Multi-Rail Infiniband Systems

APPRO INTERNATIONAL INC

APPRO

HPC Cluster Solutions

University of Tsukuba
筑波大学

Tsukuba University Cluster

- 10,784 Cores
- Quad-rail IB
- 95 TFlops

APPRO
HPC Cluster Solutions

Tsukuba University

:: Tsukuba City Japan

- Tsukuba RFP Requirements (High Points)
  - Minimum Peak Performance 80TF
  - 16 or More CPU Cores per Node
  - 2GB of Memory per Core (32GB or More per Node)
  - 64bit X86 Architecture with 2.3GHz Processors
  - 40GB/sec minimum Memory Bandwidth
  - 5GB/sec Node to Node Unidirectional (measured)
  - 10GB/sec Node to Node Full Duplex (measured)
  - 7us or Lower Latency Node to Node
  - 250GB minimum Local disk per Node
  - 400TB Global Storage with 10GB/sec aggregate BW
  - 1Kg per M maximum Floor Loading
- System + 3y Maintenance approximately $20M

OPENFABRICS ALLIANCE

APPRO HPC Cluster Solutions

Tsukuba Configured

:: As Delivered

- 95TF Peak Performance
- 638 Compute Quad Socket Quad Core Compute Nodes
- 4 Online Spares
- 10 Login Nodes with 10GbE
- 20 Sub-Management Nodes
- 2 Management Nodes
- Dual Full Bisection BW Fat-Tree Networks
- Four Mellanox ConnectX Cards per Node
- Two DDR IB Connections to each Network per Node
- 32GB of 667MHz DDR2 ECC Memory per Node
- Four 250GB Disk per Node in RAID0 Configuration
- Fault Tolerant IB and Ethernet Networks

95TF Peak

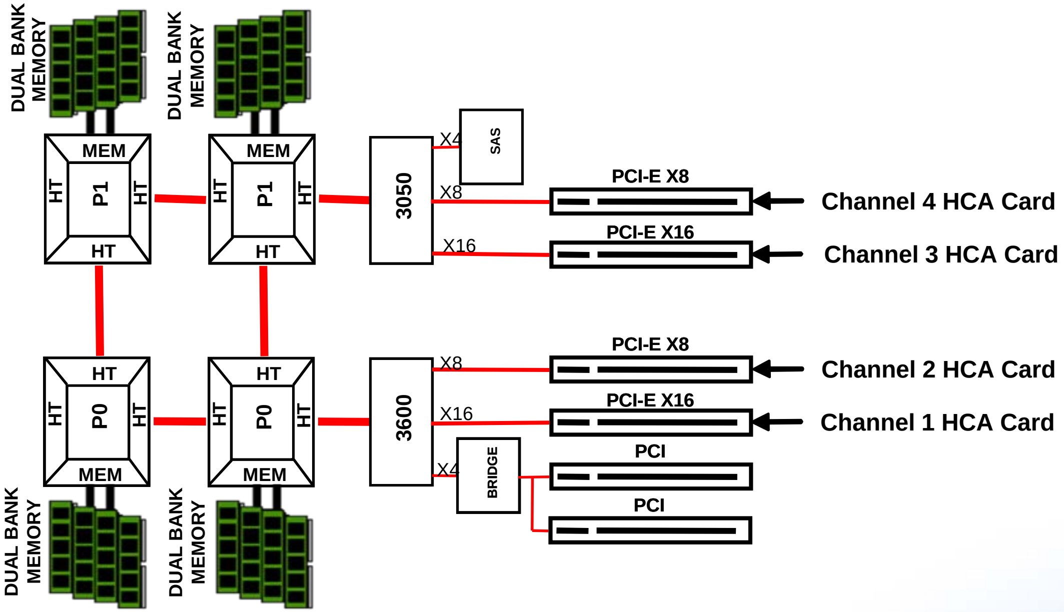
:: As Delivered

- Single 674Node Cluster with 10,784PEs
- Single Network with equal 6GB/sec between Cores
- Diskless System with Multiple Virtual Clusters
- Fault Tolerant Management Network and Mgmt Nodes
- 400TB DDN Storage with Luster
- 20 Dual Socket Nodes for OSS, Meta-Data Servers
- Total 744KVA
- 190 Tons
- 9 Scalable Units with 70+2 Nodes Each
- 1 Management and Storage SU

