JÜLICH
FORSCHUNGSZENTRUM

# Parallel I/O on JUQUEEN and JURECA

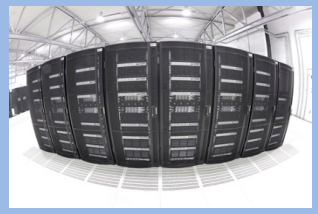
*Hardware Overview*

14.03.2016, Parallel I/O Workshop

**Michael Stephan**

# JSC Supercomputer I/O Infrastructure

JUQUEEN



JURECA

JUST

Ethernet Network

JÜLICH
FORSCHUNGSZENTRUM

# Hardware View

# JUQUEEN: Jülich's Scalable Petaflop System

- IBM Blue Gene/Q JUQUEEN
- IBM PowerPC® A2 1.6 GHz,
  - 16 cores 4way SMT per node
  - 28 racks (7 rows à 4 racks)
  - 28,672 nodes (**458,752 cores**)
- 5D torus network
- 5.9 Pflop/s peak
  5.0 Pflop/s Linpack
- Main memory: **448 TB**
- **I/O Nodes**: **248** (27x8 + 1x32)
  - **496 x 10GigE**

# JURECA: Jülich's Multi-Purpose System

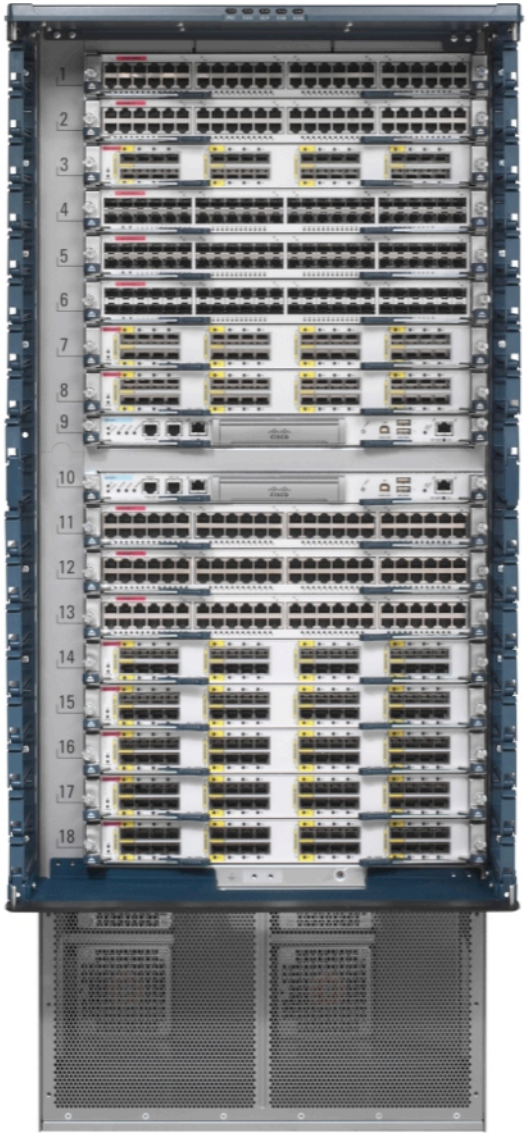
- Jülich Research on Exascale Cluster Architectures
  - Vendor: T-Platforms
- 1884 compute nodes in 34 racks
- Dual-socket Intel Xeon E5-2680 v3 Haswell CPUs
  - 24 cores @ 2.5 GHz (up to 3.3 GHz boost)
  - 480 GFlop/s peak
- RDIMM DDR4 2133 MHz
  - 128 GiB in 1605 thin nodes
  - 256/512 GiB in 196 fat nodes
- Mellanox EDR InfiniBand (100 Gb/s)
  - Two-level non-blocking fat tree
  - Two Gateway switches for storage connectivity
    - *2x 18 port FDR/40GigE gateway ⇨ 100 Gb/s to JUST*

# JUST: GPFS Storage Cluster

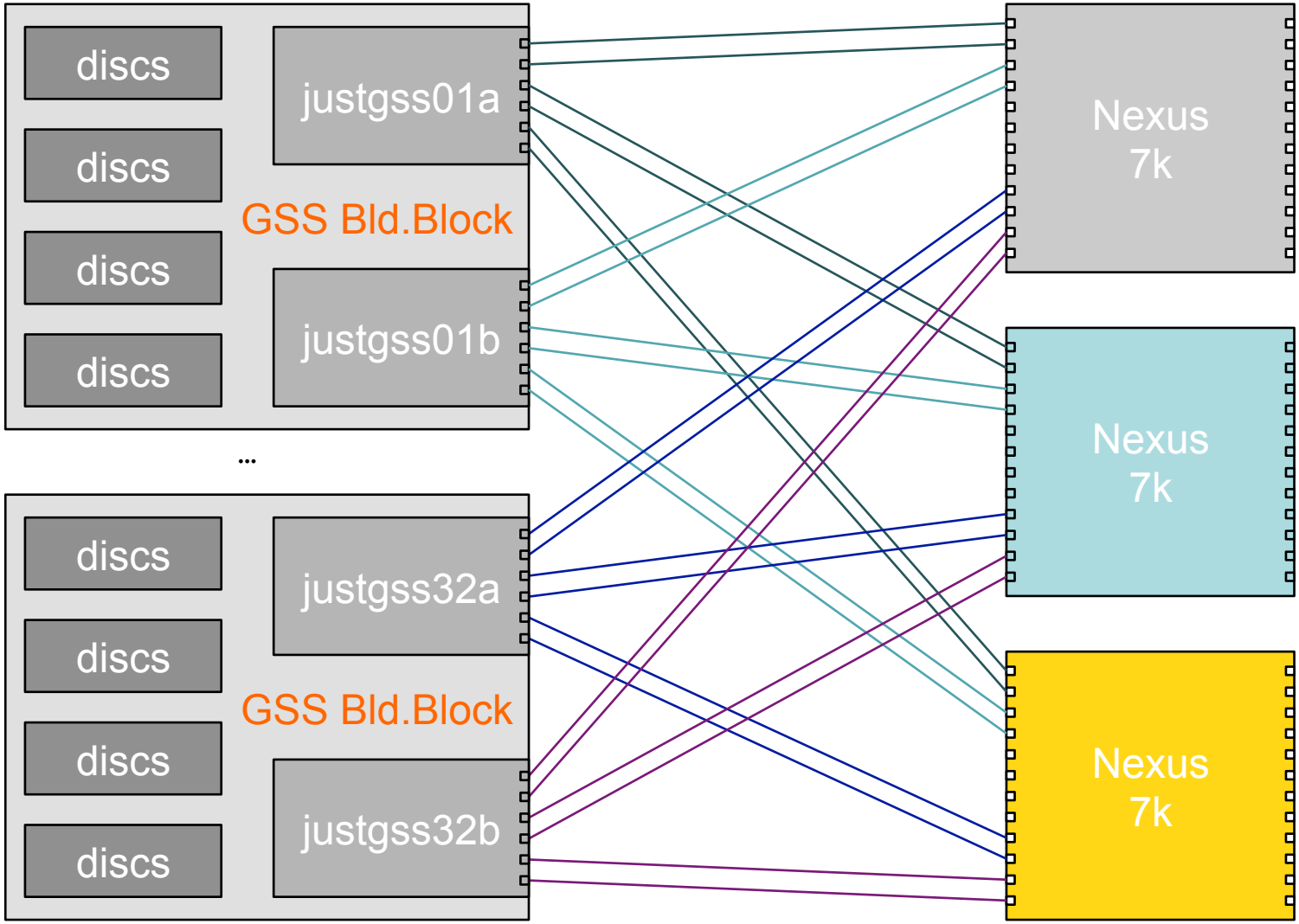
- **Ju**elich **St**orage Cluster
- **GPFS**-Filesystems
  - $WORK
    - *Capacity:* ***5.3 Pbyte,*** *up to* ***200 GB/sec***
  - $HOME (/home[a-c])
    - *Capacity: 3x* ***600 Tbyte, 12 GB/sec***
  - $ARCH (/arch, /arch2)
    - *Capacity: 2x* ***600 Tbyte, 12 GB/sec***
- 33 Building blocks:
    - *Lenovo GPFS Storage Server solution*
    - *2 x x3650 M4 server*
    - *232 NL-SAS disks (2TB)*

# Network Infrastructure

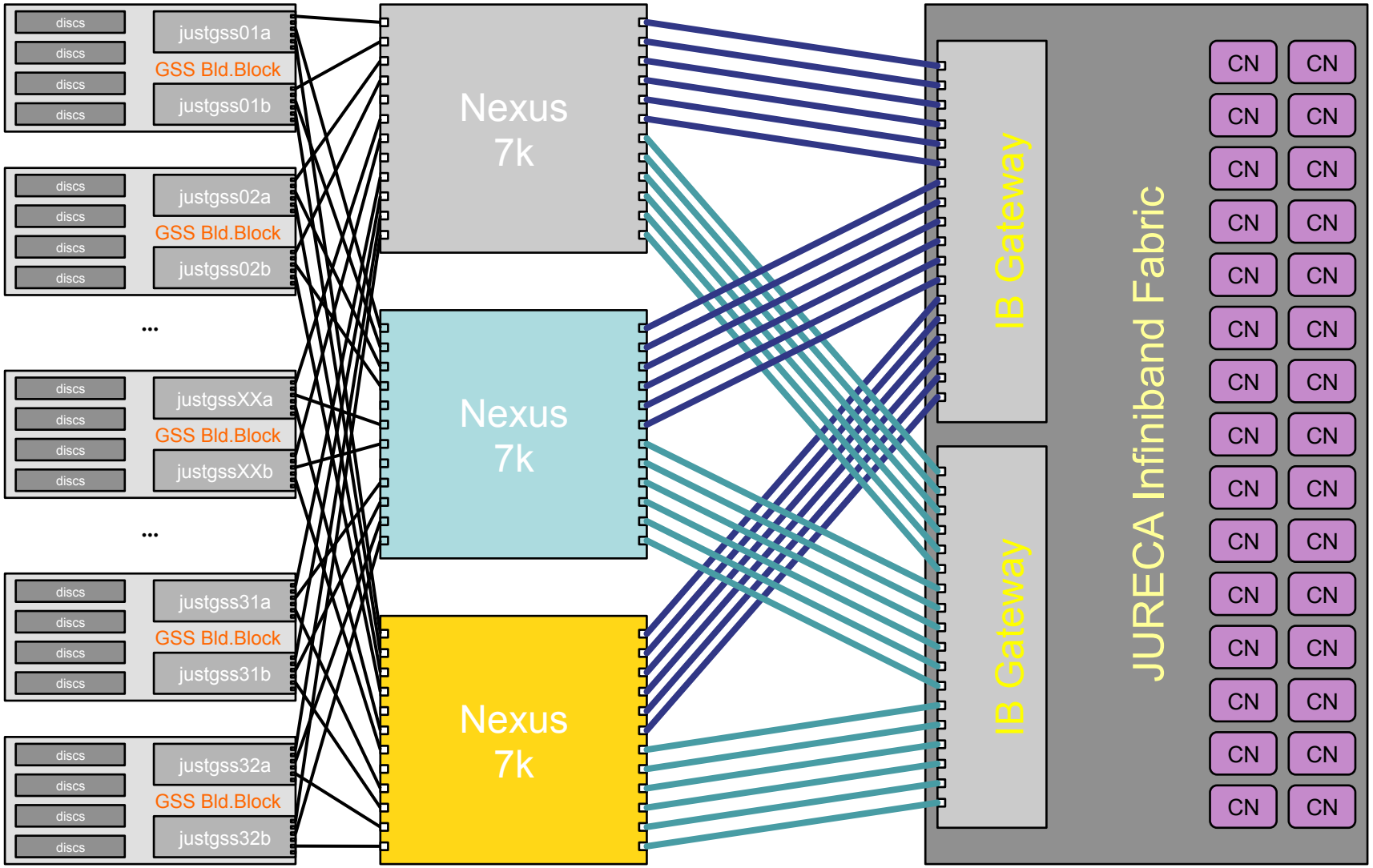
- Cisco Nexus 7018 Switch
  - Number of slots:18
  - Bandwidth per slot: 550 Gb/s
  - Maximum switching capacity: 17.6 Tb/s
  - Port density @ line-rate
    - *10GigE: 768 ports*
    - *40 GigE: 192 ports*
  - Redundant supervisors



