クラウドコンピューティング基礎論

第6回

創造情報•小林克志

ikob@acm.org

マイクロソフト、開発者向け共有サイト買収　8200億円

2018/6/5 0:45 | 日本経済新聞　電子版

【シリコンバレー＝白石武志】米マイクロソフト（ＭＳ）は４日、ソフト開発者が設計図（ソースコード）を公開・共有できるサイトを運営する米ギットハブを75億ドル（約8200億円）で買収すると発表した。ＭＳは世界で約2800万人が利用するギットハブを取り込むことで、ソフト開発者向けのクラウドサービス事業を強化する。

ギットハブは2008年の設立。スマートフォンの普及などとともに、無償公開し自由に改良できる「オープンソースソフトウエア」の開発や普及の基盤になってきた。開発者同士が情報をやりとりするソーシャル・ネットワーキング・サービス（ＳＮＳ）の機能も果たしている。

買収手続きは米国や欧州の独禁当局の認可を前提に18年中に完了する予定で、ギットハブの新たな最高経営責任者（ＣＥＯ）にはＭＳ副社長のナット・フリードマン氏が就任する。ＭＳは買収後もギットハブの利用者が米アマゾンウェブサービスや米グーグルなどのクラウドサービスも利用できるようにするなど、ギットハブの経営の独立性は維持するという。

Outline

- Administravia
- Quiz and homework review
- IEEE 802.3 LAN aka. Ethernet
  - Frame Format
  - Switch
  - Spanning Tree (Loop avoidance)
  - mapping service

Course Outline

- Administrivia
- Cloud computing
- Service reliability
- Scale-up / Scale-out
- Distributed data stores
- Global services
- Datacenter networkings (1)
- Datacenter networkings (2)
- Network performance
- User experiences
- Network latencies
- Advanced topics

Class Information

- Provided by Web page:
  http://www.ci.i.u-tokyo.ac.jp/~ikob/lecture/2018-fcloud
  - Includes report submissions/roll calls/materials.
  - An authorization is required for access:
    User: cloud
    Pass: cloud!2018

Today’s quiz

- You want to lure, or invite big datacenter to the city of your birth*.
  Tell the technical advantages of your city for datacenter business.

- Bonus points will be given excellent answers which convince me even with a difficult location.
  (*) If many datacenters are already working in your city, you can adopt “Naha, Okinawa” instead.

- Submit your answers in Japanese or in English via the course web.

本日のクイズ

- 自身の出生地に巨大データセンターを誘致したい*。その都市の技術的優位性を示せ。
- 困難な立地で説得力のある回答には加点する

(*) すでに多くのデータセンターが自身の出生地で稼働している場合は、「那覇、沖縄」に代えても良い。

- 講義 Web フォームから記入すること。

Wikipedia より

Today’s Assignment

- Amazon Web Services provides a lot kind of database services.

1. Read documents for “AWS RDS Multi-Availability Zone(Multi-AZ) Deployment”. Tell the Multi-AZ advantages over ordinary RDS. In addition, explain the consideration requirements on Multi-AZ when designing services.
   - https://aws.amazon.com/rds/details/multi-az/
2. Pick one AWS database service other than RDS. Discuss pros and cons of it against RDS Multi-AZ deployment from the viewpoint of Brewer’s CAP theorem.

- Submit your answers in Japanese or in English via the course web.

