SUPERCOMPUTER ARCHITECTURE: PRESENT AND FUTURE

Moscow, 19.07.2012

Anton Korzh
Head of ResearchLab , T-Platforms
Anton.korzh@t-platforms.ru

Outline

- Supercomputer architectures: theory
- Lomonosov architecture
- Exascale architecture

Architecture of supercomputers

Large numbers of CPUs connected together (parallelism)

- Either shared memory (SMP)
- Either coherent NUMA-system
- Either non-coherent NUMA-system
- Either distributed memory system

From user perspective

- Bunch of nodes with batch system
- Each node either multicore, or has accelerator
- MPI, each node connected to each node

Connecting thousands of nodes

- Direct Networks
  - Each node has a network card with integrated NIC and switch
  - Number of network ASICs is linear to number of nodes
  - Common topology is a torus/mesh for a low-radix
  - High-radix considers to use dragonfly-like topologies
- Indirect Networks
  - Switches and NICs are separated from each other
  - Fat-tree/Clos as common topology
  - Number of switches grow exponentially
- Considerations:
  - Cost (switches, cables)
  - Performance (diameter, bandwidth, bisection)

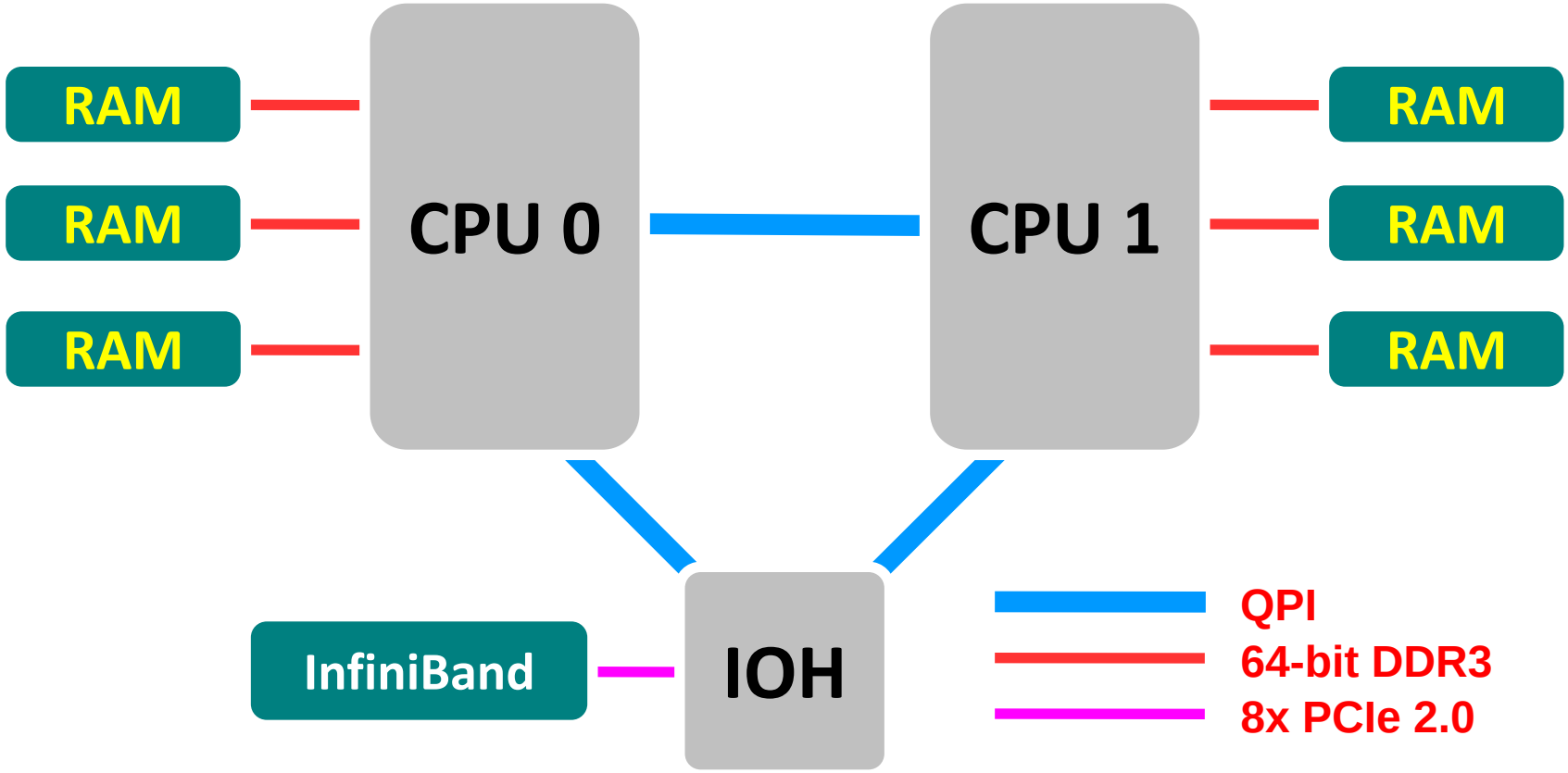