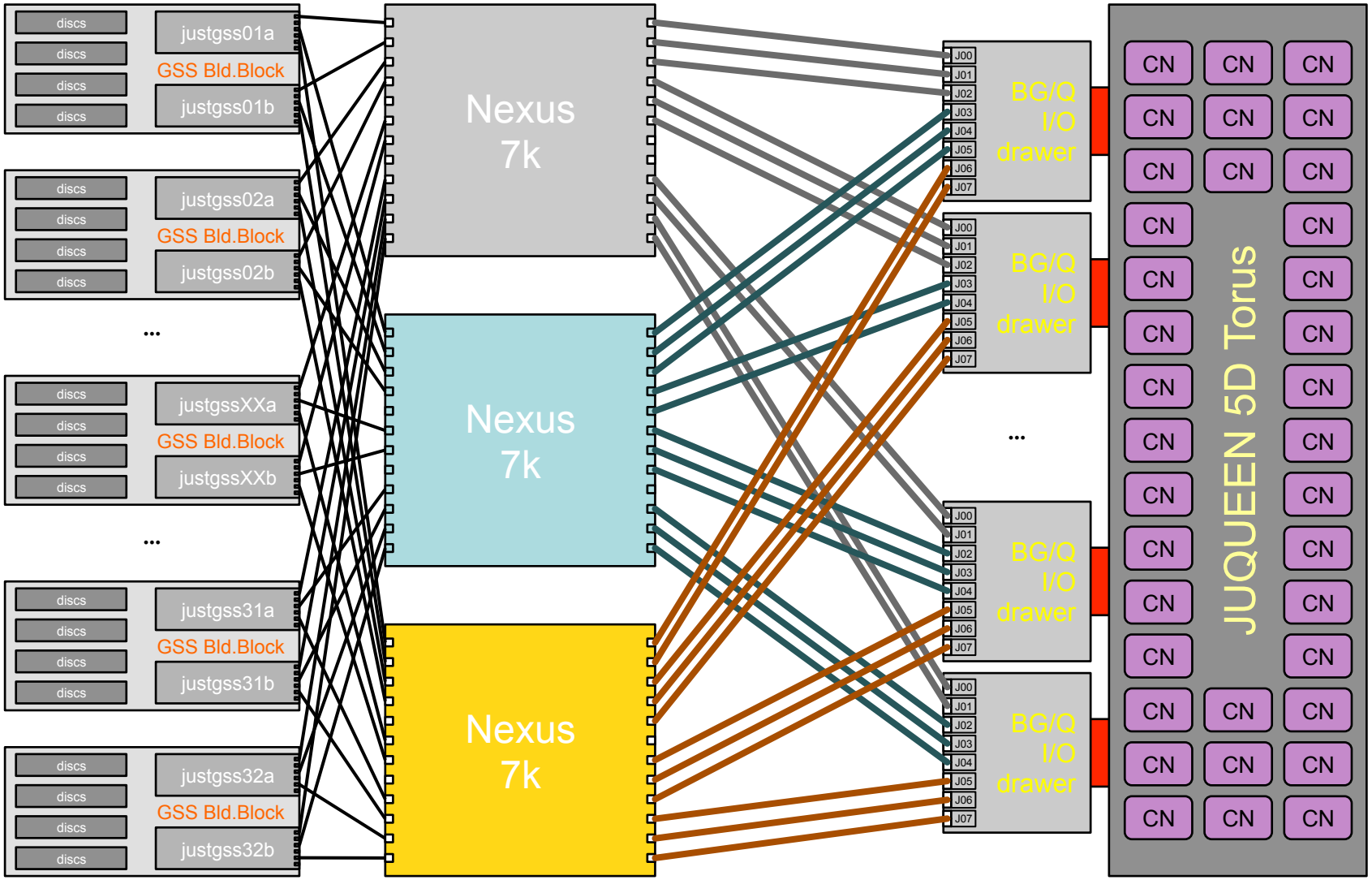
# Network Infrastructure - JUST



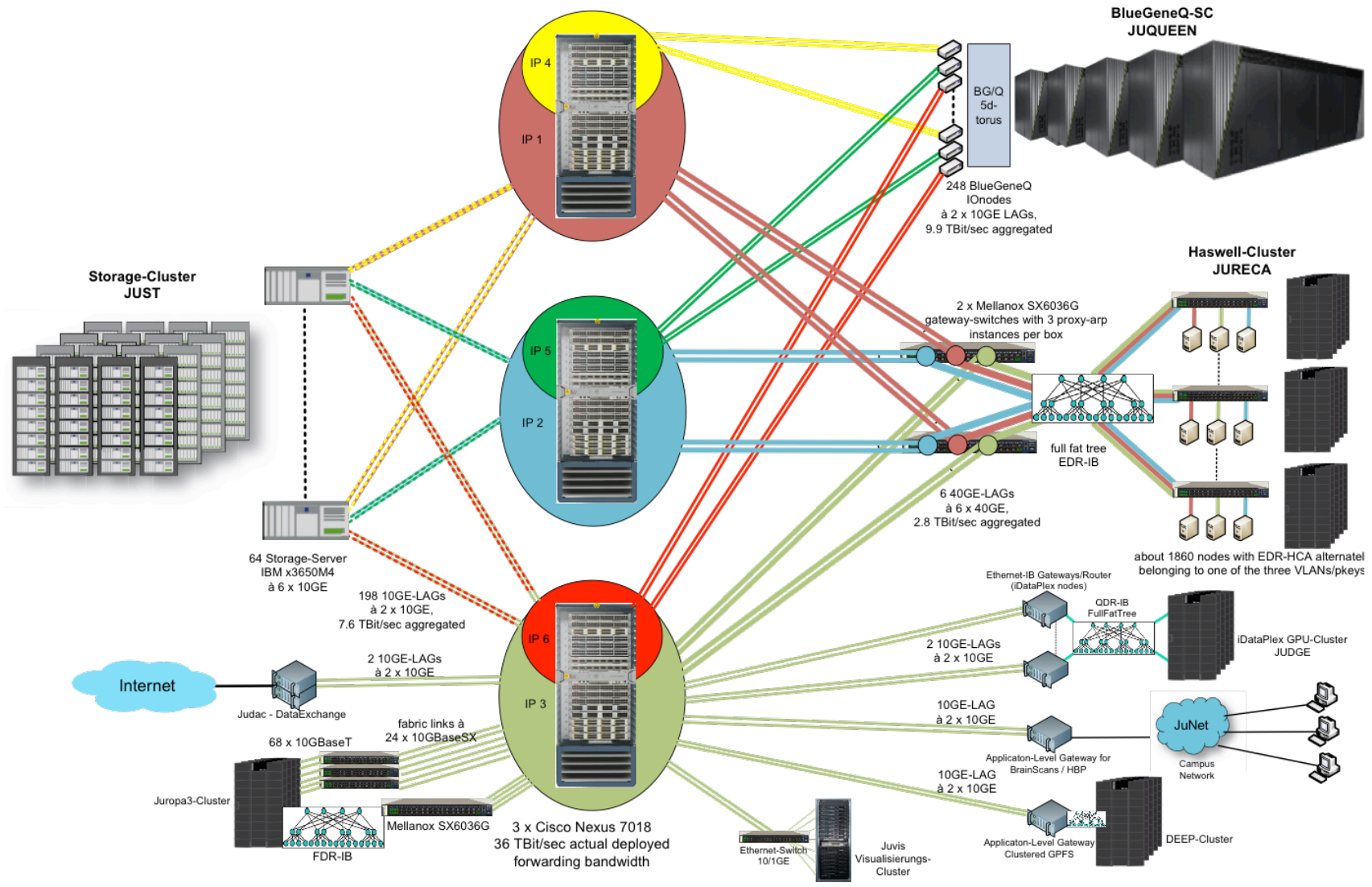
# Network Infrastructure - JURECA



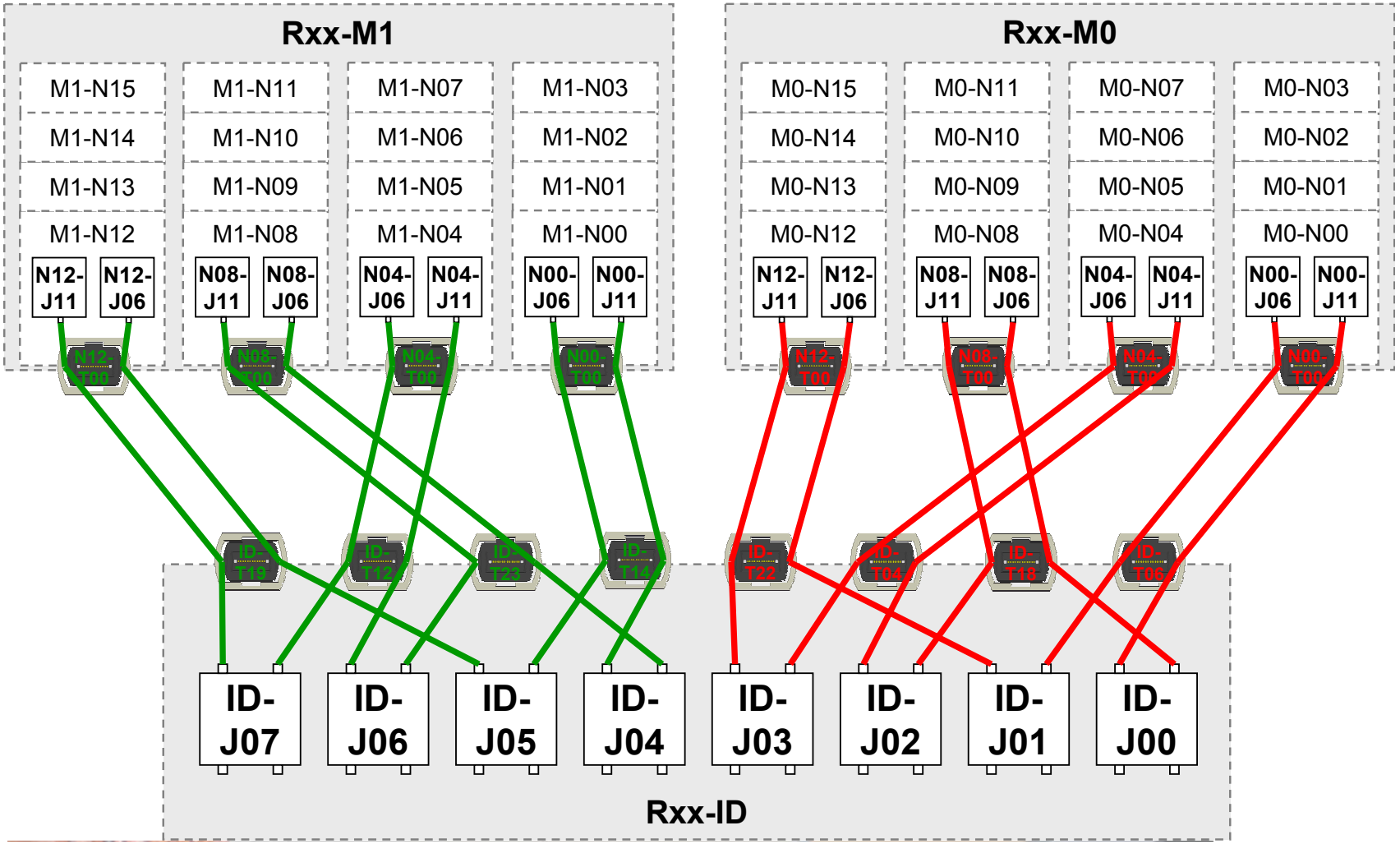
