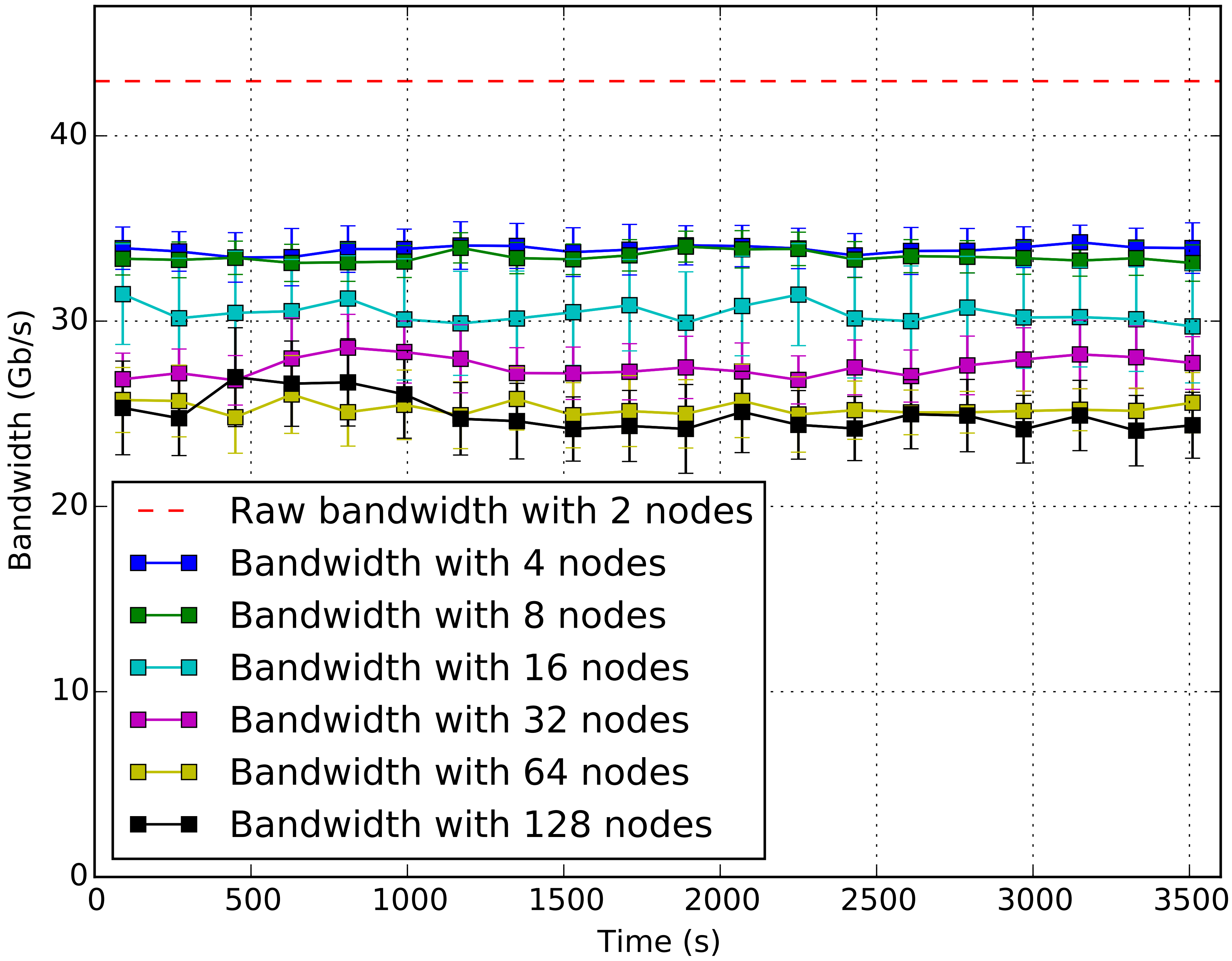
- For the 4-, 8-, 16- and 32-node tests we managed to allocate the entire node (16 cores)

- For the 64- and 128-node tests we allocated half node (8 cores)
    - Otherwise the waiting time would have been too long

