13th ANNUAL WORKSHOP 2017

TERA'S DATA NETWORK

FROM STORAGE CLUSTER TO MULTI PURPOSE IB EDR NETWORK

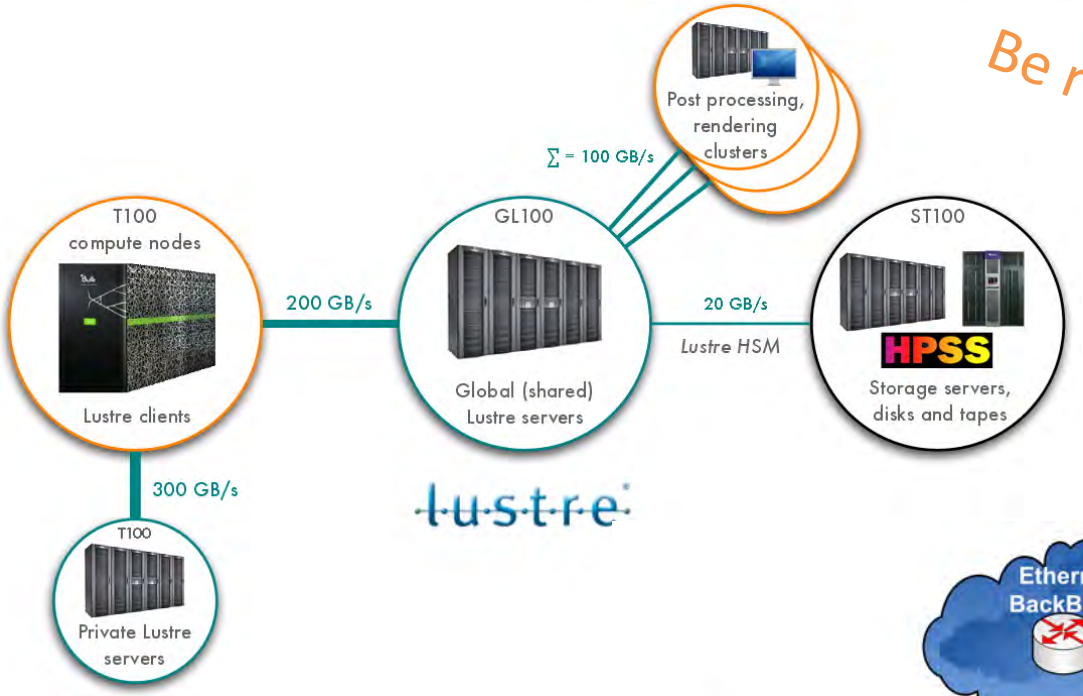
Jérôme David, Network Engineer

Commissariat à l'Energie Atomique

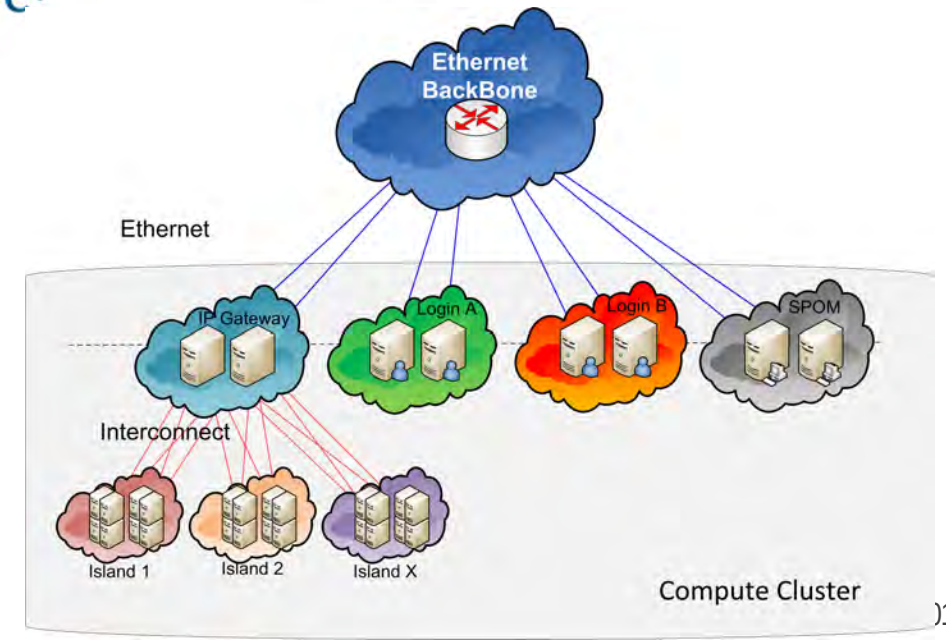
March 31, 2017

HISTORY : TERA100 STORAGE NETWORK

- **Private Lustre Storage (Scratch)**
  - 5 PB - 300 GB/s
  - Within Cluster Fabric
- **Global Lustre Storage (Store)**
  - 20 PB - 200 GB/s
  - Dedicated Fabric – Lustre Router
  - HSM Functionality with HPSS
  - Post Processing Clusters access
- **Network IB QDR**
  - Voltaire 4700 : 324 ports
  - UFM running with Tara
- **Direct Ethernet attachment**
  - Spoms Nodes
  - Login Nodes
  - Ip Gateway



Be ready for 1 Pflop !



NEEDS FOR TERA1K

- **Storage (~2 TB/s)**
  - Dedicated per cluster (4 FS)
  - Shared for all (7 clusters minimum)
  - Lustre HSM (Flash + HDD, HPSS)
- **InterCluster communication**
  - Sub cluster administration
  - Future : Login cluster
  - Future : In situ visualization
- **Hardware Standardization of « X86 Service nodes »**
  - Lustre Router
  - IO-proxy (9p)
  - IP Gateway
- **BackBone interconnection**
  - IB / Ethernet Gateway
- **Using Chassis Switches**
  - Bad history with fabric maintenance
  - Better cost / performance ratio

Now enjoy 20+ PFlops !

- ➔ Need QoS
- ➔ Need Sriov
- ➔ Need network segmentation
- ➔ Need ethernet interconnection
- ➔ Need more than 648 ports
- ➔ Need routing validation

ANSWER : MULTIPURPOSE EDR NETWORK

- **2x 648 ports chassis**
  - 36 ports for interconnection (2 leaf switches)
  - 48 leaf switches for resources
- **Lustre HSM for shared storage**
  - Flash at 1.2TB/s
  - HDD at 200 GB/s
  - HPSS at 30 GB/s
- **Dedicated storage per cluster**
  - Up to 450 GB/s per cluster
- **Hardware Standardization of « X86 Service nodes »**
  - Interface on cluster Interconnect
  - Interface on RTHP
  - No more direct Ethernet
- **IB / Ethernet Gateway**
  - 2 clusters of Mellanox SX6012
  - Up to 160 Gb/s of BackBone access
- **Extensive flow classification**
  - QoS in the whole Fabric

