Evaluating Infiniband Based Networking Solutions for HEP/NP Data Processing Applications

Alexandr Zaytsev
alezayt@bnl.gov

BNL, USA
RHIC & ATLAS Computing Facility

Outline

- Motivation
  - HEP/NP Specific
  - RACF Specific
- 4X FDR Infiniband technology evaluation for RACF
  - Phase I (2013Q2-3)
  - Phase II (2013Q3-4)
  - Phase III (2013Q4-2014Q1)
- 4X FDR IB vs 10 GbE detailed price/performance comparison
  - Cost of network hardware and cable infrastructure
  - Taking into account the Infiniband routing/congestion control effects
  - Demonstrating interconnect topology driven optimizations
- Summary & Conclusions
- Q & A

Motivation for the Study (HEP/NP Specific)

- There is an increased interest within the HEP/NP community in using both dedicated and opportunistic HPC resources for handling the data processing tasks that earlier were handled exclusively by the dedicated HTC resources (such as WLCG sites)
  - One can go one step further and not only use the existing HPC resources for HEP/NP applications, but adopt some pieces of the HPC technology in the HTC world
- There are at least two important examples of a success of such an approach that are dealing with the networking aspects of HPC/HTC systems integration:
  - NSC/SCN Project (using IPoIB/4X DDR/QDR Infiniband for running data processing jobs of HEP experiments at BINP since 2010):
    - "Use of the Virtualized HPC Infrastructure of Novosibirsk Scientific Center for Running Production Analysis for HEP Experiments at BINP" (2012): http://indico3.twgrid.org/indico/getFile.py/access?contribId=12&sessionId=47&resId=6&materialId=slides&confId=44
  - Jefferson Lab (extensive experience with 4 generations of 4X Infiniband equipment, recycling lower grade SDR/DDR equipment for HTC purposes):
    - "JLab Scientific Computing: Theory HPC & Experimental Physics" (2014): http://indico.cern.ch/event/274555/session/10/contribution/50/material/slides/0.pdf

Interconnect Solutions Available

Parallel Data Processing Farms

HPC Clusters & Supercomputers

Ethernet

1 GbE
10 GbE
40 GbE
100 GbE

Copper
Optical

Fibre Channel

4 Gbps
8 Gbps
16 Gbps

Optical

***Obsolete* Custom Made**

Myrinet

Quadrics

Scalable Coherent Interface (SCI)

…

Infiniband

SDR
DDR
QDR
FDR
EDR
HDR
…

1X
4X
12X

Copper (passive)
Copper (active)
Optical (active)

Bleeding Edge

Gemini, Dragonfly (Cray)

IBM: direct optical CPU interconnect

Tofu (K Computer)

…

Variety of Interconnect Topologies Available

Parallel Data Processing Farms

Up until recently the only economically feasible way of using the HPC networking equipment in HTC system was only through recycling

HPC Clusters & Supercomputers

Ethernet

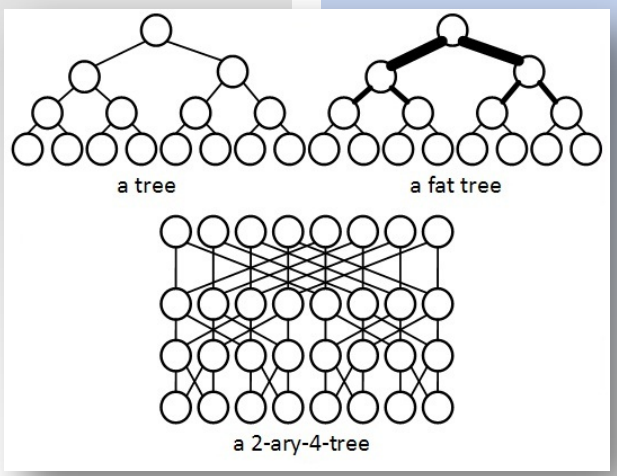
P2P
Tree
Star / Fabric

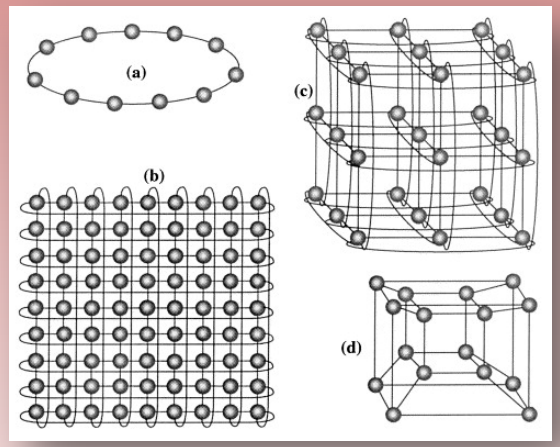
Fibre Channel

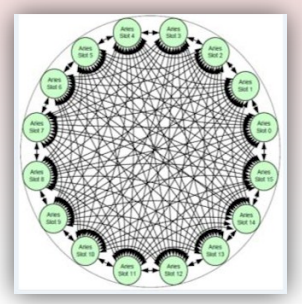
P2P
Loop
Star / Fabric

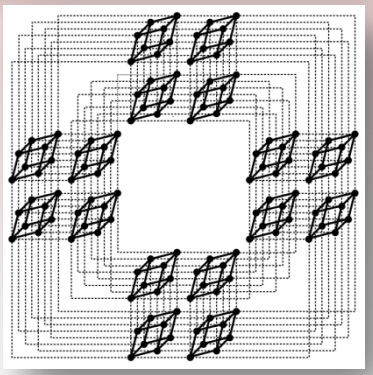
Infiniband & Custom Made (tightly-coupled systems)

P2P
Tree / Fat Tree
Star / Fabric
N-dim Hypercube
N-dim Mesh
N-dim Torus
Dragonfly
…







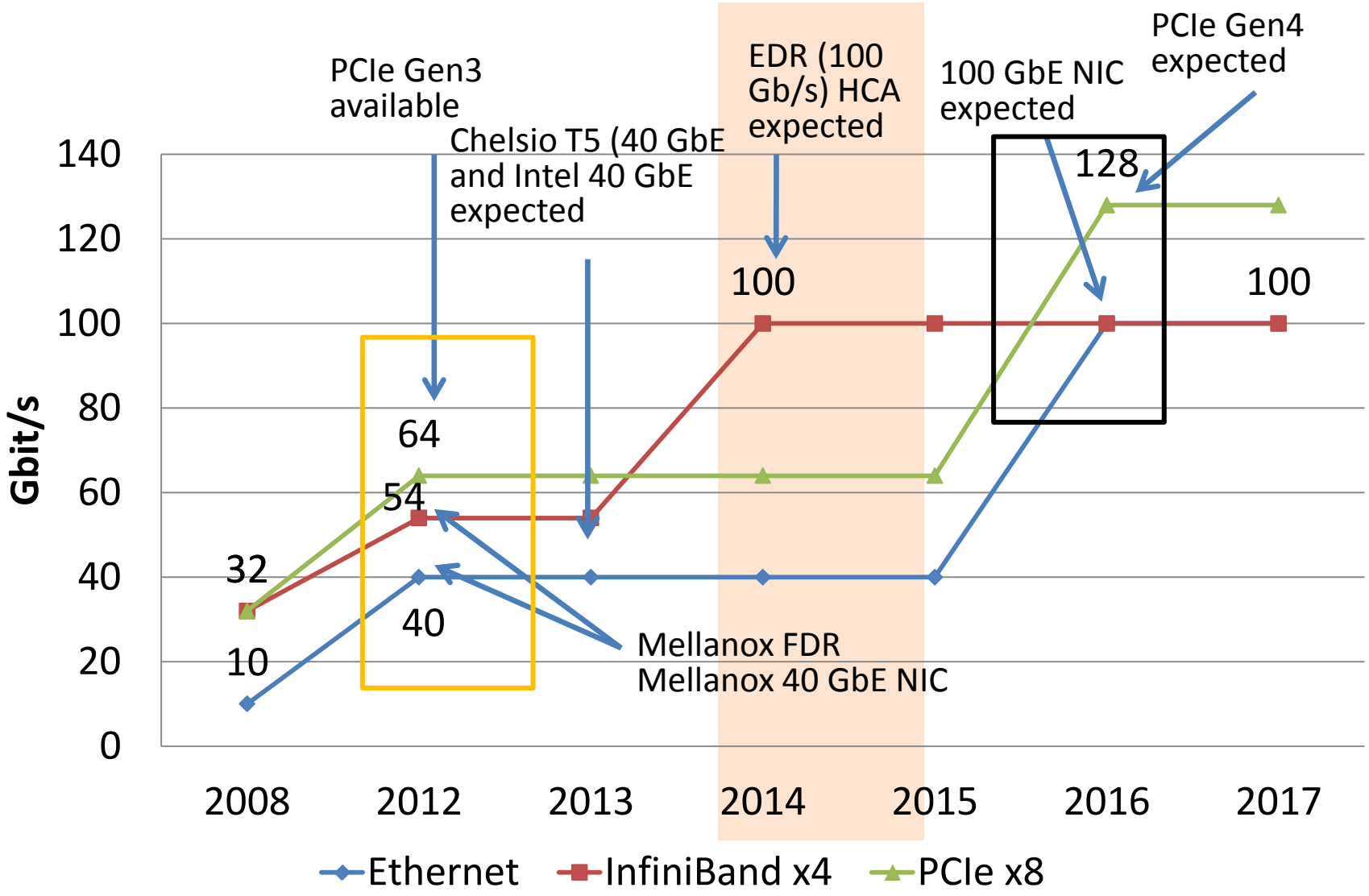


Motivation for the Study (RACF Specific)