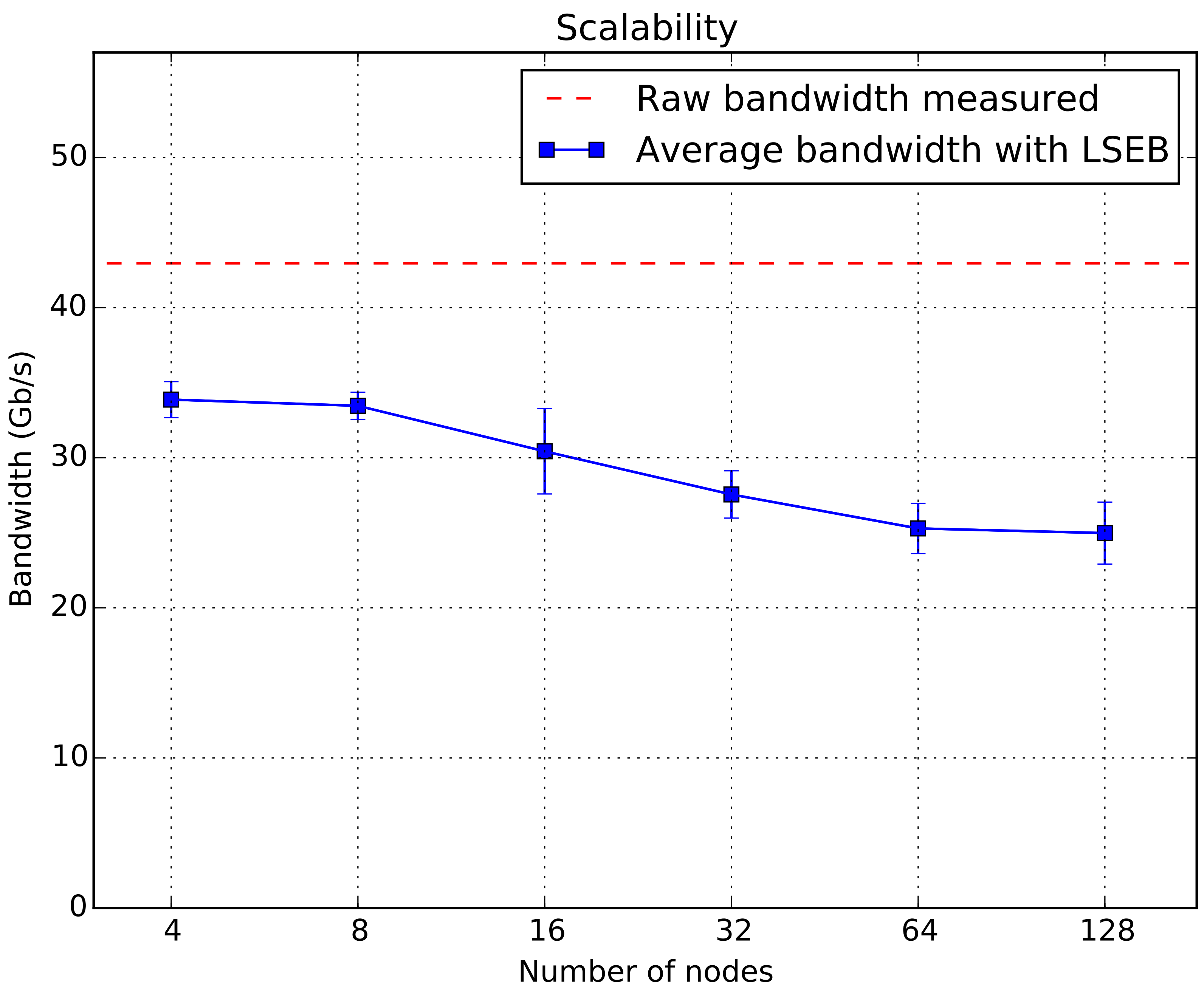
- About the size of multi-fragments:
    - The size of each fragment is ~220 bytes
      (24 bytes of Event Header + ~200 of payload)
    - The number of fragments in a multi-fragment is 600
    - The total size of a multi-fragment is ~128KB

- The bandwidth measurements that will be shown take into account only InfiniBand traffic (not the local communication between RU and BU on the same node)

