

ScaleIO on Tornado++

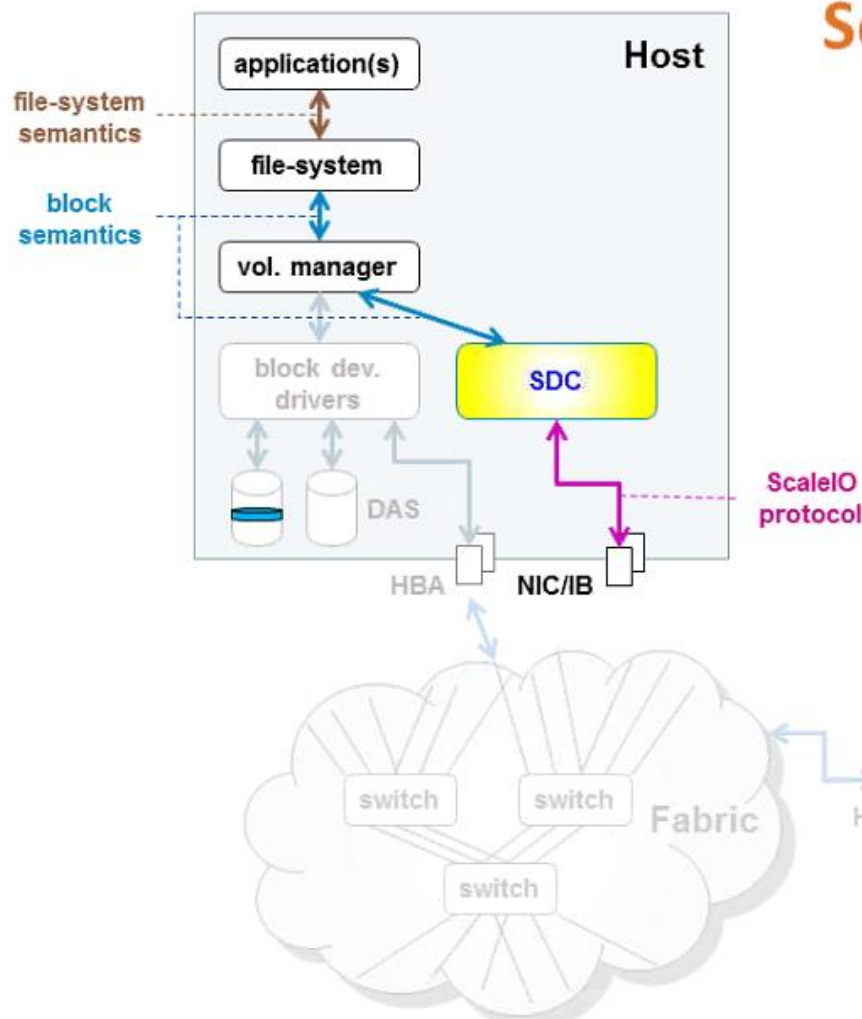
Citi Internal

April 20th 2015

Core, fundamental features

- **Converged**
 - Run alongside applications
- **Highly Scalable**
 - Scale to 100s or 1000s of nodes
 - High I/O parallelism
 - Scale linearly w/applications (eliminate I/O bottlenecks)
- **Elastic**
 - Add, move, remove nodes or disks “on the fly”
 - Auto-rebalance
- **Resilient**
 - 2 copy “mesh” mirroring
 - Fast, auto-rebuild
 - Extensive failure handling / HA
- **Platform Agnostic (physical or virtual)**
 - Runs on any x86 server (windows will be supported very soon) and ARM
 - Bare-metal or virtual (i.e. ESX, Xen, Hyper-V)
 - SSDs/PCIe or HDDs
 - Agnostic to network speed
- **Ease of management**
 - Storage as just another application
 - Any IT admin can manage entire datacenter stack
- **Partitioning / Tiering / Multi-Tenancy**
 - Protection-domains
 - Storage pools for multi-tiering
 - Bandwidth/IOPs limiter
- **Snapshots** (writeable)





ScaleIO Data Client (SDC)

