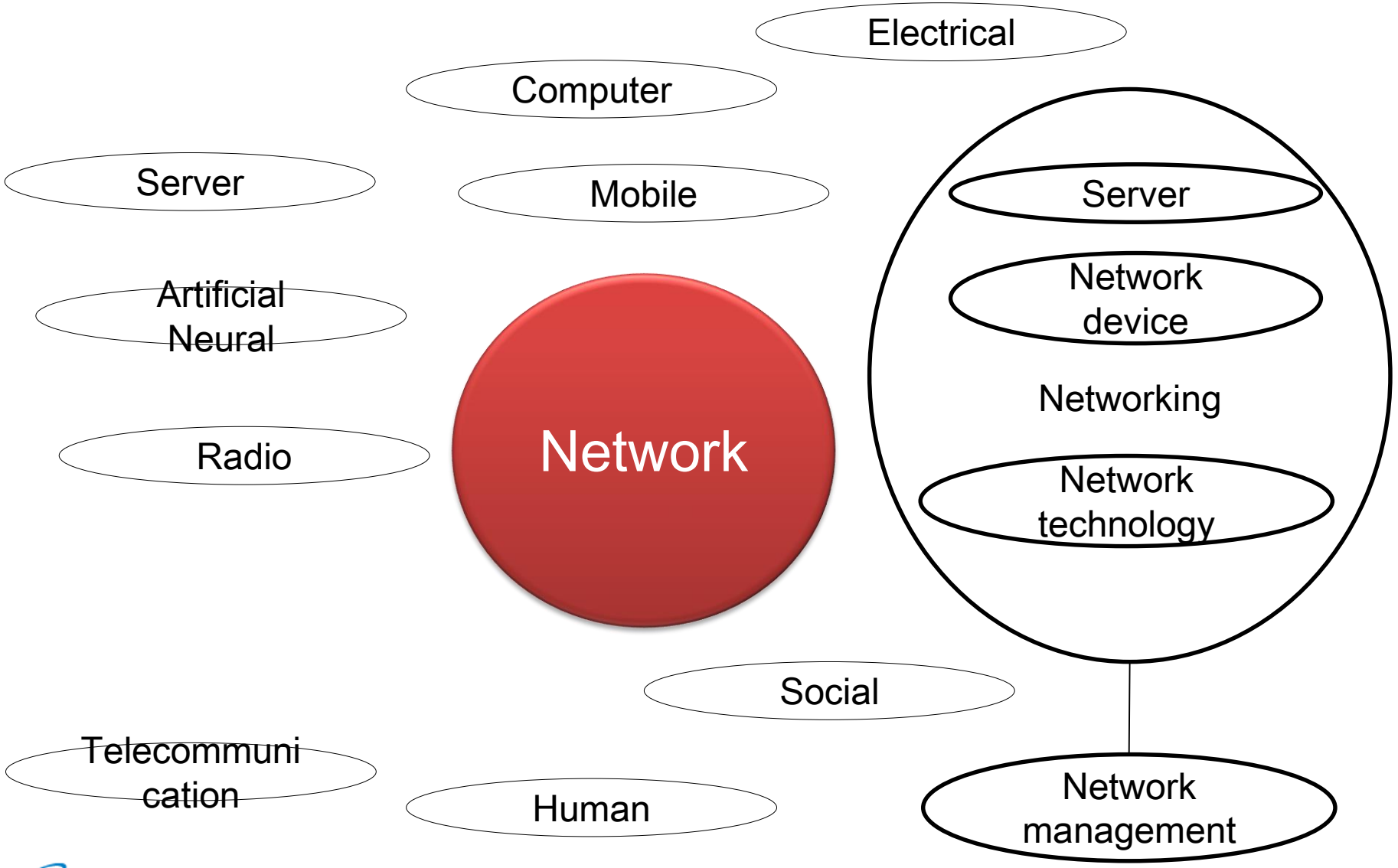
Network Basic

GSDC data computing school Day-3

Jin Kim

2016. 12. 28

Network Category



Agenda

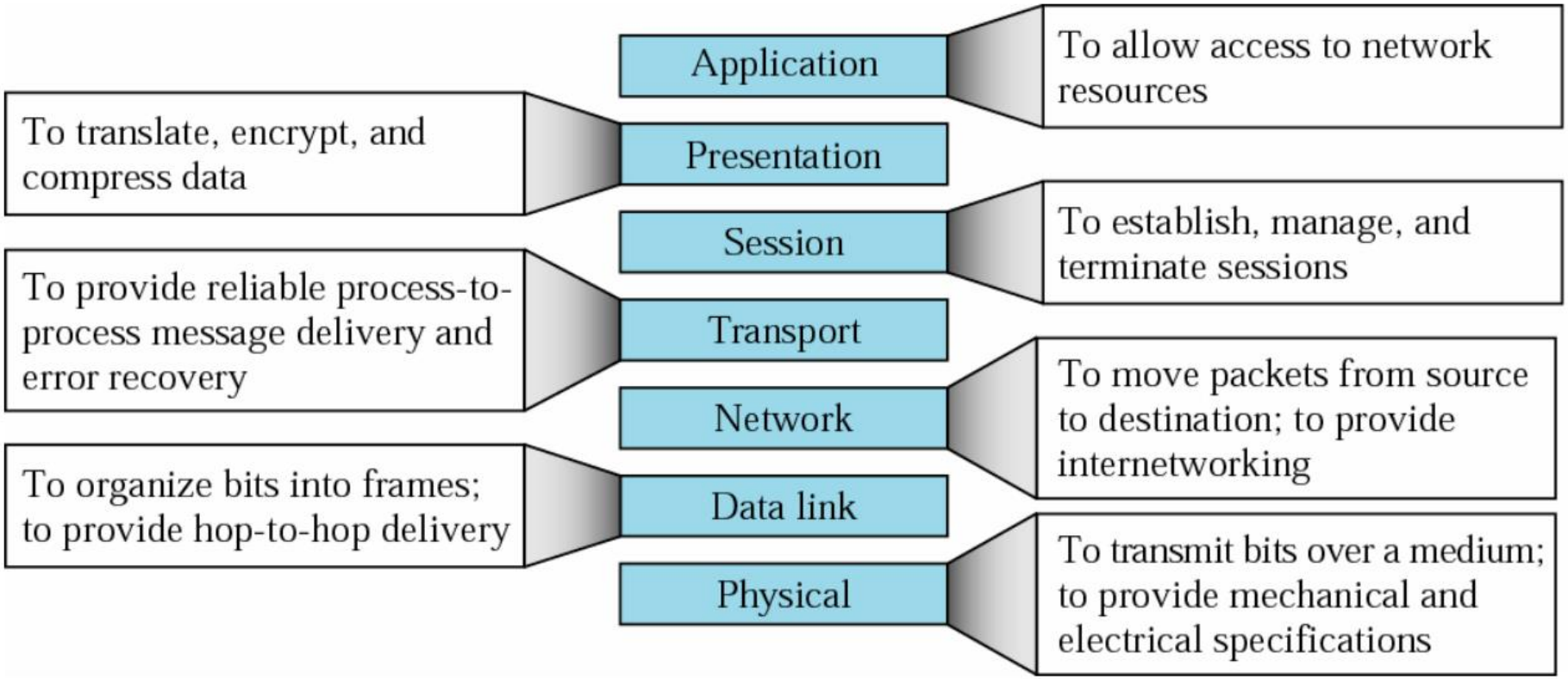
- Network Basic
- Datacenter Network
- Network Technology
- ~~Appendix~~
  - ~~TCP tuning (refer. GEANT)~~

- Network Basic
  - OSI 7 Layer
  - TCP/IP
  - CSMA
  - Congestion Control

ISO-OSI reference model

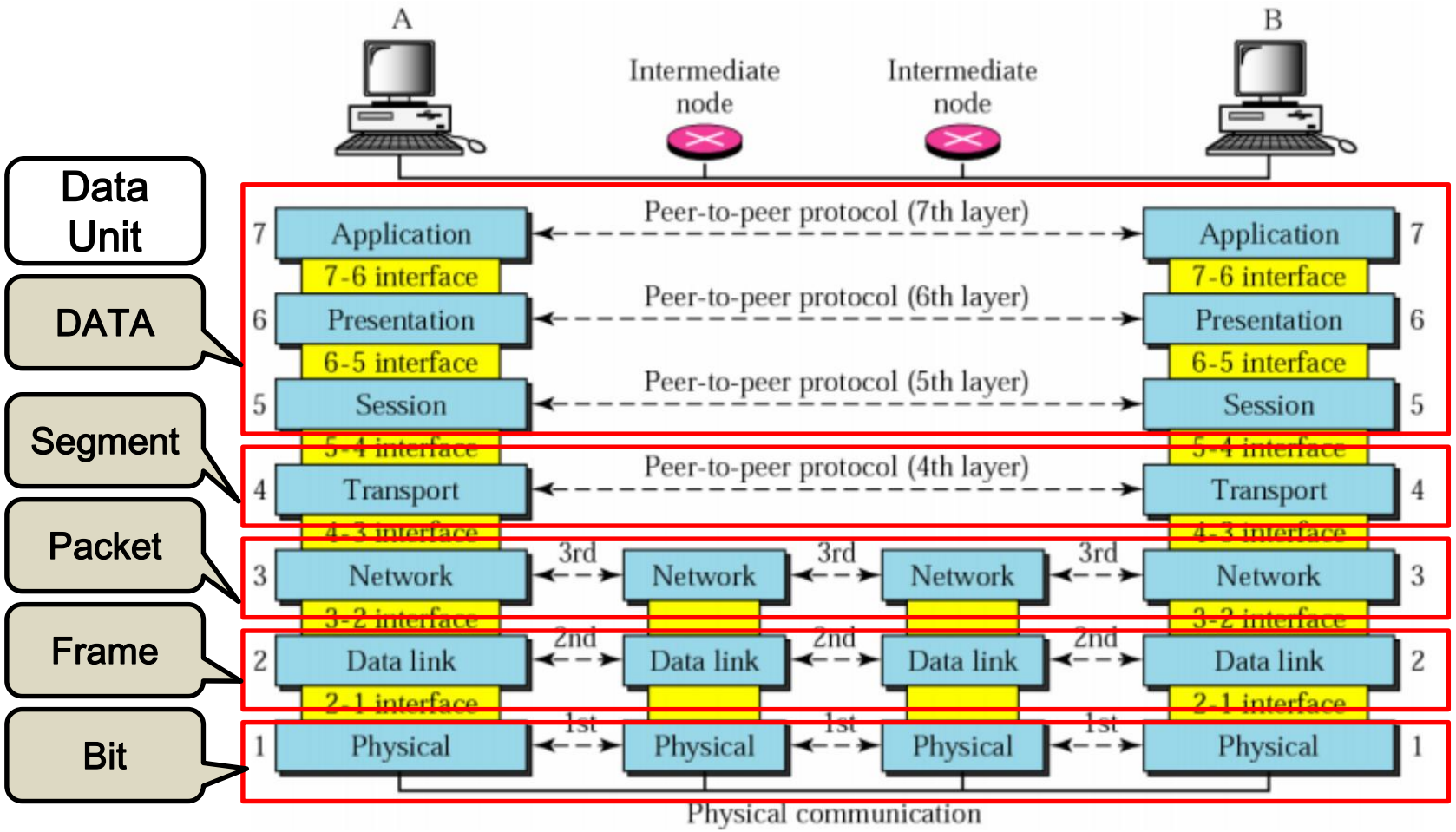
- International Standards Organization
  - Open Systems Interconnection reference model is a framework for connecting computers on a network
- Motivation
  - Reduce the complexity of networking software
  - Support various protocols

OSI 7 Layer(Open Systems Interconnection, 1/3)