Mellanox TECHNOLOGIES

lustre

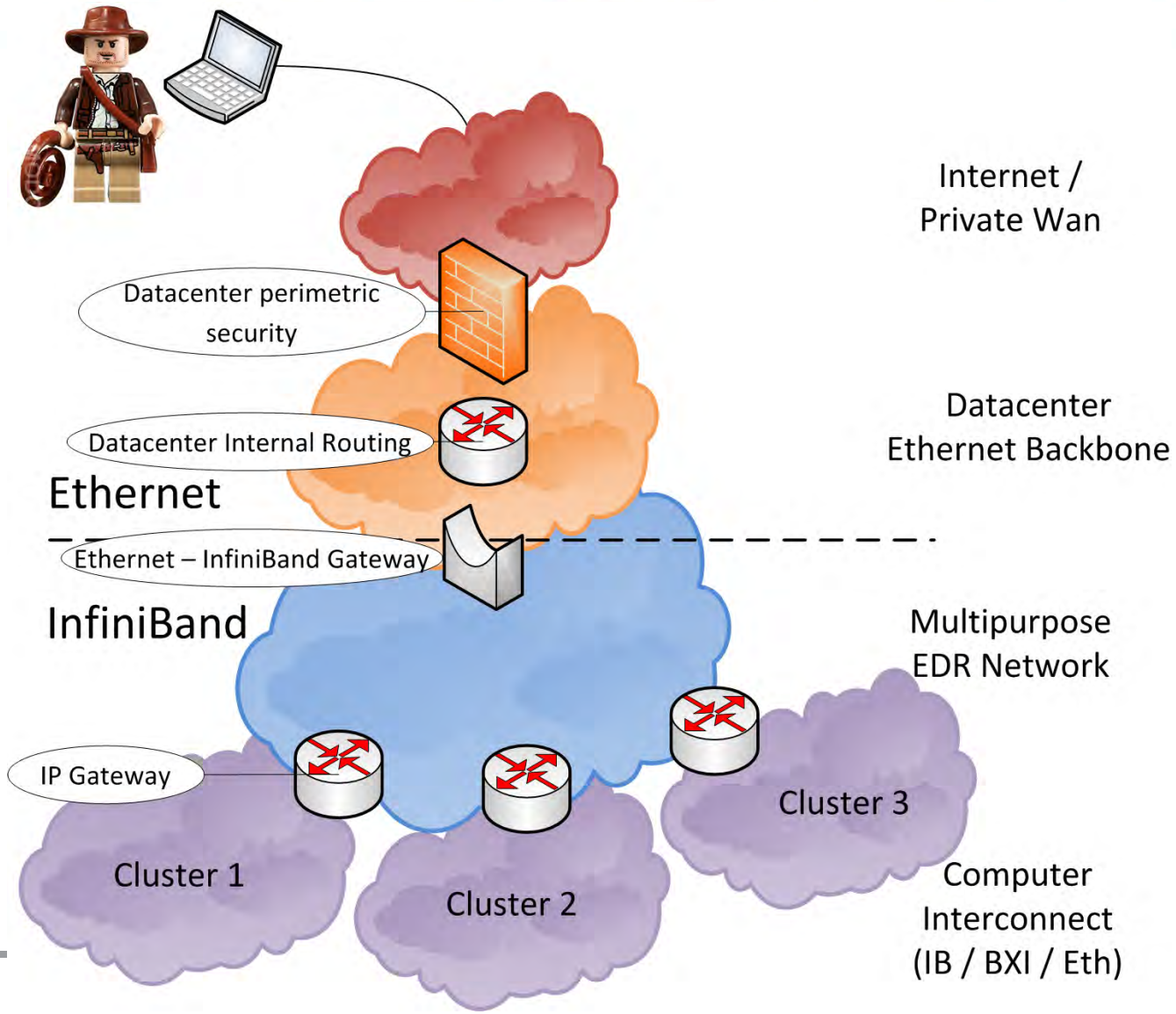
HPSS
High Performance Storage System

SwitchX-2

Award-winning acronym : RTHP
Réseau Très Hautes Performances

ANSWER : NEW NETWORK HIERARCHY

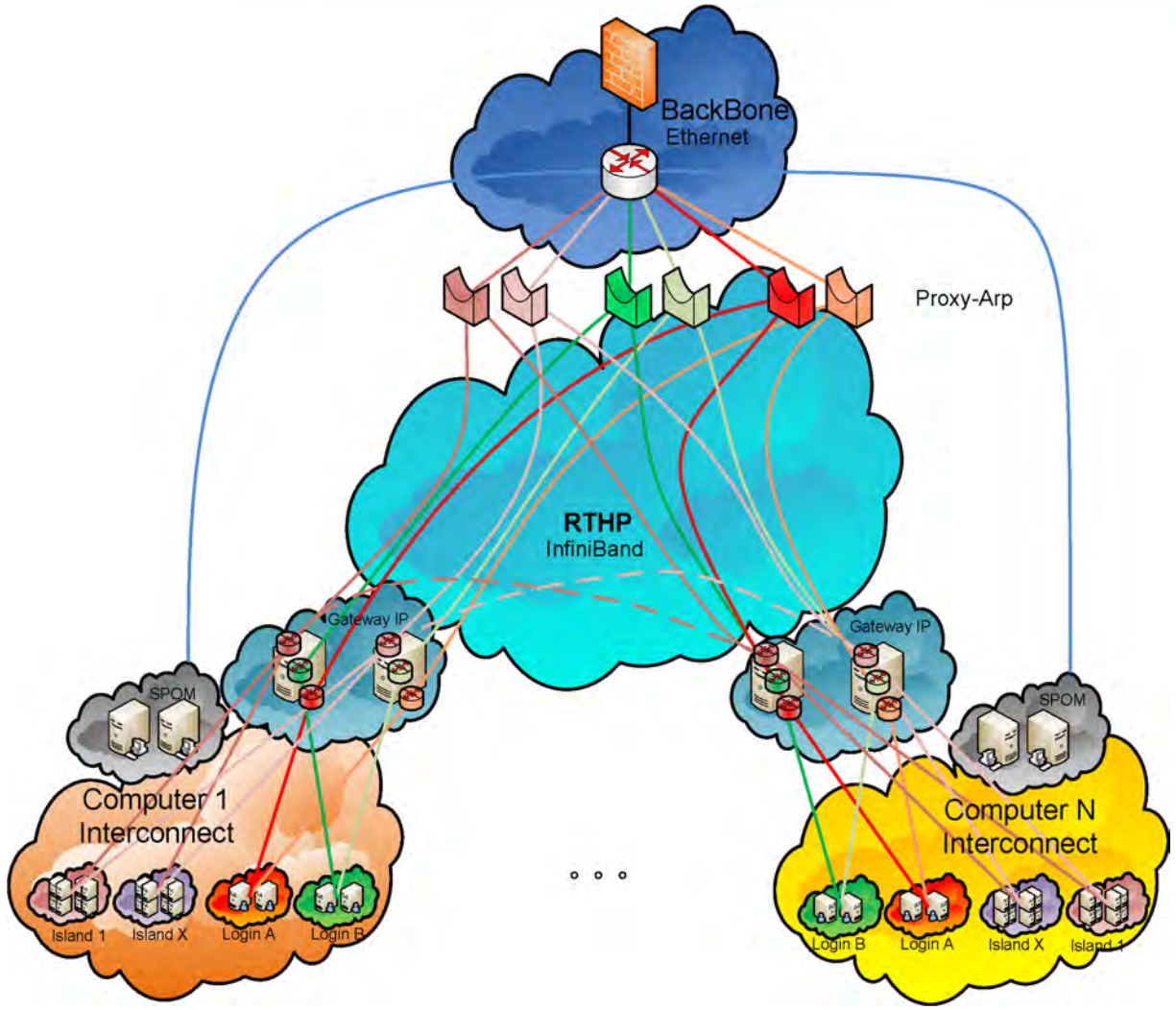
- **Taking advantage of a datacentric base…**
  - All clusters connected to the Data network
  - Shared storage
- **…to set up a new network hierarchy…**
  - Clusters are only connected to the Data network
  - Data network is connected to traditional BackBone
- **… thus enabling new functionalities**
  - BackBone access on IB
    - For compute clusters through IP gateway
    - Direct for storage cluster
    - Maintain network segmentation over IB
  - Inter-cluster communication
    - IpoIB today
    - Cluster federation
    - Login cluster perspective
    - Service cluster perspective



CONNECTING CLUSTER TO THE BACKBONE

BACKBONE INTERCONNECTION

- **Ethernet to InfiniBand Gateway**
  - Mellanox SX6012 with Gateway
  - Proxy-arp per vlan/pkey
  - Ensure Ip connectivity
- **IP Gateway on « X86 Service Node »**
  - A routing table per resource group
  - IP router running BGP with Bird
  - Multihomed BGP per resource group
- **Flow differentiation**
  - User interactive flow
  - Cluster production flow
  - Inter-cluster flow
- **Network security**
  - Traditional Ethernet/IP firewall



CONNECTING CLUSTERS TO THE STORAGE

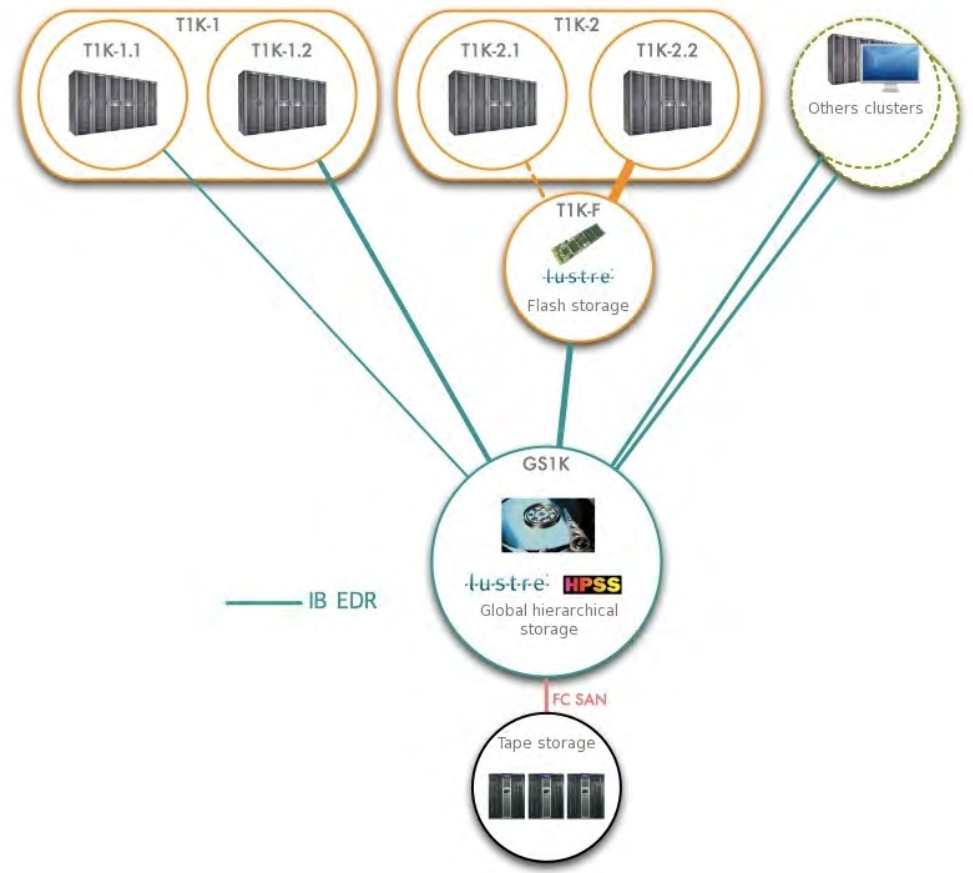
CONNECTING STORAGE

- **IO Routers on « X86 Service nodes »**
  - Lnet Router per cluster
  - ~250 Lnet routers
  - Lnet router mutualized for
    - Dedicated storage
    - Shared storage