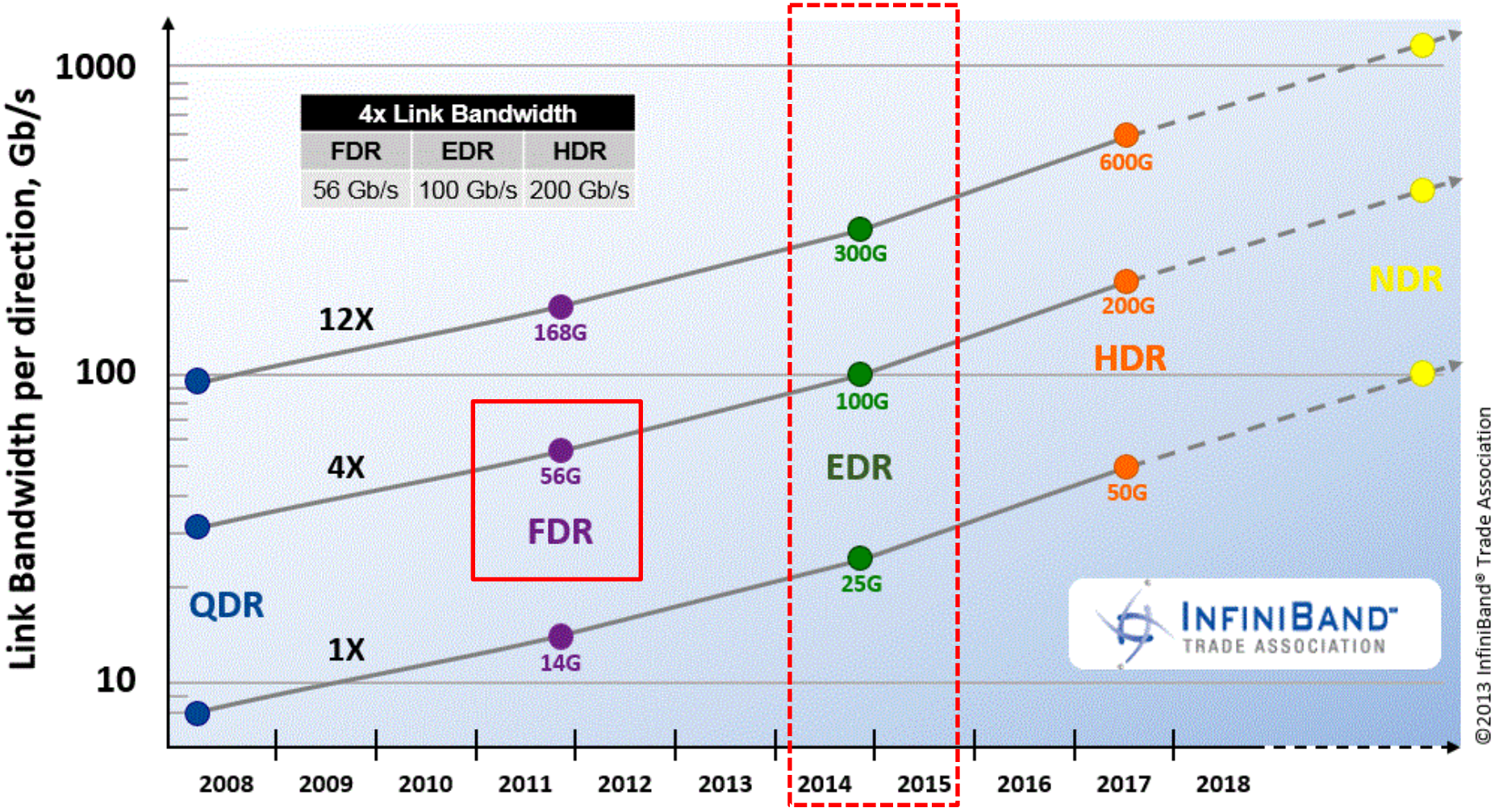
- Up to this moment (2014Q3) all three major farms deployed at the RACF facility (computing farm of the BNL ATLAS Tier-1 site, STAR and PHENIX computing farms) are all making use of 1 GbE based networking solutions at the distribution layer (single 1 GbE copper connection to every compute node)
- Since RHIC farms have dCache distributed storage deployed on top of the compute nodes and the storage capacity of each node have now reached tens of terabytes, 1 GbE networking uplink represents a severe I/O bottleneck for them (no such problem observed on the BNL ATLAS farm though)
- Furthermore, the storage access pattern is quite different comparing the STAR and PHENIX farms:
  - STAR mostly uses direct access to the input files over the network: steady streaming mode during the large scale production runs
  - PHENIX copies the input files to the local disk cache before processing begins, thus creating a surges of traffic across the entire network fabric, especially at the beginning of every production run (data staging can take more than 1 hour sometimes)
- The initiative for choosing the next generation networking solution for RHIC data processing farms (to be deployed in 2015-2020) has started in 2013 with the 10 GbE technology as the first obvious candidate; the 4X FDR Infiniband (56 Gbps/port) was soon recognized as a possible alternative that could deliver even better price/performance ratio
- Since PHENIX farm was known to suffer most from the network bandwidth restrictions, it was chosen as a target test environment for the Infiniband technology evaluation

Endpoint Interface Capabilities / Limitations

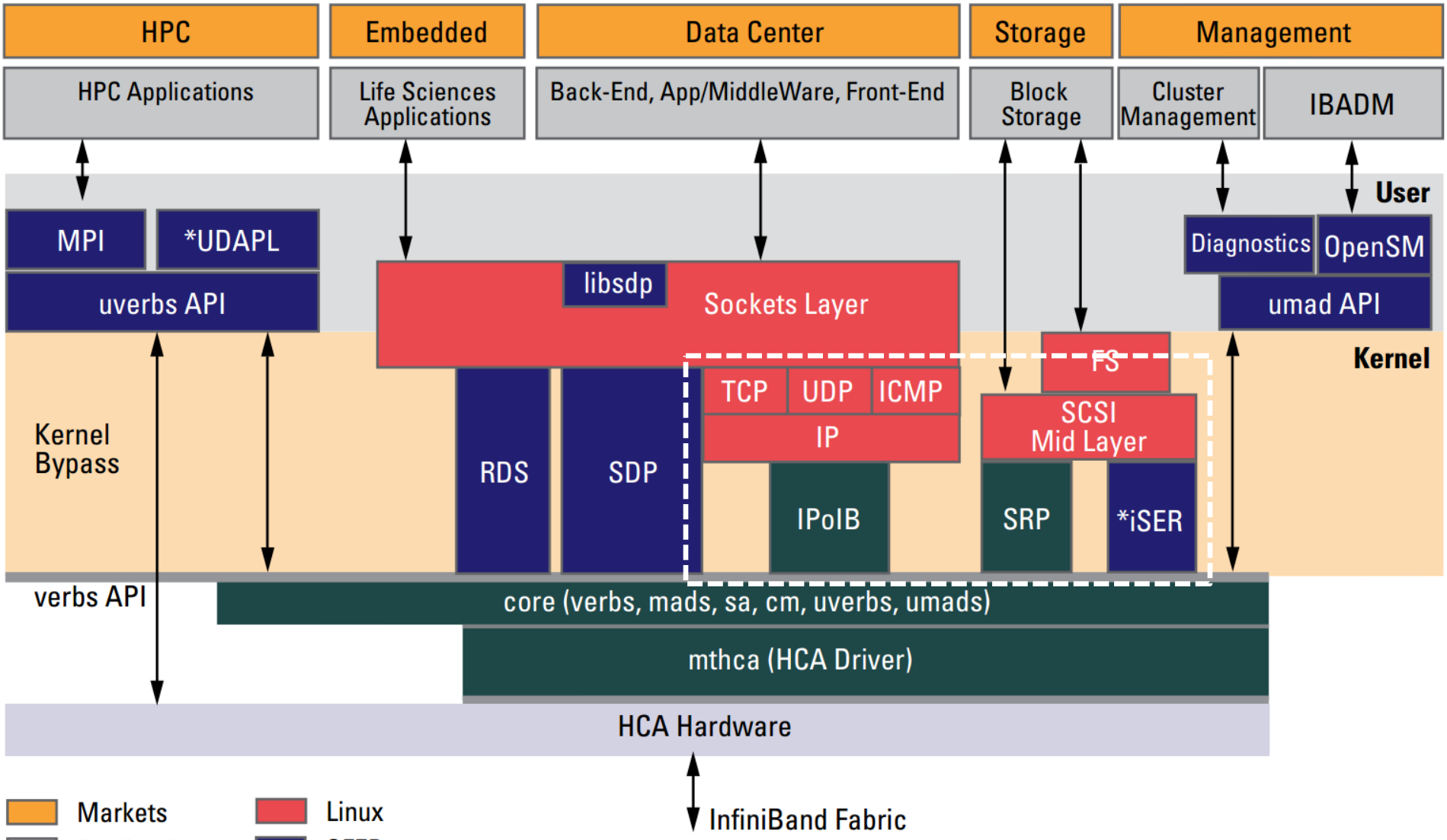
Maximum NIC/HCA Performance Available vs Time



