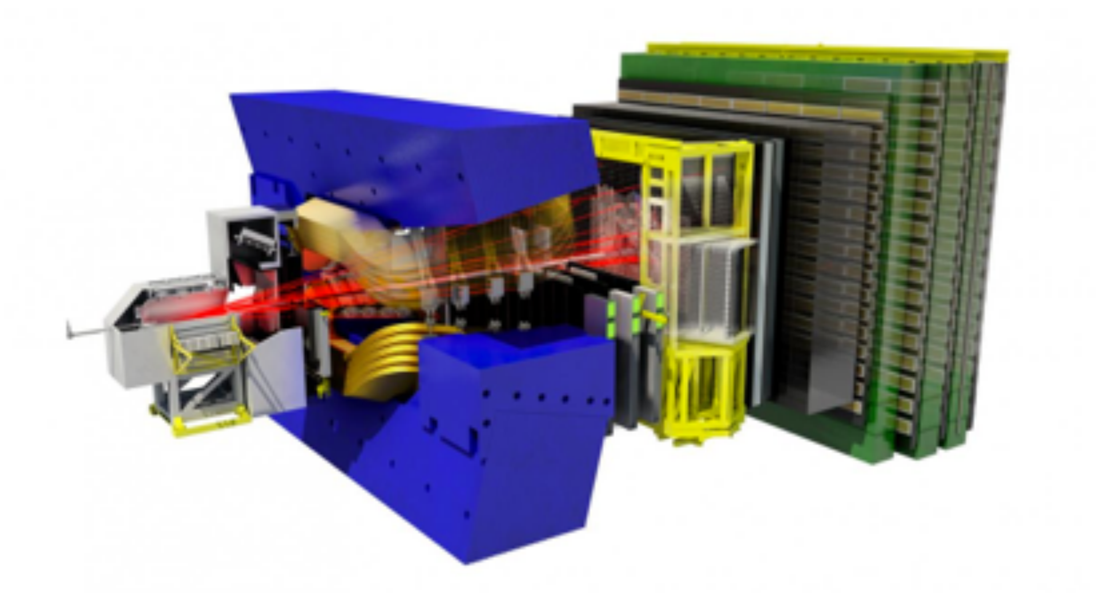
The InfiniBand based Event Builder implementation for the LHCb upgrade

