RTB HOUSE

10.06.2024

Data Centers and Networks

Piotr Jaczewski,
Piotr Kowalczyk

RTB HOUSE


RTB HOUSE

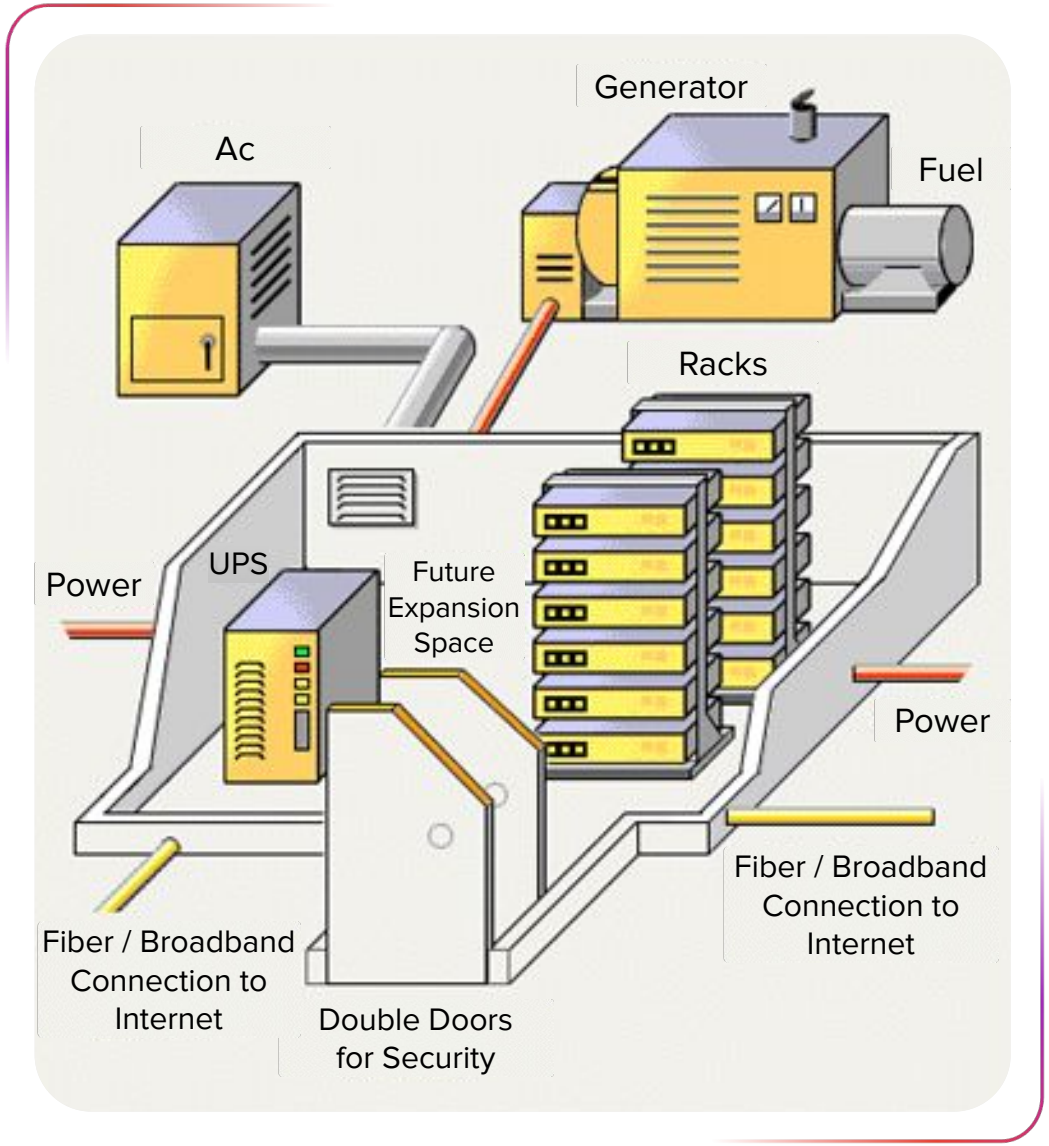

RTB HOUSE


What is
a Data Center?

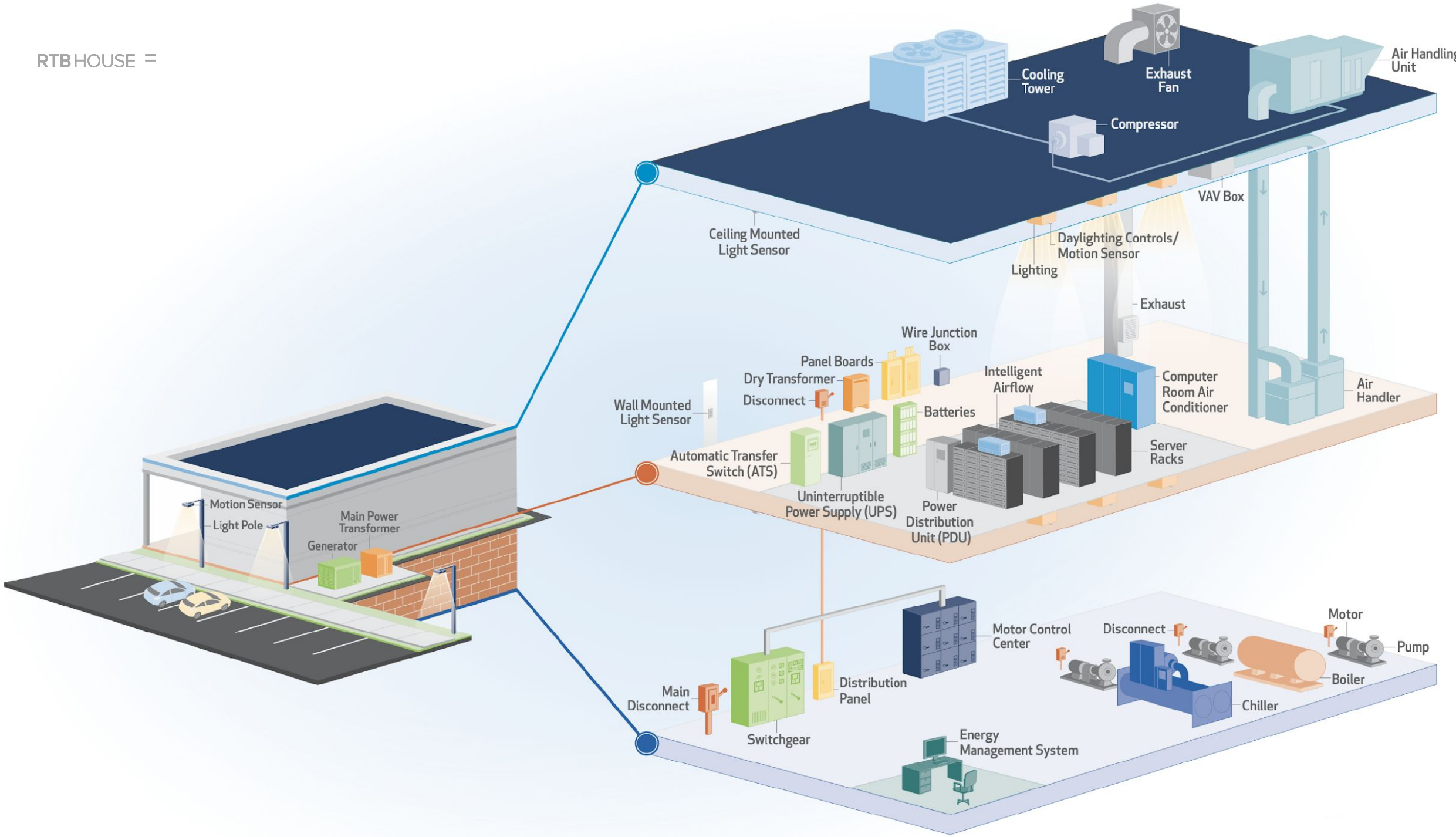
RTB HOUSE

Every Data Center building blocks:

- Space & Location
- Electrical mains power
- Backup power & UPS
- Cooling
- Racks
- Servers
- Networking
- People

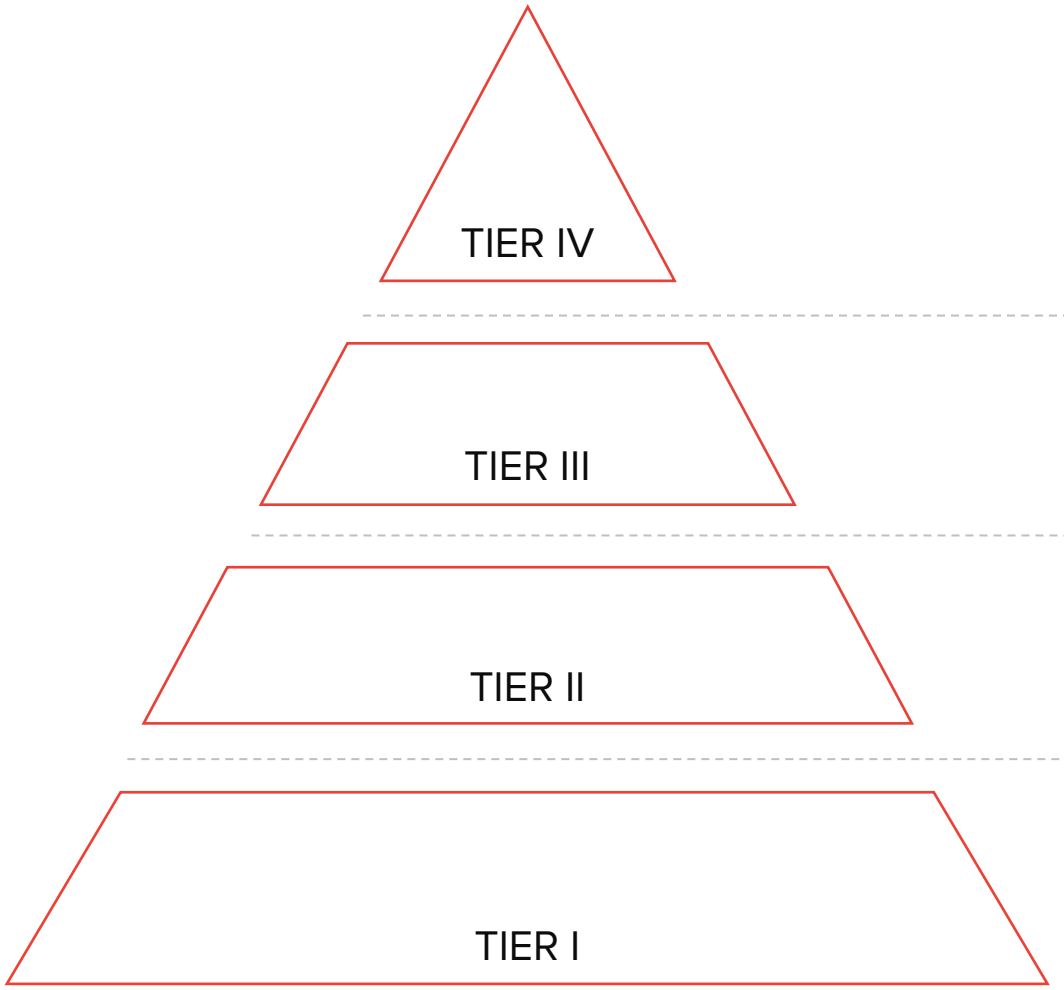


RTB HOUSE


Cooling Tower
Exhaust Fan
Air Handling Unit
Compressor
VAV Box
Ceiling Mounted Light Sensor
Daylighting Controls/ Motion Sensor
Lighting
Exhaust
Wire Junction Box
Panel Boards
Dry Transformer
Disconnect
Wall Mounted Light Sensor
Intelligent Airflow
Computer Room Air Conditioner
Air Handler
Batteries
Automatic Transfer Switch (ATS)
Server Racks
Uninterruptible Power Supply (UPS)
Power Distribution Unit (PDU)
Motion Sensor
Light Pole
Main Power Transformer
Generator
Motor Control Center
Disconnect
Motor
Pump
Main Disconnect
Distribution Panel
Boiler
Chiller
Switchgear
Energy Management System

RTB HOUSE

Data Center Tiers



Enterprise Businesses
99.995% Uptime
26.3 Minutes Downtime Per Year
2N + 1 Fully Redundant
96 Hour Power Outage Protection

Large Businesses
99.982% Uptime
1.6 Hours Downtime Per Year
N + 1 Fault Tolerant
72 Hour Power Outage Protection

Medium - size Businesses
99.749% Uptime
22 Hours Downtime Per Year
Partial Redundancy
in Power and Cooling

Small Businesses
99.671% Uptime
28.8 Hours Downtime Per Year
No Redundancy

High Availability SLA

90% -> 36.5 days per year

99% -> 3.65 days per year

99,9% -> 8.76 hours per year

99,99% -> 52.56 minutes

99,999% -> 5.26 minutes

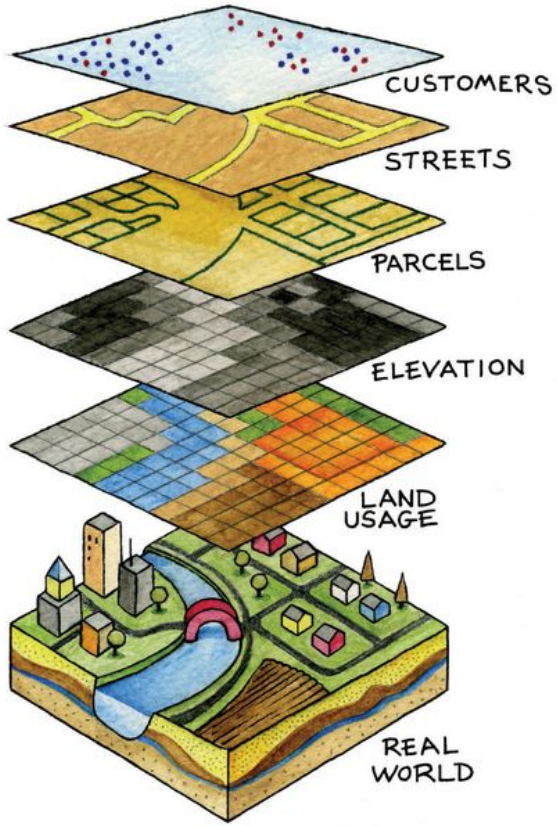
Crucial question:
How many minutes of
downtime per year
can you afford???

Uptime Institute®


What about Networking
in a Data Center?