Comput Nodes

:: IO Was a Challenge

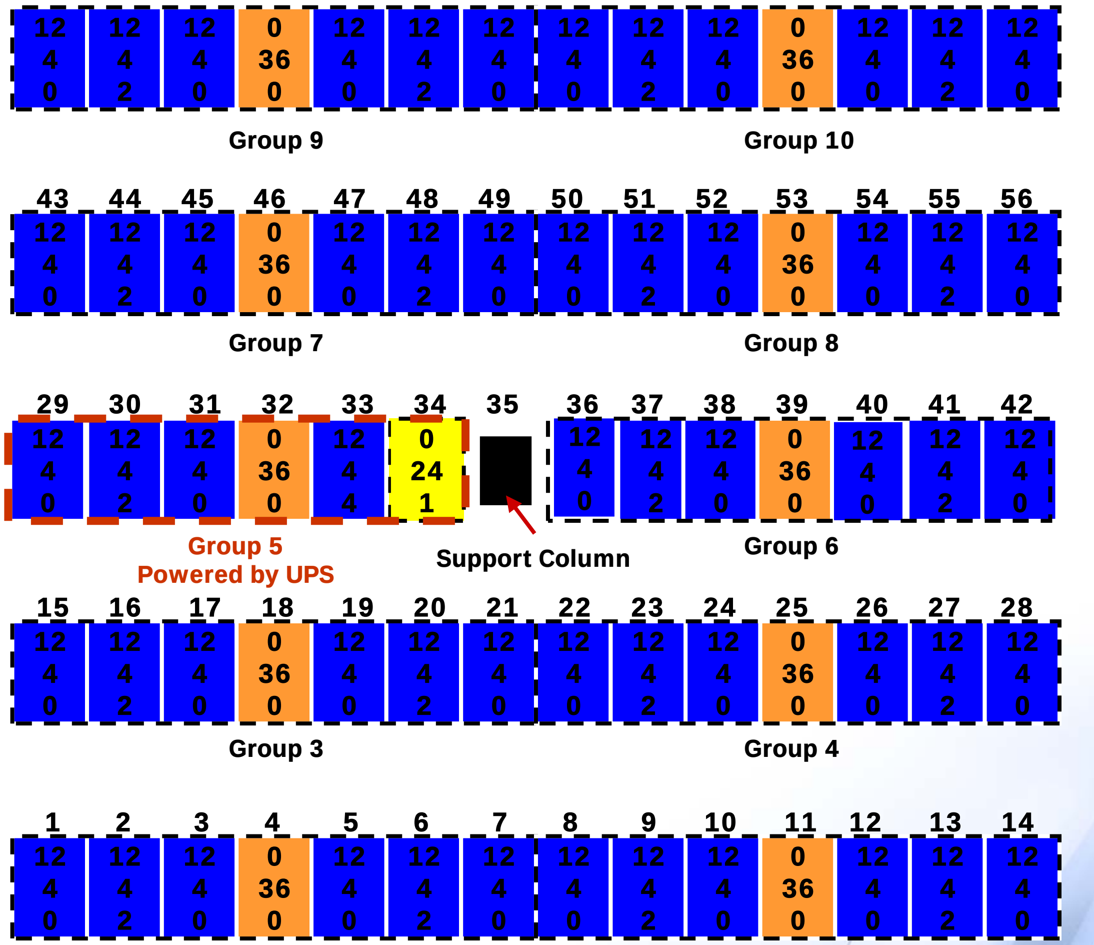
I/O Configuration for Quad-Socket 2U Servers