Lomonosov supercomputer, outside

T-BLADE 2
Major building block, Front view

T-BLADE 2
Rear view

T-BLADE 2
Hot plug blades

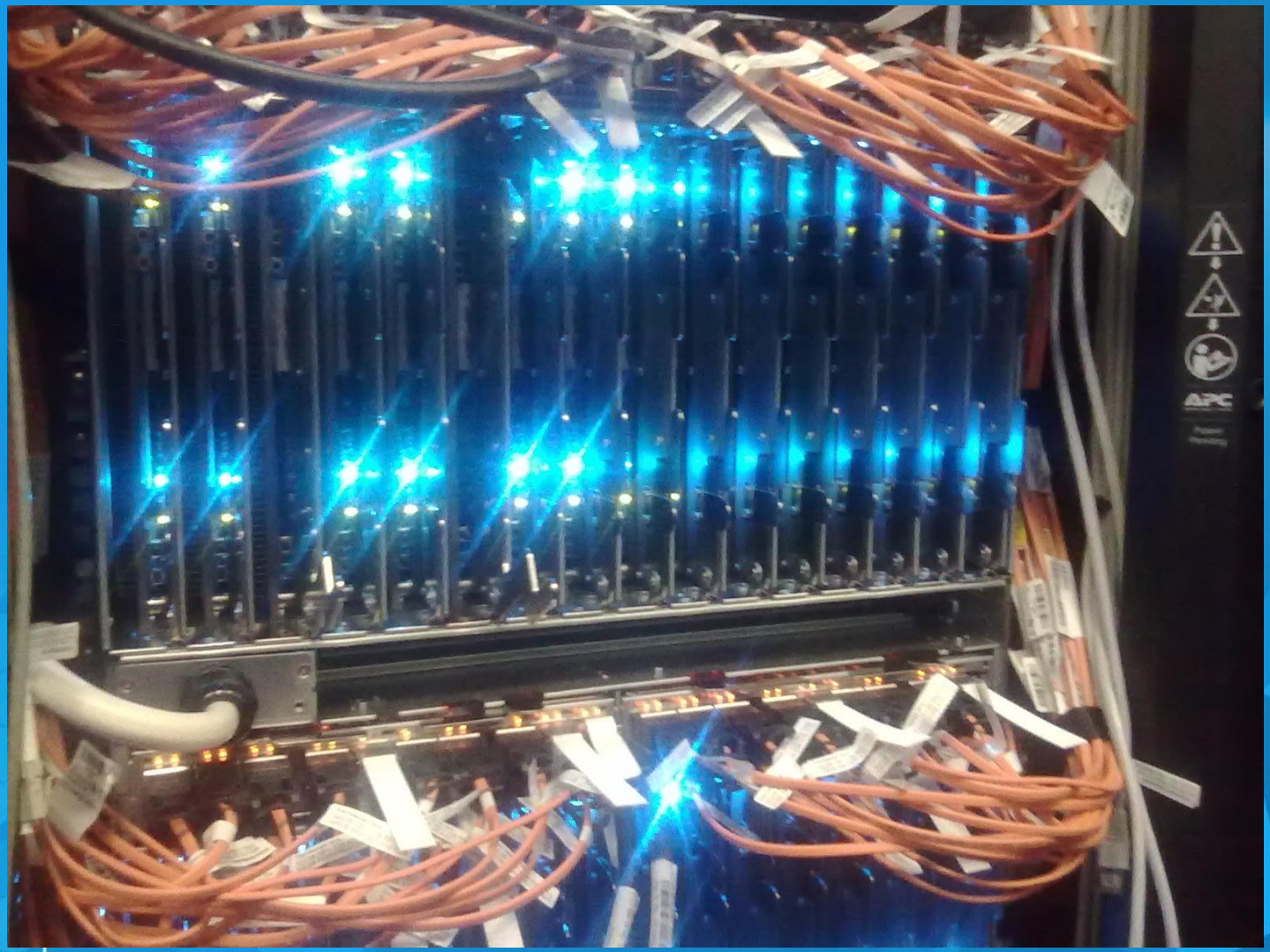
T-BLADE 2 node
Logical scheme



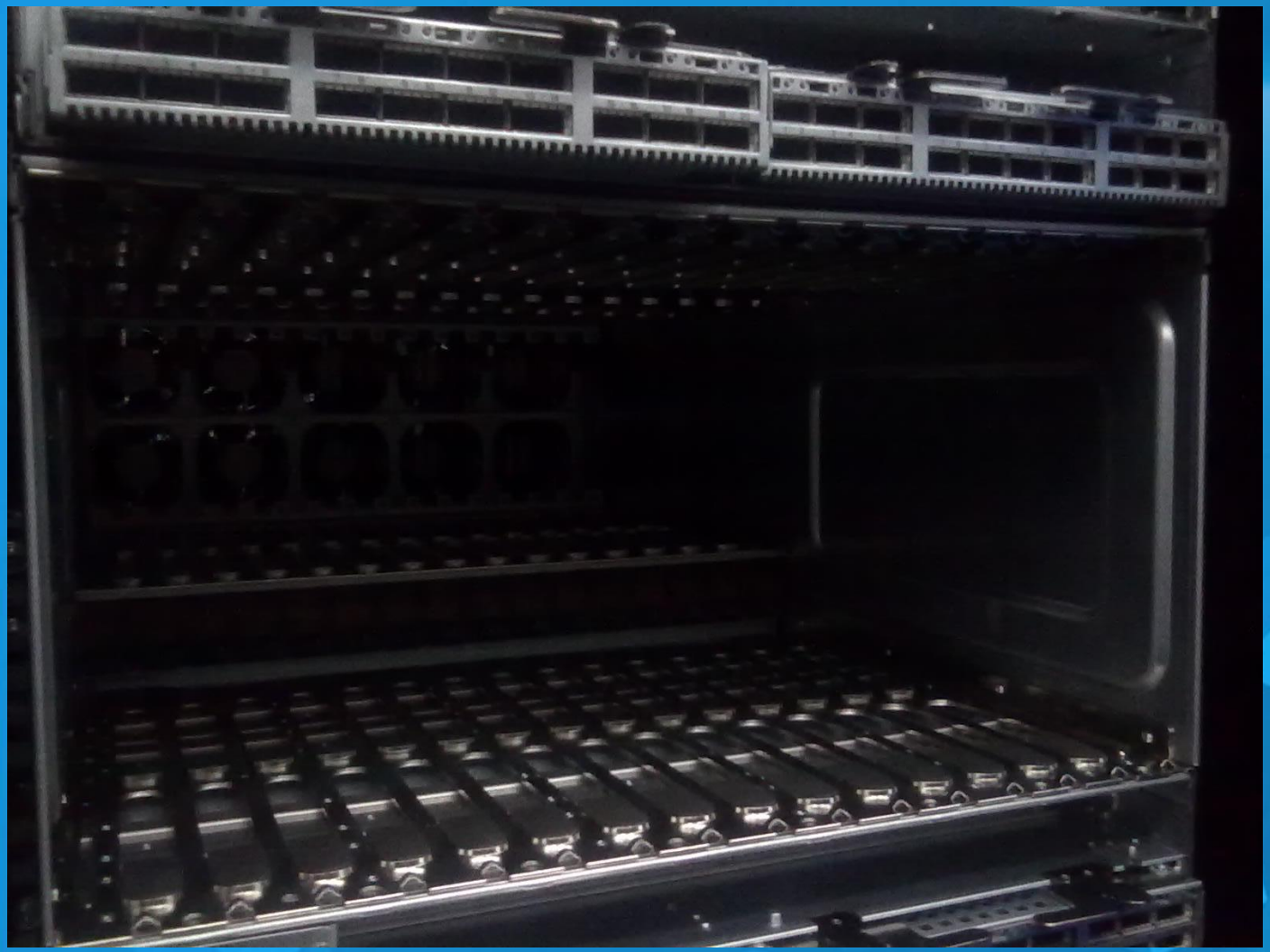
T-BLADE 2
PCBs

T-PLATFORMS

T-BLADE 2
Working enclosure



T-BLADE 2
Empty enclosure



Rack distribution in Lomonosov supercomputer

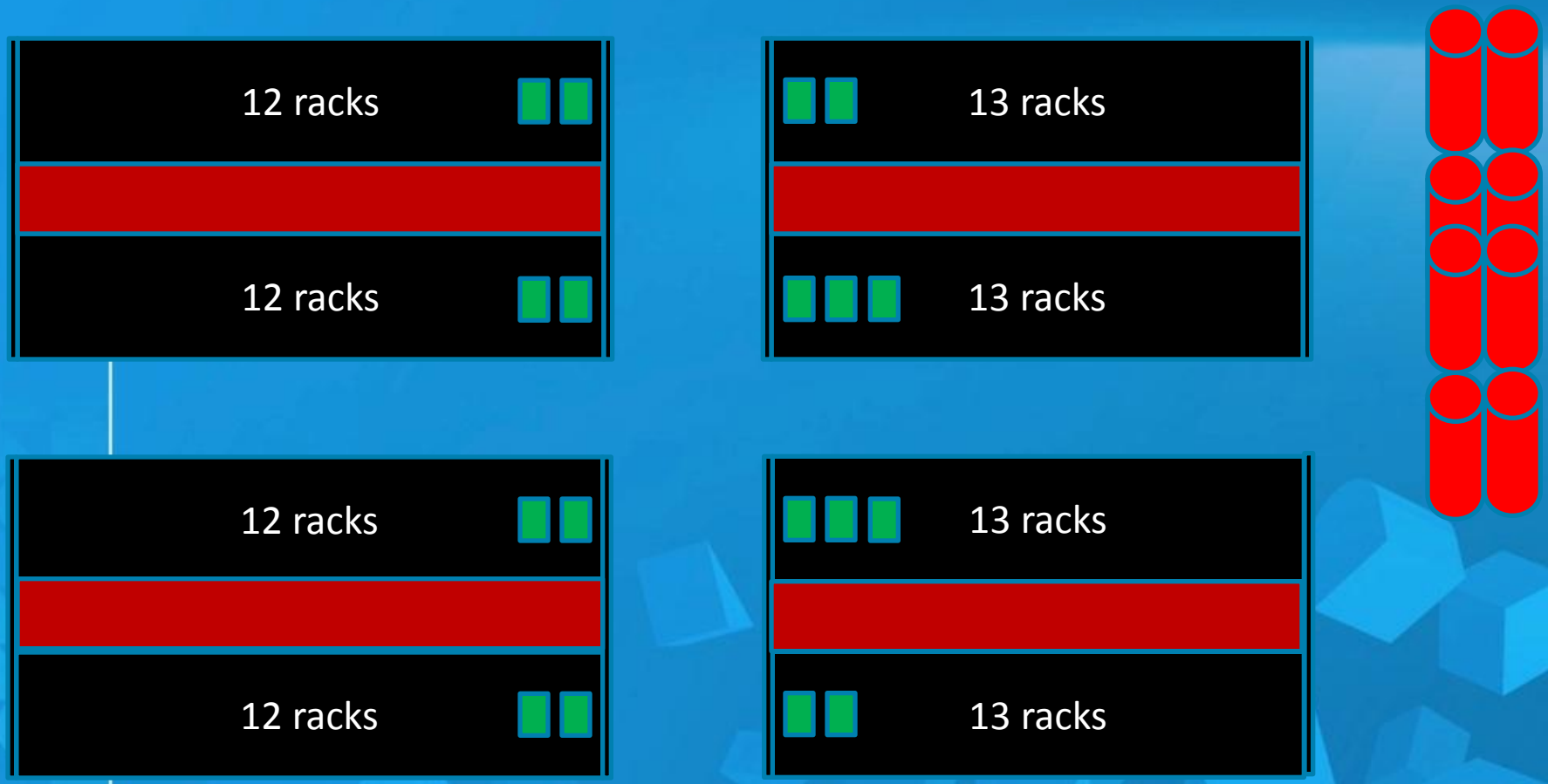
Rack distribution

x86+GPU -- 30+11
Infiniband switches – 18
Storage – 8
Management/Service/misc – 3

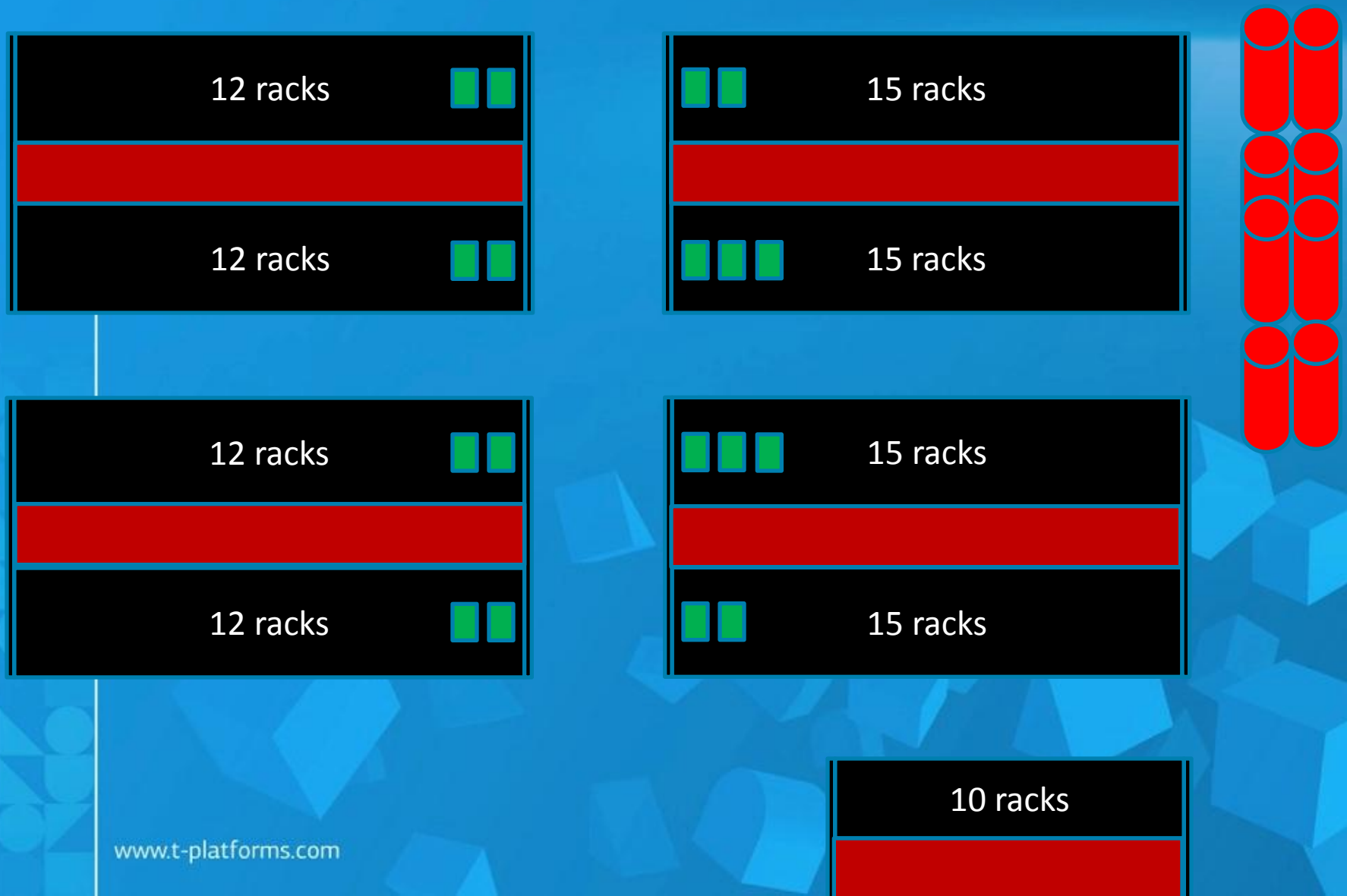
x86 rack (30pc) (42U)