DC networks can be perceived as layers of:

- Hardware
- Software



RTB HOUSE

Hardware layer - Cables

Copper based - Direct Attach Copper (DACs)

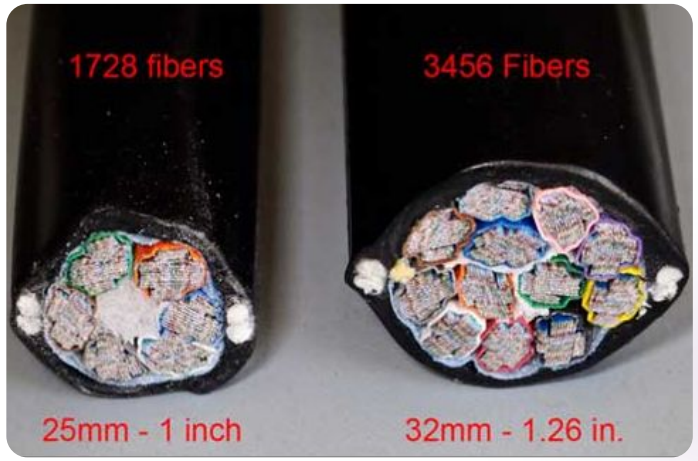
Advantages:

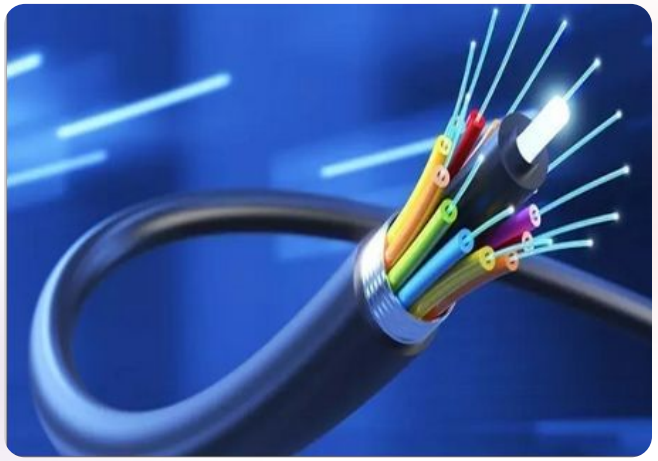
- fairly cheap

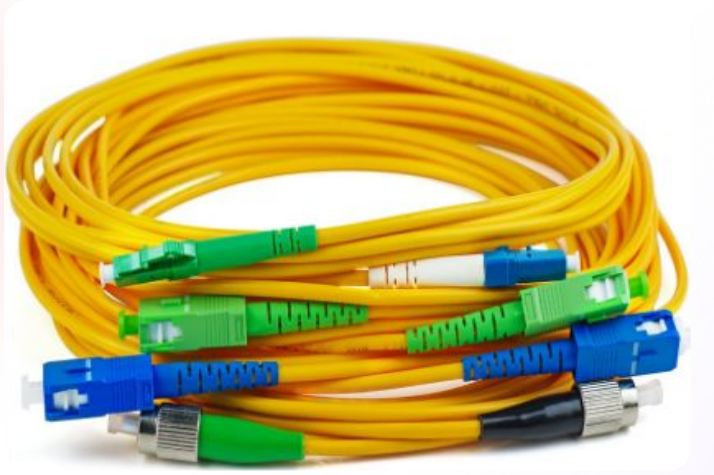
Disadvantages:

- length - maximum ~5 meters (100G)
- bend radius
- relatively easy to destroy
- heavy weight

Hardware layer - Cables







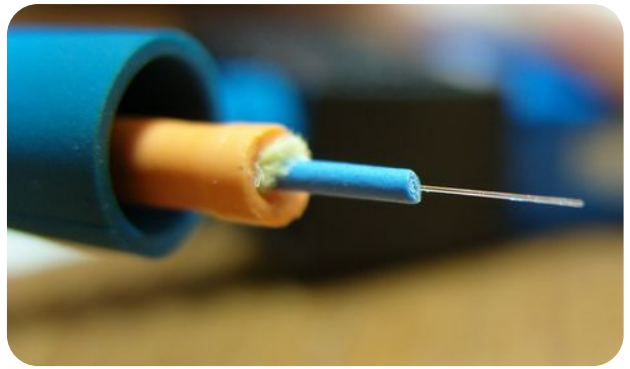
Silicon based - fiber optic

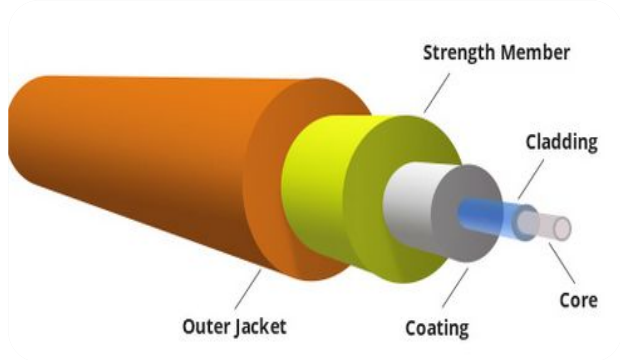
RTB HOUSE

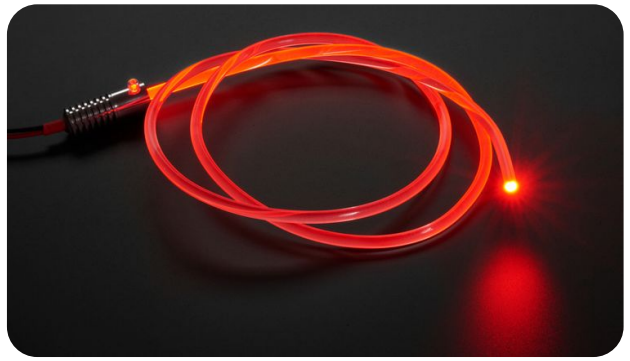
Hardware layer - Cables

Silicon based - fiber optic

- higher bandwidth
- lesser signal attenuation (80km without repeater; copper twisted pair and coaxial requires typically ~5km per repeater)
- immune to electromagnetic interference
- resistant to corrosion
- lightweight
- resistant to tapping

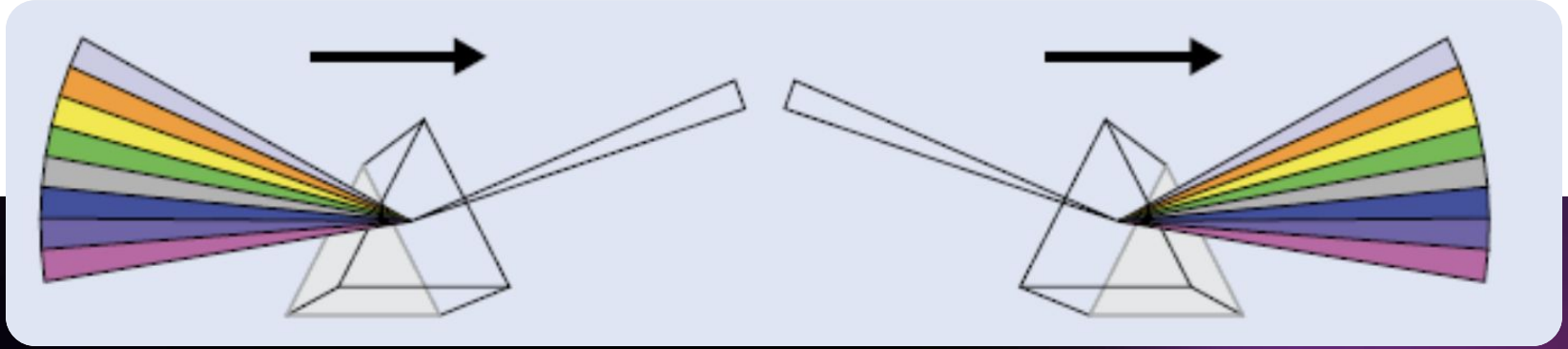






DWDM = Dense Wavelength Division Multiplexing

More lambdas over one fiber pair:

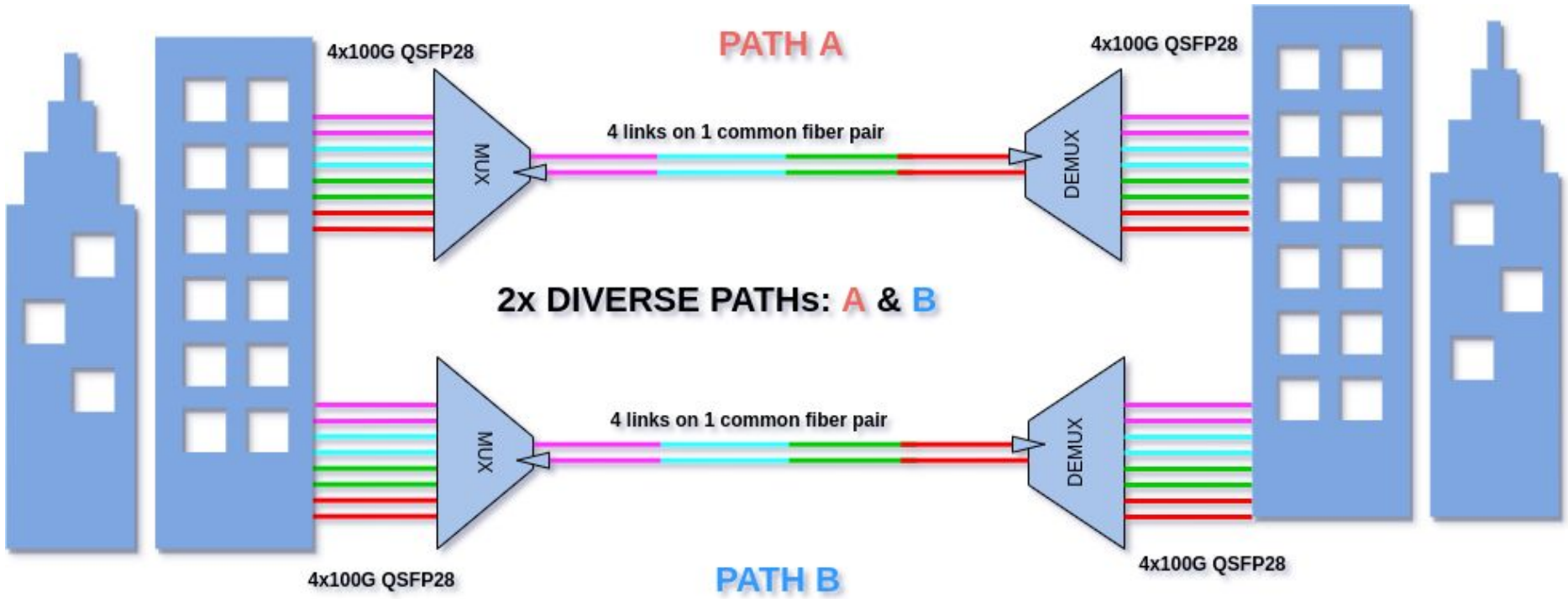


At Ingress: Multiple optical signals of differing wavelengths are combined to form a single optical signal

At Egress: A single optical signal is refracted to separate multiple optical signals of differing wavelengths

State-of-the-art DWDMs can transfer up to 192 **lambdas** (channels) over **one** fiber pair

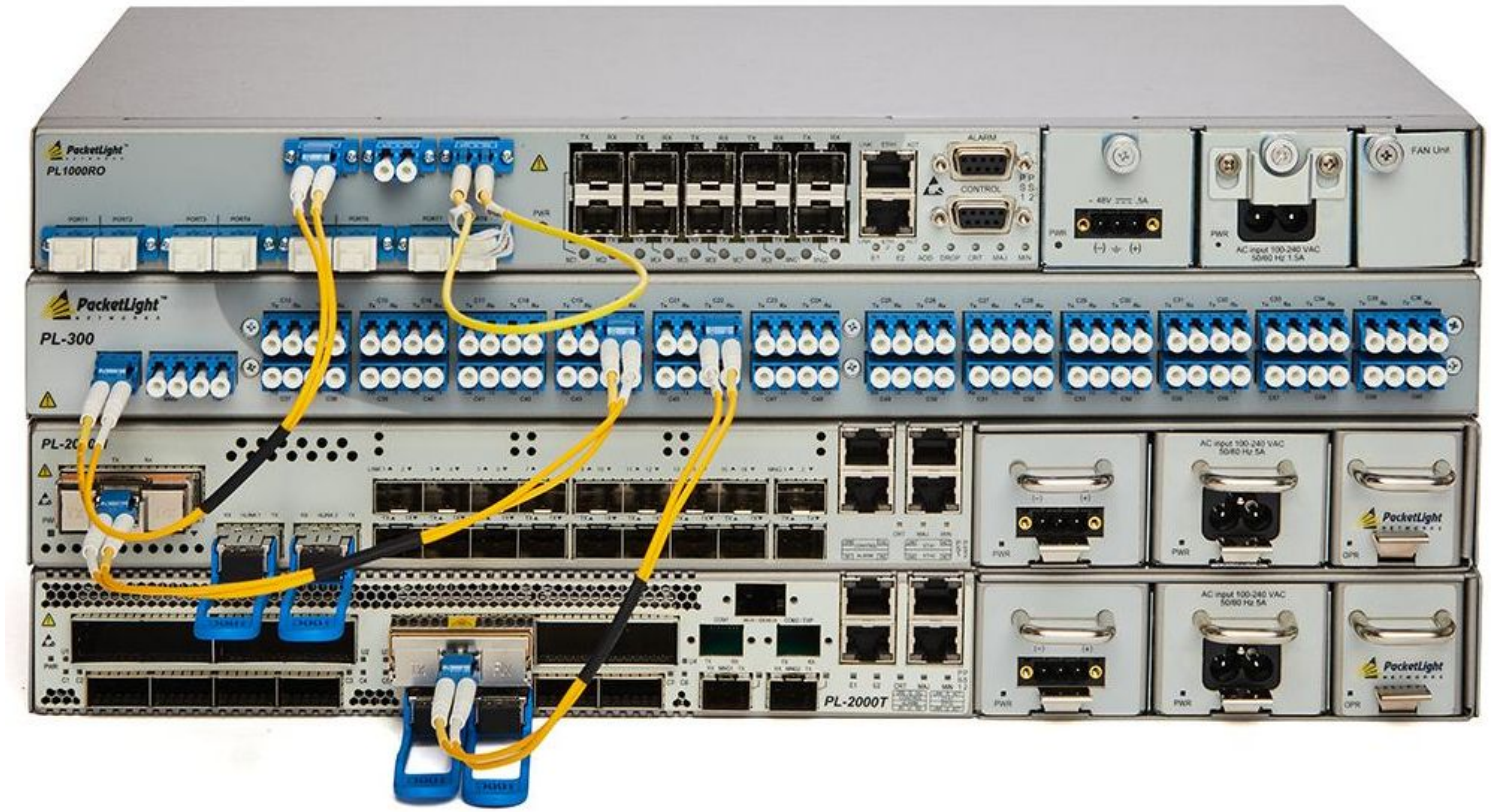
DWDM topology example



State of the art DWDMs can transfer up to 192 **lambdas** (channels) over **one** fiber pair

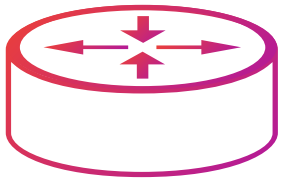
RTB HOUSE

DWDM equipment



A **RTB House** DWDM equipment can transfer up to 4 x 400Gbit/s [**1.6Terabit/s**] over long distances.

