OpenStack Swift Reference Designs

This document contains four OpenStack Swift reference designs: small, medium, and large.

The three designs have separate high level specifications and architecture diagrams but all re-use a common set of bill of materials, racking rules, and network plug suggestions.

Small	Medium	Large
Integrated Proxy. 24 object server limit.	Dedicated proxy nodes.	Dedicated proxy and dedicated meta-data nodes.

Guidelines for choosing between small and medium

Storage size:

Small is limited to a maximum of 24 object servers. If you need more storage than can fit in 24 object servers you should choose medium.

Background:

Swift small contains exactly 3 Swift proxies which run on the 3 controllers. There are no horizontal scaling guidelines going beyond 3 controllers. Given the horizontal scaling rule of thumb of 1 proxy server to 8 object servers you are limited to a maximum of 24 object servers.

Performance:

Depending on your object storage workload characteristics you may find that the proxy servers become the bottleneck due to either the workload or the sharing of controller server resources between the control plane services and the Swift proxy service. Additionally, depending on the workload you may need more than 3 proxies to handle 24 object servers. If either of these issues becomes a factor, moving to Swift medium with its dedicated Swift proxy nodes would alleviate the issue.

Guidelines for choosing between medium and large

Cost savings:

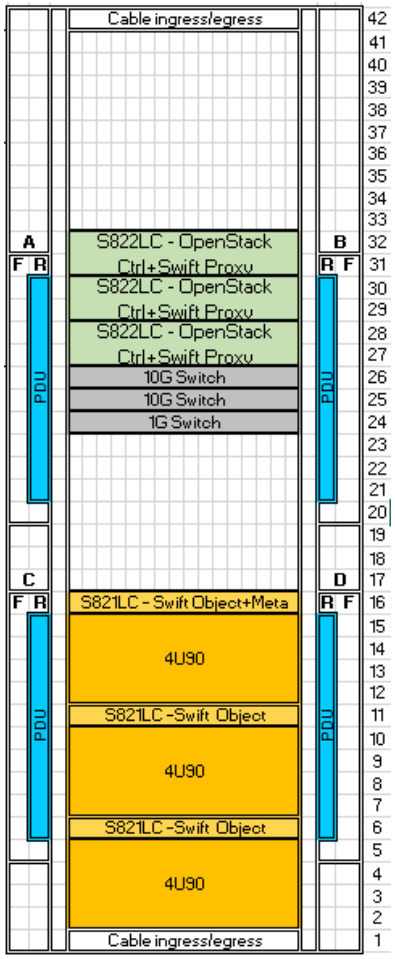
As you scale the medium architecture horizontally, given workload specifics you may begin to have under utilized SSDs which are used to hold the account and container Swift rings. At some point you hit a tipping point where it is more cost effective to host the account and container rings with their associated SSDs in dedicated metadata servers. You would then scale the metadata servers horizontally with a rule of thumb ratio of 1 metadata server to 6 object servers. The exact point you when you hit this cost savings threshold is dependent upon server and SSD pricing.

Performance:

The object storage workload specifics could favor large with its dedicated metadata servers before the cost savings threshold is hit. For example, if the workload has an extremely high number of users and containers but lower raw object storage needs, and the workload is doing a lot of account and container lookup, the large configuration with its dedicated metadata servers may be a better fit.

Small Swift Cluster

Swift Small – Base Config– High Level Specification Sheet



OpenStack Software Stack:
Ubuntu 16.04 (all nodes)
Openstack Newton

OpsMgr + Horizon DashBoard
-Nagios Core
- ELK Stack (Elasticsearch, Logstash, Kibana)

Network : (HA – with Bonding)
2 x Mellanox SX1410 (8831-S48)
1 x Lenovo G8052 (7120-48E)

Rack:
QTY: 1
SlimRack 7965-94Y (Standard 19" rack)
PDUs x 4: Each node should have 2 power cords cabled to two different PDUs

OpenStack Controller & Proxy:
QTY: **3**

Per Server Config: (Briggs 8001-22C) (2U)
20 Cores (2.92 Ghz), 128 GB,
1 x 4TB SATA HDD
1 x 2-Port 10G NIC (Intel 10G/Mellanox)

