

Data Center Network

Jin Kim



facebook

December 2010

Agenda

- Network
- Basic Knowledge
- Data Center Network

Network

- History

- 1960`s : transmit **bits** across a communication medium
- 1970`s : transmit **packets** across a communication medium
- 1980`s : provide communication **services** across a series of interconnected network
- 1990`s : provide **high-speed, broadband communication services** to support high-performance computing and multimedia applications across the globe?
- 2000`s~ : more, more

Internet

- History

- 1969 : ARPANET
- 1971 : ARPANET (23 hosts)
- 1973 : ARPANET (England-Norway)
- 1982 : term “Internet”
- 1992 : Internet Society is chartered. World Wide Web released by CERN



Common communication tasks

- Data encoding
- Signal generation
- Synchronization
- Error detection and correction
- Flow control
- Multiplexing
- Addressing
- Routing
- Message formatting
- Security
- System mgmt

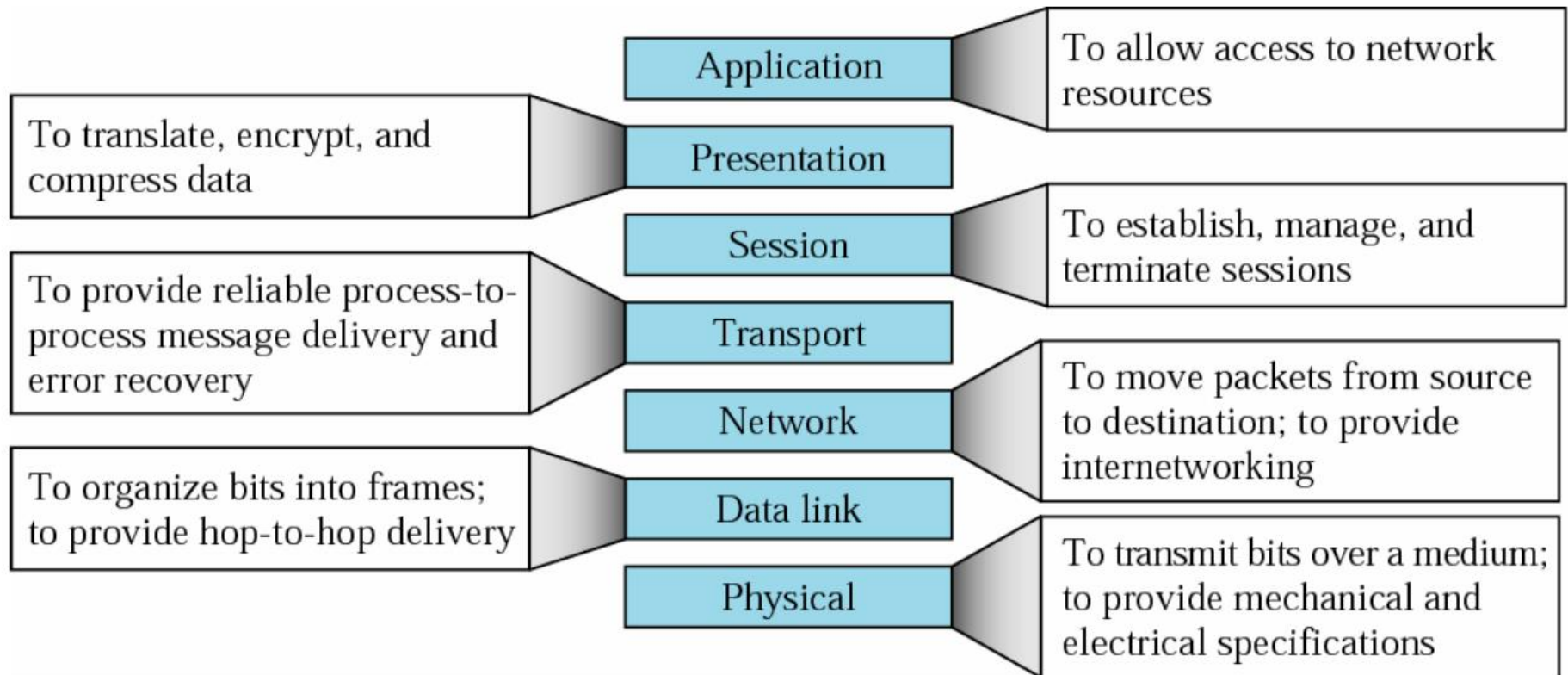
Types of communication networks

- Local Area Network (LAN)
- Metropolitan Area Network (MAN)
- Wide Area Network (WAN)

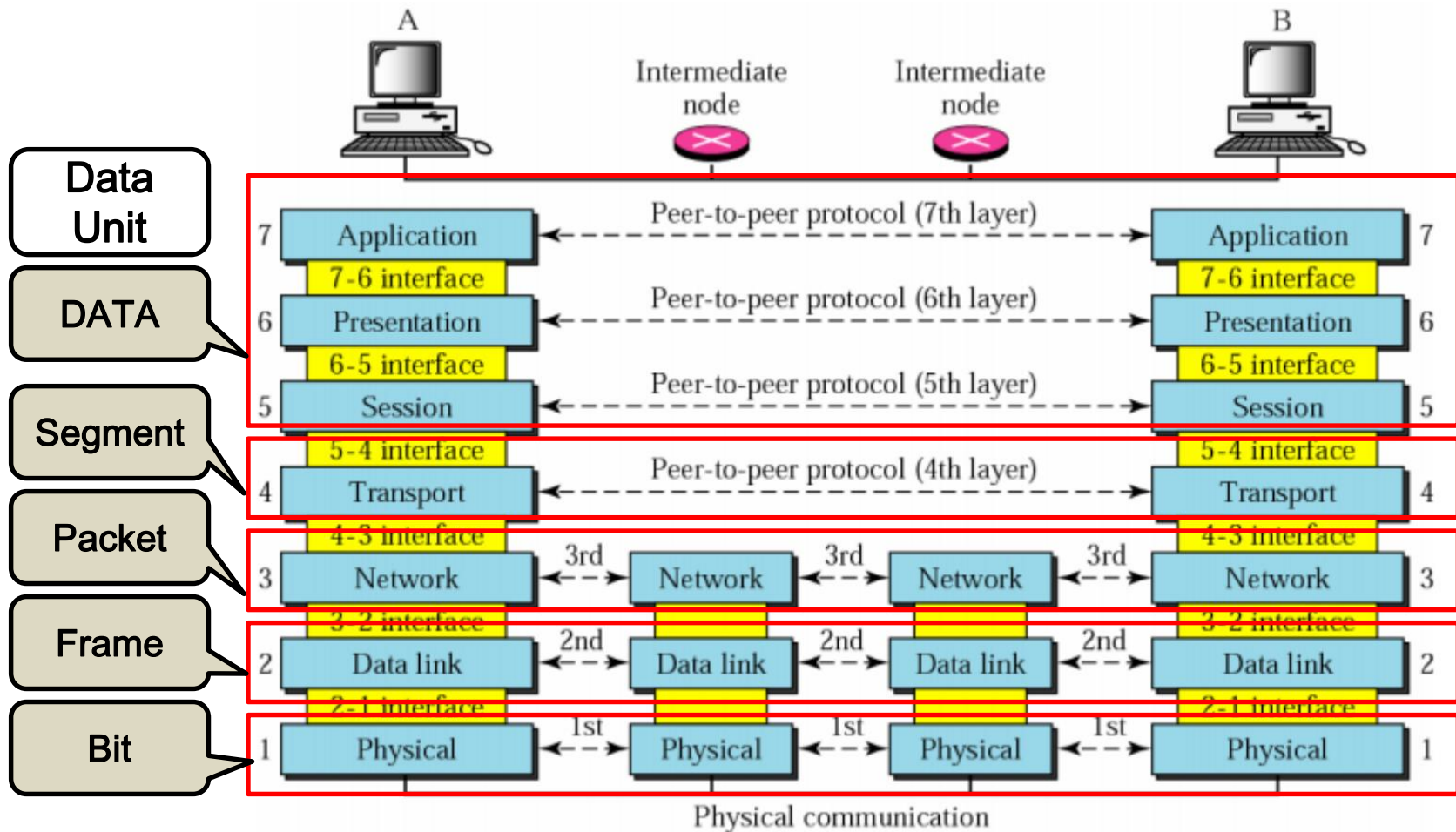
ISO-OSI reference model

- International Standards Organization
 - Open Systems Interconnection reference model is a framework for connecting computers on a network
- Motivation
 - Reduce the complexity of networking software
 - Support various protocols

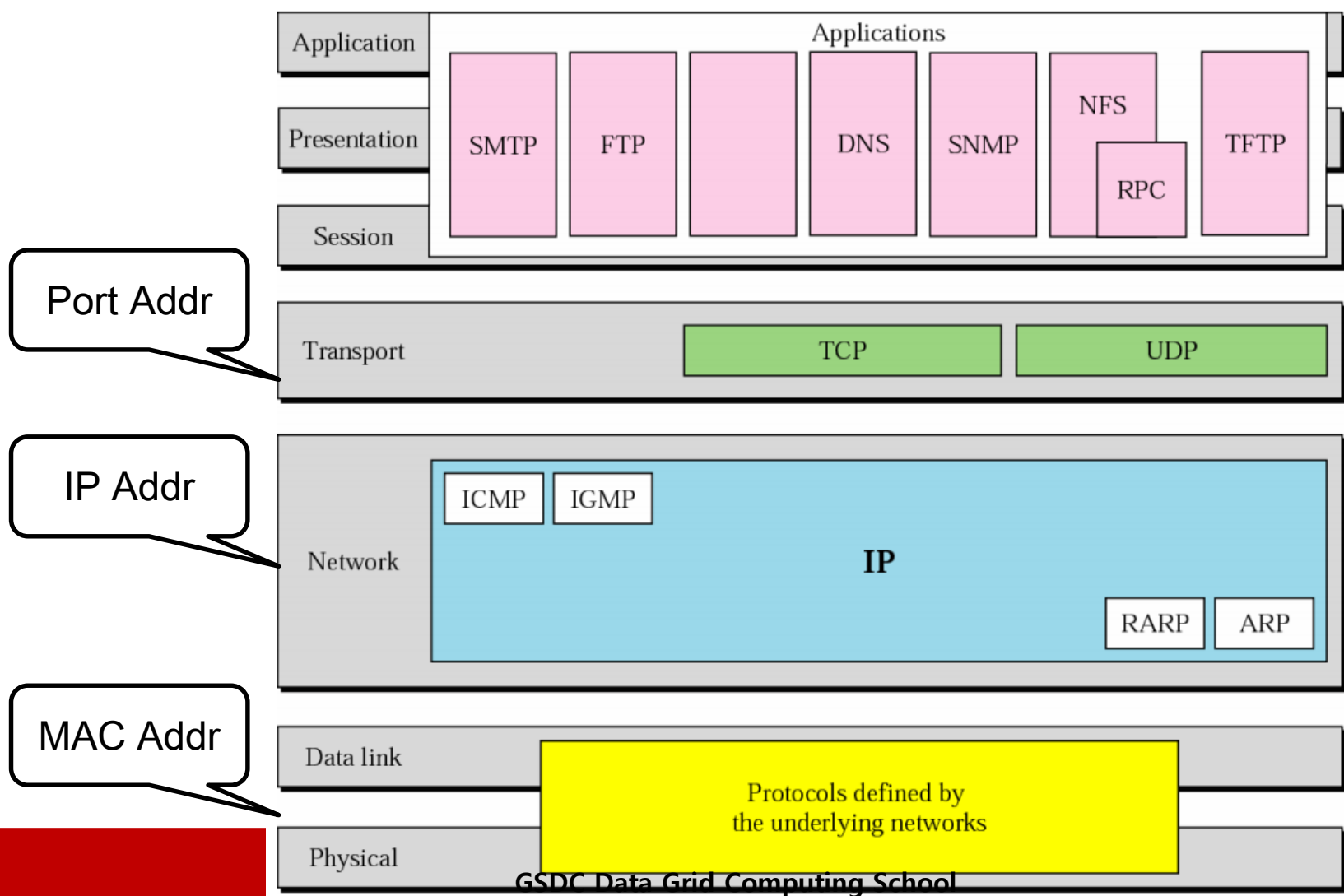
OSI 7 Layer



OSI 7 Layer(2/3)