5 enclosures TBlade2-XN (2CPU+ 12GB RAM)
1 enclosure TBlade1.1 (2CPU+24GB+2 HDD)
1 dual Cell BE server

Hot aisles, 100 racks



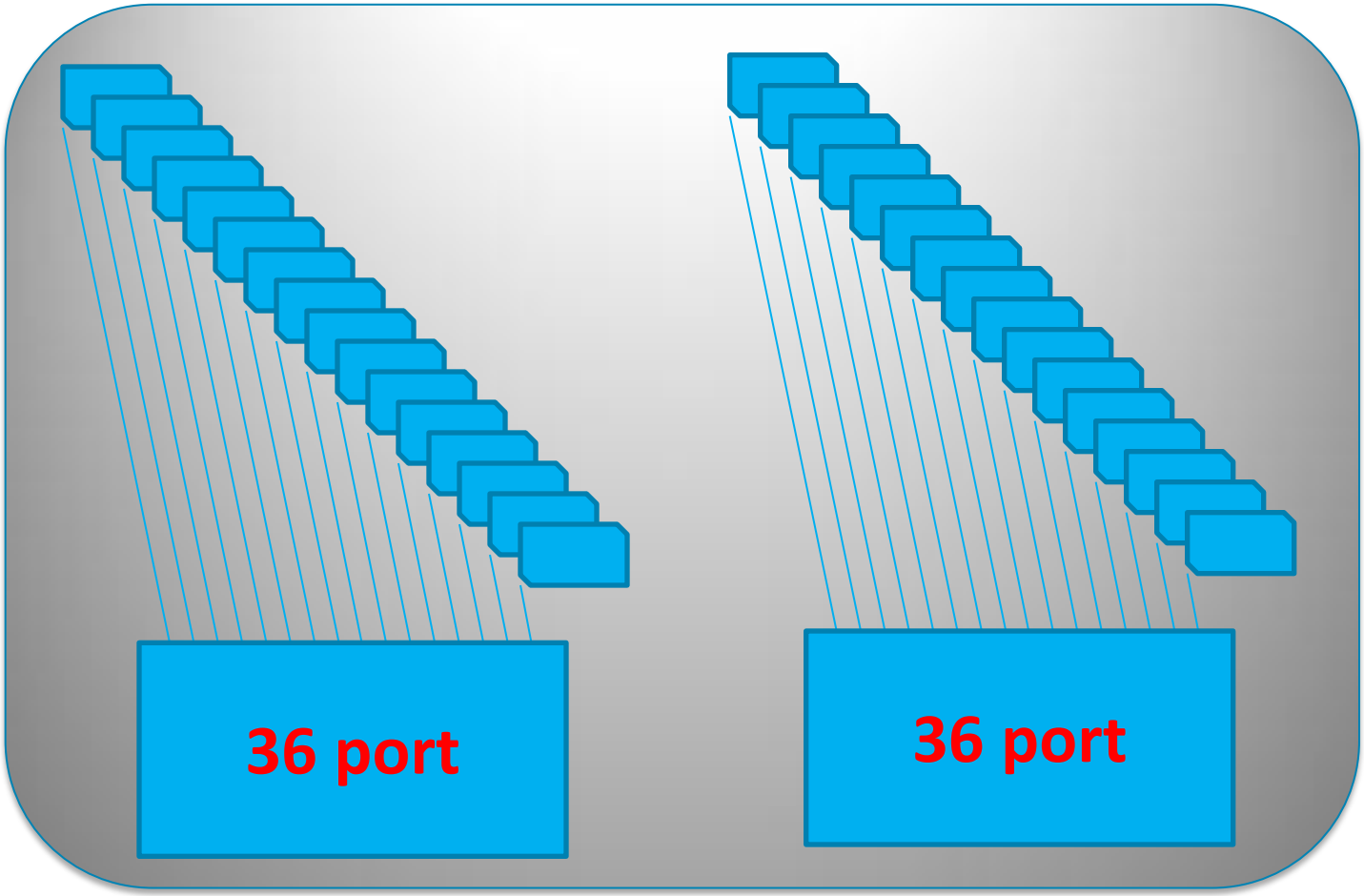
Hot aisles, 118 racks



Networks in Lomonosov

- System network(IB)
- Service network (Eth100)
- Management network (Eth1000)
- Custom barrier network
- Global interrupt network