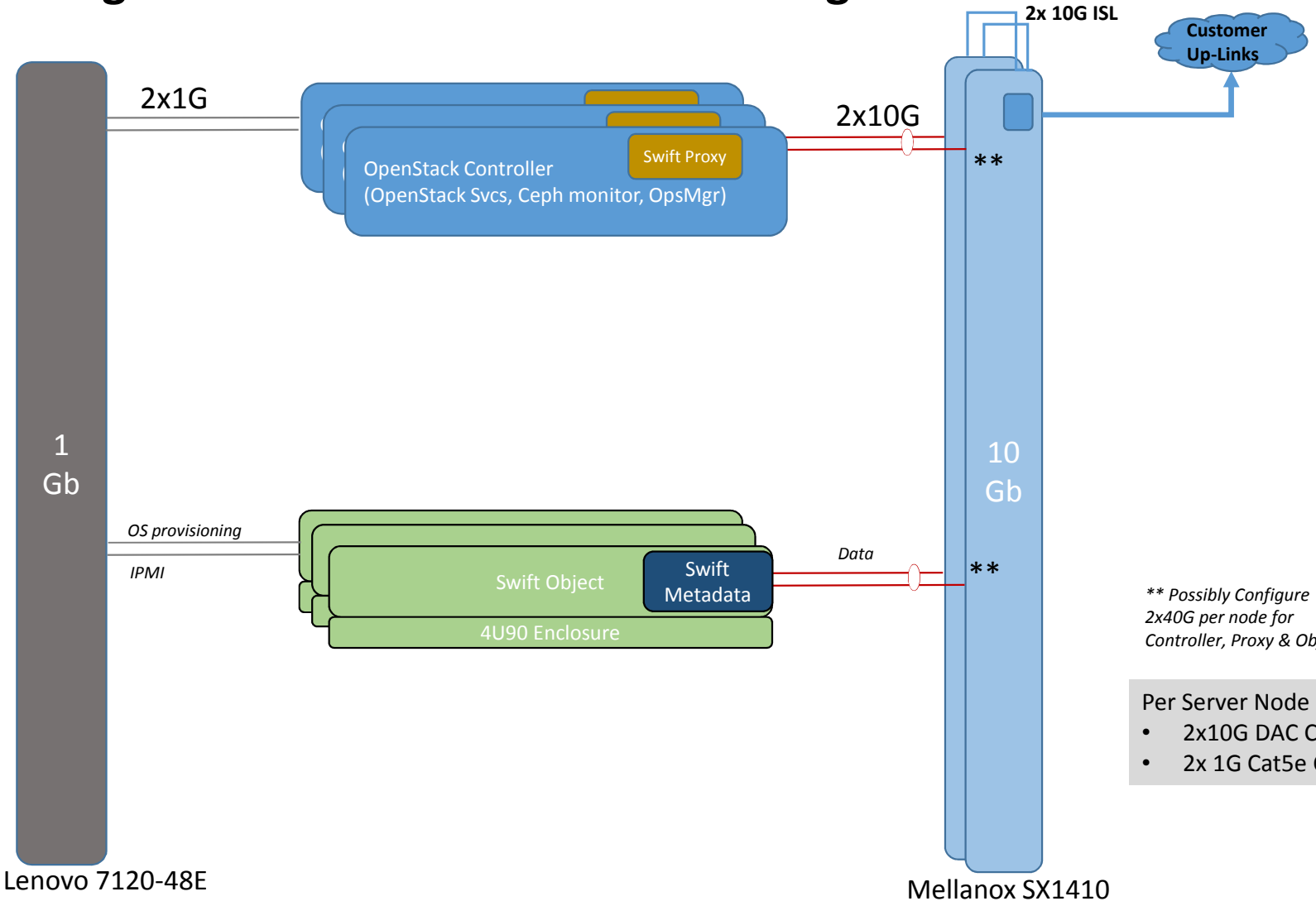
Swift Object /MetaData
QTY: **3**

Per Server Config: (Stratton 8001-12C) (1U)
16 Cores (2.3Ghz), 128GB
- (OS) 1 x 4TB SATA HDD + 4 x 240 GB SSDs
- 1 x 2-Port 10G NIC (Intel/Mellanox)
- 1 x External SAS (8 port SAS3) LSI 3008 based

Expansion Drawer (4U) :
90 LFF JBOD Storage SMC PN SE-946ED-R2KJBOD
90 LFF – 2TB SAS HDDs

****Notes:**
a) Proc + Memory config may need to be altered based on actual performance requirements

Swift Small - High Level Network Architecture Diagram

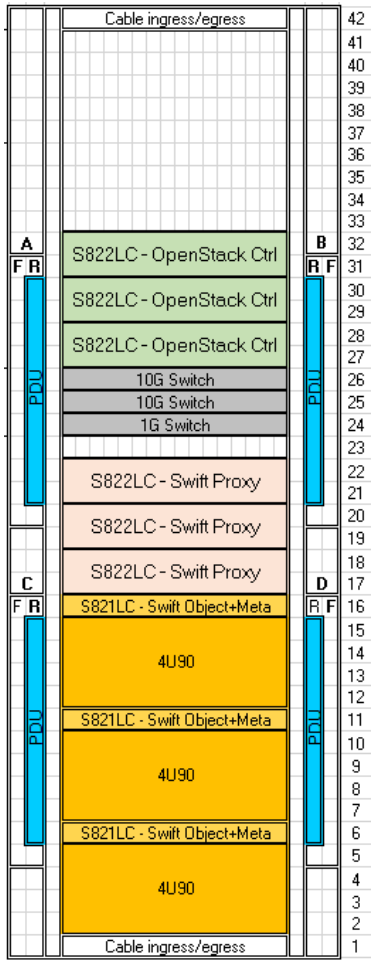


*** Possibly Configure 2x40G per node for Controller, Proxy & Object*

Per Server Node
- 2x10G DAC Cables
- 2x 1G Cat5e Cables

Medium Swift Cluster

Swift Medium– Base Config– High Level Specification Sheet



OpenStack Software Stack:
Ubuntu 16.04 (all nodes)
Openstack Newton

OpsMgr + Horizon DashBoard
-Nagios Core
- ELK Stack (Elasticsearch, Logstash, Kibana)

Network : (HA – with Bonding)
2 x Mellanox SX1410 (8831-S48)
1 x Lenovo G8052 (7120-48E)

Rack:
QTY: 1
SlimRack 7965-94Y (Standard 19" rack)
PDUs x 4: Each node should have 2 power cords cabled to two different PDUs

OpenStack Controller:
QTY: 3

Per Server Config: (Briggs 8001-22C) (2U)
20 Cores (2.92 Ghz), 128 GB,
1 x 4TB SATA HDD
1 x 2-Port 10G NIC (Intel 10G/Mellanox)