Matteo Manzali
INFN - Università degli Studi di Ferrara

The LHCb experiment

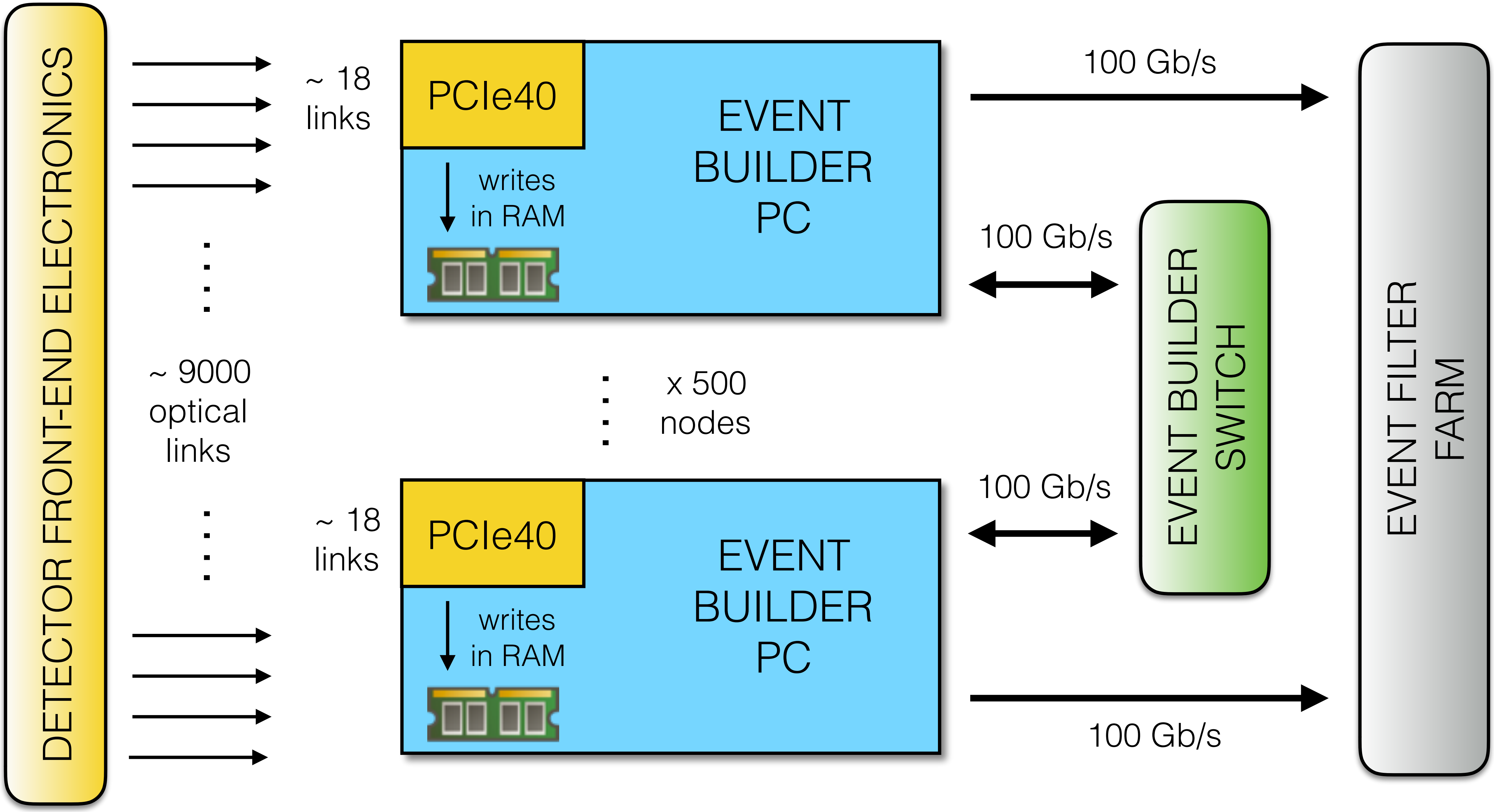
- The LHCb experiment is one of the four large experiments based at CERN



- A major upgrade is scheduled in the 2018-2020 period:
  - Upgrade of the detector
  - Upgrade of the Data Acquisition system (DAQ)

	2015		2018
Event size	65 KB	→	100 KB
Event rate	1 MHz	→	40 MHz
Aggregate bandwidth	520 Gb/s	→	32 Tb/s