Exposes ScaleIO shared block volumes to the application

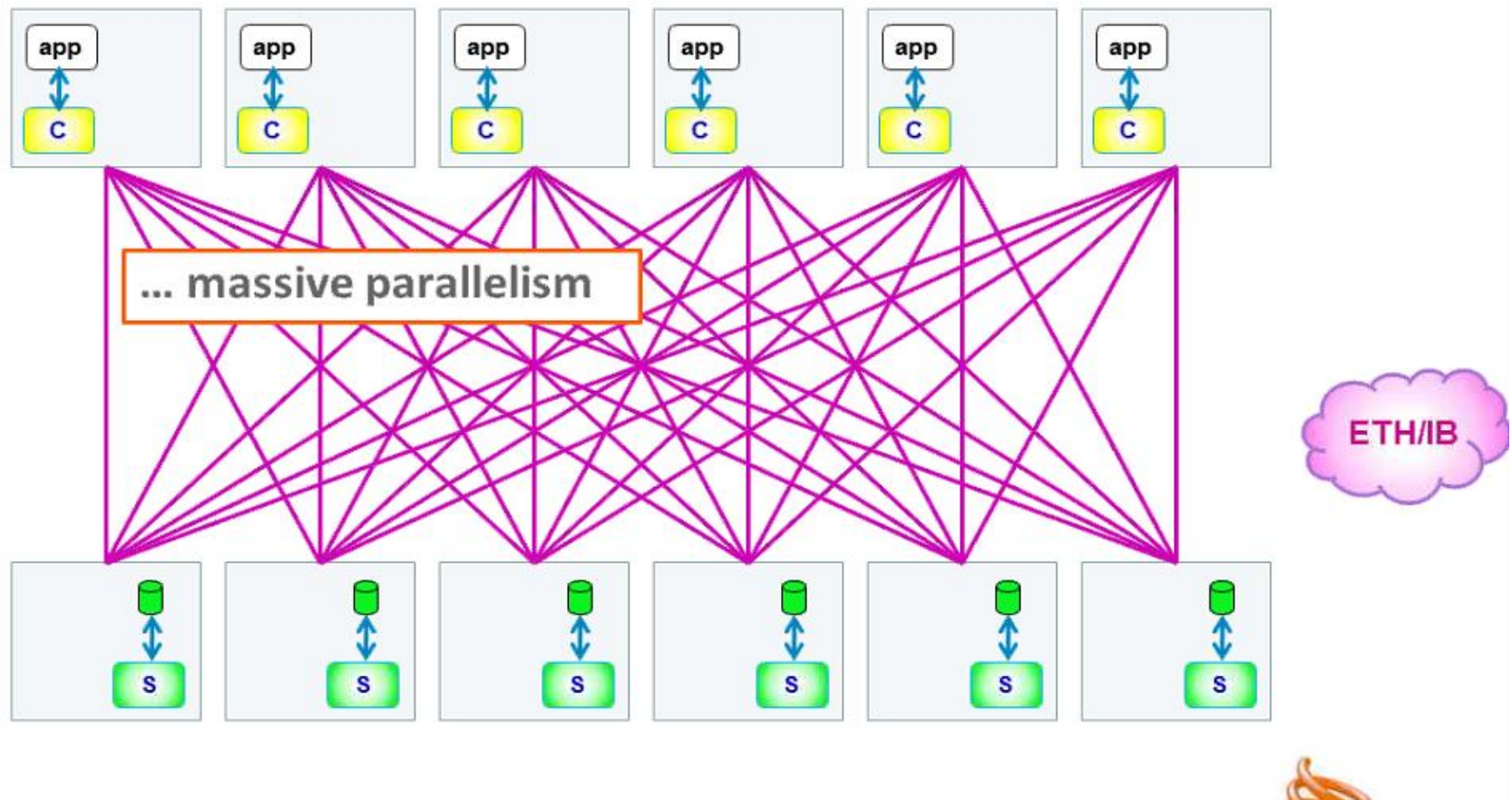
Notes

Access to OS partition may still be done "regularly"

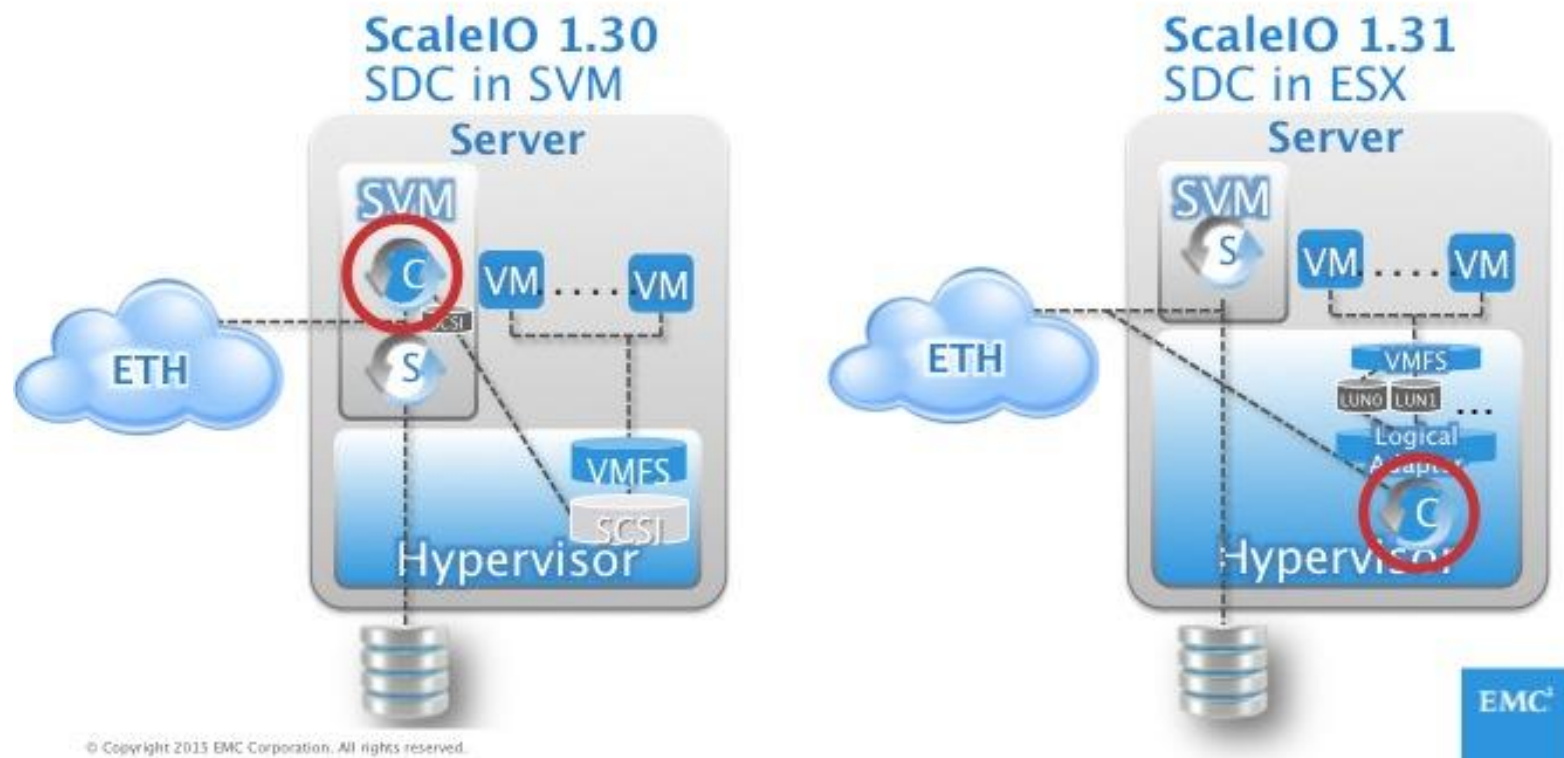
ScaleIO data client (SDC) is a block device driver