Swift Object /MetaData
QTY: 3

Per Server Config: (Stratton 8001-12C) (1U)
16 Cores (2.3Ghz), 128GB
- (OS) 1 x 4TB SATA HDD + 4 x 240 GB SSDs
- 1 x 2-Port 10G NIC (Intel/Mellanox)
- 1 x External SAS (8 port SAS3) LSI 3008 based

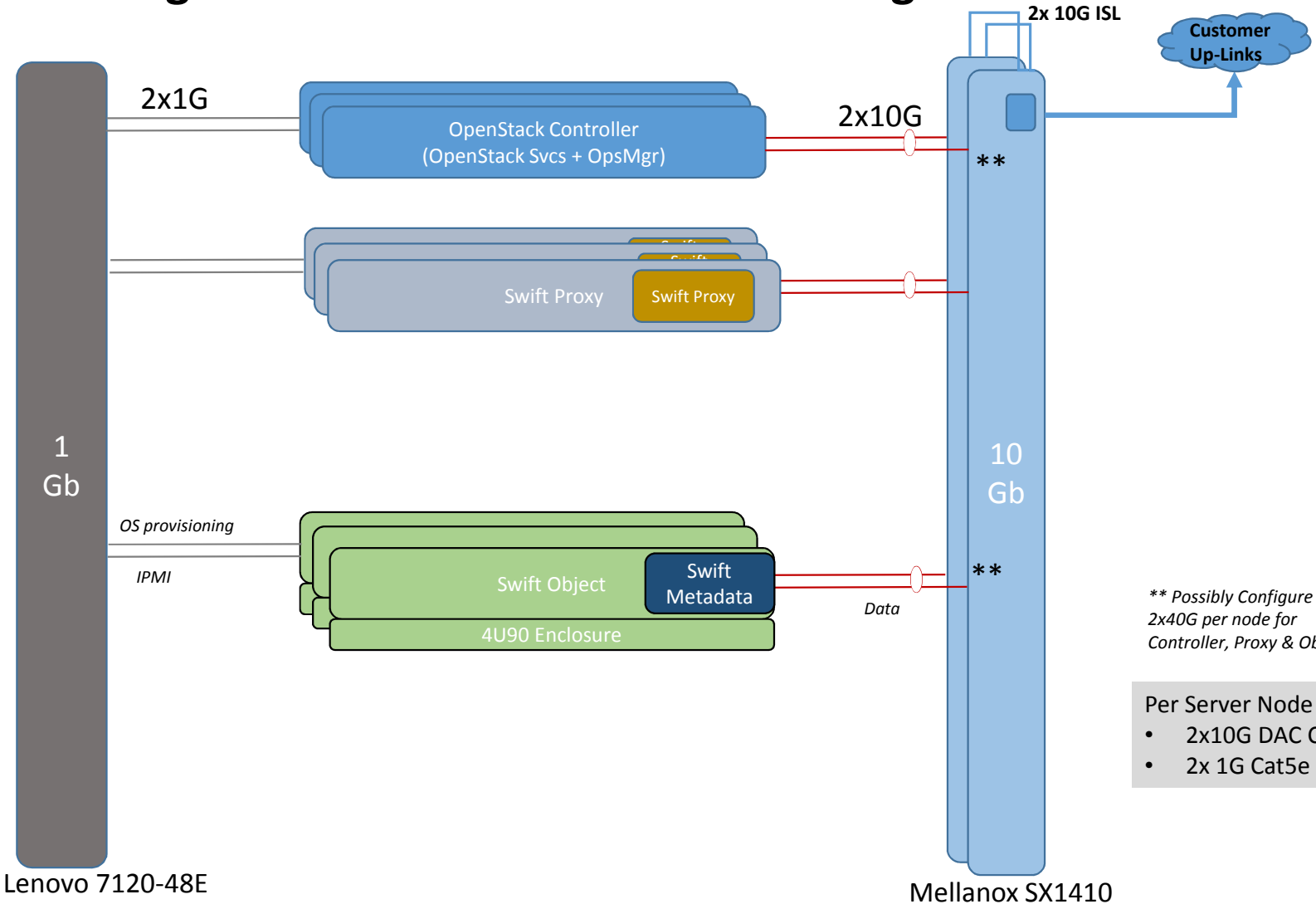
Expansion Drawer (4U) :
90 LFF JBOD Storage SMC PN SE-946ED-R2KJBOD
90 LFF – 2TB SAS HDDs

Swift Proxy:
QTY: 3

Per Server Config: (Briggs 8001-22C) (2U)
20 Cores (2.92Ghz), 256GB
1 x 4TB SATA HDD
1 x 2-Port 10G NIC (Intel 10G/Mellanox)

****Notes:**
a) Proc + Memory config may need to be altered based on actual performance requirements

Swift Medium - High Level Network Architecture Diagram

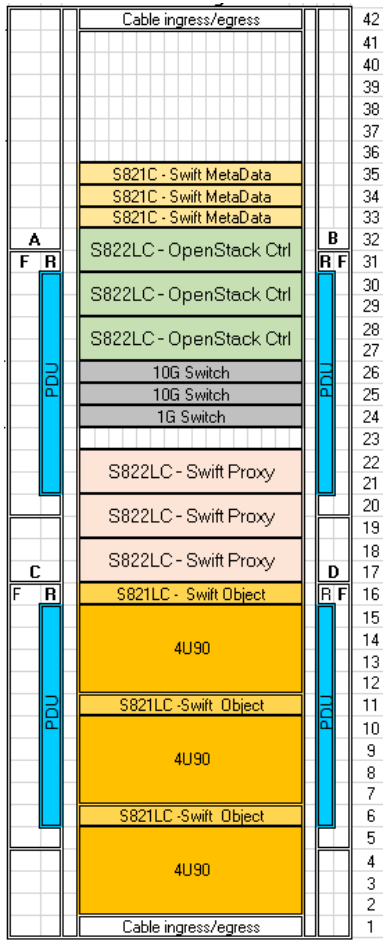


*** Possibly Configure 2x40G per node for Controller, Proxy & Object*

Per Server Node
- 2x10G DAC Cables
- 2x 1G Cat5e Cables

Large Swift Cluster

Swift Large – Base Config– High Level Specification Sheet



OpenStack Software Stack:
Ubuntu 16.04 (all nodes)
Openstack Newton

OpsMgr + Horizon DashBoard
-Nagios Core
- ELK Stack (Elasticsearch, Logstash, Kibana)