# Network Infrastructure - JUQUEEN



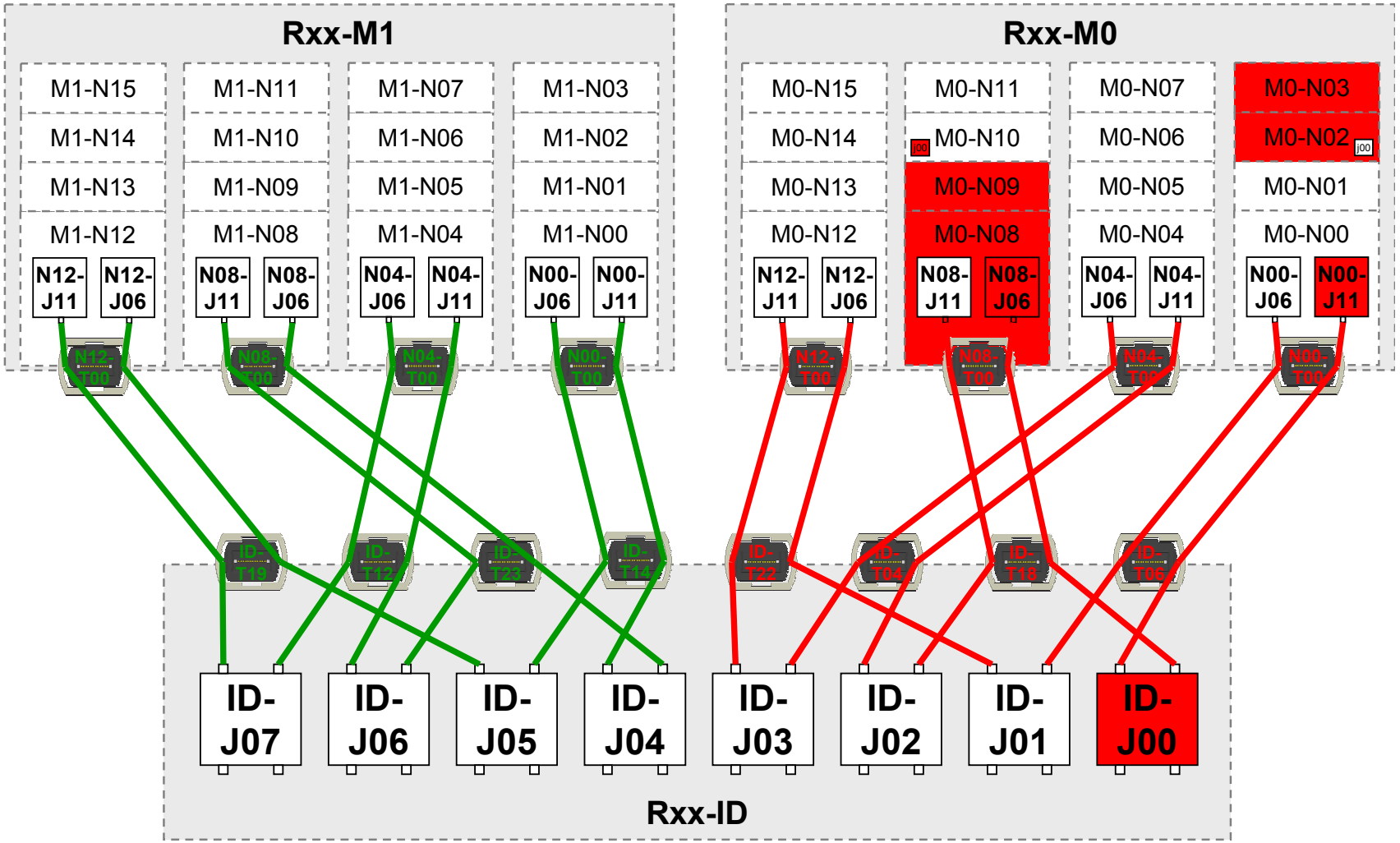
# Network Infrastructure – Final picture



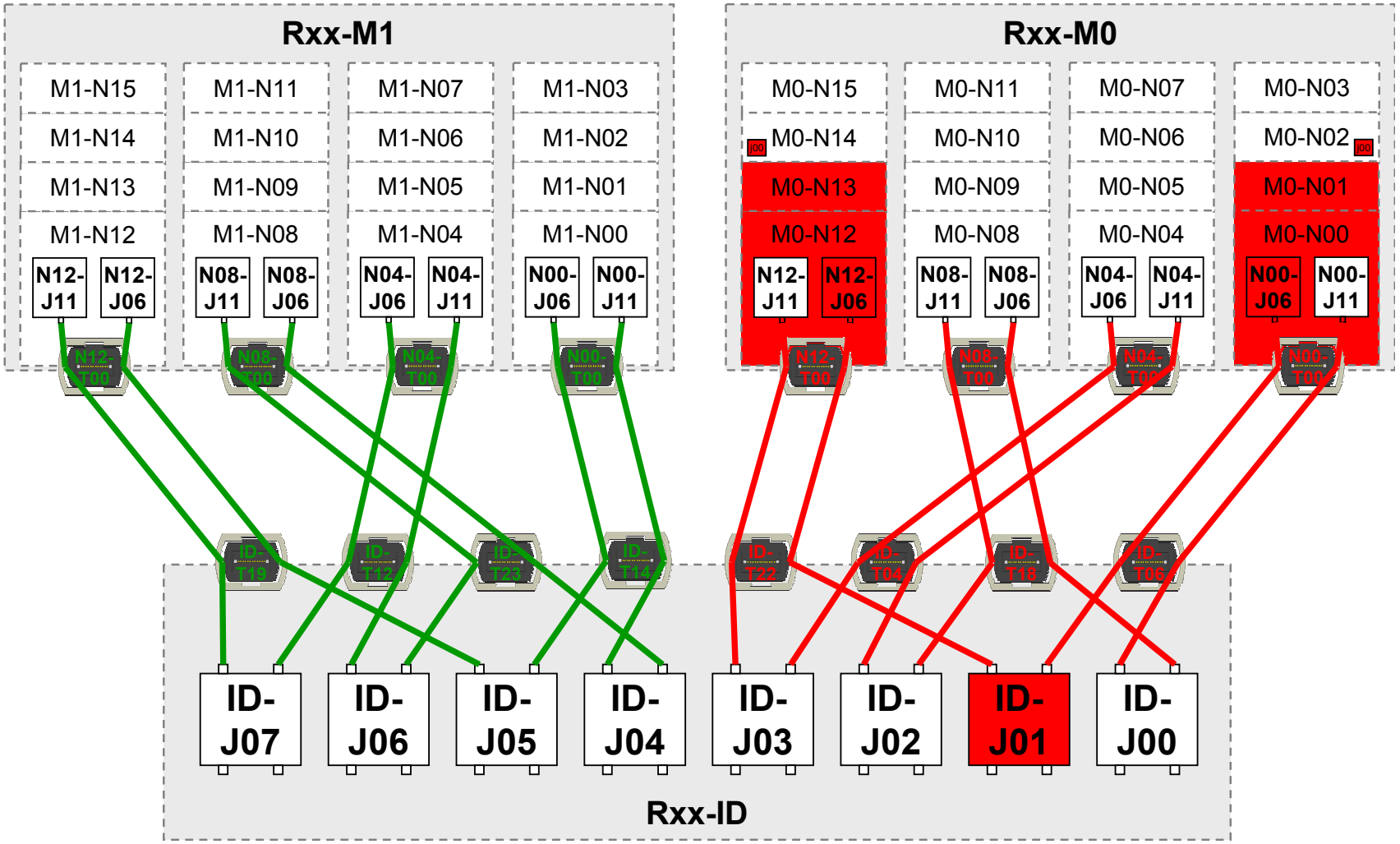
# Blue Gene/Q: I/O-node cabling (8 ION/Rack)