Large-scale tests

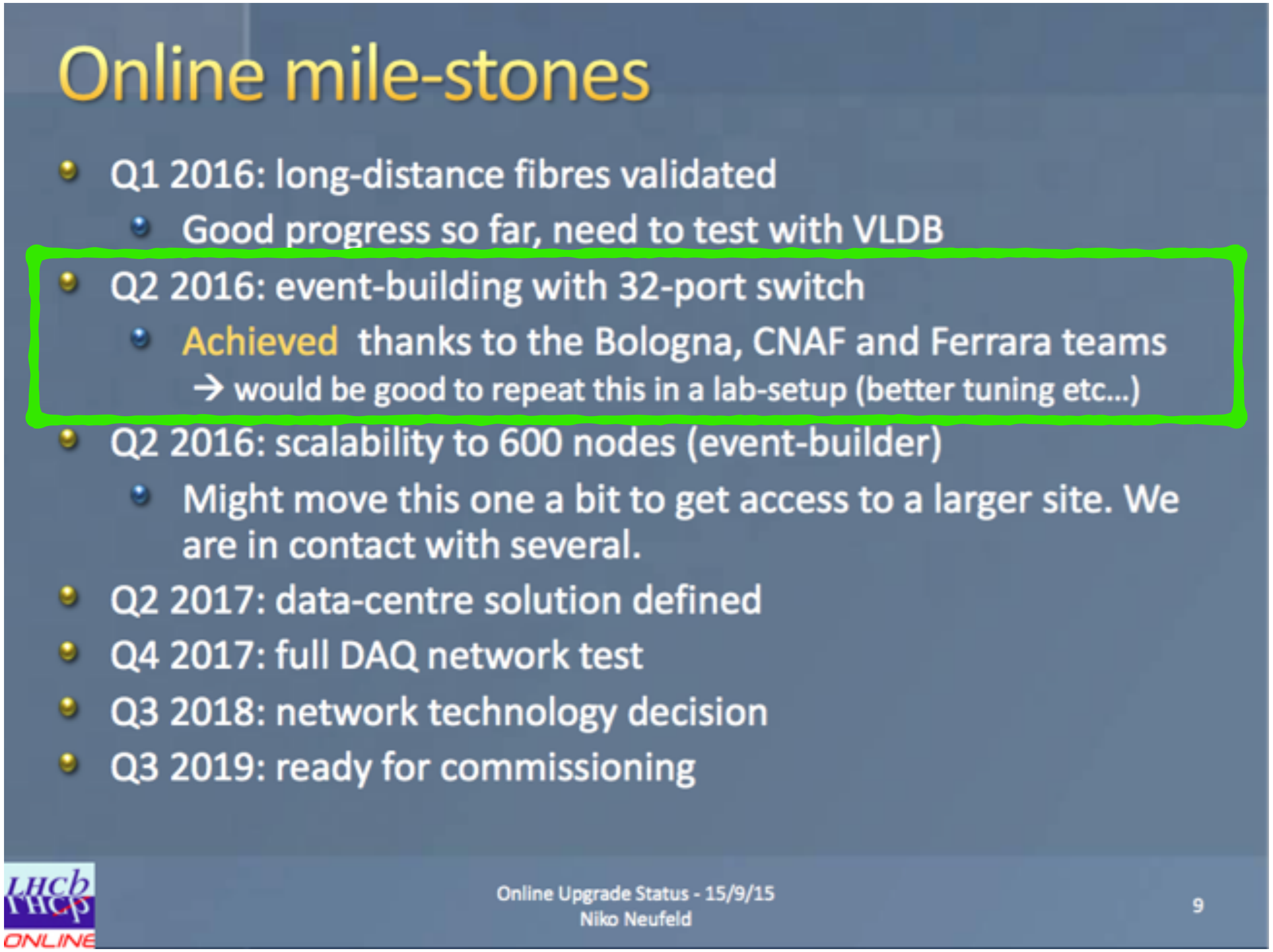


Scalability



First milestone achieved

- From the talk of Niko Neufeld on the online upgrade status at the LHCb week



Online mile-stones

- Q1 2016: long-distance fibres validated
  - Good progress so far, need to test with VLDB
- Q2 2016: event-building with 32-port switch
  - Achieved thanks to the Bologna, CNAF and Ferrara teams
    → would be good to repeat this in a lab-setup (better tuning etc...)
- Q2 2016: scalability to 600 nodes (event-builder)
  - Might move this one a bit to get access to a larger site. We are in contact with several.
- Q2 2017: data-centre solution defined
- Q4 2017: full DAQ network test
- Q3 2018: network technology decision
- Q3 2019: ready for commissioning