本日の課題

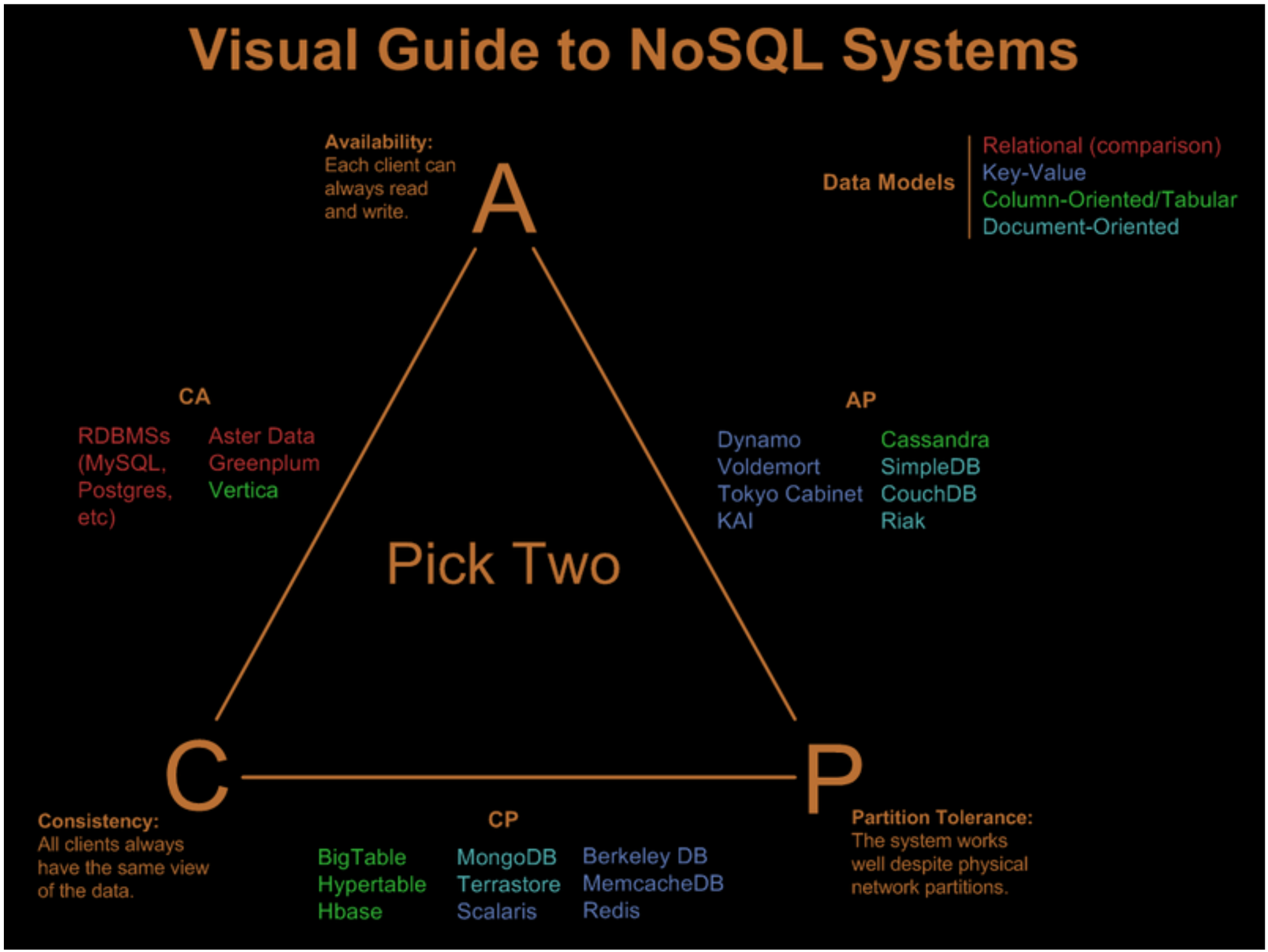
•Amazon Web Service (AWS)では多数のデータベースサービスを提供している。

1.AWS RDS Multi-Availability Zone(Multi-AZ) Deployment の技術文書を読み、AWS RDS Multi-AZ の通常の AWS RDS に対する優位点、サービス設計に当たって考慮すべき点を述べよ。

•https://aws.amazon.com/rds/details/multi-az/

2.AWS が提供する RDS 以外のデータベースサービスを一つ選択せよ。これと AWS RDS Multi-AZ の優劣を Brewer の CAP 定理の観点から議論せよ。

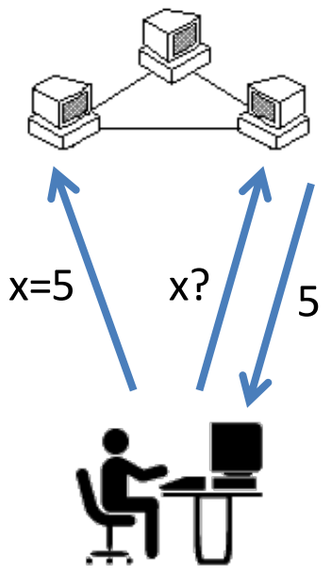
•講義 Web フォームから記入すること。



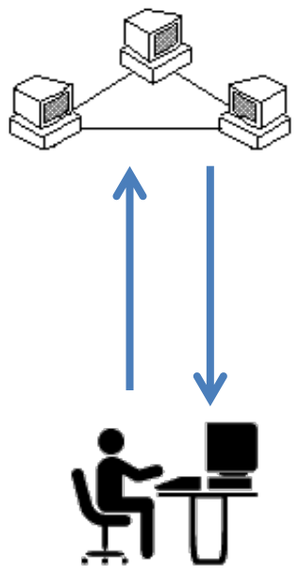
Nathan Hurst, “Visual Guide to NoSQL Systems”, 2010