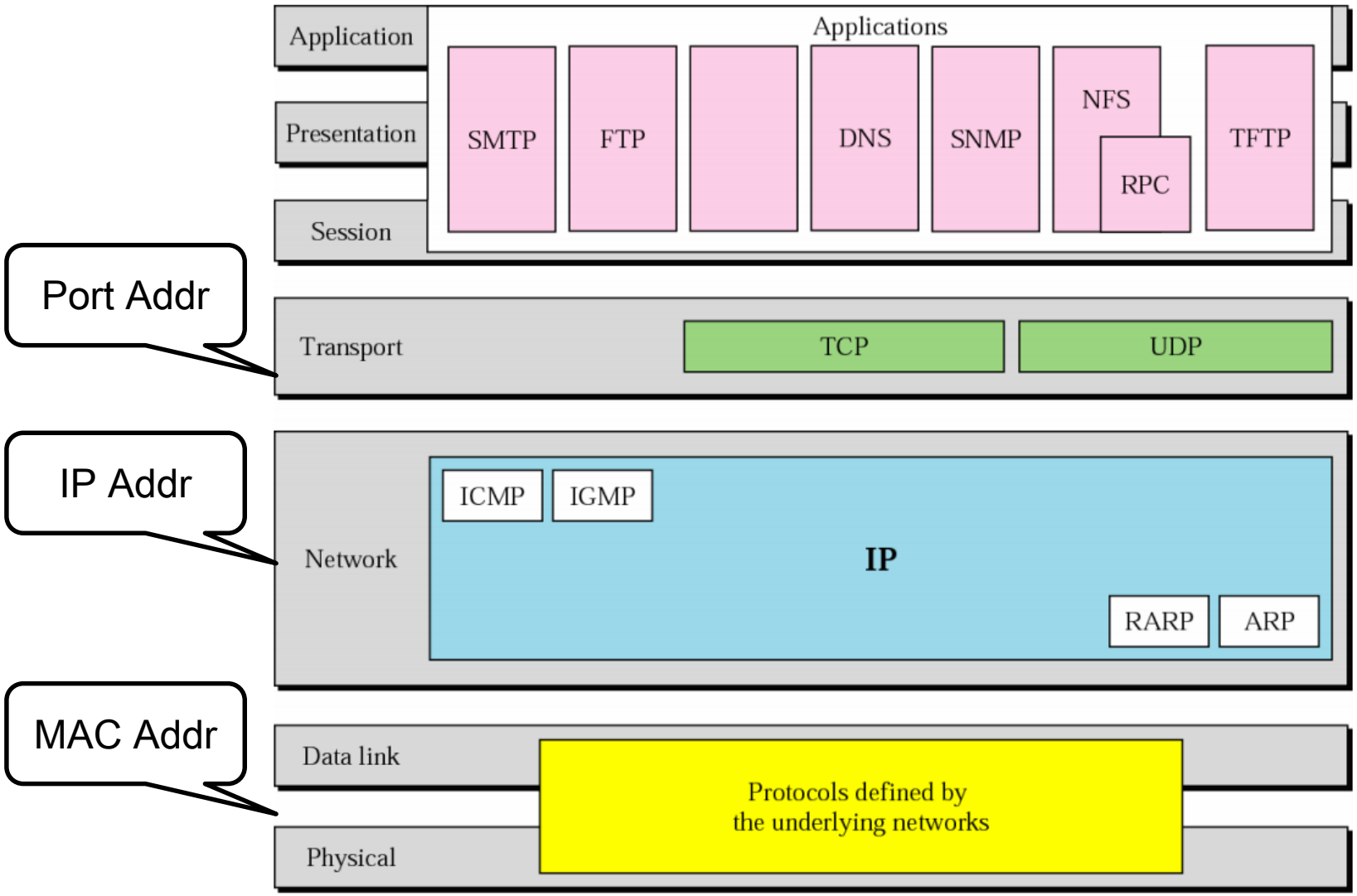
The McGraw-Hill Companies, Inc., 2000

OSI 7 Layer(2/3)



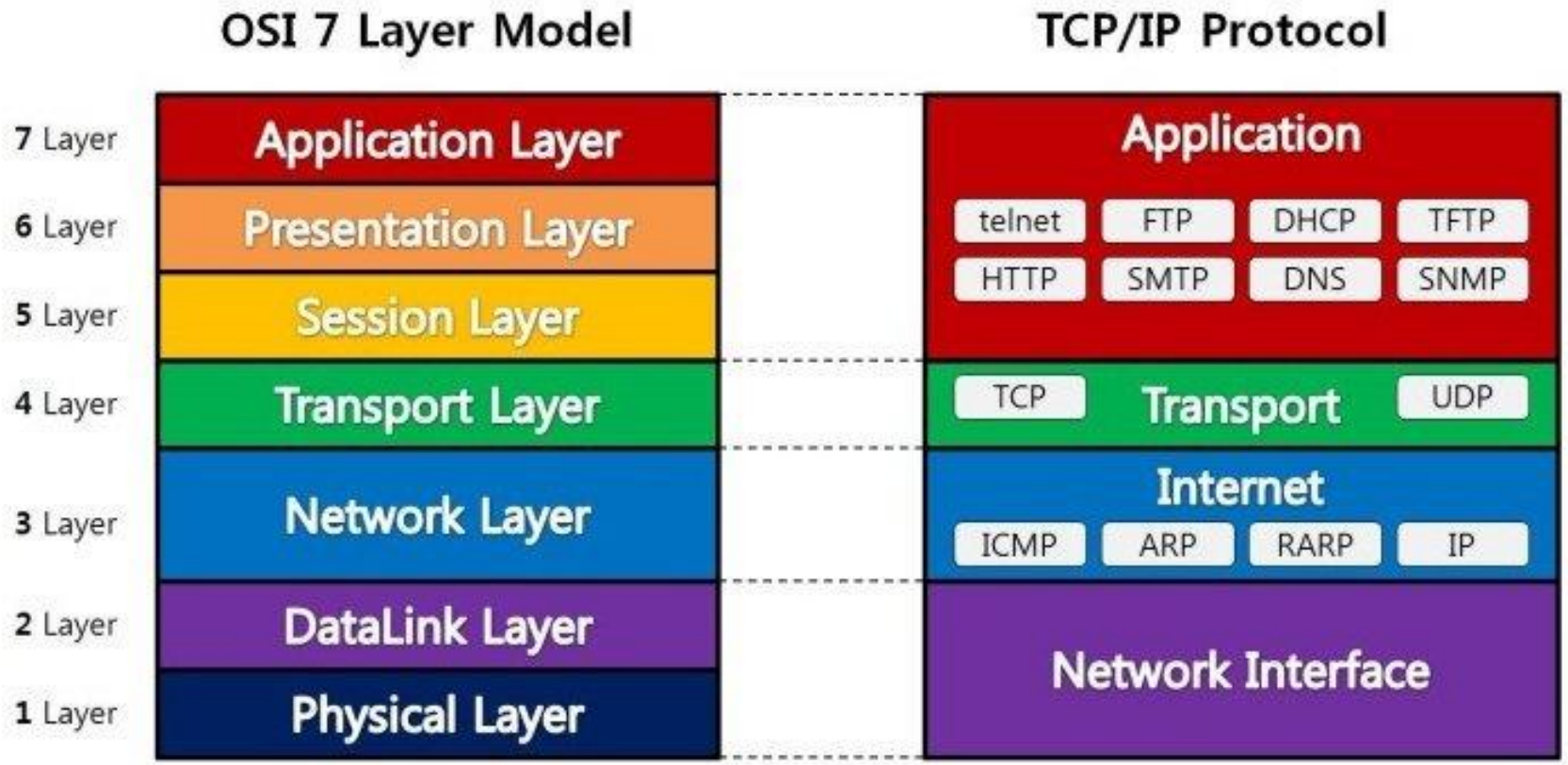
The McGraw-Hill Companies, Inc., 2000

OSI 7 Layer(3/3)

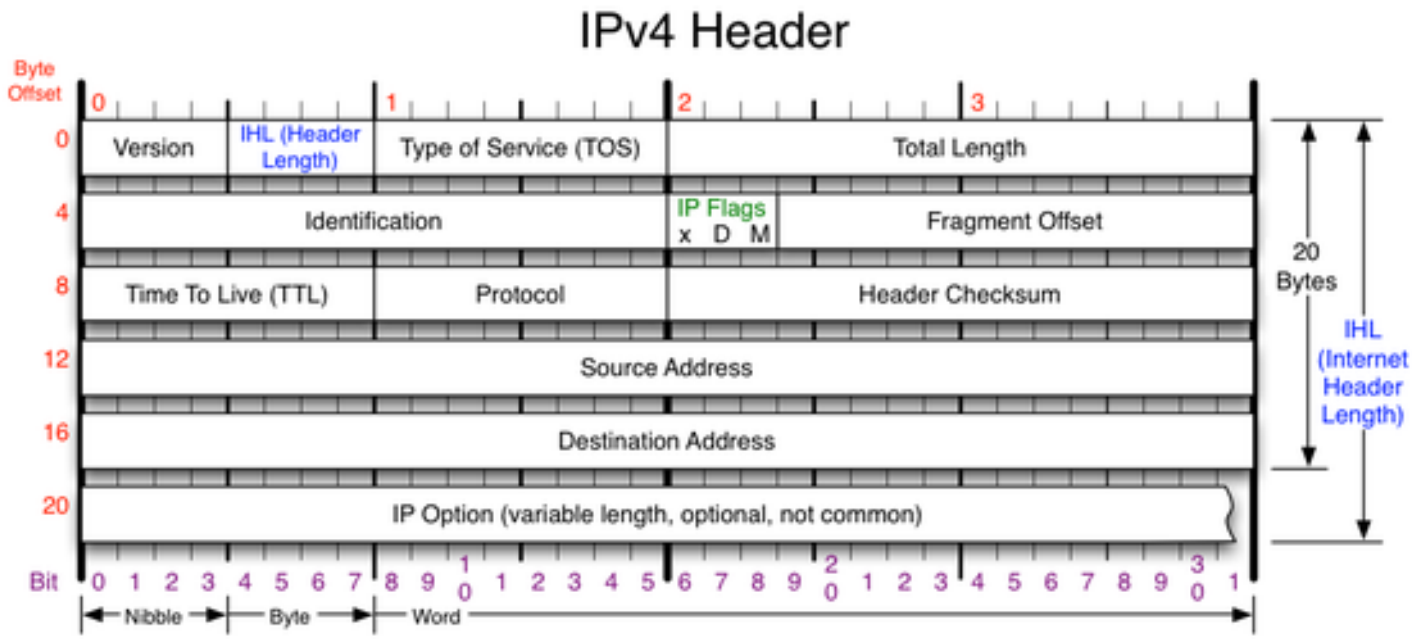


The McGraw-Hill Companies, Inc., 2000

Networking layers



IPv4 header



Version

Version of IP Protocol. 4 and 6 are valid. This diagram represents version 4 structure only.

Header Length

Number of 32-bit words in TCP header, minimum value of 5. Multiply by 4 to get byte count.

Protocol

IP Protocol ID. Including (but not limited to):

1 ICMP	17 UDP	57 SKIP
2 IGMP	47 GRE	88 EIGRP
6 TCP	50 ESP	89 OSPF
9 IGRP	51 AH	115 L2TP

Total Length

Total length of IP datagram, or IP fragment if fragmented. Measured in Bytes.

Fragment Offset

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.