Infiniband Technology Roadmap



OFED (OpenFabrics Enterprise Distribution) Protocol Stack



http://www.mellanox.com/pdf/products/software/OFED_PB_1.pdf

Parts of the Infiniband Protocol Stack That Are of Highest Interest for RACF

- IPoIB (OSI Layer 3 implementation)
  - Encapsulation of IP packets in two modes:
    - Datagram mode (unreliable, MTU limited by HCA)
    - Connected mode (MTUs up to 65520 are supported)
    - Up to 48k unicast + 16k multicast destinations in a single fabric
  - VLANs, IPv4, IPv6, ARP, DHCP are supported
  - It doesn't provide complete OSI Layer 2 support (e.g. attaching bridges)
- EoIB/FCoIB: Ethernet/FC over Infiniband (full encapsulation)
- NFS over RDMA (NFSoRDMA)
- SRP: SCSI RDMA Protocol & iSER: ISCSI Extensions for RDMA
- SDP (Sockets Direct Protocol): byte-stream transport protocol that provides TCP stream semantics

4X FDR IB Technology Evaluation (RACF)

Phase I (2013Q2-3)

- Basic functionality / iperf tests with 4X FDR single port and dual port QSFPs and two test hosts without a switch in between

Phase II (2013Q3-4)

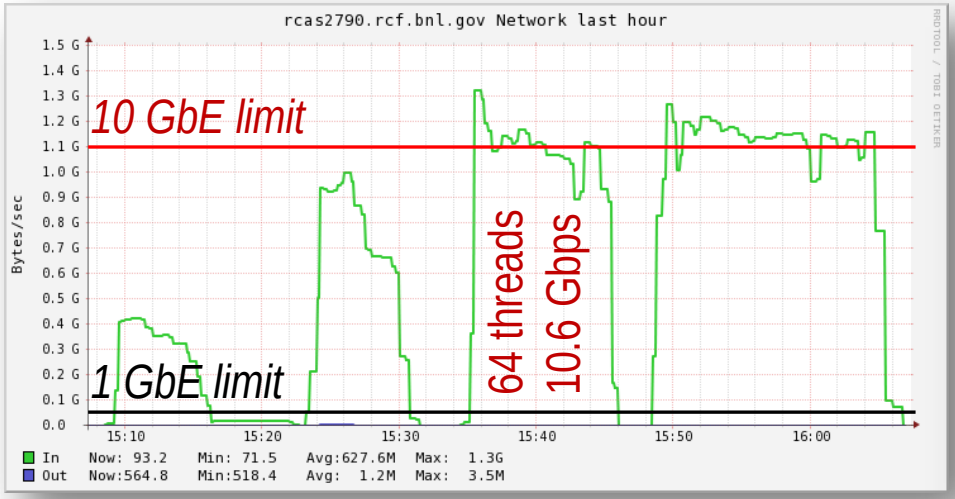
- Introducing a Mellanox managed FDR SX6036 VPI-enabled switch into a testbed (2 compute nodes, 3 HCAs, 1 switch)
- Both iperf and dCache (dccp) performance tests are performed
- Test a custom built Infiniband/Ethernet bridge
- Test a 10/40 GbE (VPI) capability of the SX6036 switch
- Study performance optimizations for IPoIB/4X FDR IB interfaces

Phase III (2013Q4-2014Q1)

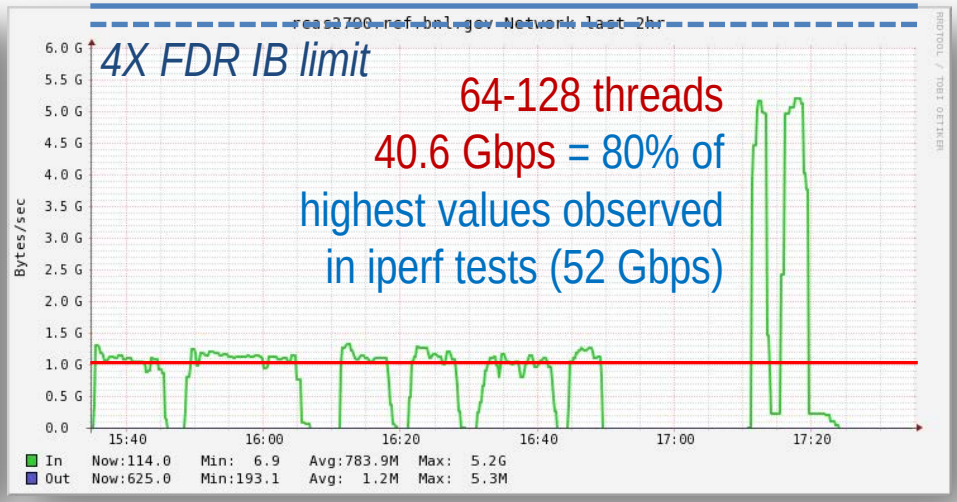
- Introduce two additional Mellanox SX6018 switches plus 28 single port 4X FDR IB HCAs into the testbed: 28 compute nodes and 3 switches (1 spine + 2 leafs); test different fabric oversubscription factors
- 40 GbE uplink is established to the existing Ethernet infrastructure via the ARP bridge configured on the SX6036 switch
- SRP & iSER functionality tests
- HCAs PXE boot functionality tests
- All 28 compute nodes are switch to default connectivity through the IPoIB interfaces running PHENIX production jobs for more than 6 months in 2014

4X FDR IB Technology Evaluation (RACF)

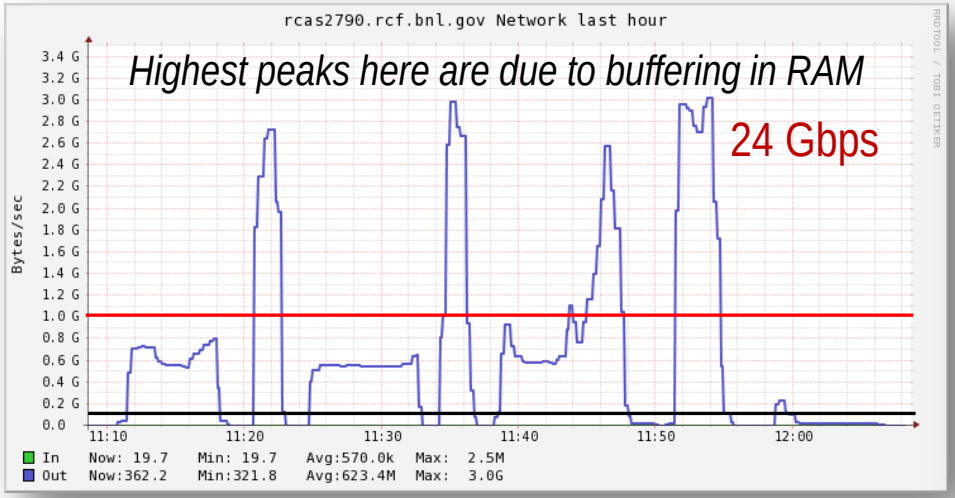
I/O Tests with Gen. 2013 of PHENIX Farm Nodes (12x SATA HDDs in HW RAID5)