Network : (HA – with Bonding)
2 x Mellanox SX1410 (8831-S48)
1 x Lenovo G8052 (7120-48E)

Rack:
QTY: 1
SlimRack 7965-94Y (Standard 19" rack)
PDUs x 4: Each node should have 2 power cords cabled to two different PDUs

OpenStack Controller:
QTY: 3

Per Server Config: (Briggs 8001-22C) (2U)
20 Cores (2.92 Ghz), 128 GB,
1 x 4TB SATA HDD
1 x 2-Port 10G NIC (Intel 10G/Mellanox)

Swift MetaData
QTY: 3

Per Server Config: (Stratton 8001-12C) (1U)
16 Cores (2.3Ghz), 128GB
- (OS) 1 x 4TB SATA HDD + 4 x 240 GB SSDs
- 1 x 2-Port 10G NIC (Intel/Mellanox)

Swift Proxy:
QTY: 3

Per Server Config: (Briggs 8001-22C) (2U)
20 Cores (2.92Ghz), 256GB
1 x 4TB SATA HDD
1 x 2-Port 10G NIC (Intel 10G/Mellanox)

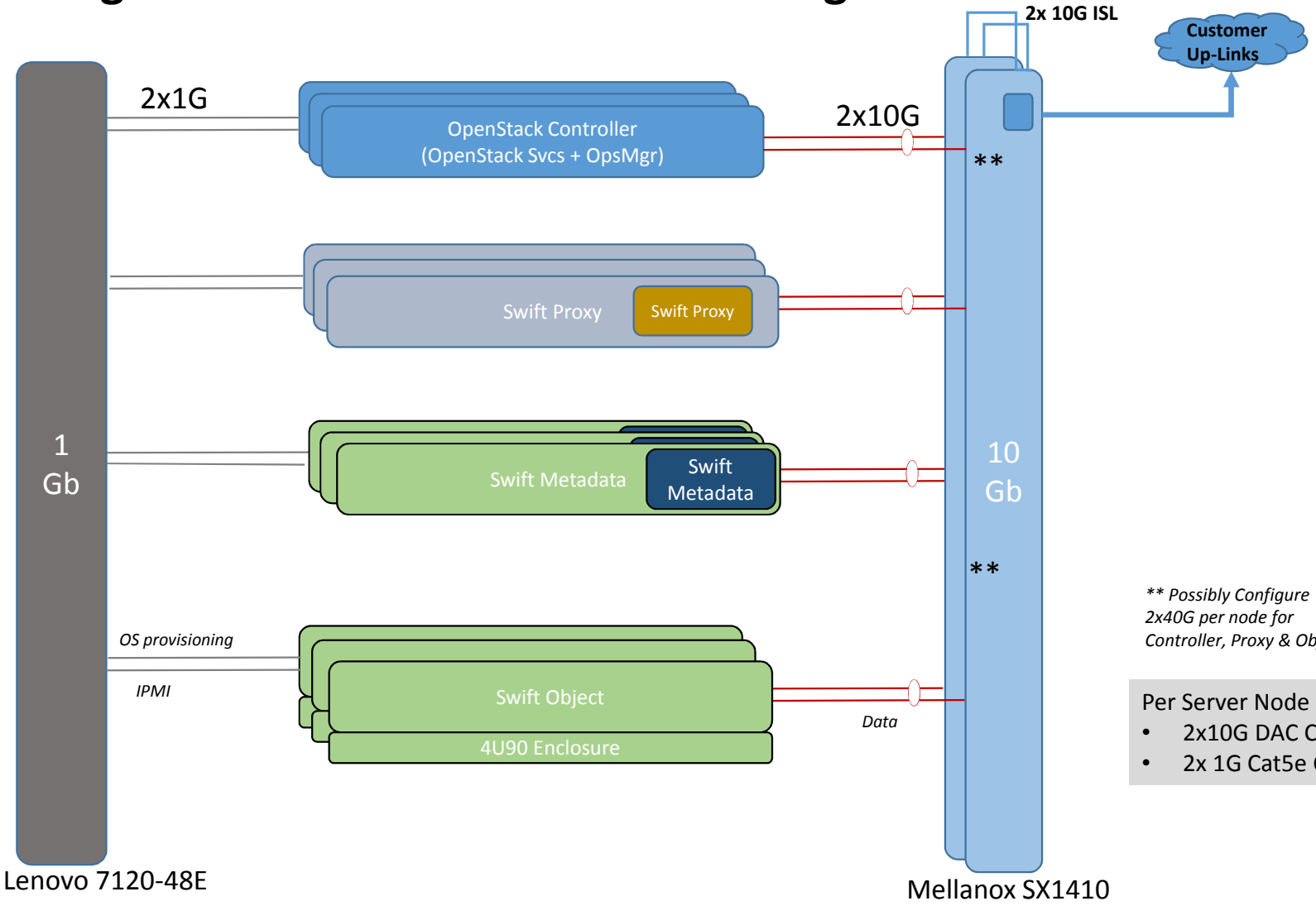
Swift Object
QTY: 3

Per Server Config: (Stratton 8001-12C) (1U)
16 Cores (2.3Ghz), 128GB
- (OS) 1 x 4TB SATA HDD + 4 x 240 GB SSDs
- 1 x 2-Port 10G NIC (Intel/Mellanox)
- 1 x External SAS (8 port SAS3) LSI 3008 based