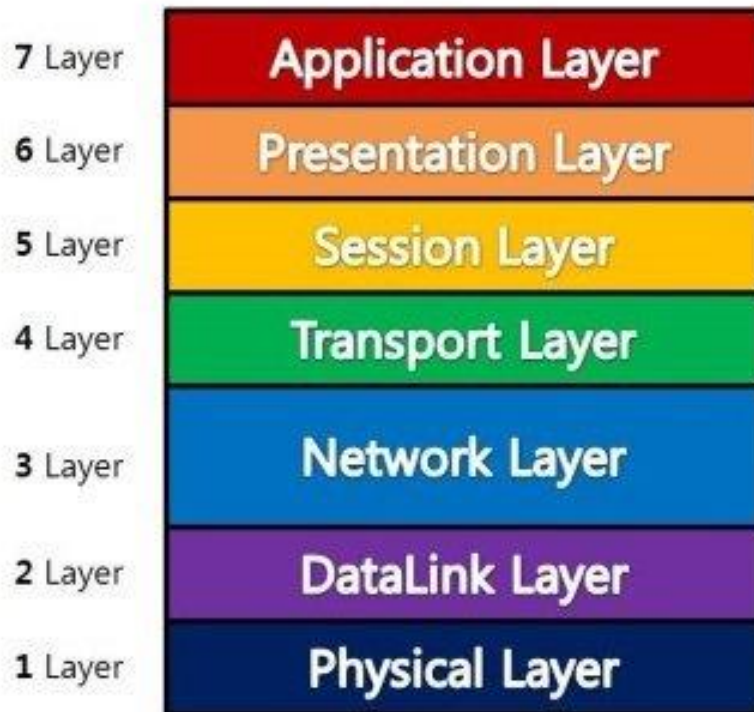


OSI 7 Layer(3/3)

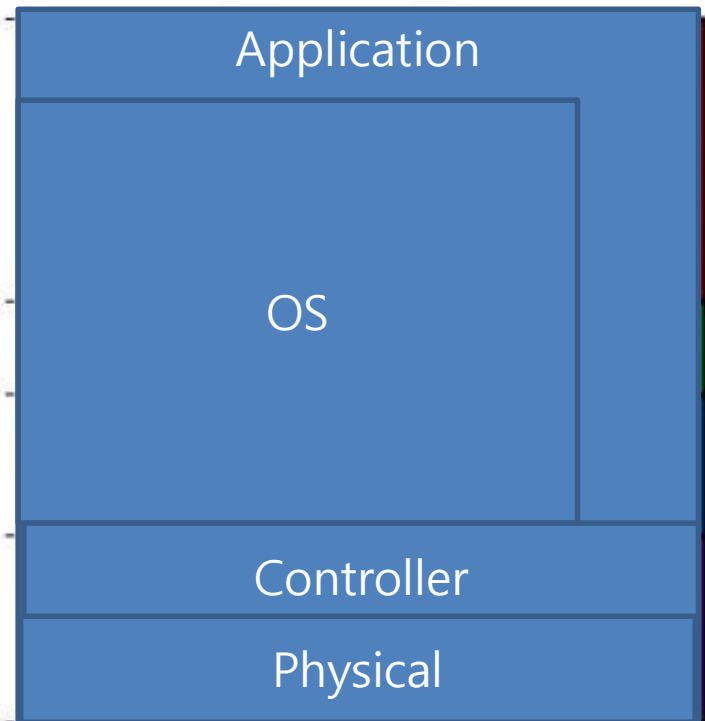


Networking layers

OSI 7 Layer Model



The Anarchistic model



Based on TCP/IP

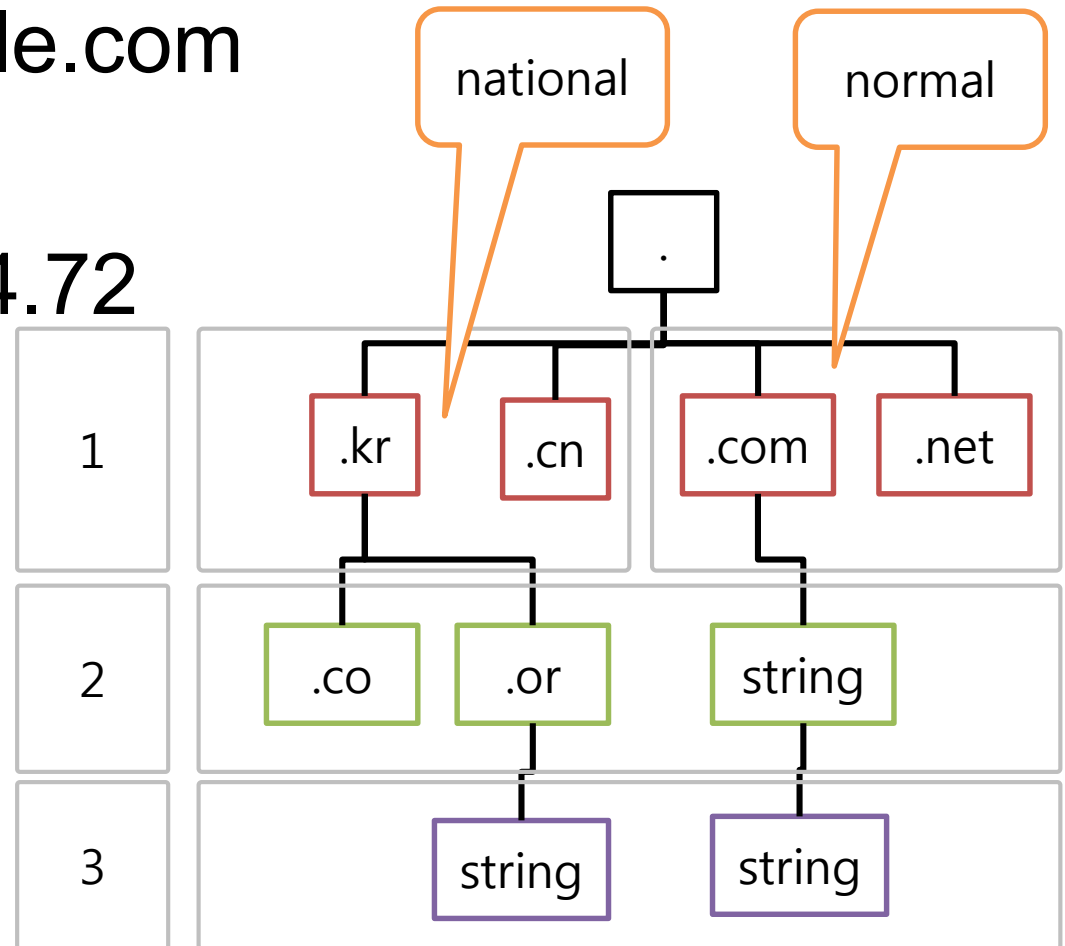
- TCP/IP is
 - A suite of protocols
 - Rules for sending and receiving data across networks
 - Addressing
 - Mgmt and verification

DNS (Domain Name System)

- `http://www.google.com`

||

- `http://74.125.224.72`



eth0: Capturing - Wireshark

File Edit View Go Capture Analyze Statistics Help

Filter: + Expression... Clear Apply

No.	Time	Source	Destination	Protocol	Info
46	139.931167	wistron_07:07:ee	broadcast	ARP	who has 192.168.1.254? Tell 192.168.1.68
47	139.931463	ThomsonT 08:35:4f	Wistron 07:07:ee	ARP	192.168.1.254 is at 00:90:d0:08:35:4f
48	139.931466	192.168.1.68	192.168.1.254	DNS	Standard query A www.google.com
49	139.975406	192.168.1.254	192.168.1.68	DNS	Standard query response CNAME www.l.google.com A 66.102.9.99
50	139.976811	192.168.1.68	66.102.9.99	TCP	62216 > http [SYN] Seq=0 Win=8192 Len=0 MSS=1460 WS=2
51	140.079578	66.102.9.99	192.168.1.68	TCP	http > 62216 [SYN, ACK] Seq=0 Ack=1 Win=5720 Len=0 MSS=1430
52	140.079583	192.168.1.68	66.102.9.99	TCP	62216 > http [ACK] Seq=1 Ack=1 Win=65780 Len=0
53	140.080278	192.168.1.68	66.102.9.99	HTTP	GET /complete/search?hl=en&client=suggest&js=true&q=m&cp=1 H
54	140.086765	192.168.1.68	66.102.9.99	TCP	62216 > http [FIN, ACK] Seq=805 Ack=1 Win=65780 Len=0
55	140.086921	192.168.1.68	66.102.9.99	TCP	62218 > http [SYN] Seq=0 Win=8192 Len=0 MSS=1460 WS=2
56	140.197484	66.102.9.99	192.168.1.68	TCP	http > 62216 [ACK] Seq=1 Ack=805 Win=7360 Len=0
57	140.197777	66.102.9.99	192.168.1.68	TCP	http > 62216 [FIN, ACK] Seq=1 Ack=806 Win=7360 Len=0
58	140.197811	192.168.1.68	66.102.9.99	TCP	62216 > http [ACK] Seq=806 Ack=2 Win=65780 Len=0

Frame 1 (42 bytes on wire, 42 bytes captured)

Ethernet II, Src: Vmware_38:eb:0e (00:0c:29:38:eb:0e), Dst: Broadcast (ff:ff:ff:ff:ff:ff)

Address Resolution Protocol (request)

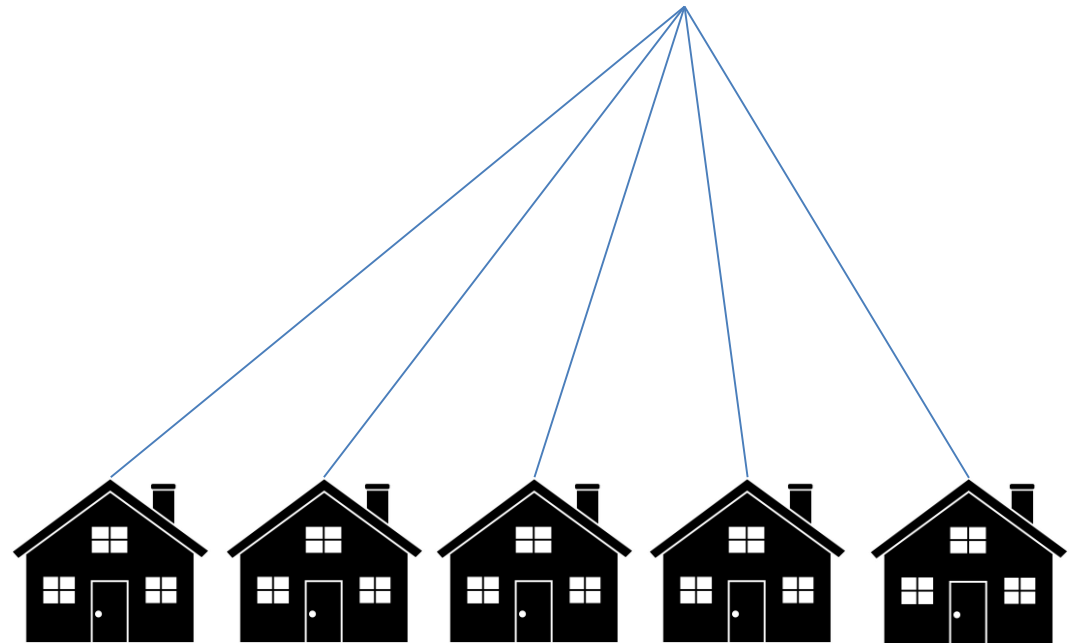
```

0000  ff ff ff ff ff ff 00 0c 29 38 eb 0e 08 06 00 01  ..... )8.....
0010  08 00 06 04 00 01 00 0c 29 38 eb 0e c0 a8 39 80  ..... )8....9.
0020  00 00 00 00 00 00 c0 a8 39 02  ..... 9.
  