2-layer configuration

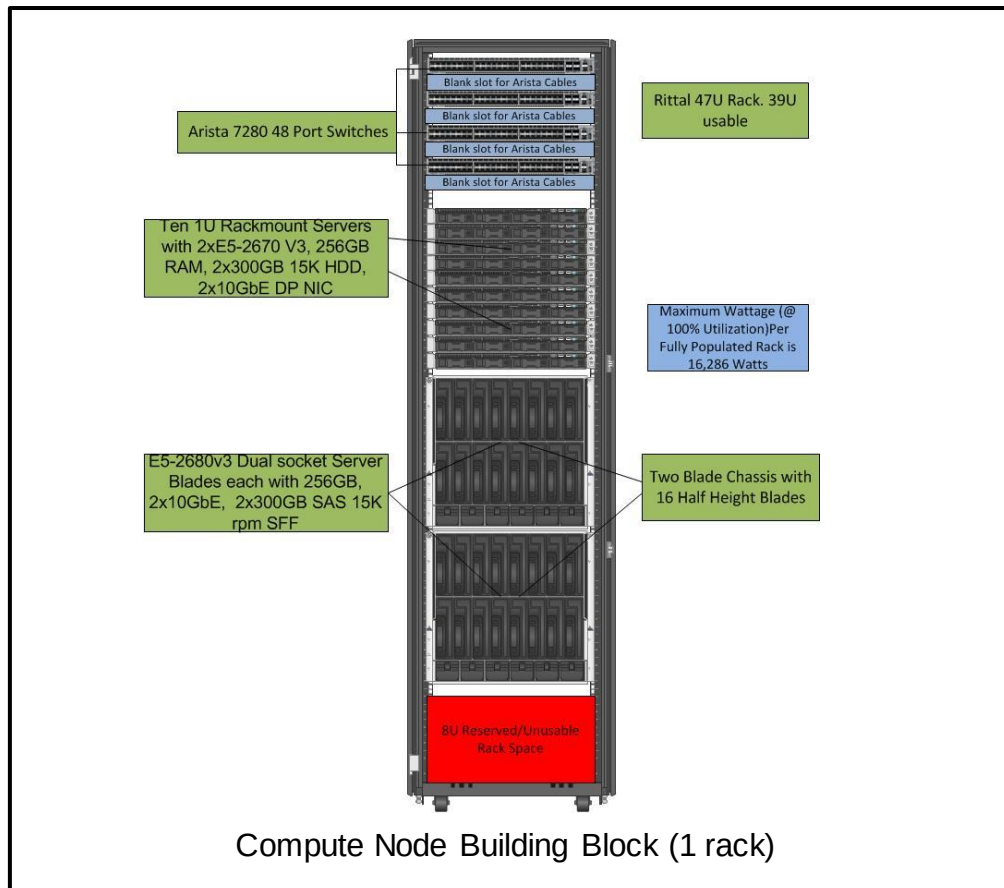


ScaleIO & VMware Environments



Citi POD Definitions (Separate Compute and Storage)

A Compute Building Block consists of a single rack of Compute, and is the smallest increment that can be added to a POD.