Upgraded DAQ design

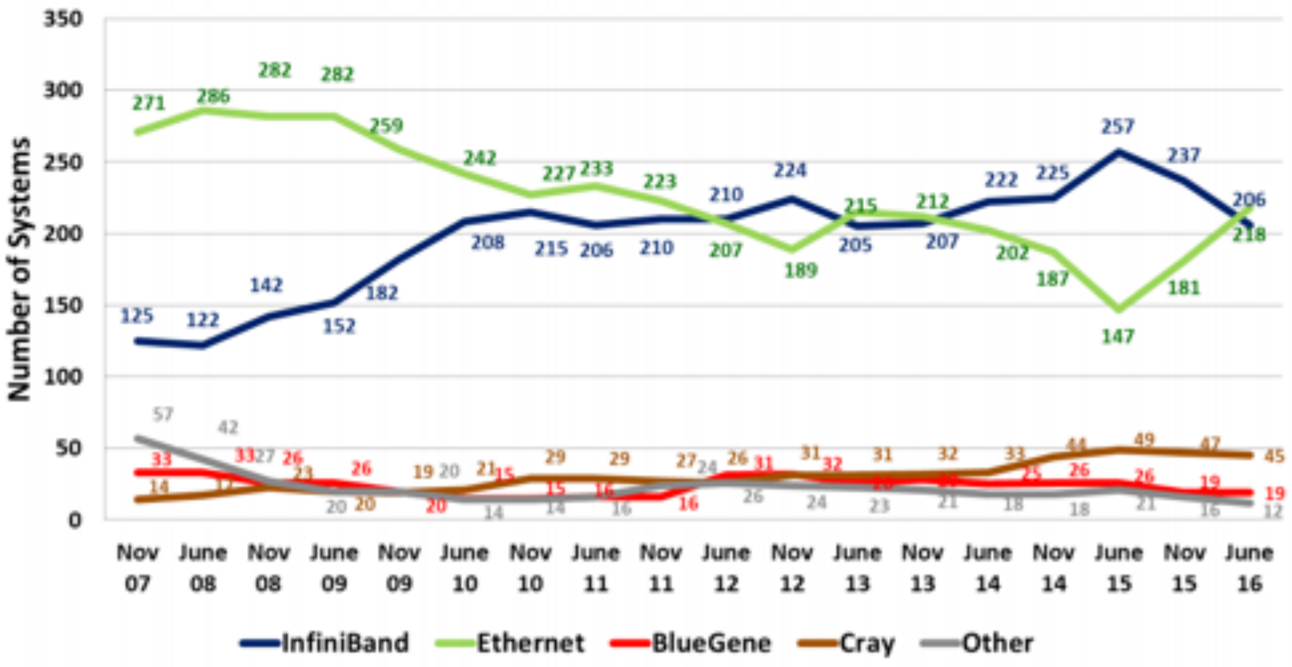


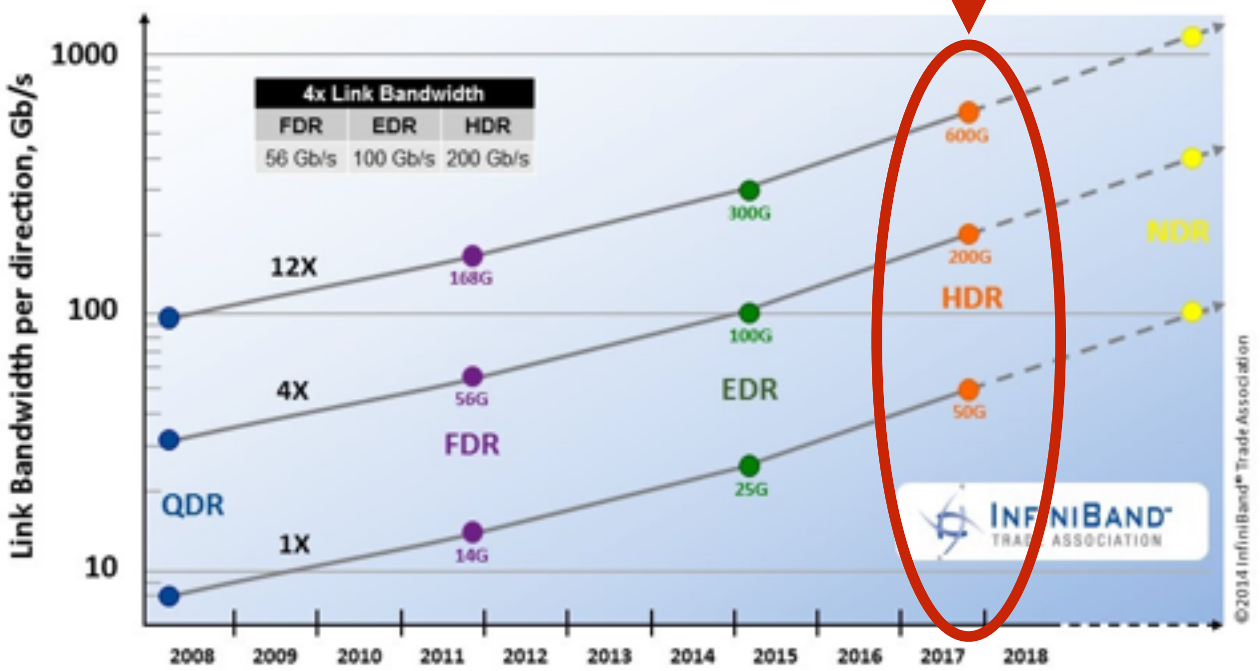
Network technologies

- Different network technologies under study by the LHCb online working group (Ethernet, InfiniBand, Intel OmniPath)