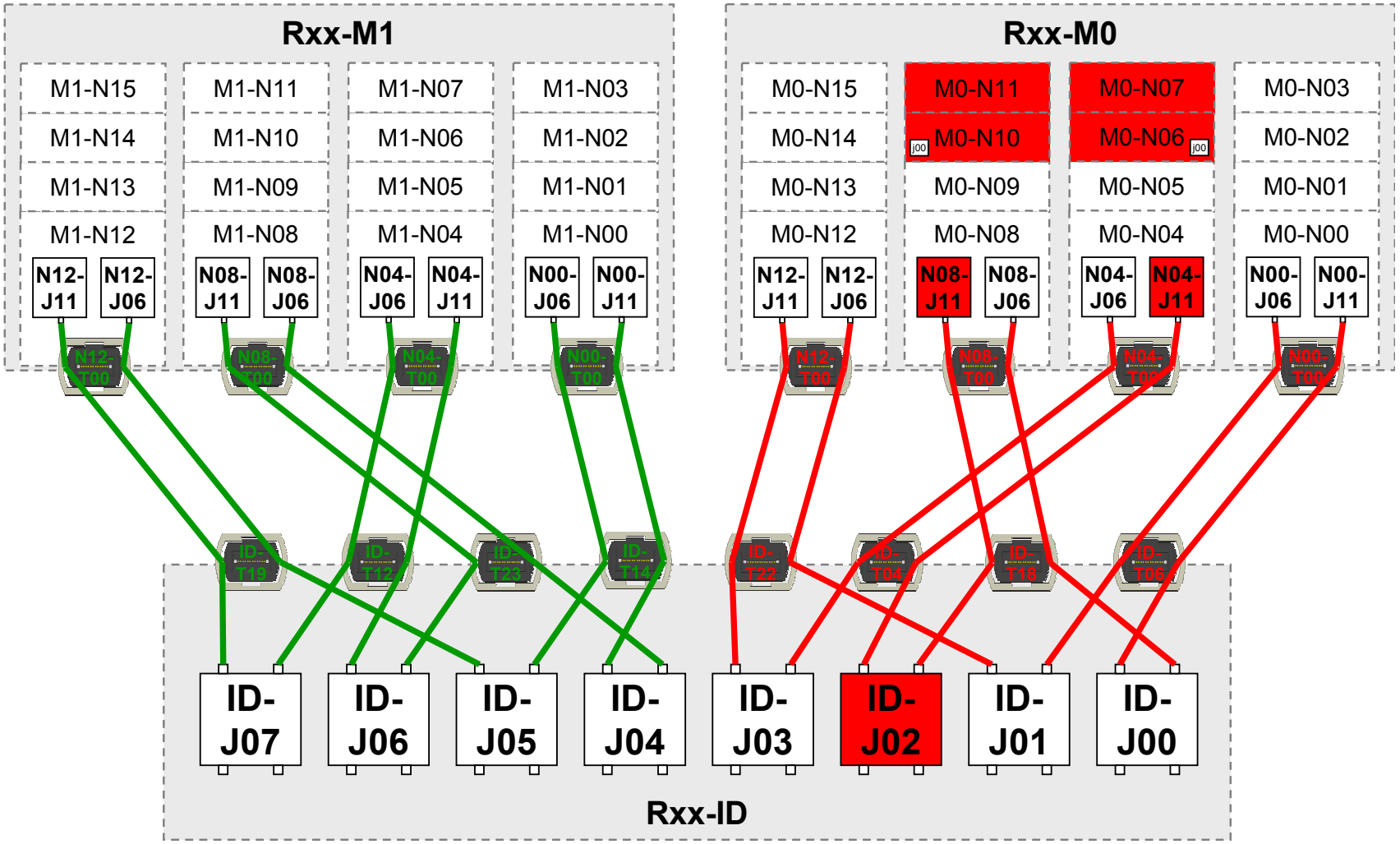
# Blue Gene/Q: I/O-node cabling (8 ION/Rack)



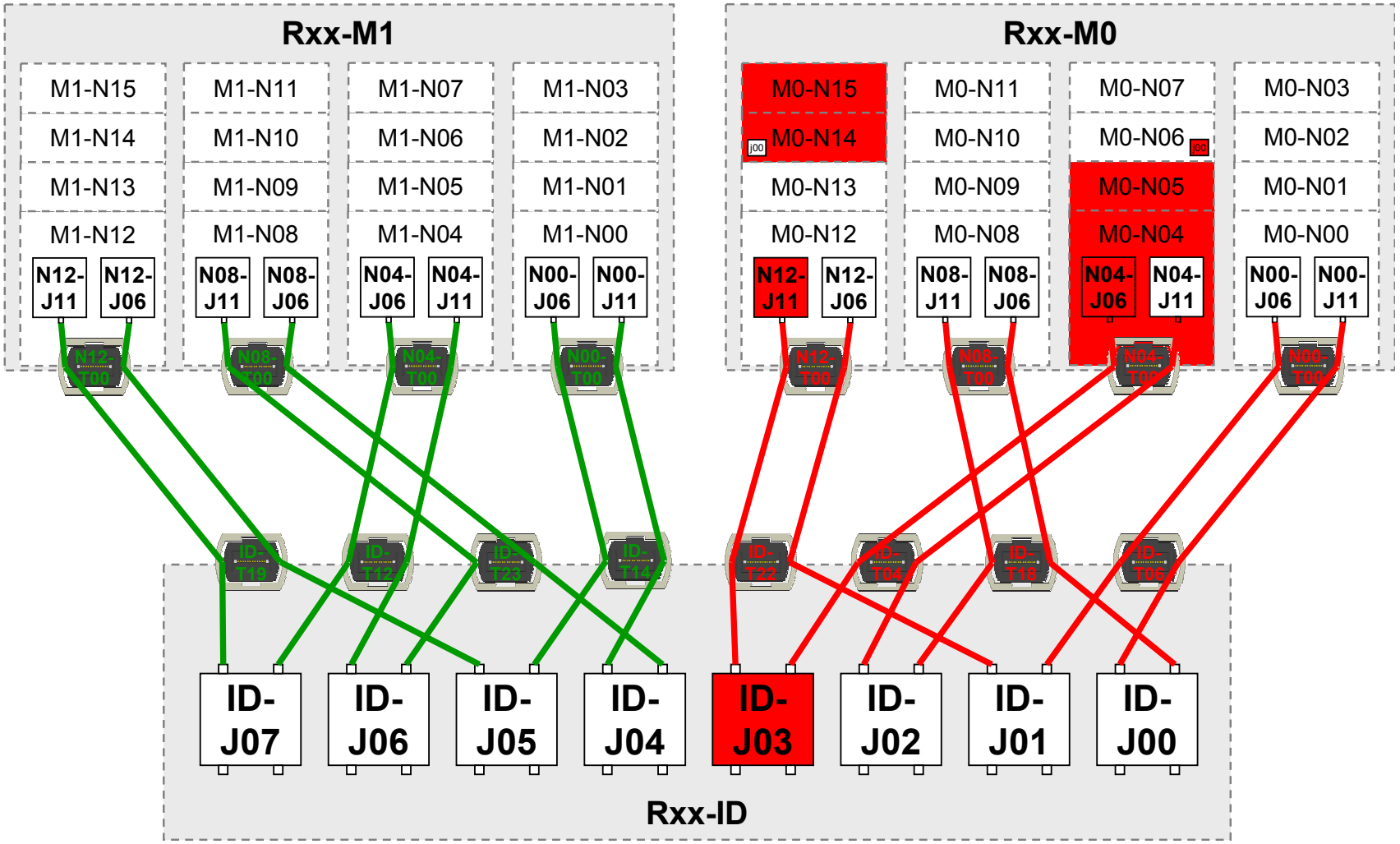
# Blue Gene/Q: I/O-node cabling (8 ION/Rack)



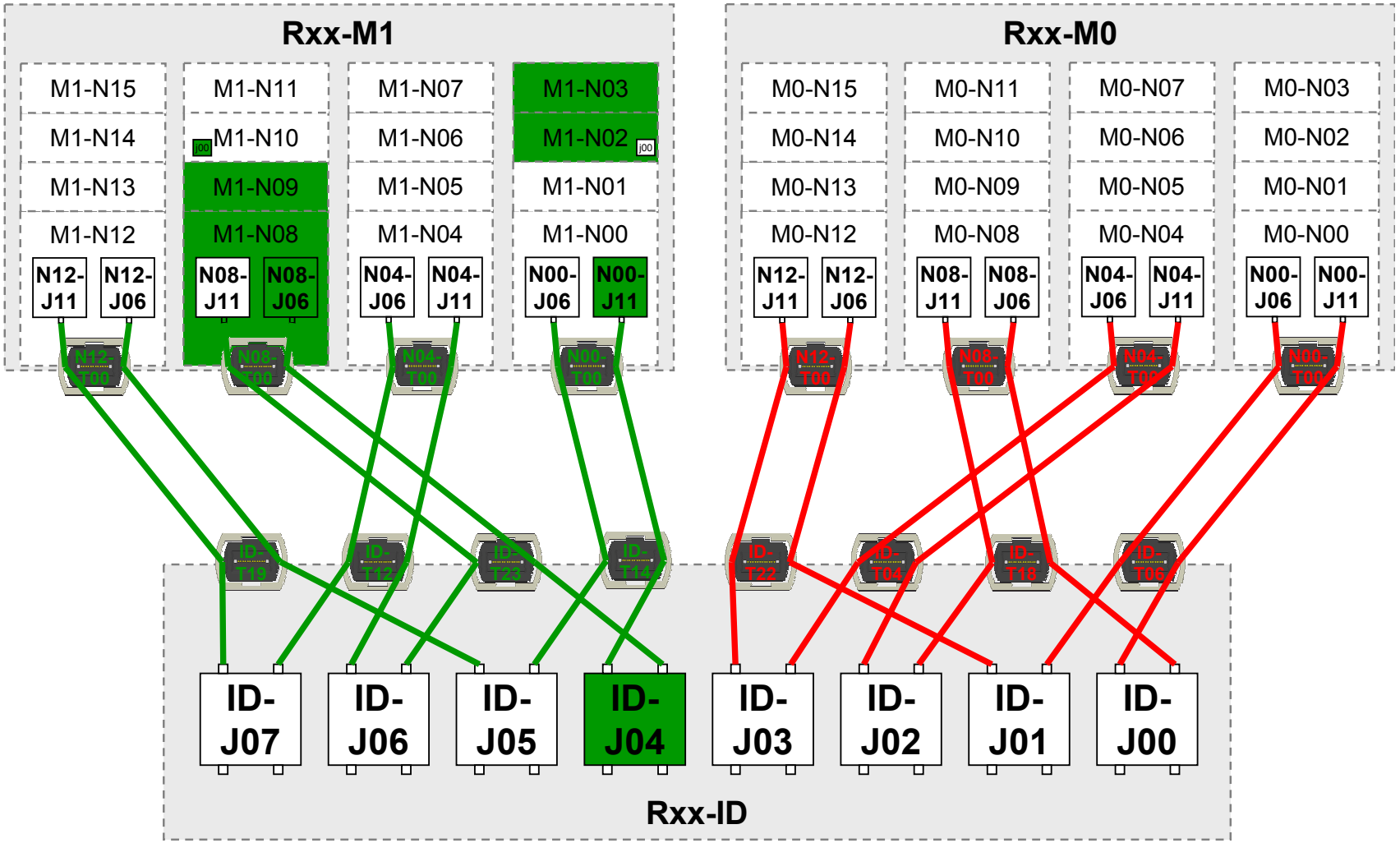
# Blue Gene/Q: I/O-node cabling (8 ION/Rack)