```

eth0: <live capture in progress> Fil... Packets: 445 Displayed: 445 Marked: 0 Profile: Default

IP

- Resolved IP
 - 127.0.0.1
 - . . . 0~2 / . . . 254~255
 - 192.....
- IPv4 / IPv6
- Allocation IP

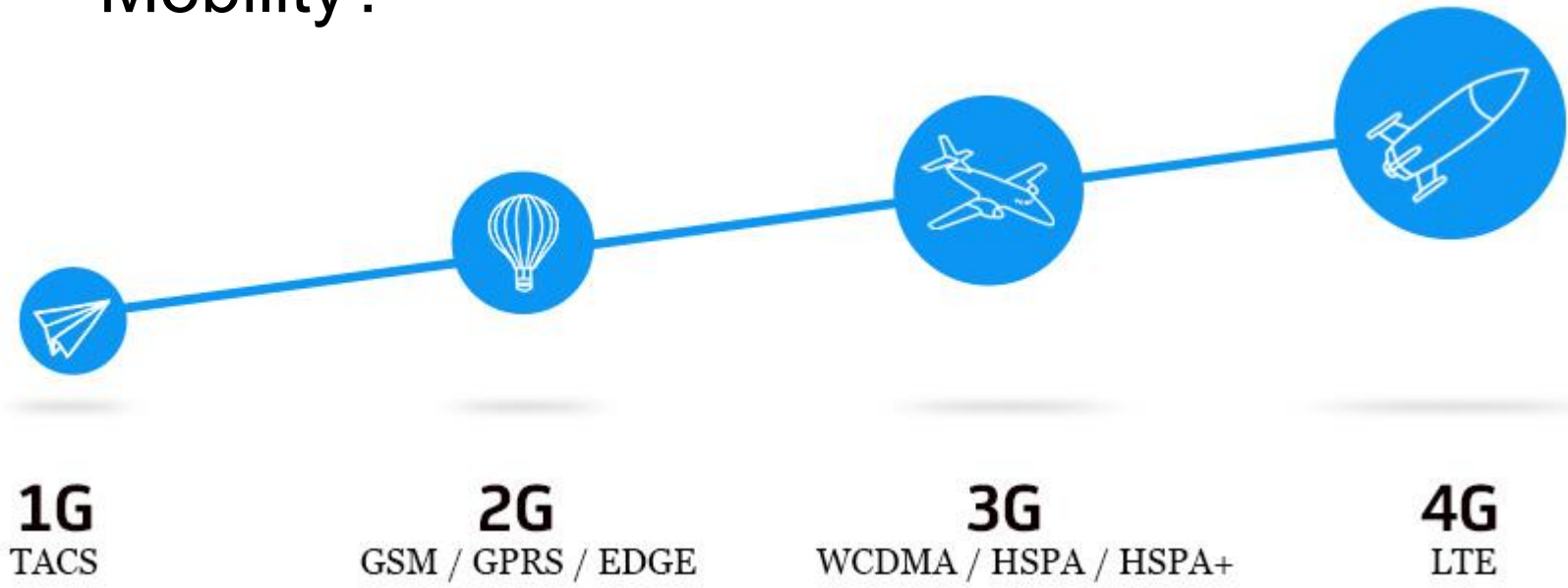


2007 LG



Else..

- Wifi vs 3,4,5 G
 - Internet connection for host (PC, mobile)
 - Mobility?

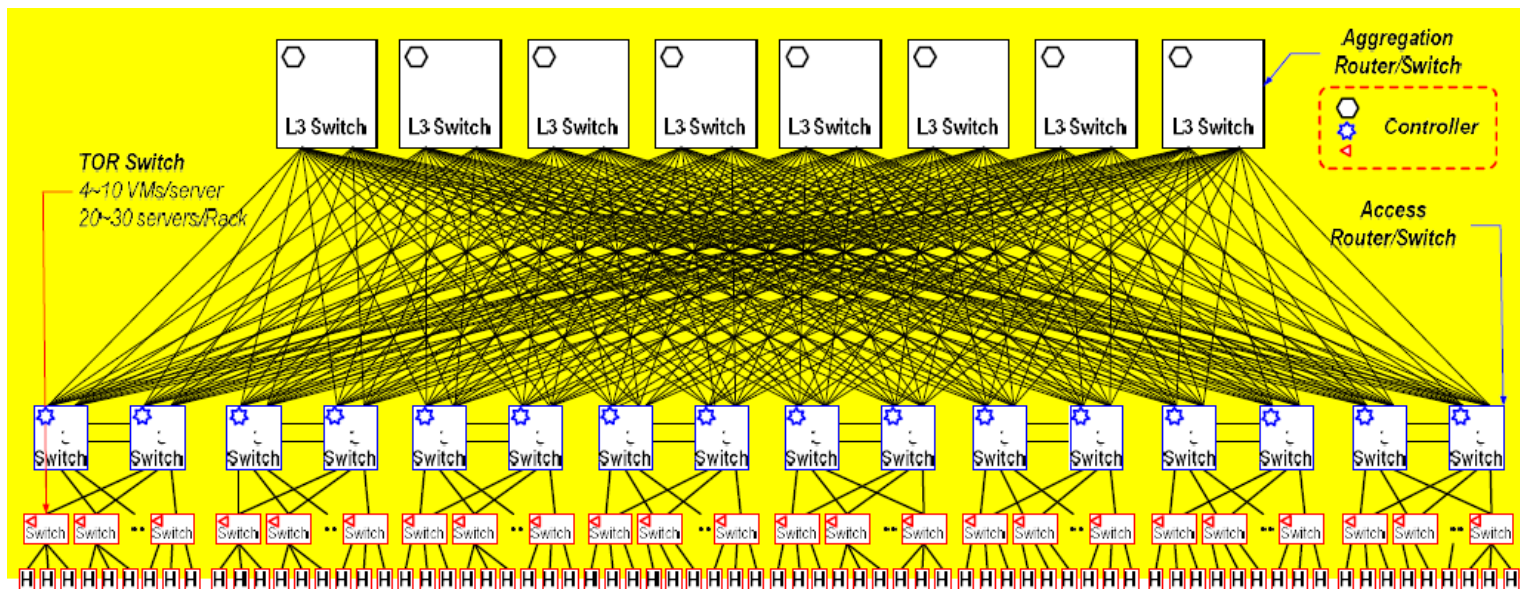


Date Center Network



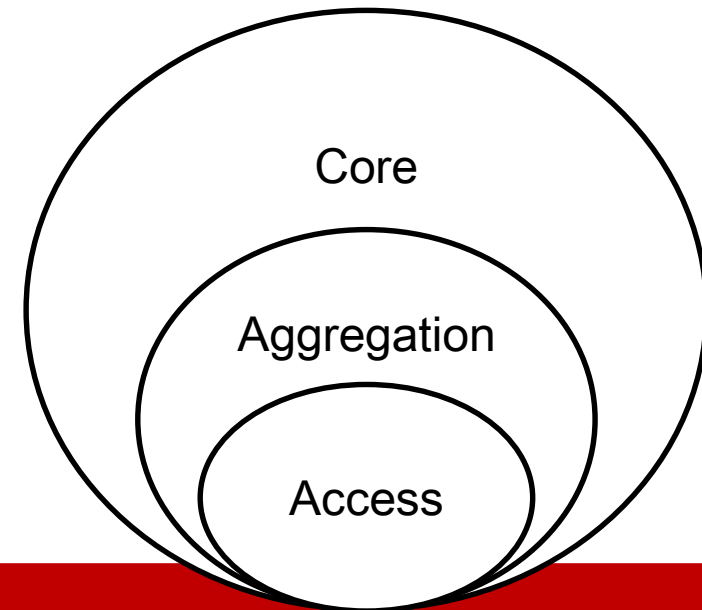
Data center

- Data center is a pool of resources (computational, storage, network) interconnected using a communication network



Datacenter

- Type
 - Three-tier
 - Fat tree: High throughput, low latency (supercomputing interconnecting)
 - Dcell (multiple NIC to connect host directly)
- Structure
 - Tree: several depth (north-south traffic)
 - Spin-leaf: 2 depth only (east-west traffic)
- Performance factor
 - Latency, throughput -> traffic pattern



Datacenter

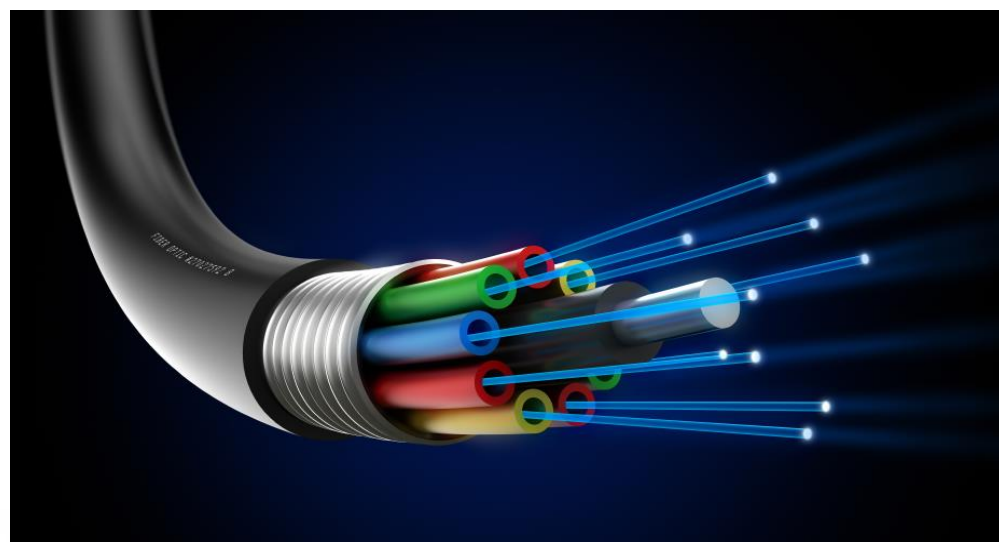
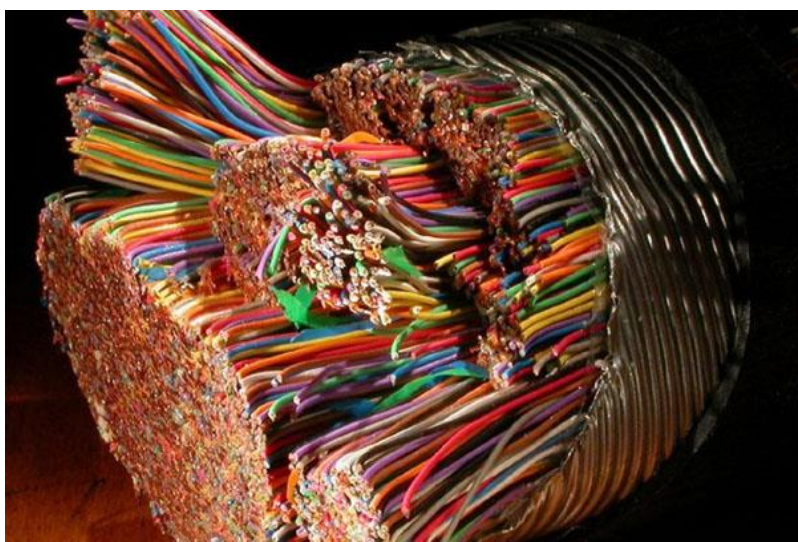
Unstructured cabling