Header Checksum

Checksum of entire IP header

IP Flags

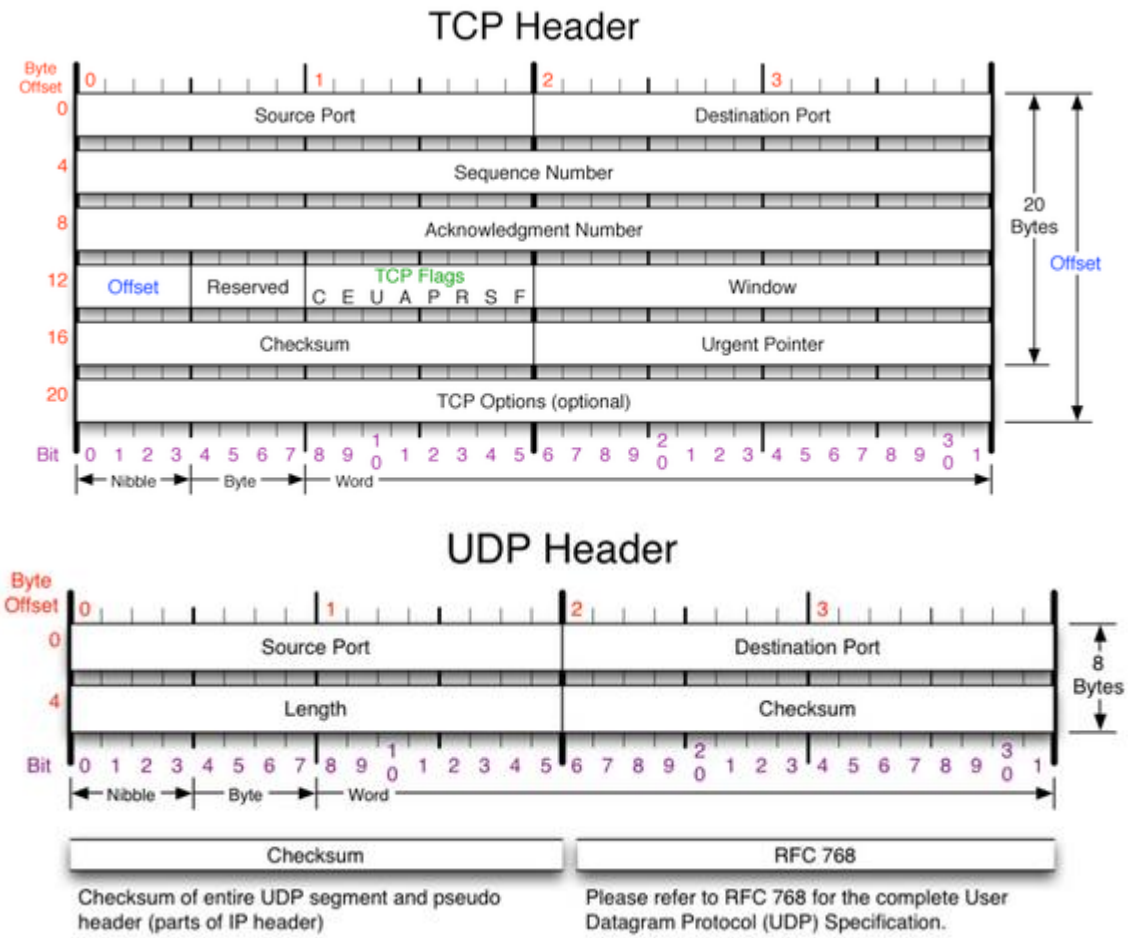
x D M

x 0x80 reserved (evil bit)
D 0x40 Do Not Fragment
M 0x20 More Fragments follow

RFC 791

Please refer to RFC 791 for the complete Internet Protocol (IP) Specification.

TCP vs UDP header



IP Structure

- IPv4
  - 32 bit address space
  - Network ID + Host ID
  - A class: 1 xxx xxxx . xxxx xxxx . xxxx xxxx . xxxx xxxx
  - B class: 10 xx xxxx . xxxx xxxx . xxxx xxxx . xxxx xxxx
  - C class: 110 x xxxx . xxxx xxxx . xxxx xxxx . xxxx xxxx
  - D class: 1110 xxxx . xxxx xxxx . xxxx xxxx . xxxx xxxx
  - Reserved IP: 127.0.0.1, x.x.x.0, x.x.x.1

Multicast
No network ID

8 bit

Binary	1000 0110	0100 1011	0111 1101	1111 1110
Decimal	134	75	125	254

IP Structure

- **Subnet Mask**
  - **To use IP address economically**
  - **CIDR (Classless Internet Domain Routing)**
- **Subnetting**
  - **Divide Host ID part**

- **Supernetting, VLSM (Variable Length Subnet Mask)**
  - **Reduce the size of routing table**

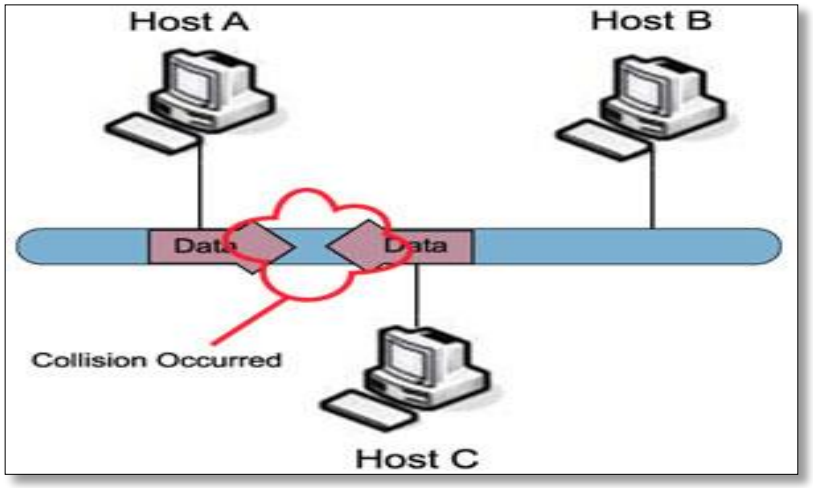
➡ Network device configuration

A class	1111 1111 . 0000 0000 . 0000 0000 . 0000 0000	255.0.0.0	/8
B class	1111 1111 . 1111 1111 . 0000 0000 . 0000 0000	255.255.0.0	/16
C class	1111 1111 . 1111 1111 . 1111 1111 . 0000 0000	255.255.255.0	/24

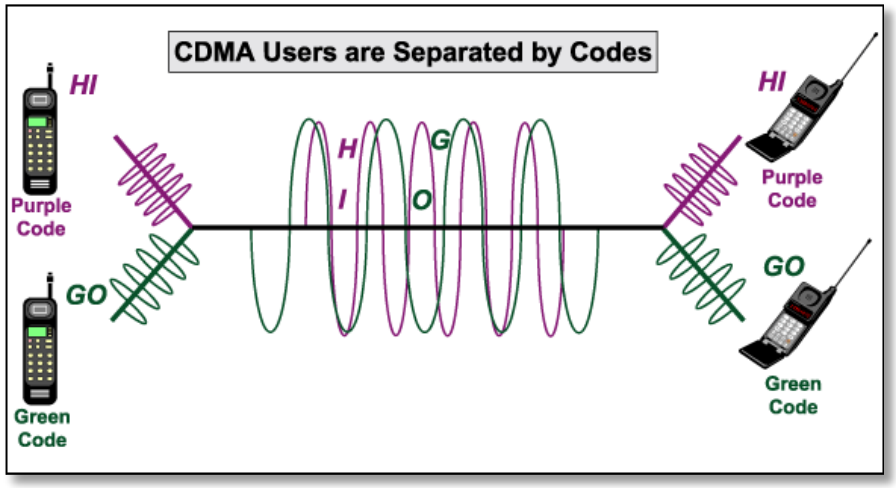
Classful

Classless

CSMA/CD



< CSMA >

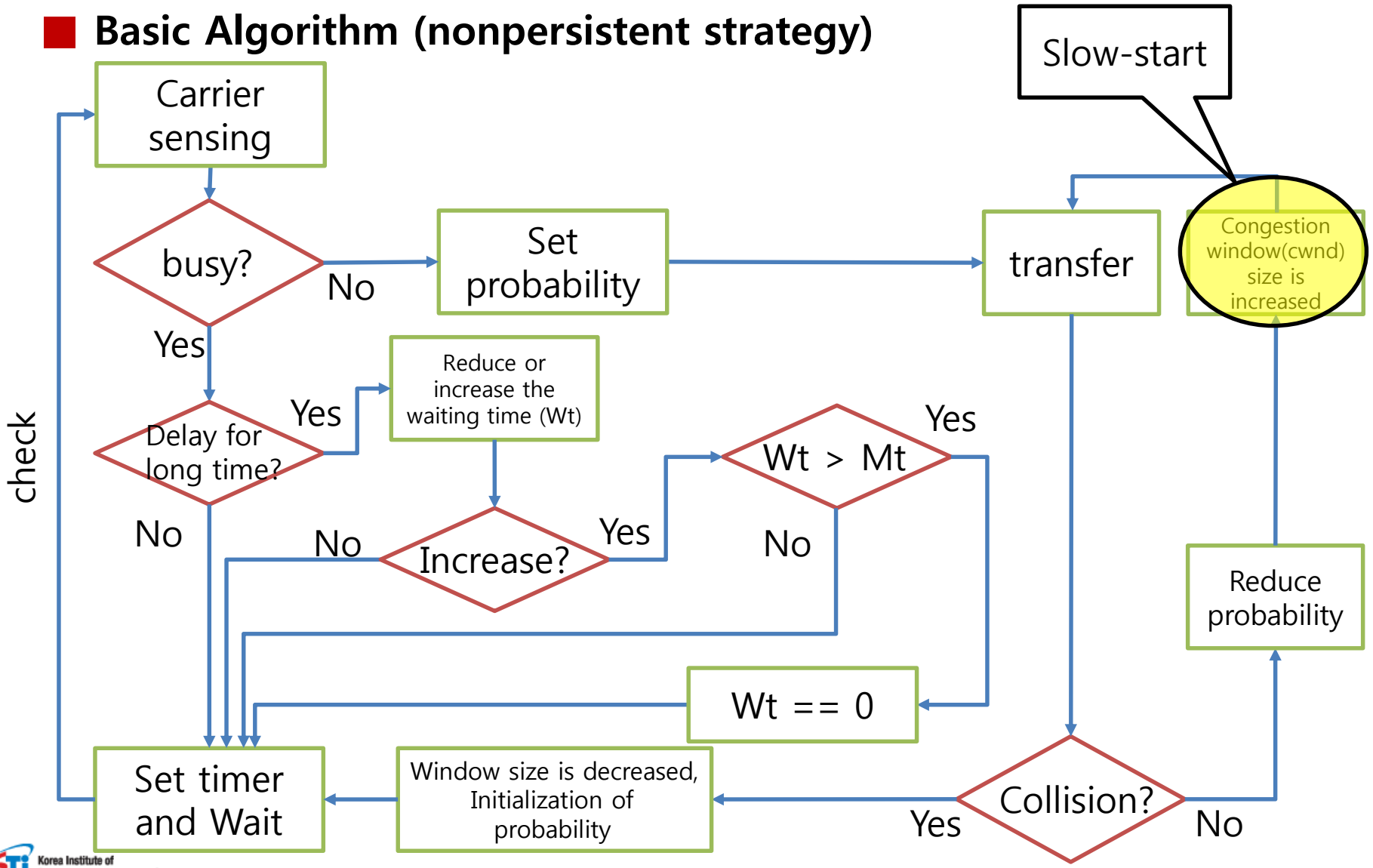


< CDMA >

- **CDMA (Code Division Multiple Access)**
  - **Separate each frequency on a media**
- **CSMA (Carrier Sensing Media Access)**
  - **The way how to use a media**
- **CD (Collision Detection)**
- **CA (Collision Avoidance)**

CSMA

Basic Algorithm (nonpersistent strategy)



TCP congestion control

```
[jkim@admin-ui ipv4]$ cat /proc/sys/net/ipv4/tcp_available_congestion_control
cubic reno
[jkim@admin-ui ipv4]$
[jkim_SYST <.2.el6.x86_64/kernel/net/ipv4*> "admin-ui.sdfarm.kr" 17:12 19-Dec-16
```