CAP Theorem

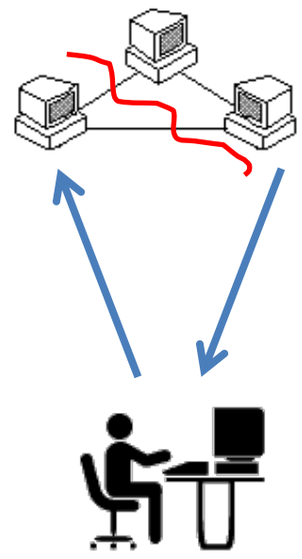
Consistency



Availability

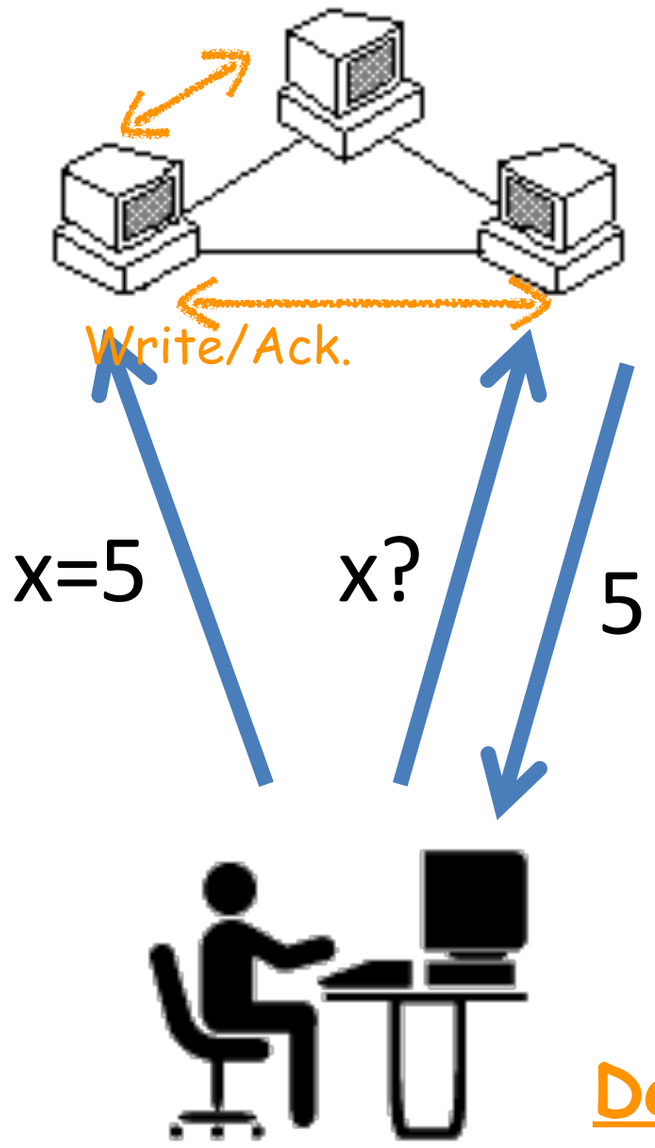


Partition tolerance



Consistency

+ Network Latency

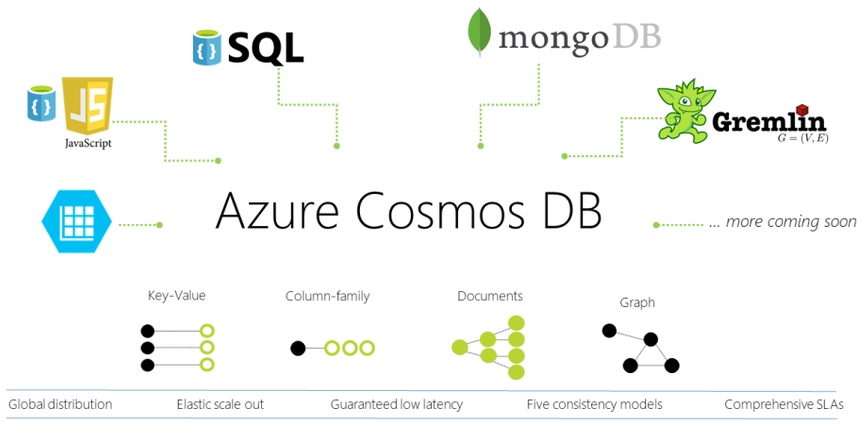


Delay due to network even in not partitioned.

remove an existing region or take a region that was previously associated with their database account offline.

Multi-model, multi-API support

Azure Cosmos DB natively supports multiple data models including documents, key-value, graph, and column-family. The core content-model of Cosmos DB's database engine is based on atom-record-sequence (ARS). Atoms consist of a small set of primitive types like string, bool, and number. Records are structs composed of these types. Sequences are arrays consisting of atoms, records, or sequences.



A globally-distributed, multi-model database service

The database engine can efficiently translate and project different data models onto the ARS-based data model. The core data model of Cosmos DB is natively accessible from dynamically typed programming languages and can be exposed as-is as JSON.

The service also supports popular database APIs for data access and querying. Cosmos DB's database engine currently supports DocumentDB SQL, MongoDB, Azure Tables (preview), and Gremlin (preview). You can continue to build applications using popular OSS APIs and get all the benefits of a battle-tested and fully managed, globally distributed database service.

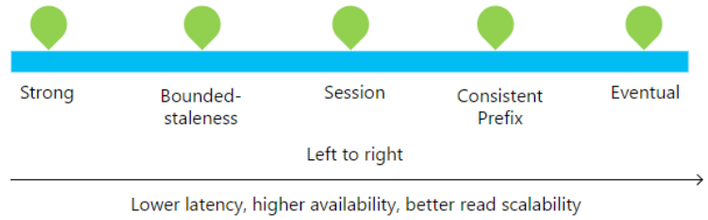
Horizontal scaling of storage and throughput

All the data within a Cosmos DB container (for example, a document collection, table, or graph) is horizontally partitioned and transparently managed by resource partitions. A resource partition is a consistent and highly available container of data partitioned by a customer specified partition-key. It provides a single system image for a set of resources it manages and is a fundamental unit of scalability and distribution. Cosmos DB is designed to let you elastically scale throughput based on the application traffic patterns across different geographical regions to support fluctuating workloads varying both by geography and time. The service manages the partitions transparently without compromising the availability, consistency, latency, or throughput of a Cosmos DB container.