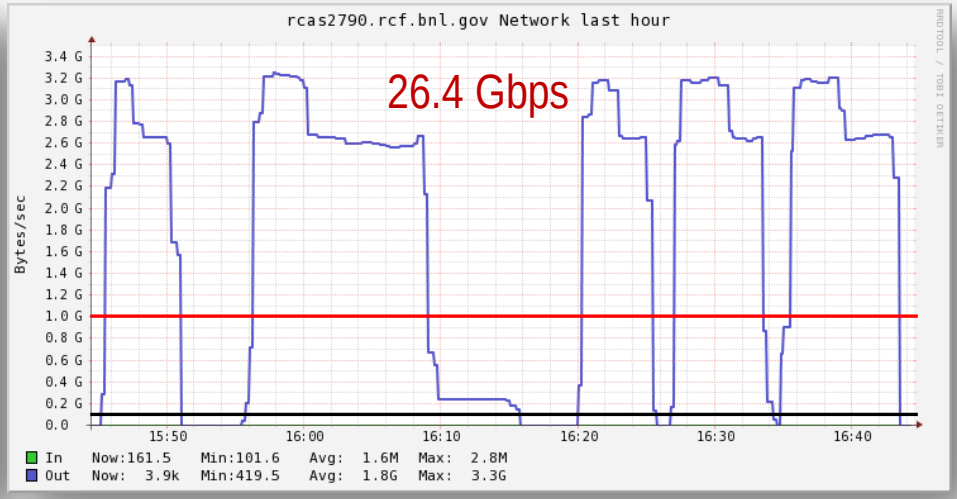
dccp write test 1 (pull from 27 nodes to 1, **local disk**)



dccp write test 2 (pull from 27 nodes to 1, **/dev/null**)



dccp read test 1 (pull on 27 nodes from 1, 1 hot file)



dccp read test 2 (pull on 27 nodes from 1, 7 hot files)

4X FDR IB vs 10 GbE Detailed Price/Performance Comparison (RACF/RHIC – 2014Q1)

- Farm layout used for comparison (projected state by 2020):
  - 1500x 2U compute nodes in total, 14 compute nodes per rack
  - 108 racks in total distributed among 3 designated locations in the RACF data center
  - 2 subfarms (PHENIX & STAR): 750 compute nodes in 27 racks in each
- Projected purchase schedule:
  - 6 purchases in 2015-2020, one purchase per year
  - Adding 250 compute nodes each year (125 for PHENIX plus 125 for STAR)
  - All spine switches + subnet manager nodes (IB) / chassis + supervisor board + switch fabric modules (10 GbE) are obtained in the first purchase
  - Leaf switches (IB) / Line cards (10 GbE) and the necessary cabling infrastructure are added on each purchase step

- 4X FDR IB specific
  - All the racks (save one in each subfarm) are grouped into pairs with one IB leaf switch in each pair
  - Most of the IB cables between the racks are routed on top of the racks on the cable trays (cost is included in the estimate)

- 10 GbE specific
  - MTP/LC patch panel(s) are added in every rack of compute nodes
  - All the cabling between the racks is done under the raised floor (no additional infrastructure needed for cabling)

4X FDR IB vs 10 GbE: Building Blocks

16x Mellanox SX6036
(managed, VPI-enabled, capable of GW/SM functionality)

54x Mellanox SX6025
(unmanaged, non-VPI)

- Switching Capacity: 4.0 Tbps
- 4X FDR IB or 40 GbE Interfaces: 36
- Fabric latency: < 0.2 µs

Mellanox
MCX353A-FCBT

- PCEe 2.0 (3.0 compliant), x8
- 8.0 GT/s max
- MPI latency: < 0.7 µs
- RTT with IPoIB/4X FDR IB: 70 µs

2x Arista 7508

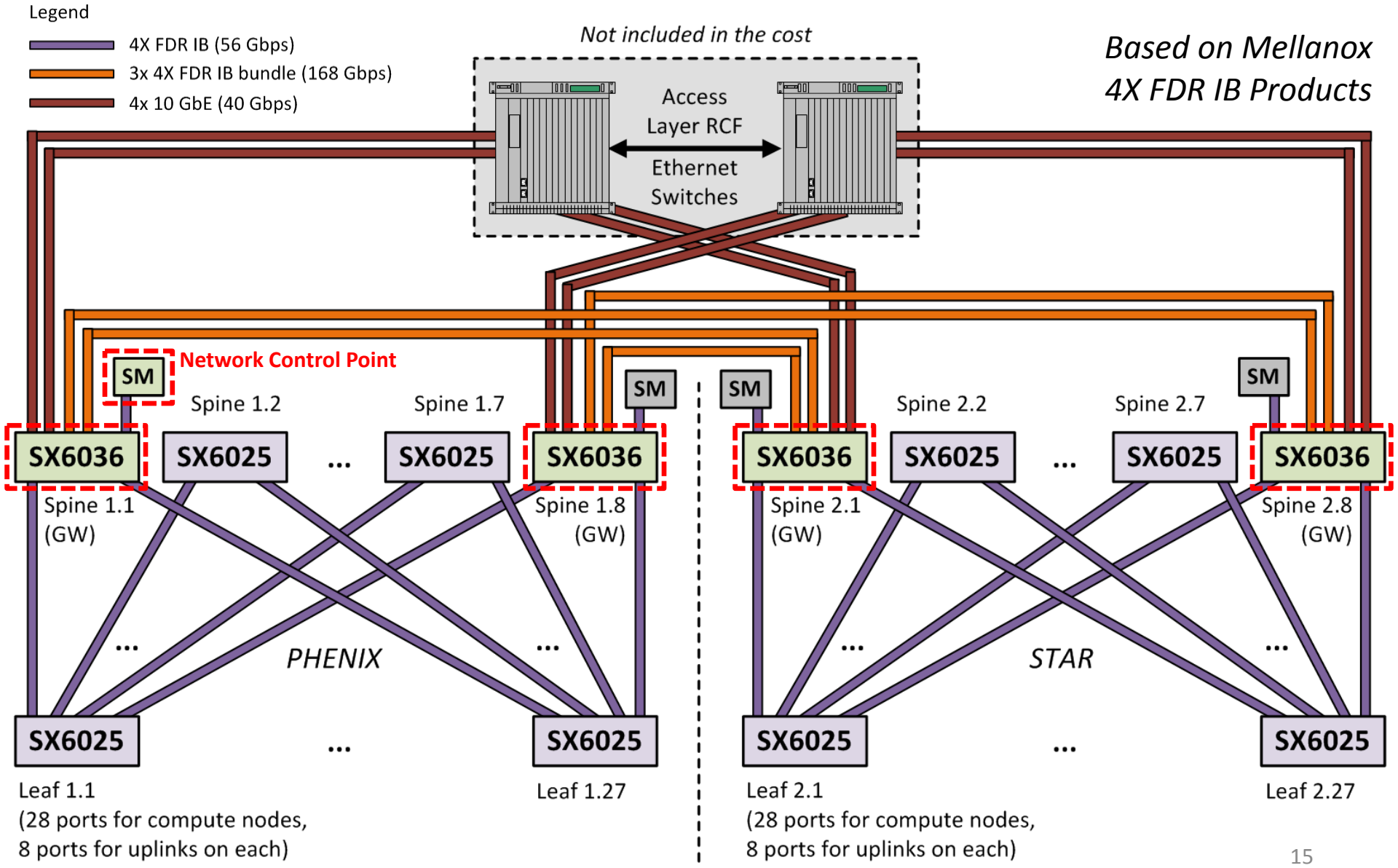
- Switching Capacity: 30 Tbps
- Linecard Capacity: 3.84 Tbps
- 10GbE Interfaces: 1152 (max)
- Fabric latency: < 4 µs

