KISTI GSDC DATACENTER NETWORK ARICHTECTURE

2021 06 08

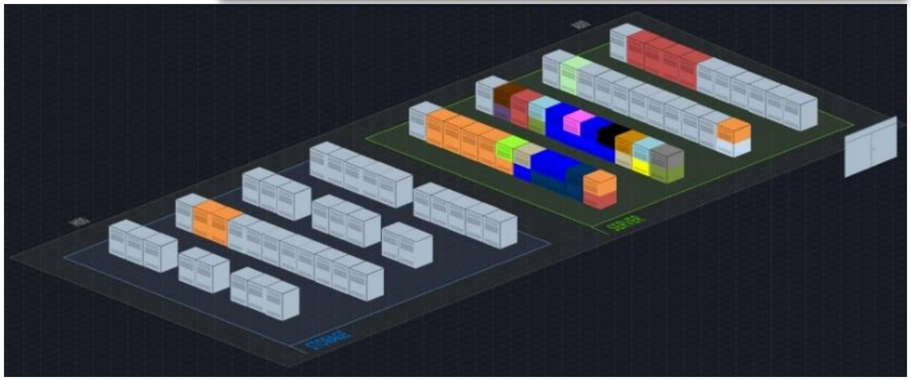
jkim@kisti.re.kr

Agenda

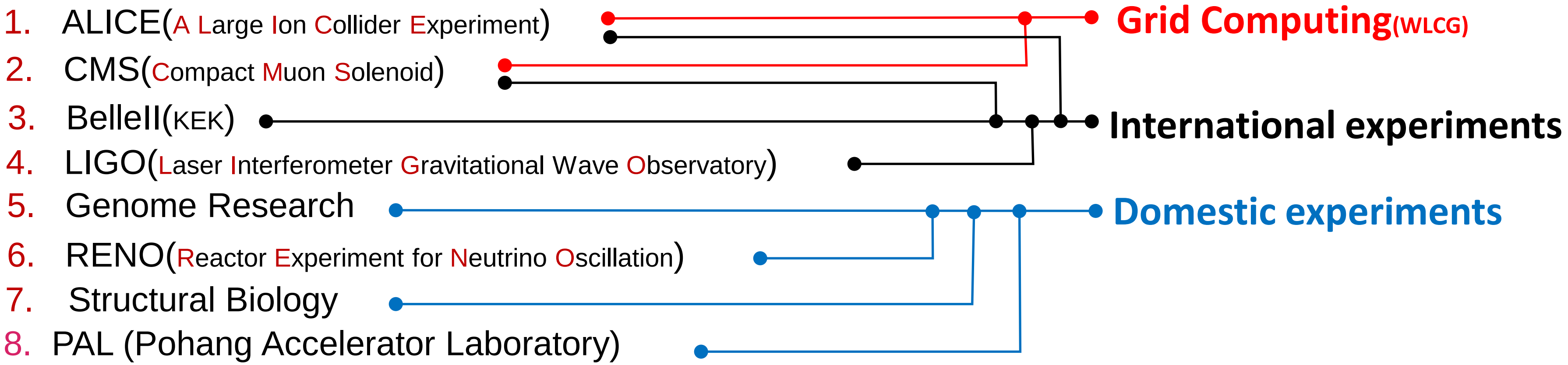
- Brief show the GSDC computing facility
- GSDC network
    - LHCOPN upgrade
    - Legacy network architecture
    - Docker supported network architecture
    - Monitoring

GSDC DATACENTER

- Computing Facility
  - # of Physical servers: 606
  - Total Computing Core: 14946
  - # of server racks: 89 (20 empty)
  - # of network switches: 101
  - Volume of storage: 18 PB



Supported Experiments



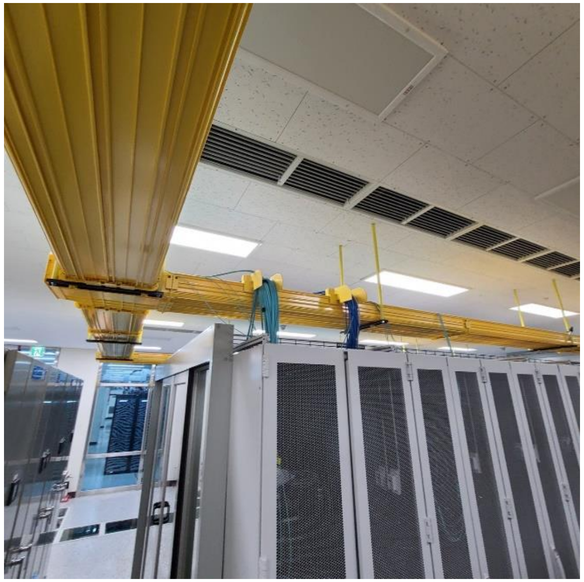
1. ALICE(A Large Ion Collider Experiment)
2. CMS(Compact Muon Solenoid)
3. BelleII(KEK)
4. LIGO(Laser Interferometer Gravitational Wave Observatory)
5. Genome Research
6. RENO(Reactor Experiment for Neutrino Oscillation)
7. Structural Biology
8. PAL (Pohang Accelerator Laboratory)

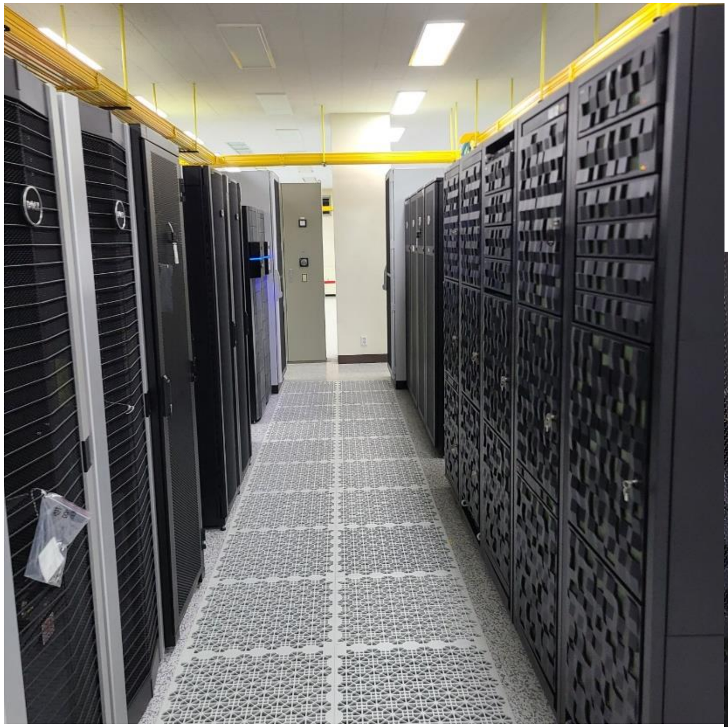
Grid Computing(WLCG)