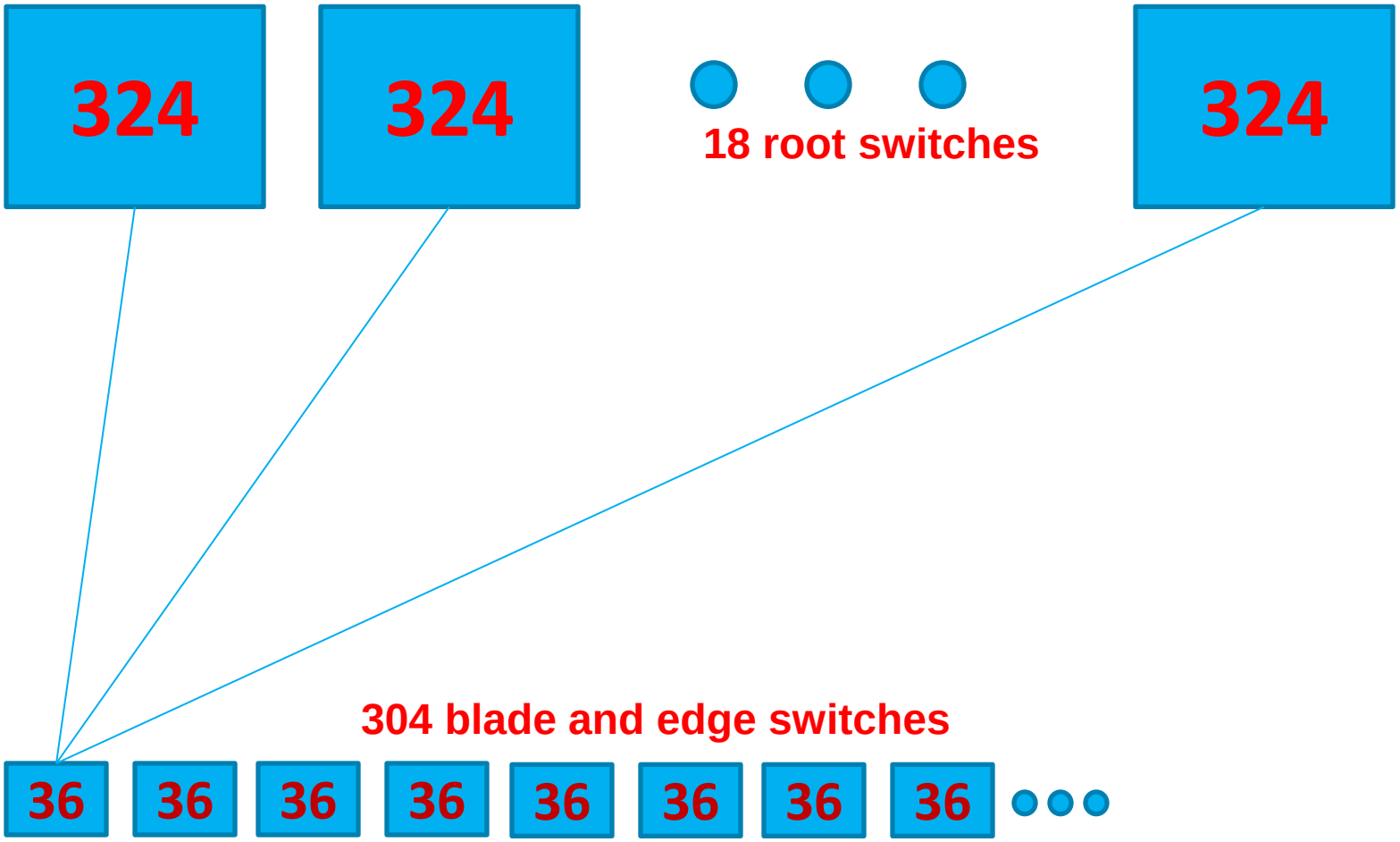
System network in enclosure



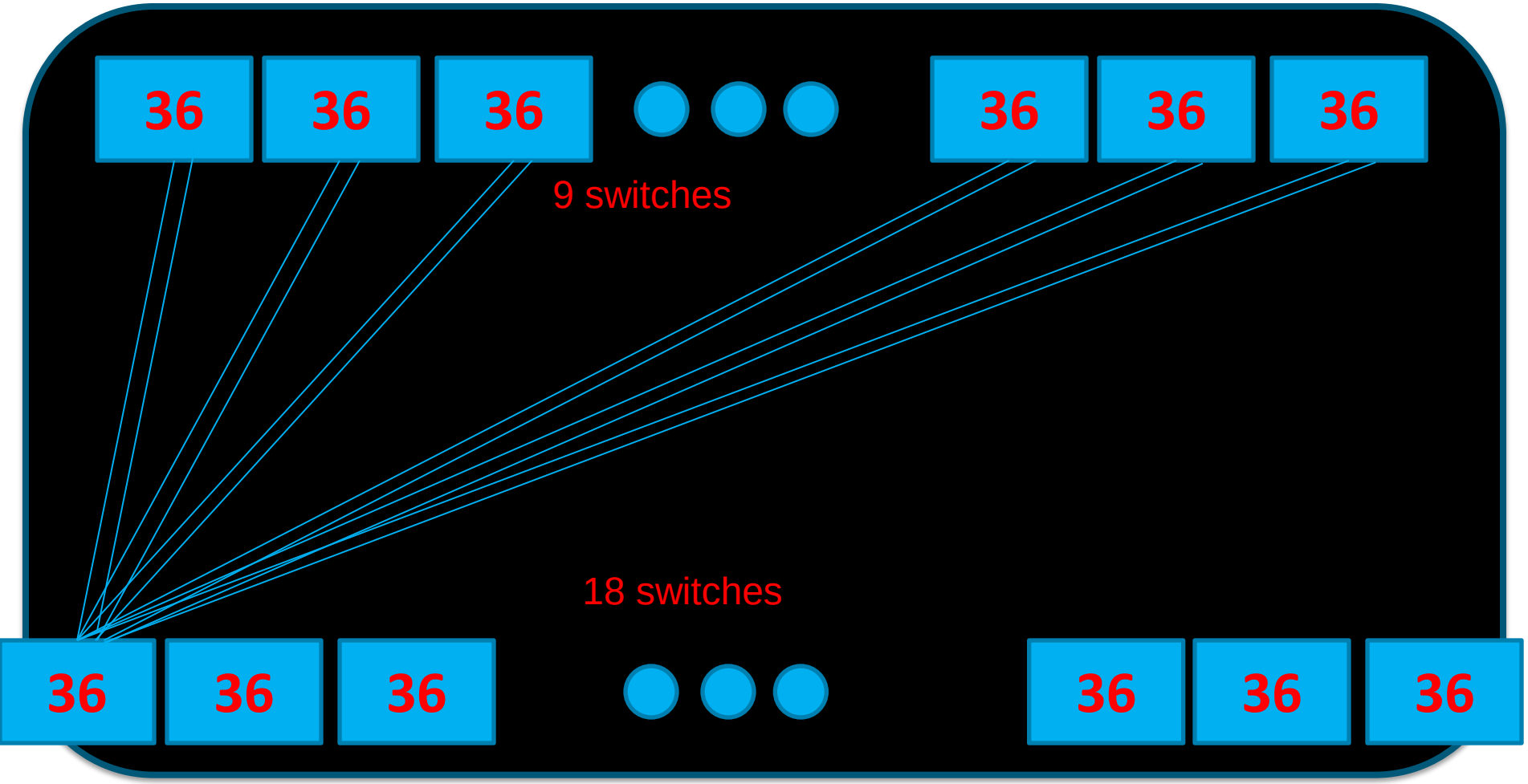
20 external ports

20 external ports

Topology


324
324
18 root switches
324
304 blade and edge switches
36
36
36
36
36
36
36
36

Root 324-port switch



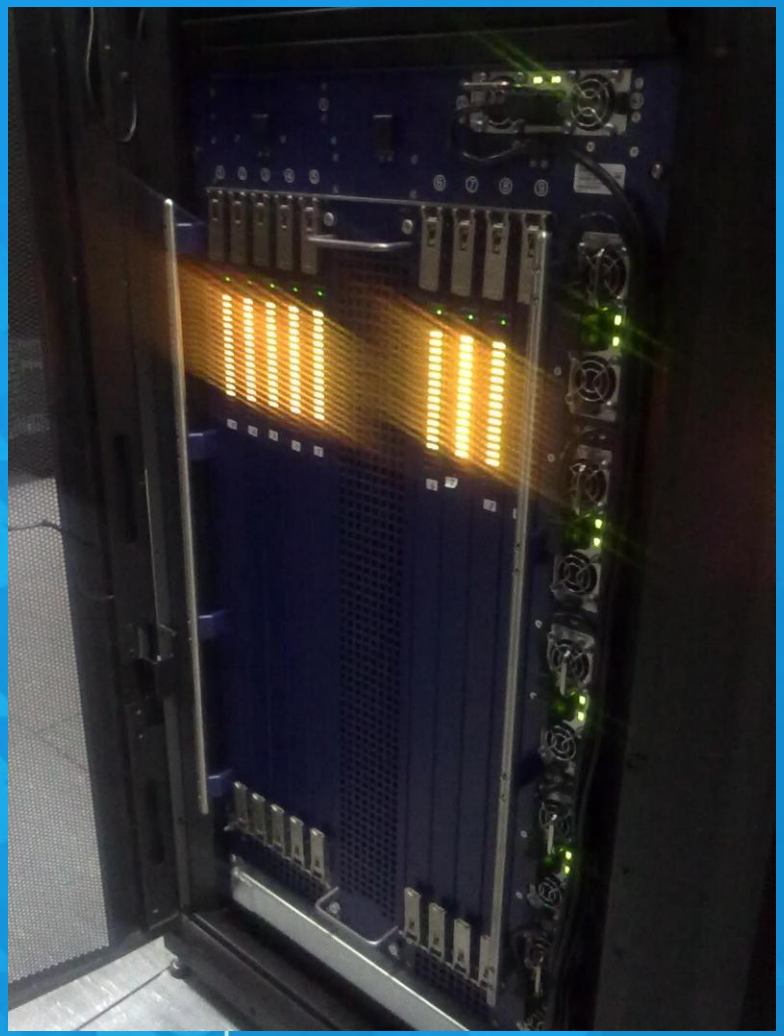