Hardware - Network equipment basics



Router

vs

Switch

- Operates on Layer 3 (Network)
- Routes packet with help of IP addresses
- Can communicate to external networks

- Operates on Layer 2 (Data link)
- Sends frames to destination based on MAC address
- Can communicate within a single network only

RTB HOUSE

Hardware layer - Switches

- Typically 100Gbit/s based but now 400Gbit/s is becoming more popular
- Many ports - as much as 128 x 100Gbit/s per switch

RTB HOUSE

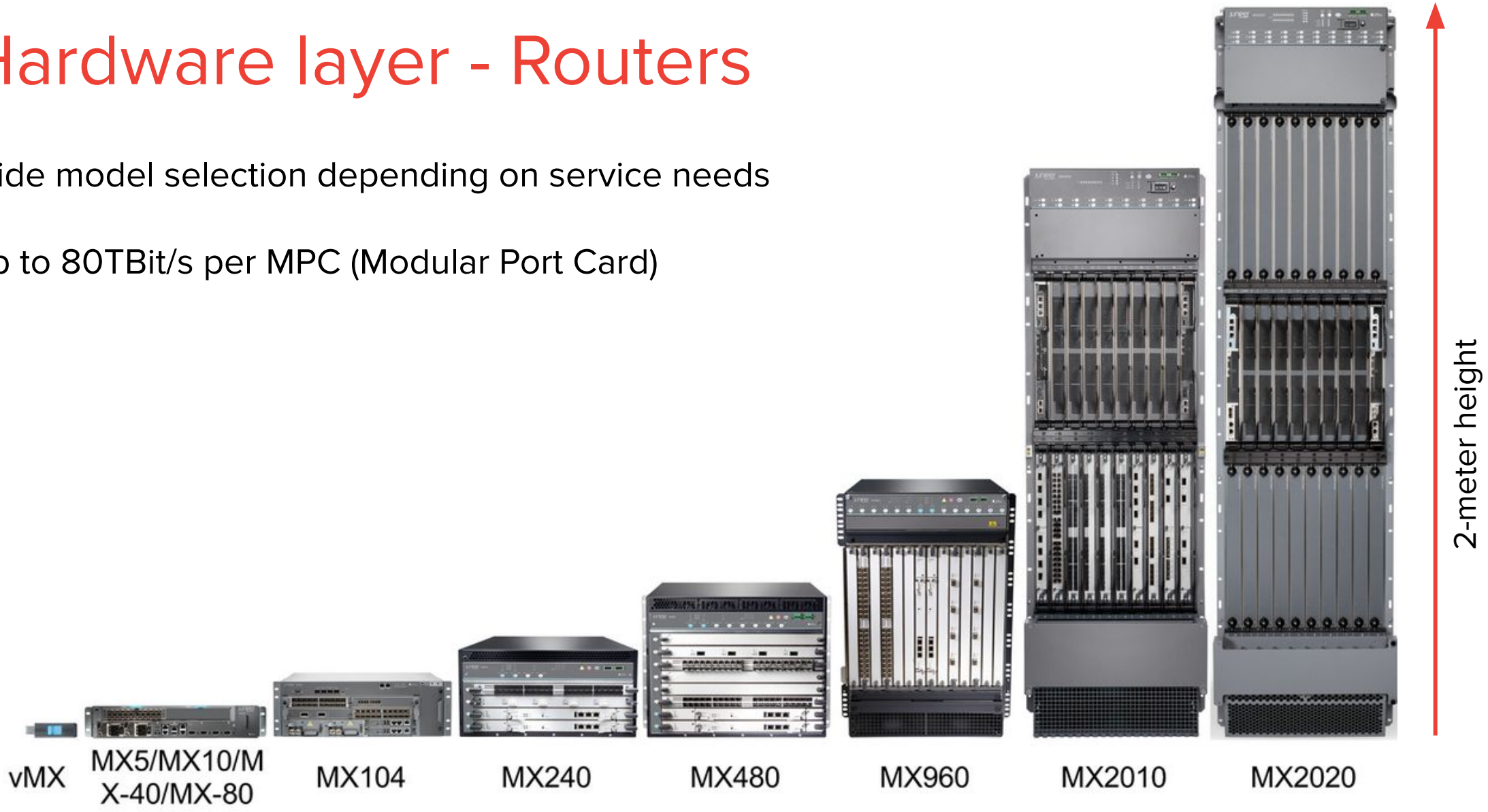
Hardware layer - Routers

At RTB House we use Juniper MX960 routers

Hardware layer - Routers

Wide model selection depending on service needs

Up to 80TBit/s per MPC (Modular Port Card)



Hardware layer - Routers

Packet forwarding throughput:

It's about how many packets per second (**PPS**) you can forward without any service degradation.

When designing a Data Center network you must take for consideration a future growth, especially **please count desired packet rate x4:**

2x for redundancy

2x for application usage growth over time

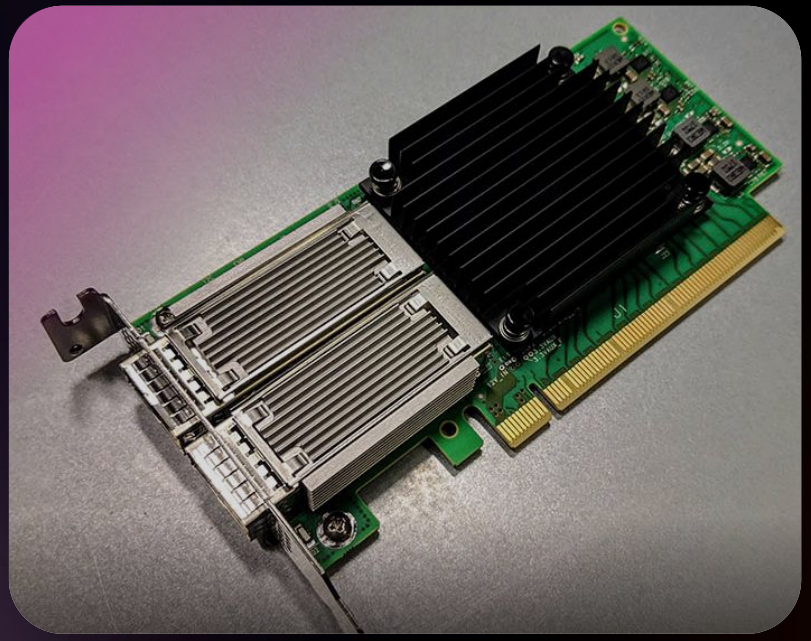
Correlation between BPS and PPS

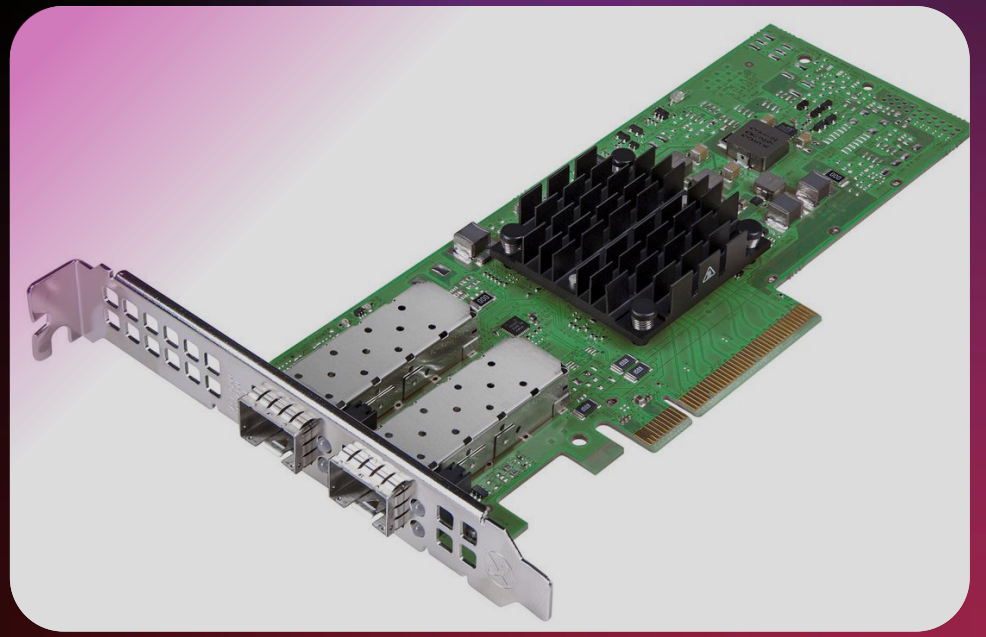
There is a difference between BPS and PPS

Speed	bits / second	bytes / second	maximum PPS
10 Mbps	10,000,000	1,250,000	14,881
100 Mbps	100,000,000	12,500,000	148,810
1 Gbps	1,000,000,000	125,000,000	1,488,095
10 Gbps	10,000,000,000	1,250,000,000	14,880,952
100 Gbps	100,000,000,000	12,500,000,000	148,809,524

RTB HOUSE

Hardware layer - Network Interface Cards





Hardware layer - Network Interface Cards

- Typically 2x10Gbit/s, 2x25Gbit/s, 2x100Gbit/s
- Modern high-end NICs (aka Smart NICs) have additional capabilities like CHECKSUM/TCP/VXLAN/SSL offloading but are 4x times more expensive.
- NIC offloading - helps host CPU to outsource some network functions to the NIC ASIC (Application Specific Integrated Circuit).

Data Center Networking Models

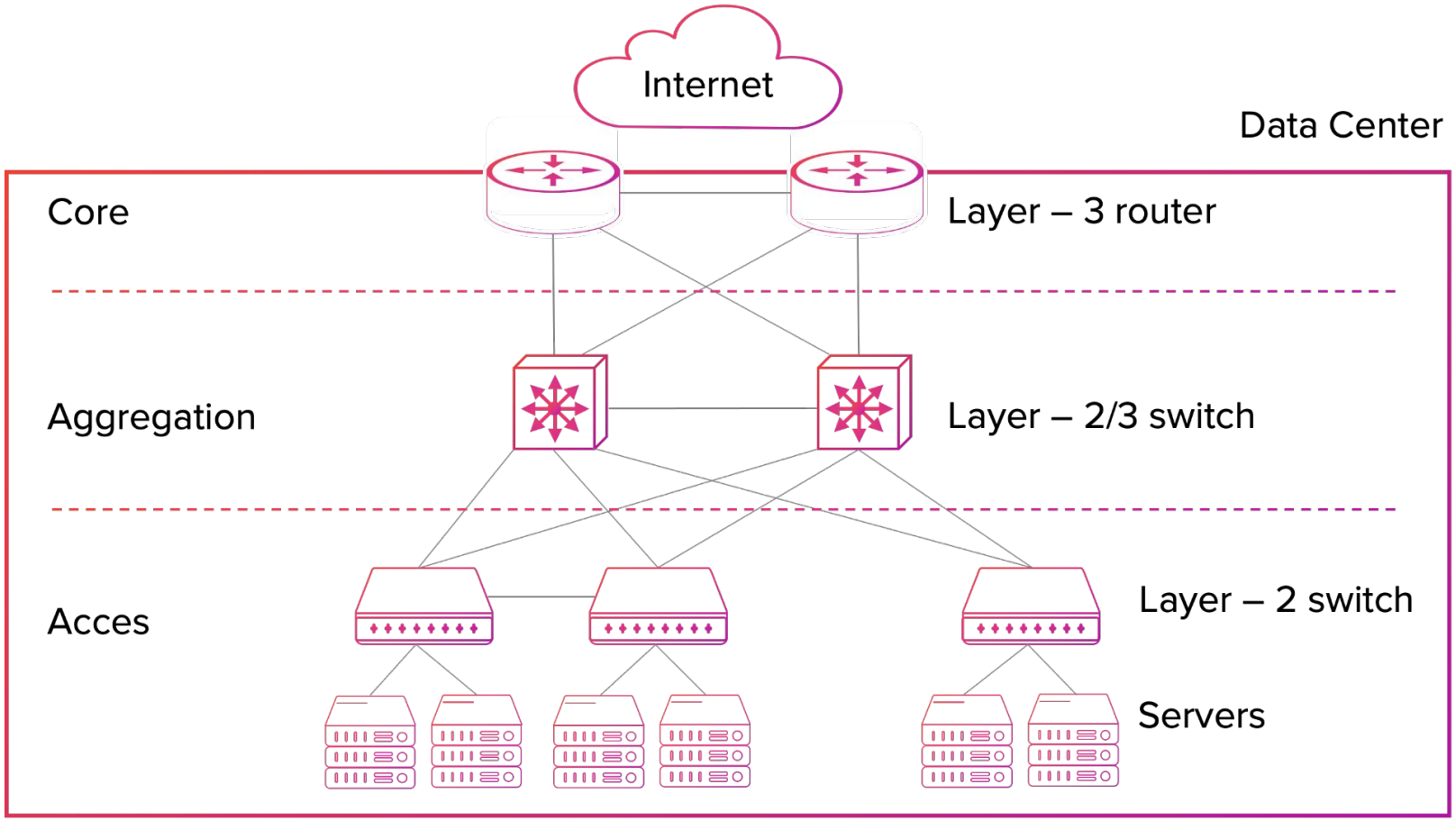
- Traditional “Layer 2” based (Ethernet based only) - past model.
- Hybrid (Ethernet VPN/VXLAN based) - current model.
- Modern “pure Layer 3” based (routing-on-the-host) - likely to be a standard in the future.

Traditional (old) Layer 2 Topology

According to the current requirements has many drawbacks:

- Cannot scale to thousands of hosts - dependent on ARP (Address Resolution Protocol) for MAC address discovery - a lot of broadcast messages
- Not loop proof - depends on long recovering STP (Spanning Tree Protocol)
- No load balancing at the network level - no ECMP
- Reduced Endpoint Mobility - difficult to transfer VMs to new physical servers

Traditional (old) Layer 2 Topology



Traditional (old) Layer 2 Topology

Challenges regarding Ethernet MAC addresses:

If a networked Layer 2 device could contain a list of all known MAC addresses, then the network node could function in much the same way as a router, forwarding frames instead of packets hop-by-hop through the network from source LAN to destination LAN. However, the MAC address is much larger than the IPv4 address currently used on the Internet backbone (48 bits compared to the 32 bits of IPv4).