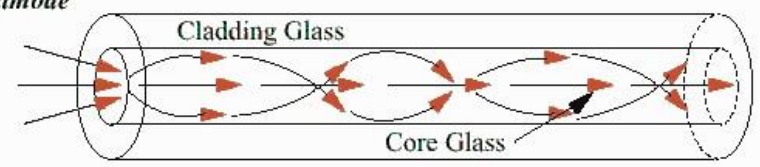
Structured cabling



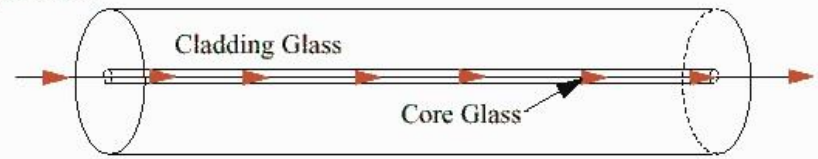
Cables



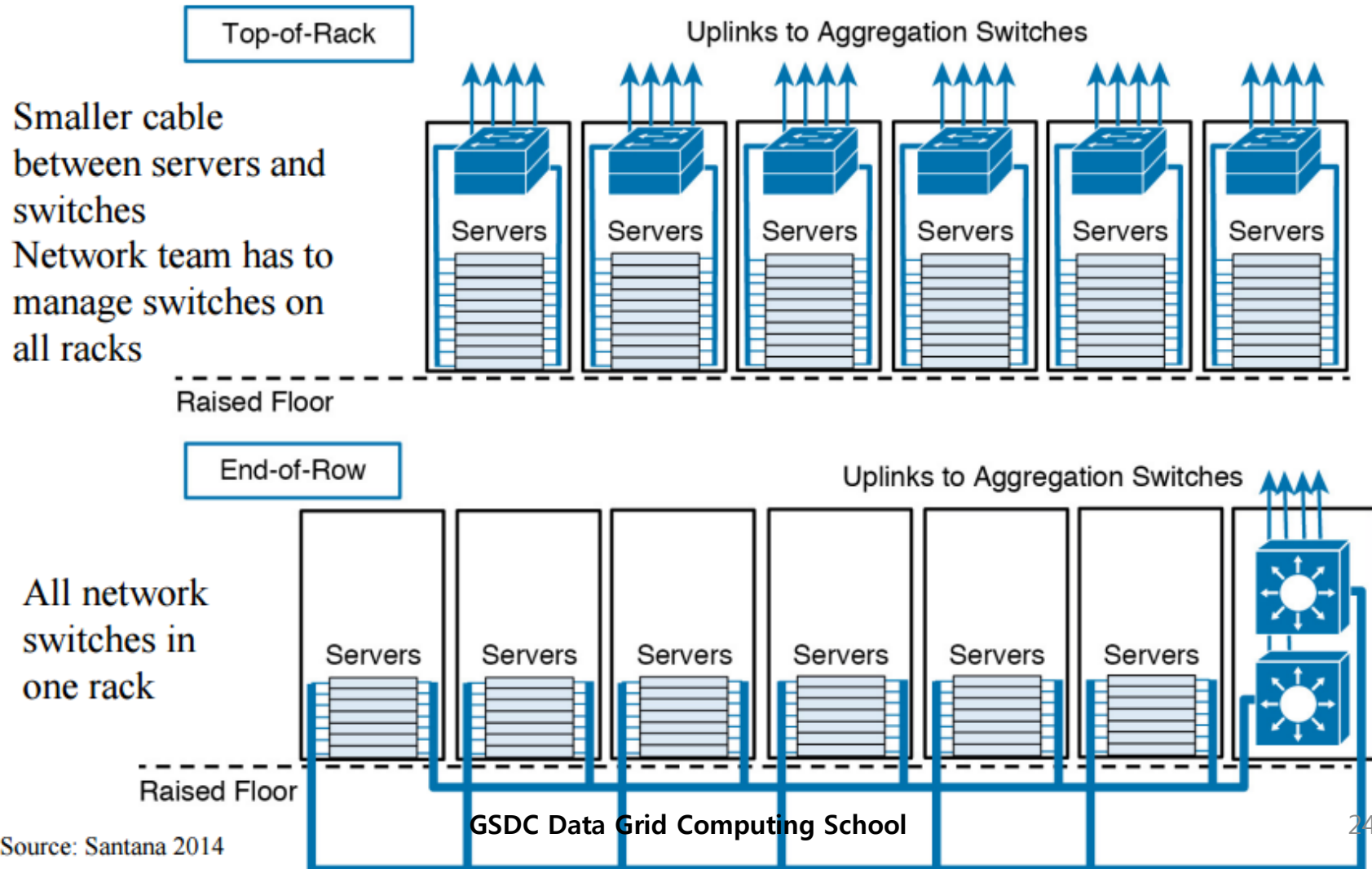
Multimode



Single-Mode

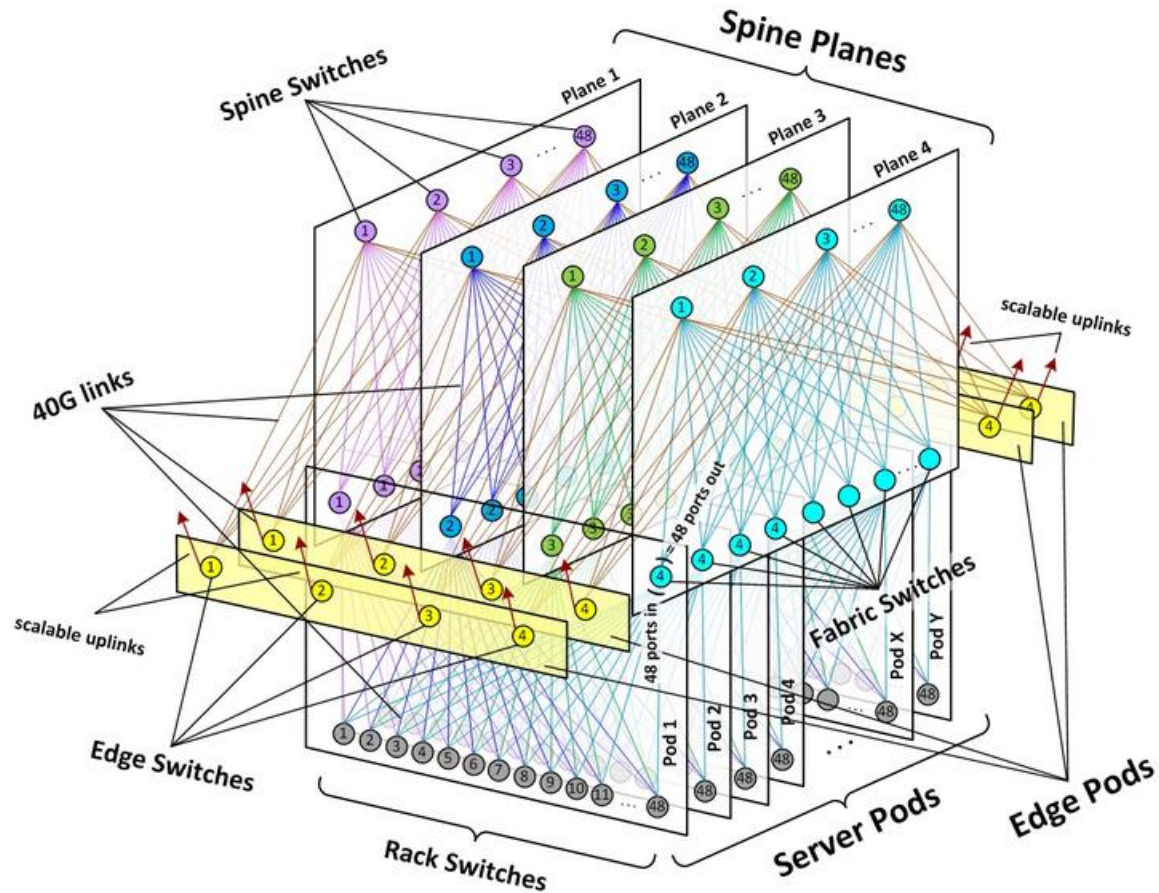


Datacenter (SW location)



Datacenter

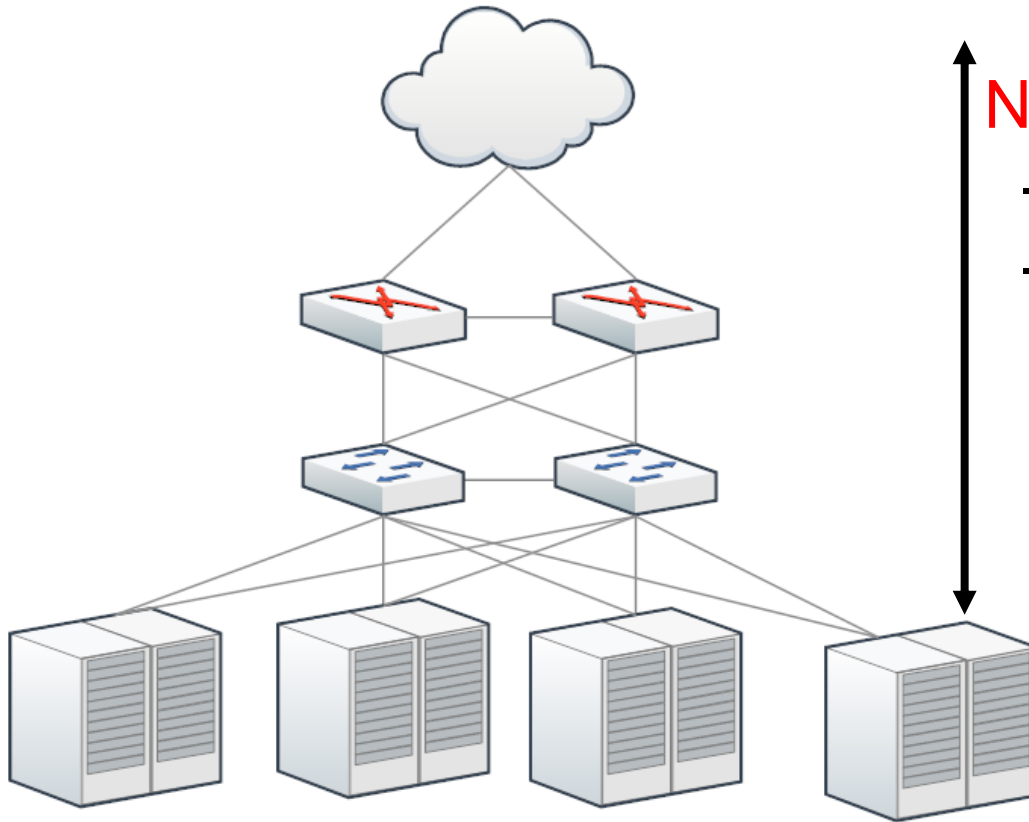
Schematic of Facebook data center fabric network topology



Characteristic

- Multi layer: latency, complexity, QoS, virtualization
- East-west vs north-south
- System add: case-by-case