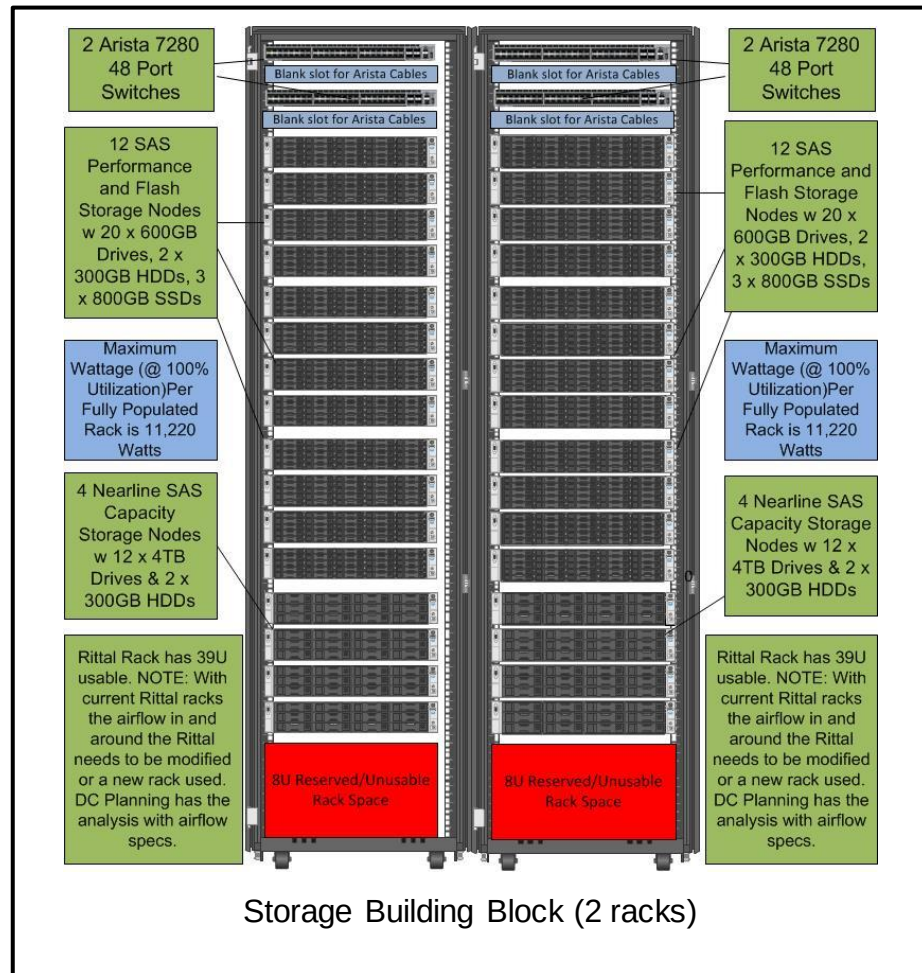


Compute Building Block

- 1 Rack
- 4 Arista 7280 Switches
- 42 Compute Servers
 - 2 blade chassis/32 Blades (768 Cores)
 - 10 Rackmount Servers (240 Cores)
 - 1,008 Cores combined

Citi POD Definitions (Separate Compute and Storage)

A Storage Building Block consists of a pair of Storage racks, and is the smallest increment that can be added to a POD.