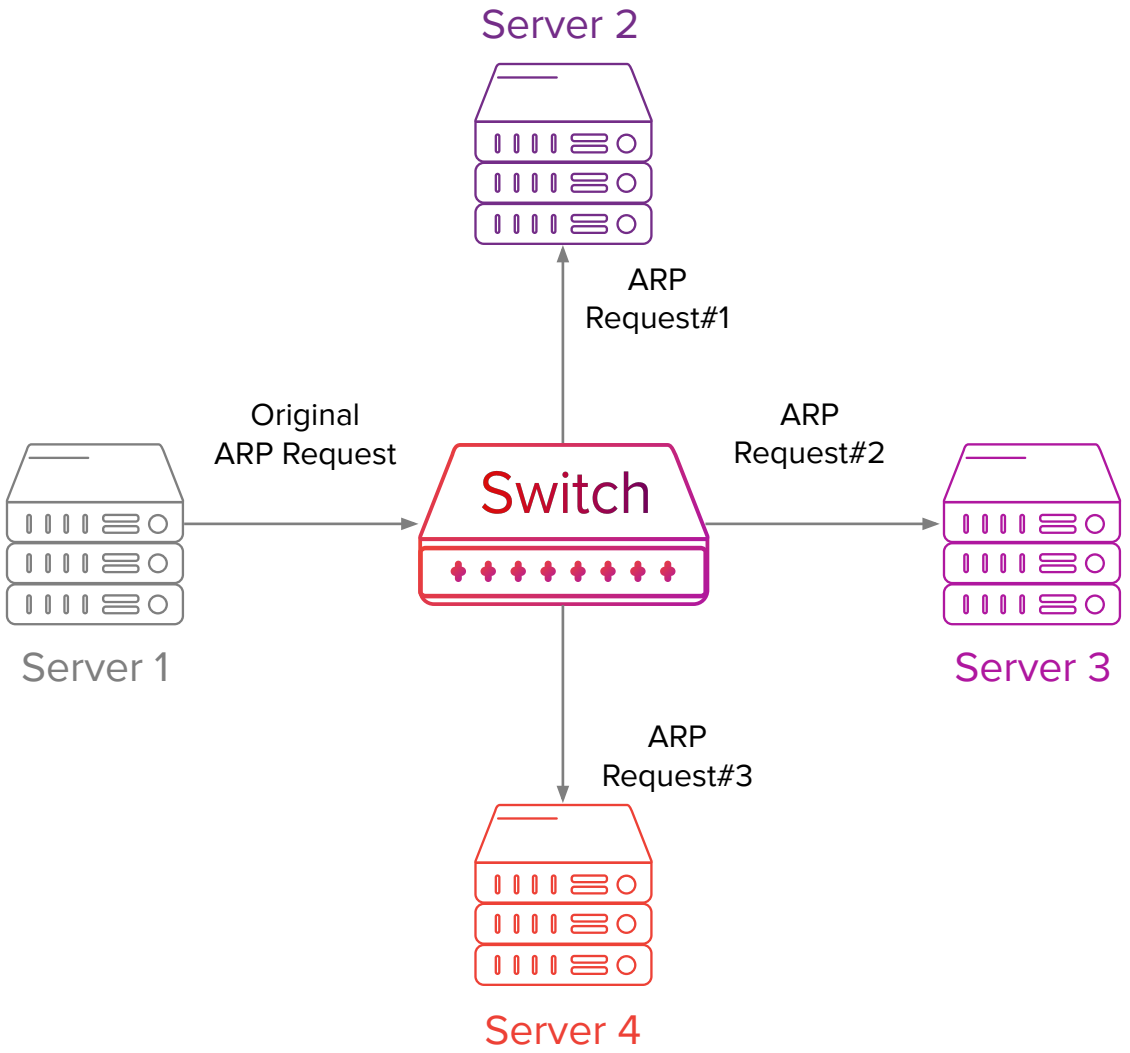
The MAC address has no “network organization” like the IPv4 or IPv6 address, so a Layer 2 network node would have to potentially store every conceivable MAC address in memory for next-hop table lookups. Instead of tables of about 125,000 entries, every Layer 2 network node would be required to store billions of entries.

ARP requests - flood of frames

Too many ARP request are a big issue with 1000s hosts in a single network.
End hosts are frequently asking for neighbor MACs which leads to multiple unnecessary communication flows that saturate every switch port in big broadcast domain.

When the network has 1000s of ports, each ARP request is broadcasted (copied) to each one of 1000s ports.

Remember: In theory, ARP would have to run an address resolution for each outgoing IP packet before transmitting. However, this would significantly increase the required bandwidth. For this reason, address mappings are stored in a table, the so-called ARP cache, as the protocol discovers them. But the cache has to be purged eventually, what results in more ARP requests.


Server 2
ARP
Request#1
Original
ARP Request
Switch
ARP
Request#2
Server 1
Server 3
ARP
Request#3
Server 4

```
15:10:43.195376 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.202 tell 192.168.130.2, length 46
15:10:43.259532 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.217 tell 192.168.130.2, length 46
15:10:44.059459 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.210 tell 192.168.130.2, length 46
15:10:44.155602 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.211 tell 192.168.130.2, length 46
15:10:44.155602 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.215 tell 192.168.130.2, length 46
15:10:44.155602 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.209 tell 192.168.130.2, length 46
15:10:44.219543 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.202 tell 192.168.130.2, length 46
15:10:44.283516 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.217 tell 192.168.130.2, length 46
15:10:45.083812 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.210 tell 192.168.130.2, length 46
15:10:45.179518 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.211 tell 192.168.130.2, length 46
15:10:45.179518 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.215 tell 192.168.130.2, length 46
15:10:45.179518 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.209 tell 192.168.130.2, length 46
15:10:45.243508 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.202 tell 192.168.130.2, length 46
15:10:45.307431 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.217 tell 192.168.130.2, length 46
15:10:46.111501 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.210 tell 192.168.130.2, length 46
15:10:46.203604 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.209 tell 192.168.130.2, length 46
15:10:46.203604 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.215 tell 192.168.130.2, length 46
15:10:46.203604 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.211 tell 192.168.130.2, length 46
15:10:46.267430 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.202 tell 192.168.130.2, length 46
15:10:46.331471 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.217 tell 192.168.130.2, length 46
15:10:47.131611 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.210 tell 192.168.130.2, length 46
15:10:47.227638 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.209 tell 192.168.130.2, length 46
15:10:47.227638 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.211 tell 192.168.130.2, length 46
15:10:47.231474 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.215 tell 192.168.130.2, length 46
15:10:47.291649 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.202 tell 192.168.130.2, length 46
15:10:47.355444 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.217 tell 192.168.130.2, length 46
15:10:47.919247 58:9e:c6:2e:e8:36 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.1 tell 192.168.130.145, length 46
15:10:48.155469 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.210 tell 192.168.130.2, length 46
15:10:48.251507 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.215 tell 192.168.130.2, length 46
15:10:48.251507 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.209 tell 192.168.130.2, length 46
15:10:48.251695 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.211 tell 192.168.130.2, length 46
15:10:48.315460 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.202 tell 192.168.130.2, length 46
15:10:48.379558 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.217 tell 192.168.130.2, length 46
15:10:49.179579 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.210 tell 192.168.130.2, length 46
15:10:49.275479 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.215 tell 192.168.130.2, length 46
15:10:49.275568 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.211 tell 192.168.130.2, length 46
15:10:49.275695 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.209 tell 192.168.130.2, length 46
15:10:49.339481 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.202 tell 192.168.130.2, length 46
15:10:49.403552 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.217 tell 192.168.130.2, length 46
15:10:49.963356 00:12:16:9b:6b:79 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.160 tell 192.168.130.160, length 46
15:10:50.299583 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.209 tell 192.168.130.2, length 46
15:10:50.299583 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.211 tell 192.168.130.2, length 46
15:10:50.299714 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.215 tell 192.168.130.2, length 46
15:10:50.363586 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.202 tell 192.168.130.2, length 46
15:10:51.323569 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.215 tell 192.168.130.2, length 46
15:10:51.323569 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.209 tell 192.168.130.2, length 46
15:10:51.323569 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.211 tell 192.168.130.2, length 46
15:10:51.387678 40:62:31:09:06:b2 > ff:ff:ff:ff:ff:ff, ethertype ARP (0x0806), length 60: Request who-has 192.168.130.202 tell 192.168.130.2, length 46
```

Issues and Bottlenecks in L2 networks (STP)

Layer 2 networking is also founded on the premise of a single path between any two points; that is the foundation of the STP (Spanning Tree Protocol).

Through the years, the industry has come up with mechanisms such as bonding and MLAG to improve upon this. However, they are all still striving with the single path principle.

Issues and Bottlenecks in L2 networks (redundancy)

Two protocols to improve bandwidth and redundancy in L2 networks:

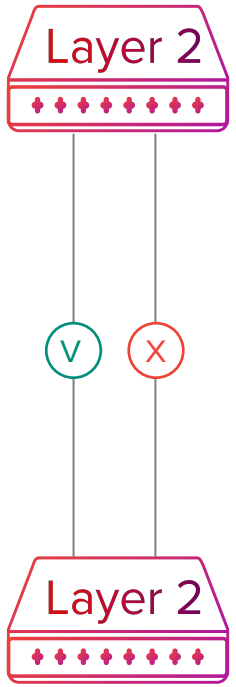
LACP

(Link Aggregation Control Protocol) – host is connected with two links forming 1 logical link to a switch.

MLAG

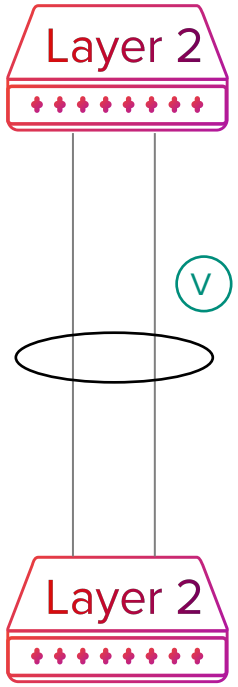
(Multichassis Link Aggregation) – host is connected with two links to 2 switches acting as a one logical switch.

Standalone



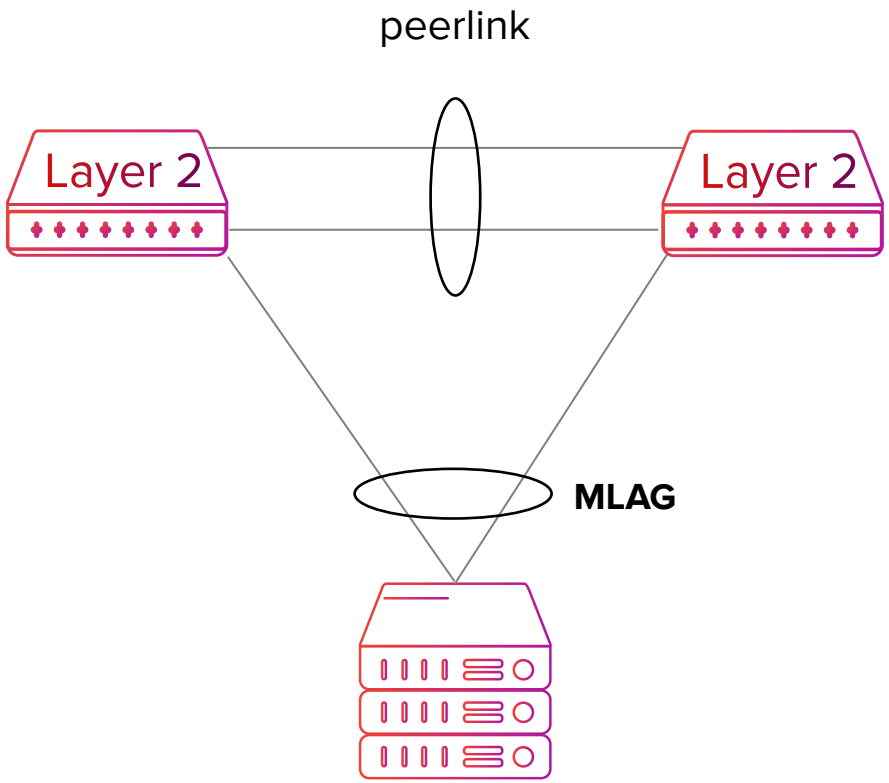
STP blocks second link

LAG



LAG = **Li**nk**AG**gregation

BUNDLE LINKS SO THEY BEHAVE
LIKE ONE SINGLE LINK



FULL REDUNDANCY: TWO SWITCHES & TWO LINKS

RTB HOUSE

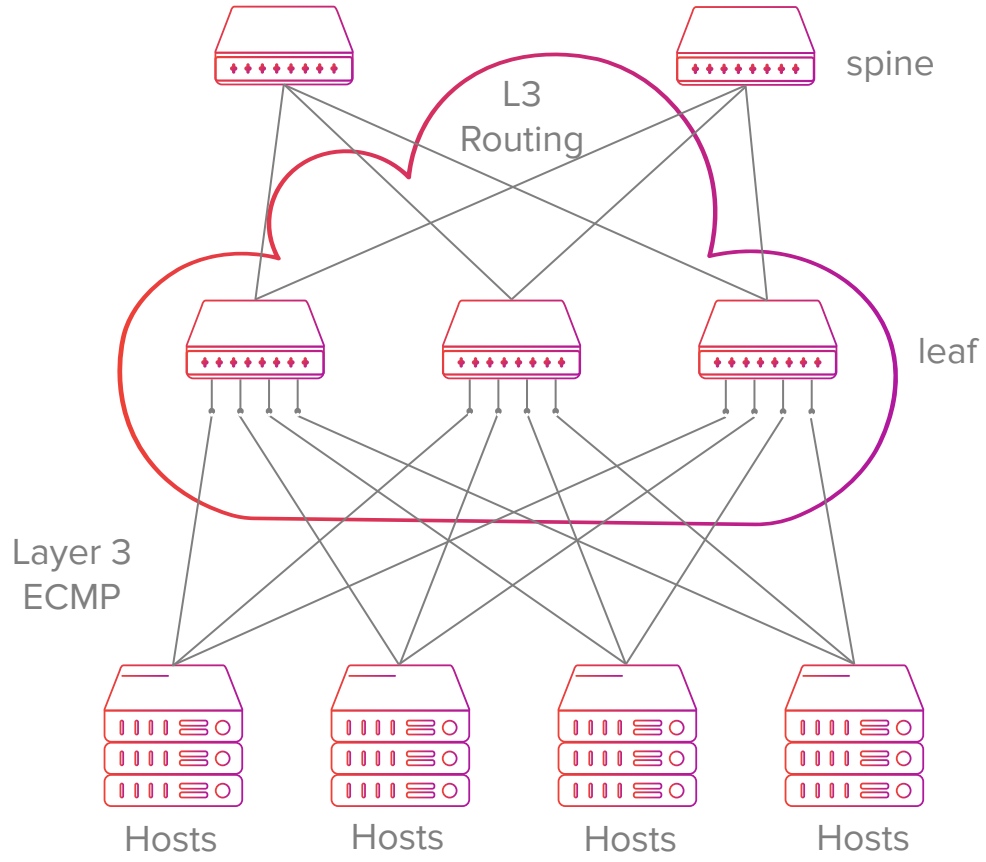
Server Racks:
Host connection to multiple leaf switches