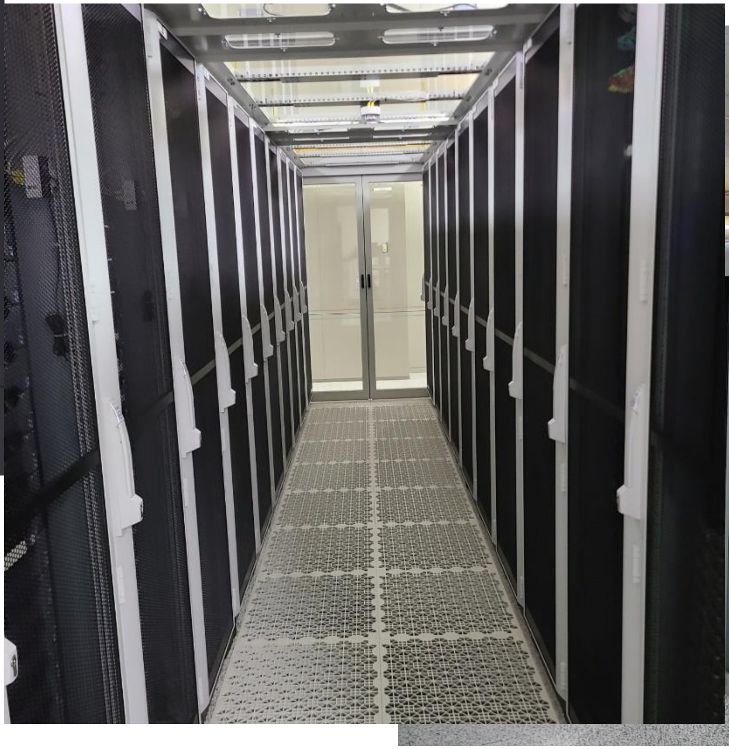
International experiments

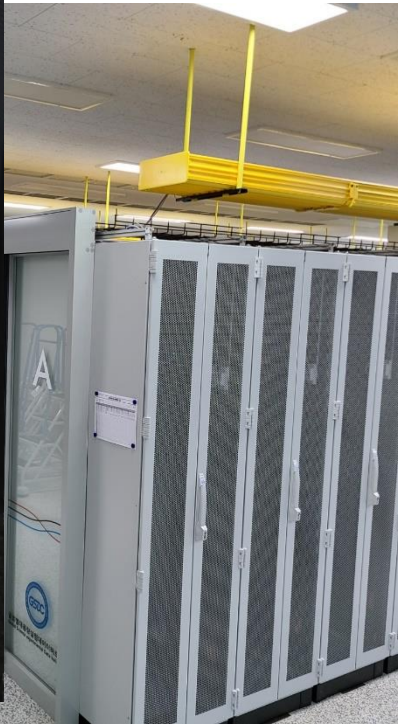
Domestic experiments

Computing Room

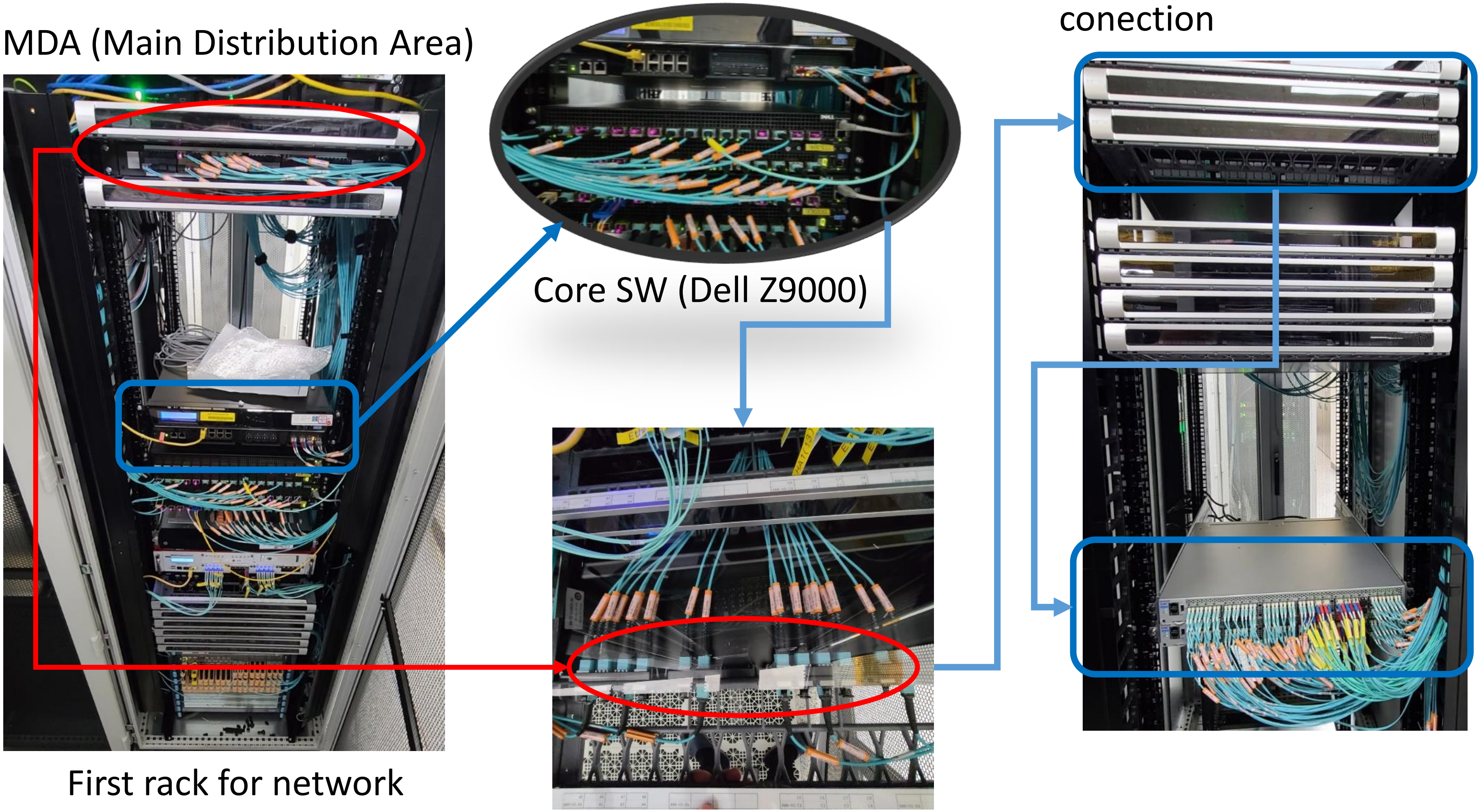








Connecting between containment



MDA (Main Distribution Area)