Traffic pattern



North-south traffic

- internet connecting
- SW-SW, SW-Router

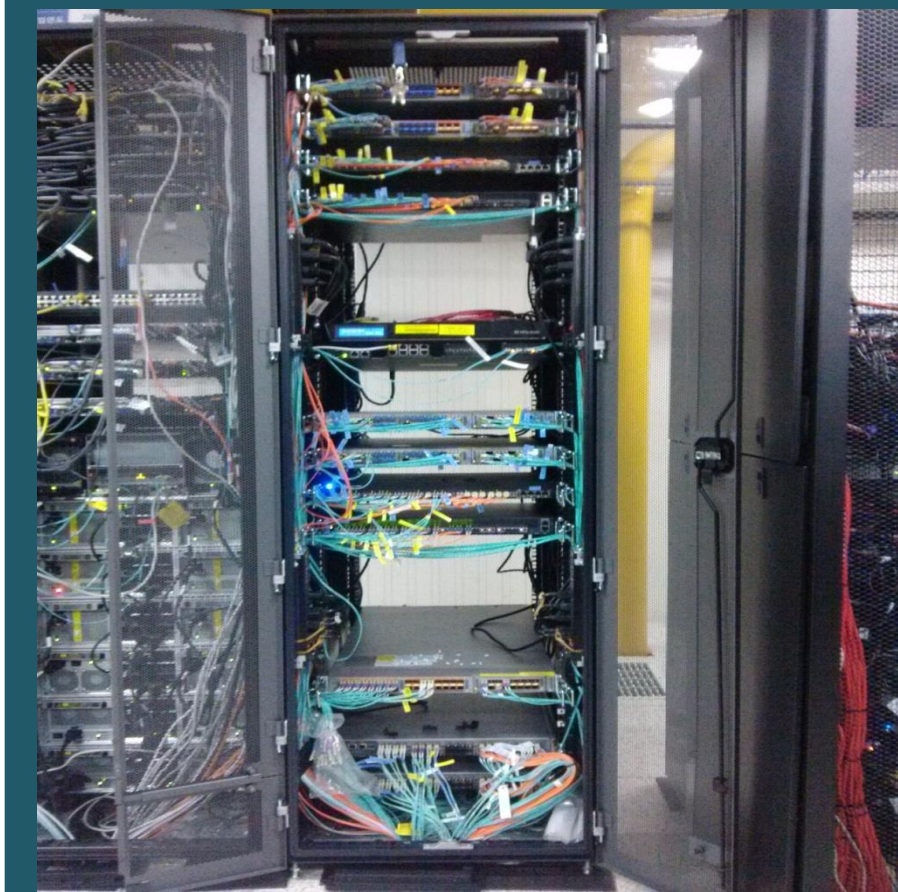
East-west traffic (computing with storage connection)

- Host-Host, Host-Storage

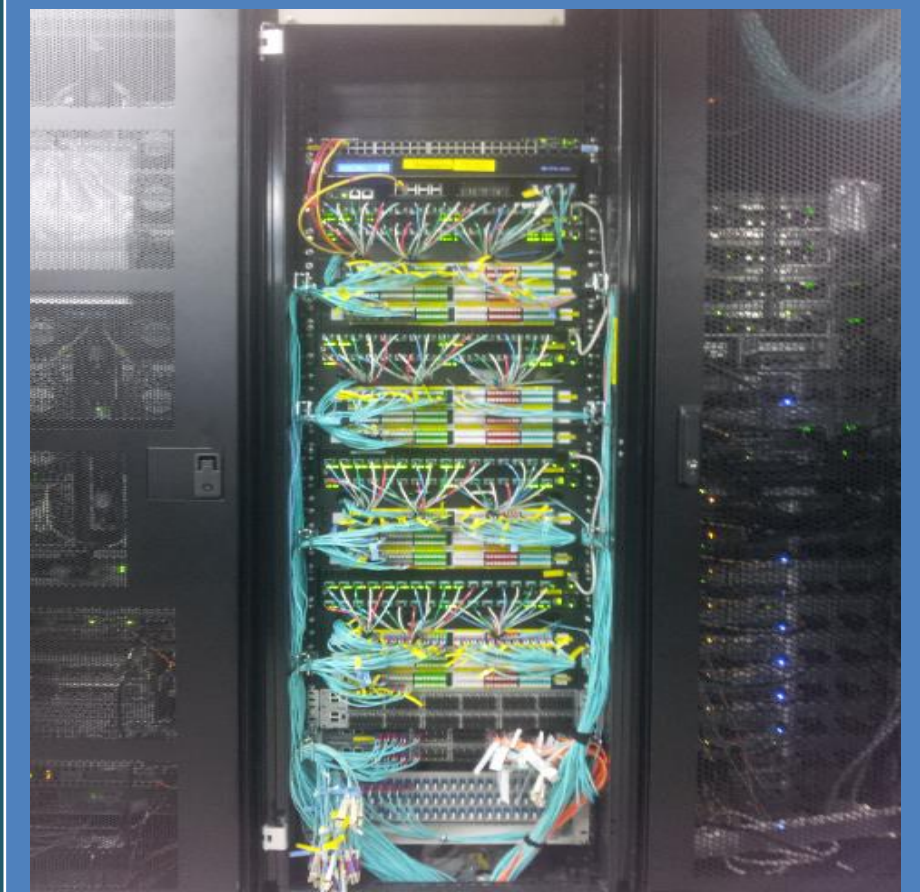
GSDC DCN

Cable hell

Main rack switch 설치 전 (Public, Private)

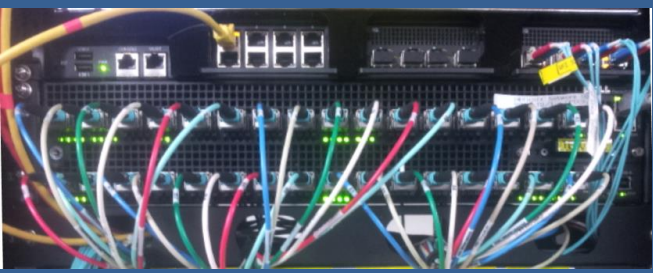

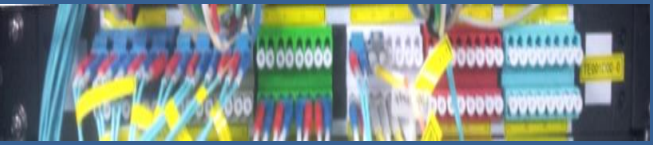



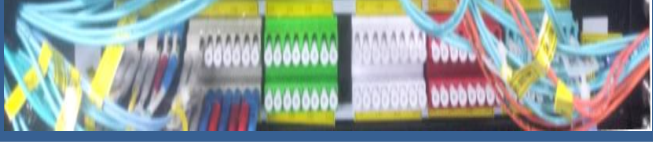
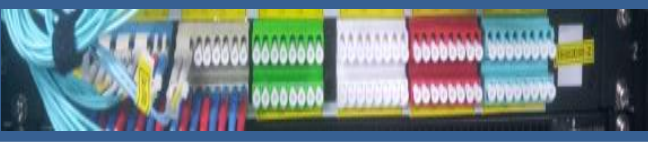


Main rack switch 설치 후 (External, Internal)



10G Switch and Patch panel

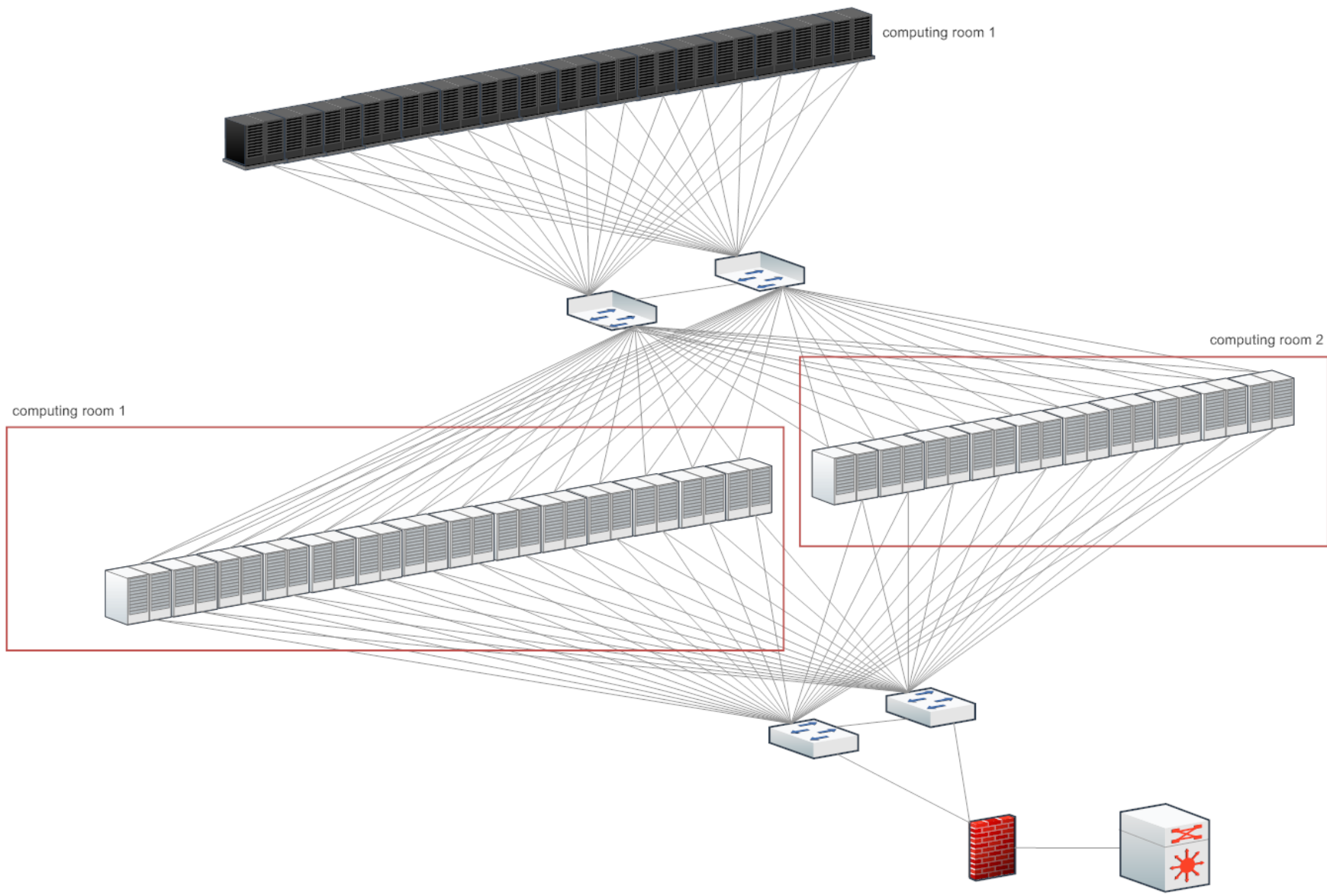
Force10 Z9000(External Main Switch)