Host to switch the old L2-way :

- Limits redundancy - MLAG can only connect to 2 switches
- Proprietary solution/vendor lock-in

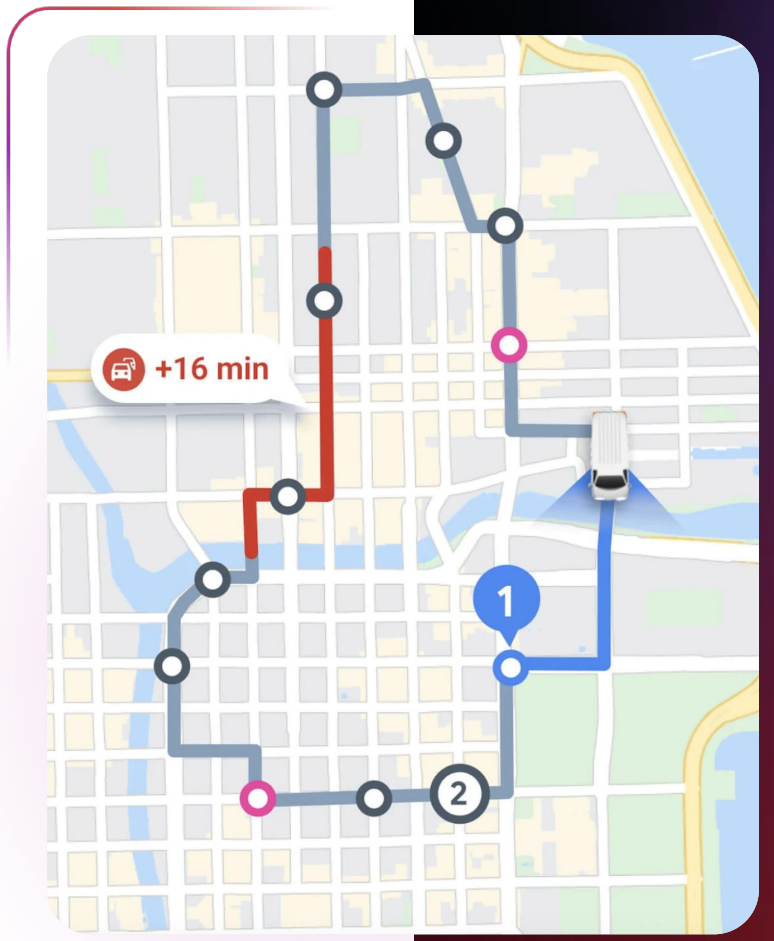
Host to switch the new L3-way

- Routing on the host
- Vast redundancy (3 leafs and more)
- Upgrade leaf switches easily (use routing protocols metrics to redirect traffic)
- Eliminates STP loops/MLAG troubleshooting



What is routing?

Routing is the process of selecting a path for traffic in a network across multiple distinct networks.

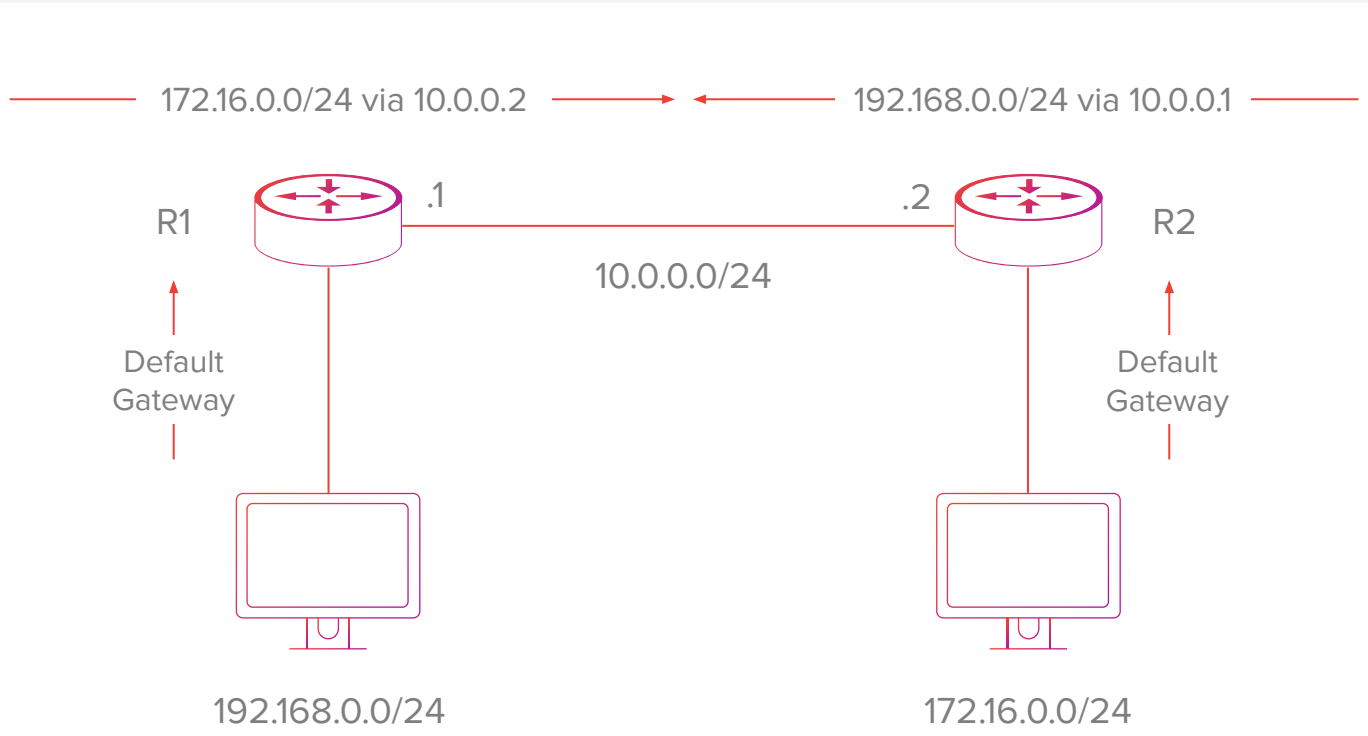


What types of routing do we have?

- Static
- Dynamic
  - ➢ Link state based protocols:
    - OSPF - Open Shortest Path First
    - IS-IS - Intermediate System to Intermediate System
  - ➢ Distance-vector based protocols:
    - RIP v1/v2
    - EIGRP - Cisco proprietary
    - BGP - Border Gateway Protocol

Static routing

On all routers in networks we must add
a manual path to other networks:

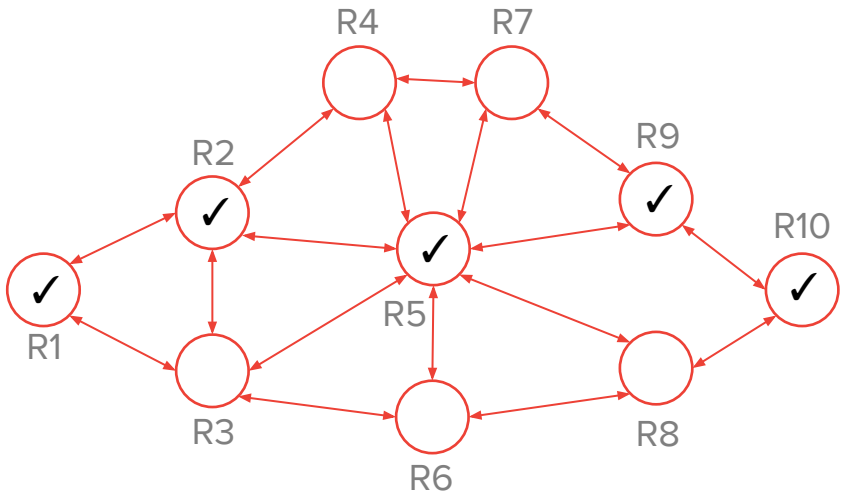


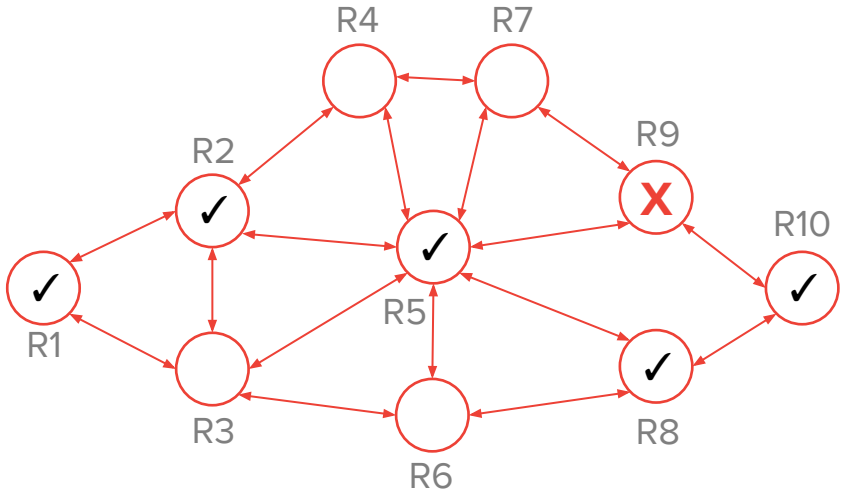
RTB HOUSE

Dynamic routing

Information about routes/prefixes are exchanged dynamically between nodes and circuits.

We do not have to change manually paths when for example one of the nodes is broken.





Network design at RTB House

In all Data Centers we have the following networking assets:

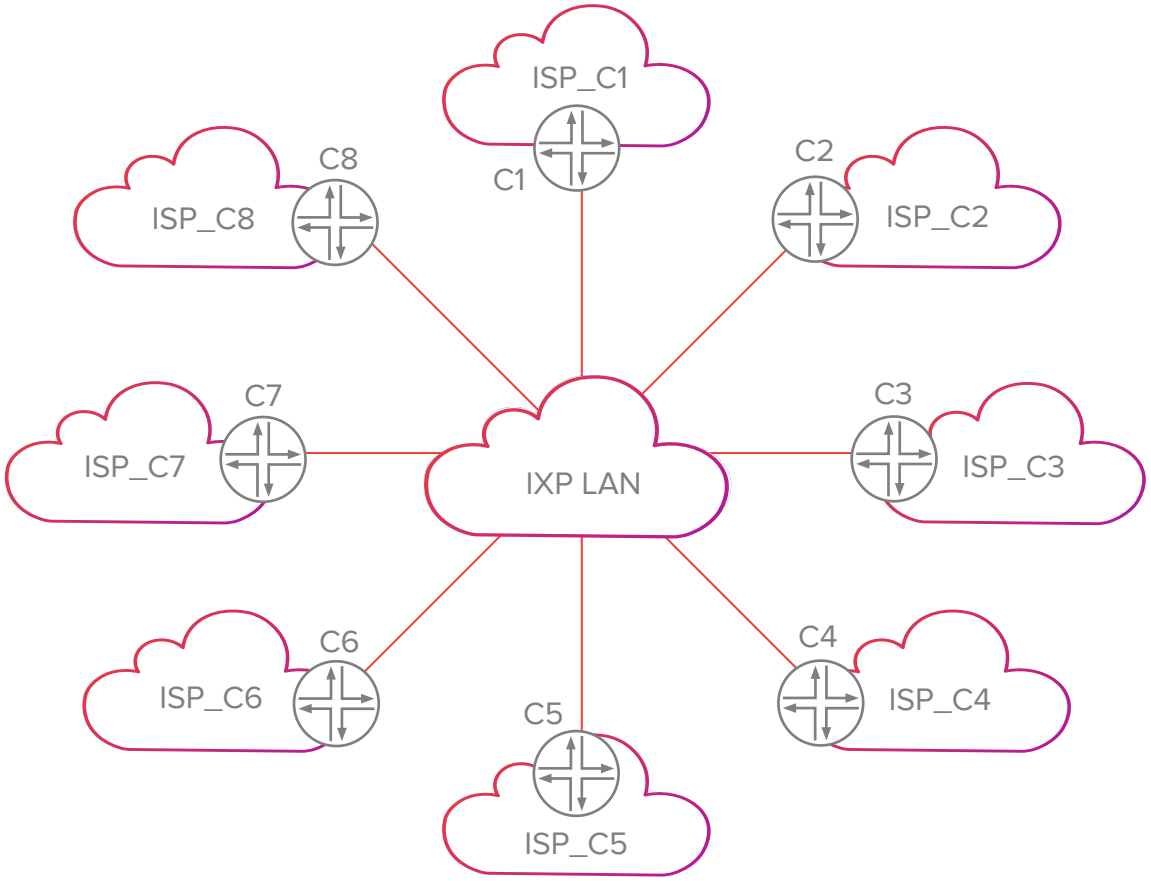
Physical:

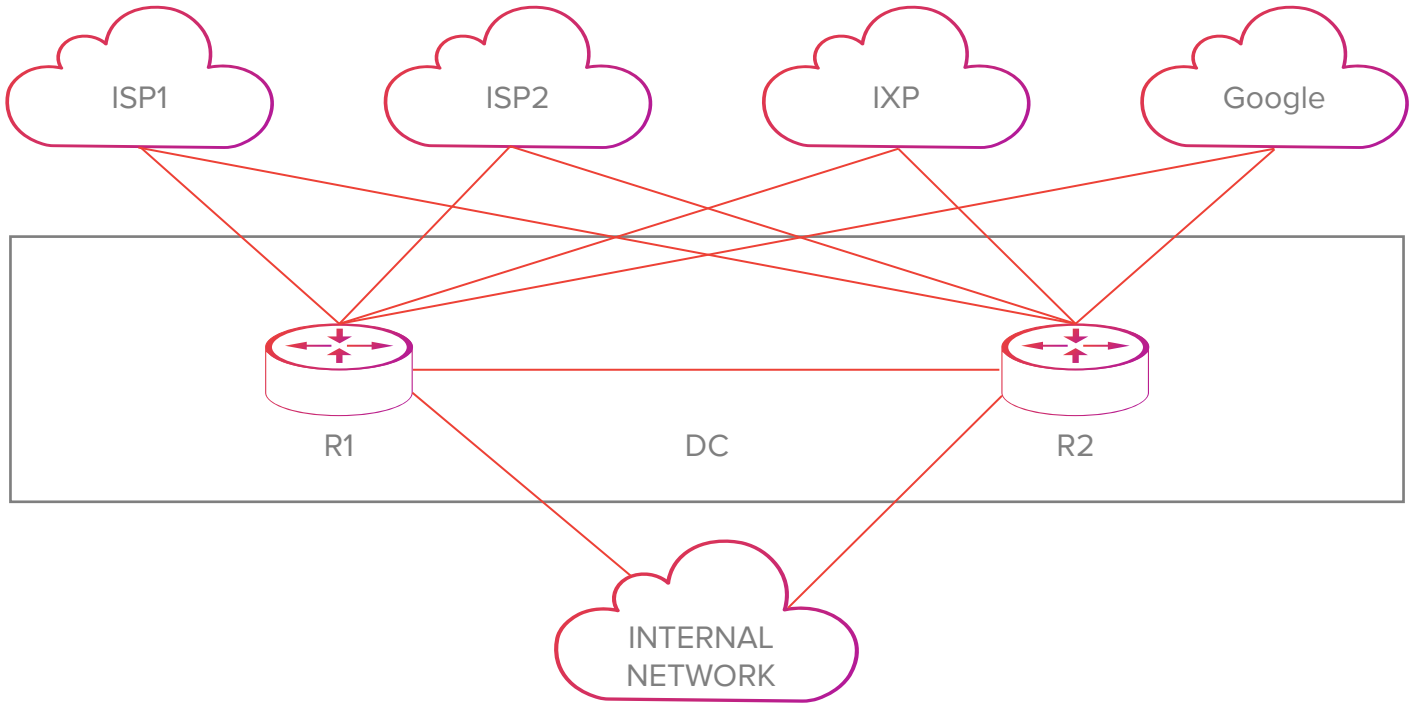
- 2 routers
- 4 x 100Gbit/s connections to ISP (Internet Service Provider)
- 2 x 100Gbit/s to IXP (Internet Exchange Point)
- 2 x 100Gbit/s to Google network

Logical:

- Own AS
  (Autonomous Number – unique for organization) number per DC – required by BGP
- Own IPv4 and IPv6 prefixes

Internet eXchange Point




ISP1
ISP2
IXP
Google
R1
DC
R2
INTERNAL
NETWORK

What routing protocols are used in RTB House network?