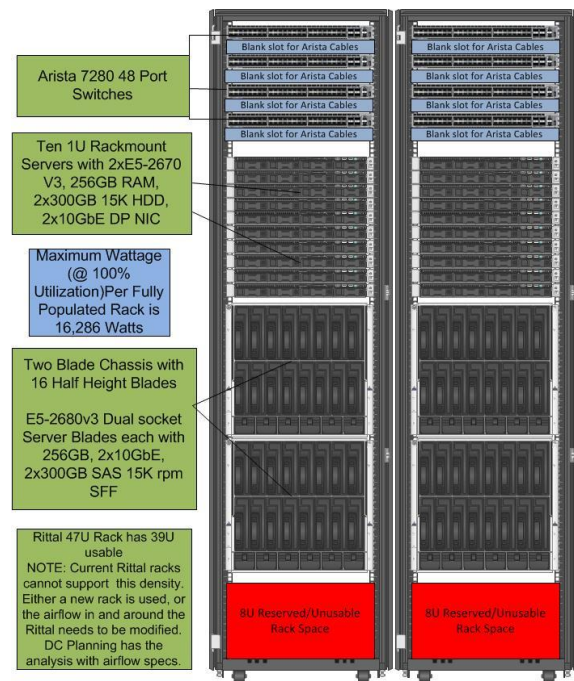


Storage Building Block

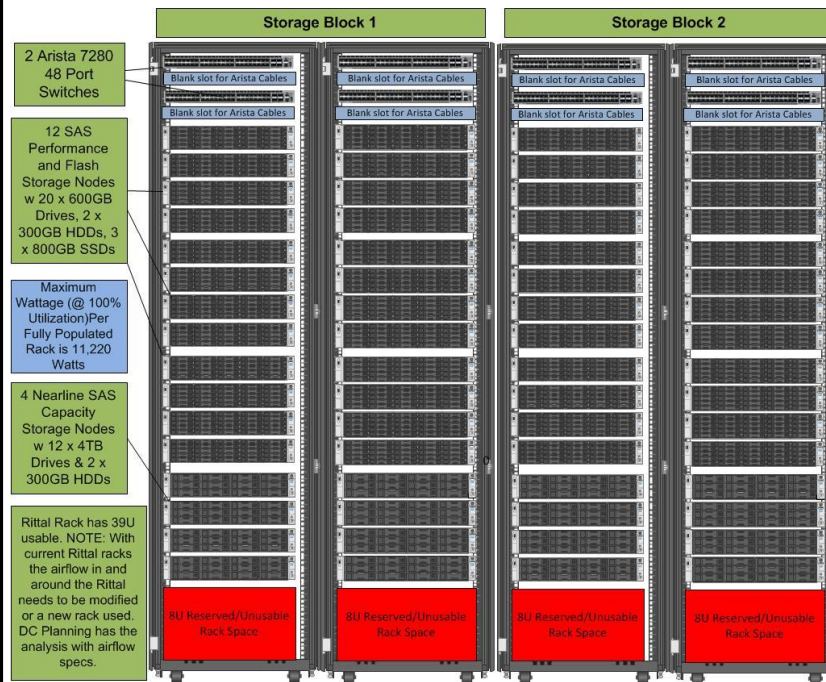
- 2 Racks
- 4 Arista 7280 Switches (2 in each rack)
- 32 Storage Servers
 - 24 Servers with 20 x 600GB 10K SAS Drives, and 3 x 800GB SAS SSD Drives
 - 8 Servers with 12 x 4TB 7.2K SAS Drives
- Total Combined RAW Storage is 729.6 TB:
 - 57.6 TB of Flash SSD
 - 288 TB of SAS (600GB Drives)
 - 384 TB of SAS (4TB Drives)

Citi POD Definition (Separate Compute and Storage)

- A Compute Node Building Block consists of 1 Compute Rack.
- A Storage Node Building Block consists of 2 Storage Racks which are cross connected.
 - Minimum feasible POD configuration requires 2 Compute Blocks and 2 Storage Blocks.



2 Compute Node Building Blocks (2 racks)



2 Storage Node Building Blocks (4 racks)

POD Design

- 6 Racks (2 compute, 4 Storage)
- 16 Arista 7280 Switches
- 84 Compute Servers
 - 4 blade chassis/64 Blades (1,536 Cores)
 - 20 Rackmount Servers (480Cores)
 - 2,016 Cores combined
- 64 Storage Nodes, Total Combined RAW Storage is 1,459.2 TB:
 - 115.2 TB of Flash SSD
 - 576 TB of SAS (600GB Drives)
 - 768 TB of SAS (4TB Drives)

ScaleIO performance estimates for Tornado++

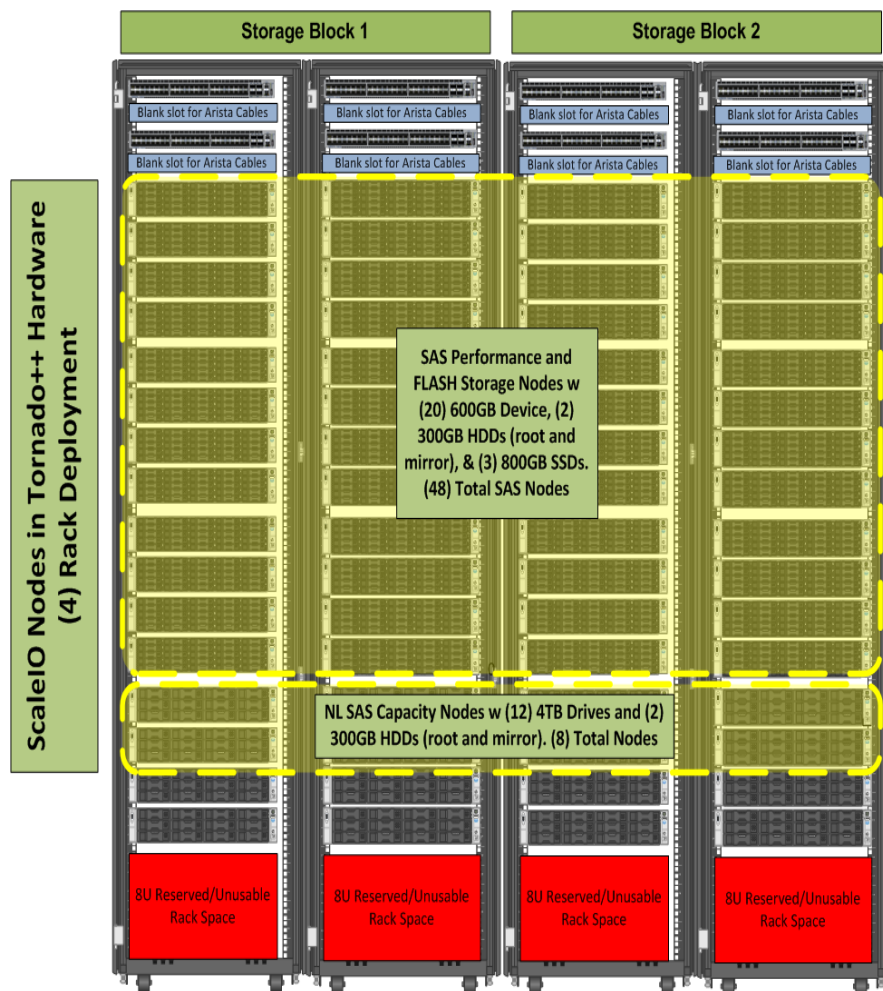
ScaleIO on Tornado++						
Tier	Racks in pod	Servers/rack	Disks/server	Total disks per pod	Max POD IOPS Host IOPS (50% Read / 50% write)*	Sustained Host IOPS/usable TB (50% Read / 50% write)*
800GB SSD	4	12	3	144	Up to 320,000**	Up to 70,000**
600GB SAS	4	12	20	960	100,000	440
4TB NL-SAS	4	2***	12	96	4160	30