- InfiniBand standard will reach 200 Gb/s (HDR) at the end of 2017

- Low CPU utilization with RDMA (Remote Direct Memory Access)
  - remote memory access without involving OS and CPUs





4x Link Bandwidth		
FDR	EDR	HDR
56 Gb/s	100 Gb/s	200 Gb/s

The Large Scale Event Builder

- The Large Scale Event Builder (LSEB) is an Event Builder software prototype based on the InfiniBand interconnect technology

- Communication relies on the OFED verbs library over InfiniBand
  - Based on RDMA
  - Busy polling

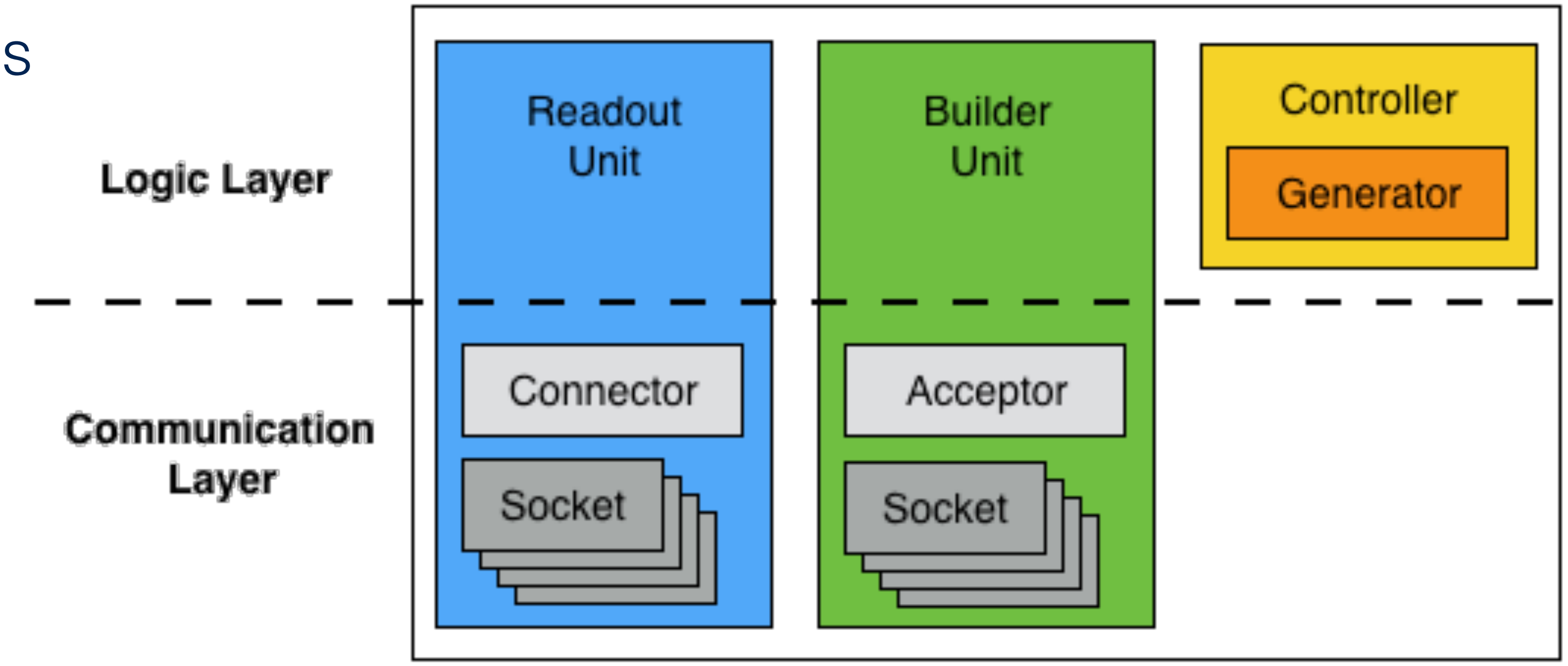
- Few LSEB highlights:
  - C++ and Boost libraries
  - source code available on GitHub (https://goo.gl/Er3rfV)
  - ~3400 lines of code
  - no central scheduler