Expansion Drawer (4U) :
90 LFF JBOD Storage SMC PN SE-946ED-R2KJBOD
90 LFF – 2TB SAS HDDs

****Notes:**
a) Proc + Memory config may need to be altered based on actual performance requirements

Swift Large - High Level Network Architecture Diagram



*** Possibly Configure 2x40G per node for Controller, Proxy & Object*

Per Server Node
- 2x10G DAC Cables
- 2x 1G Cat5e Cables

Common Suggested Racking Rules, Server Bill of Materials, and Networking Diagrams

Suggested Racking Rules

Reserved for accessibility

U37-U41 may be used for aggregation switches or additional compute nodes

Reserved **U24-U26** for **In-rack** Network Switches

Start at U2and go UP

Common Rule:

--Place Pod / same MTM / Similar Function together

--Place Heavier MTMs first starting at U2

--Observe native MTM unique racking requirements

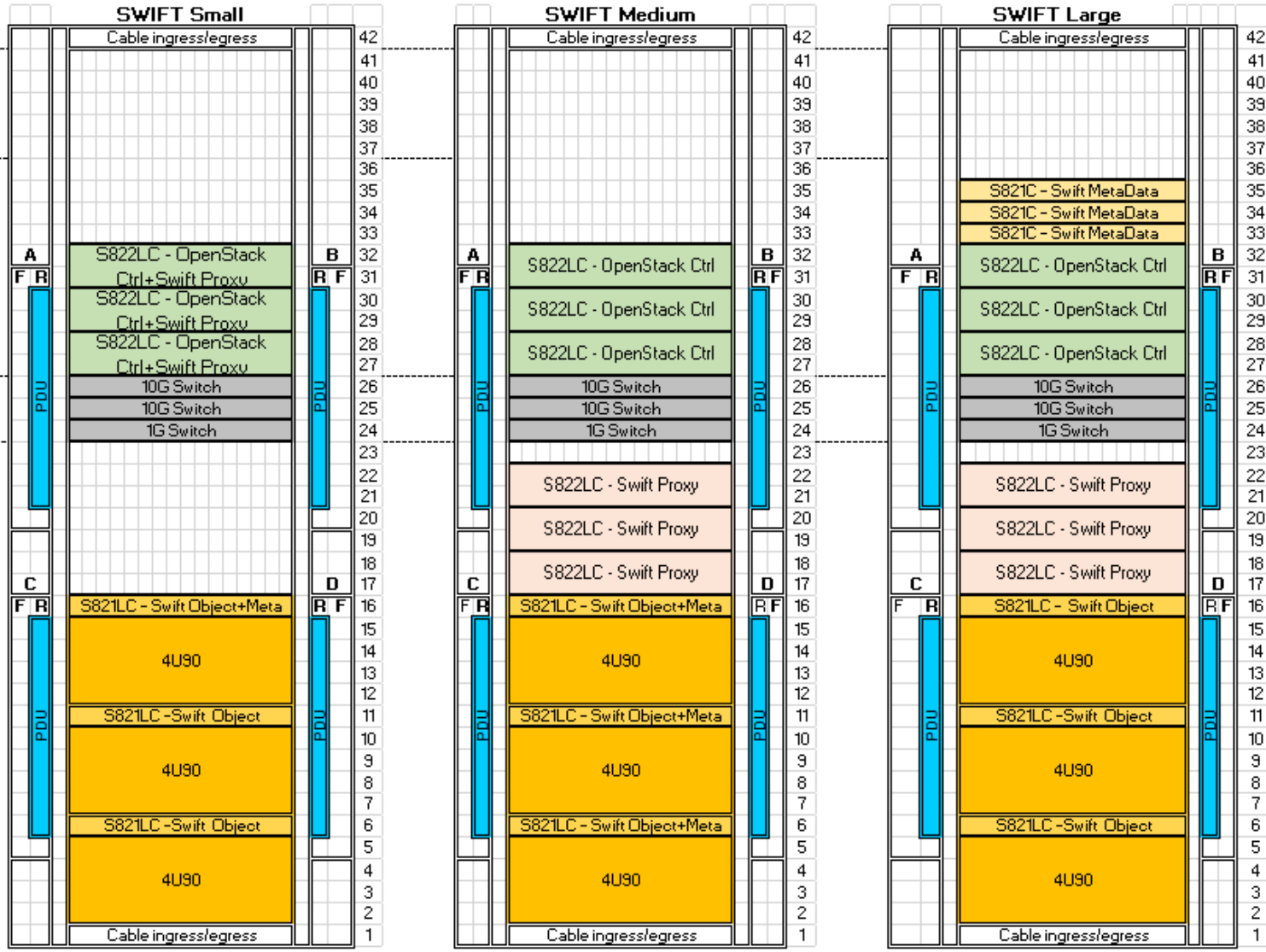
Swift Racking Rules:

-- Recommend (0 -18) 2U Servers per Rack

-- Swift Object are placed in the lower part of the rack

-- Then Proxy 2U servers, follow by 1U MetaData and Controller servers

Reserved for accessibility



Swift Object and Metadata Server BOMs

MT	Model	Description	Mfg Config #1	Min	Max	Comments
S812C Server Config =Swift Object / Metadata						
8001	12C	S821LC (8001)	2	1	**	
	Solution ID	Solution Specify Code (for grouping only)	1	1	1	n/a
	Pod Type	Login Server Specify Code	1	1	1	n/a
	Processor	8-core POWER8 2.328 GHz	2	1	2	
	Memory	EKM2 (PS) 16GB DDR4 MEMORY DIMM	8	4	16	
	Bezel	EKB4 2S base system with LFF high-function drive midplane (NVMe driv	1	1	1	
	Storage Adapter	Integrated Sata controller	1	1	1	Build-in HDDs : Integrate SATA controller + Optional SAS /RAID Controller
		EKAD Storage Adapter SAS-3 3008 Chipset 8 Ports external for 1U	1	1	1	Optional - Exteral SAS adapter for Expansion SAS drawer
	Disks	EKDB 4TB 3.5" SATA HDD	1	0	2	OS Boot Disk
		EKS1 240 GB, SFF SATA SSD; 1.2 Disk Writes Per Day (DWPD) kit	4	4	4	If SAS drive is selected, please choose Bezel Assembly to match drive size (.5" or 3.5" and SAS controller
	NVmE PCI		0	4	2	
	GPU		0	0	1	
	HDD Drawer	90 LFF JBOD Storage 90 LFF – 2TB SAS HDDs	1	1	1	Supermicro CSE-946ED-R2KJBOD 4U Rackmount https://www.supermicro.com/products/chassis/4u/946/SC946ED-R2KJBOD
S812C Server (Base config) -- Required Inter-connect						
Required for Mfg Genesis	Network Adapter	EKA2 PCIe3 2-port 10 GbE SFP+ Adapter, based on Intel XL710	1	1	3	**(Required) For High Speed Network**
			0	0	3	Section IO device (optional)
	Power	EKLJ (PS #6665) PWR CBL DRWR TO IBM PDU, 2.8m (9.2ft), 250V/10A, IEC320/C13, IEC320/C20	2	2	2	Select Proper Line cord if not connected to IBM PDU
	Cables	CAT5E SWITCH CABLE, BLUE (2M)	1	1	*	**(Required) For OS 1G Network (Recommended 2M length min)**
		CAT5E SWITCH CABLE, GREEN (2M)	1	1	*	**(Required) For IPMI 1G Network (Recommended 2M length min)**
		EKC1 3M- Active Twinax cable	2	2	*	**(Required) For High Speed Network (Recommended 2M length min)**
	Misc	No rack integration	1	1	1	
		Country specific FCs (keyboards, language groups) are selectable	1	1	1	User select
		Shipping and Handling	1	1	1	User select

Swift Proxy and OpenStack Controller BOMs

MT	Model	Description	Mfg Config #1	Min	Max	Comments
822C Server Config : Swift Proxy and OpenStack controller						
8001	22C	ServerConfig- S822C	**3**	**3**	**	This section Defined the **Common config of the Server node** (in group servers) -- **Next Section :** Defined any unique config that you may need (Optional)
	Processor	EKP5 10-core POWER8 2.92 GHz	2	1	2	
	Memory	EKM2 (PS) 16GB DDR4 MEMORY DIMM	8	4	16	Note: 128GB should be used for the controller and 256GB for the standalone proxy
	Bezel	EKB5 (PS) 2S BRIGGS LFF DIRECT ATTACH FAB ASSEMBLY	1	1	1	Need to Choose drive assemply to match your Disks (LFF/SFF) and Controler type (SAS)
	Storage Adapter	Integrated Sata controller	1	1	1	Build-in HDDs : Integrate SATA controller + Optional SAS /RAID Controller
			0	0	1	Optional - Exteral SAS adapter for Expansion SAS drawer
	Disks	EKDB 4TB 3.5" SATA HDD	1	0	2	OS Boot Disk
			0	0	4	If SAS drive is selected, please choose Bezel Assembly to match drive size (.5" or
	NVmE PCI		0	4	2	
	GPU		0	0	1	
822C Server (Base config) -- Required Inter-connect						
Required for Mfg Genesis	Network Adapter	EKA2 (PS) INTEL 82599ES 2-PORT SFP+ 10G GEN2 x8 STANDARD	1	1	3	**(Required) For High Speed Network**
			0	0	3	Section IO device (optional)
	Power	EKLJ (PS #6665) PWR CBL DRWR TO IBM PDU, 2.8m (9.2ft), 250V/10A, IEC320/C13, IEC320/C20	2	2	2	Select Proper Line cord if not connected to IBM PDU
	Cables	CAT5E SWITCH CABLE, BLUE (2M)	1	1	*	**(Required) For OS 1G Network (Recommended 2M length min)**
		CAT5E SWITCH CABLE, GREEN (2M)	1	1	*	**(Required) For IPMI 1G Network (Recommended 2M length min)**
		EKC1 3M- Active Twinax cable	2	2	*	**(Required) For High Speed Network (Recommended 2M length min)**
	Misc	Country specific FCs (keyboards, language groups) are selectable	1	1	1	User select
		Shipping and Handling	1	1	1	User select