IS-IS

connection only between routers in Data Center (faster convergence)

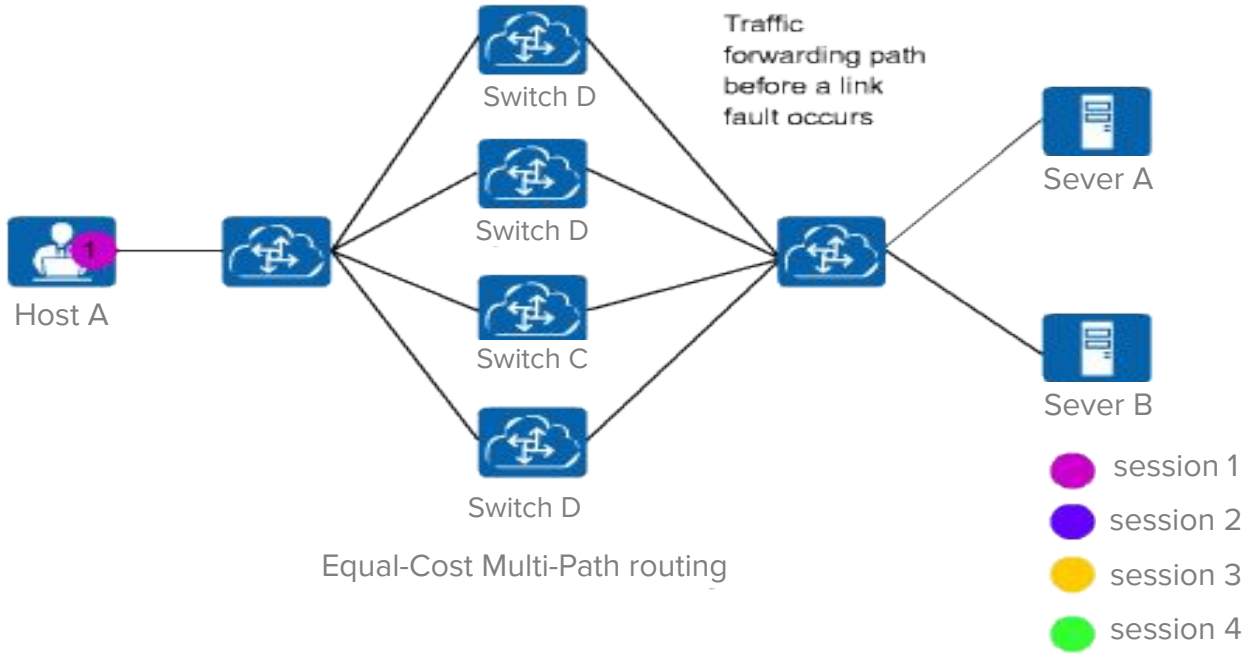
BGP

connection to ISPs (an industry standard) and Load Balancers (ECMP)

Why BGP?

- Flexibility - designed to easily add new functionalities
- Is the de facto standard in connection with other large networks like ISP, IXP, etc.
- Load balancing - due to the use of the ECMP routing strategy.

Equal Cost Multi Path Routing

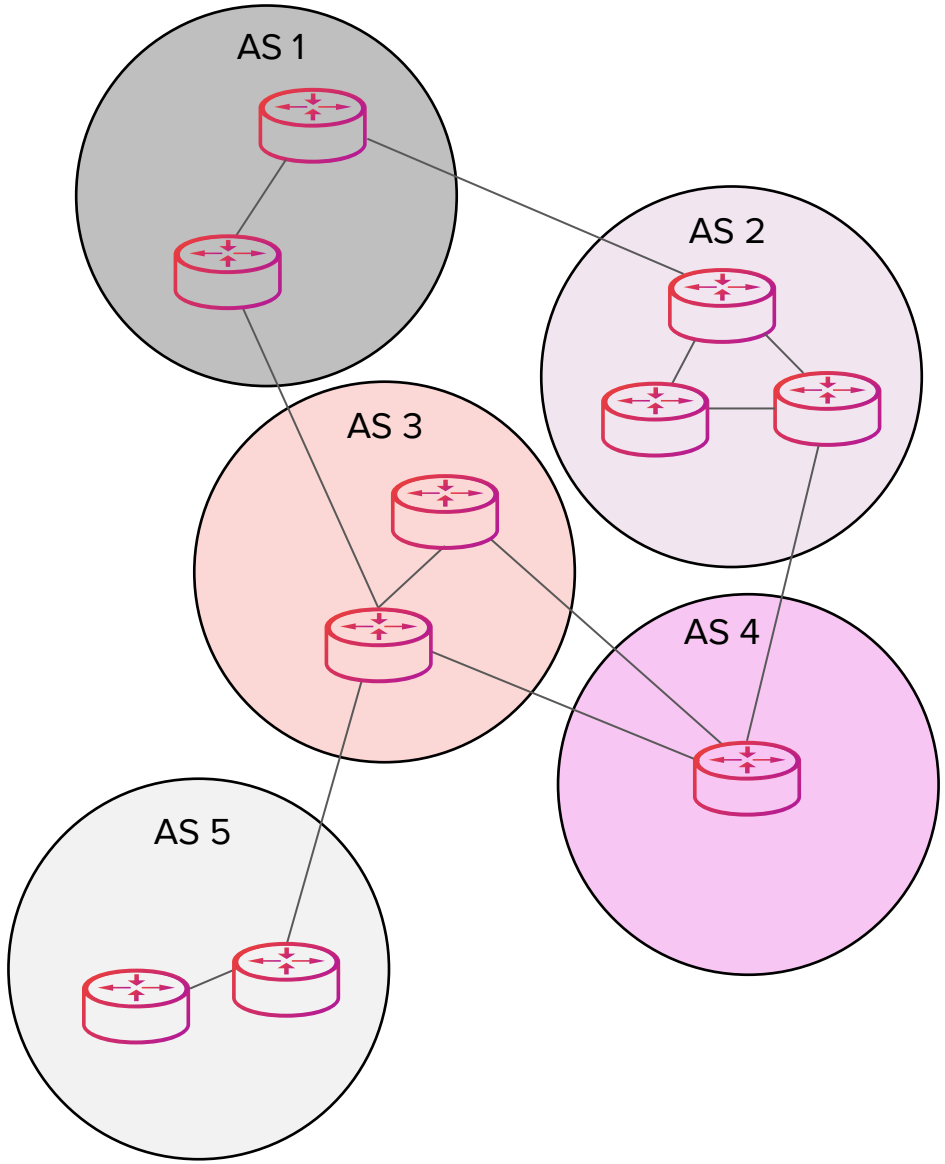


Equal-Cost Multi-Path routing

RTB HOUSE

BGP characteristics

- BGP is a distance - vector protocol
- Each network which uses BGP protocol must have its individual AS number (assigned by IANA)
- Routes are calculated based on the length of the path to the destination prefix



Example - Facebook network outage

Some facts about network at Facebook:

Dedicated in-house built global networks in many countries in the world	Proprietary systems to control traffic between Data Centers	Uses BGP protocol to communicate with all Data Centers	Dedicated DNS servers which were advertised via BGP

Example - Facebook network outage

The course of outage:

During routine maintenance they ceased to advertise DNS prefixes (due to human error) to global routing table.

Users couldn't resolve names - facebook.com was inaccessible

They coudn't rollback this change because internal FB systems were using the same DNS servers ;)

They had to send engineers on site to the DCs to have them debug the issue and restart the systems.

https://blog.cloudflare.com/october-2021-facebook-outage/

https://engineering.fb.com/2021/10/05/networking-traffic/outage-details/

Modern Network Topology Architecture in Data Center

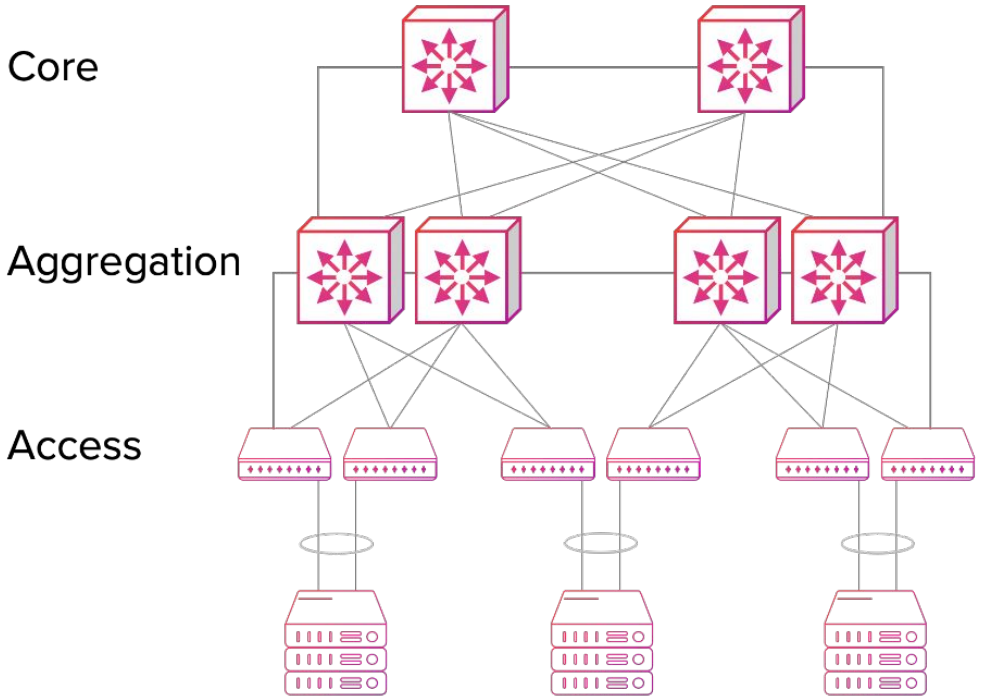
Spine Layer

backbone of the network similar to the core layer in traditional three-tier design. Each Layer 3 port is connected to the underlying Layer 2 leaf switch.

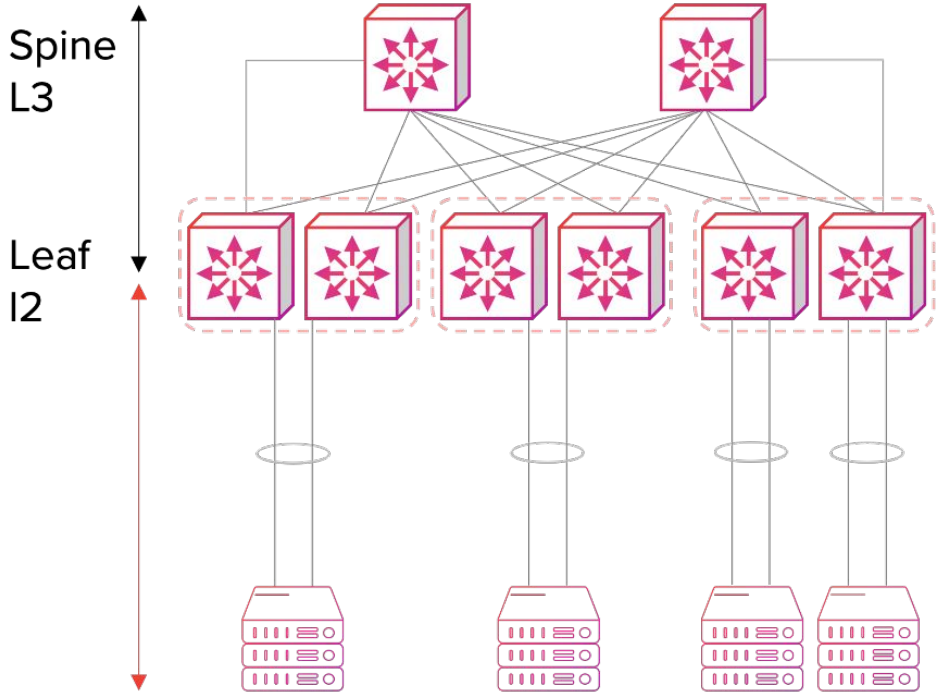
Leaf Layer

connects to end devices like servers. It is similar to the access layer in traditional network topology design.

Traditional 3-Tier Architecture



2-Tier Spine-Leaf Architecture



Spine-Leaf Topology - benefits

Improved Redundancy

As opposed to traditional three-tier architecture where access layer switches connect to only two uplink distribution switches, every leaf switch connects to every spine switch. And instead of Spanning Tree Protocol (STP), we implement pure layer 3 routing mainly based on BGP or OSPF.

Increased Bandwidth

By implementing ECMP, we have the ability to use multiple active links instead of one what increases bandwidth. With STP, only one link is active and the other links are blocked.

Improved Scalability

In the event of oversubscription, we can add a spine switch and connect it to every leaf switch. If the port density is a concern, we can add a leaf switch and connect it to every spine switch.

Spine-Leaf Topology - benefits

Lower Costs

fixed-configuration switches unlike modular switches, have a fixed number of ports and are usually not expandable. Many spine-leaf networks use fixed-configuration switches.

Low Latency and Congestion Avoidance

With having only a maximum of two hops between any source and destination nodes, we can make a more direct traffic path, which improves performance and reduces bottlenecks. The only exception is when the destination is on the same leaf switch.