Online Upgrade Status - 15/9/15
Niko Neufeld

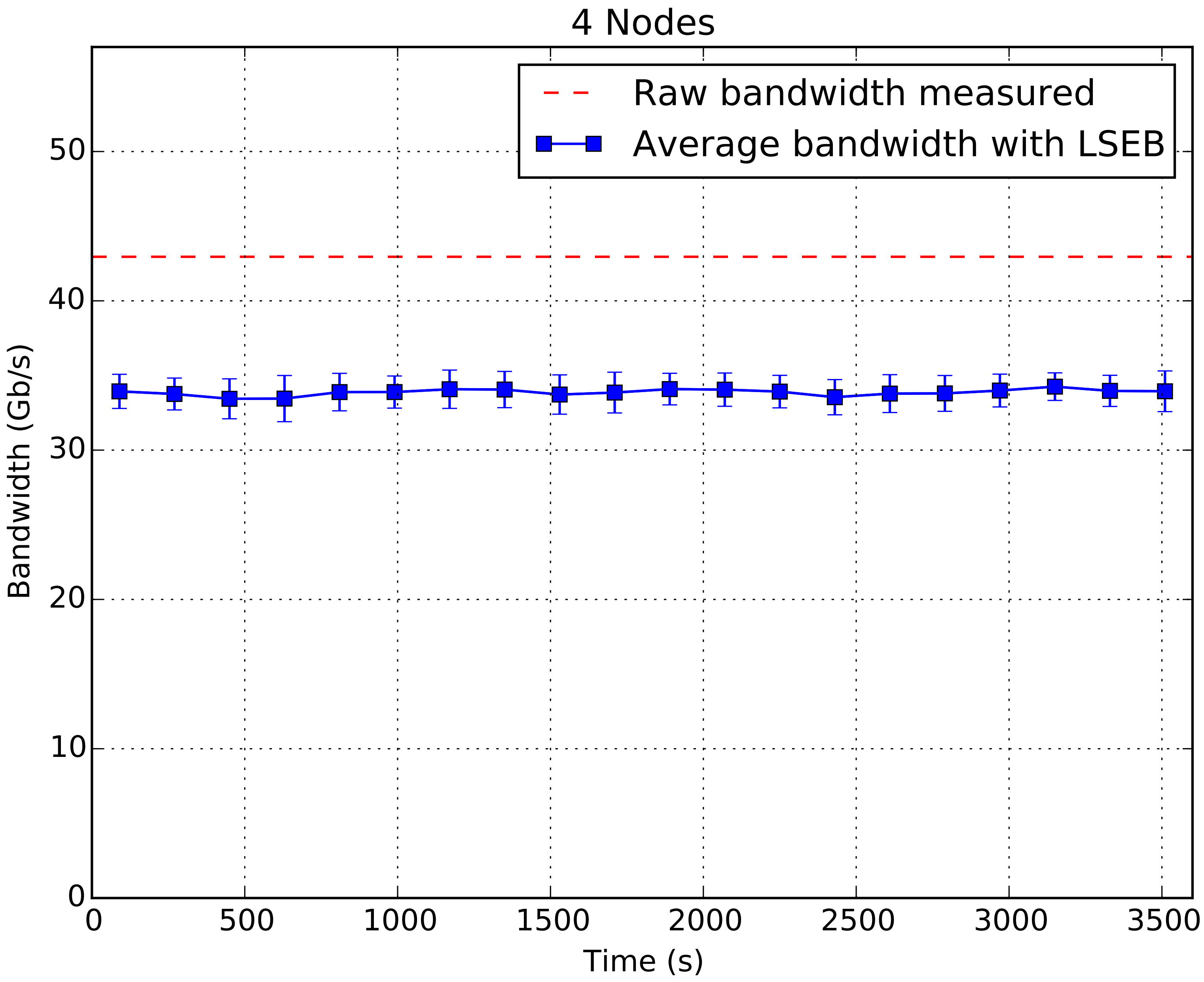
9

Conclusions and future works

- The software is stable over time and also it scales relatively well up to 128 nodes, reaching an overall aggregate traffic of 3,2 Tb/s.
- A proper tuning of the system seems to be needed to reach the max theoretical bandwidth.
- Further in development to guarantee fault tolerance and processes monitoring.
- Collaboration with LHCb online group ongoing.