18x18=324 external ports

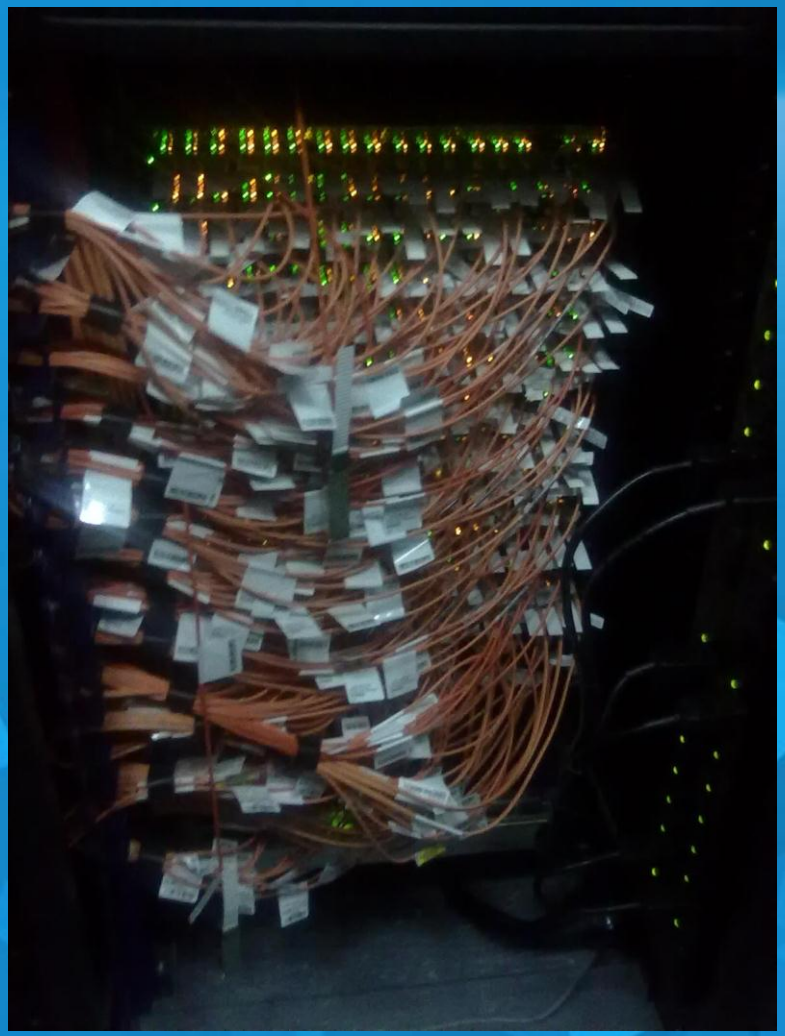
System network

- Root 324-port switches (18 pc)
  - Consists of 18+9 36-port switch ASICs
- 36-port switches in enclosures
- 4 additional edge 36-port switches
- Intrarack cables: copper
- Interrack cables: fiber