The Large Scale Event Builder

- A LSEB process is mainly composed of two distinct logical components: the Readout Unit (RU) and the Builder Unit (BU)

- Each RU:
  - receives the event fragments from a generator
  - ships them to the receiving BU in a many-to-one pattern

- Each BU:
  - gathers event fragments
  - generates full events

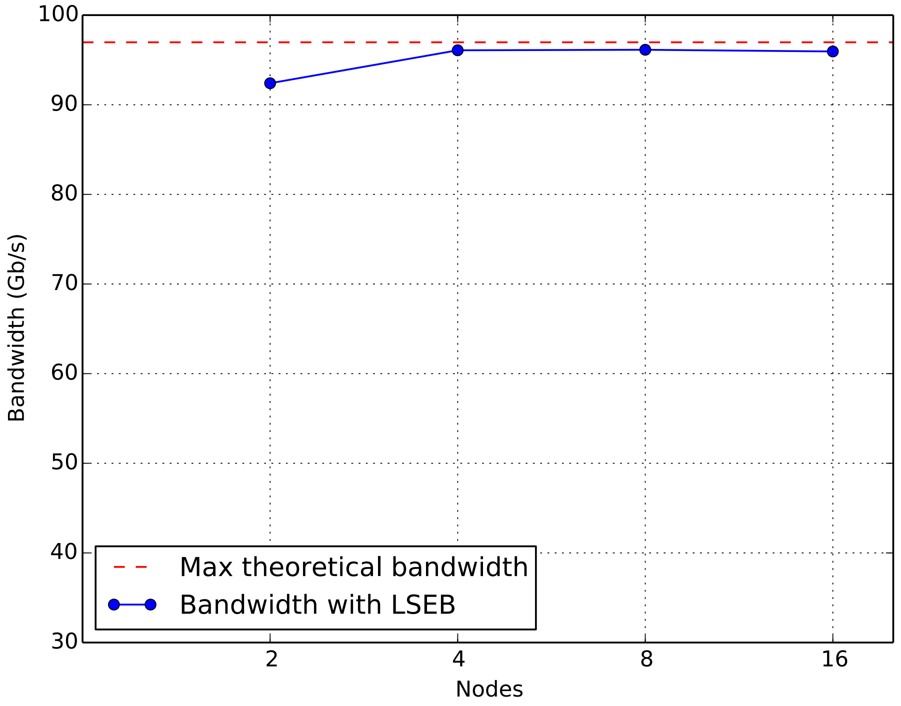


Performance tests

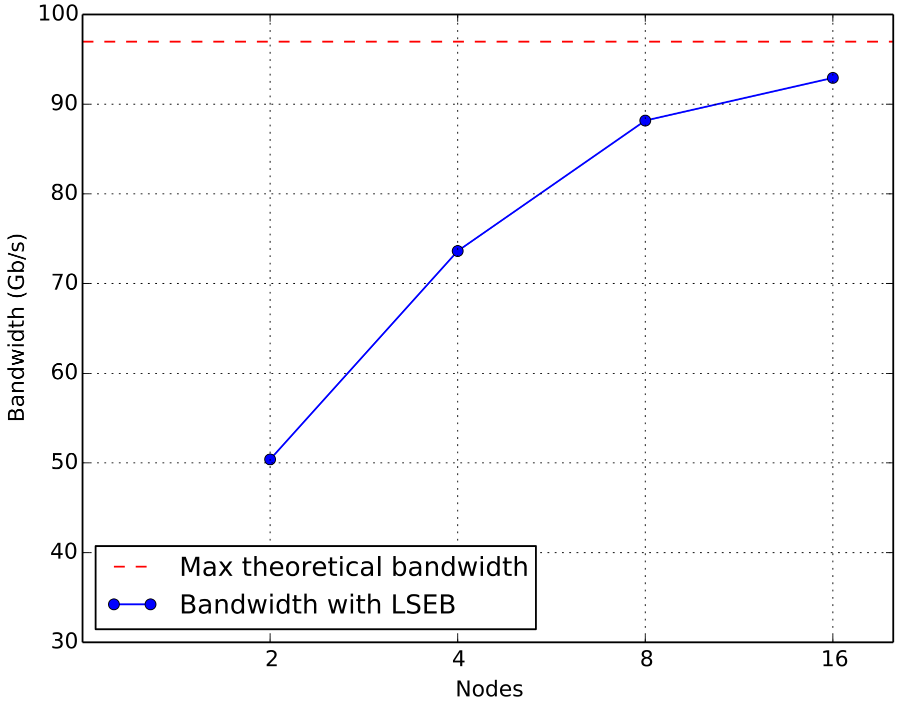
- Cluster size: 16 nodes
- Processors: 2 x 14-cores Intel Broadwell 2.60 GHz (Intel Xeon E5-2690 v4)
- Connectivity: InfiniBand EDR (Mellanox MT_2180110032)
  - Theoretical throughput: 96.97 Gb/s (100 Gb/s - 64/66b)
- Exclusive access to the whole cluster (root access)
- Two different versions of LSEB tested:
  - local traffic handled separately
  - local traffic through network device