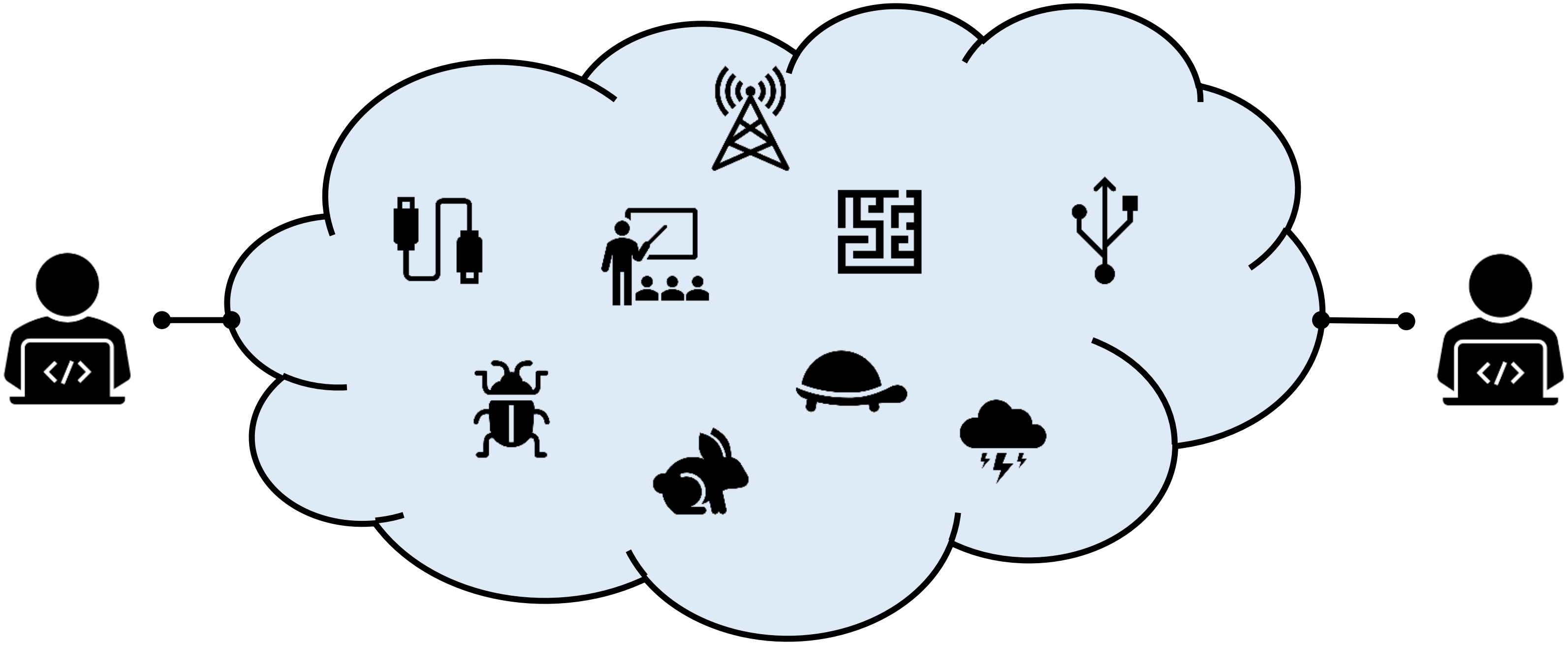
First rack for network

Core SW (Dell Z9000)

HDA-ToR/Aggregation SW conection

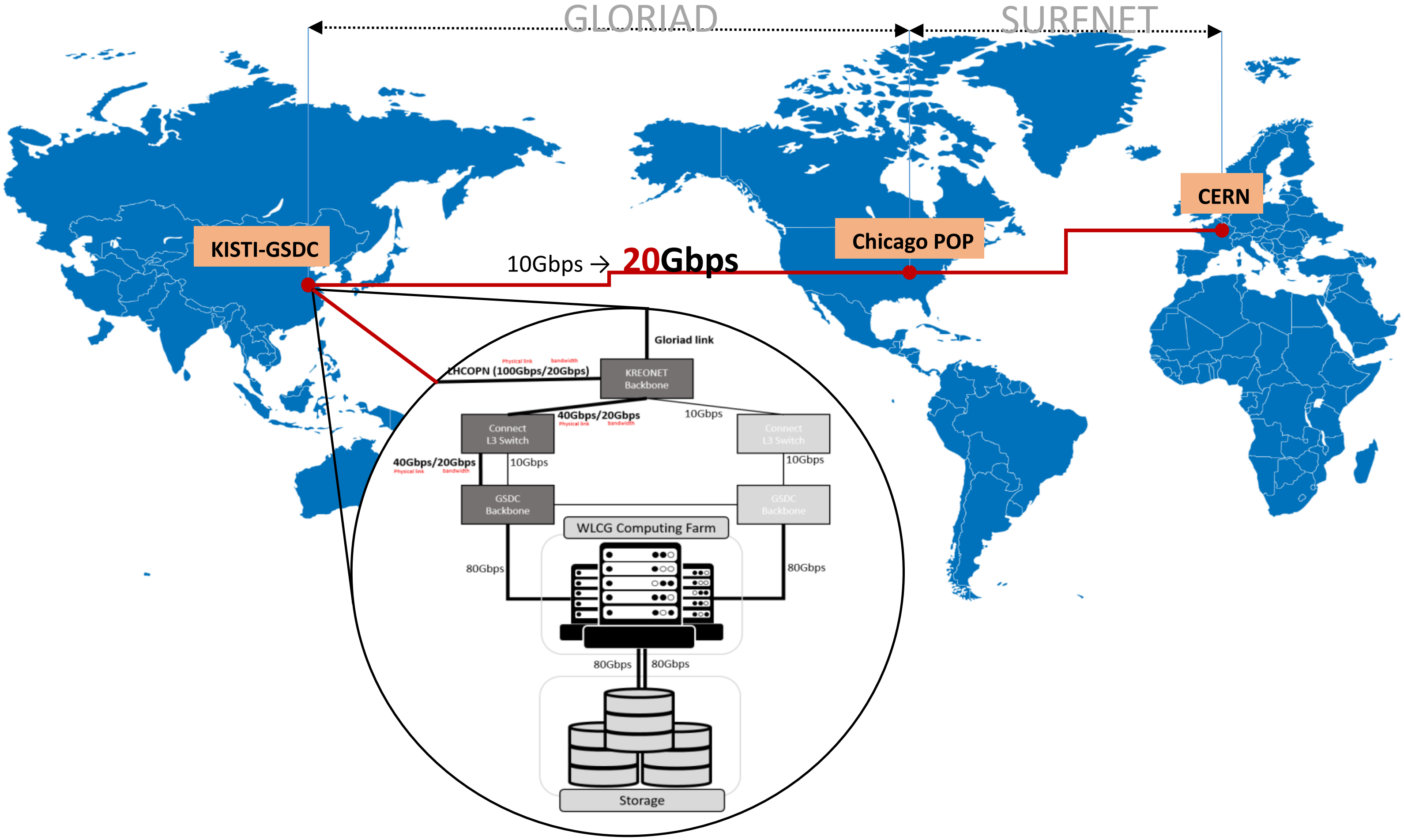
MDA-HDA(Horizonal Distribution Area) connection