Floor Layout

:: 1kg/sq m is not very dense

Proposed Floor Layout with 600kg/m^2 Floor Loading Limit



57	58	59	60	61	62	63	64	65	66	67	68	69	70
12 4 0	12 4 2	12 4 0	0 36 0	12 4 0	12 4 2	12 4 0	12 4 0	12 4 2	12 4 0	0 36 0	12 4 0	12 4 2	12 4 0

Group 9

Group 10

43	44	45	46	47	48	49	50	51	52	53	54	55	56
12 4 0	12 4 2	12 4 0	0 36 0	12 4 0	12 4 2	12 4 0	12 4 0	12 4 2	12 4 0	0 36 0	12 4 0	12 4 2	12 4 0

Group 7

Group 8

29	30	31	32	33	34	35	36	37	38	39	40	41	42
12 4 0	12 4 2	12 4 0	0 36 0	12 4 4	0 24 1	■	12 4 0	12 4 2	12 4 0	0 36 0	12 4 0	12 4 2	12 4 0

Group 5
Powered by UPS

Support Column

Group 6

15	16	17	18	19	20	21	22	23	24	25	26	27	28
12 4 0	12 4 2	12 4 0	0 36 0	12 4 0	12 4 2	12 4 0	12 4 0	12 4 2	12 4 0	0 36 0	12 4 0	12 4 2	12 4 0

Group 3

Group 4

1	2	3	4	5	6	7	8	9	10	11	12	13	14
12 4 0	12 4 2	12 4 0	0 36 0	12 4 0	12 4 2	12 4 0	12 4 0	12 4 2	12 4 0	0 36 0	12 4 0	12 4 2	12 4 0