Performance tests

- Local traffic handled separately:
    - most intuitive solution
    - BW with 16 nodes: 95.95 Gb/s
    - Reached 98.95 % of the max theoretical BW

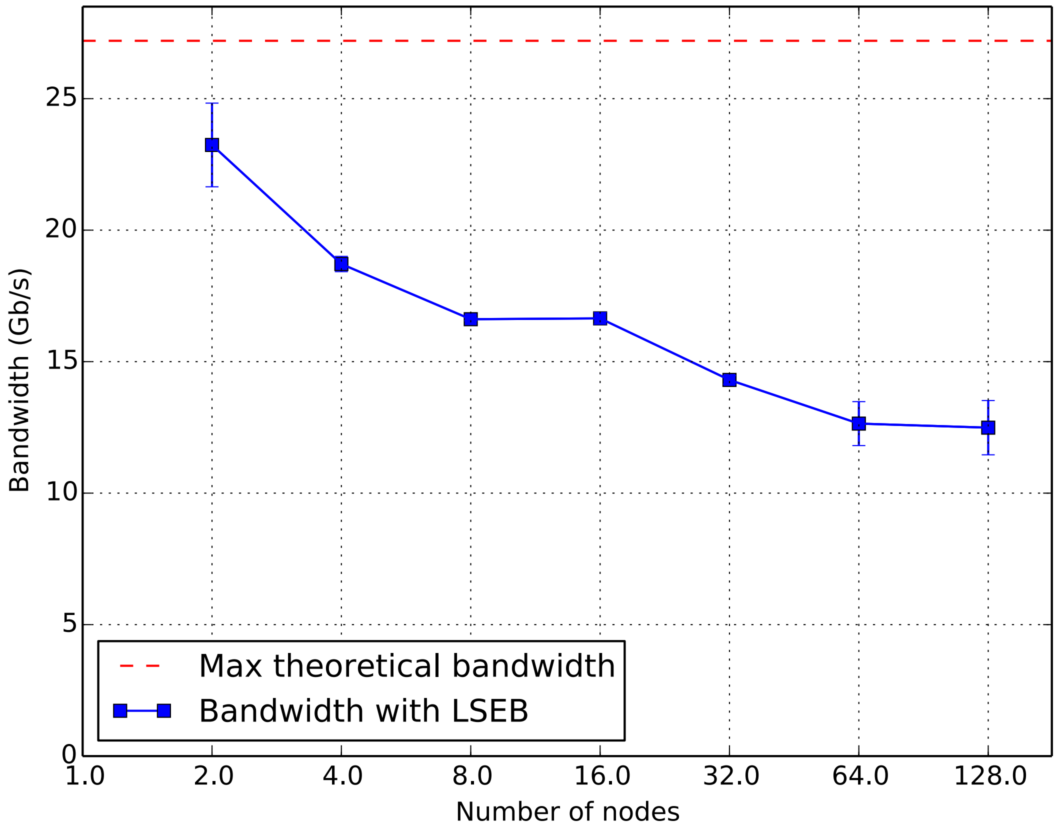


- Local traffic through network device:
    - simpler logic
    - BW gap is due to missing local BW measurements
    - BW with 16 nodes: 92.93 Gb/s
    - Reached 95.83 % of the max theoretical BW