male-female MPO connection

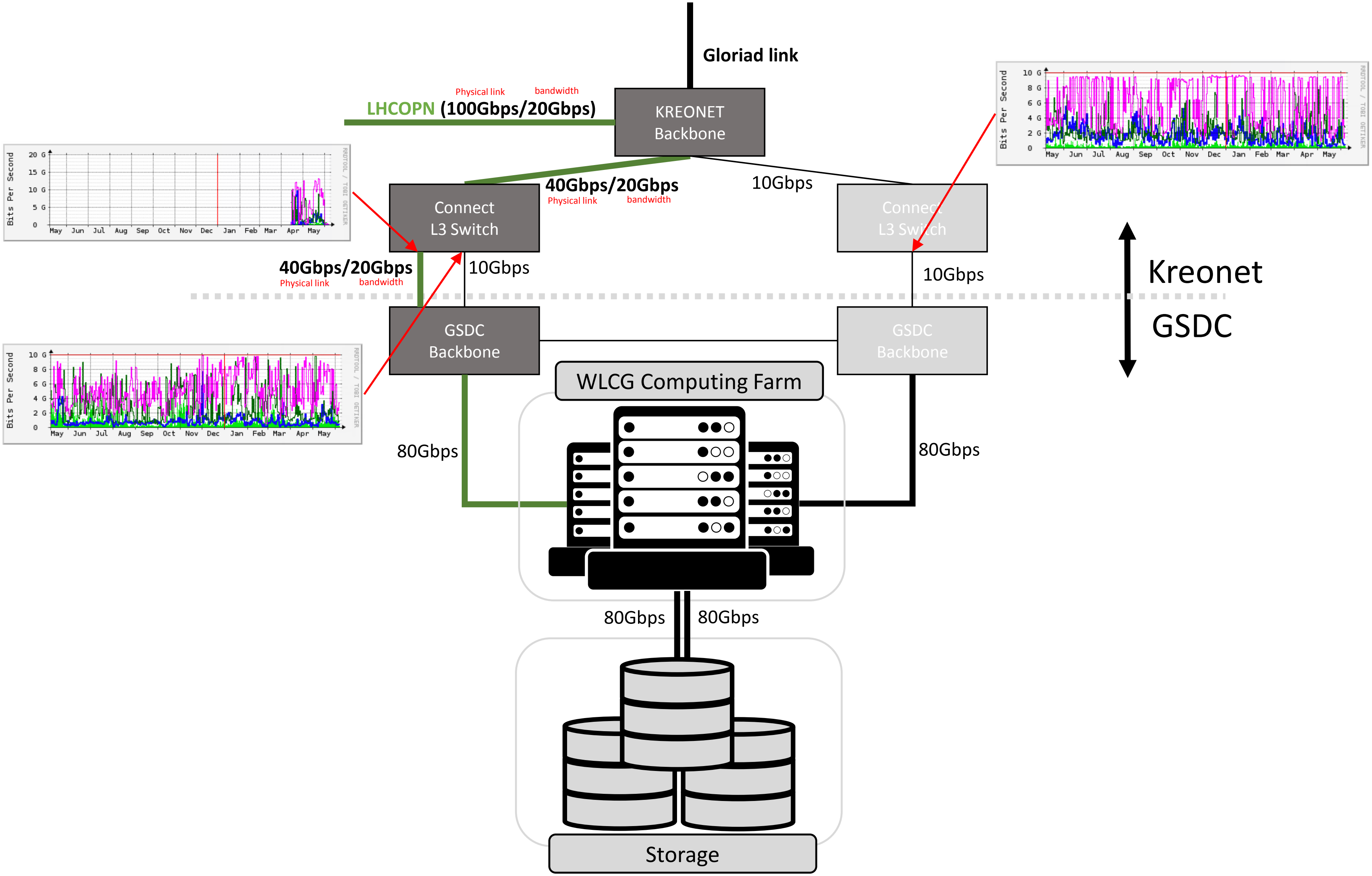


GSDC NETWORK

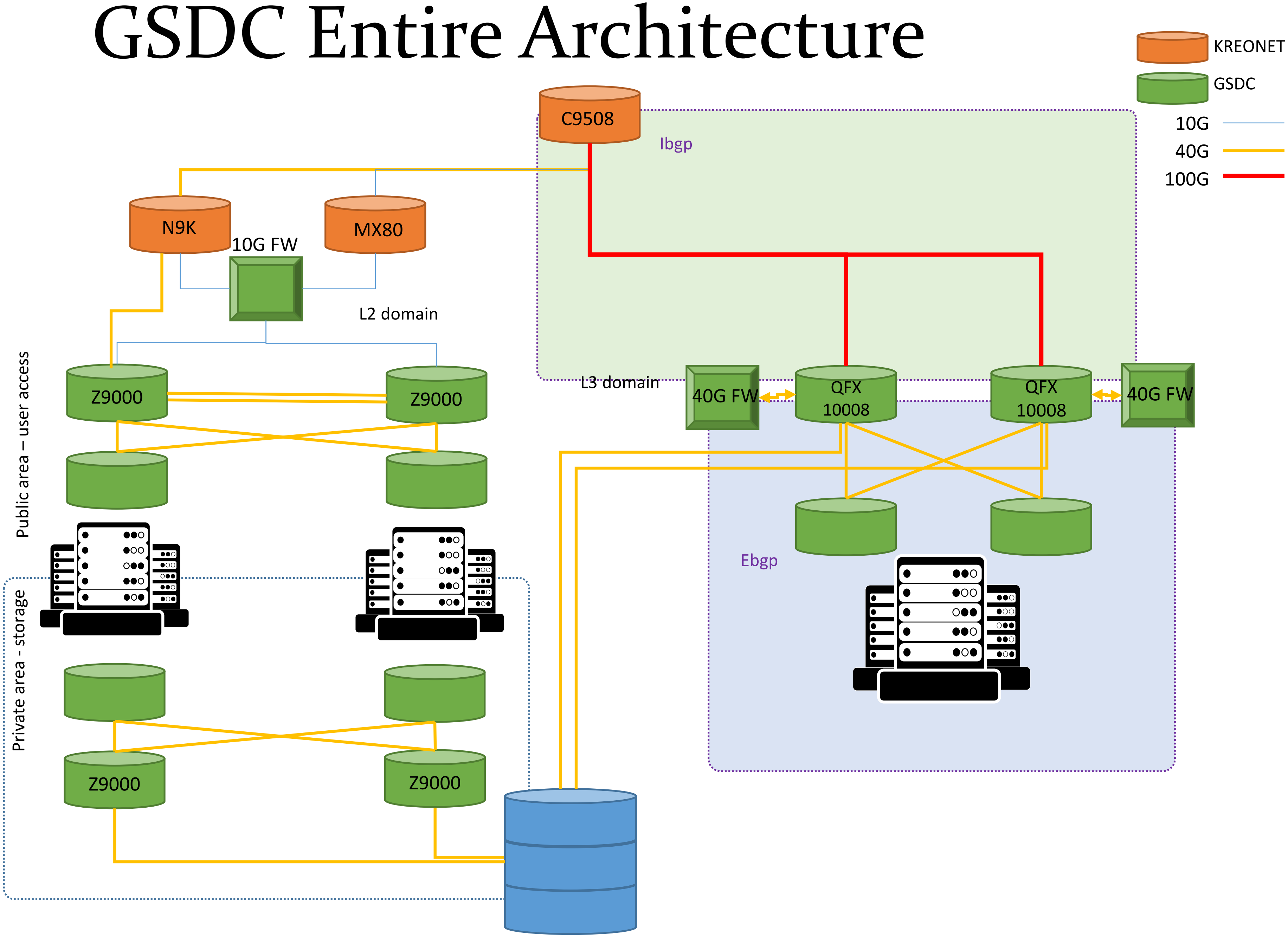
LHCOPN 20Gbps Upgrade (2021.04.14)



Kreonet-GSDC Network connection



GSDC Entire Architecture



Network requirements to support the docker