Root switches

T-PLATFORMS





Upgrade phases, Lomonosov

1. T500 (414 TF, 2009)
2. T500+ (510 TF, 2010)
3. T1000 (1370 TF, 2011)
4. T1000+(1700 TF, 2012)

GPU upgrade

Announcing New Tesla Systems

IBM BladeCenter

#1 HPC Provider
196 of Top 500

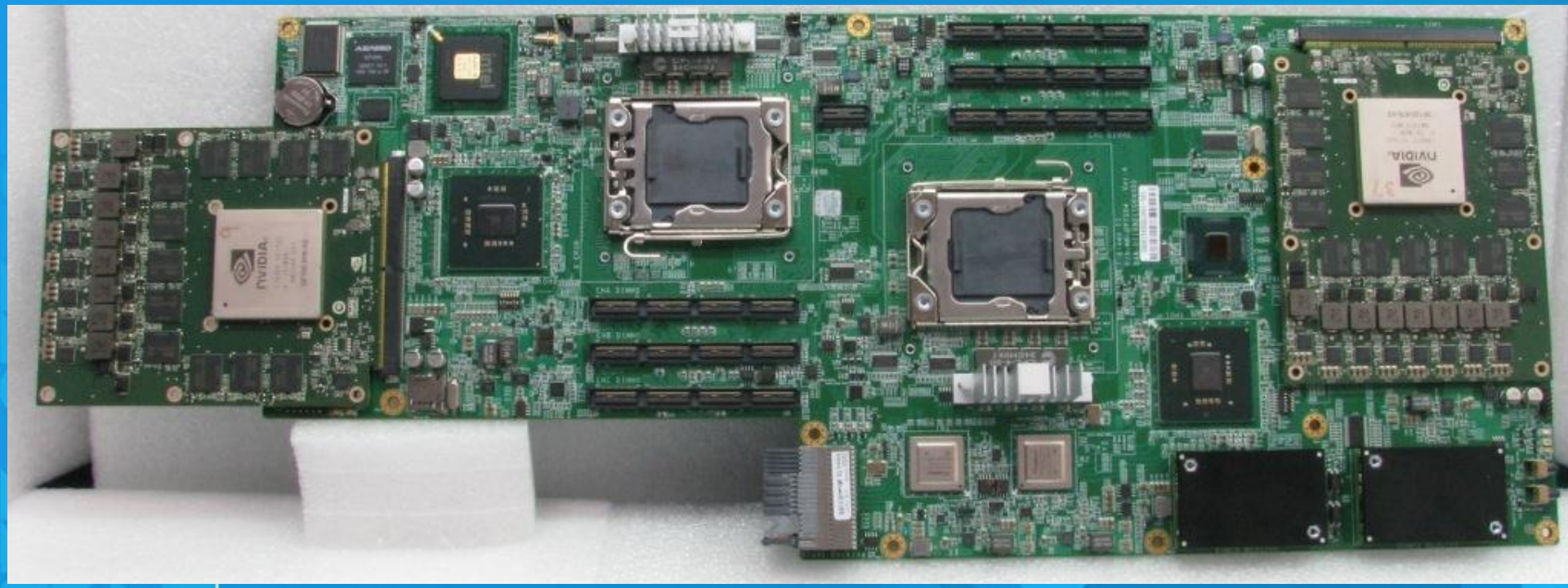
T-Platforms TB2

Lead HPC Provider in Russia
50% of FLOPS of Russia Top 50

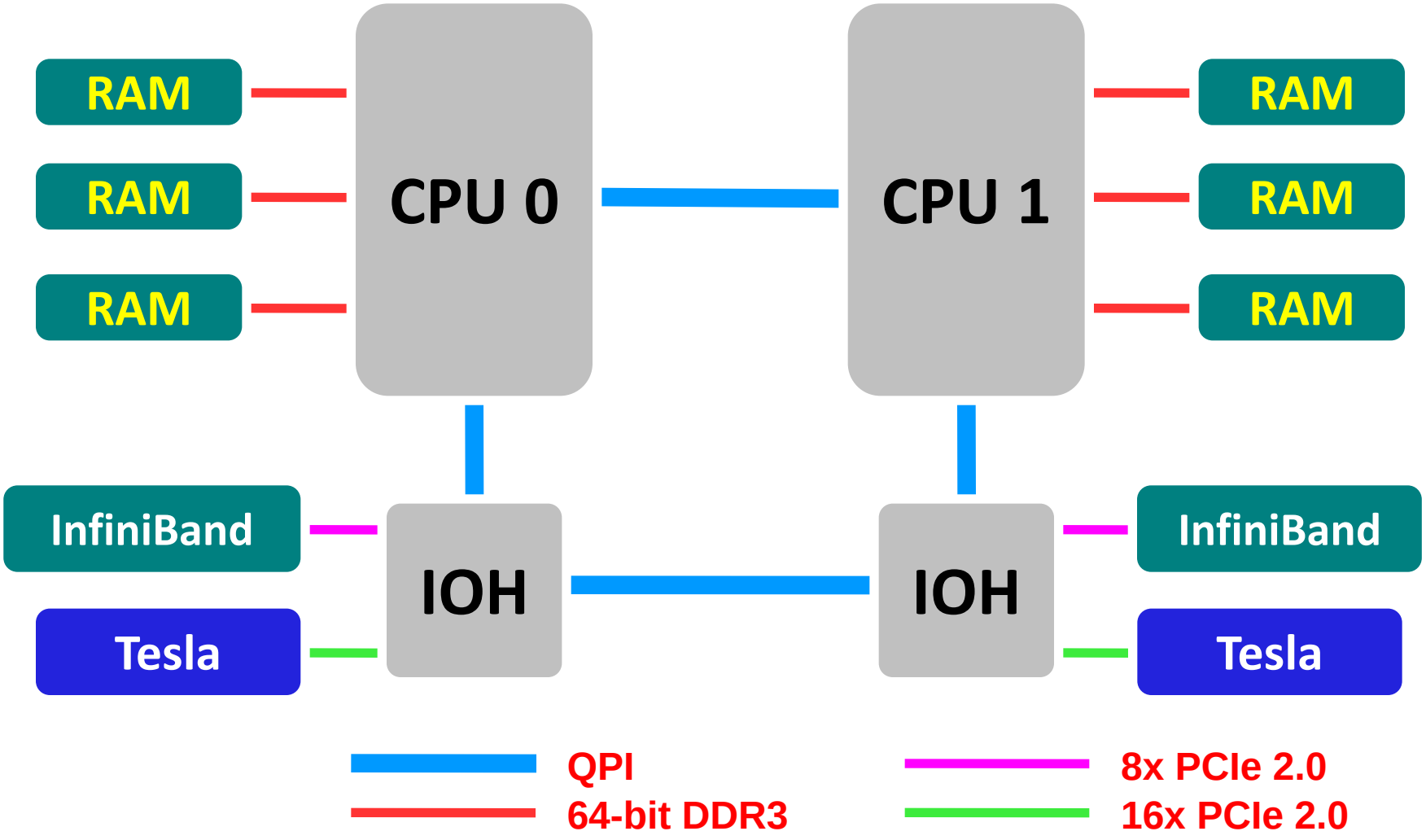
Cray XE6

The Supercomputing Co.
10 of Top 50

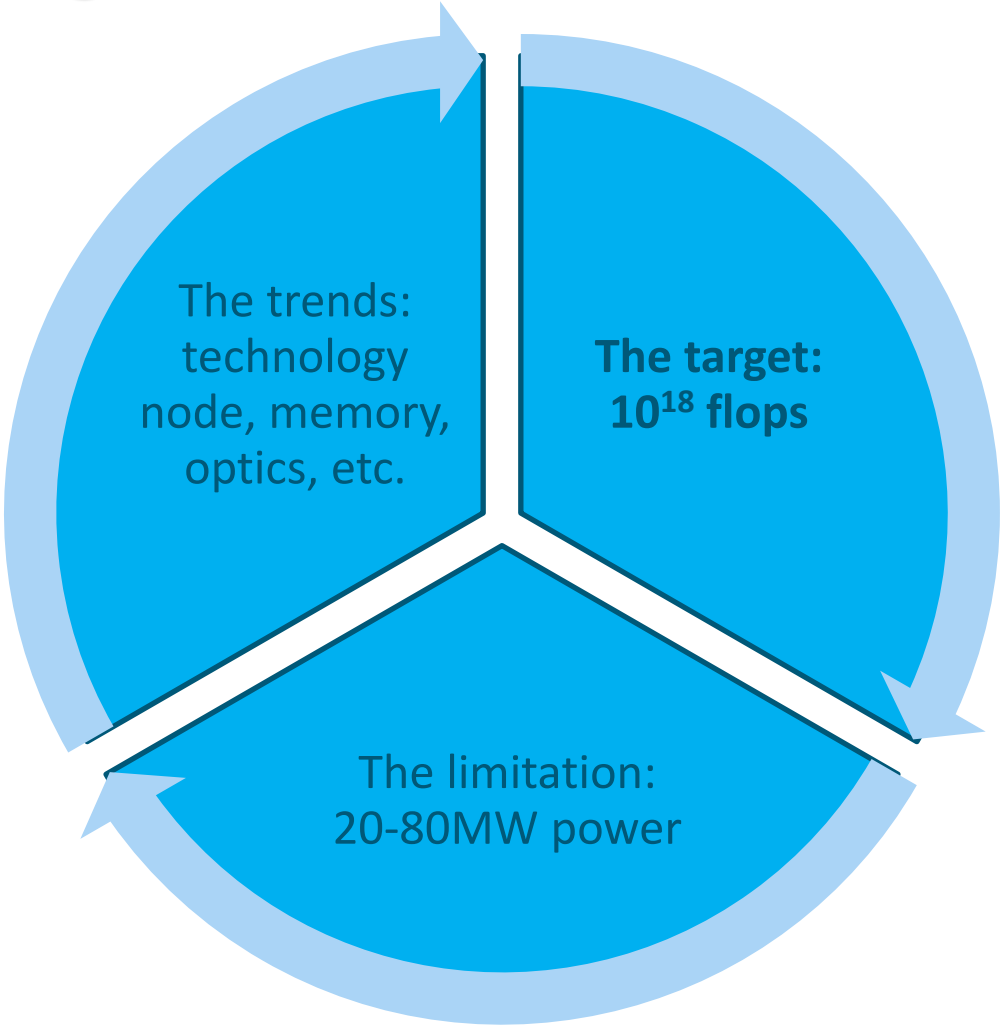
TB2-TL



TB2-TL logical scheme



The Exascale Challenge: Architecturing the Future



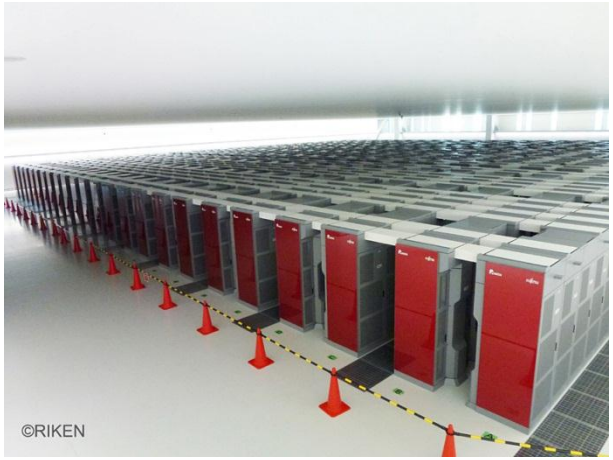
Trends and Requirements

System Peak [PF]	1000
Power [MW]	20-80
System memory [PB]	32-64
GB RAM/Core	0.1-0.5
Node Performance [GF]	1000-10000
Cores/Node	1000-10000
Node Memory BW [GB/s]	400-4000
Number of nodes	100000-1000000
Total concurrency	$O(10^9)$
MTTI	O(1 day)

Source: EESI Final Report, November 2011

Part 1: System at a Glance

- Few example of existing systems:
- K Computer:
  - Specially designed building
  - 50x60m room
  - About 900 racks
  - Do we want such monster for exascale?
- BlueGene/Q
  - 318m^2
  - 96 racks
  - Looks much better…



Part 1: System at a Glance

- Why we need to increase the density and reduce the system area?
- Main reason: Interconnect
  - Topology and cabling
  - Latency
  - Power consumption
- The latency problem
  - 1m ≈ 3.3ns delay
  - Let's assume that point-to-point latency between two adjacent topology nodes is about 300-500ns (which is reasonable)
  - Then 100m cable => 67-100% latency increase
  - Especially critical for low-diameter topologies

Part 1: System at a Glance

- Topology and cabling problem
  - N-dimensional torus
    - Pros: easy cabling; short cable lengths
    - Cons: only two dimensions may have big benefit from racks grid; the network diameter is huge; low bisectional bandwidth
  - N-dimensional flattened butterfly
    - Pros: easy cabling; small diameter
    - Cons: requires more ports on the router
  - Dragonfly
    - Pros: low diameter; high bisectional bandwidth
    - Cons: difficult cabling; long cable distances
- Power problem
  - Shorter cables may require lower power transceivers

Part 1: System at a Glance

- The limitation is the rack power consumption
- Let’s assume 50MW/EF
- The reasonable configuration may look like the following:
  - 256 racks
  - 16x16 rack grid
  - Approximately 200KW per rack
  - Maximum X-Y distance the nodes is about 50-60m
  - 3.9PF rack performance

Part 2: The Rack