Arista
7500E-12CM

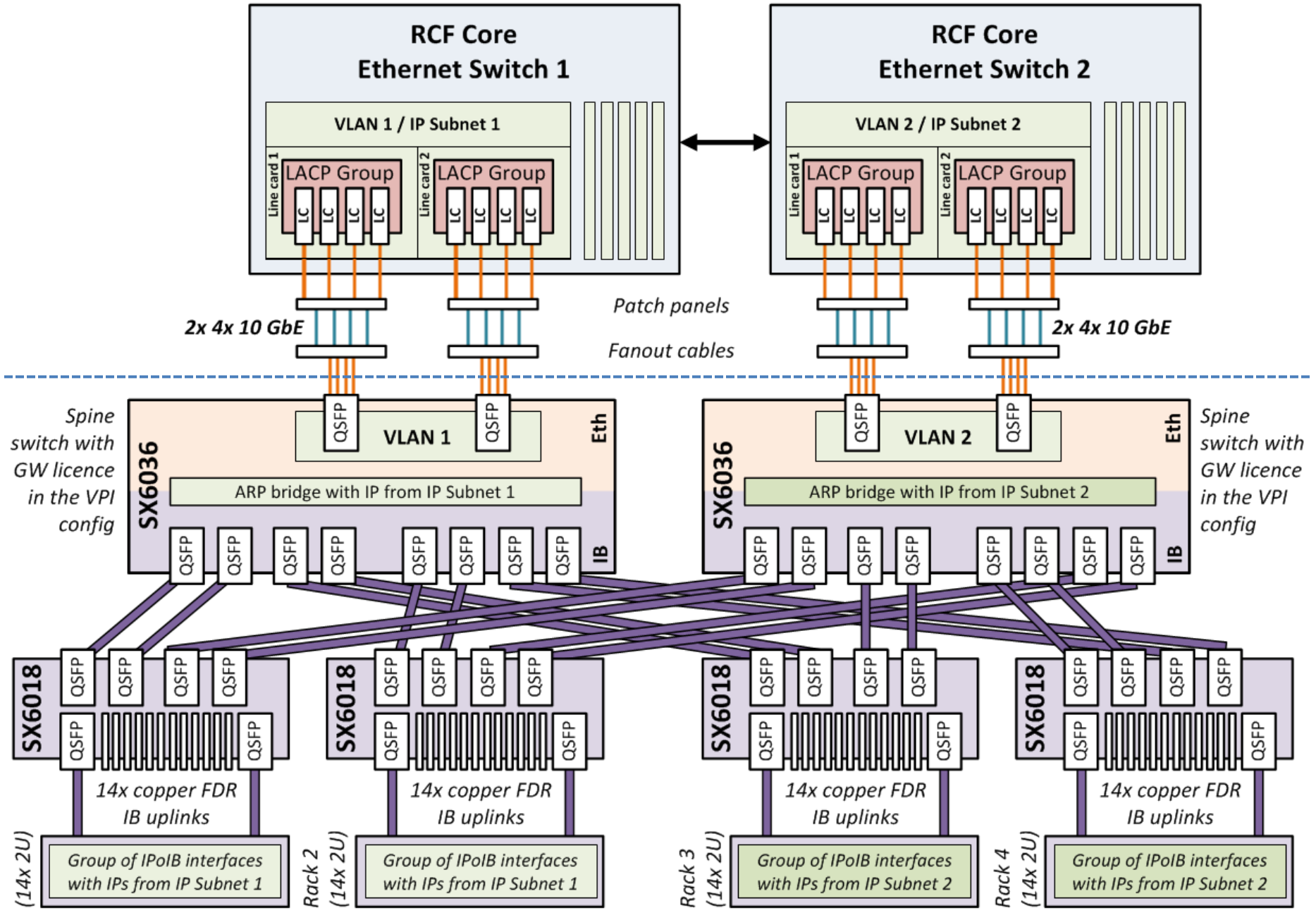
- Forwarding Rate: 1.8 Bpps
- Port Buffer: 18 GB

- Dual port 10 GbE LOM card
- Typical short distance RTT: 200 µs

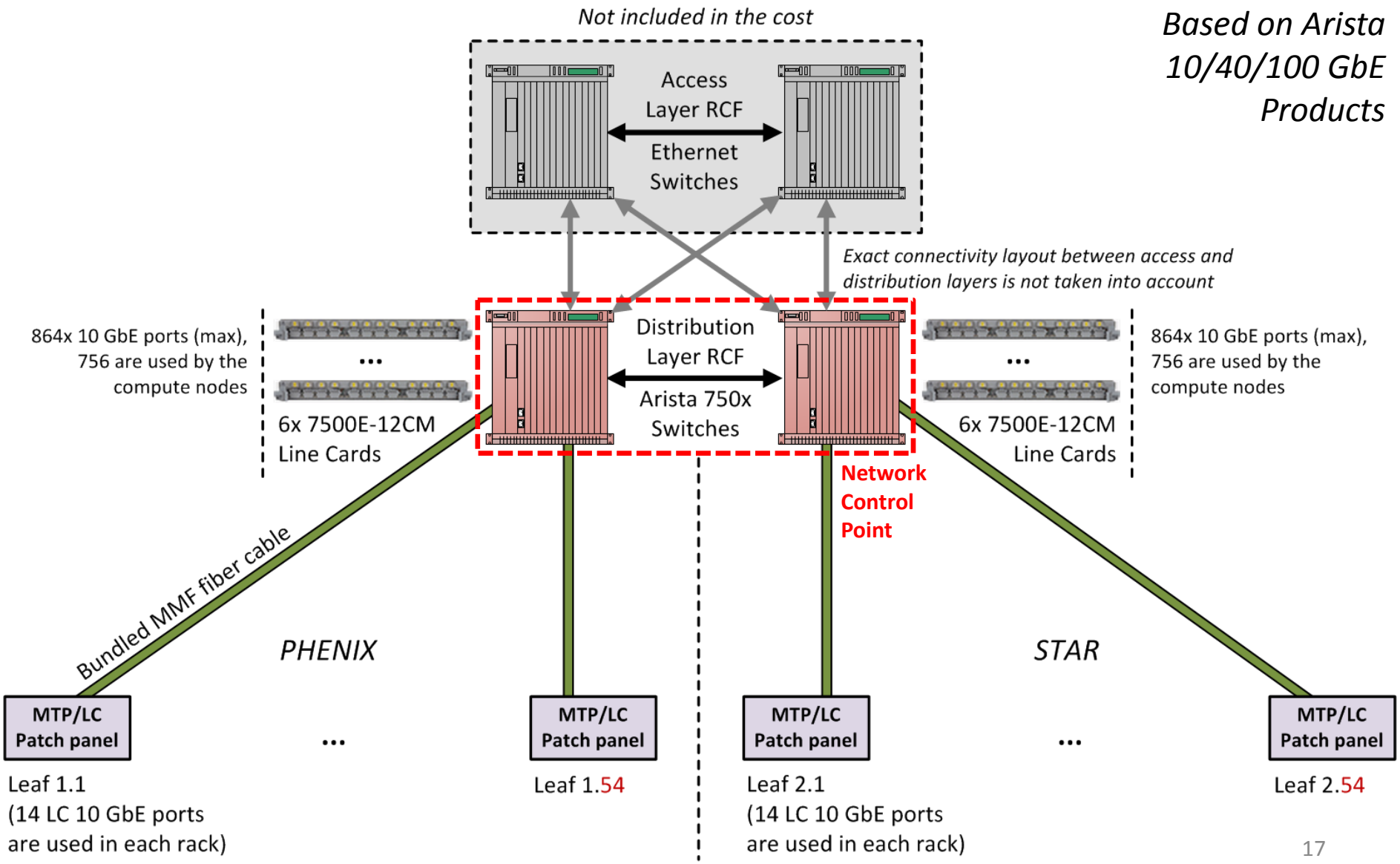
4X FDR IB Based Solution: Network Layout



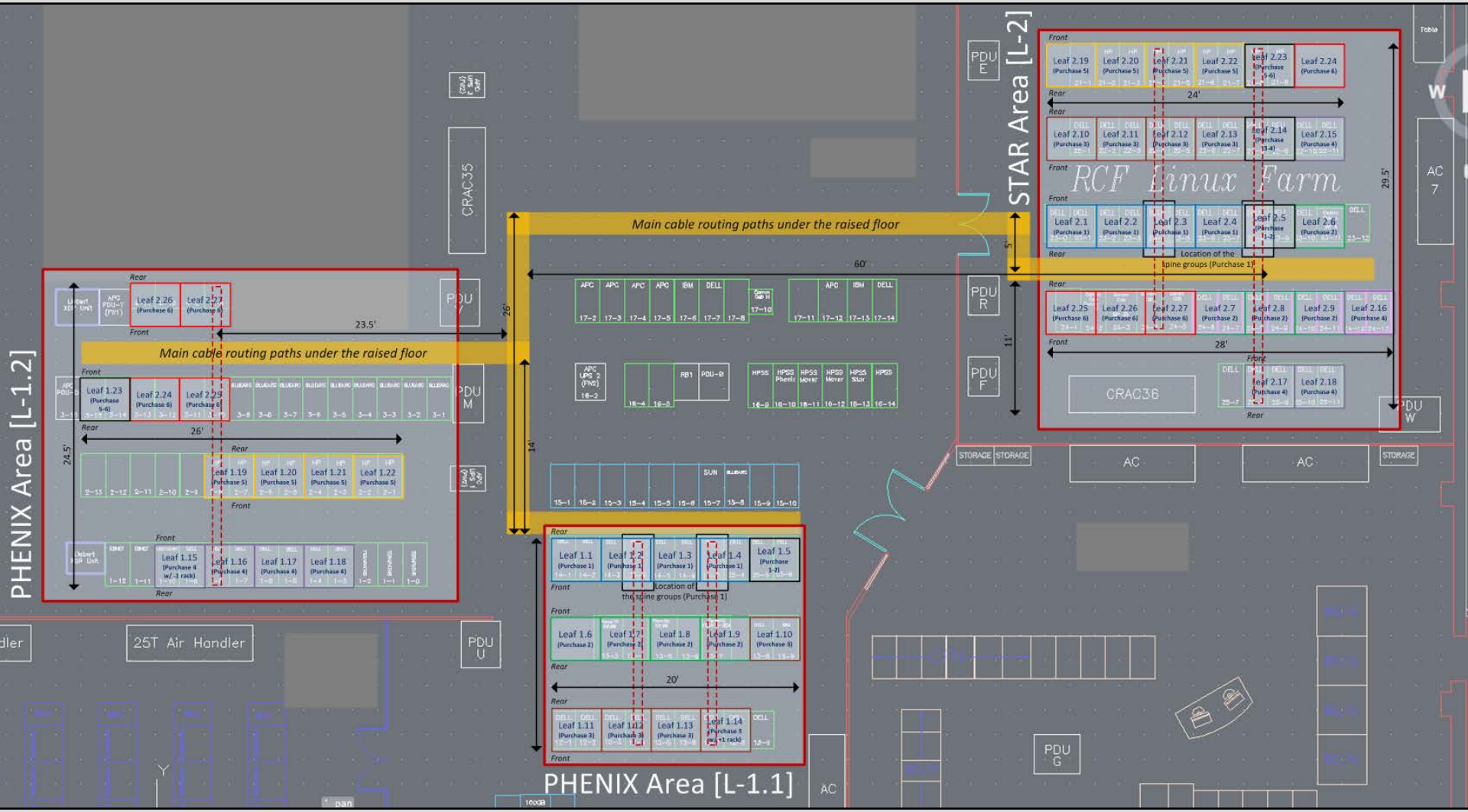
4X FDR IB: VLANs in the Mellanox VPI World (4x 14 nodes)