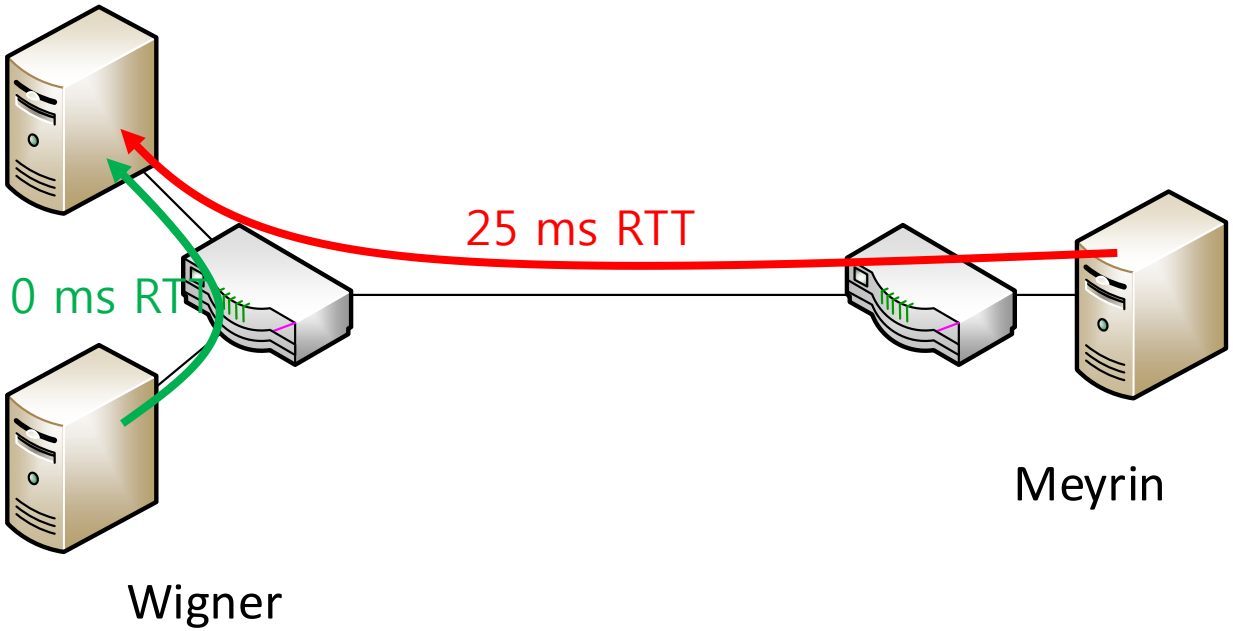
- **Congestion control algorithm**
  - reno
  - BIC
  - CUBIC
  - Scalable
  - Compund TCP

$$B_{max} = \frac{W_{max}}{RTT} = \frac{4MB}{25ms} \approx 1.28\ Gbps$$

```
[jkim@admin-ui ipv4]$ cat /proc/sys/net/ipv4/tcp_rmem
4096    87380   4194304
[jkim@admin-ui ipv4]$
[jkim_SYST <.2.el6.x86_64/kernel/net/ipv4*> "admin-ui.sdfarm.kr" 17:17 19-Dec-16
```

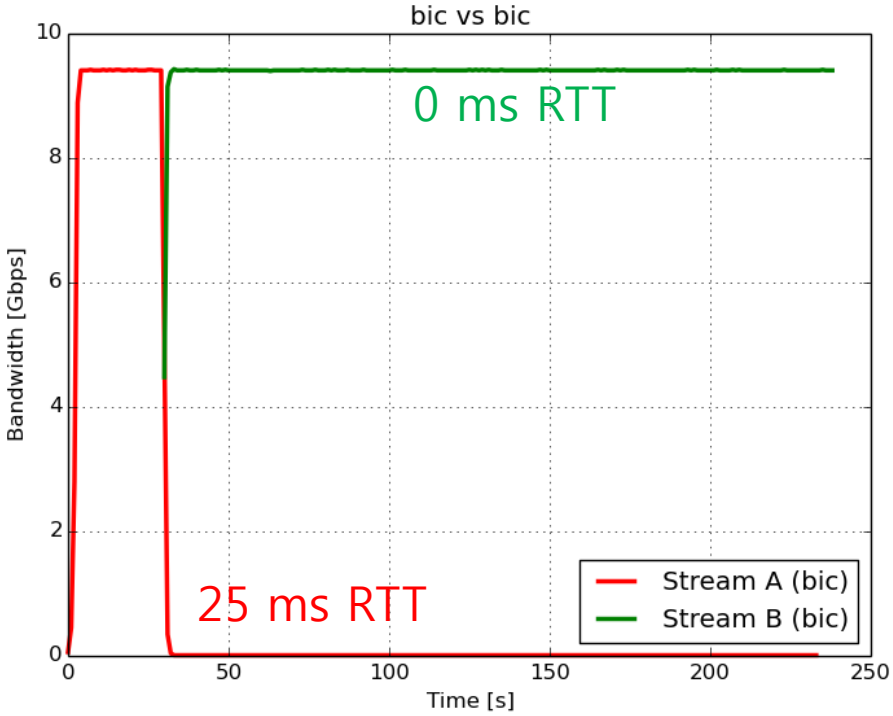
Cubic action (1/2)

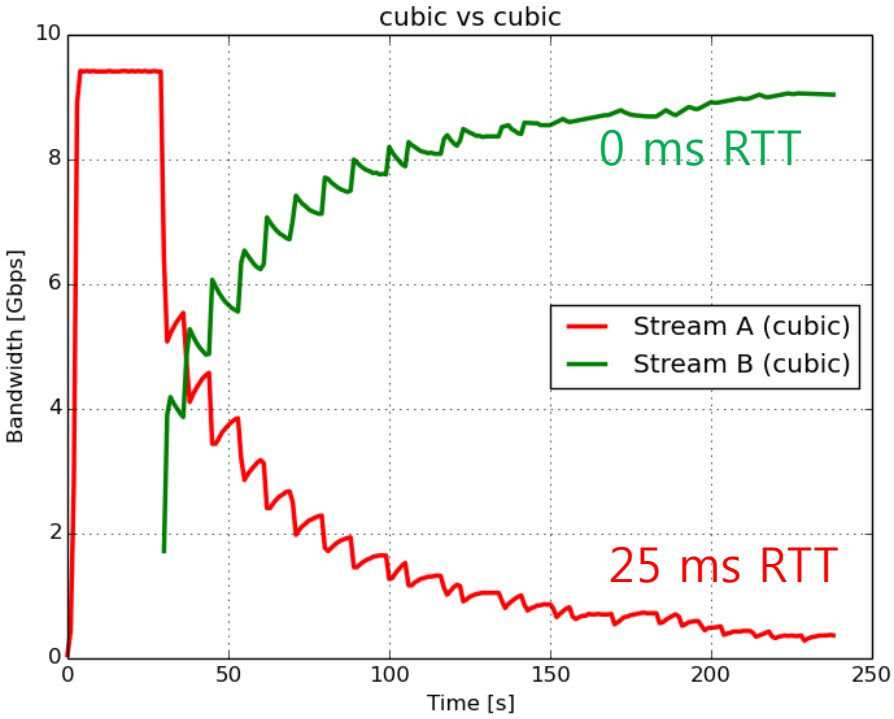
- It has better RTT fairness properties



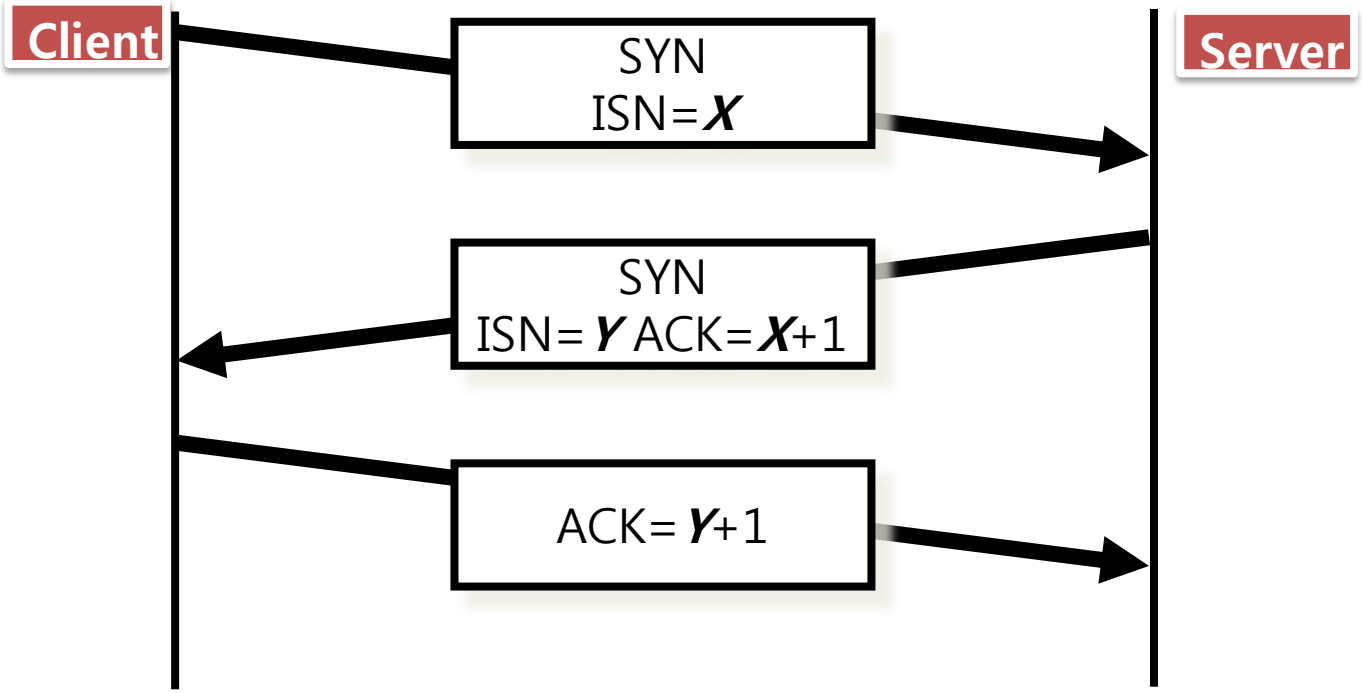
Cubic action (2/2)

RTT fairness test:





TCP 3 way handshaking



- **DDoS (Distributed Denial of Service)**
  - **TCP = Syn flooding**
  - **UDP = bandwidth consumption**
  - **HTTP = web server overload**

WLCG Global Network

- Role of networks in WLCG
  - Computer networks are an essential component of the WLCG
  - Data analysis in LHC will need more network bandwidth between any pair of sites

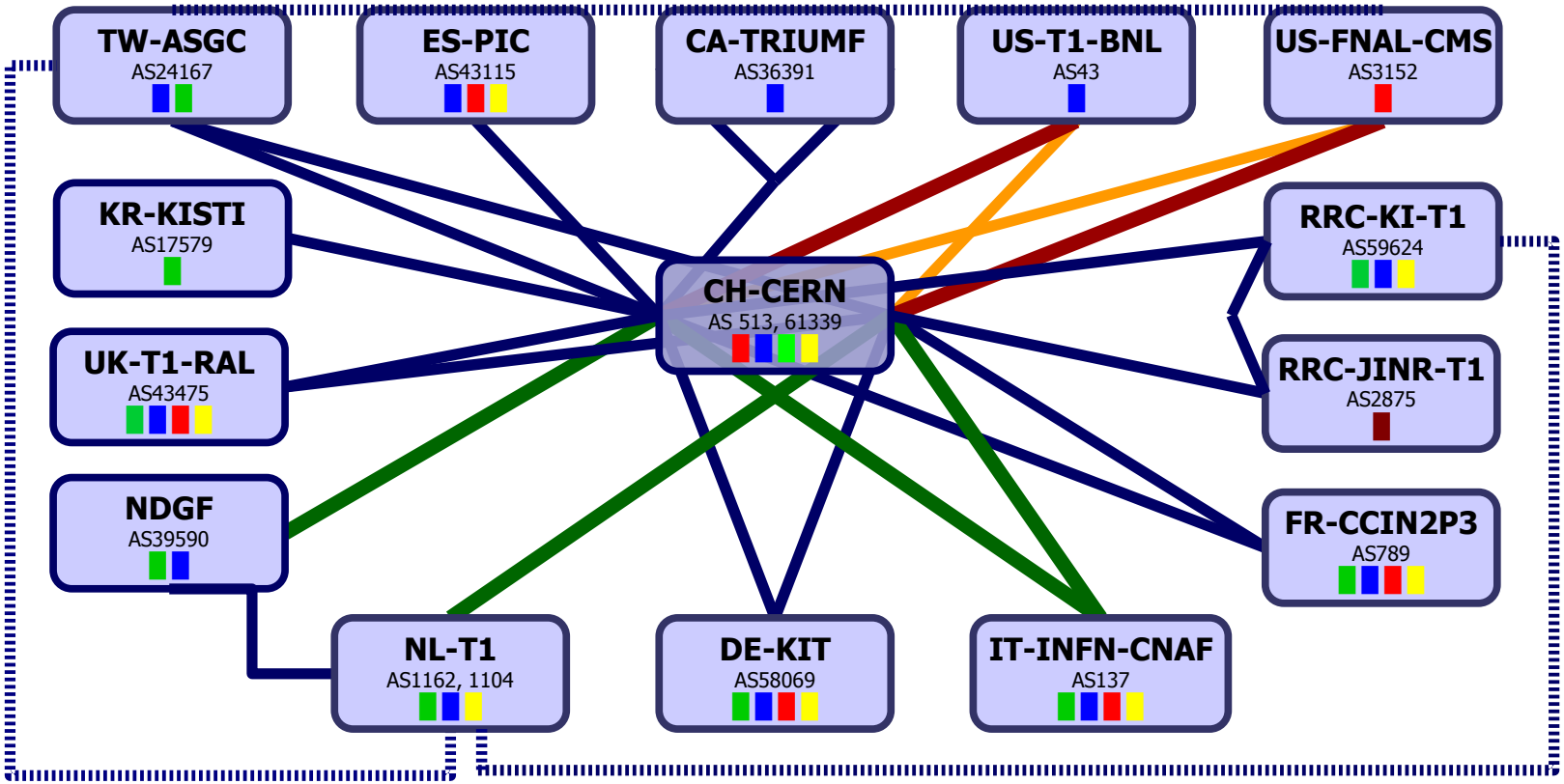
- Two dedicated, private data network have been built for WLCG:
  - LHCOPN (tier0-tier1)
  - LHCONE (tier1-tier2)