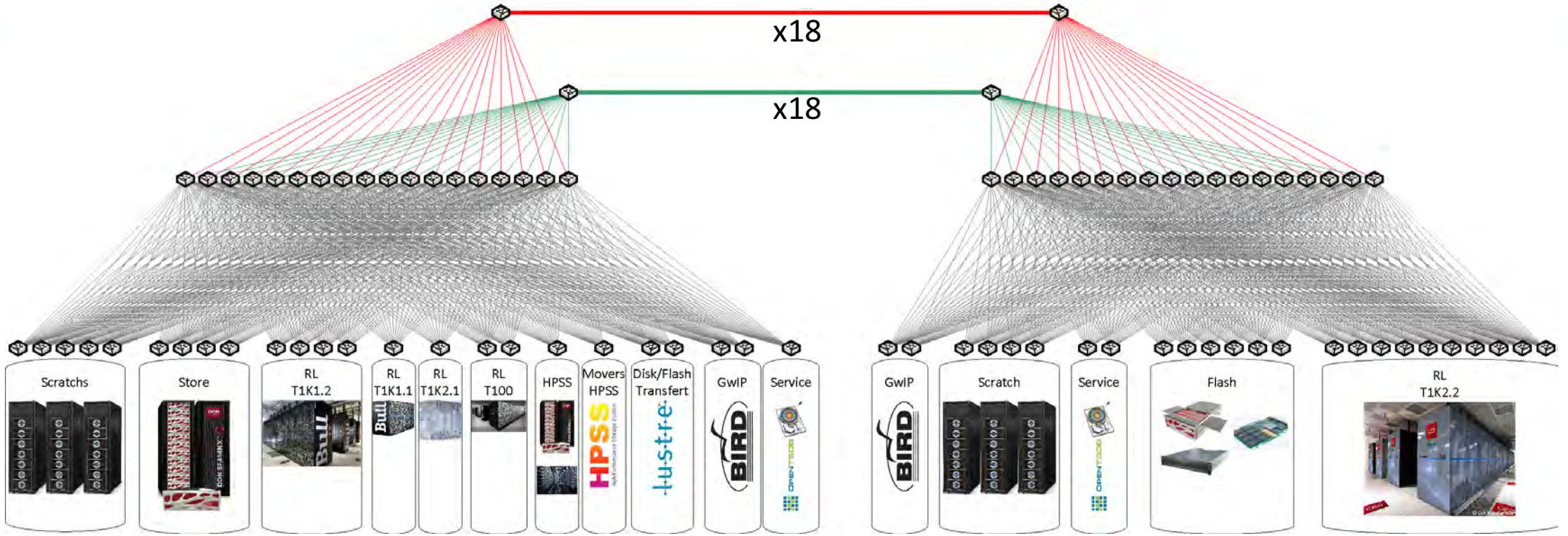
- **Transfers Nodes**
  - Lustre HSM
    - 2 levels (Flash/HDD)
  - HPSS
    - 2 levels (HDD/tape)

- **Flow differentiation**
  - Latency for MDS access
  - For Dedicated Storage
  - For Shared Storage
    - Based on Source, Destination, Partition