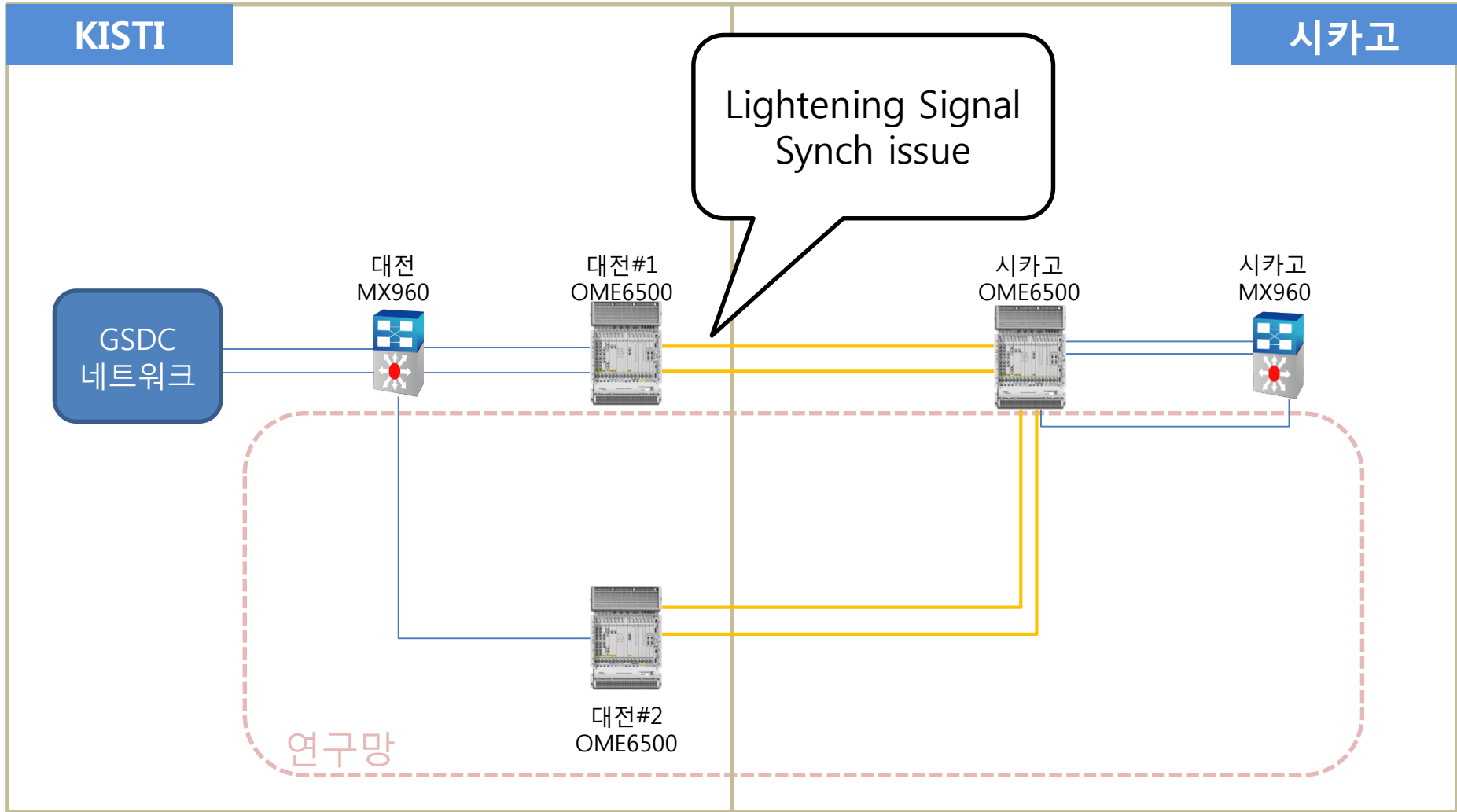
<p>R00E01 (VLT Primary)</p>		<p>R00E02 (VLT Secondary)</p>	
<p>패치패널#1</p>		<p>패치패널#1</p>	
<p>패치패널#2</p>		<p>패치패널#2</p>	
<p>패치패널#3</p>		<p>패치패널#3</p>	

GSDCN(LAN)

internal
area
(storage)

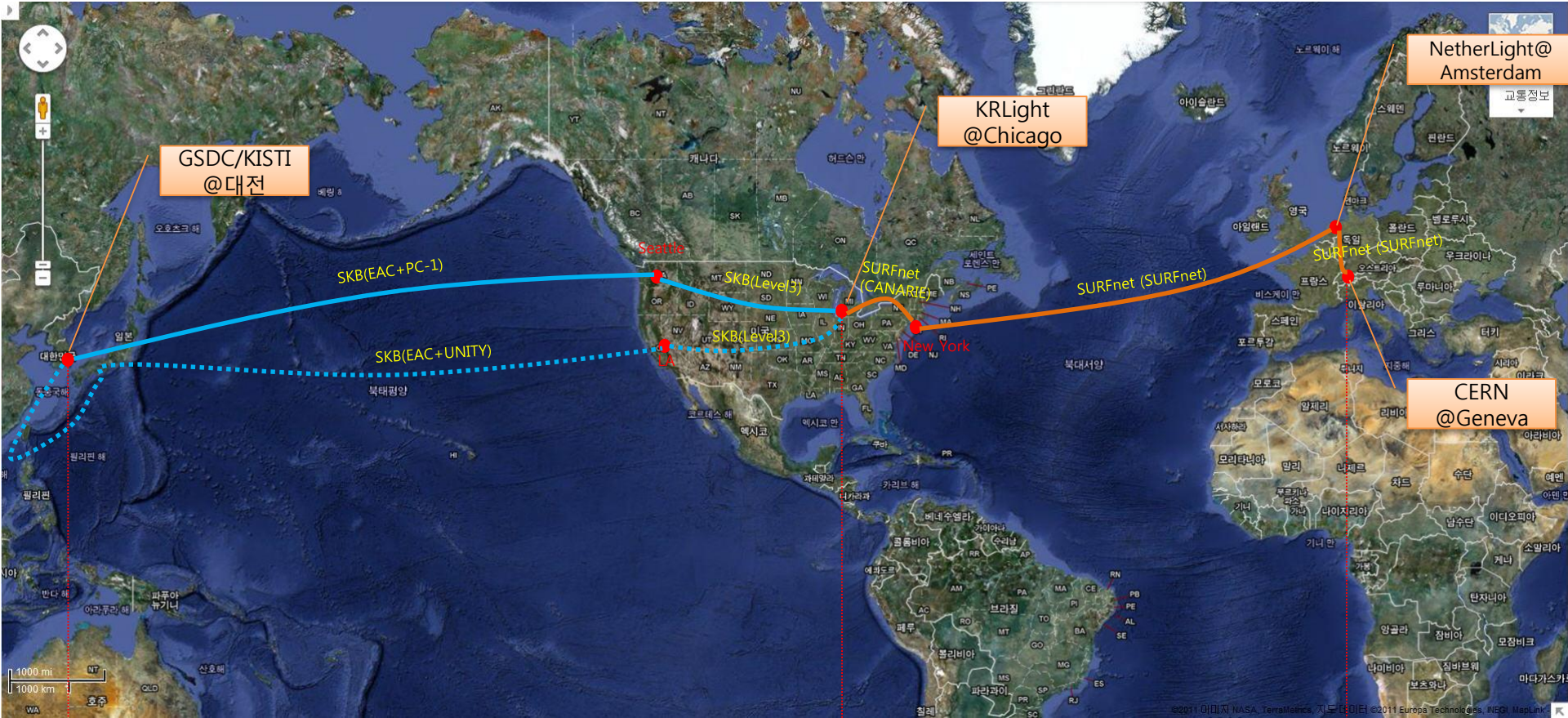
external
area
(computing)





FAST NETWORK?