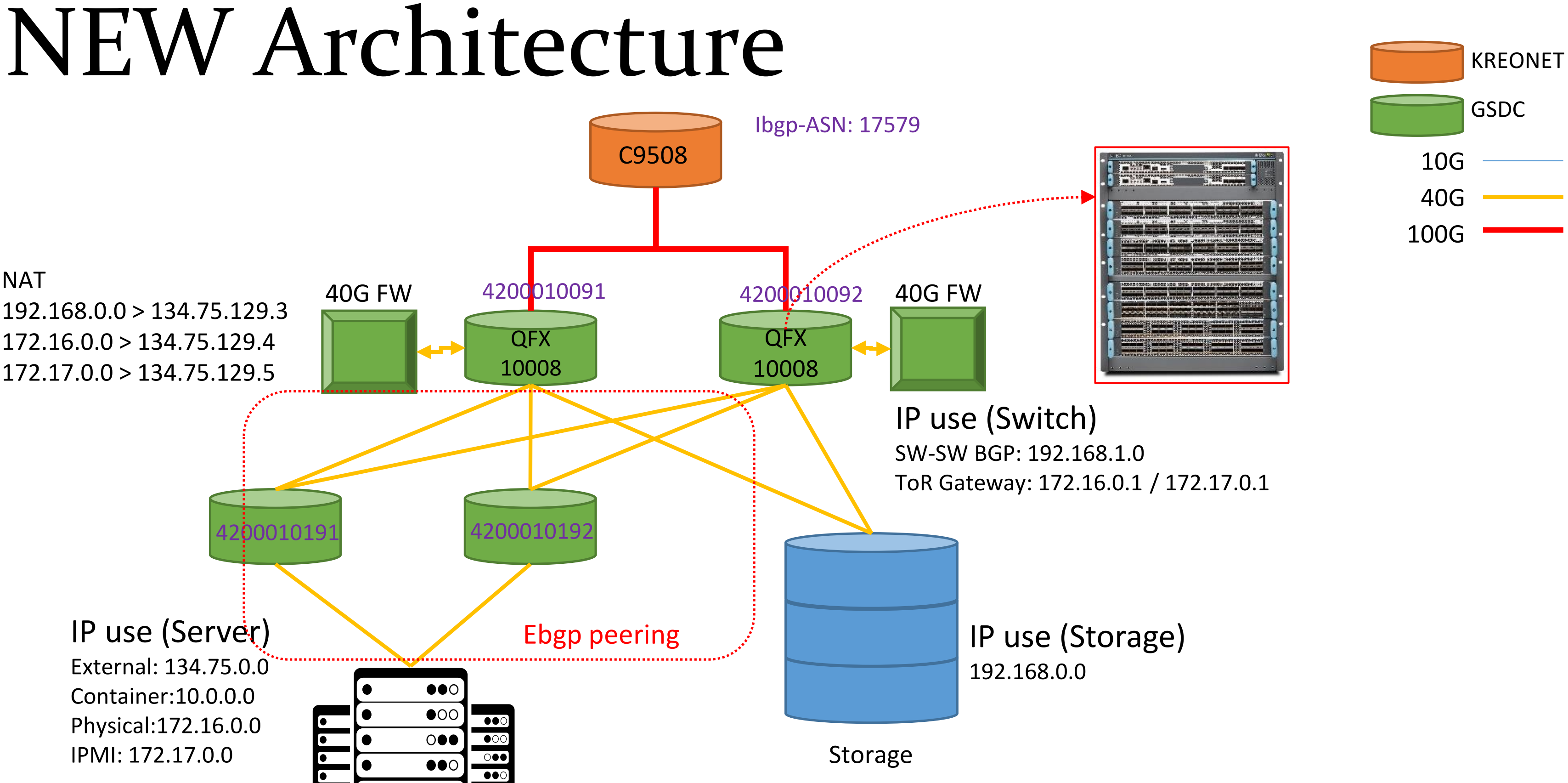
- Internal L3 routing (eBGP) – IP fabric control plane
- The server does not need to know the network architecture because the routing engine is existed in own that server. Compute server acts as a router.