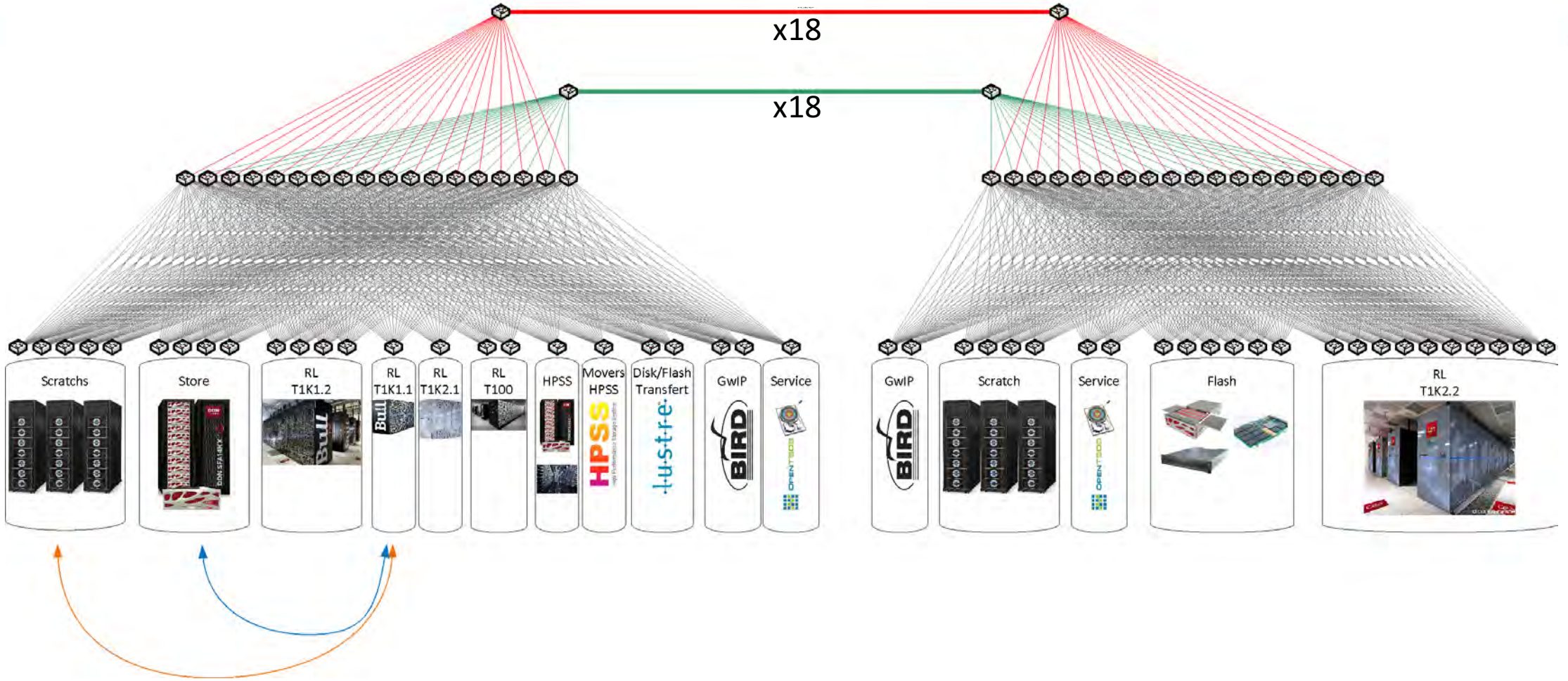
➔ Using 8 VLs



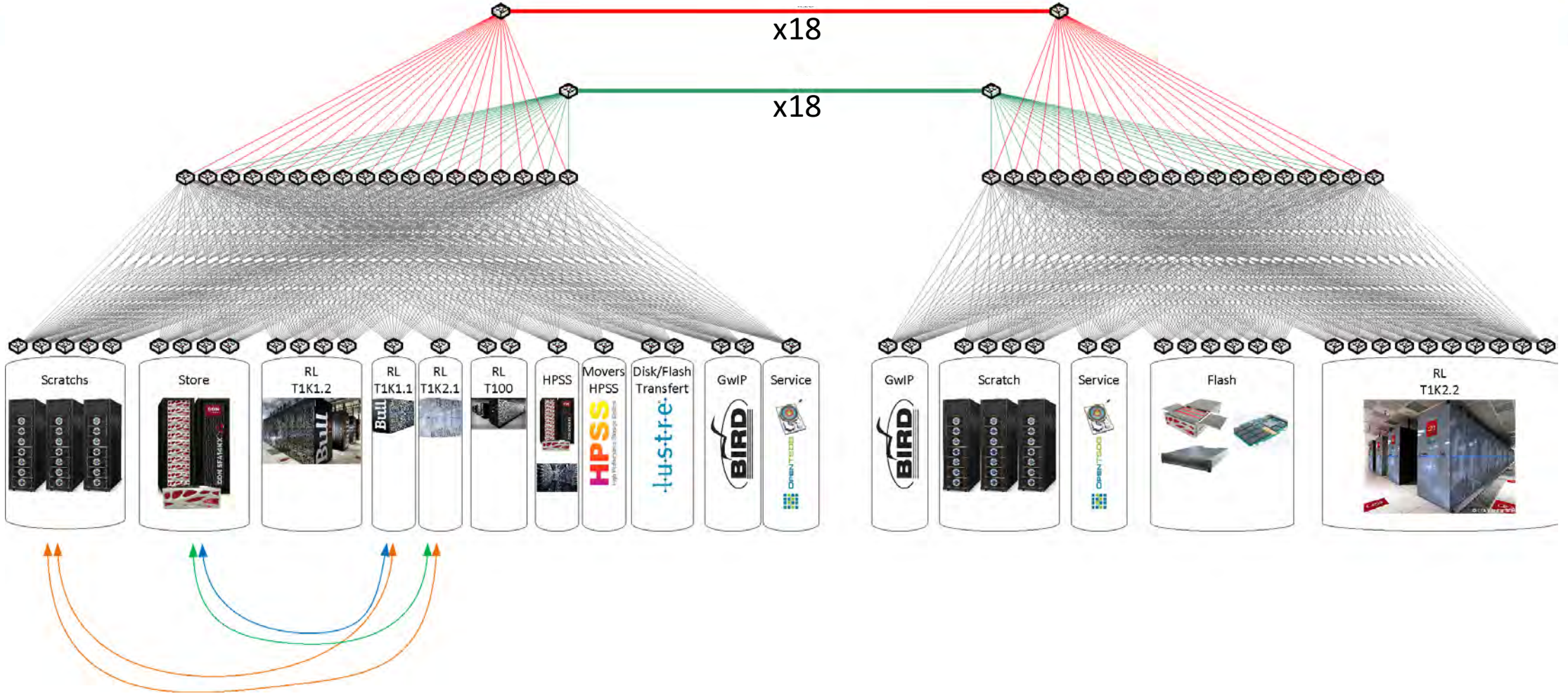
TOPOLOGY – FLOW MAP



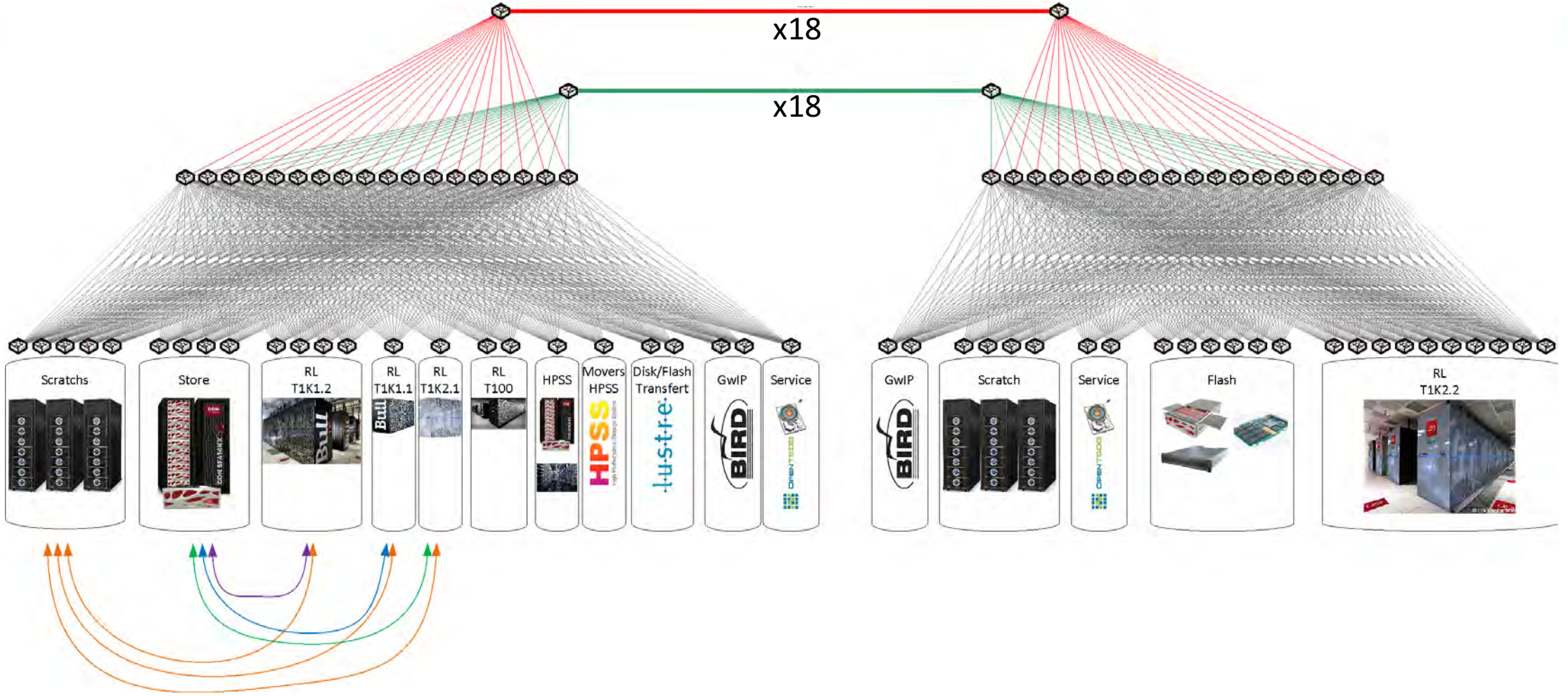
TOPOLOGY – FLOW MAP



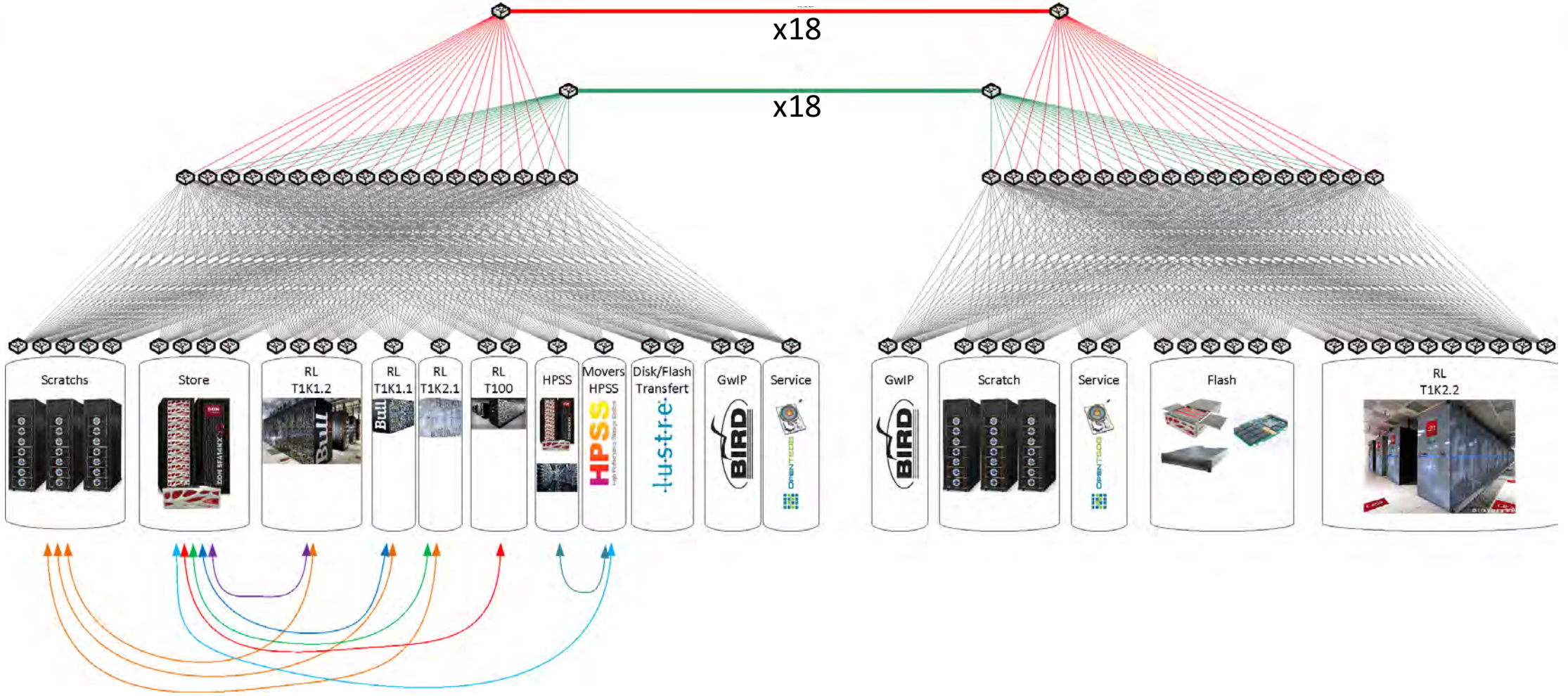
TOPOLOGY – FLOW MAP



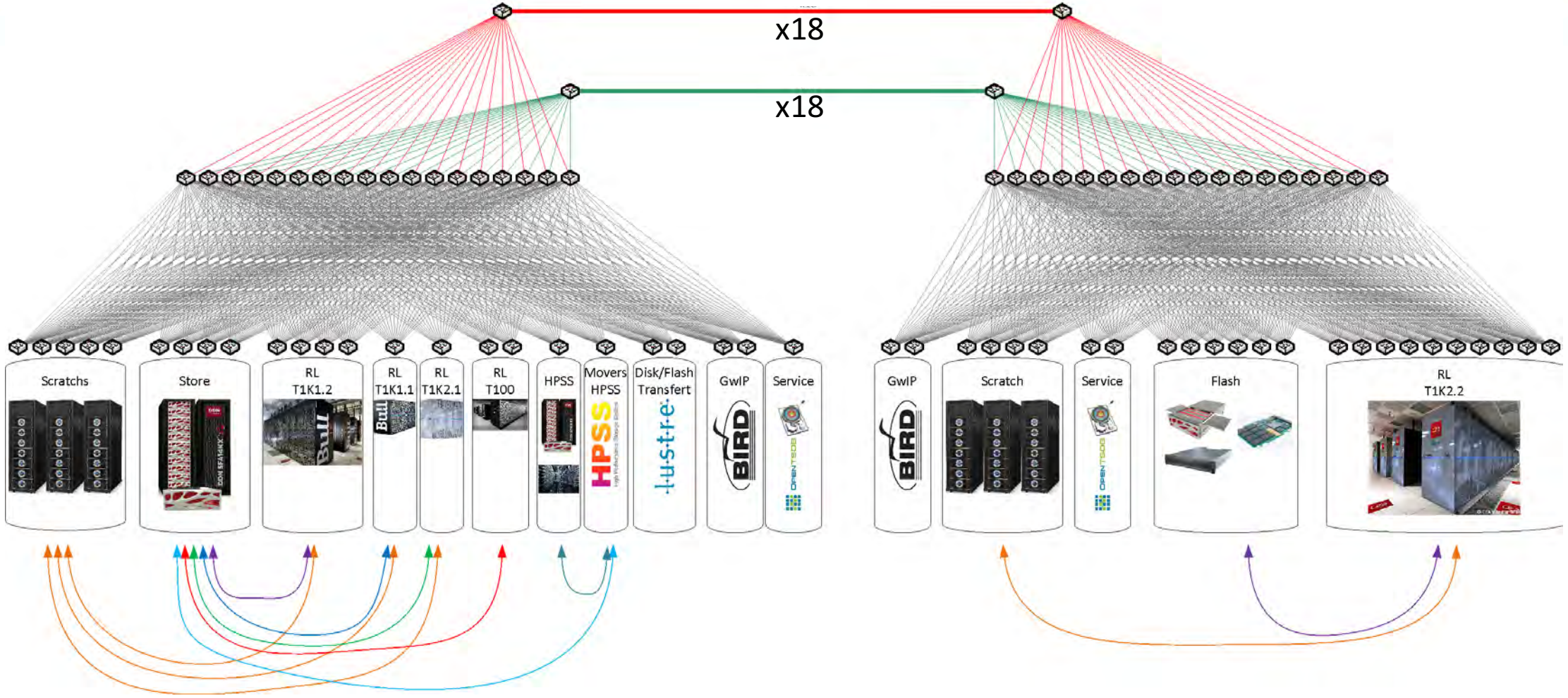
TOPOLOGY – FLOW MAP



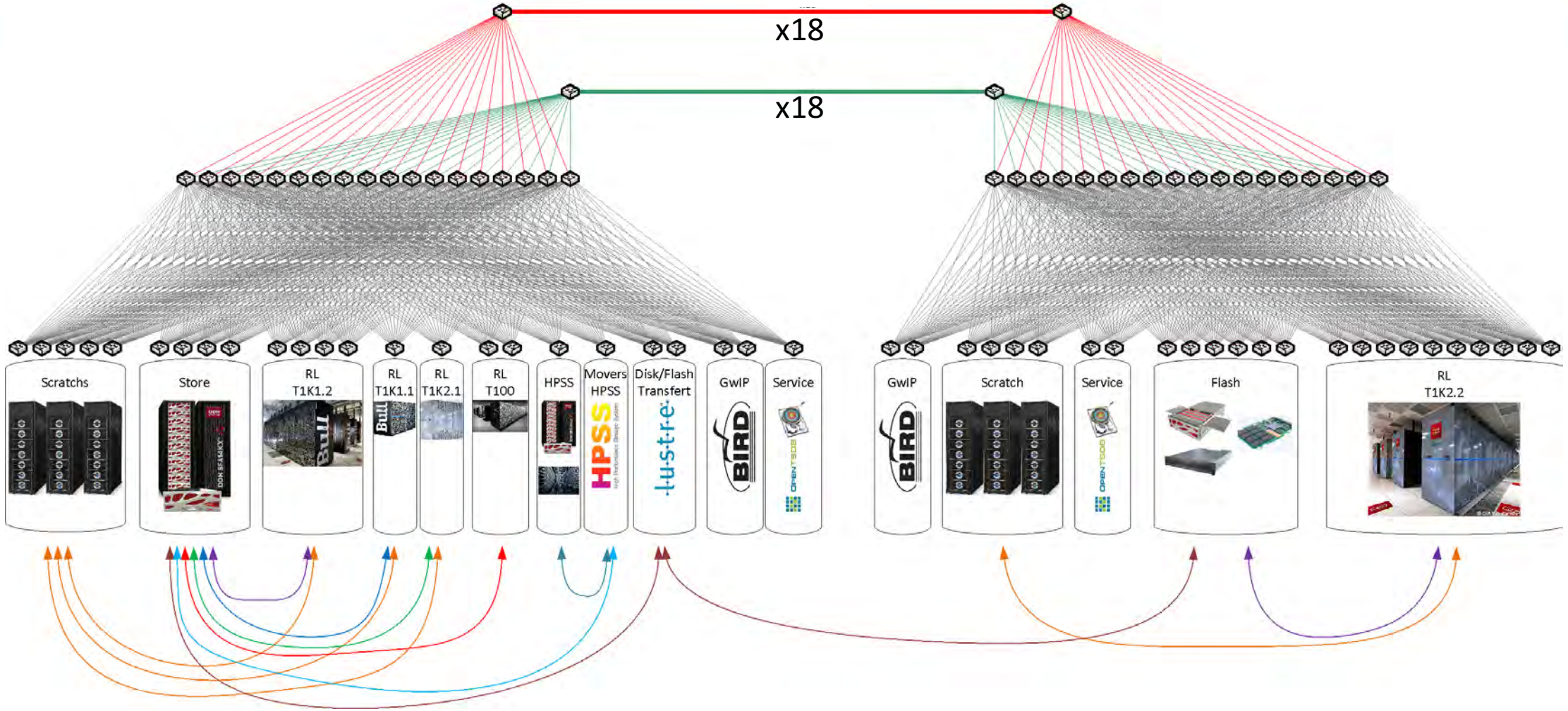
TOPOLOGY – FLOW MAP



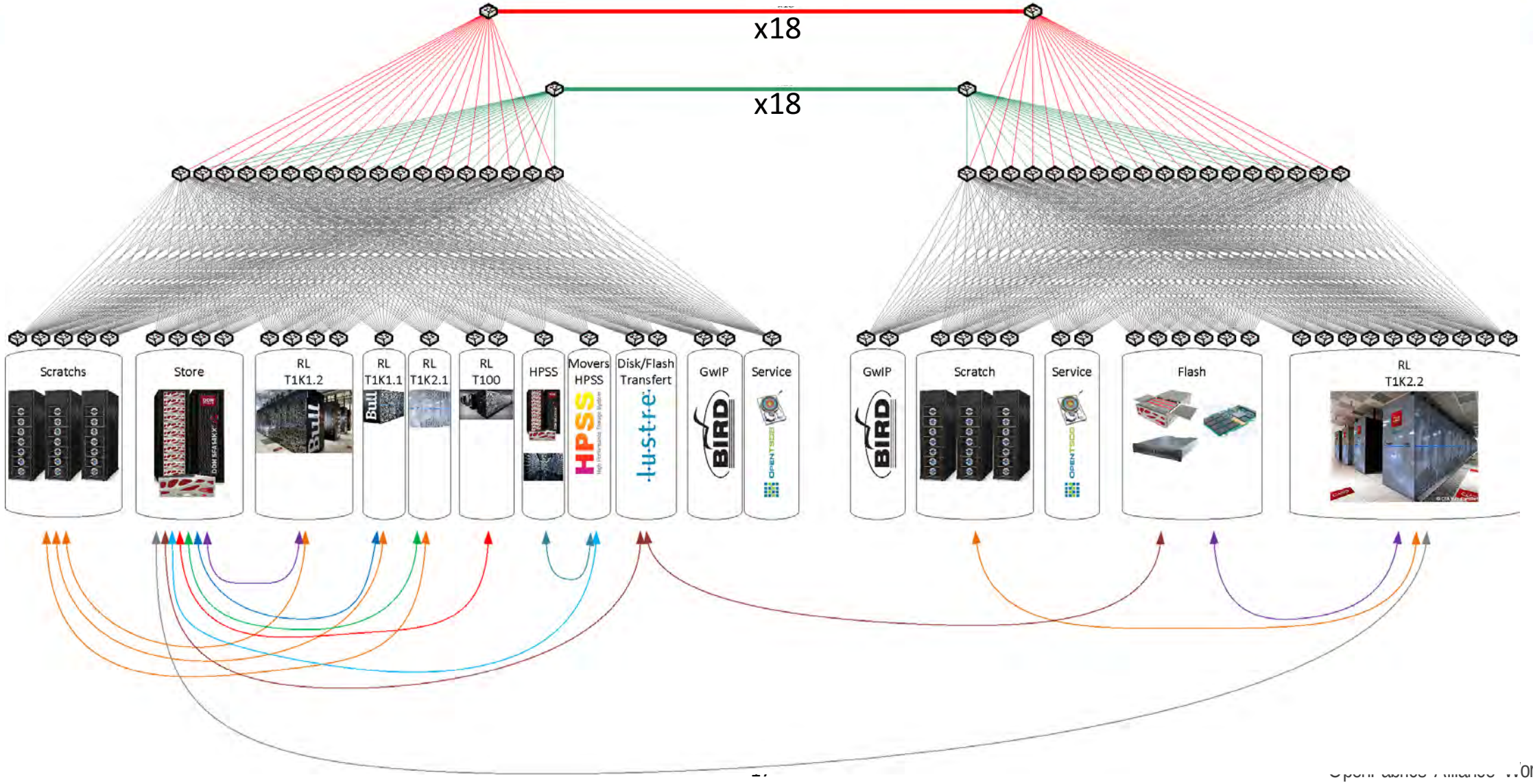