Group 1

Group 2

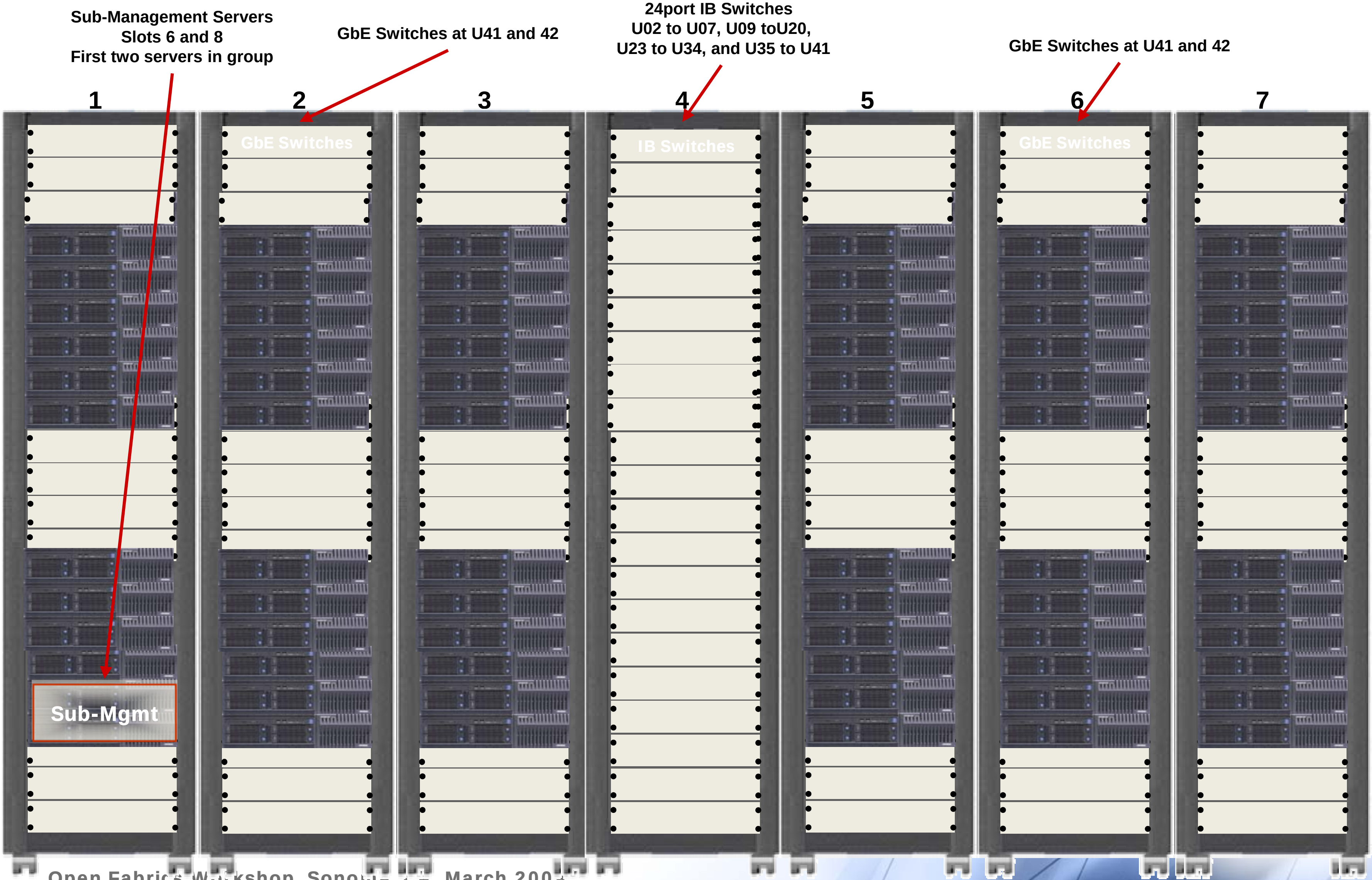
OPENFABRICS ALLIANCE

APPRO HPC Cluster Solutions

Scalable Unit

Group 1

:: 70 Compute Nodes plus 2 Mgmt Nodes



Sub-Management Servers