* Assuming 40% read hits to SAS tier. 10% spare capacity. 85% utilization of remaining storage. Other bottlenecks may exist.

** Estimated for small IOs. Not tested.

*** Assumes only 2 of the 4 NL-SAS nodes in each rack is used for ScaleIO capacity-optimized tier

ScaleIO on Tornado++ Hardware



The ScaleIO environment would have the following capacity for each Tier.

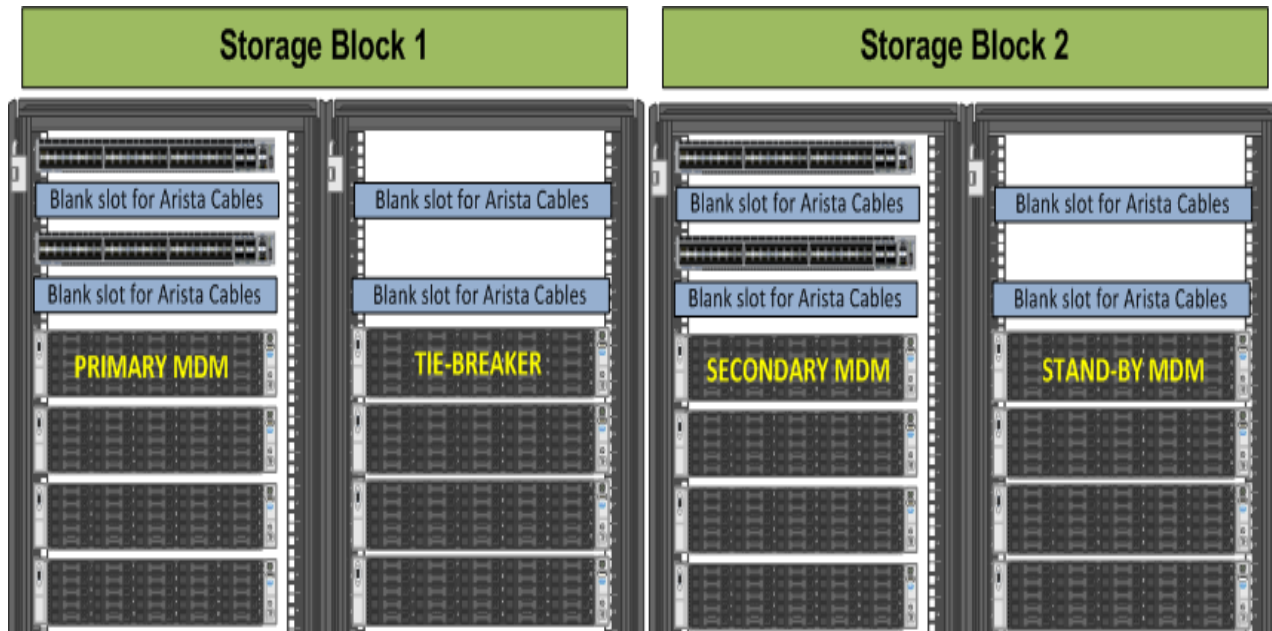
SSD Capacity – (144) 800GB SSDs
= 112.5 TB Raw SSD