LHCOPN and LHCONE

- Private network connecting Tier0 and Tier1s
  - Reserved to LHC data transfers and analysis
  - Single and bundled long distance 10G and 100G ethernet link
  - Star topology
  - BGP routing: communities for traffic engineering, load balancing
  - Security: only declared IP prefixes can exchange traffic

- Open network connecting Tier1s and Tier2s
  - Serving any LHC sites according to their needs and allowing them to grow
  - Sharing the cost and use of expensive resources
  - A collaborative effort among research & education network providers
  - Traffic separation: no clash with other data transfer, resource allocated for and funded by the HEP community
  - Trusted peers: common security policies

LHCOPN 2016. 11. 30.



Korea Institute of Science and Technology Information

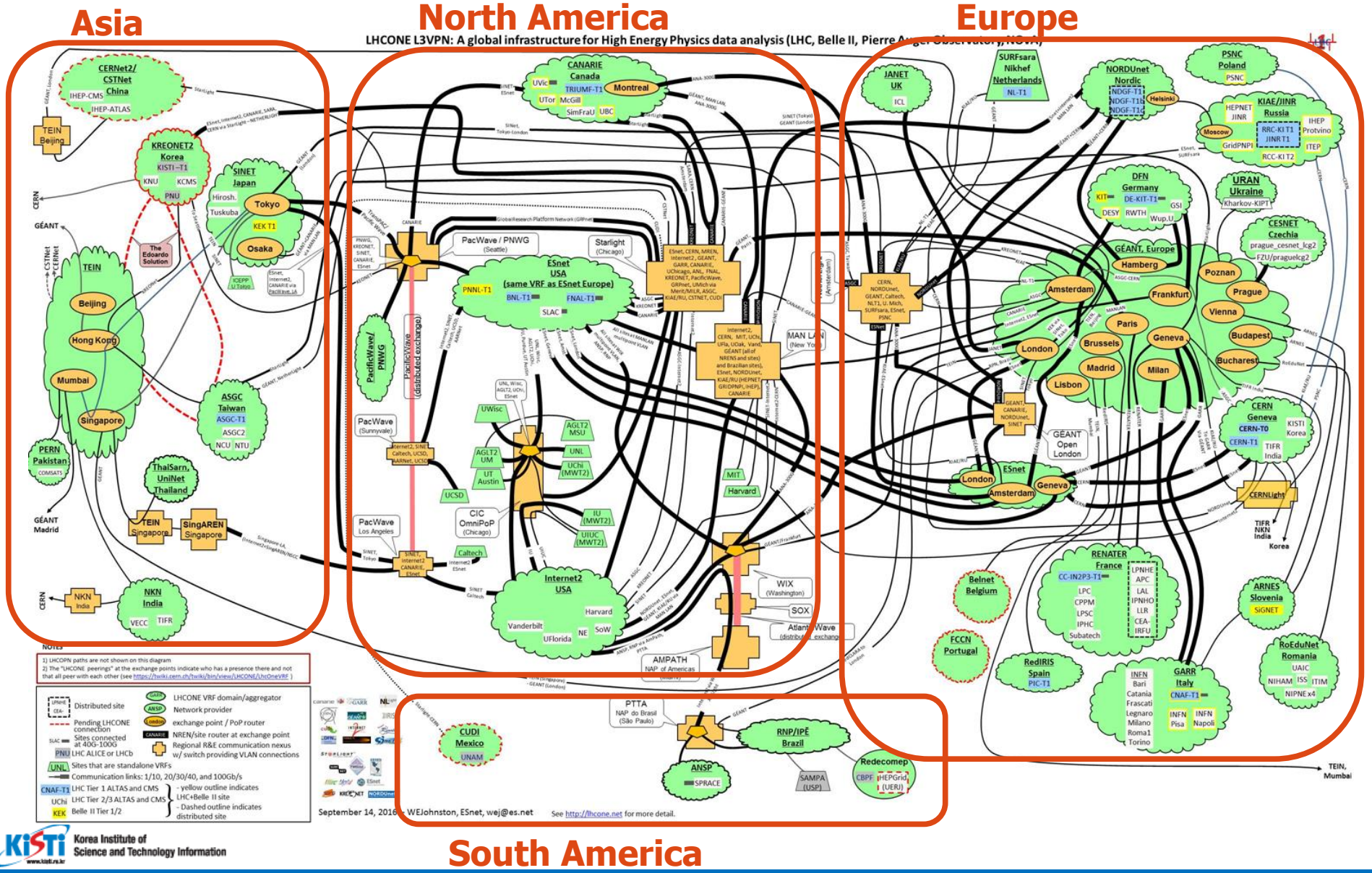
T0-T1 and T1-T1 traffic
T1-T1 traffic only

10Gbps
20Gbps
40Gbps
100Gbps

= Alice = Atlas = CMS = LHCb

edoardo.martelli@cern.ch 20161010

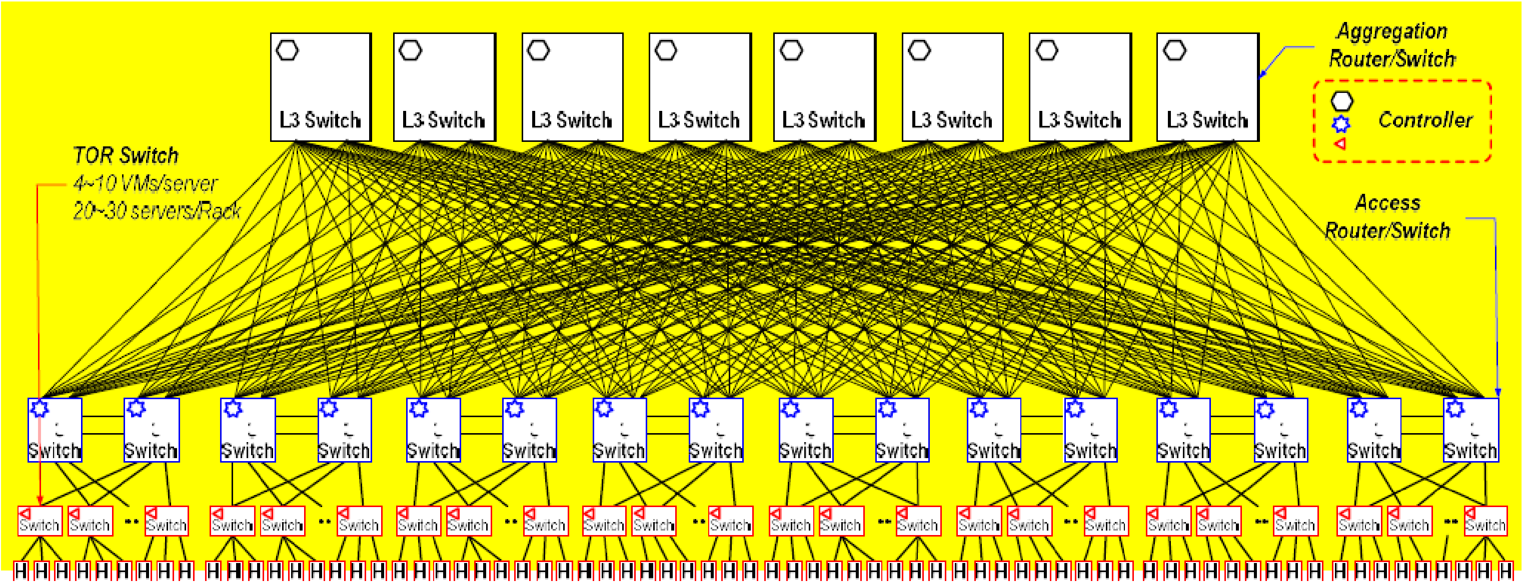
LHCONE 2016. 11. 30



Data center network

Data center

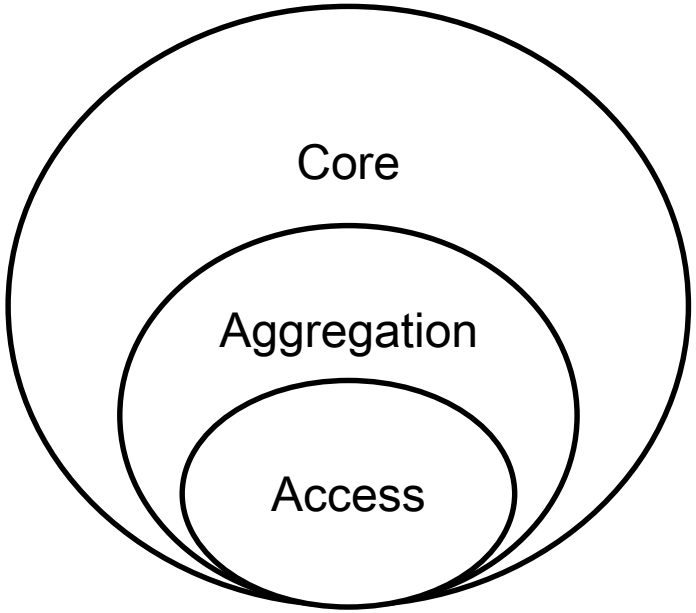
- **Data center is a pool of resources(computational, storage, network) interconnected using a communication network**



Korea Institute of Science and Technology Information

Datacenter (1/4)

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



Datacenter (2/4)

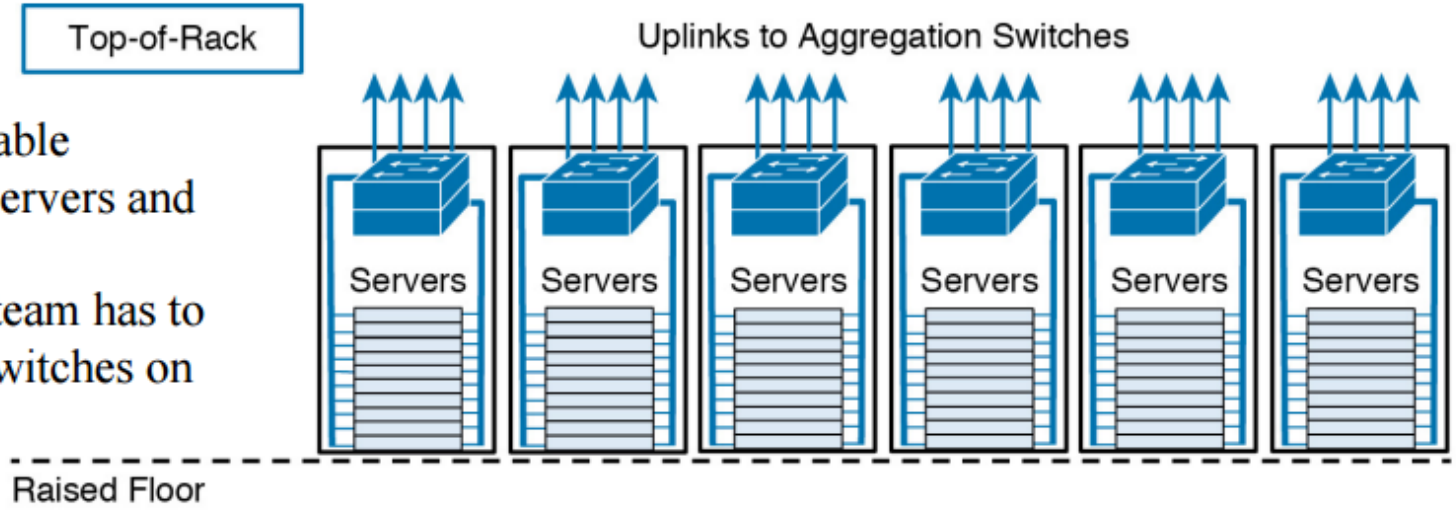
Unstructured cabling

Structured cabling

Datacenter (SW location, 3/4)