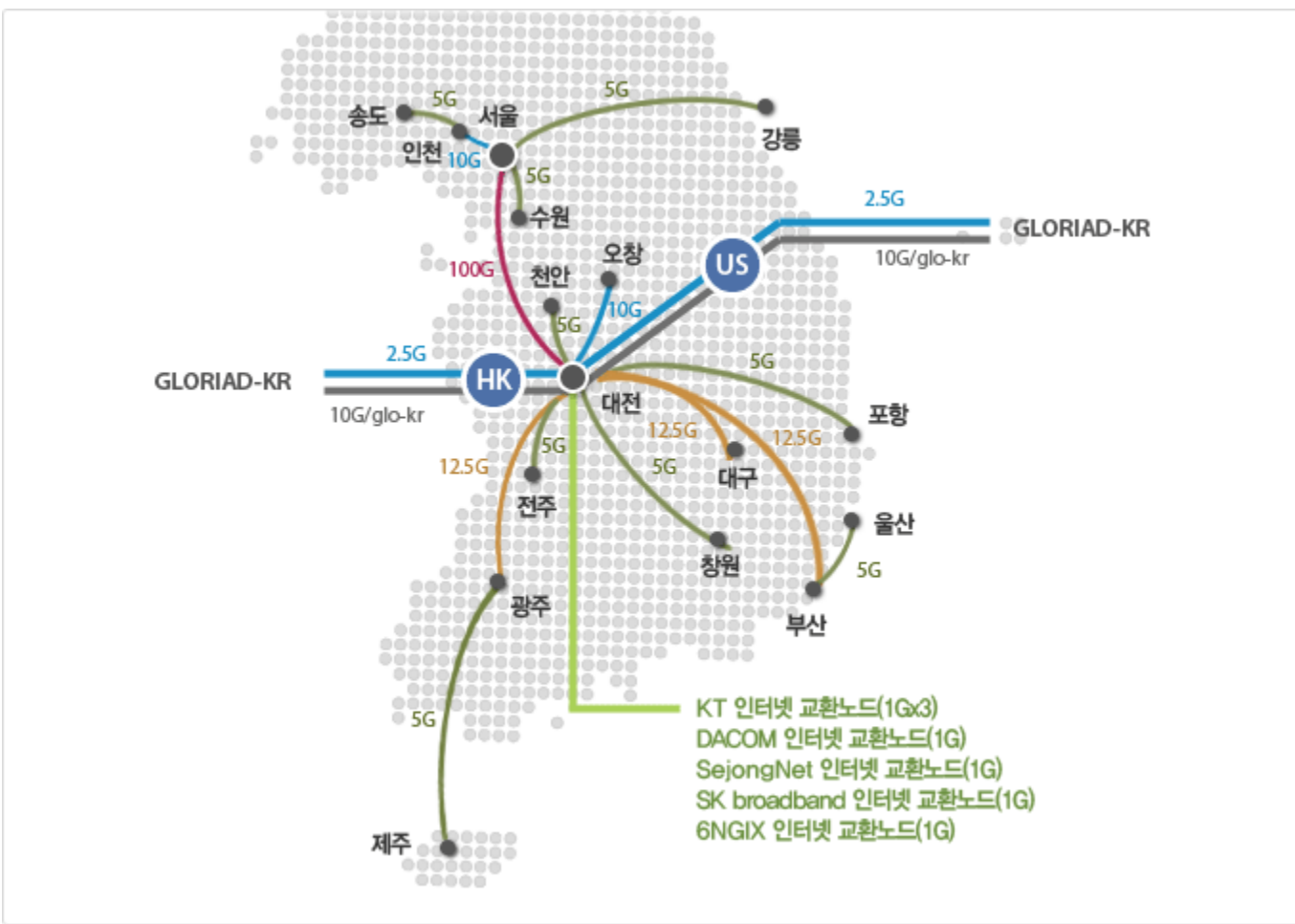
GSDC/KISTI – CERN 10G LHCOPN



SKB

SURFnet

KREONET



GLORIAD

USA-RUSSIA-CHINA-KOREA-NETHERLANDS-CANADA-DENMARK-FINLAND-ICELAND-NORWAY-SWEDEN-INDIA-EGYPT-SINGAPORE



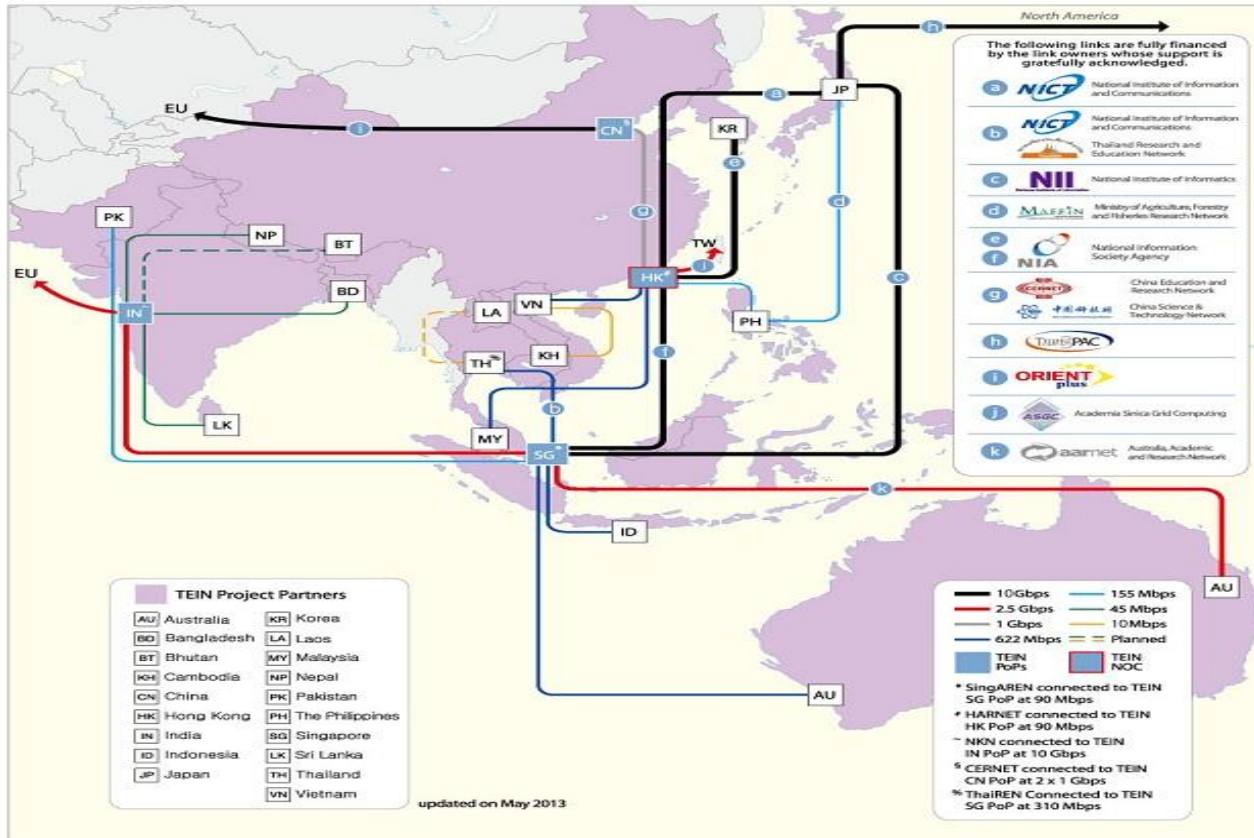
GLORIAD
CURRENT
2013

Global Ring Network for Advanced Applications Development



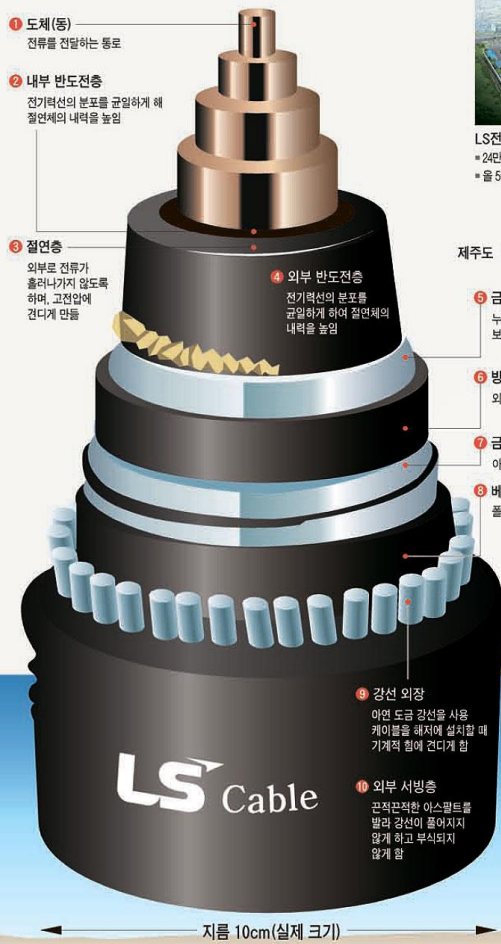
TEIN

Connecting Europe and Asia's research and education communities



Wan connection

해저케이블의 구조



해저케이블의 테크놀로지



- 금속시스(납)**
누설전류를 집지로 흘러보내 케이블을 보호하고 절연체에 물이 안 들어가게 함
- 방식층(폴리에틸렌)**
외부 충격으로부터 코어를 보호하는 역할
- 금속 보강층**
이전 도금 강철 테이프를 같이 코어를 보강함
- 베딩층**
폴리에스터 테이프를 감아줘 케이블 코어를 보호

진도 변환소

- 발전소에서 생산된 교류(AC)를 직류(DC)로 전환 (장거리 송전과 전력 손실을 줄이기 위해)

케이블 설치선
(영국에 하루 임대료 1억원, 8000t)

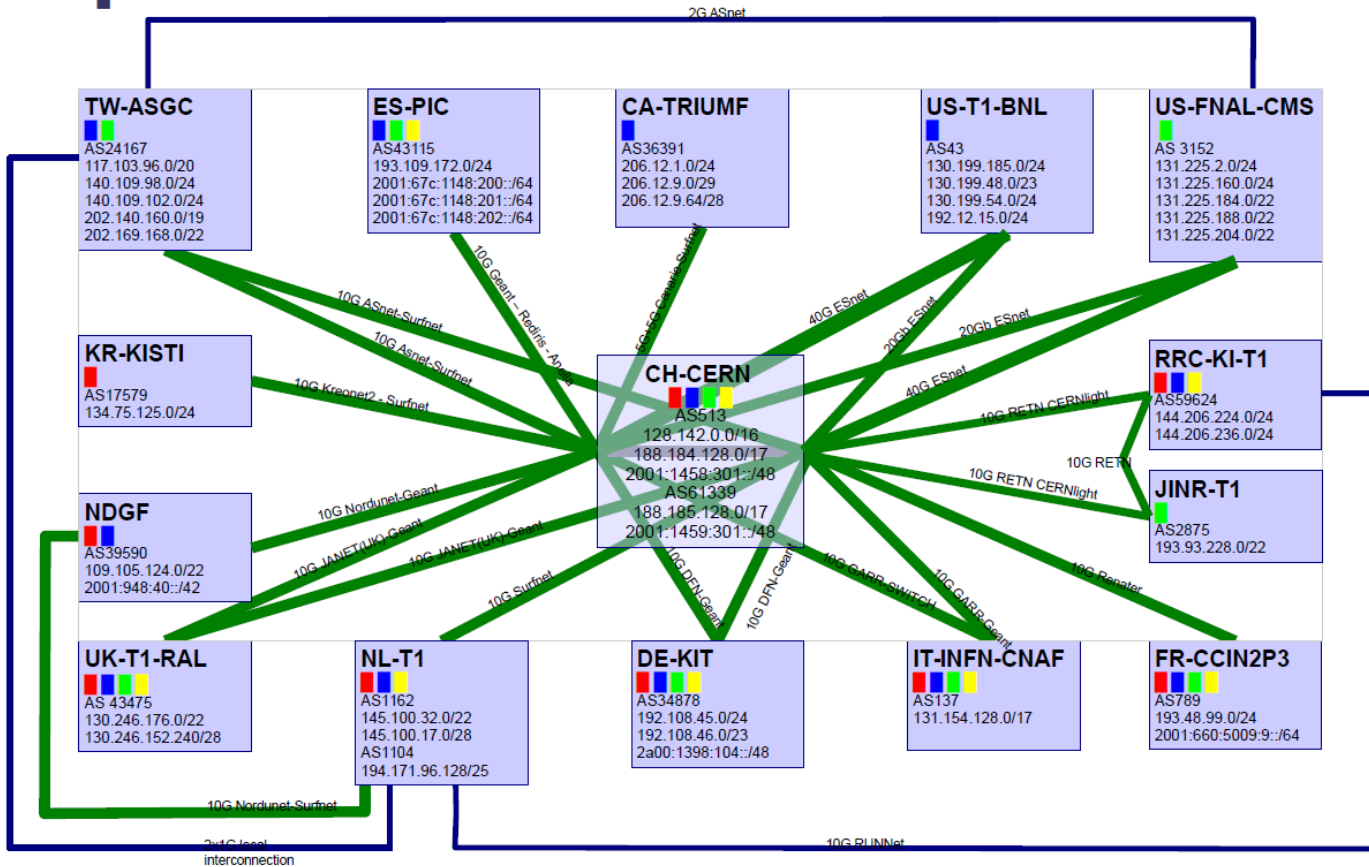


Submarine cable



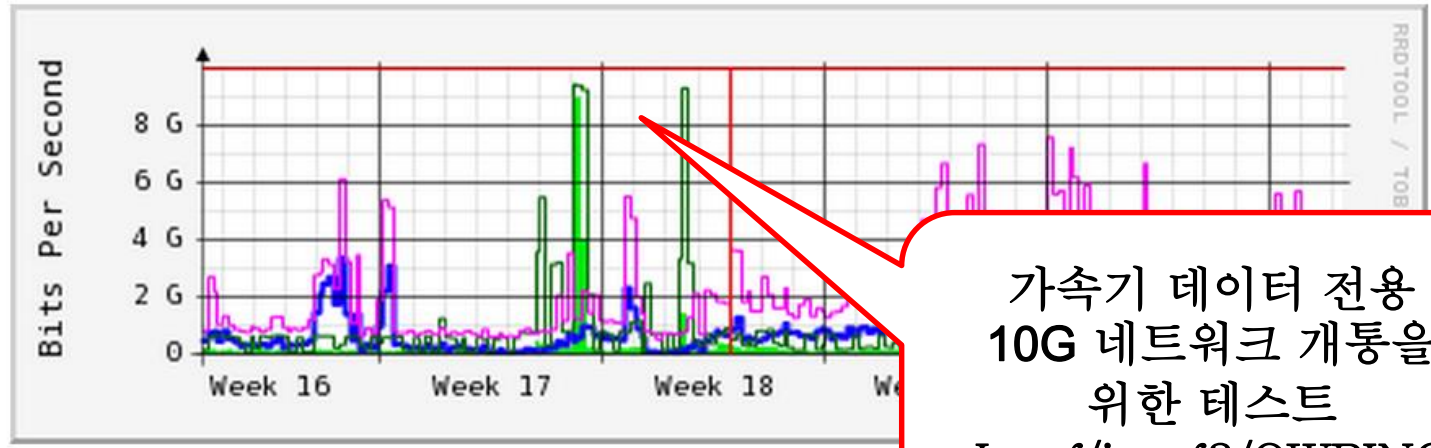
LAN and WAN

Map LHCOPN



LHCOPN data transfer

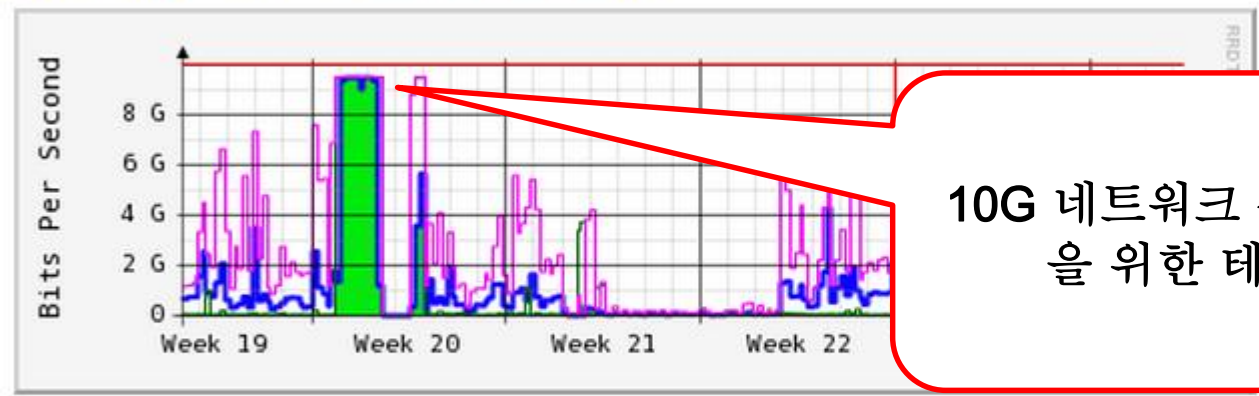
'Monthly' Graph (2 Hour Average)