TOPOLOGY – FLOW MAP



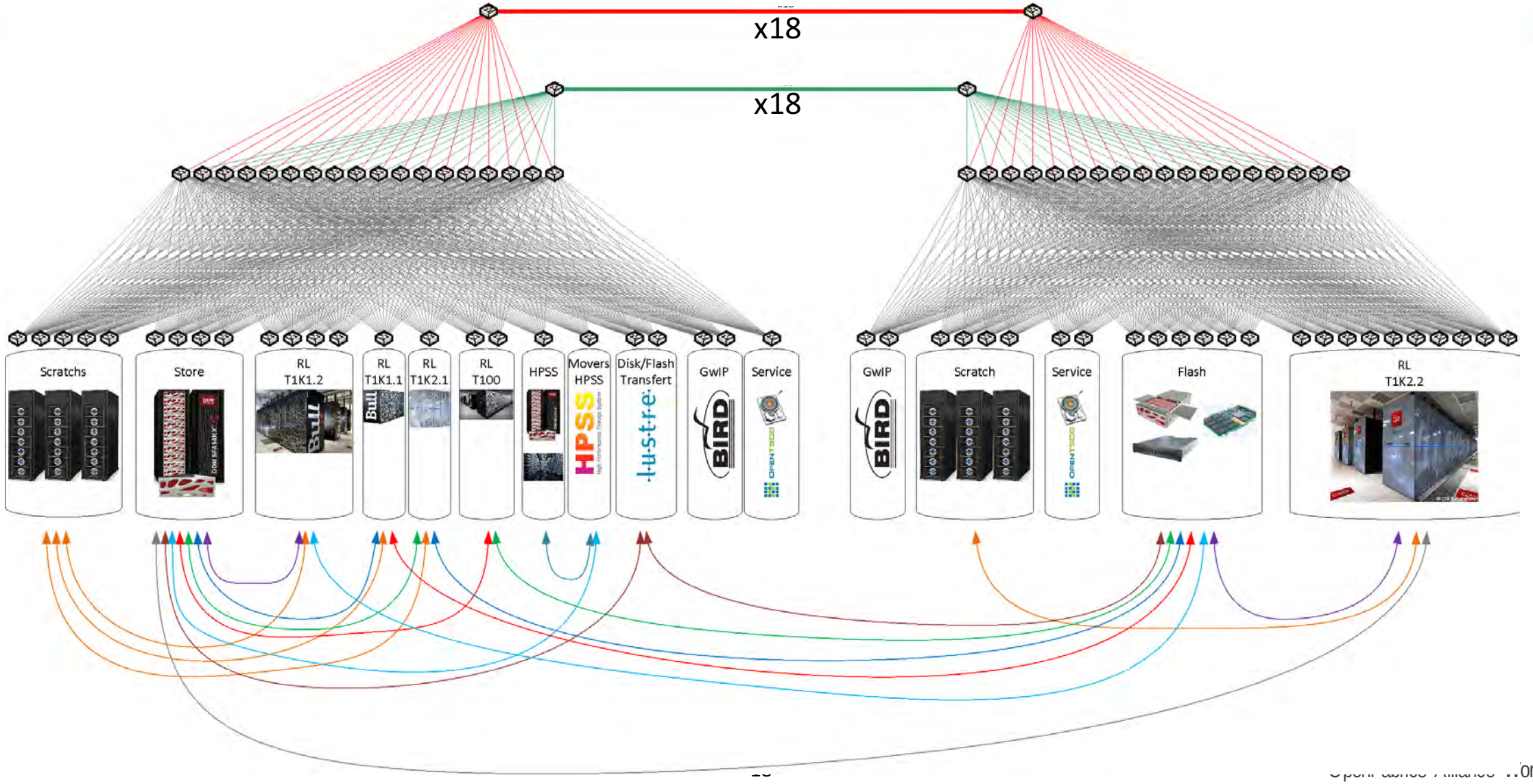
TOPOLOGY – FLOW MAP



TOPOLOGY – FLOW MAP



TOPOLOGY – FLOW MAP



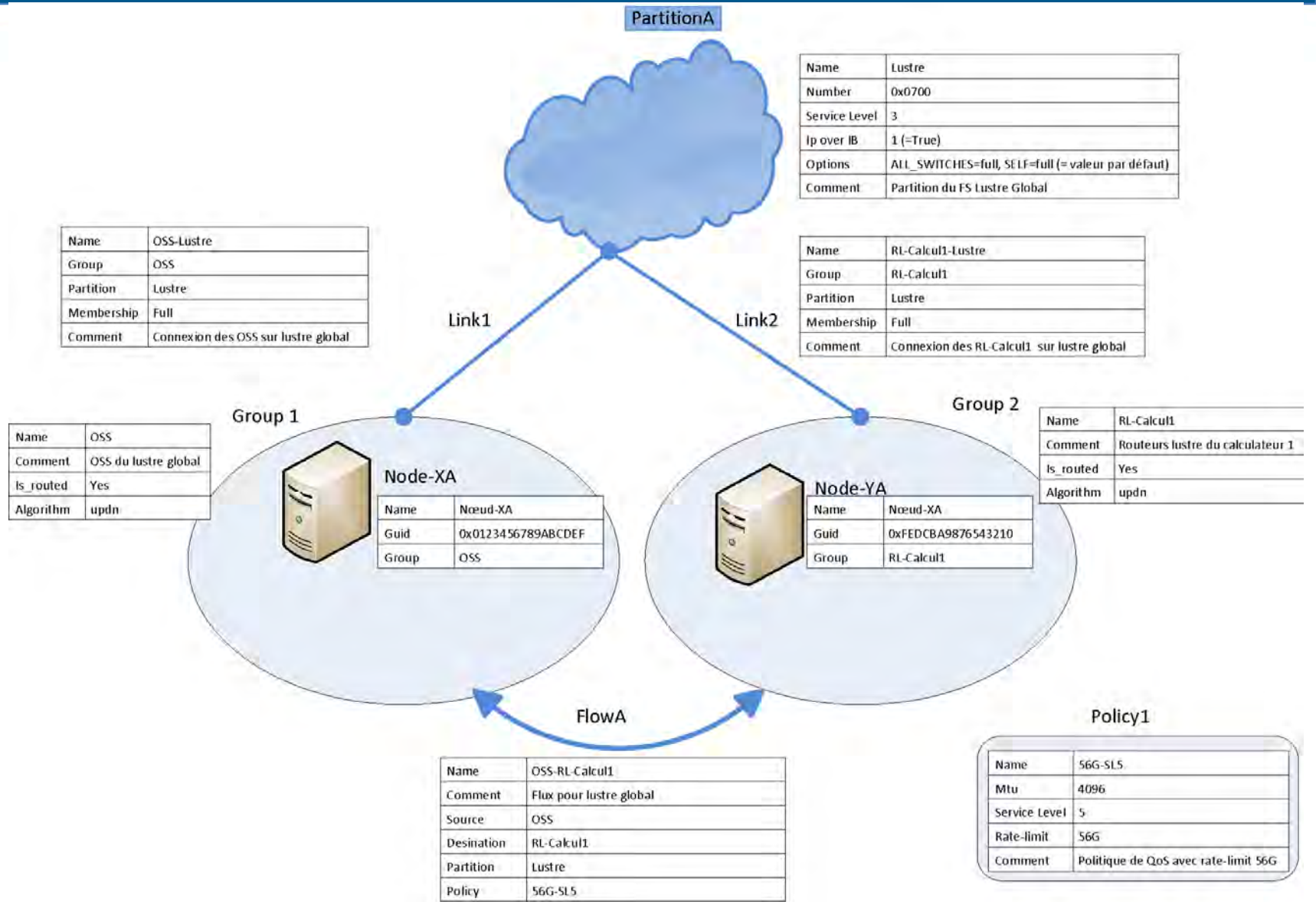
FABRIC CONFIGURATION

FABRIC CONFIGURATION

- **Lots of configuration**
  - Partitions membership
    - partitions.conf
  - QoS definition
    - qos-policy.conf
  - Root switches
    - root_guid_file
  - Proxy-arp consideration
    - io_guid_file
  - Routing configuration (for routing_chain)
    - port_groups_policy_file
    - topo_policy_file.cfg
    - chains_policy_file

- **Tool required !**
  - Higher level description
  - Prevent error in config file
  - Allow non expert to do some configuration

- **Starting with IbManager**
  - Python script using python-rdma
  - Discover nodes on the fabric
    - Still need to manually enter VMs (sriov)
  - Generating configuration files