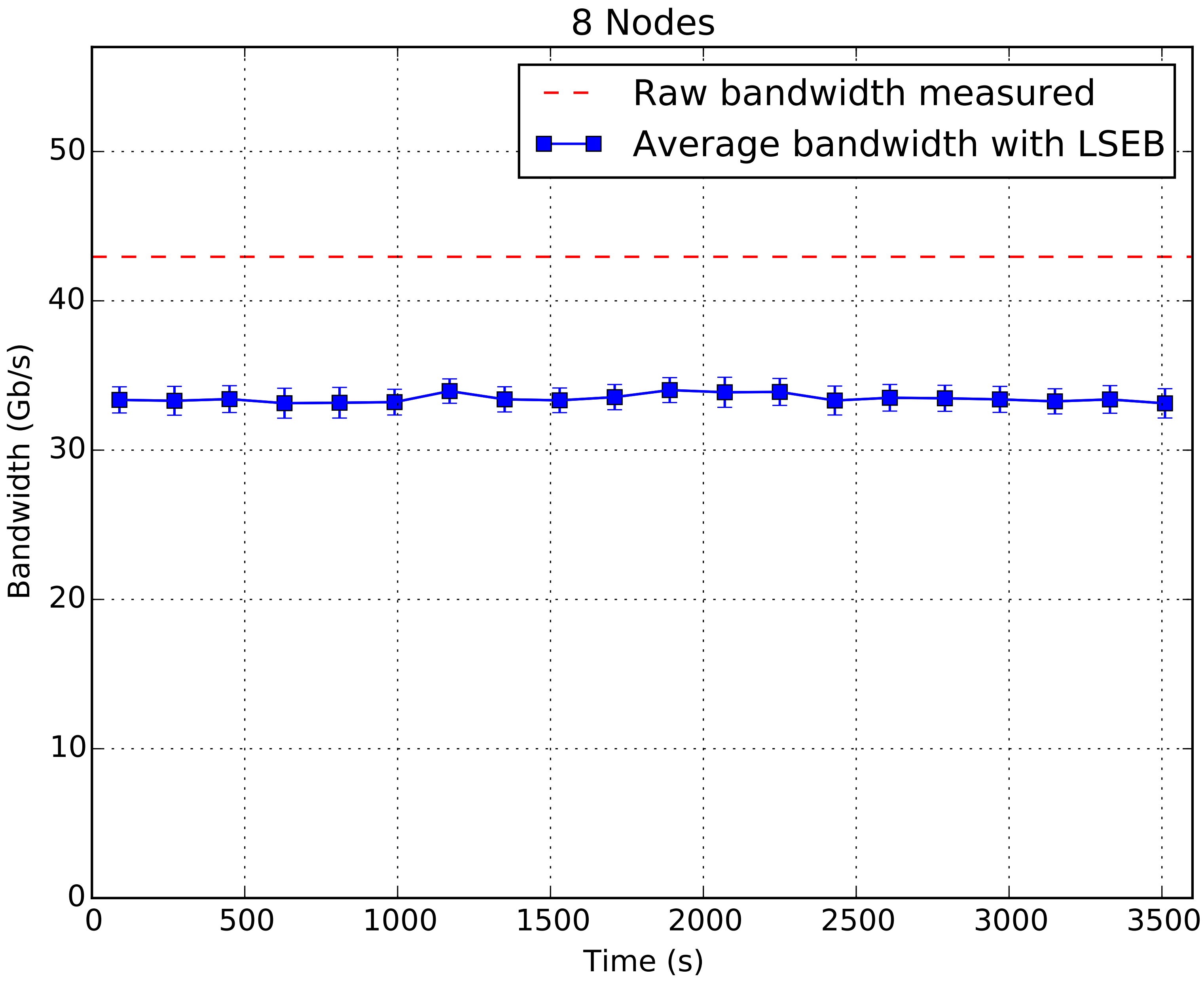
BACKUP

Run with 4 nodes



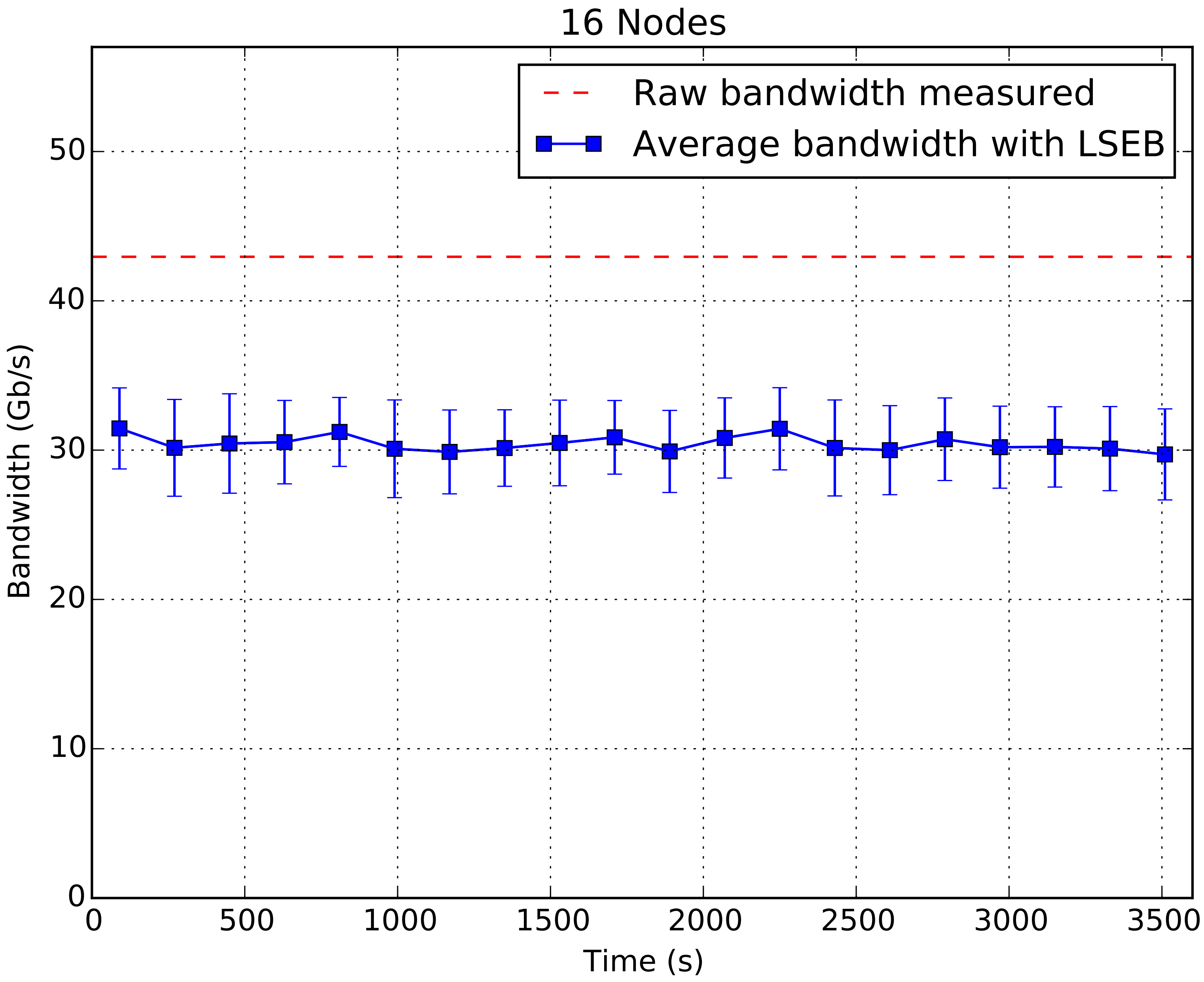
- Cores allocated: 16/16
- Size of buffer: ~128KB
- Average bw: 33.87 Gb/s