SAS Capacity – (960) 600GB SAS Disks
= 562.5 TB Raw SAS

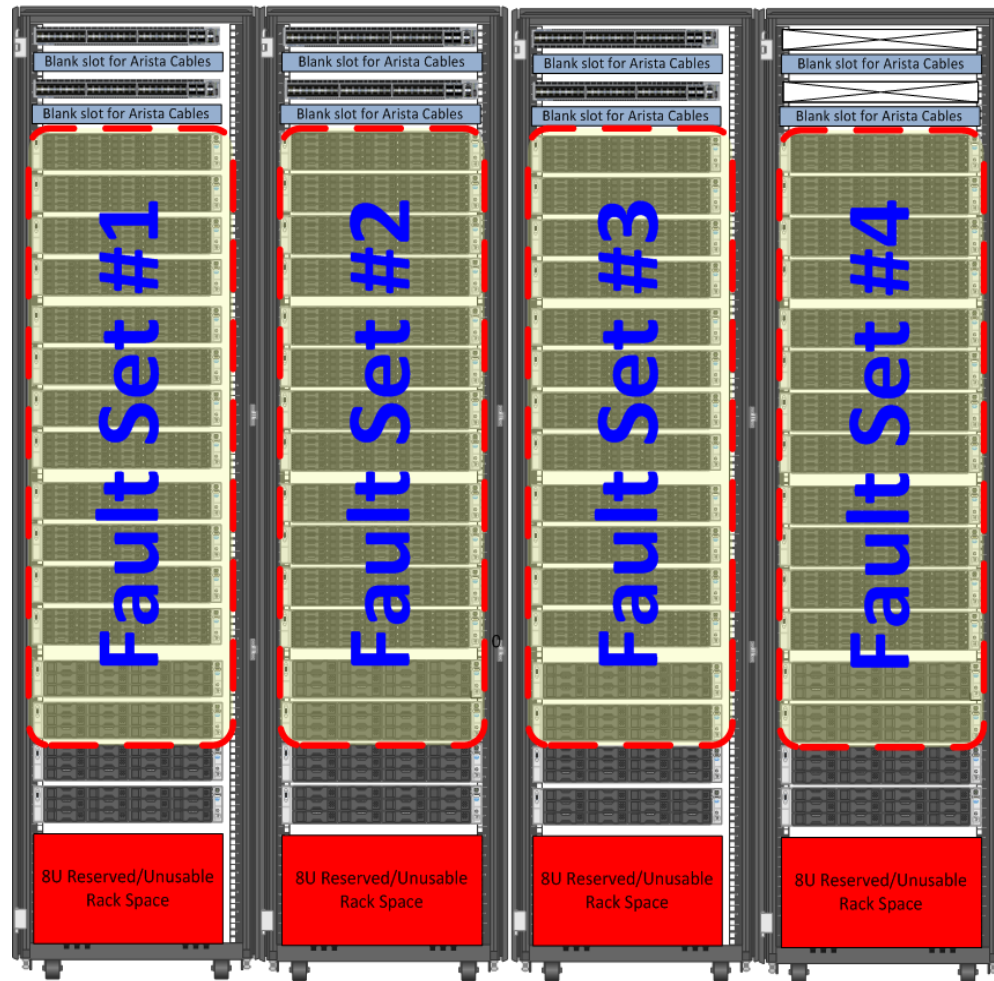
NL-SAS Capacity – (96) 4TB NS-SAS Disks
= 384 TB Raw NL-SAS

When calculated, the total raw capacity of ScaleIO running on Tornado++ hardware will be ~1.03 PB. When factoring in mirroring and spare capacity, the usable total space across all pools is roughly ~480TB.