Scalability tests

- Cluster size: ~500 nodes
- Processors: 2 x 8-cores Intel Haswell 2.40 GHz (Intel Xeon E5-2630 v3)
- Connectivity: InfiniBand QDR (QLogic InfiniPath_QLE7340)
    - Theoretical throughput: 27.20 Gb/s (from datasheet)
- Missing optimization settings on nodes
- Non-performant network (constant presence of external jobs running)
- LSEB scales up to 128 nodes reaching 60% of the max theoretical BW

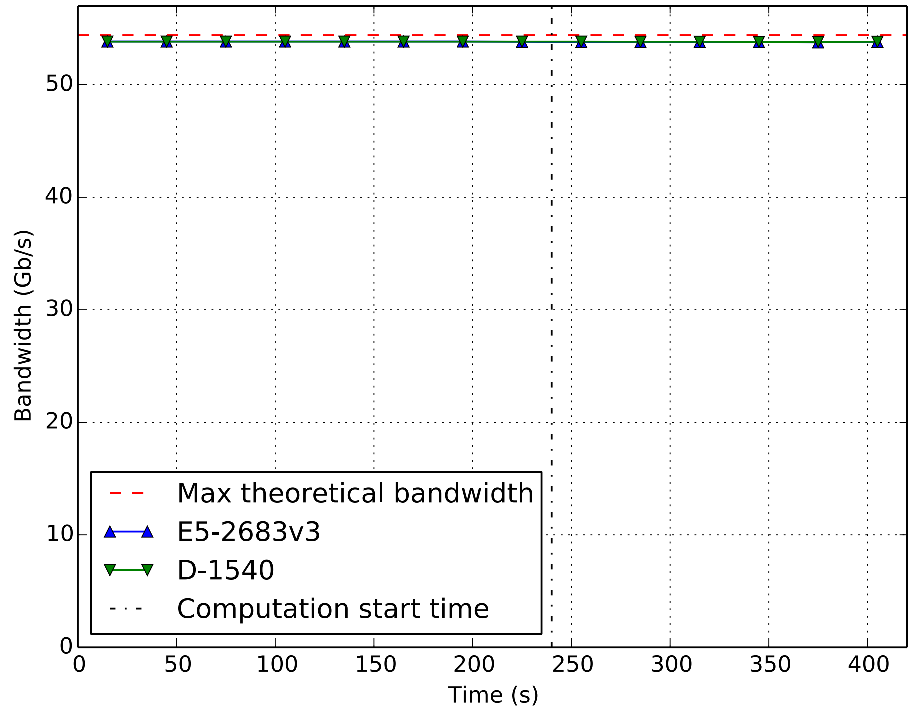


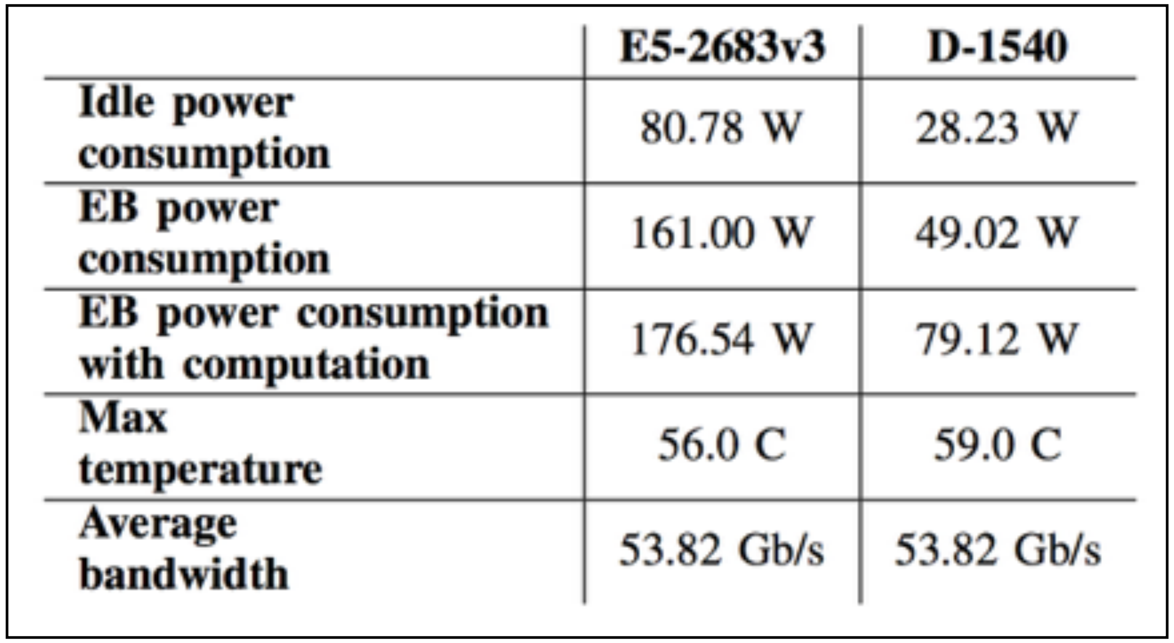
Matteo Manzali - INFN - University of Ferrara