Run with 8 nodes



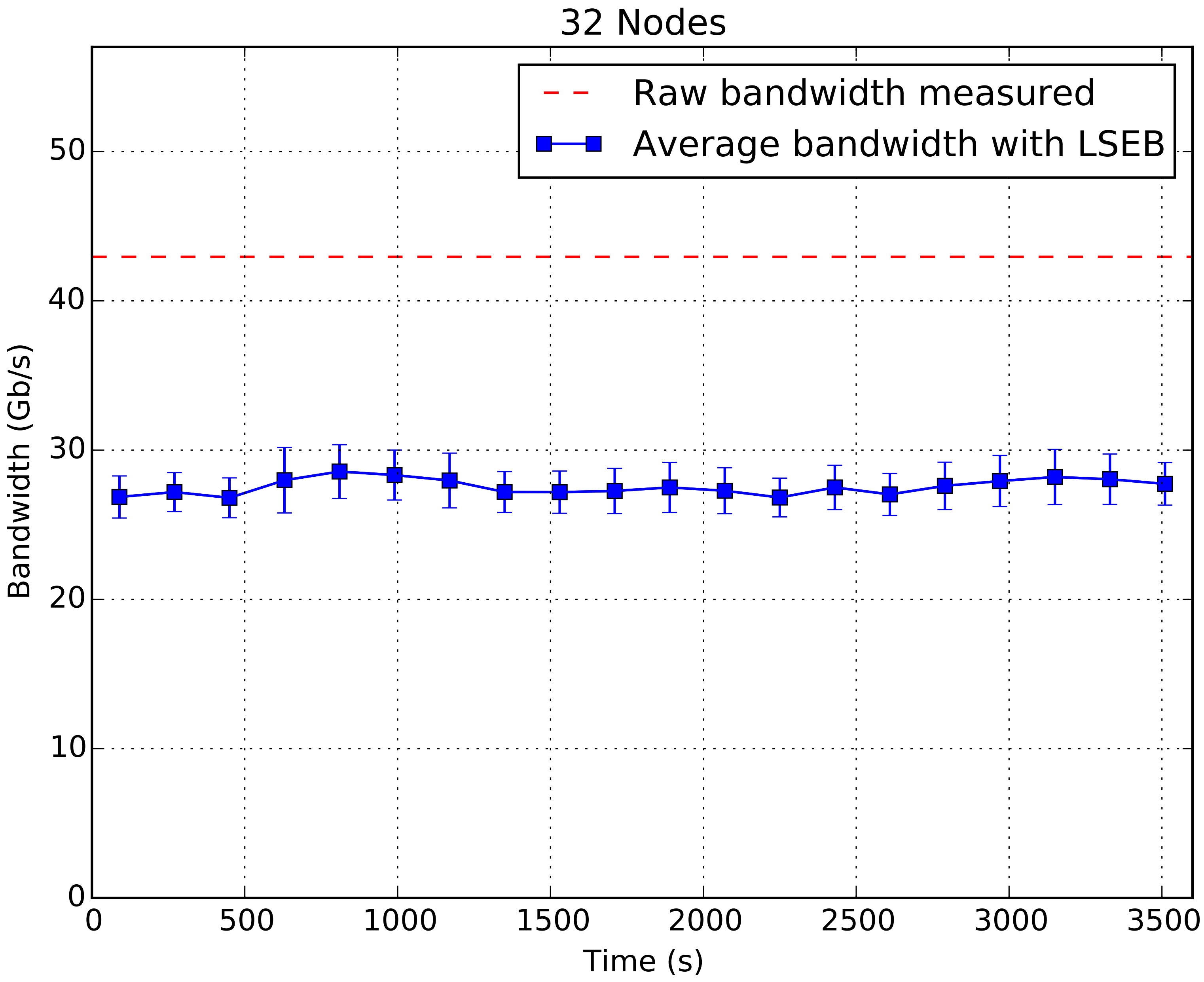
- Cores allocated: 16/16
- Size of buffer: ~128KB
- Average bw: 33.46 Gb/s