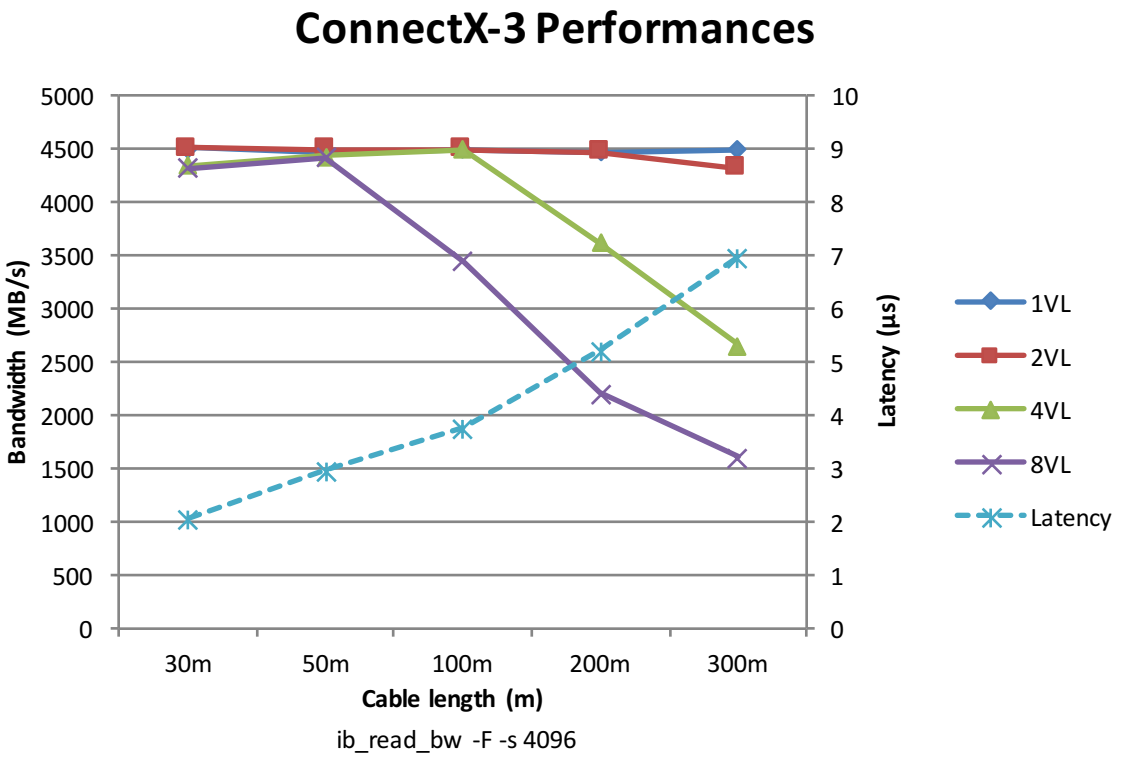
FABRIC VERIFICATION

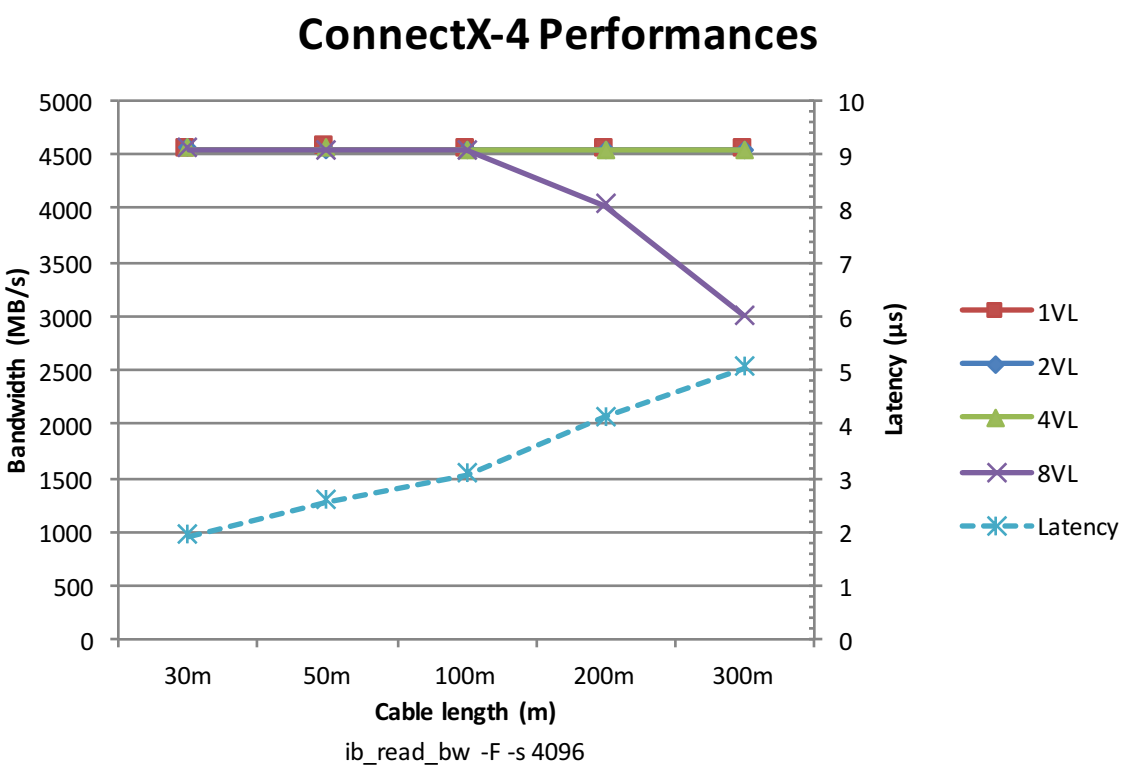
- **Verification enlighted some problems**
  - Sriov was not able to cross proxy-arp
  - Rate-limit feature was not implemented on ConnectX4 Hca
  - Advanced QoS policy return errors when overlapping partition SL
  - Only 4vl were available on ConnectX4 Hca
  - Out of memory problem on nodes crossing the proxy-arp while restarting the SM or proxy-arp
  - QoS was not respected on inter-switch links
  - Sriov was disabling QoS on the card (only one VL supported)
  - Limited membership was not working on default pkey for ConnectX4 Hca

- **Issues have been fixed by Mellanox**
  - Integration of theses fixes may be complicated … and take some time
    - Several manufacturers on the fabric (Atos-Bull/Seagate/DDN…)
    - Several OS on the fabric (SCS, Ocean, SFA-Os)

FABRIC VERIFICATION : QOS COST

- **QoS with long link experience**
  - Using 300m FDR links showed us some performance problems
  - With ConnectX3 and Connect4 Hca
- **Test several cable length at FDR speed**

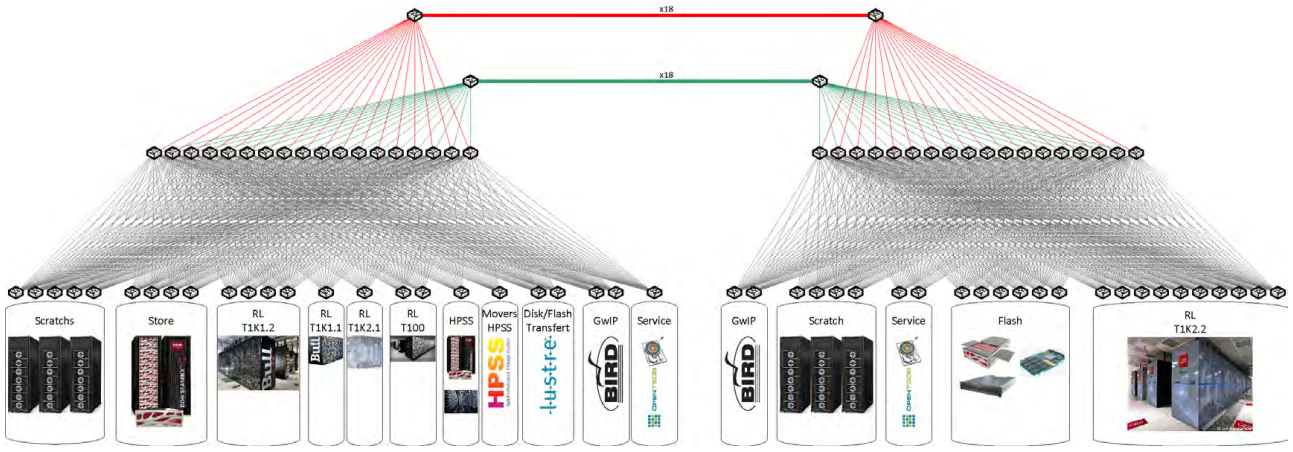




TOPOLOGY VALIDATION

TOPOLOGY VALIDATION

- **Interconnecting two chassis through leaf switches**
  - First time for us
  - Define the interconnection size on the BW need
    - Not less, but not a lot more
    - 36 EDR links on the interconnect
    - 250 GB/s generated by 34 transfers nodes (main need)