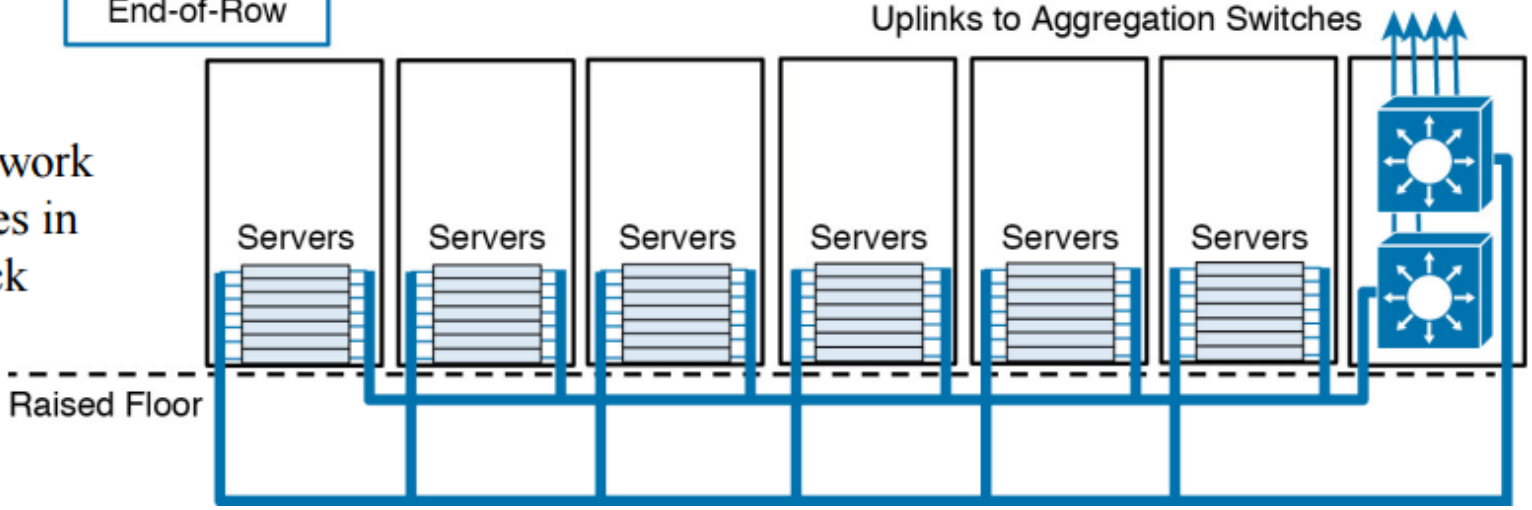
Top-of-Rack

Smaller cable between servers and switches
Network team has to manage switches on all racks



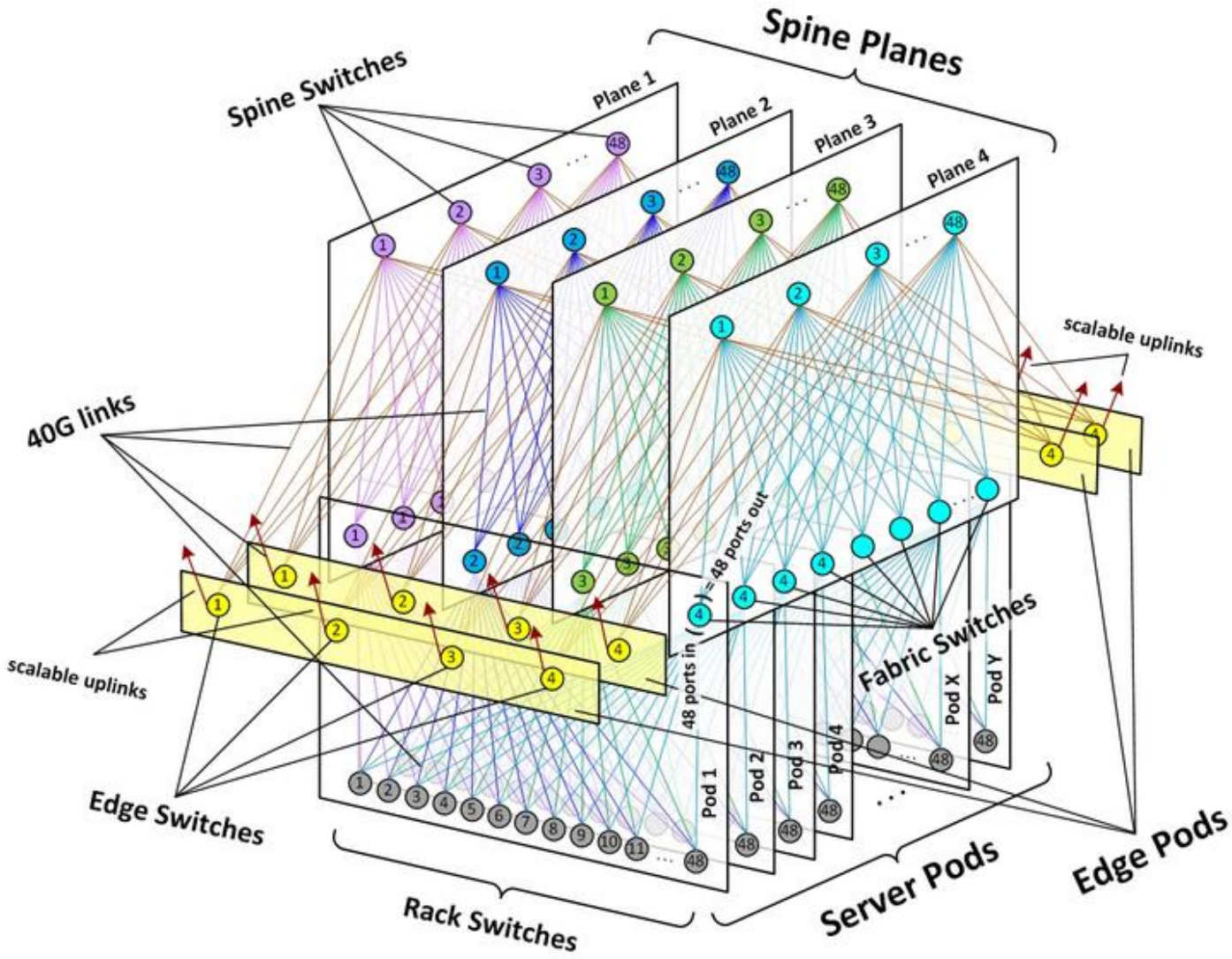
End-of-Row

All network switches in one rack



Datacenter (4/4)

Schematic of Facebook data center fabric network topology



Storage : NAS, SAN

- NAS (Network Attacked Storage)
  - File sharing device based on IP connection
  - Data transfer: TCP/IP
  - Remote file service: SMB(CIFS), NFS
  - Data share : NFS, SMB, FTP
  - Use
    - Server and storage integration
    - Heterogeneous environment for file access
    - Easy to management
    - Extentionable
    - Data protection and security

- SAN (Storage Area Network)
  - Specialized, dedicated high speed network joining servers and storage, including disks, disk arrays, tapes, etc.
  - High capacity, high availability, high scalability, ease of configuration, ease of reconfiguration
  - Fiber channel is the de facto SAN networking architecture, although other network standards could be used

- Fibre channel
  - Is well established in the open systems environment as the underlining architecture of the SAN

More..

- Channel and network
- High speed, low latency
- Topology
  - Point-to-point
  - FC-AL (arbitrated loop)
  - Switched fabric

	channel	network
relation	Master-slave	host-host
throughput	high	low
Processing load	small	high
distance	short	long

Network technology

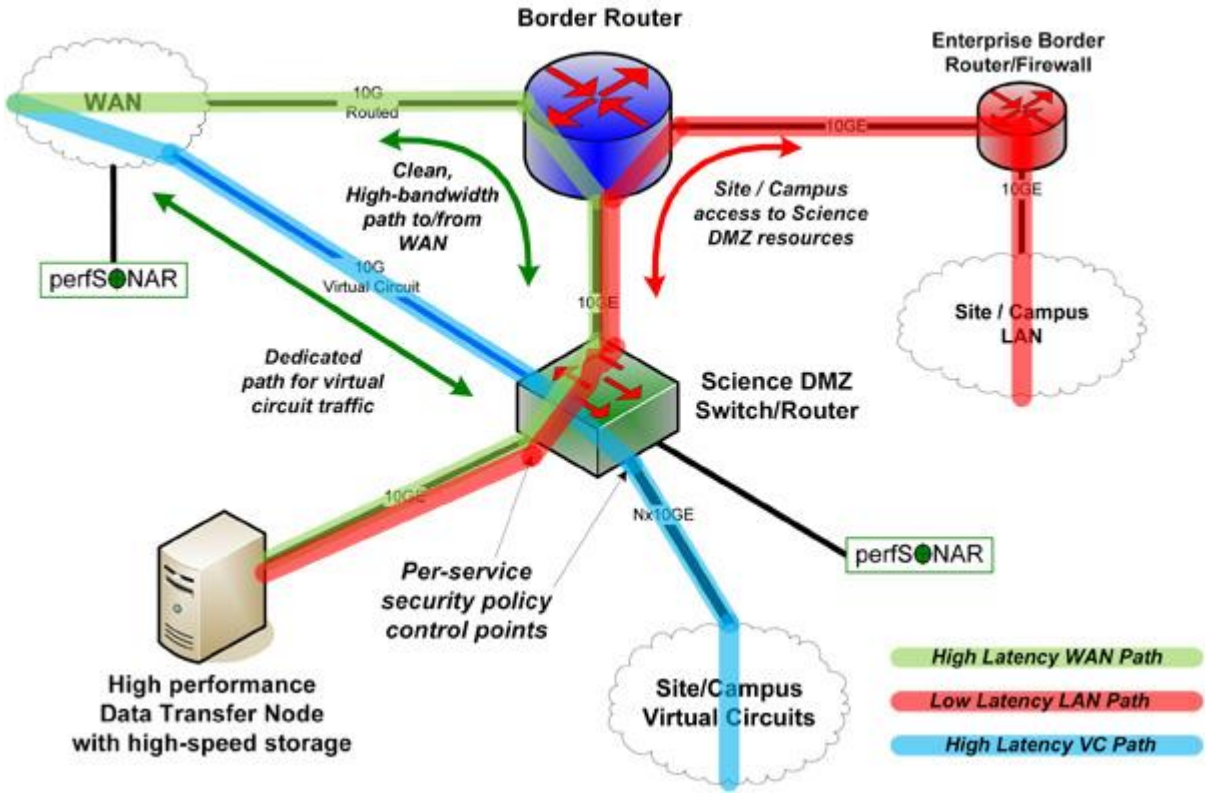
- Network Technology
  - Science DMZ (refer. Esnet)
  - SDN/NFV
  - Bluetooth, WIFI, 3/4/5 G network
  - Long Range network