Run with 16 nodes



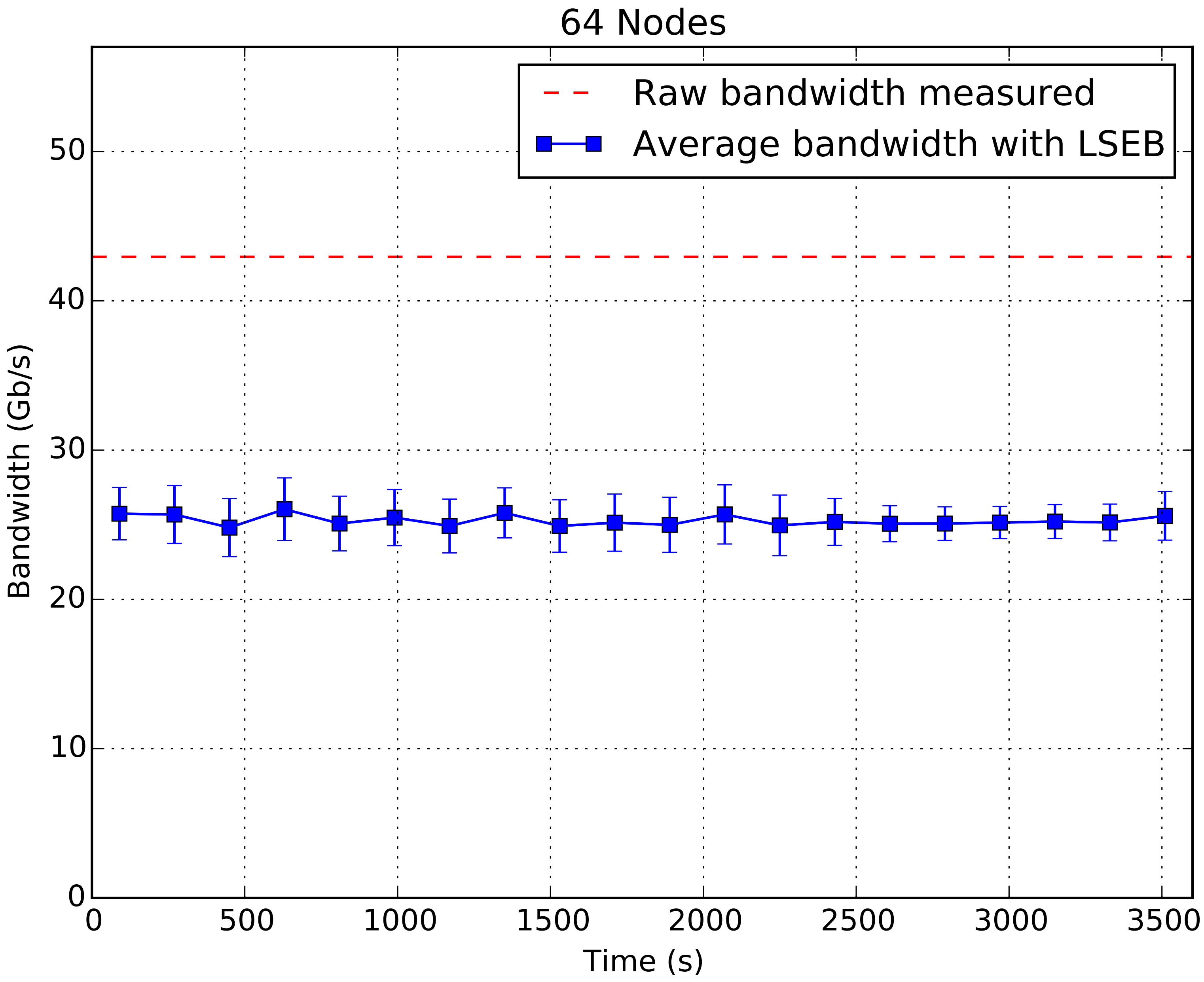
- Cores allocated: 16/16
- Size of buffer: ~128KB
- Average bw: 30.43 Gb/s