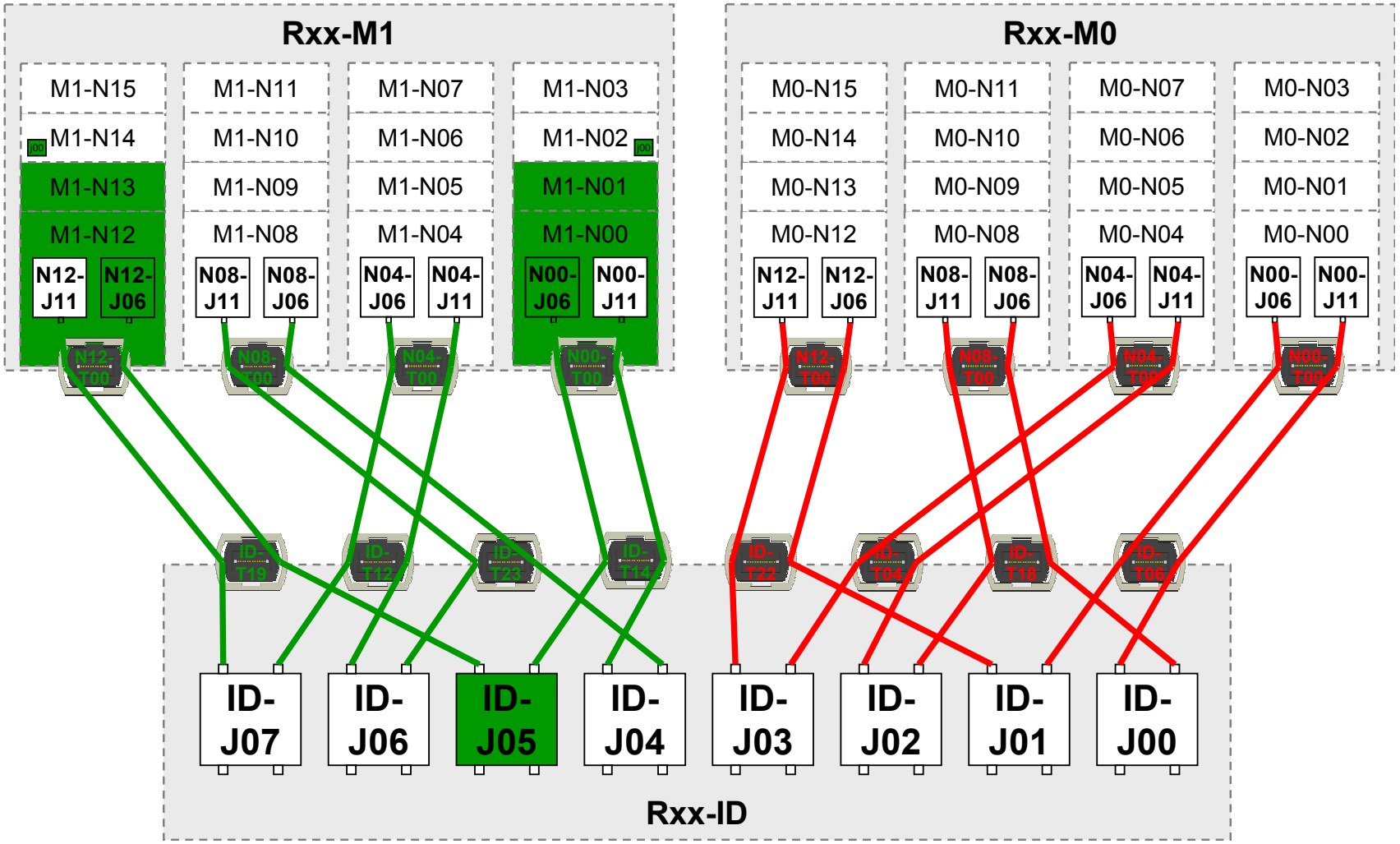
# Blue Gene/Q: I/O-node cabling (8 ION/Rack)



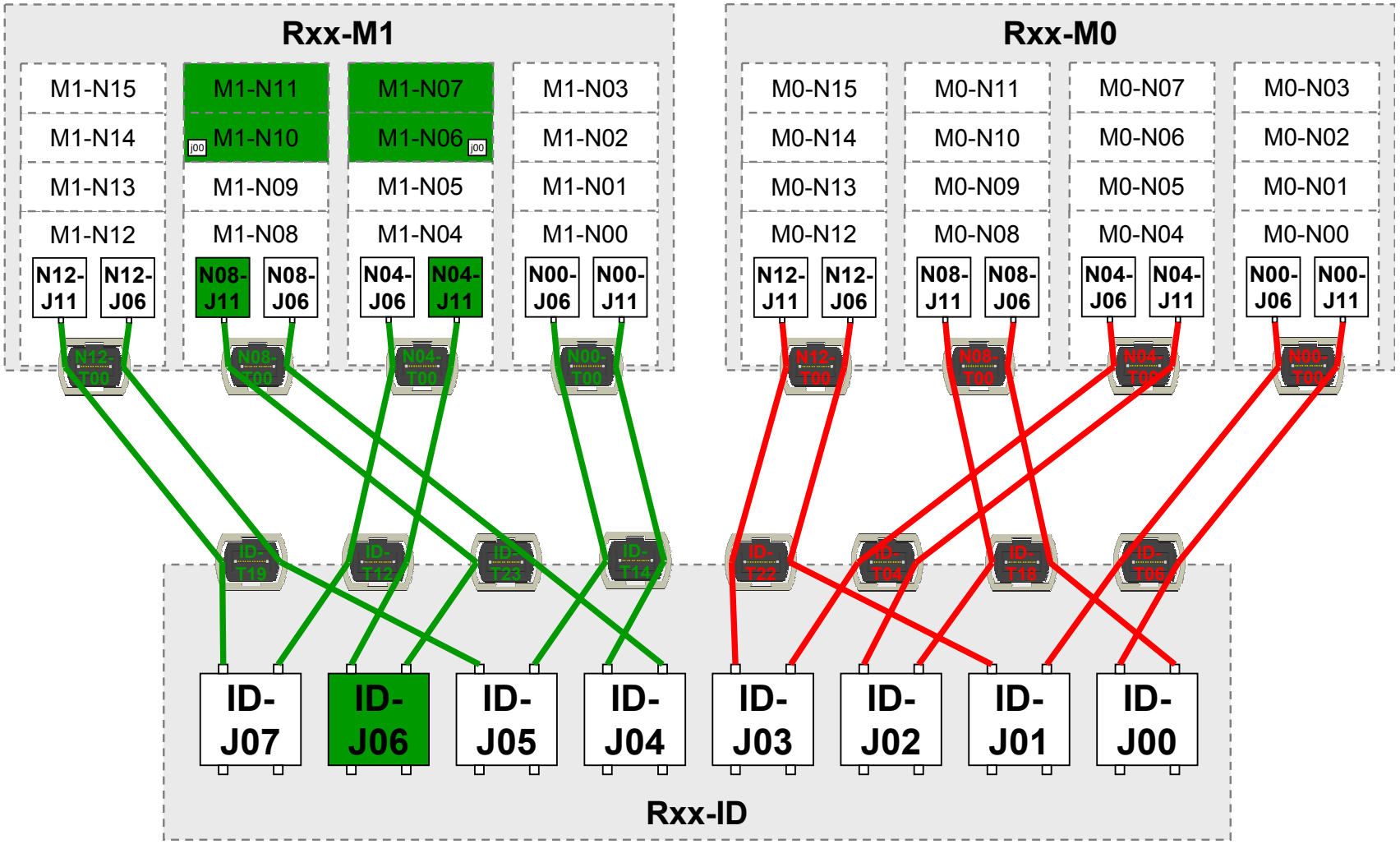
# Blue Gene/Q: I/O-node cabling (8 ION/Rack)



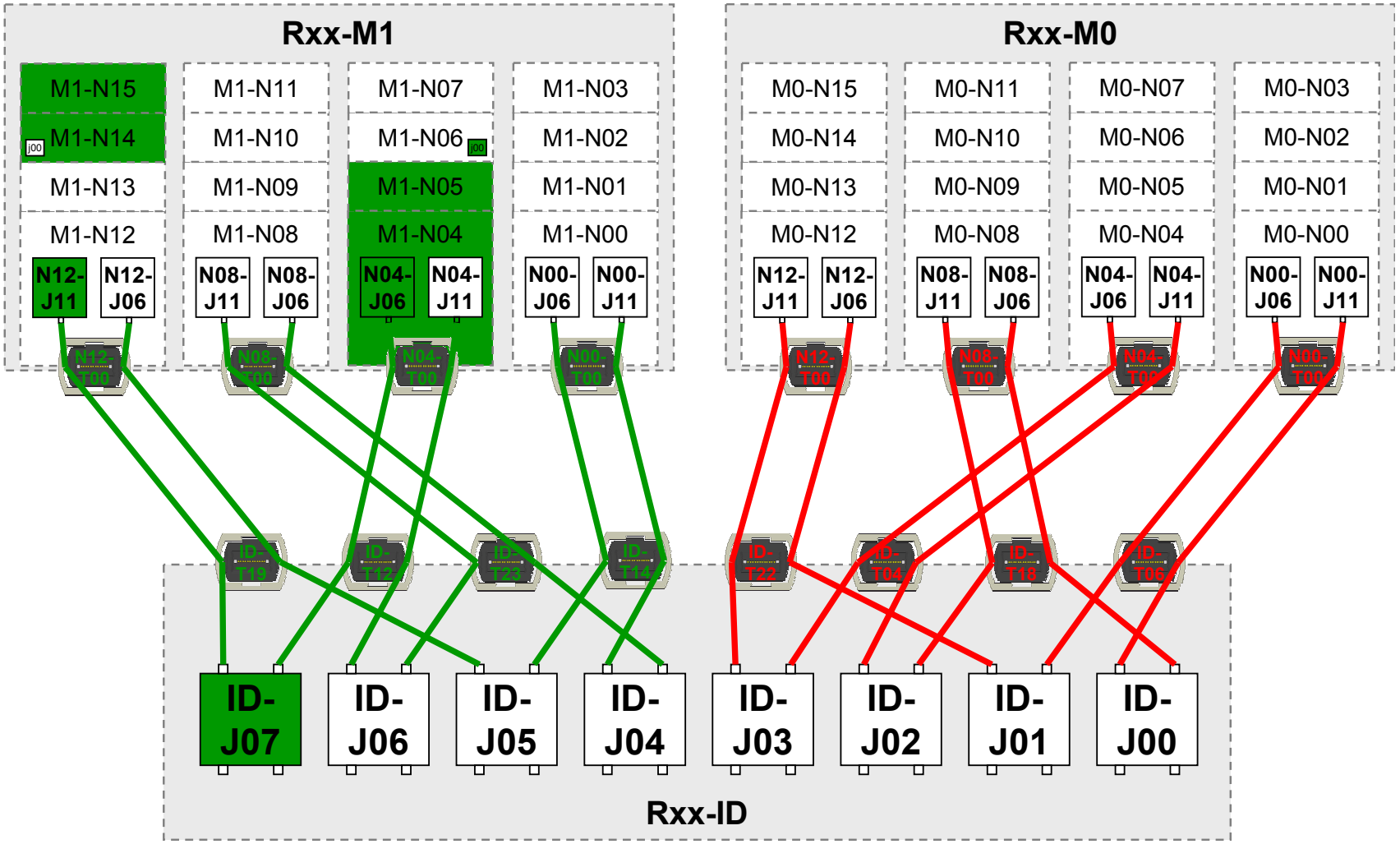
# Blue Gene/Q: I/O-node cabling (8 ION/Rack)