Science DMZ

- **Background**
  - **The data mobility performance requirements for data intensive science are beyond what can typically be achieved using traditional methods**
    - **Default host configurations (TCP, FS, NICs)**
    - **Converged network architectures designed for commodity traffic**
    - **Conventional security tools and policies**
    - **Legacy data transfer tools (e.g. SCP)**
    - **Wait-for-trouble-ticket operational models for network performance**
  - **The science DMZ model describes a performance-based approach**
    - **Dedicated infrastructure for wide-area data transfer**
      - Well-configured data transfer hosts with modern tools
      - Capable network devices
      - high-performance data path which does not traverse commodity LAN
    - **Proactive operational models that enable performance**
      - Well-deployed test and measurement tools (perfSONAR)
      - Periodic testing to locate issues instead of waiting for users to complain
    - **Security posture well-matched to high-performance science applications**

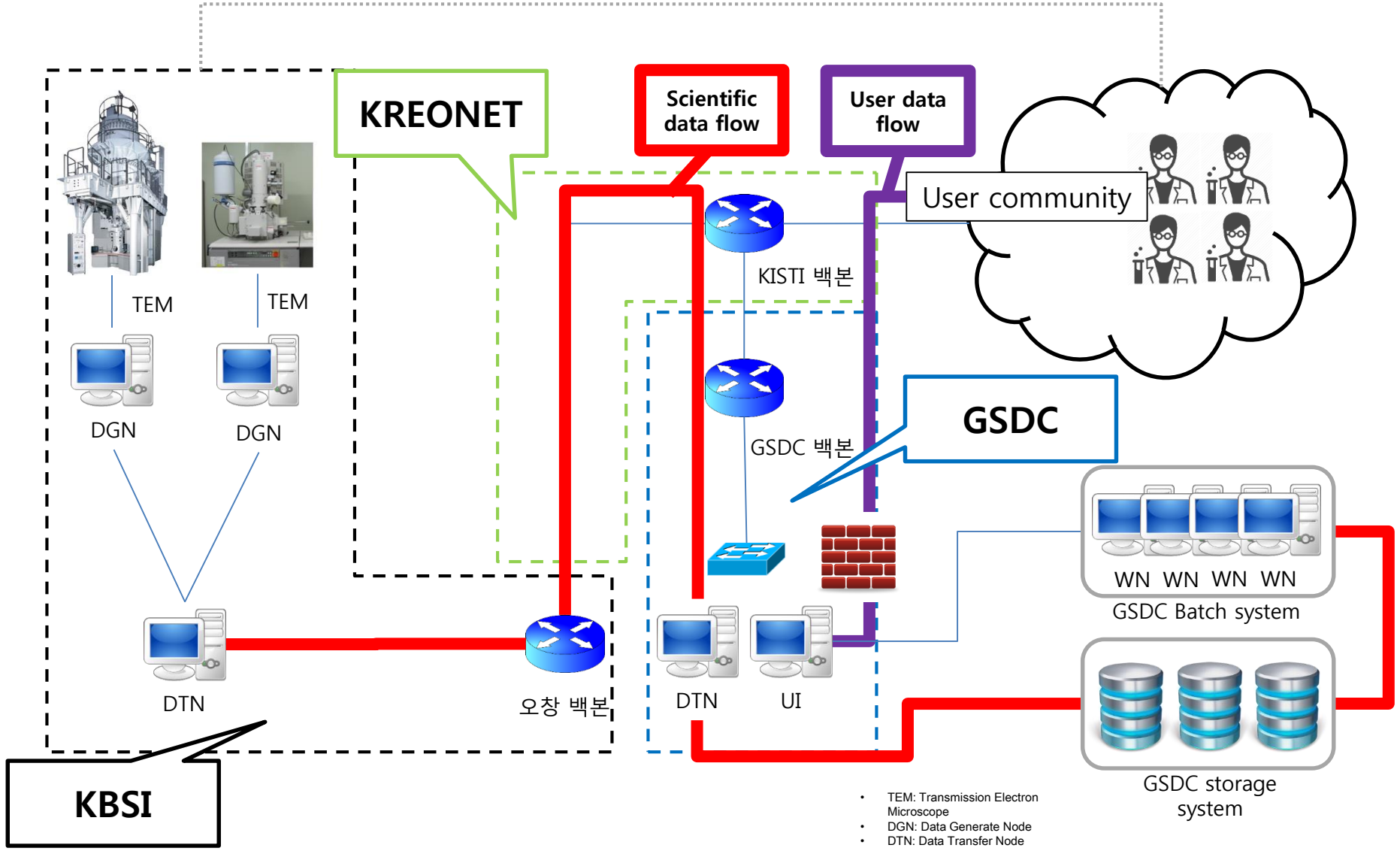
Esnet network Engineering Group, "Science DMZ security, 2013.1.15

Science DMZ



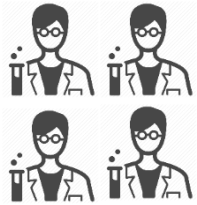
https://fasterdata.es.net/science-dmz/science-dmz-architecture/

Example (GSDC-KBSI)



TEM

TEM



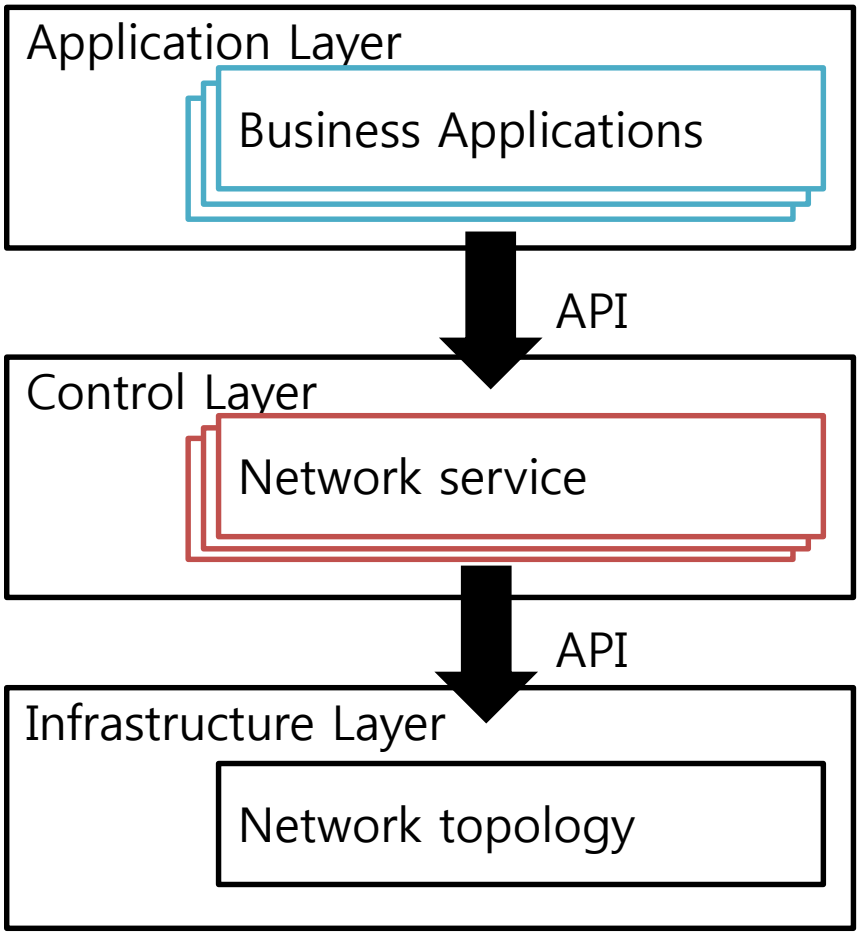
GLOBAL SCIENCE EXPERIMENTAL DATA HUB CENTER
GSDC
국가대용량실험데이터허브센터

Korea Institute of Science and Technology Information

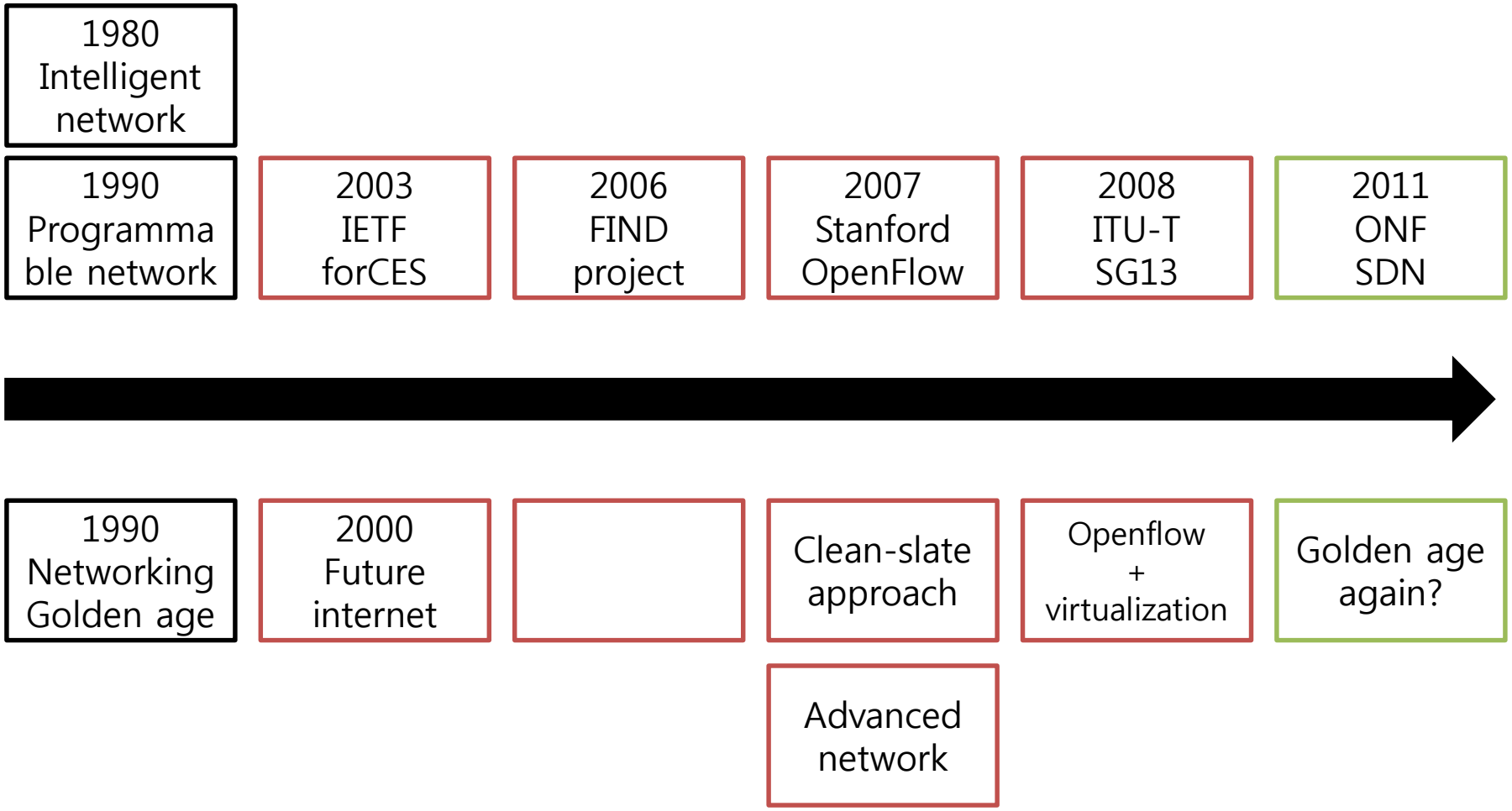
Software Define Network

- **Legacy problem**
  - **Packet switching**
  - **I`m sorry that we made the network as that way**
    – Prof. Kilnam Chon, 2016.11.23.