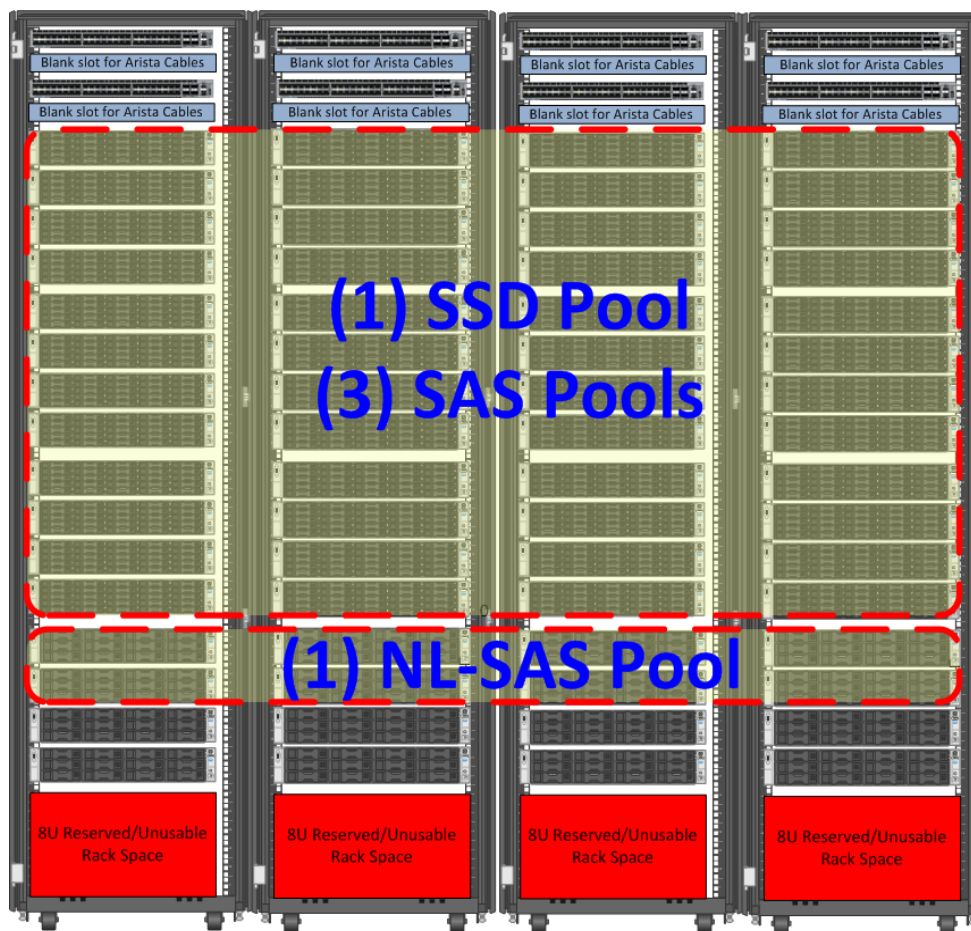
MDM and Tie-Breaker Placement in Racks



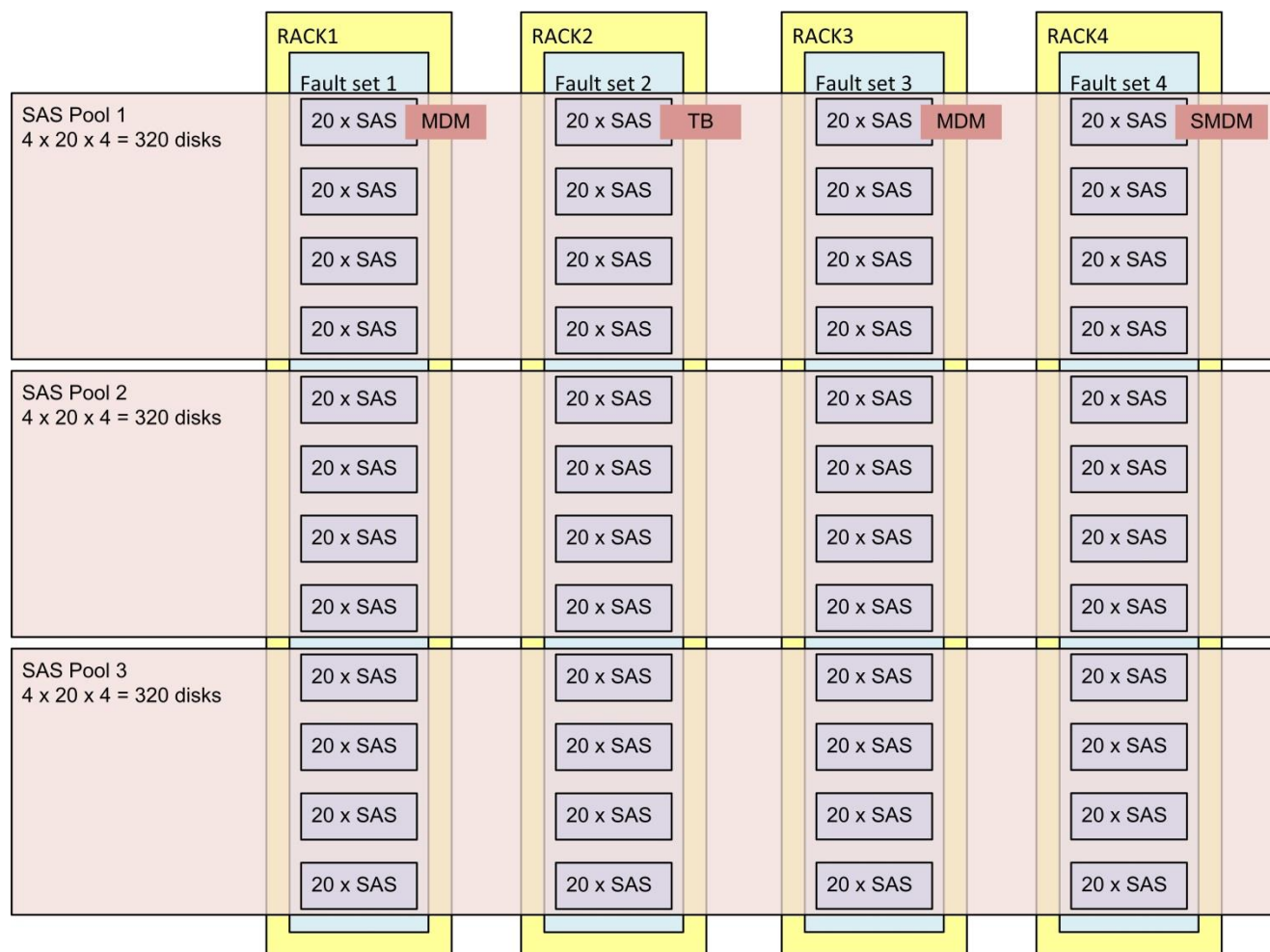
Fault Sets layout in Tornado++



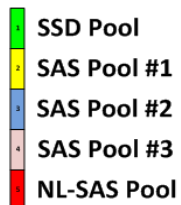
Data Layout on Pools



Example of SAS pool layout



Total disks = 12 x 80 = 960 disks
Total pools = 12



ScaleIO Network Configuration on Tornado++