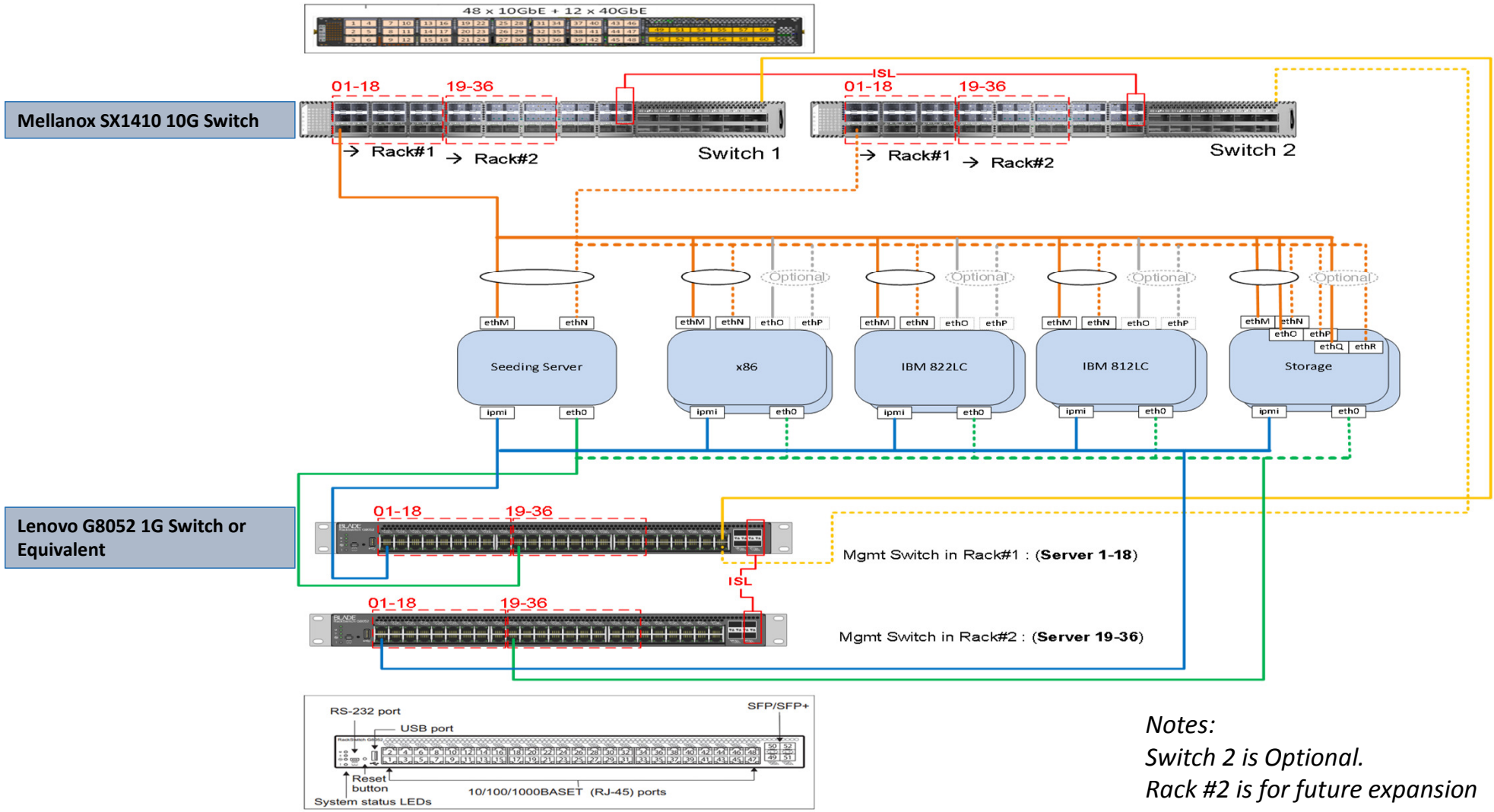
Network Switch BOMs

	MT	Model	FC	Description	
1G Mgmt (Based)	7120	48E		Lenovo G8052 1GbE Switch (48x 10GbE ports + 4x 10GbE ports)	1
			1118	CAT5E SWITCH CABLE, 3M, YELLOW	1
			6577	PWR CBL, DRWR TO IBM PDU, MFG SEL LENGTH, 200-240V/10A, IEC320/C13, IEC320/C14	2
				Include all existing FCs; except FCs 0010, 0011, 0712, 0714, EGSx, EHKx, EHLA, 4649 (Rack Integration Services), and 0456 (Customer Specified Placement); do not include these FCs.	

10G Data Network	8831	S48		Mellanox 1410 ..10GB Switch (48x10G + 12x40G)	2
			EDT6	1U AIR DUCT FOR S48	1
			EN01	1m DAC cable SFP+ to SFP+	1
				Include all existing FCs; except FC 4649, FC 0456 (Customer Specified Placement) and ESC1 (Shipping & Handling), do not include these FCs	1