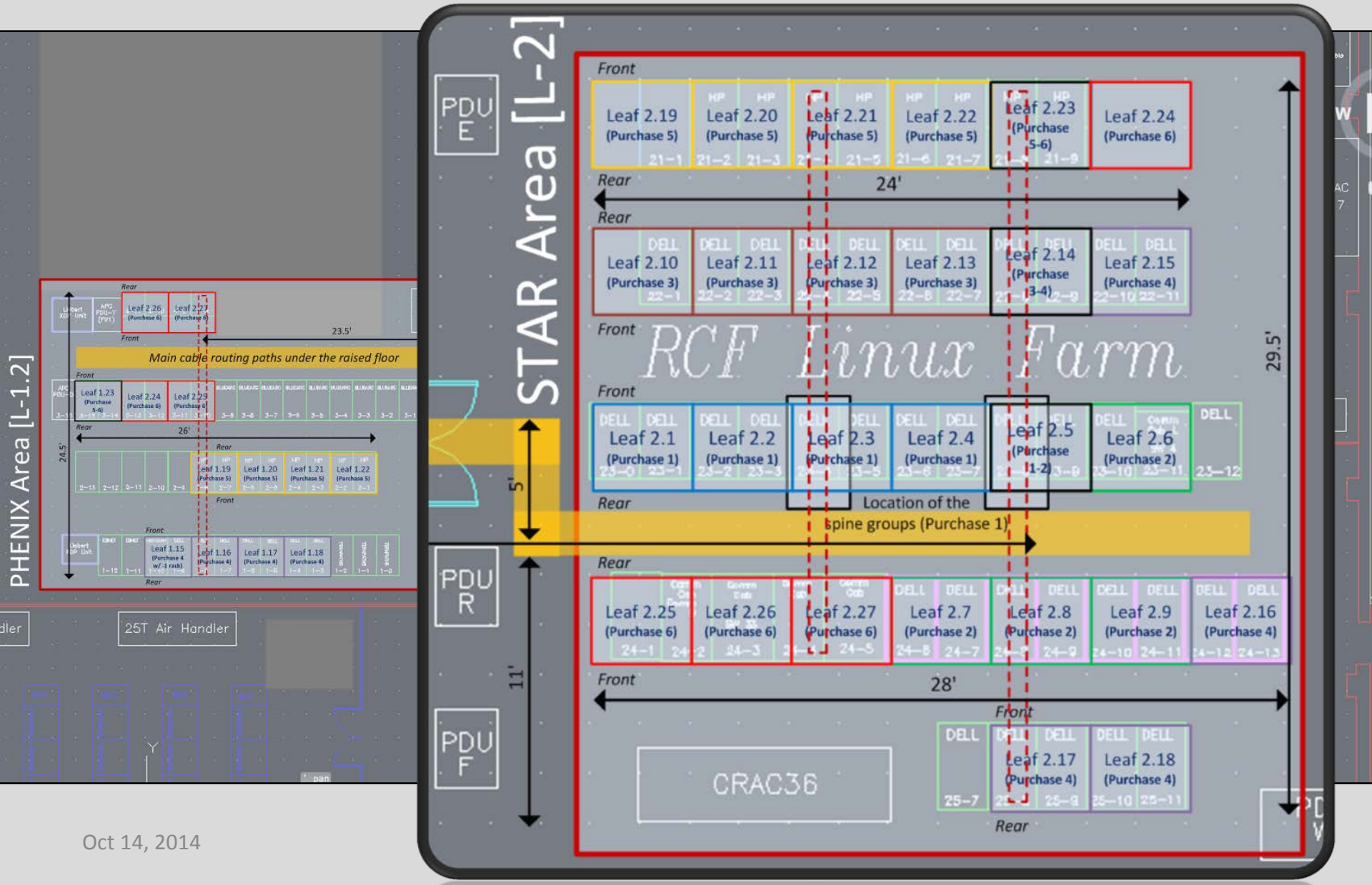
10 GbE Based Solution: Network Layout



4X FDR IB Based Solution: Cabling Layout



4X FDR IB Based Solution: Cabling Layout



4X FDR IB vs 10 GbE Detailed Price/Performance Comparison (RACF/RHIC – 2014Q1)

10 GbE (line speed)

- Sources for price estimates:
  - **Arista (central switches)**
  - RACF NetAdmin team (cabling systems)
  - RACF Linux Farm team, Dell (server side equipment)
- Network bandwidth delivered to the compute nodes:
  - **10 Gbps** in all scenarios and for all access patterns
- 10 GbE solution is **5%** cheaper by the integral cost per compute node as of Feb 2014
  - *The integral price of 10 GbE solution has dropped by 7% as of Oct 2014, but we didn't repeat the whole comparison process (yet)*

4X FDR IB (3.5x oversubscr. tree)

- Sources fir price estimates:
  - **Mellanox, Dell (IB equipment)**
  - Legrand (cabling systems)
- Network bandwidth delivered to the compute nodes:
  - **16 Gbps** guaranteed for the random access pattern across the entire farm
  - **56 Gbps** for communications within the same rack group (same leaf switch)
  - **23 Gbps** weighted average with ≈20% traffic localized on the level of the rack groups (leaf switches) across the farm
- 4X FDR IB solution delivers factor of **1.6x** more guaranteed bandwidth to the compute nodes (**2.3x** if job locality is exploited in the case of the RACF PHENIX farm)

4X FDR IB solution based on factor of 3.5x oversubscribed tree topology was 52% more price/performance efficient per unit of bandwidth delivered to a compute node as of Feb 2014 (factor of 2.2x more price performance efficient if job localization mechanism is in place)

4X FDR IB vs 10 GbE Detailed Price/Performance Comparison: Cost Structure & Upgrades

10 GbE (line speed)

- Major cost contributions:
  - Ethernet switches: **56%**
  - NICs (dual SFP+ port Dell 10 GbE LOM NIC + one 10 GbE transceiver + short fiber): **36%**
  - Optical fiber and cabling infrastructure: **8%**
- Potential upgrades:
  - Since the dual port 10 GbE NICs are deployed on all the nodes in the first place, there is always a possibility of adding the second 10 GbE links to every compute node (factor of 2x increase of bandwidth available)
  - Additional transceivers, 10x line cards, one more switch chassis and extra cable infrastructure would increase the cost of the solution by 63%

4X FDR IB (3.5x oversubscr. tree)

- Major cost contributions:
  - IB switches: **43%**
  - HCAs (single QSFP port Mellanox 4X FDR IB HCA): **40%**
  - IB cables (optical and copper) and cabling infrastructure : **17%**
- Potential upgrades:
  - Starting from 2015Q3 there is a possibility to use 4X EDR IB equipment (100 Gbps per port) in the purchases 2-6, thus increasing the bandwidth, available within each rack group by a factor of 1.8x (no additional cabling infrastructure required)
  - The price of EDR equipment as of 2016Q1 is expected to be on the same level as the FDR equipment in 2014Q1, so it would not affect the current price estimate much (±10%)

4X FDR IB Technology Evaluation (RACF)

Interconnect Topology Driven Optimizations: Traffic Localization

- The job localization mechanism was developed for the RACF PHENIX farm in HTCondor that takes into account the input data placement in the PHENIX dCache deployed on the compute nodes in order to steer jobs closer to the data
- Two level hierarchy was imposed on top of the PHENIX farm provided with 1 GbE interconnect with a simple central star topology that would mimic the two level tree topology of the 4X FDR IB fabric
  - The goal is to demonstrate what level of benefit could be reached by using the real production jobs of PHENIX experiment (first Anatrain and later on – CRS as well) without re-arranging data in dCache
  - Each rack becomes a member of the adjacent pair of the racks to be provided with a single leaf IB switch
- Three levels of job locality are recognized by the mechanism:
  - Machine level locality (input data are on the same host with the job, network traffic eliminated completely)
  - Rack group level locality (all the traffic goes through the non-blocking local leaf switch and doesn't reach the spine switches)
  - No locality (traffic needs to go through the spine switches)