Calico Network model

- The AS per rack model
- The AS per compute server model

- eBGP between spine-leaf
- iBGP between spine-leaf

NEW Architecture

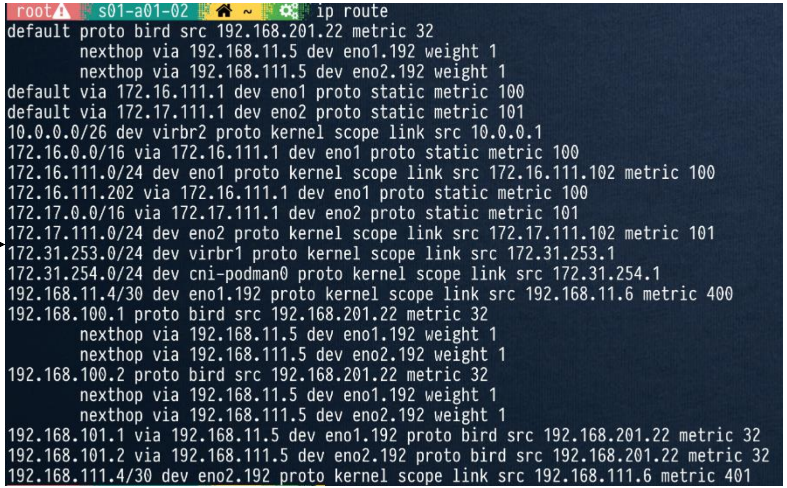


NAT
192.168.0.0 > 134.75.129.3
172.16.0.0 > 134.75.129.4
172.17.0.0 > 134.75.129.5

IP use (Switch)
SW-SW BGP: 192.168.1.0
ToR Gateway: 172.16.0.1 / 172.17.0.1

IP use (Server)
External: 134.75.0.0
Container:10.0.0.0
Physical:172.16.0.0
IPMI: 172.17.0.0

IP use (Storage)
192.168.0.0



```
 root  s01-a01-02  ~  ip route
default proto bird src 192.168.201.22 metric 32
        nexthop via 192.168.11.5 dev eno1.192 weight 1
        nexthop via 192.168.111.5 dev eno2.192 weight 1
default via 172.16.111.1 dev eno1 proto static metric 100
default via 172.17.111.1 dev eno2 proto static metric 101
10.0.0.0/26 dev virbr2 proto kernel scope link src 10.0.0.1
172.16.0.0/16 via 172.16.111.1 dev eno1 proto static metric 100
172.16.111.0/24 dev eno1 proto kernel scope link src 172.16.111.102 metric 100
172.16.111.202 via 172.16.111.1 dev eno1 proto static metric 100
172.17.0.0/16 via 172.17.111.1 dev eno2 proto static metric 101
172.17.111.0/24 dev eno2 proto kernel scope link src 172.17.111.102 metric 101
172.31.253.0/24 dev virbr1 proto kernel scope link src 172.31.253.1
172.31.254.0/24 dev cni-podman0 proto kernel scope link src 172.31.254.1
192.168.11.4/30 dev eno1.192 proto kernel scope link src 192.168.11.6 metric 400
192.168.100.1 proto bird src 192.168.201.22 metric 32
        nexthop via 192.168.11.5 dev eno1.192 weight 1
        nexthop via 192.168.111.5 dev eno2.192 weight 1
192.168.100.2 proto bird src 192.168.201.22 metric 32
        nexthop via 192.168.11.5 dev eno1.192 weight 1
        nexthop via 192.168.111.5 dev eno2.192 weight 1
192.168.101.1 via 192.168.11.5 dev eno1.192 proto bird src 192.168.201.22 metric 32
192.168.101.2 via 192.168.111.5 dev eno2.192 proto bird src 192.168.201.22 metric 32
192.168.111.4/30 dev eno2.192 proto kernel scope link src 192.168.111.6 metric 401
```

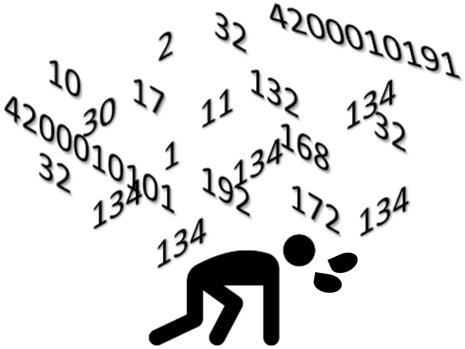
Number purgatory

IP Consumption

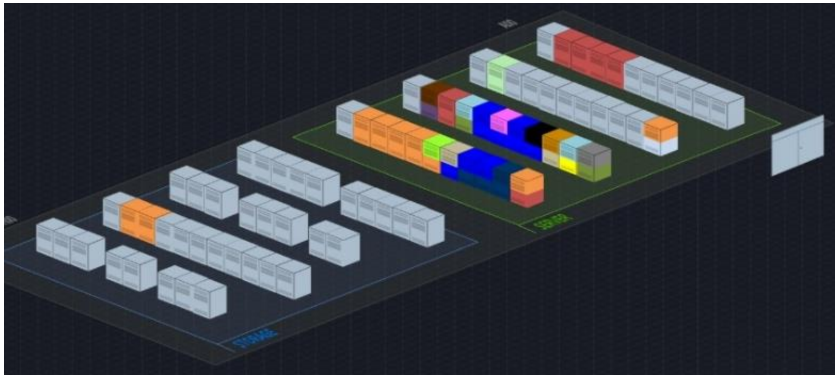
- Network device
  - Loopback (BGP establish) – 192.168.0.0/32
  - Interlocking (BGP peering) – 192.168.011.0/30
  - System migration vlan - 192
- Server
  - Loopback(1) - 192.168.20.0/32
  - Interlocking(2) – 192.168.0.0/30
  - Container – 10.0.0.0/8
  - Mgmt container(2) – 172.30.0.0/16
  - Physical NIC(2) – 172.17.0.0/16, 172.16.0.0/16
  - IPMI(1,NIC 1 share) – 172.16.0.0/16
- NAT
  - 134.75.129.0/24
- Public IP
  - 134.75.132.0/24
- Available IP range
  - Private : 192.168.0.0, 10.0.0.0/8, 172.17.0.0/16, 172.16.0.0/16
  - Public : 134.75.129.0/24, 134.75.130.0/24, 134.75.131.0/24, 134.75.132.0/24, 134.75.133.0/24