NOTE: 1m DAC SFP+ to SFP+ cables provide interpeer link connections

Network Plug Rule - Sample



Notes:
Switch 2 is Optional.
Rack #2 is for future expansion

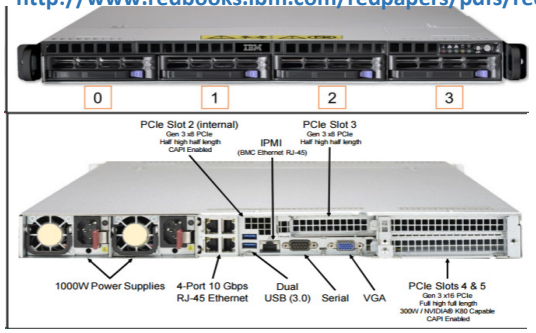
Network Plug P2P Label -- Sample

Server PCI Slot Placement

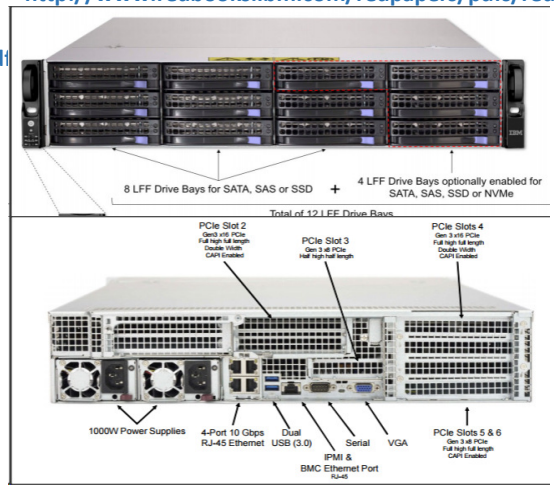
8001-12C/22C Stratton/Briggs

	adapter	PCI slot	Port	Cabling
Primary NIC	10GbE	slot 3	T1	yes
			T2	yes
Optional NIC	10GbE	slot 4	T1	
			T2	
Mgmt-OS	1GbE	LOM	T1	yes
BMC	1GbE	LOM	impi	yes

MTM: 8001-12C
http://www.redbooks.ibm.com/redpapers/pdfs/redp5406.pdf



MTM: 8001-22C
http://www.redbooks.ibm.com/redpapers/pdfs/redp5407.pdf



Cable P2P Label for D_TOR#1-2

		10GbE	10GbE	10GbE	10GbE	1GbE	1GbE
		D_TOR_1	D_TOR_1	D_TOR_2	D_TOR_2	M_TOR_1	M_TOR_1
Server #	Name <opt>	P2P Data network Cable Label	P2P Data network Cable Label	P2P Data network Cable Label	P2P Data network Cable Label	P2P Mgmt RJ4-5 Cable Label	P2P IPMI RJ-45 Cable Label
1		1A/SVR1/slot 3/T1 <> D_TOR_1/Port1		1A/SVR1/slot 3/T2 <> D_TOR_2/Port1		1A/SVR1/LOM/T1 <> M_TOR_1/Port1	1A/SVR1/LOM/impi <> M_TOR_1/Port19
2		1A/SVR2/slot 3/T1 <> D_TOR_1/Port2		1A/SVR2/slot 3/T2 <> D_TOR_2/Port2		1A/SVR2/LOM/T1 <> M_TOR_1/Port2	1A/SVR2/LOM/impi <> M_TOR_1/Port20
3		1A/SVR3/slot 3/T1 <> D_TOR_1/Port3		1A/SVR3/slot 3/T2 <> D_TOR_2/Port3		1A/SVR3/LOM/T1 <> M_TOR_1/Port3	1A/SVR3/LOM/impi <> M_TOR_1/Port21
4		1A/SVR4/slot 3/T1 <> D_TOR_1/Port4		1A/SVR4/slot 3/T2 <> D_TOR_2/Port4		1A/SVR4/LOM/T1 <> M_TOR_1/Port4	1A/SVR4/LOM/impi <> M_TOR_1/Port22
5		1A/SVR5/slot 3/T1 <> D_TOR_1/Port5		1A/SVR5/slot 3/T2 <> D_TOR_2/Port5		1A/SVR5/LOM/T1 <> M_TOR_1/Port5	1A/SVR5/LOM/impi <> M_TOR_1/Port23
6		1A/SVR6/slot 3/T1 <> D_TOR_1/Port6		1A/SVR6/slot 3/T2 <> D_TOR_2/Port6		1A/SVR6/LOM/T1 <> M_TOR_1/Port6	1A/SVR6/LOM/impi <> M_TOR_1/Port24
7		1A/SVR7/slot 3/T1 <> D_TOR_1/Port7		1A/SVR7/slot 3/T2 <> D_TOR_2/Port7		1A/SVR7/LOM/T1 <> M_TOR_1/Port7	1A/SVR7/LOM/impi <> M_TOR_1/Port25
8		1A/SVR8/slot 3/T1 <> D_TOR_1/Port8		1A/SVR8/slot 3/T2 <> D_TOR_2/Port8		1A/SVR8/LOM/T1 <> M_TOR_1/Port8	1A/SVR8/LOM/impi <> M_TOR_1/Port26
9		1A/SVR9/slot 3/T1 <> D_TOR_1/Port9		1A/SVR9/slot 3/T2 <> D_TOR_2/Port9		1A/SVR9/LOM/T1 <> M_TOR_1/Port9	1A/SVR9/LOM/impi <> M_TOR_1/Port27
10		1A/SVR10/slot 3/T1 <> D_TOR_1/Port10		1A/SVR10/slot 3/T2 <> D_TOR_2/Port10		1A/SVR10/LOM/T1 <> M_TOR_1/Port10	1A/SVR10/LOM/impi <> M_TOR_1/Port28
11		1A/SVR11/slot 3/T1 <> D_TOR_1/Port11		1A/SVR11/slot 3/T2 <> D_TOR_2/Port11		1A/SVR11/LOM/T1 <> M_TOR_1/Port11	1A/SVR11/LOM/impi <> M_TOR_1/Port29
12		1A/SVR12/slot 3/T1 <> D_TOR_1/Port12		1A/SVR12/slot 3/T2 <> D_TOR_2/Port12		1A/SVR12/LOM/T1 <> M_TOR_1/Port12	1A/SVR12/LOM/impi <> M_TOR_1/Port30

MLAG IPL connections are D_TOR_1 port 37 to D_TOR_2 port 37 and D_TOR_1 port 38 to D_TOR_2 port 38