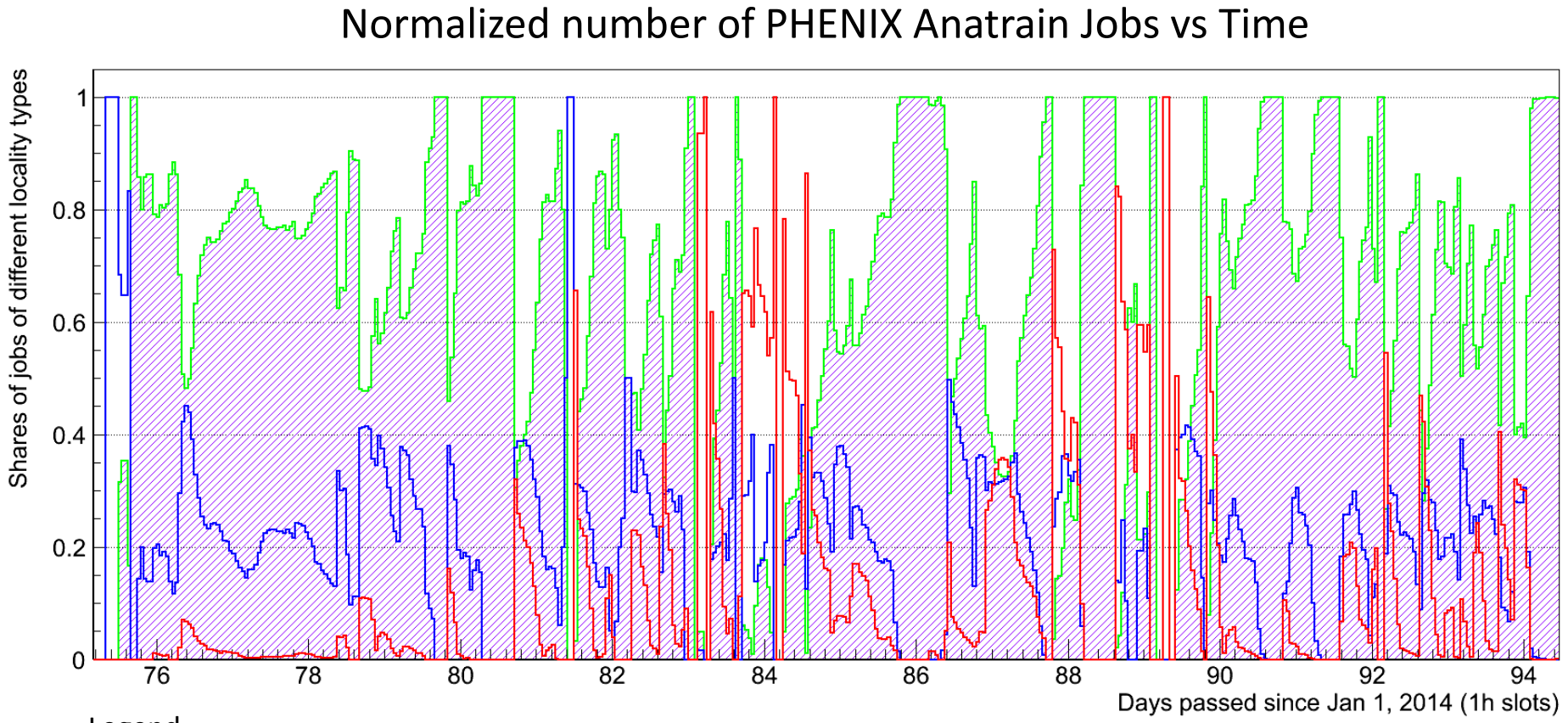
4X FDR IB Technology Evaluation (RACF)

Network Traffic Localization: Implementation Details

- File placement (dCache PNFS IDs) harvested from nodes and placed in a dedicated central database that contains
  - Map of files to machines (client-side name to ID translation)
  - Machines to racks and racks to rack groups mappings
- Host rack-group locality advertised to Condor via STARTD_CRON
- RANK-statement used to steer jobs towards machines where their input files are
  - Slight increase in negotiation time, upgraded hardware to compensate
  - Several thousand matches per hour with possibly unique RANK statements
- The mechanism is used by PHENIX in production since Jan 2014, and during that period the following results were obtained
  - Machine level localization: > 80% on an empty farm, ≈13% on a full farm
  - Rack group localization (inclusive of machine level localization): > 90% on an empty farm, ≈20% on a full farm
  - Even though the mechanism was implemented for demonstration purposes, it actually deliverers about 10% traffic reduction on average due to the machine level localization even in the existing flat network environment
  - These result are applicable to any interconnect system with an oversubscribed tree topology (e.g.: the mechanism would give the same benefits 40 GbE TOR switches)

4X FDR IB Technology Evaluation (RACF)

Traffic Localization: Results (Zoom Into a 1 Month Long Subperiod)



Legend

- Machine level locality
- Rack group level locality (exclusive of the machine level locality)
- No locality (filling on the green histogram)

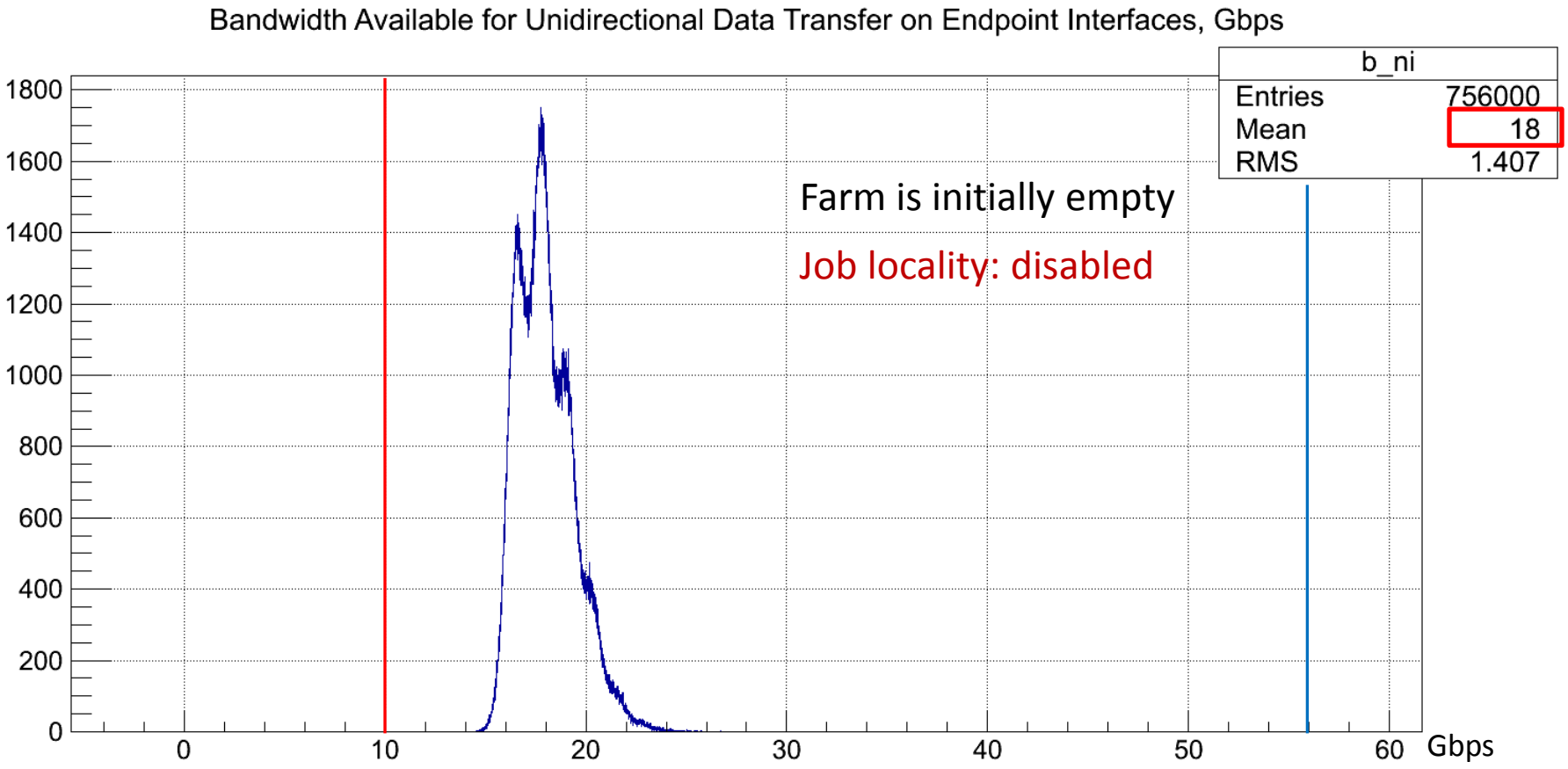
4X FDR IB Technology Evaluation (RACF)