가속기 데이터 전용
10G 네트워크 개통을
위한 테스트
Iperf/iperf3/OWPING

o/s (0.4%)
o/s (1.4%)

'Monthly' Graph (2 Hour Average)

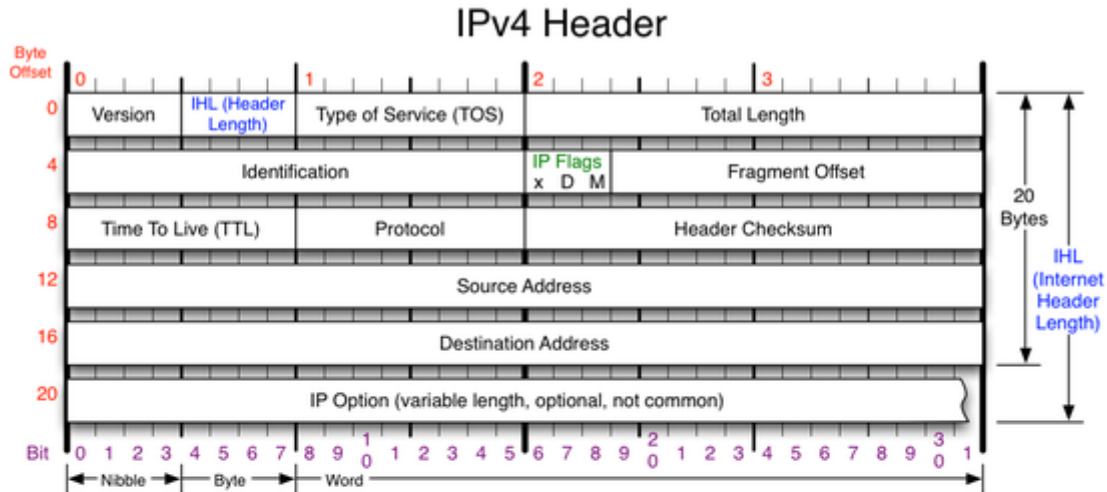


10G 네트워크 성능 측정
을 위한 테스트

Max In: 9479.4 Mb/s (94.8%) Average In: 438.6 Mb/s (4.4%) Current In: 13.4 Mb/s (0.1%)
Max Out: 9479.4 Mb/s (94.8%) Average Out: 931.6 Mb/s (9.3%)

Q & A

IPv4 header

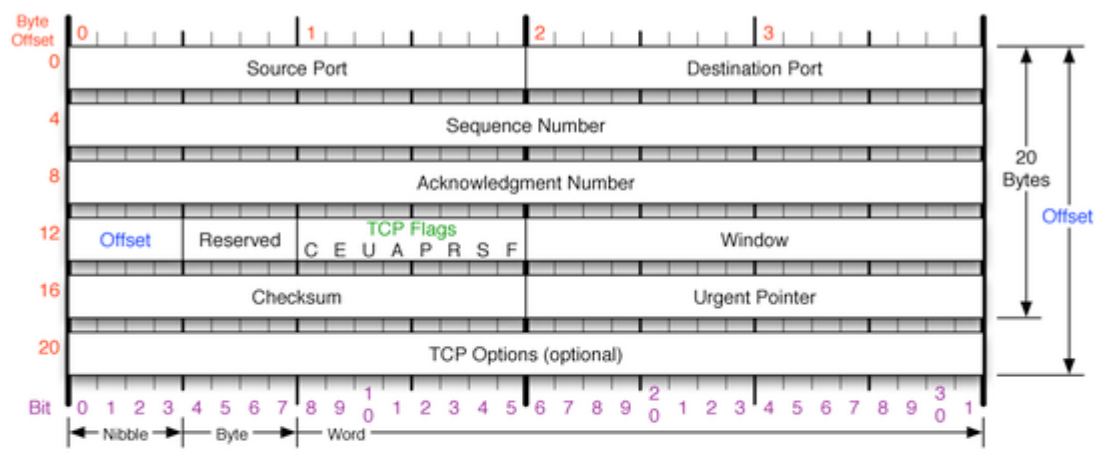


<p>Version</p> <p>Version of IP Protocol. 4 and 6 are valid. This diagram represents version 4 structure only.</p>	<p>Protocol</p> <p>IP Protocol ID. Including (but not limited to):</p> <table border="0"> <tr> <td>1 ICMP</td> <td>17 UDP</td> <td>57 SKIP</td> </tr> <tr> <td>2 IGMP</td> <td>47 GRE</td> <td>88 EIGRP</td> </tr> <tr> <td>6 TCP</td> <td>50 ESP</td> <td>89 OSPF</td> </tr> <tr> <td>9 IGRP</td> <td>51 AH</td> <td>115 L2TP</td> </tr> </table>	1 ICMP	17 UDP	57 SKIP	2 IGMP	47 GRE	88 EIGRP	6 TCP	50 ESP	89 OSPF	9 IGRP	51 AH	115 L2TP	<p>Fragment Offset</p> <p>Fragment offset from start of IP datagram. Measured in 8 byte (2 words, 64 bits) increments. If IP datagram is fragmented, fragment size (Total Length) must be a multiple of 8 bytes.</p>	<p>IP Flags</p> <table border="0"> <tr> <td>x</td> <td>D</td> <td>M</td> </tr> </table> <p>x 0x80 reserved (evil bit) D 0x40 Do Not Fragment M 0x20 More Fragments follow</p>	x	D	M
1 ICMP	17 UDP	57 SKIP																
2 IGMP	47 GRE	88 EIGRP																
6 TCP	50 ESP	89 OSPF																
9 IGRP	51 AH	115 L2TP																
x	D	M																
<p>Header Length</p> <p>Number of 32-bit words in TCP header, minimum value of 5. Multiply by 4 to get byte count.</p>	<p>Total Length</p> <p>Total length of IP datagram, or IP fragment if fragmented. Measured in Bytes.</p>	<p>Header Checksum</p> <p>Checksum of entire IP header</p>	<p>RFC 791</p> <p>Please refer to RFC 791 for the complete Internet Protocol (IP) Specification.</p>															

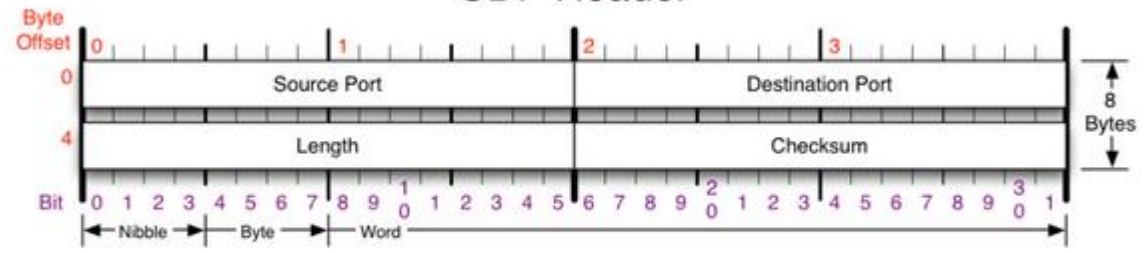
Copyright 2008 - Matt Baxter - mjb@fatpipe.org - www.fatpipe.org/~mjb/Drawings/

TCP vs UDP header

TCP Header



UDP Header



Checksum

Checksum of entire UDP segment and pseudo header (parts of IP header)

RFC 768

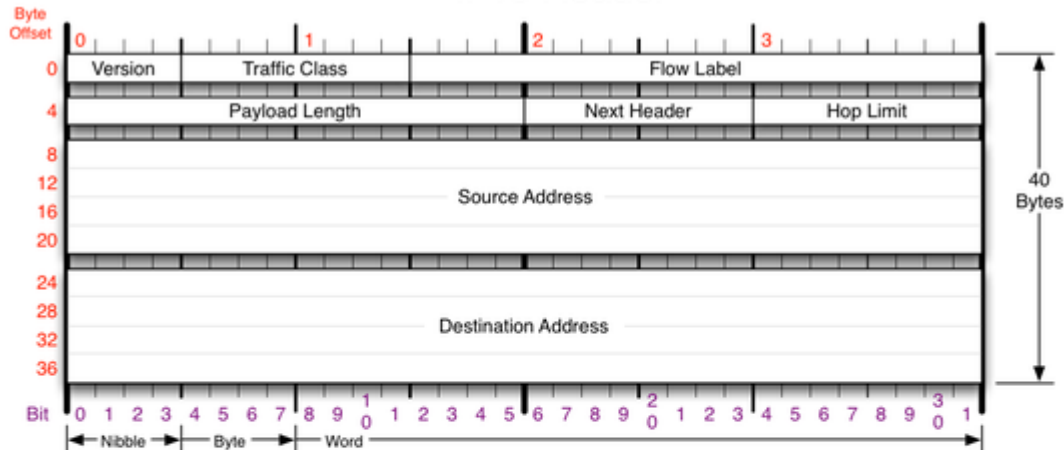
Please refer to RFC 768 for the complete User Datagram Protocol (UDP) Specification.

LLDP and IPv6 Header

LLDP Ethernet frame structure

Preamble	Destination MAC	Source MAC	Ethertype	Chassis ID TLV	Port ID TLV	Time to live TLV	Optional TLVs	End of LLDPDU TLV	Frame check sequence
	01:80:c2:00:00:0e, or 01:80:c2:00:00:03, or 01:80:c2:00:00:00	Station's address	0x88CC	Type=1	Type=2	Type=3	Zero or more complete TLVs	Type=0, Length=0	

IPv6 Header



Version Version of IP Protocol. 4 and 6 are valid. This diagram represents version 6 structure only.	Payload Length 16-bit unsigned integer. Length of the IPv6 payload, i.e., the rest of the packet following this IPv6 header, in octets. Any extension headers are considered part of the payload.	Next Header 8-bit selector. Identifies the type of header immediately following the IPv6 header. Uses the same values as the IPv4 Protocol field.	Hop Limit 8-bit unsigned integer. Decremented by 1 by each node that forwards the packet. The packet is discarded if Hop Limit is decremented to zero.
Traffic Class 8 bit traffic class field.	Source Address 128-bit address of the originator of the packet.	Destination Address 128-bit address of the intended recipient of the packet (possibly not the ultimate recipient, if a Routing header is present).	RFC 2460 Please refer to RFC 2460 for the complete Internet Protocol version 6 (IPv6) Specification.
Flow Label 20 bit flow label.			