# Blue Gene/Q: I/O-node cabling (8 ION/Rack)



# Blue Gene/Q: I/O-node cabling (8 ION/Rack)



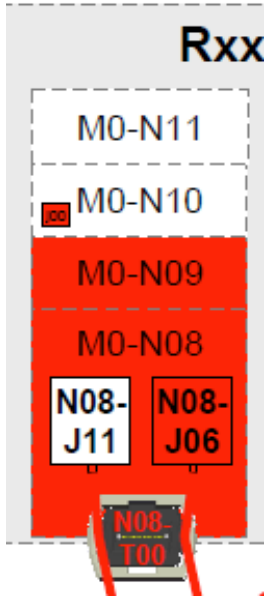
# BGQ: Tasks connected to same I/O-node

- MPIX_Calls available on BG/Q
  (see `http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/UserInfo/MPIextensions.html`)
- Communicator: All tasks belonging to same I/O Bridge Node

```
FORTRAN: MPIX_PSET_SAME_COMM_CREATE(INTEGER pset_comm_same,
                                    INTEGER ierr)
C: #include <mpix.h>
   int MPIX_Pset_same_comm_create( MPI_Comm *pset_comm_same )
```

- Usage: implementation of own I/O strategy
  (One file per I/O-bridge)
- Passing new communicator to SIONlib paropen-Call (as local communicator)

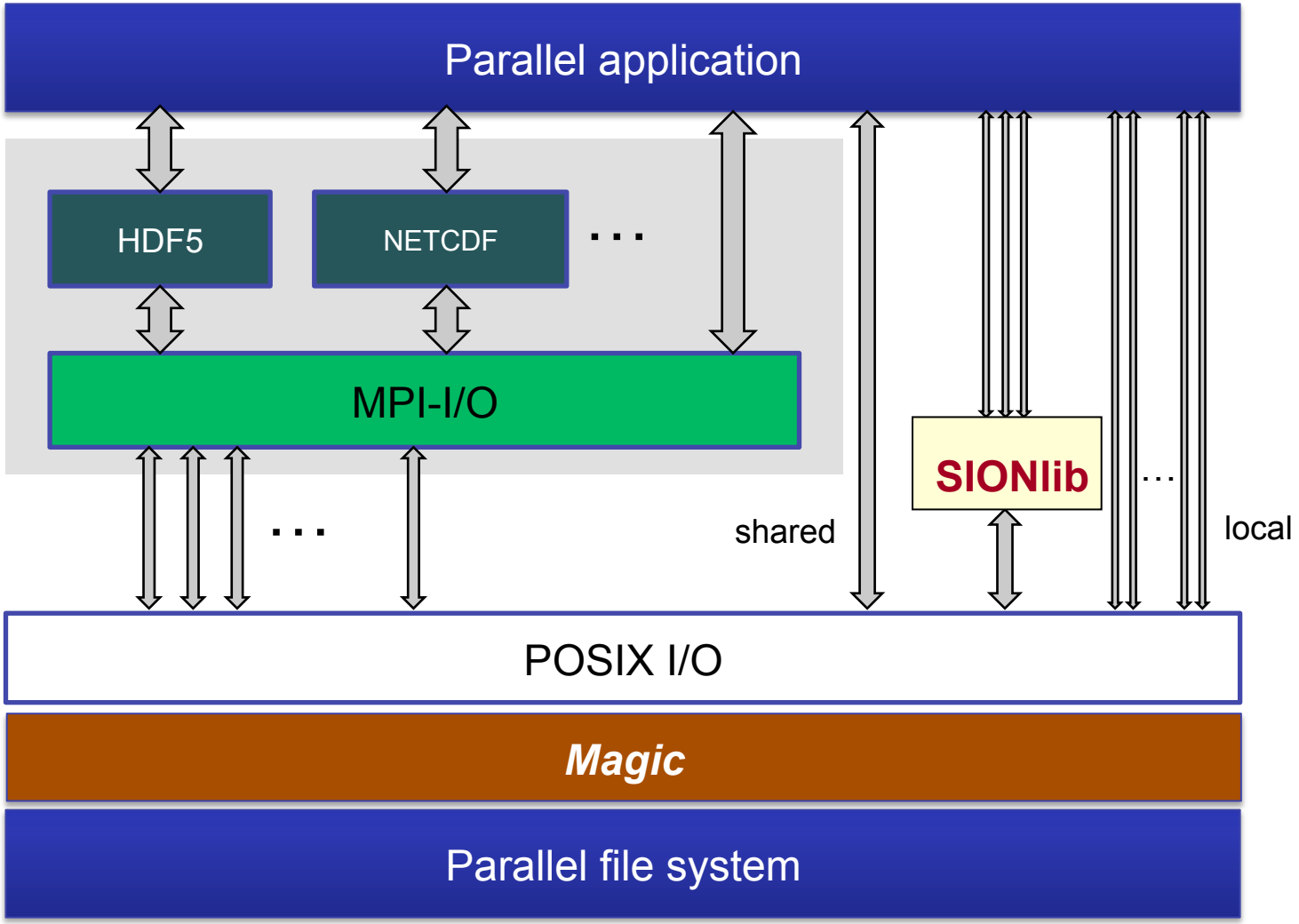
```
…
sid=sion_paropen_mpi( filename , “bw“,
                      &numfiles, &chunksize,
                      gcom, &lcom, &fileptr, ...);
...
```