Low power tests

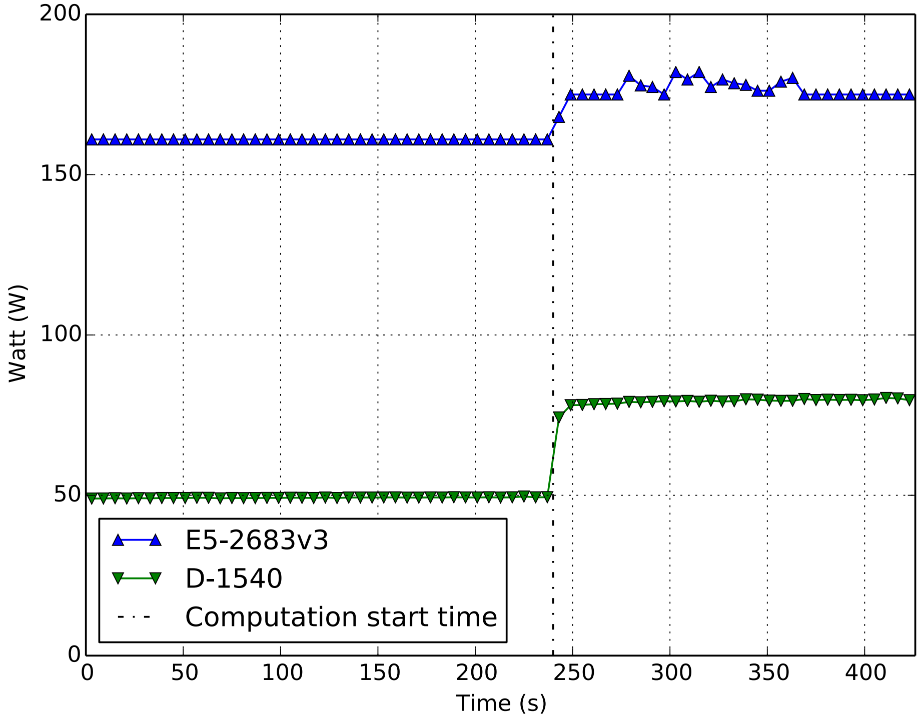
- 2 nodes connected back-to-back with **InfiniBand FDR** (54.3 Gb/s)

- Comparison between:
    - **Xeon D-1540** (low power x86)
    - **Xeon E5-2683v3** (standard server)





	E5-2683v3	D-1540
Idle power consumption	80.78 W	28.23 W
EB power consumption	161.00 W	49.02 W
EB power consumption with computation	176.54 W	79.12 W
Max temperature	56.0 C	59.0 C
Average bandwidth	53.82 Gb/s	53.82 Gb/s



Conclusions

- Implementation of an Event Builder software based on IB interconnect (LSEB)
- Performance and scalability tests performed on different clusters:
    - tests on a small cluster (16 nodes) show that the IB EDR standard can cope with the Event Builder requirements
    - scalability up to 128 nodes has been proven (even without optimisation settings)
- Secondary activity: first investigations on x86 low power architectures

Thank
You