Spine-Leaf Topology - limitations

Amount of Cables

We need to run more cables since each leaf must be connected to every spine device.

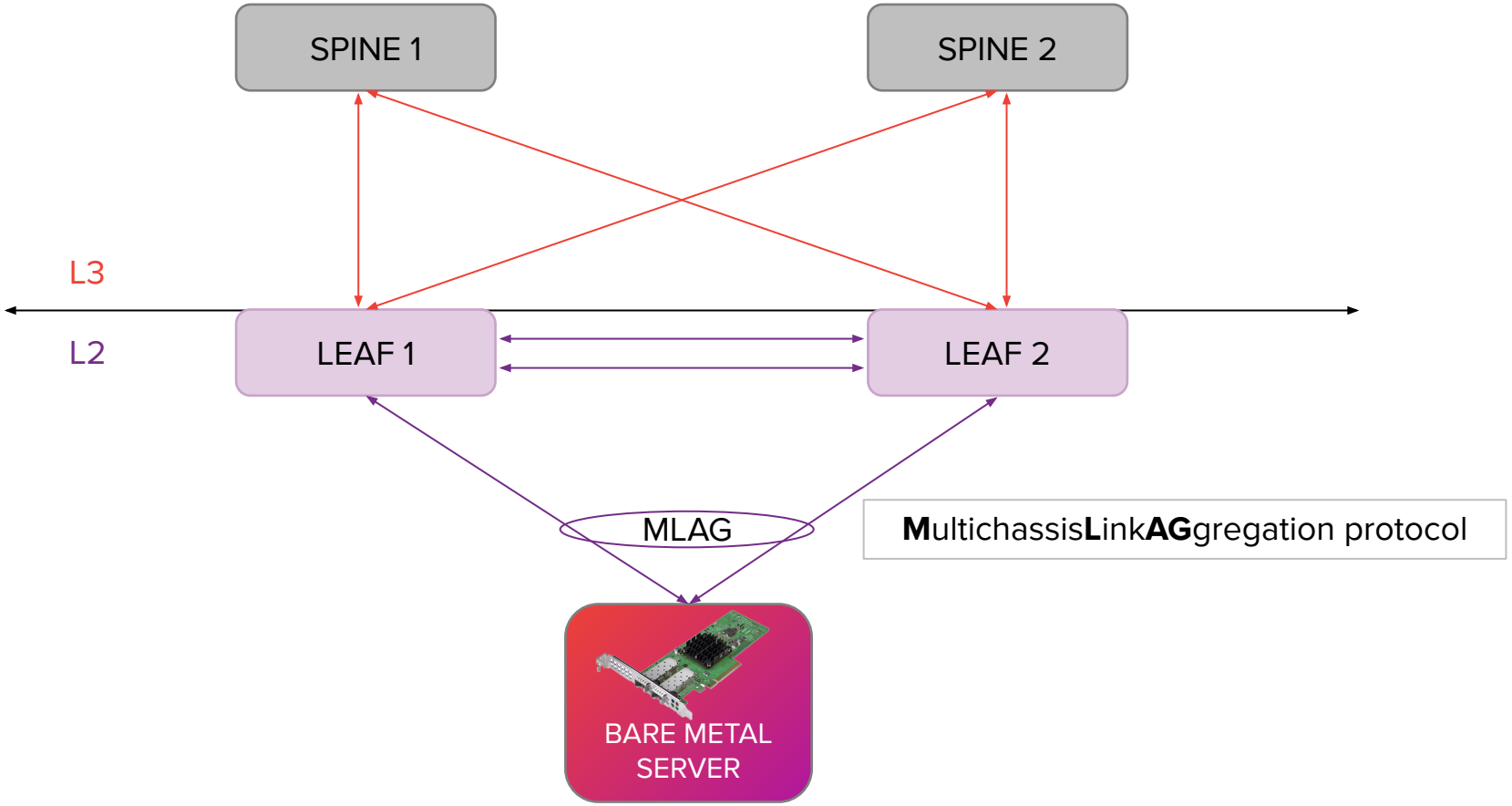
Limited Hosts

The number of hosts that we can support is limited. Spine port counts can restrict the number of leaf switch connections.

Oversubscription

practice of committing more network bandwidth to devices connected to that network than what is physically available. Typical Oversubscription ratio in DC networks is 2.4:1 (e.g. 48x10G / 200G uplink)

Full Redundancy



RTB HOUSE

Overlay and Underlay routing in Data Center Networks

Underlay

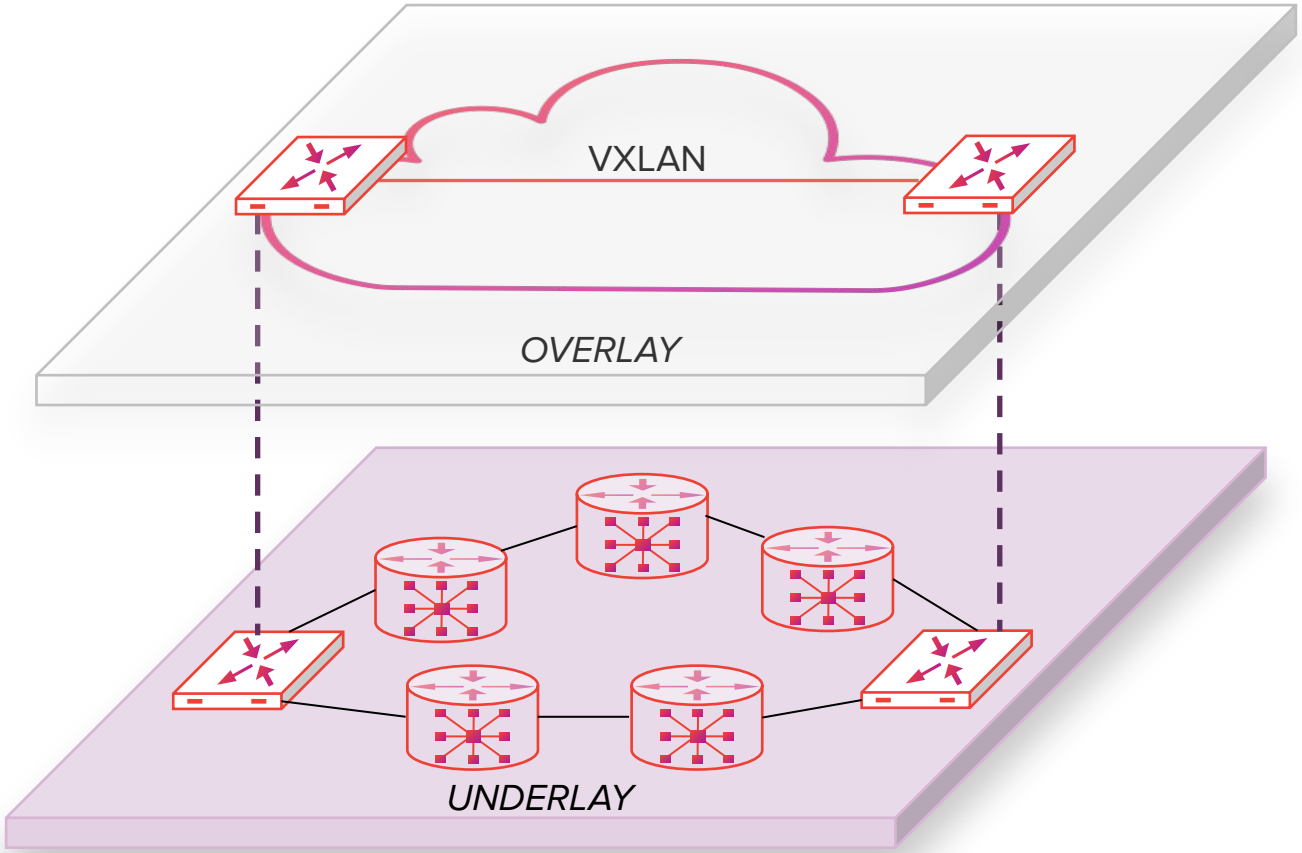
network is the **physical** network responsible for the delivery of packets like DWDM, L2, L3, or Internet, etc.

Overlay

is a **logical** network that uses **network virtualization** to build connectivity on top of physical infrastructure using **tunneling** encapsulations such as VXLAN, IPSec.

New paradigm:

You can have multiple overlay networks built on top of one underlay!


VXLAN
OVERLAY
UNDERLAY

VXLAN

Virtual Extensible Local Area Network is a protocol for running a Layer 2 network and stretching it **over a Layer 3 network**, which can be referred to as a VXLAN segment or tunnel by utilizing **MAC-in-UDP** encapsulation.

There can be created up to 16 million VXLANs on single physical (underlay) network due to 2^24 VNI (VXLAN Network Identifier) header length.

All of that leads to increased scalability.

VXLAN header encapsulated within UDP header

14B -18B

6	DST MAC
6	SRC MAC
4	VLAN (OPTIONAL)
2	ETH TYPE 0x8000

20B

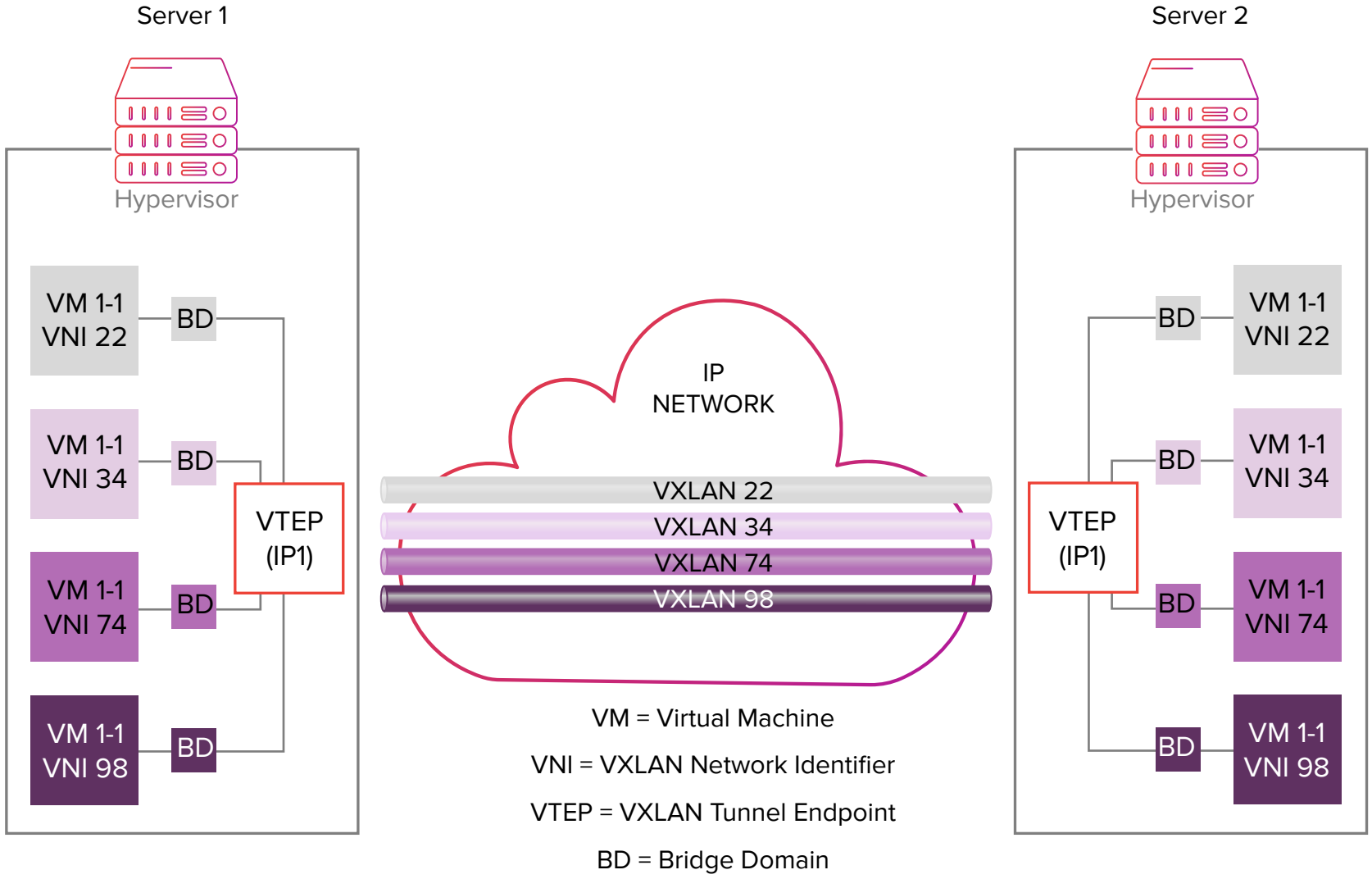
9	IP HDR DATA
1	PROTO: UDP
2	CHSUM
4	SRC IP: MY VTEP
4	DST IP: DST VTEP

8B

2	SRC PORT
2	VXLAN PORT
2	UDP Length
2	CKSUM 0x0000

8B

1	FLAGS RRRR1RRR
3	RESERVED
3	VNI
1	RESERVED


Server 1
Hypervisor
VM 1-1
VNI 22
BD
VM 1-1
VNI 34
BD
VM 1-1
VNI 74
BD
VM 1-1
VNI 98
BD
VTEP
(IP1)
IP
NETWORK
VXLAN 22
VXLAN 34
VXLAN 74
VXLAN 98
Server 2
Hypervisor
BD
VM 1-1
VNI 22
BD
VM 1-1
VNI 34
VTEP
(IP1)
BD
VM 1-1
VNI 74
BD
VM 1-1
VNI 98
VM = Virtual Machine
VNI = VXLAN Network Identifier
VTEP = VXLAN Tunnel Endpoint
BD = Bridge Domain

VXLAN - benefits

- Network Segmentation/Multitenancy - the overlay networks are completely isolated.
- VXLAN can utilize Layer 3 routing - equal-cost multipath (ECMP).
- Major improvement over the VLAN Layer 2 - not affected by STP, 16 million VNIs vs 4096 VLAN IDs.

Network Disaggregation

There are technologies that allow networking software and hardware vendor decoupling.

The benefits of network disaggregation:

1

Hardware and software independence.