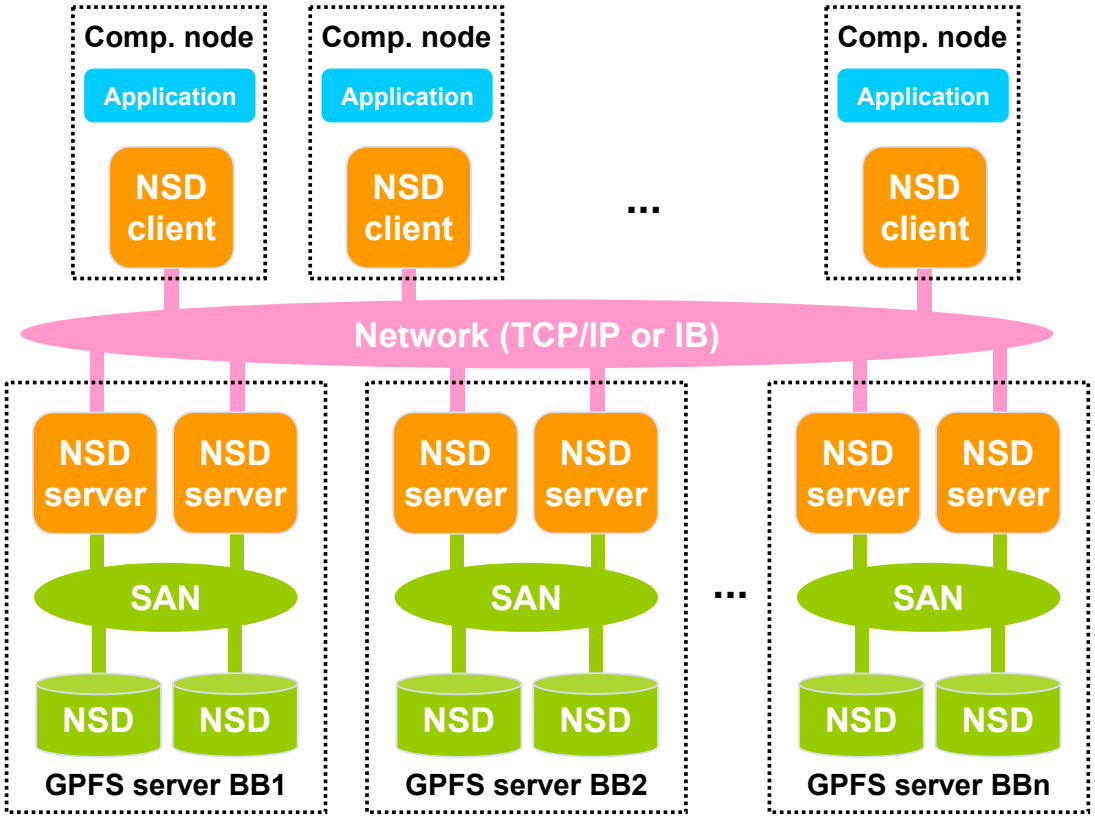
JÜLICH
FORSCHUNGSZENTRUM

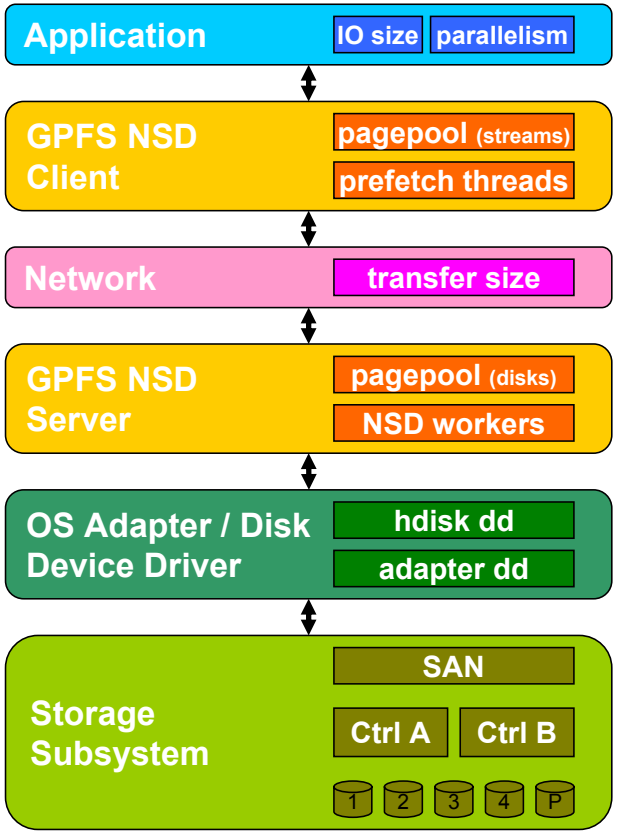
# Software View

# Application View to Parallel I/O

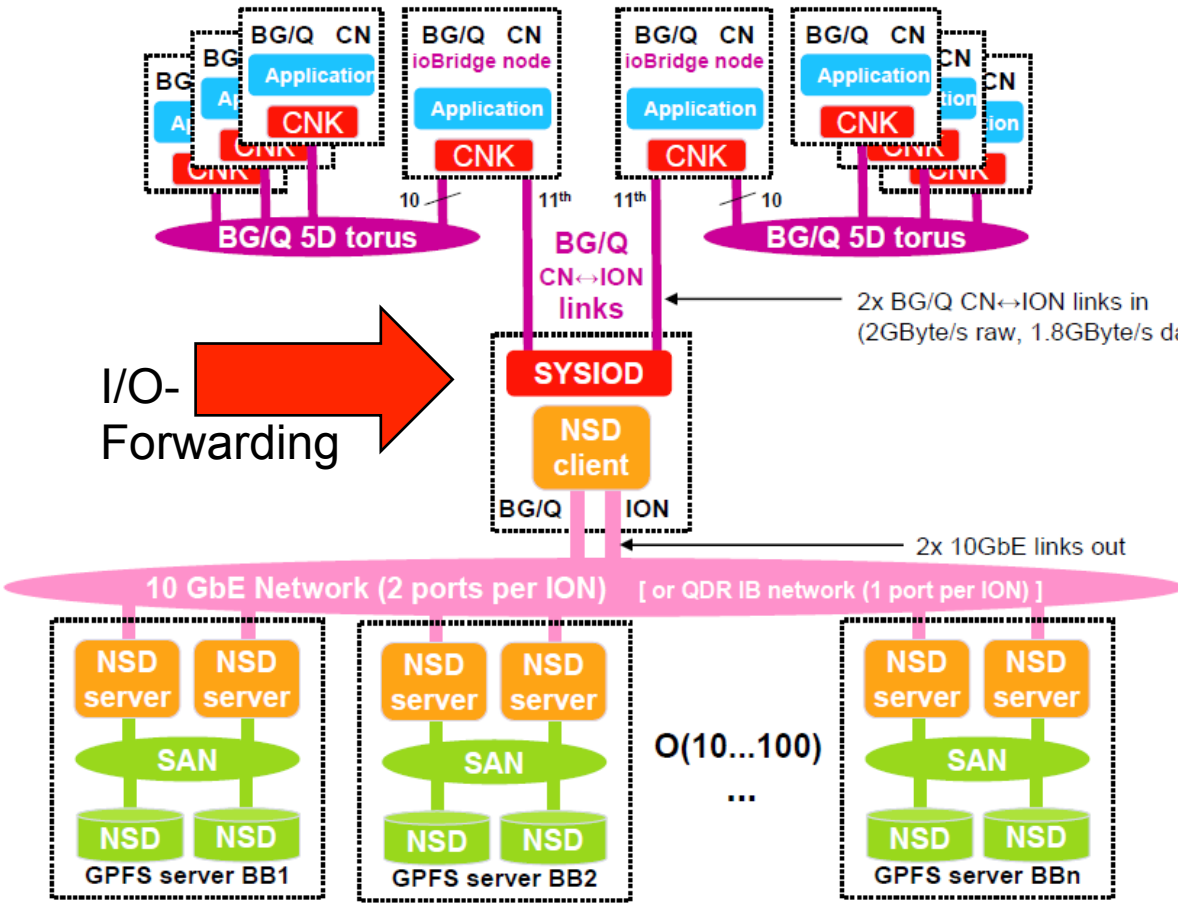


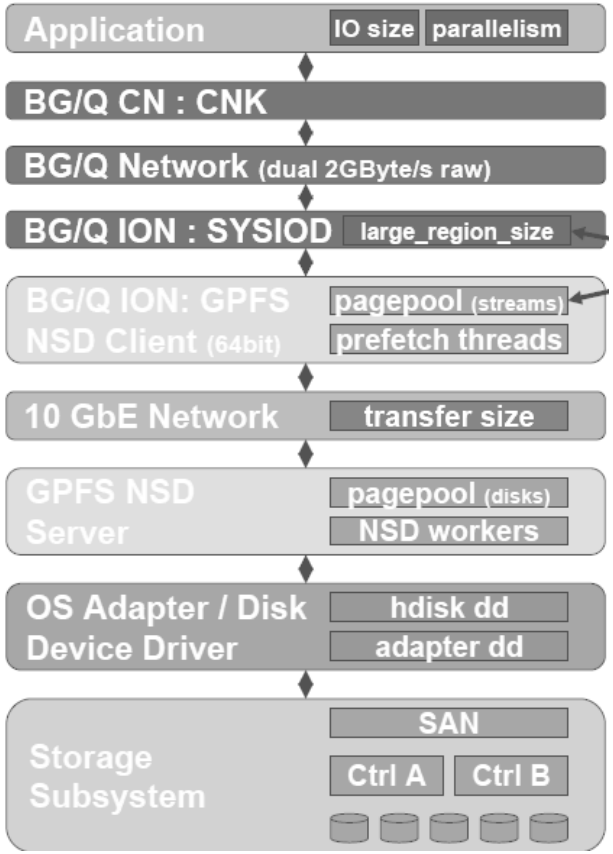
# File I/O to GPFS on JURECA



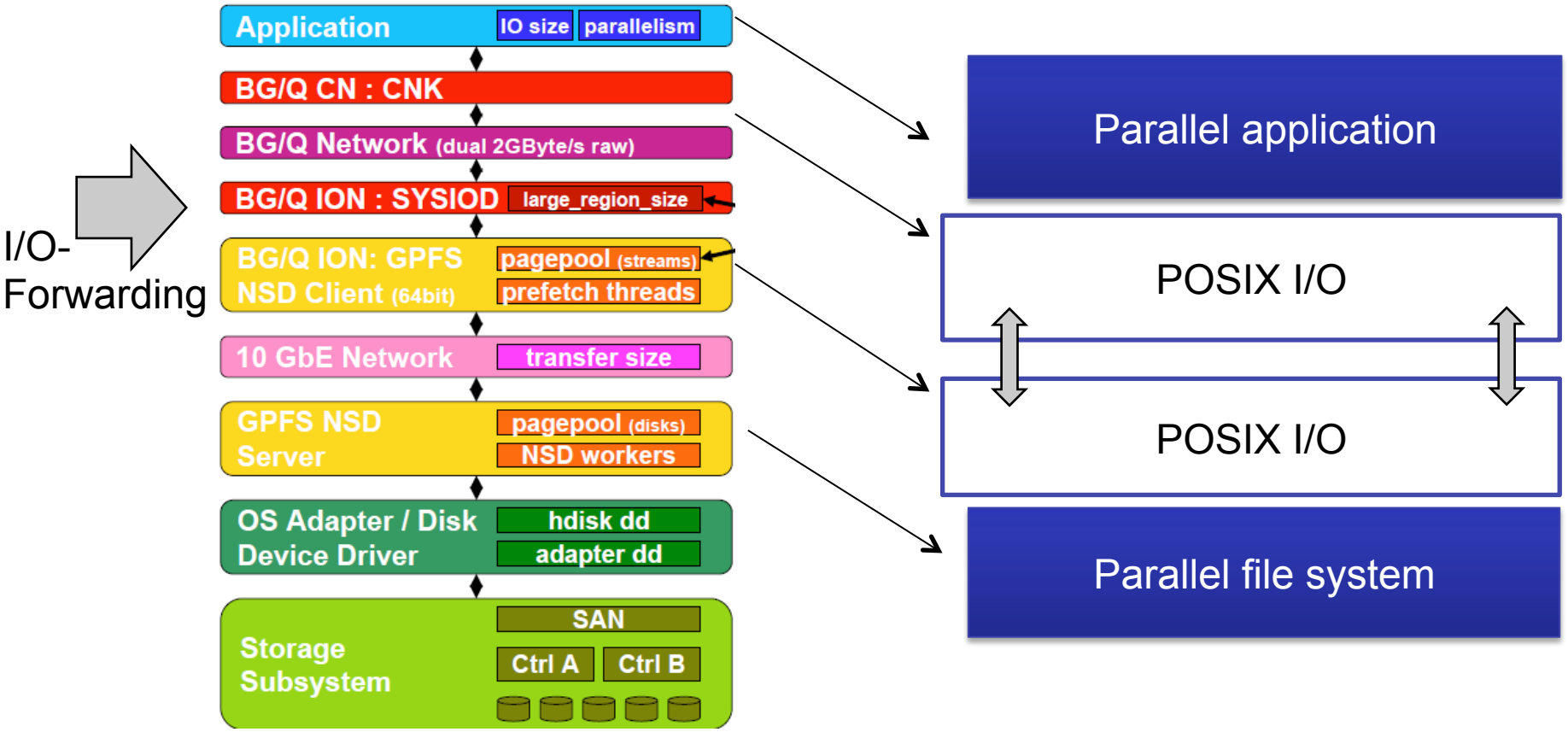



# File I/O to GPFS on JUQUEEN(I)

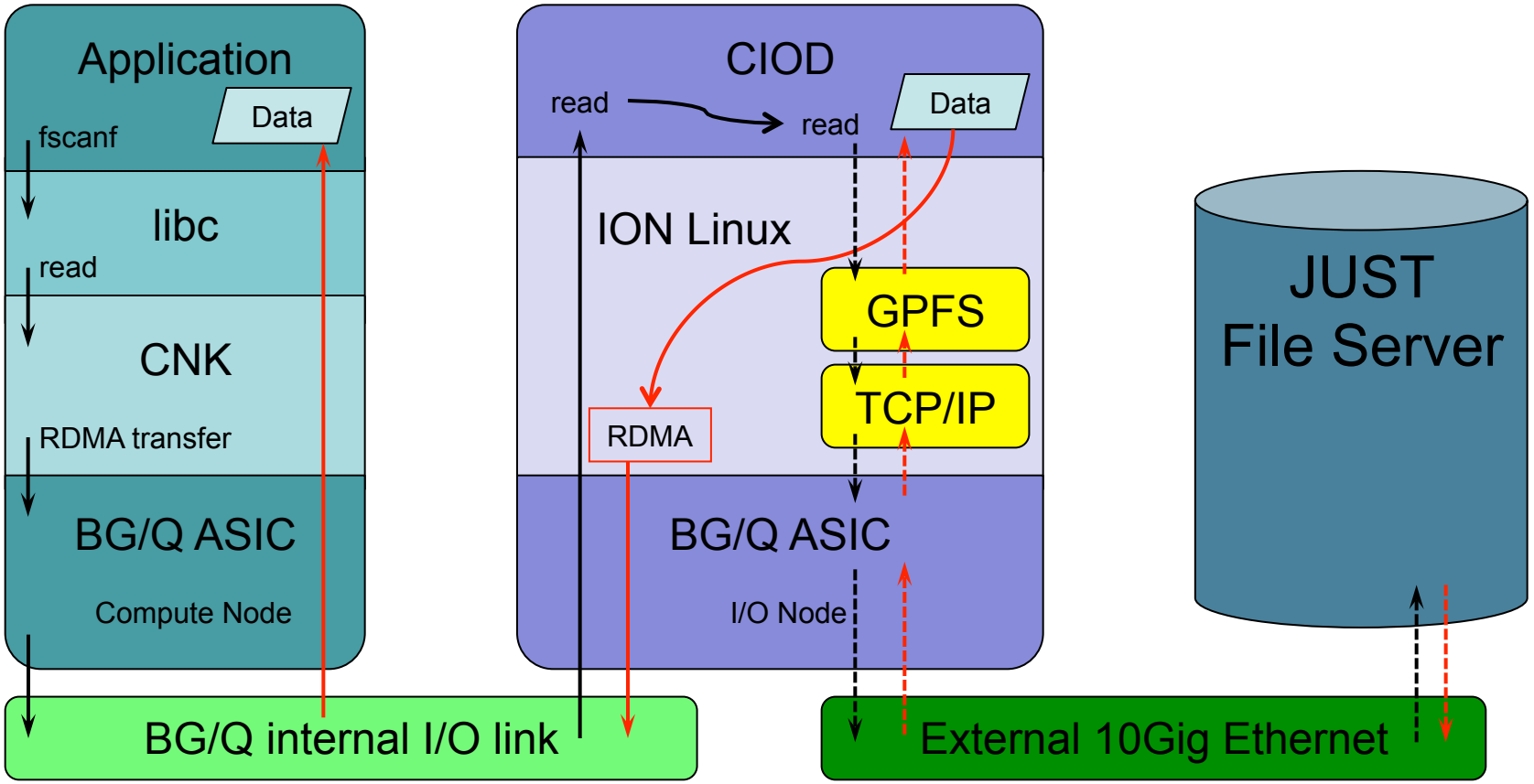


# File I/O to GPFS on JUQUEEN(II)



# File I/O to GPFS on JUQUEEN(III)



JÜLICH
FORSCHUNGSZENTRUM

# GPFS Storage Server

# GPFS Storage Server – Building Block

x3650 M4 Server

JBOD
Disk Enclosure

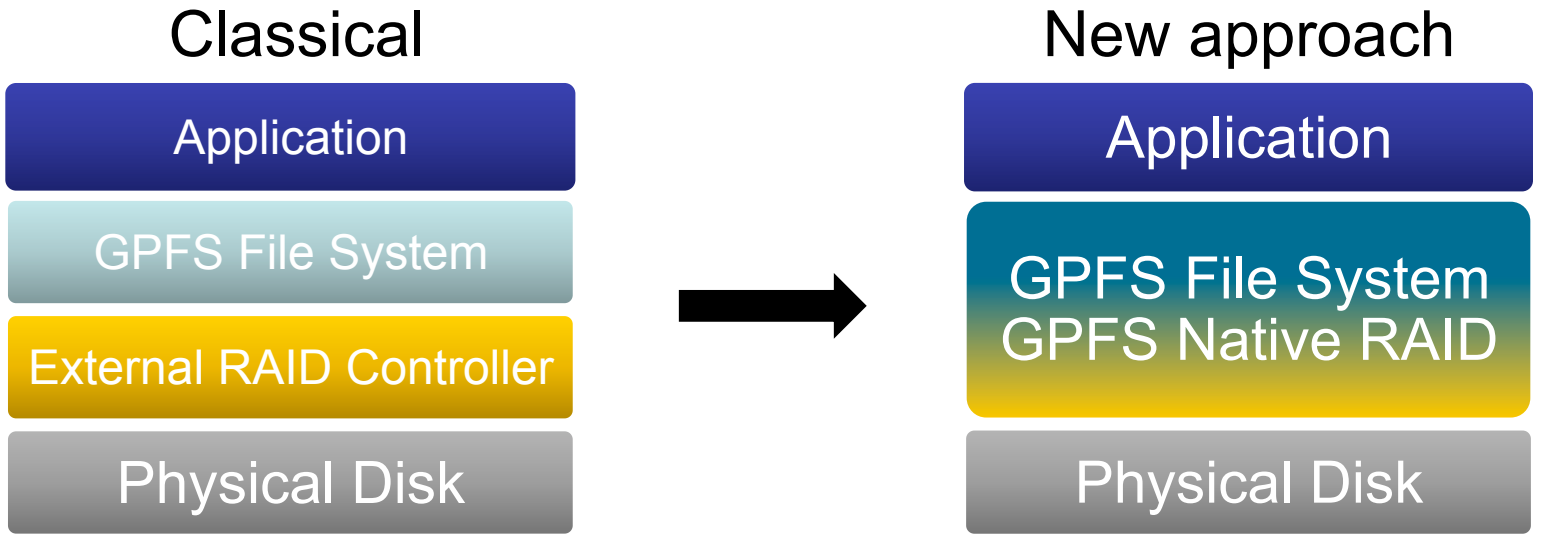
***GSS 24: Light and Fast***
2 x3650 servers +
4 JBOD 20U rack

***GSS 26: HPC Workhorse***
2 x3650 servers +
6 JBOD Enclosures, 28U

# GPFS Storage Server – Building Block

| GSS Model | JBODs (60 slots) | Disk Size | Gross Capacity | Spare Capacity | 8+2P Net Capacity | 8+3P Net Capacity |
|---|---|---|---|---|---|---|
| GSS-24 | 4 | **2 TB** | 464 TB (4*58*2TB) | 2 disks per DA (one DA has 58 disks) | **358 TB** (4*56*2TB*(8/10)) | **326 TB** (4*56*2TB*(8/11)) |
| | | **3 TB** | 696 TB (4*58*3TB) | | **538 TB** (4*56*3TB*(8/10)) | **489 TB** (4*56*3TB*(8/11)) |
| | | **4 TB** (GSS v1.5) | 928 TB (4*58*4TB) | | **717 TB** (4*56*4TB*(8/10)) | **652 TB** (4*56*4TB*(8/11)) |