Slots 6 and 8
First two servers in group

GbE Switches at U41 and 42

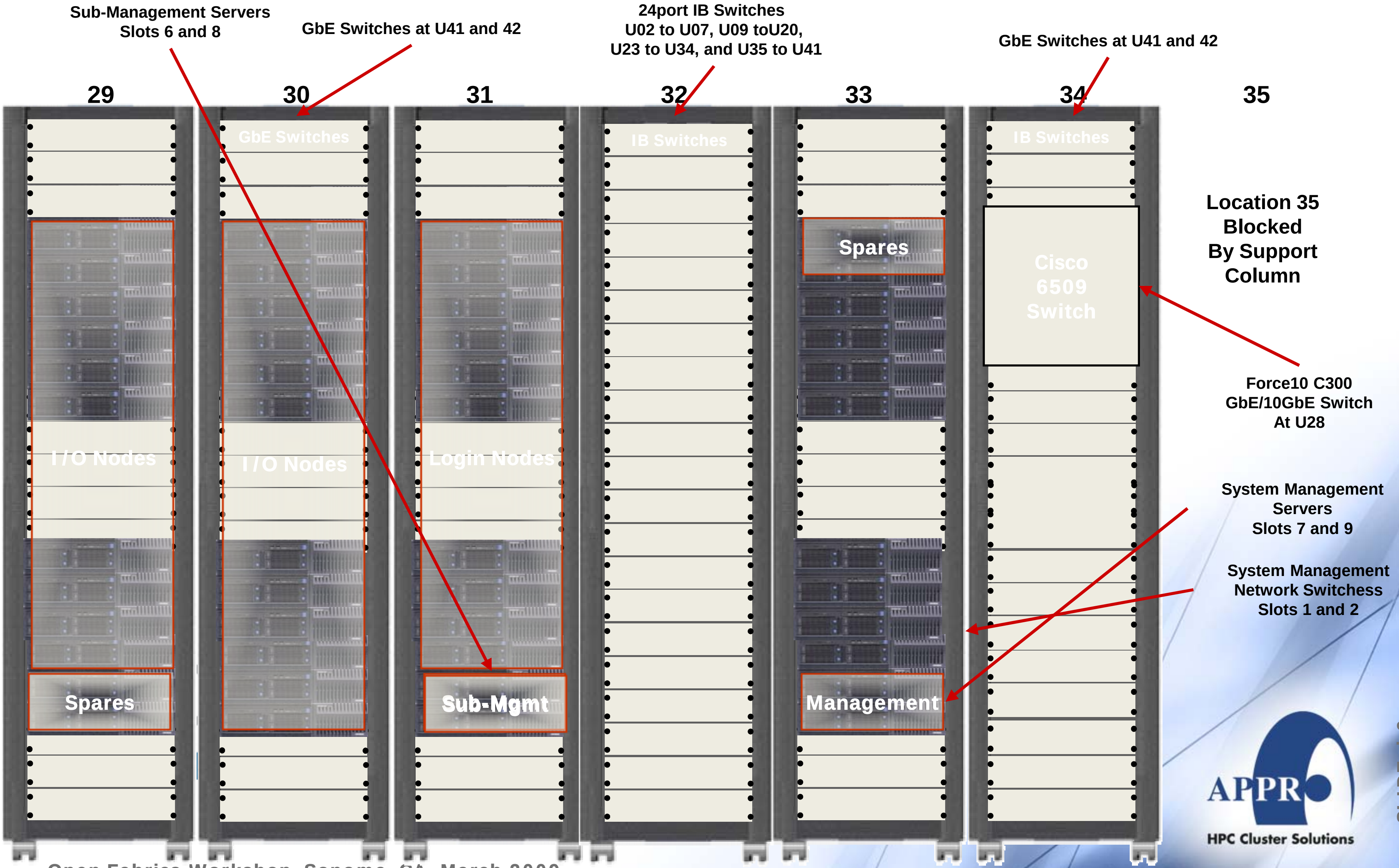
24port IB Switches
U02 to U07, U09 toU20,
U23 to U34, and U35 to U41