32 bit ASN

- Each device in same rack has one ASN
- Gen role
  - 42000+Containement+Rack+device
  - Ex) 42000 1 01 91



Interlocking IP pair



Max # of ToR SW : 40

192.168.
3rd octet

Rack #	1	2	3	4	5	6	7	8	9	10
A line	11		13		15		17		19	
	111		113		115		117		119	
B line	21		23		25		27		29	
	121		123		125		127		129	
C line	31		33		35		37		39	
	131		133		135		137		139	
D line	41		43		45		47		49	
	141		143		145		147		149	

4th octet

A01up SW-Server: 192.168.011. / 30

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
SW IP	1	5	9	13	17	21	25	29	33	37	41	45	49	53	57	61	65	69	73	77	81	85	89	93	97	101	105	109	113	117	121	125	129	133	137	141	145	149	153	157
server IP	2	6	10	14	18	22	26	30	34	38	42	46	50	54	58	62	66	70	74	78	82	86	90	94	98	102	106	110	114	118	122	126	130	134	138	142	146	150	154	158

A01under SW-Server: 192.168.111. / 30

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
sw IP	1	5	9	13	17	21	25	29	33	37	41	45	49	53	57	61	65	69	73	77	81	85	89	93	97	101	105	109	113	117	121	125	129	133	137	141	145	149	153	157
server IP	2	6	10	14	18	22	26	30	34	38	42	46	50	54	58	62	66	70	74	78	82	86	90	94	98	102	106	110	114	118	122	126	130	134	138	142	146	150	154	158

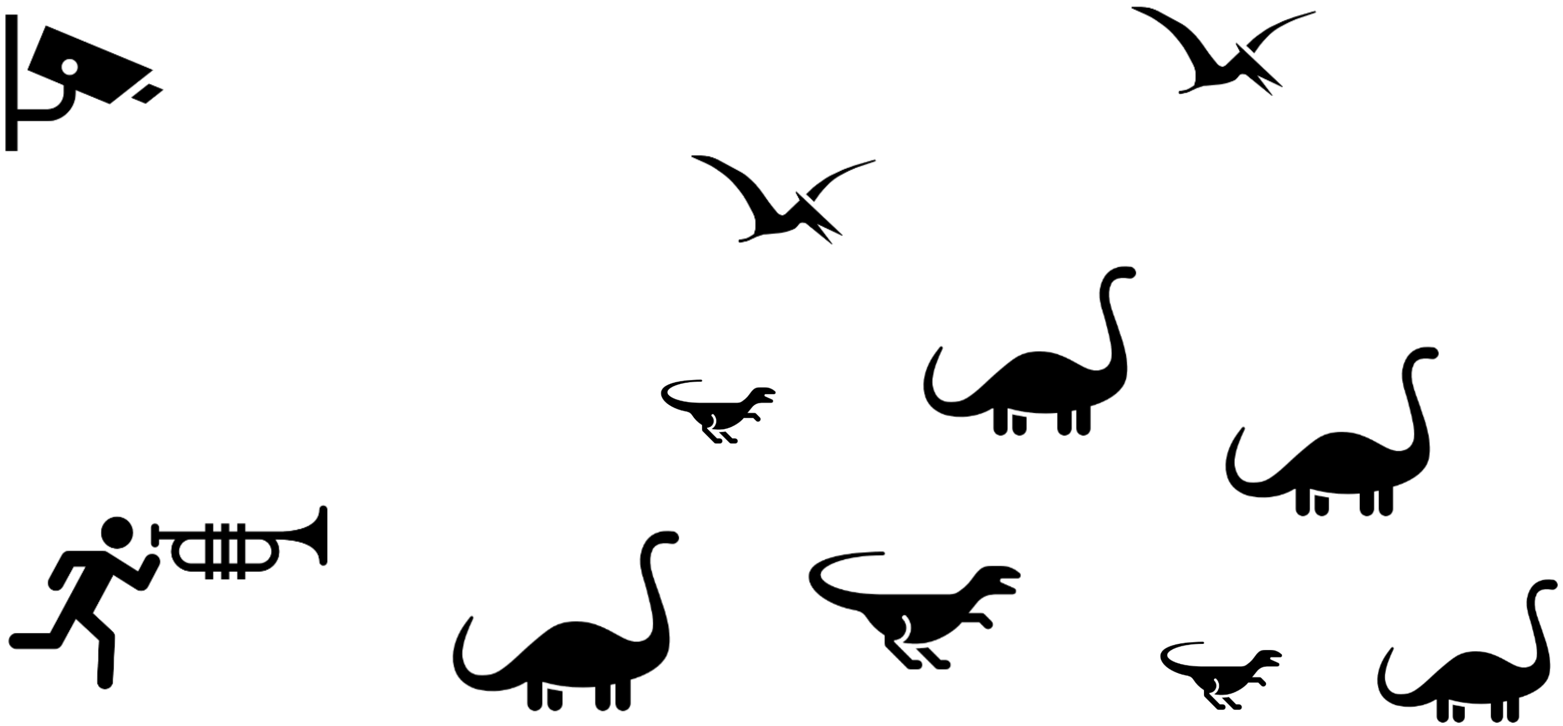
What we want

- Dynamic networking
  - Each container should communicate.
  - Network isolation is good for network manager.
  - If we use network pod using L2 or else, then the tunneling is used.
  - As the result of tunneling, network overhead is increased. (usable bandwidth is reduced)

- Limitation of datacenter size
  - The scale of gsdc is BIG?? Small?? I have no idea.