2

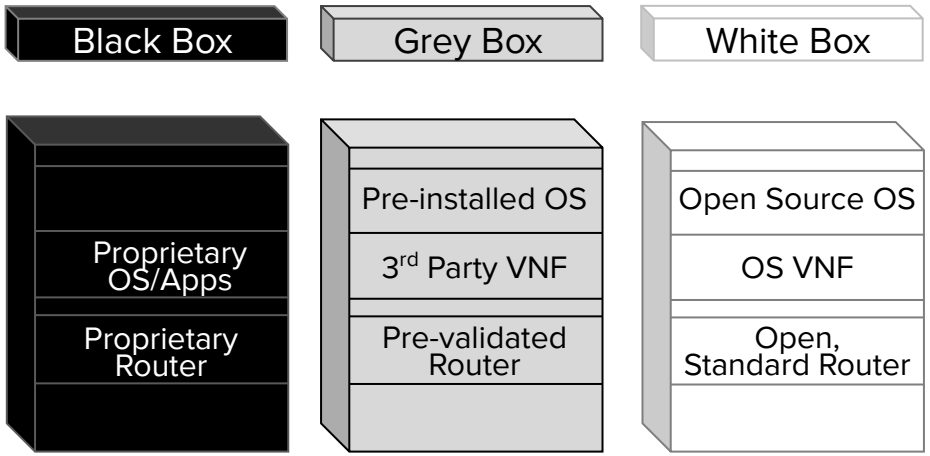
Increased scalability.

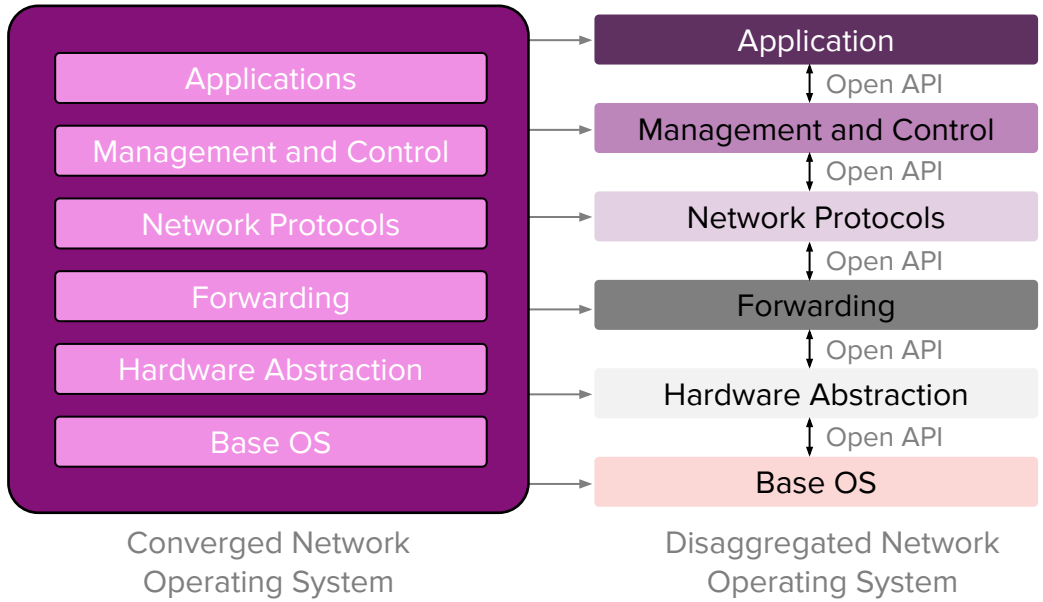
3

Open source.

4

Agility with a flexible management framework.


Black Box
Grey Box
White Box
Proprietary OS/Apps
Proprietary Router
Pre-installed OS
3rd Party VNF
Pre-validated Router
Open Source OS
OS VNF
Open, Standard Router


Applications
Management and Control
Network Protocols
Forwarding
Hardware Abstraction
Base OS
Converged Network Operating System
Application
Open API
Management and Control
Open API
Network Protocols
Open API
Forwarding
Open API
Hardware Abstraction
Open API
Base OS
Disaggregated Network Operating System

SONiC Network Operating System

At RTB House we use the SoNiC® operating system
for disaggregation purposes:
https://sonic-net.github.io/SONiC/

Features:

1

All configuration is done via **JSON, REST, CLI**

2

Or via declarative management tools

3

Unified, friendly syntax

4

One JSON to rule all networking configuration! ;)

SONiC Software for Open Networking in the Cloud

configuration and managment tools

Jenkins

ANSIBLE

kubernetes

SWARM

puppet

CHEF

1st party

New

More apps

SNMP

BGP

DHCP

IPv6

New

SONiC

Database

Platform

SWSS

Utility

TeamD

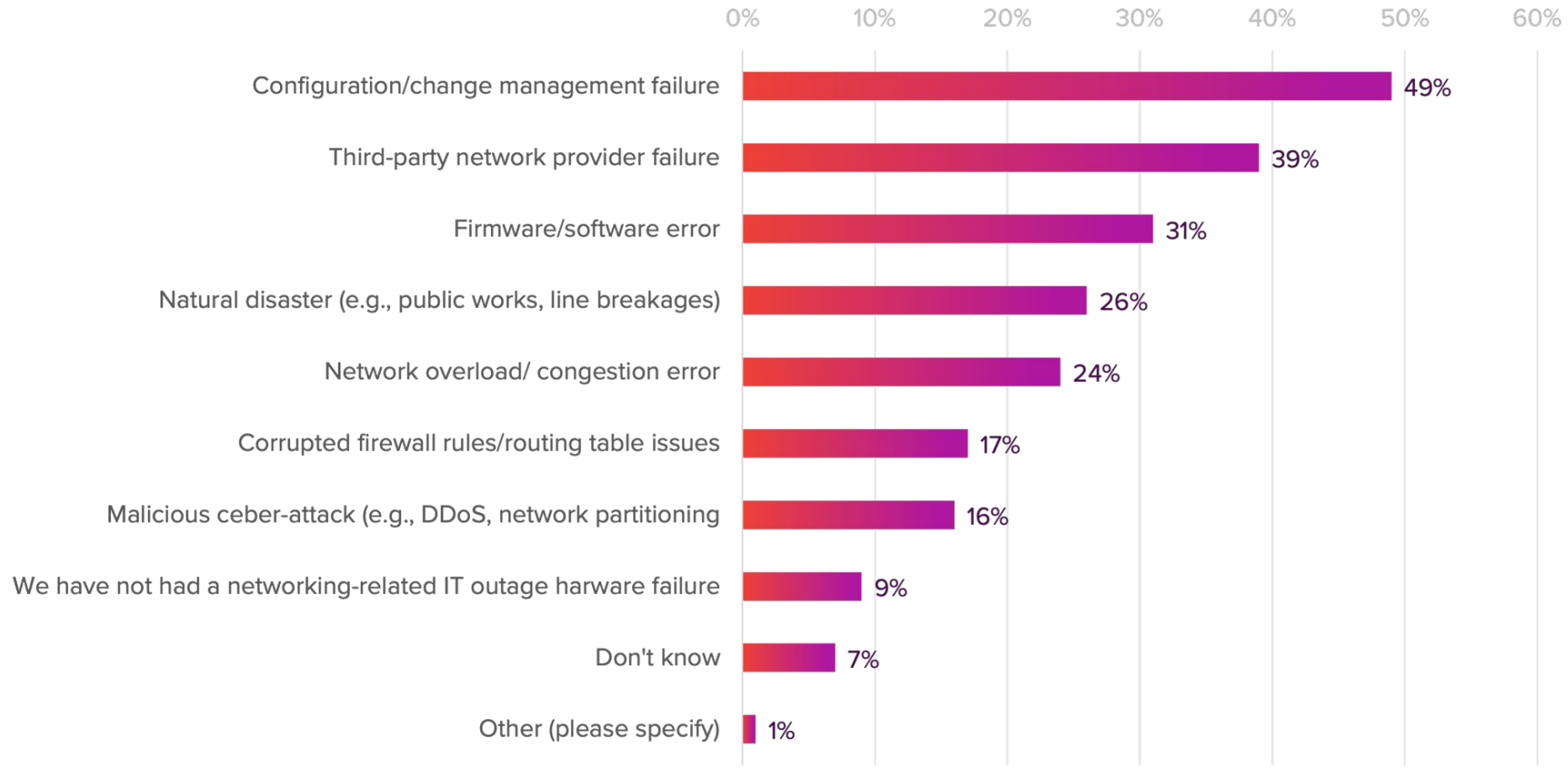
LLDP

RedisDB

SYNCD

Linux

Switch Abstraction Interface (SAI)



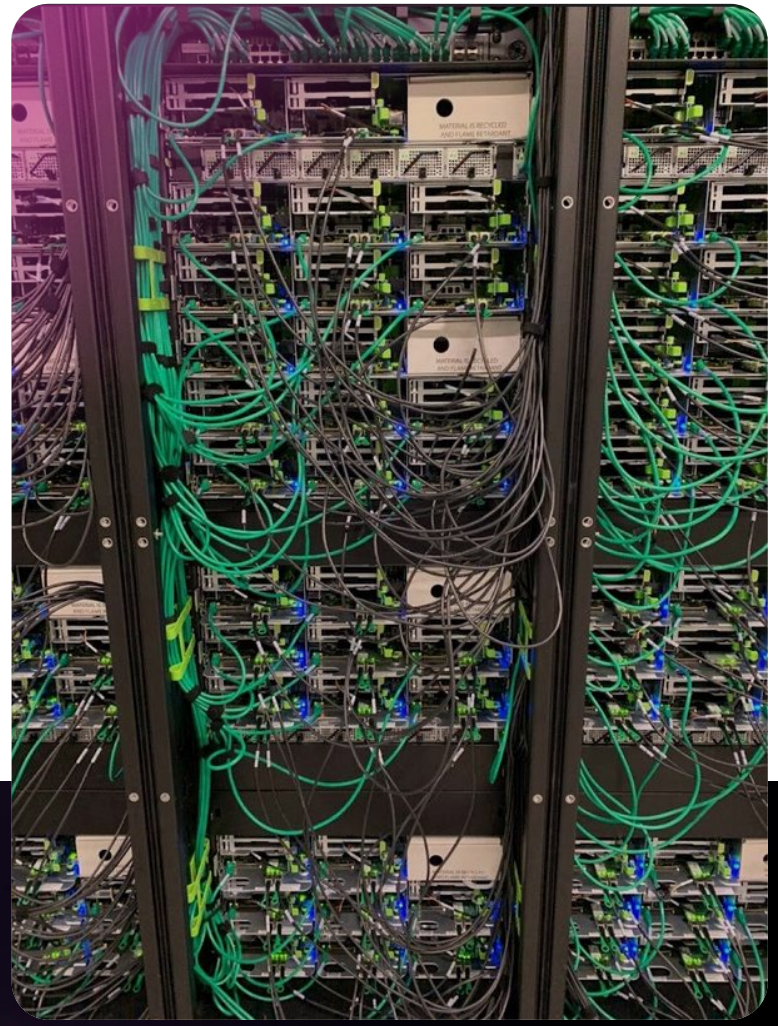
From the following list, which are the most common root causes for the networking-related IT outages at your organization's data centers over the past three years (Choose no more than three)

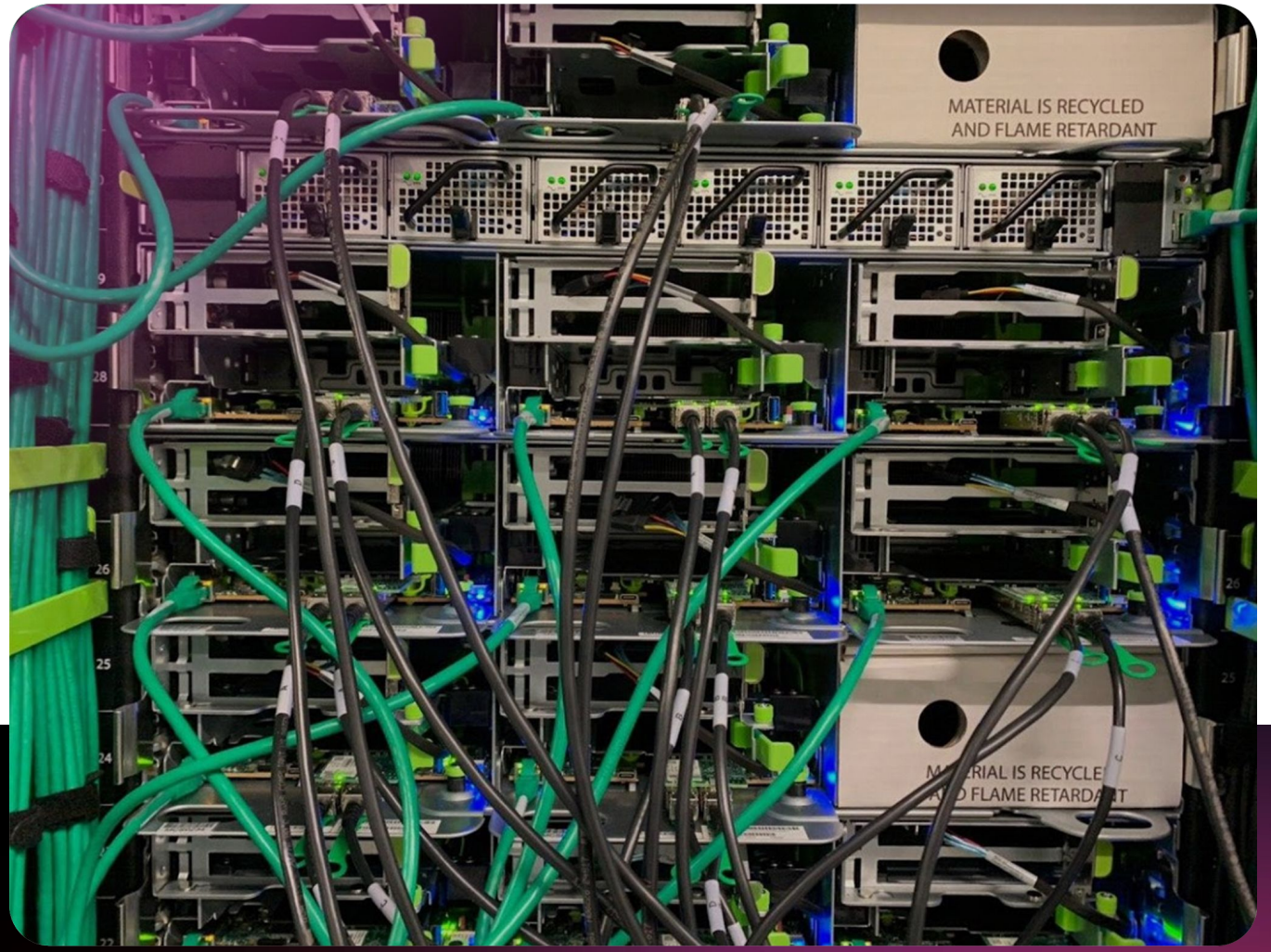
Source: Uptime Institute Data Center Resiliency survey – January 2021 (n=153)

Uptime Institute® INTELLIGENCE

RTB HOUSE

And now some photos from the RTB House Data Center




MATERIAL IS RECYCLED
AND FLAME RETARDANT

RTB HOUSE

RTB HOUSE

https://www.youtube.com/watch?v=Xi_niFopp8o

RTB HOUSE

Thank you.

Piotr Jaczewski,
Piotr Kowalczyk