GbE Switches at U41 and 42

Scalable Unit

Group 5

:: Mgmt Nodes, Storage Nodes, Spares



APPRO
HPC Cluster Solutions

Cable Layout

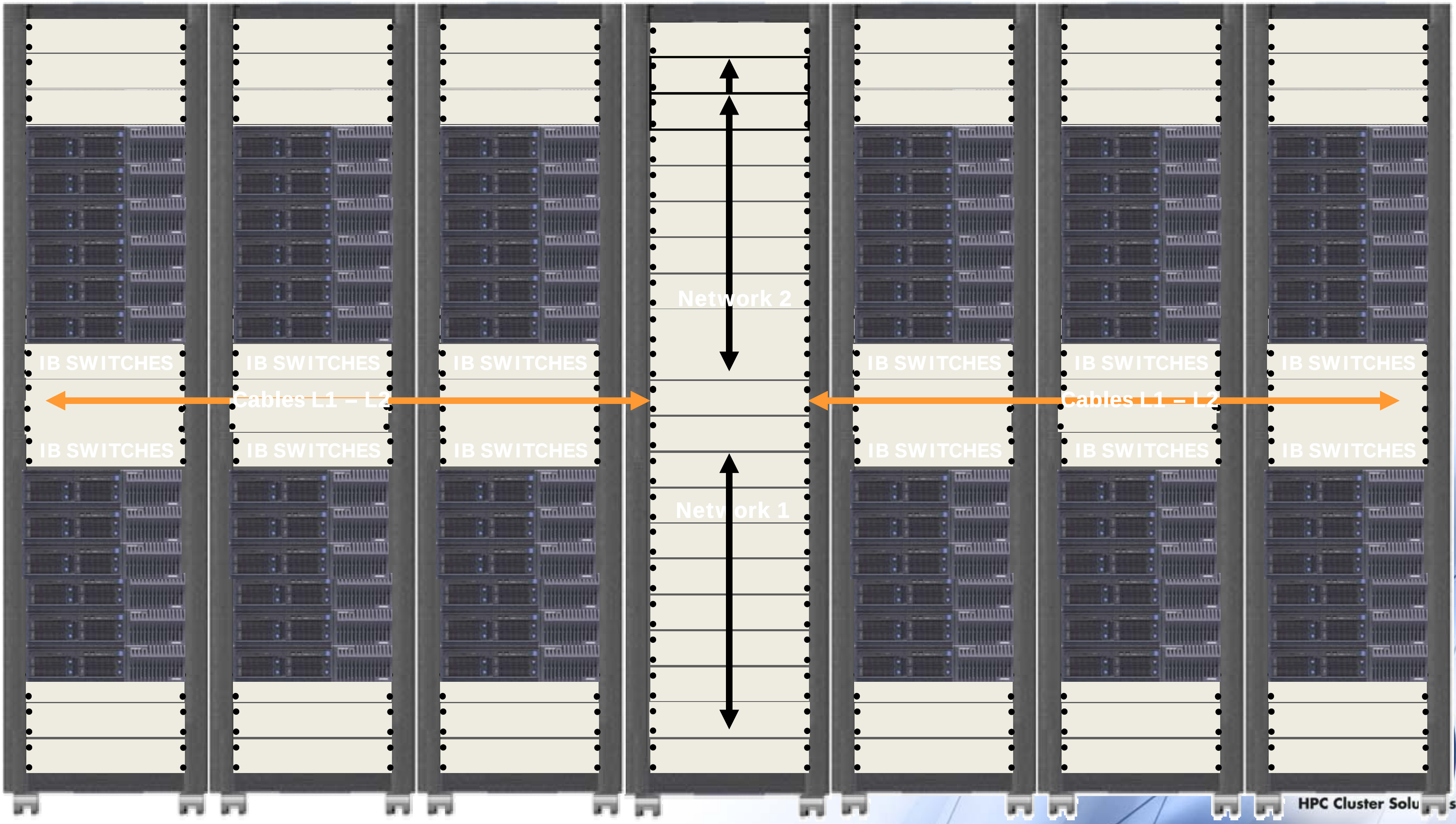
Equipment Rack Configuration with 600kg/m^2 Floor Loading Limit