Run with 32 nodes



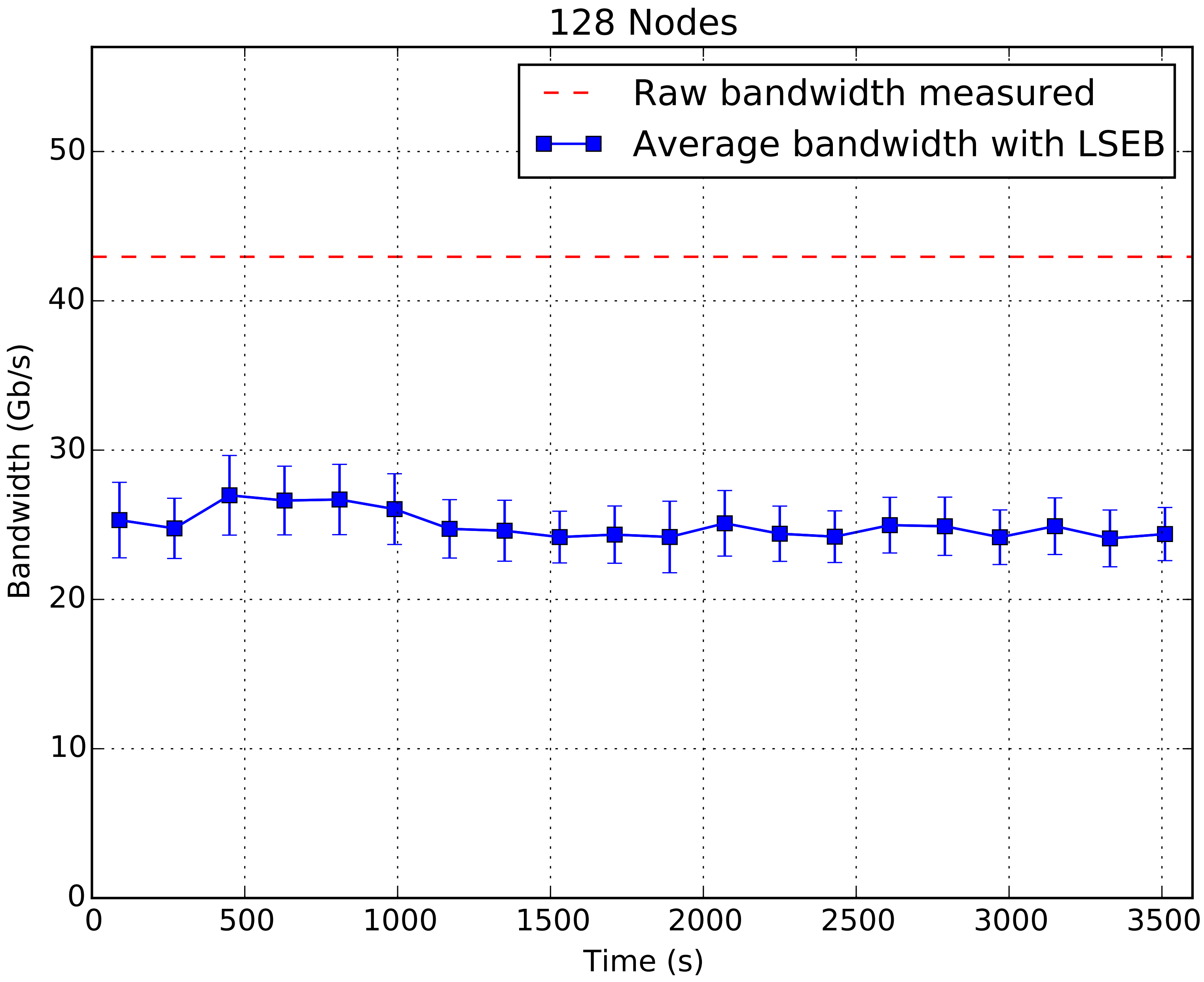
- Cores allocated: 16/16
- Size of buffer: ~128KB
- Average bw: 27.55 Gb/s

Run with 64 nodes



- Cores allocated: 8/16
- Size of buffer: ~128KB
- Average bw: 25.29 Gb/s

Run with 128 nodes



- Cores allocated: 8/16
- Size of buffer: ~128KB
- Average bw: 24.98 Gb/s