- Compute node vs. topology node
- Modularity
- Intra-rack topology

- BlueGene/Q: 1024 nodes per rack, already very dense
- With better integration 2048 compute nodes is feasible, but more is unlikely
- Topology nodes:
  - Using the torus, 1 compute node = 1 topology node (router is integrated into the processor)
  - For other topologies the number of ports is the limiting factor, 4-8 nodes per high-radix router looks feasible, router is a separate chip

Part 2: The Rack

- From the mechanical prospective more than 256 replaceable units looks like a maximum for a reasonable size cabinet
- Water cooling is assumed for 200KW
- For advanced topologies like dragonfly 1 rack = 1 group in fully connected graph

Part 3: The Unit

- With 256 units per rack we have:
  - 15.26TF performance
  - 1KW power consumption
  - Up to 8 compute nodes per unit

- Some technology assumptions:
  - We assume that the compute node has memory and interconnect integrated into a single package, non-volatile memory is separate
  - We assume that non-volatile memory may consume up to 20% of overall power budget
  - We assume that link speed (single lane) will be at least 40Gbps

- Then we may have the following variants (next slides)

Part 3: The Unit

- Variant A
  - 2TF per node with about 100W power budget, torus topology, router is integrated into the package (10x performance comparing with BG/Q)
  - BG/Q has 2GB/s bandwidth per link, 10 x 2GB/s = 20GB/s = 160Gbps = 4 x 40Gbps links
  - Number of nodes will be 512K comparing to 96K in BG/Q, but (comparing with BG/Q) we may extend the 5th dimension or add yet another one

Part 3: The Unit

- Variant B
  - 2-4TF per compute node and one high-radix router per 4-8 nodes
  - Power budget for high-radix router chip(s) in a range of 100-200W
  - Node power budget is about 75W (2TF node) or 150W (4TF node)
  - 48-64 ports on a router
  - Intra-rack topology: flattened butterfly (2-tier 16x16 with 256 router chips and 3-tier 8x8x8 with 512 router chips)
  - Inter-rack topology: dragonfly

Summary

- Possible high-level architectures of potential exascale system are described
- Does it look feasible?
- Yes, with the current roadmaps of heterogeneous architectures development we will likely achieve the necessary performance within the required power budget
- Some advanced packaging like 3D memory stacking and package integration is required, but the recent introduction of HMC technology shows that it’s not only possible, but ready for production

T-PLATFORMS

THANK YOU!