:: Mostly Short Copper Cables

48 Cables from each compute rack can be routed inside the racks at slots 20 to 23

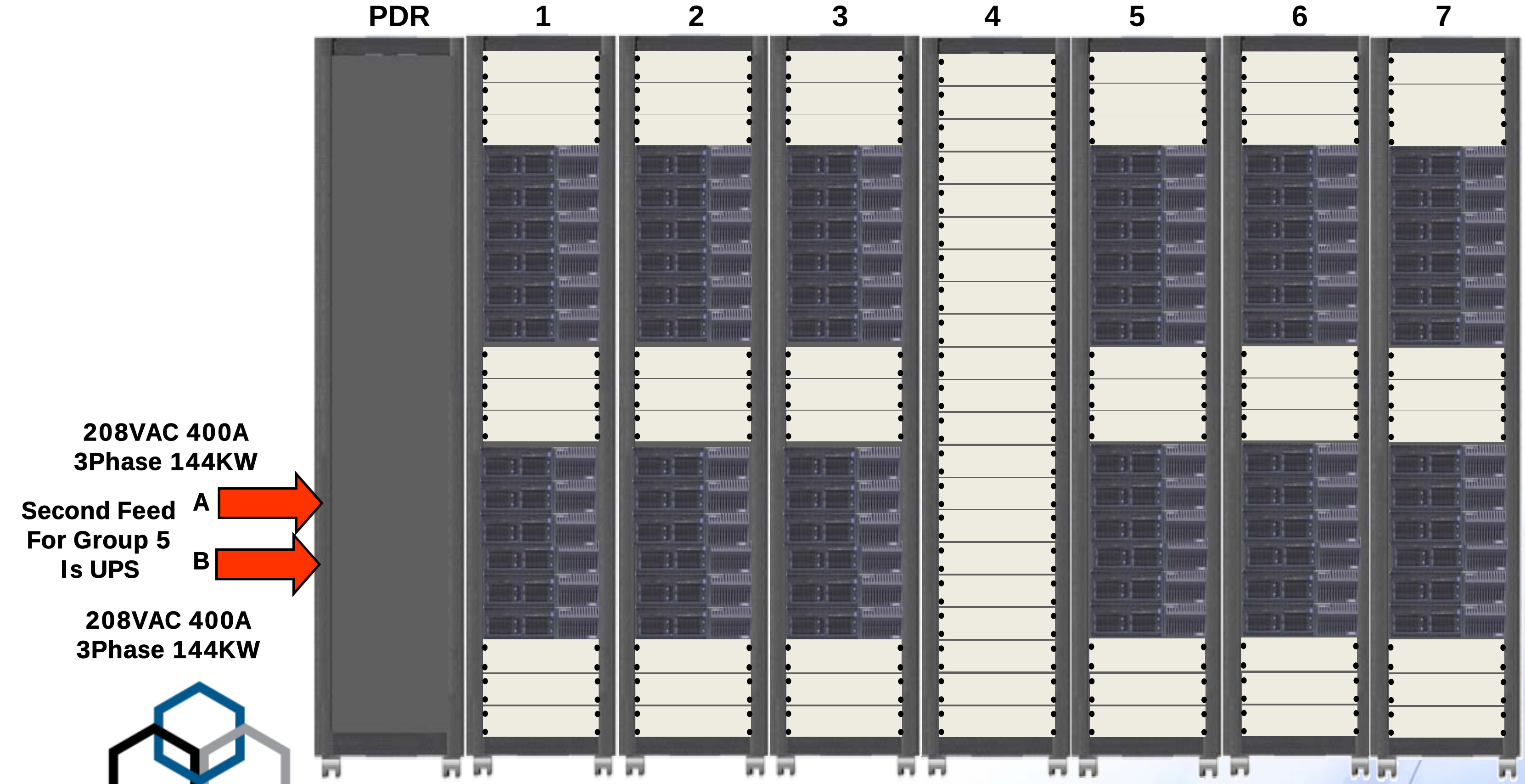
IB switches mounted to front of rack to allow as much space on cable side as possible