| GSS-26 | 6 | **2 TB** | 696 TB (6*58*2TB) | | **538 TB** (6*56*2TB*(8/10)) | **489 TB** (6*56*2TB*(8/11)) |
| | | **3 TB** | 1044 TB (6*58*3TB) | | **806 TB** (6*56*3TB*(8/10)) | **733 TB** (6*56*3TB*(8/11)) |
| | | **4 TB** (GSS v1.5) | 1392 TB (6*58*4TB) | | **1075 TB** (6*56*4TB*(8/10)) | **977 TB** (6*56*4TB*(8/11)) |

# GPFS Native RAID - Motivation



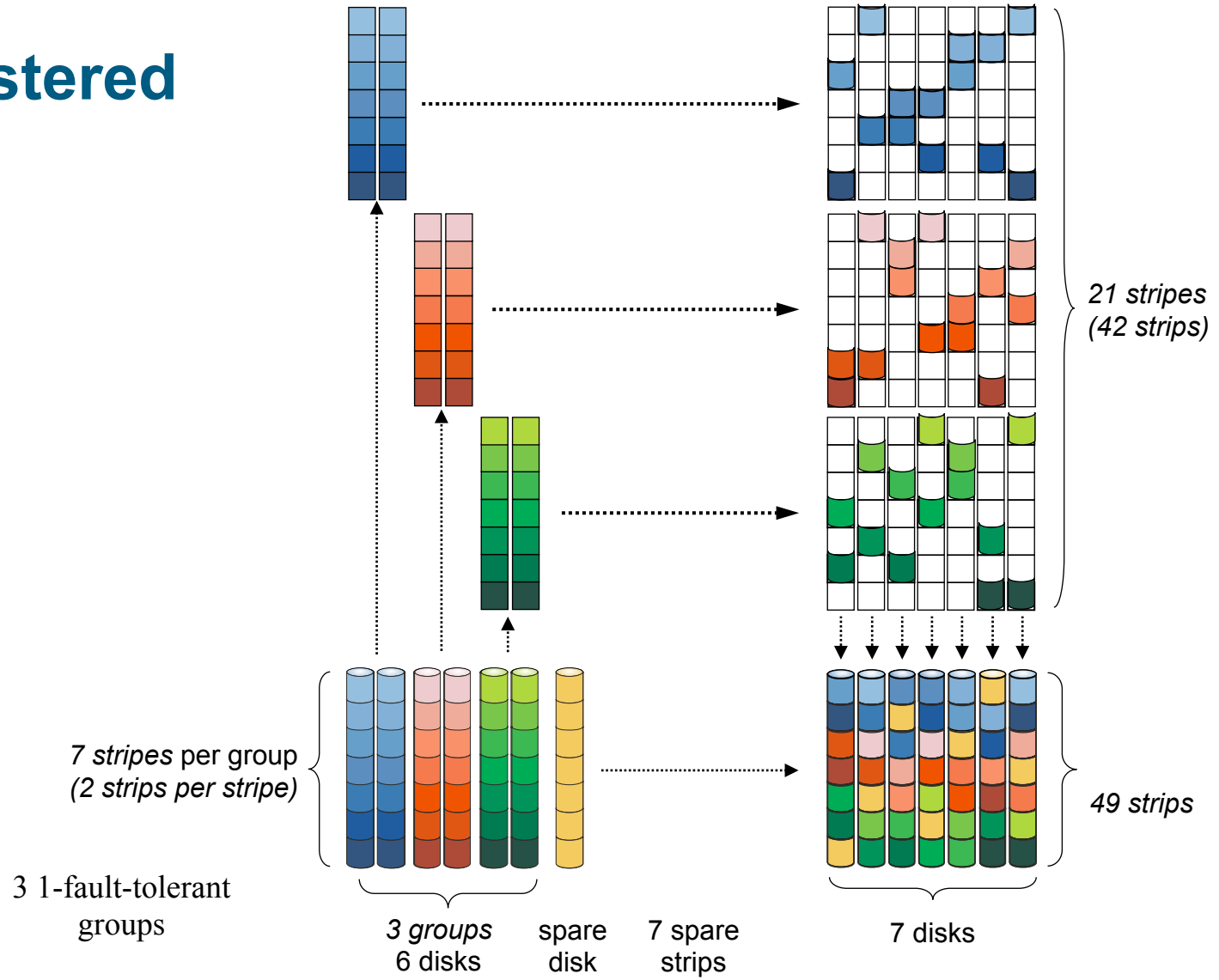

- Performance degradation on disk rebuild
- Silent data corruption

- Fast disk rebuild using Declustered RAID
- End-to-End data integrity (checksums, version no)

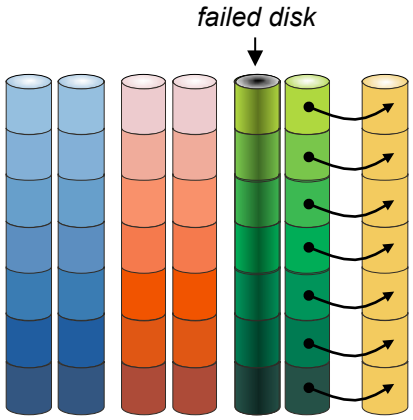
# GPFS Native RAID (GNR) - Features

- End-to-end checksums against silent data corruption
- Declustered RAID
- RAID codes:
  - 3-way/4-way replication
  - 8+2p/8+3p Reed Solomon
- Recovery Groups (failover)
- Disk hospital
  - Diagnoses errors/faults in storage subsystem

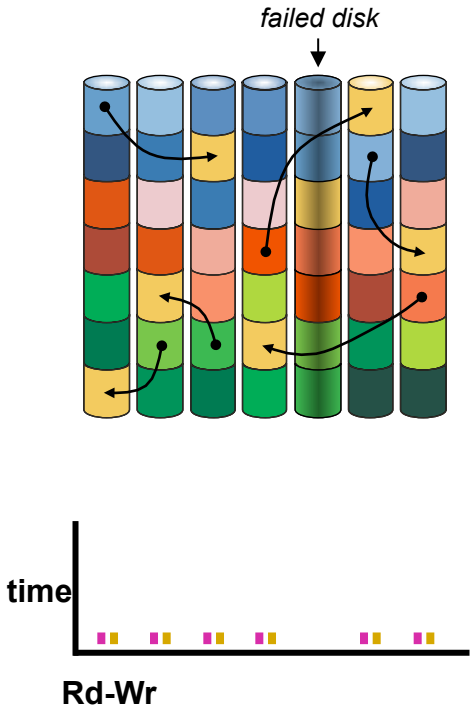
# Declustered RAID



# Declustered RAID Rebuild Example - Single Fault




Disk failure causes disk rebuild
➔ Volume degraded for a long time
➔ performance impact for file system




Disk failure causes strips rebuild
➔ all disc involved
➔ Volume degraded for a short time
➔ minimized performance impact