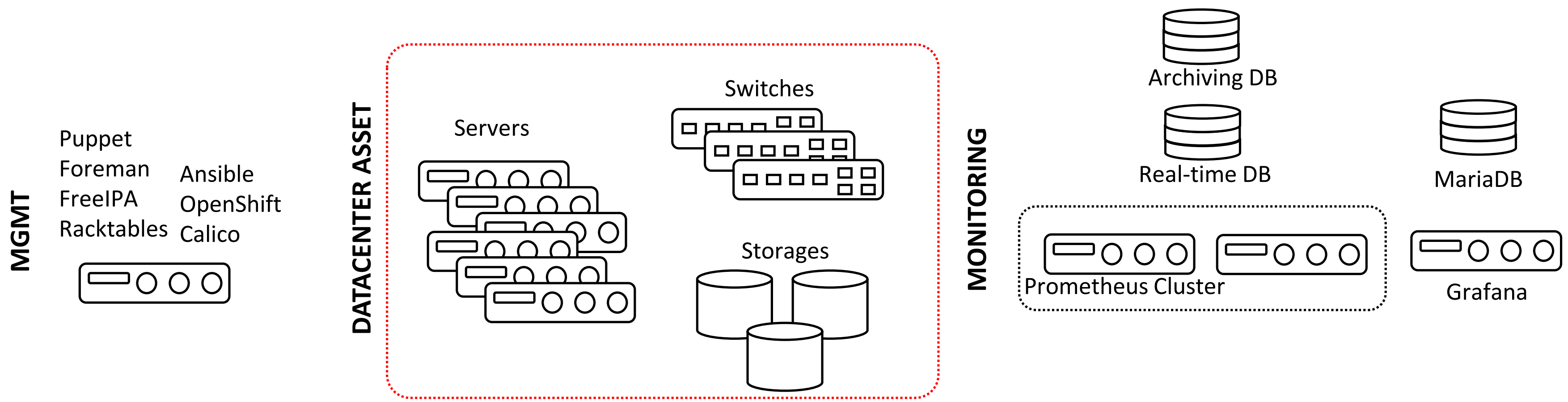
- Limitation of IP use (hard to use public DNS)
  - Some of services should use fixed public IP such as UI and grid servers.

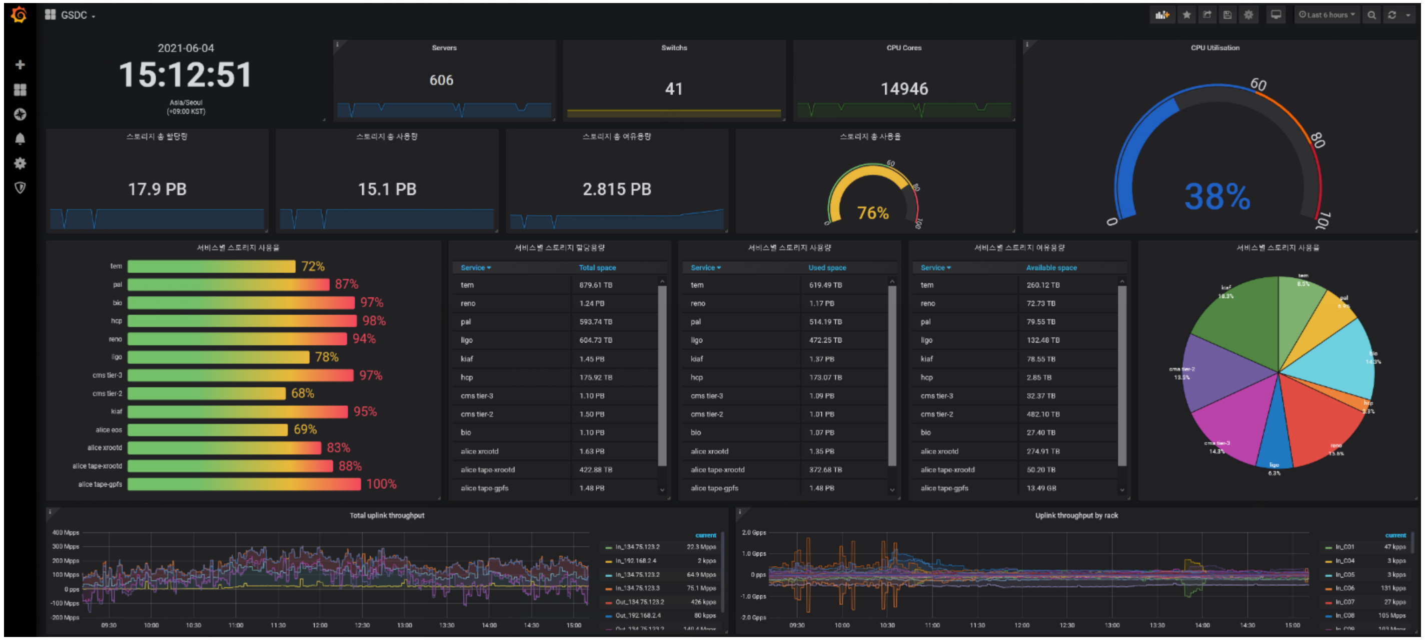
- Physical server maintain(container move, NAT limitation)
  - In the situation of server maintain, the services should be move to other physical machine ASAP.



MGMT & MONITORING

Monitoring





서비스별 스토리지 할당용량

Service	Total space
tem	879.61 TB
reno	1.24 PB
pal	593.74 TB
ligo	604.73 TB
kiaf	1.45 PB
hcp	175.92 TB
cms tier-3	1.10 PB
cms tier-2	1.50 PB
bio	1.10 PB
alice xrootd	1.63 PB
alice tape-xrootd	422.88 TB
alice tape-gpfs	1.48 PB

서비스별 스토리지 사용량

Service	Used space
tem	619.49 TB
reno	1.17 PB
pal	514.19 TB
ligo	472.25 TB
kiaf	1.37 PB
hcp	173.07 TB
cms tier-3	1.09 PB
cms tier-2	1.01 PB
bio	1.07 PB
alice xrootd	1.35 PB
alice tape-xrootd	372.68 TB
alice tape-gpfs	1.48 PB

서비스별 스토리지 여유용량

Service	Available space
tem	260.12 TB
reno	72.73 TB
pal	79.55 TB
ligo	132.48 TB
kiaf	78.55 TB
hcp	2.85 TB
cms tier-3	32.37 TB
cms tier-2	482.10 TB
bio	27.40 TB
alice xrootd	274.91 TB
alice tape-xrootd	50.20 TB
alice tape-gpfs	13.49 GB

Thank you