- **SDN**
  - **Emerging network architecture**
  - **Separate control and data plane**
  - **Characteristic**
    - **Directly programmable**
    - **Agile**
    - **Centrally managed**
    - **Programmatically configured**
    - **Open standards/vendor neutral**

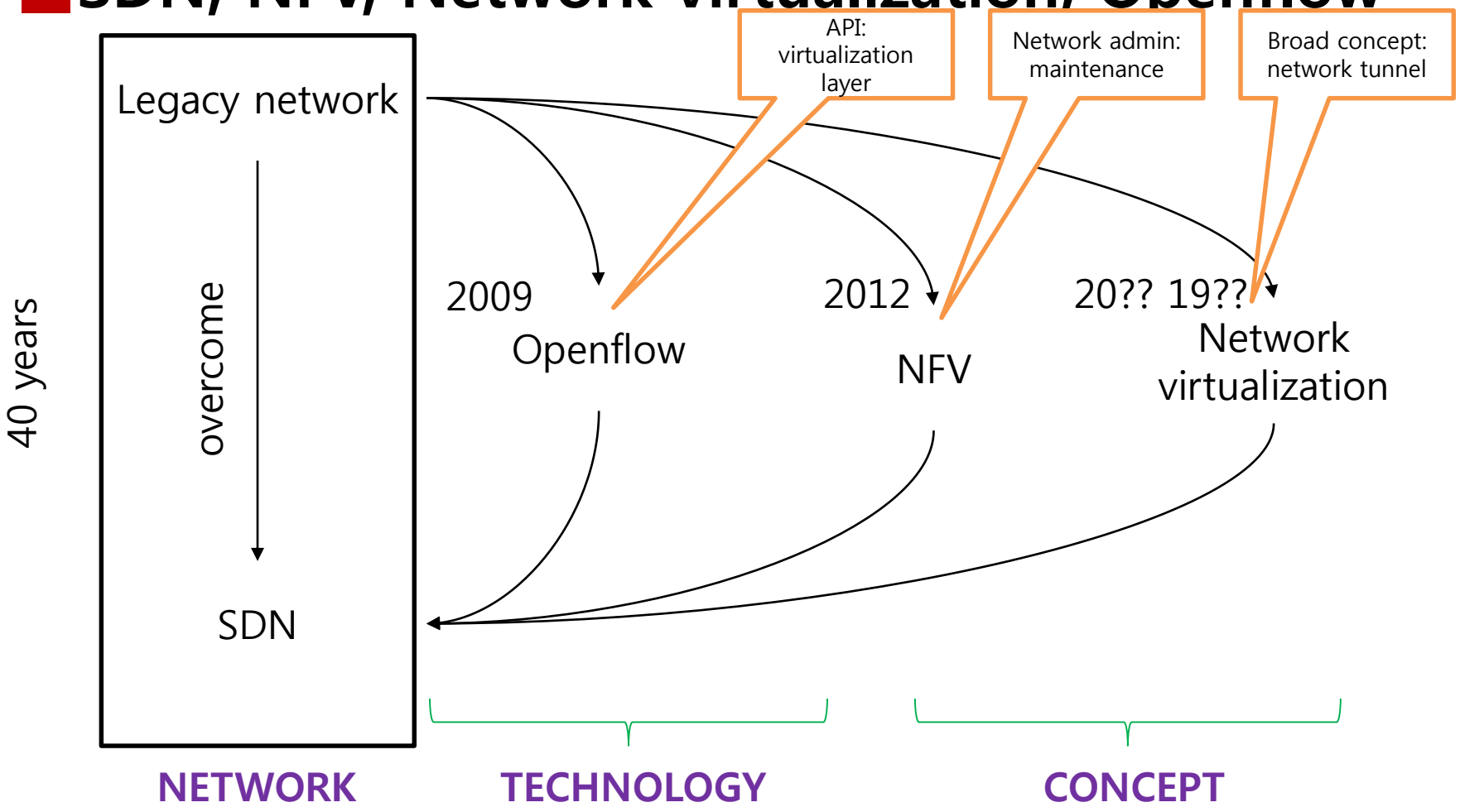


History of SDN



Confusion! Confusion! Confusion!

SDN, NFV, Network virtualization, Openflow



Bluetooth

Bluetooth

- **Viking Harald Bluetooth**
  - **10centry, Denmark + Norway**

- **history**
  - **1994 Ericson try to connect mobile phone and peripherals**
  - **Low power consumption(100mW), cheap**
  - **1998 SIG(Special Interest Group):**
    ericson, nokia, IBM, Toshiba, Intel join
  - **IEEE 802.15.1 standard**
  - **2402, 2480 / 2400, 2483.5 MHz**

version	MAX speed	MAX range
3.0	25 Mbit/s	
4.0	25 Mbit/s	200 feet (60m)
5	50 Mbit/s	800 feet (240m)

WIFI

- Naming
  - Wireless + Fidelity
  - 1997 2Mpbs
  - 1999 11Mbps (IEEE 802.11 x
  - IEEE 802.11
    - 802.11 b : 2.4GHz , 11Mbps
    - 802.11 a/g : 5 GHz/ 2.4 GHz, 54Mbps
    - 802.11n : 2.4/5GHz , 150Mbps(600Mbps)
    - 802.11ac : 5GHZ , 6.9Gbps
  - Origin.....EAP (Extensible Authentication Protocol) authentication...
- WiFi travel
  - ISP -> Modem -> Router(AP) -> Extender ☜

Wireless communication networks

- **Evolution**
  - **1st generation / 1981**
    - **Cellular communication**
    - **voice**
  - **2nd generation / 1991**
    - **EU: GSM (TDMA)**
    - **USA: CDMA**
    - **14.4 ~ 64 kbps**
    - **Voice, SMS**
  - **3nd generation / 2002**
    - **EU: WCDMA**
    - **USA: CDMA 2000**
    - **144 kbps ~ 2Mbps**
    - **Voice, internet, video call**
  - **4th generation / 2008**
    - **EU: LTE / LTE-A / 광대역 LTE-A / 3band LTE-A**
    - **USA: Wibro / WiMax**
    - **100Mbps**
    - **Multimedia communication**
  - **5th generation / ?**

Bell Lab

Qualcom

IMT-2000

Slow moving: 1Gbps
Fast moving: 100mbps

5 Generation network

5G Requirements

Data rates	1-10 Gbps / 100s of Mbps
Capacity	36TB / 500 GB /month/user
Spectrum	Higher frequencies & flexibility
Energy	~10% of today's consumption
Latency reduction	~ 1ms
D2D capabilities	NSPS, ITS, resilience, …
Reliability	99.999% within time budget
Coverage	>20 dB of LTE
Battery	~10 years
# devices per area	300.000 per access node

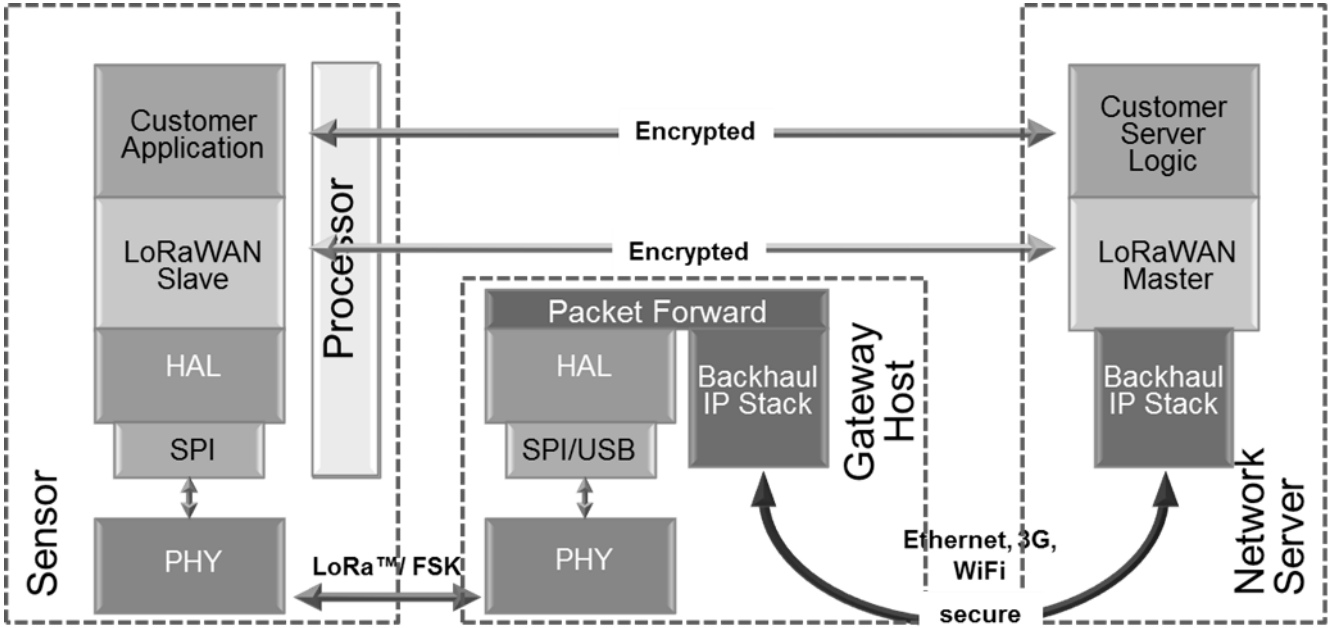
Ultra-dense networks

Ultra Reliable Comm.

Massive Machines

LoRa

- LoRa alliance
- Long-Range sub-GHz Module
  - Mesh, star structure
  - Low power consumption
  - 330Kbps
  - 21 Km range
  - Low cost



https://www.lora-alliance.org/What-Is-LoRa/Technology

I can live alone well...☺

Communication..What should I do?

Local network switch(L2) turns up!

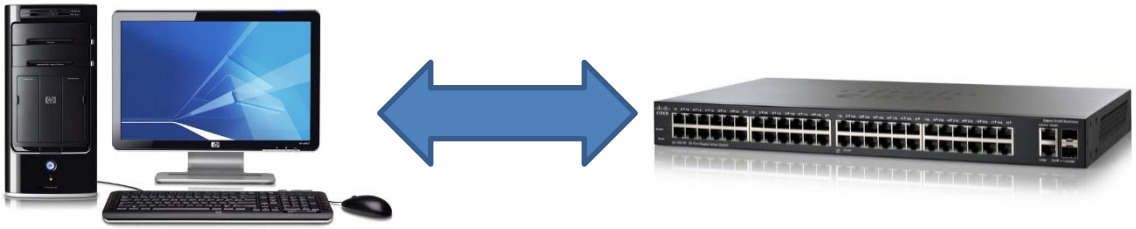
SERVER

1. ARP packet(MAC): A->B
2. ARP reply: B->A
3. TCP 3way hand shaking
4. Connection establish

L2 SWITCH

1. When A send pkt, sw learns a`s MAC in MAC table (L2)
2. To find b`s MAC, search MAC table
3. There is no B`s MAC, then broadcast A`s ARP pkt
4. SW know which port is connected by B

Other network. What should I do?



network router (L3) turns up!

IP routing table

1. Router already know the directions of each IP pkt

```
ip classless (default)
ip route 192.168.1.0 255.255.255.0 10.10.10.2!
```

Verifying Configuration

To verify that you have properly configured static routing, enter the show ip route command and look for static routes signified by the "S."

You should see verification output similar to the following:

```
Router# show ip route
Codes: C - connected, S - static, R - RIP, M - mobile, B - BGP
D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
E1 - OSPF external type 1, E2 - OSPF external type 2
i - IS-IS, su - IS-IS summary, L1 - IS-IS level-1, L2 - IS-IS level-2
ia - IS-IS inter area, * - candidate default, U - per-user static route
o - ODR, P - periodic downloaded static route

Gateway of last resort is not set

10.0.0.0/24 is subnetted, 1 subnets
C 10.108.1.0 is directly connected, Loopback0
S* 0.0.0.0/0 is directly connected, FastEthernet0
```

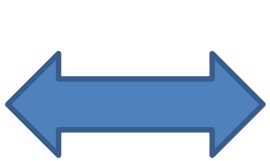
KiSTi Korea Institute of Science and Technology Information

Other other network. What should I do?

L3 routing (OSPF etc)

1. Router already know the directions of each IP pkt
2. If there is no routing path, pkt goes to default routing path

```
ip classless (default)
ip route 192.168.1.0 255.255.255.0 10.10.10.2!
```



foreign network. What should I do?

LAN-WAN communication (LHCOPN)


QUESTIONS