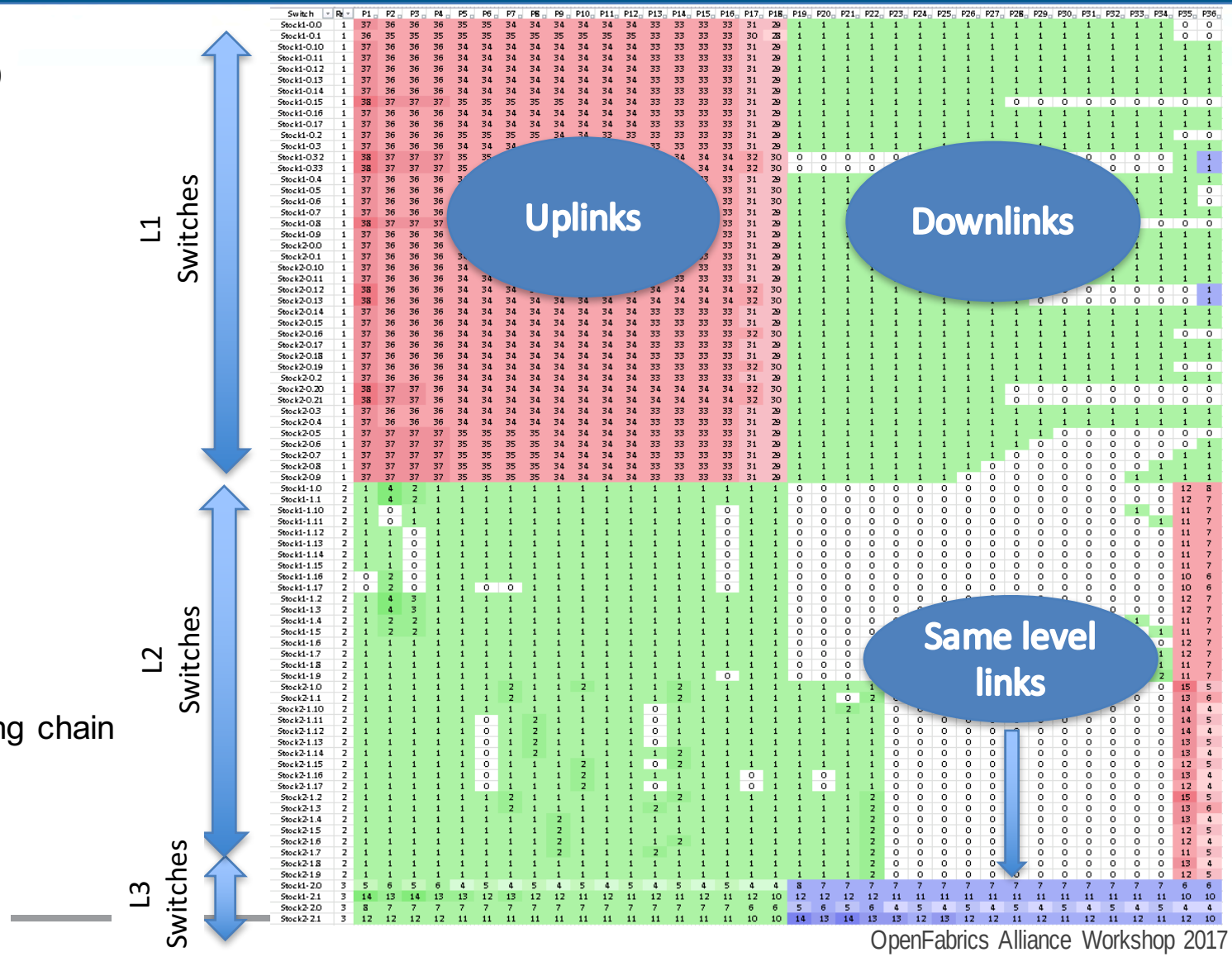
- **Listen to the system engineers**
  - They have a lot of imagination
  - They have new needs to be addressed

- **Need routing validation**
  - Create topology to use with ibsim
  - Opensm accept the topology
  - Routing seems to be fine
  - Get some path overlapping on top switches when considering a group of interest

ROUTING VALIDATION

- **Validation Tool**
  - Had a previous work of path illumination (clean lft)
    - Based initially on ibgraph from CalculQuebec
  - Need to consider routing group by group
  - Need a way to display the results
    - Currently using a spreadsheet
      - 1 tab per group
      - 1 line per switch
      - 1 column per port
    - Gradient coloring of cell
      - Link type (uplink/downlink)
      - Route count

- **Workflow for routing validation**
  - IbSim (patched to work with python-rdma)
  - IbManager (with python-rdma sim branch)
    - Defining group within topology
    - Defining routing engine per group if using routing chain
  - Opensm
  - Routing analysis tool, eyes and brain

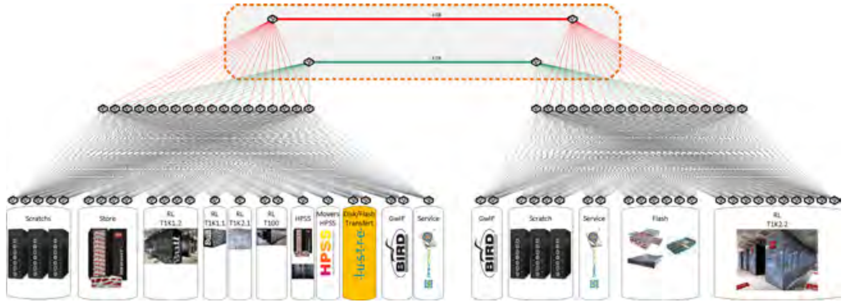


ROUTING VALIDATION : FTREE

- **Big picture seems ok on top level**
  - Paths are well balanced on links

Stock1-2.0	3	11	9	9	9	9	9	9	9	9	8	8	8	8	8	8	8	7	6	9	9	9	9	9	9	9	9	8	8	8	8	8	8	8	8	8	8
Stock1-2.1	3	11	9	9	9	9	9	9	9	9	8	8	8	8	8	8	8	6	5	10	10	10	10	10	10	10	10	10	10	10	9	9	9	9	9	9	9
Stock2-2.0	3	11	11	11	11	11	11	11	11	11	9	8	5	5	5	5	5	5	6	9	9	9	9	9	9	9	9	8	8	8	8	8	8	8	8	8	8
Stock2-2.1	3	11	11	11	11	11	11	11	11	11	11	10	8	8	8	8	8	6	7	9	9	9	9	9	9	8	8	8	8	8	8	8	8	8	8	8	8

- **Closer look on group of interest (disk/flash transfers)**
  - Paths overlapping on top switches
  - May or may not happened depending on switch guid (50% chance)



Stock1-2.0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Stock1-2.1	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Stock2-2.0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Stock2-2.1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0

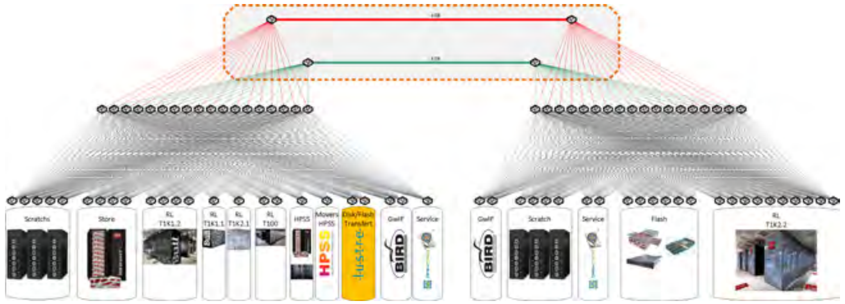
- **Two solutions**
  - Define group of interest in io_guid_file
    - 1 group routed separately
  - Use routing chain feature from Mellanox Opensm
    - Define multiple groups with separated routing

ROUTING VALIDATION : ROUTING CHAIN

- **Big picture seems less ok on top level**
  - Paths are not equally balanced on links

Stock1-2.0	3	5	6	5	6	4	5	4	5	4	5	4	5	4	5	4	5	4	4	8	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	6	6
Stock1-2.1	3	14	13	14	13	13	12	13	12	12	11	12	11	12	11	12	11	12	10	12	12	12	12	11	11	11	11	11	11	11	11	11	11	11	11	10	10
Stock2-2.0	3	8	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	6	6	5	6	5	6	4	5	4	5	4	5	4	5	4	5	4	5	4	4
Stock2-2.1	3	12	12	12	12	11	11	11	11	11	11	11	11	11	11	11	11	10	10	14	13	14	13	13	12	13	12	12	11	12	11	12	11	12	11	12	10

- **Main group is routed with ftree as today**
  - Resources on this group will not cross the interconnection



- **Closer look on group of interest (disk/flash transfers)**
  - Paths equally balanced on each level

Stock1-2.0	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Stock1-2.1	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Stock2-2.0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
Stock2-2.1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0

  - Routing will not be the source of congestion here
  - QoS will manage interflow concurrency

CONCLUSION 1/2

- **Importance of simulation before deployment**
  - Could be a long process
  - Step by step validation
    - Topology
    - Routing
  - Enhancement with traffic simulation

- **Routing analysis is crucial**
  - Design topologies
  - Answer the sysadmin questions
    - Why is my performance bad ?
    - Why performance doesn't scale ?

  → Okay when routing is easy enough to draw it

  → Adaptive routing may add complexity

  → IB Routers are interesting; what about load sharing between fabrics

- **Fabric configuration tool**
  - Describe the topology easily
  - Allow network segmentation with no effort
  - Guaranty the QoS
  - Can evolve with the SM and features

CONCLUSION 2/2

- **Deployment under progress (2nd phase, T1K2.2, T1KF)**
  - Project challenging for the team
  - Solid experience in fabric configuration
  - New network hierarchy opening horizon for exascale

- **Currently using maximum resources of implemented QoS**
  - 8VL maximum available on ConnectX hardware
    - Will find use case for more !
  - Suffer of performance issue with long cables
    - Dynamic buffer allocation for hca ?

- **InfiniBand/Ethernet Gateway**
  - Proxy-arp could improve with some features
    - ECMP would permit more agility in design
    - Diffserv tagging (InfninBand SL <-> Ethernet DSCP)

OPENFABRICS ALLIANCE

13th ANNUAL WORKSHOP 2017

THANK YOU

Jérôme David

Commissariat à l'Energie Atomique

Commissariat à l'énergie atomique et aux énergies alternatives
Centre DAM-Île de France, 91297 Arpajon Cedex
T. +33 (0)1 69 26 40 00,
Etablissement public à caractère industriel et commercial , RCS Paris B 775 685 019