Longest cable is 5m and can be 28ga



Power Layout

:: PDU at End of Row



208VAC 400A
3Phase 144KW

Second Feed
For Group 5
Is UPS

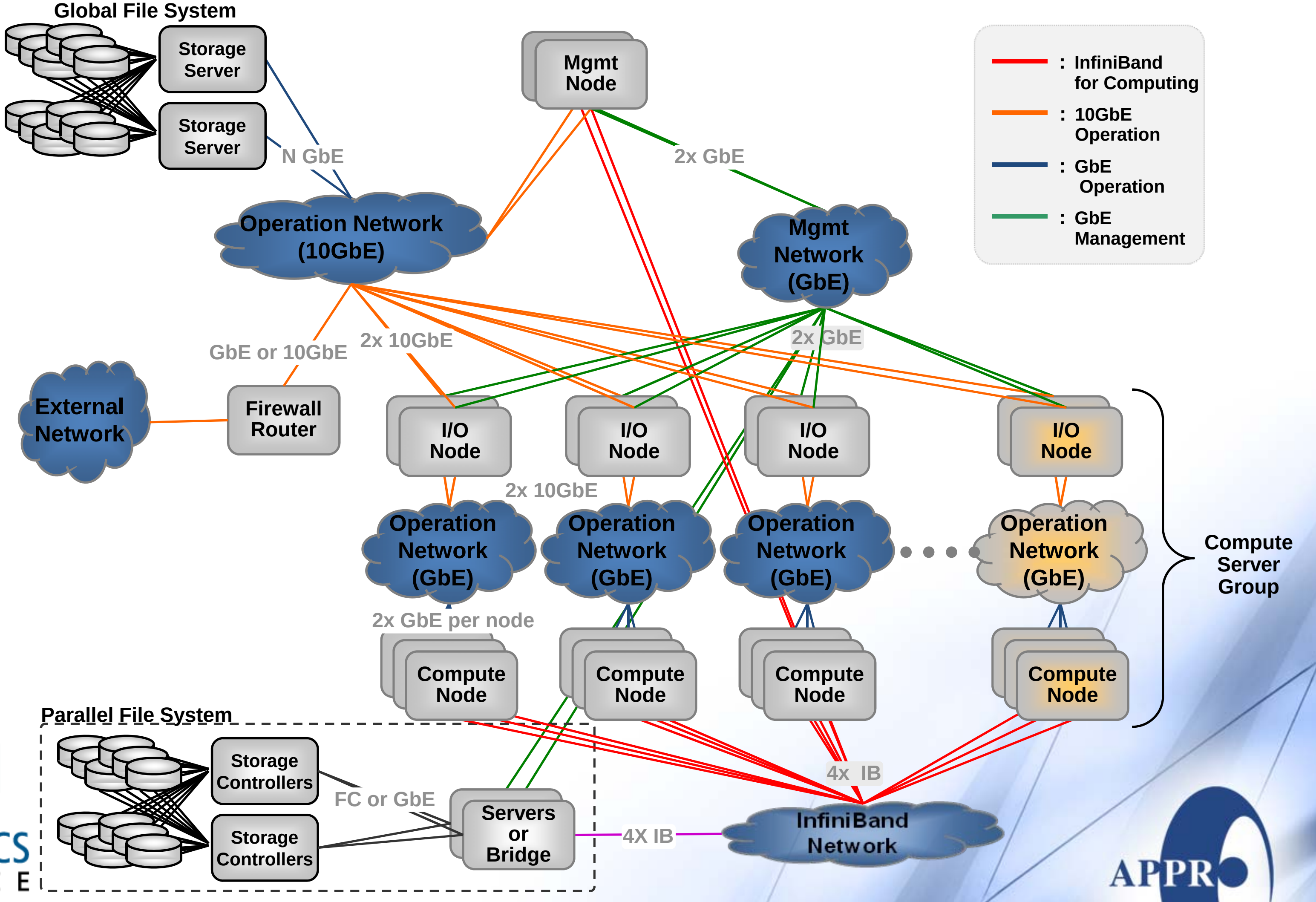
208VAC 400A
3Phase 144KW

OPENFABRICS ALLIANCE

APPRO
HPC Cluster Solutions

Mgmt Architecture

:: Diskless with Stock OS and MVC



OPENFABRICS ALLIANCE

APPRO HPC Cluster Solutions

Why Multi-Rail

- Reliablility
  - Single Rail Inifiniband is very good
- Improved Bandwidth for Multi-Core Nodes
  - Parallel Operation
- Lower Latency for Multi-Core Nodes
  - Parallel Operation is easier than making the single channel latency lower.
  - Keeps messages from multiple cores from queuing up
- Faster Processors Demand Faster Fabrics
- Faster Processors Demand Faster IO
- IB has more band for the buck than 10GbE

OPENFABRICS ALLIANCE

APPRO
HPC Cluster Solutions

Problems

- Must be able to operate channels like trunks
- Transmit on any and receive on any
- Retransmit on second channel after failure
- Should be able to assign cores to channels
- Should be able to use one or more channels per core
- Should be able to re-sequence messages
- Protect against last byte problem
- Multiple Channels should be as transparent to user as possible
- Should be able to use shared receive queues.