Taking Infiniband Routing / Congestion Control Into Account

- Two kinds of congestion can occur in a simply connected Infiniband fabric provided with static topology such as fat or oversubscribed tree topology:
  - Congestion due to the limited bandwidth along the established paths between two endpoints: being dealt with quite efficiently by the internal IB transmission control mechanisms
  - Congestion due to the specifics of the static routing algorithms (such as MINHOP) selecting which link out of a bundle of links between leaf/spine switches in the fabric are used in each case
    - Difficult to avoid without dynamic routing, much easier to deal with in the fabrics with higher level of connectivity, such N-dimensional torus/hypercube: http://htor.inf.ethz.ch/publications/img/hoefler-tokyotech12.pdf
    - One of the possible solutions is via LID re-assignment across the fabric http://www.cse.wustl.edu/ANCS/2007/slides/acmOlrando.pdf but it yet remains to be demonstrated that it can be used without network service interruption on a large fabric such as one we need (1500 endpoints)
    - Since this second type of congestion is unavoidable in our IB layout, we created a simple piece of simulation software to estimate its affect on the fabric performance in the case of a PHENIX-like farm (written in C++ with ROOT used for visualization)

4X FDR IB Technology Evaluation (RACF)

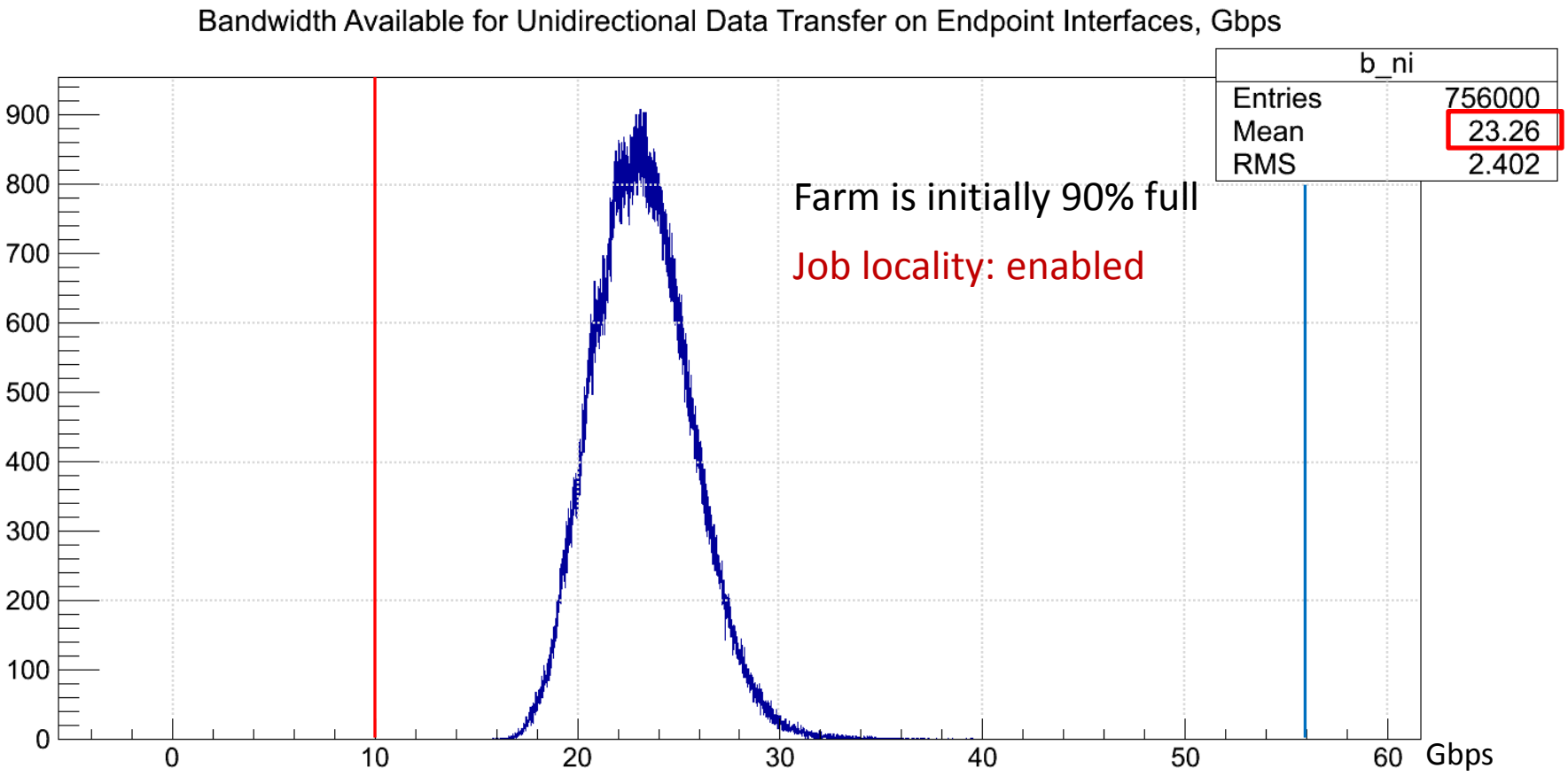
Estimating Effects of Congestion Induced by Static Routing



All the jobs are single core and each jobs pulls one data file from another node in the farm that is randomly selected. The farm is initially empty and is being filled completely on each repetition. (This corresponds to the PHENIX Anatrain start on the empty farm with locality mechanism disabled.)

4X FDR IB Technology Evaluation (RACF)

Estimating Effects of Congestion Induced by Static Routing



All the jobs are single core and each jobs pulls one data file from another node in the farm that is randomly selected. Locality mechanism is enabled but only works in 10% of all attempts (This corresponds to the on-going PHENIX Anatrain start on the fully occupied farm with locality mechanism enabled while the blocks of 10% of job slots are being re-scheduled as the previously started jobs end.)

Summary & Conclusions (Technical Aspects)

- With the recent advances of VPI technology and the increased interest of the HEP community in using the HPC resources for the data processing, Infiniband technology became a competitive alternative to the Ethernet technology traditionally used in the HTC systems (such as parallel data processing farms in HEP/NP)
- A comprehensive functionality and performance evaluation of 4X FDR Infiniband technology has been performed in RACF by using the Mellanox VPI-enabled equipment in 2013Q2-2014Q2
  - A factor of 1.5x price/performance (per unit of bandwidth) benefit of the IPoIB/4X FDR Infiniband networking solution provided with the factor of 3.5x oversubscribed tree topology compared to a conventional 10 Gbps Ethernet networking solution based on Arista 7500 series central switches (one of the most efficient 10/40/100 GbE solutions available on the market) was demonstrated as of Feb 2014
  - Prices of both solutions are gradually dropping, though a substantial drop in the price of Mellanox FDR equipment is expected once the EDR equipment is released next year
- The ways how an oversubscribed tree networking topology can be used efficiently by a HEP data processing farm provided with a distributed storage (such as dCache) deployed on its compute nodes and HTCondor batch system were explored with the PHENIX computing farm:
  - 60-80% peak network traffic localization on the level of a rack group is achievable (20% - on a full farm)
  - A factor of 2.2x price/performance (per unit of bandwidth) benefit can be reached by using job localization within the environment of the RHIC PHENIX farm
- 4X FDR Infiniband technology was also successfully adopted by the RACF Ceph based object storage system (please refer to a dedicated talk entitled “Ceph Based Storage Systems for RACF” in the Storage & Filesystems track of this conference)

Summary & Conclusions (Technical Aspects)

- An important transition is on-going over the last years in the HTC worlds (following the similar transition that already happened in the HPC):
  - The highest level of efficiency in building and operating a large scale computing systems can no longer be achieved by optimizing individual components of such a system (such as compute nodes, network, storage subsystems, user software)
  - Such systems must be optimized as a whole taking into account the specifics of the user driven workflow, thus sysadmin & netadmin tasks are no longer well separated
- Infiniband (the most used interconnect among the TOP-500 machines as of 2014Q3), together with IPoIB and the VPI technology (as provided by Mellanox) gives a perfect example of an HPC technology that can be seamlessly integrated into the existing HTC facilities resulting in a great deal of price/performance benefits, especially if the user workflow optimizations are possible
  - It can also be the key for the HEP/NP community to gain access to the opportunistic CPU cycles of more than 44% of largest supercomputer sites across the globe, and not only the TOP-10 machines that are currently receiving most of attention
- Even though both PHENIX and STAR experiments at BNL eventually decided to go after a 10 GbE based solution as their next generation networking systems (for non-technical reasons), the results of this work are quite generic and could be used by any other HTC site willing to explore the HPC-driven networking technology

H. Subramoni *et al.*, Design of a Scalable InfiniBand Topology Service to Enable Network-Topology-Aware Placement of Processes, 2012
http://conferences.computer.org/sc/2012/papers/1000a076.pdf