Multiple, well-defined consistency models

Commercial distributed databases fall into two categories: databases that do not offer well-defined, provable consistency choices at all, and databases which offer two extreme programmability choices (strong vs. eventual consistency). The former burdens application developers with minutia of their replication protocols and expects them to make difficult tradeoffs between consistency, availability, latency, and throughput. The latter puts a pressure to choose one of the two extremes. Despite the abundance of research and proposals for more than 50 consistency models, the distributed database community has not been able to commercialize consistency levels beyond strong and eventual consistency.

Cosmos DB allows you to choose between five well-defined consistency models along the consistency spectrum – strong, bounded staleness, session, consistent prefix, and eventual.



The following table illustrates the specific guarantees each consistency level provides.

Consistency Levels and guarantees

CONSISTENCY LEVEL	GUARANTEES
Strong	Linearizability
Bounded Staleness	Consistent Prefix. Reads lag behind writes by k prefixes or t interval
Session	Consistent Prefix. Monotonic reads, monotonic writes, read-your-writes, write-follows-reads
Consistent Prefix	Updates returned are some prefix of all the updates, with no gaps
Eventual	Out of order reads

You can configure the default consistency level on your Cosmos DB account (and later override the consistency on a specific read request). Internally, the default consistency level applies to data within the partition sets which may be span regions.

Guaranteed service level agreements

Cosmos DB is the first managed database service to offer 99.99% SLA guarantees for availability, throughput, low latency, and consistency.

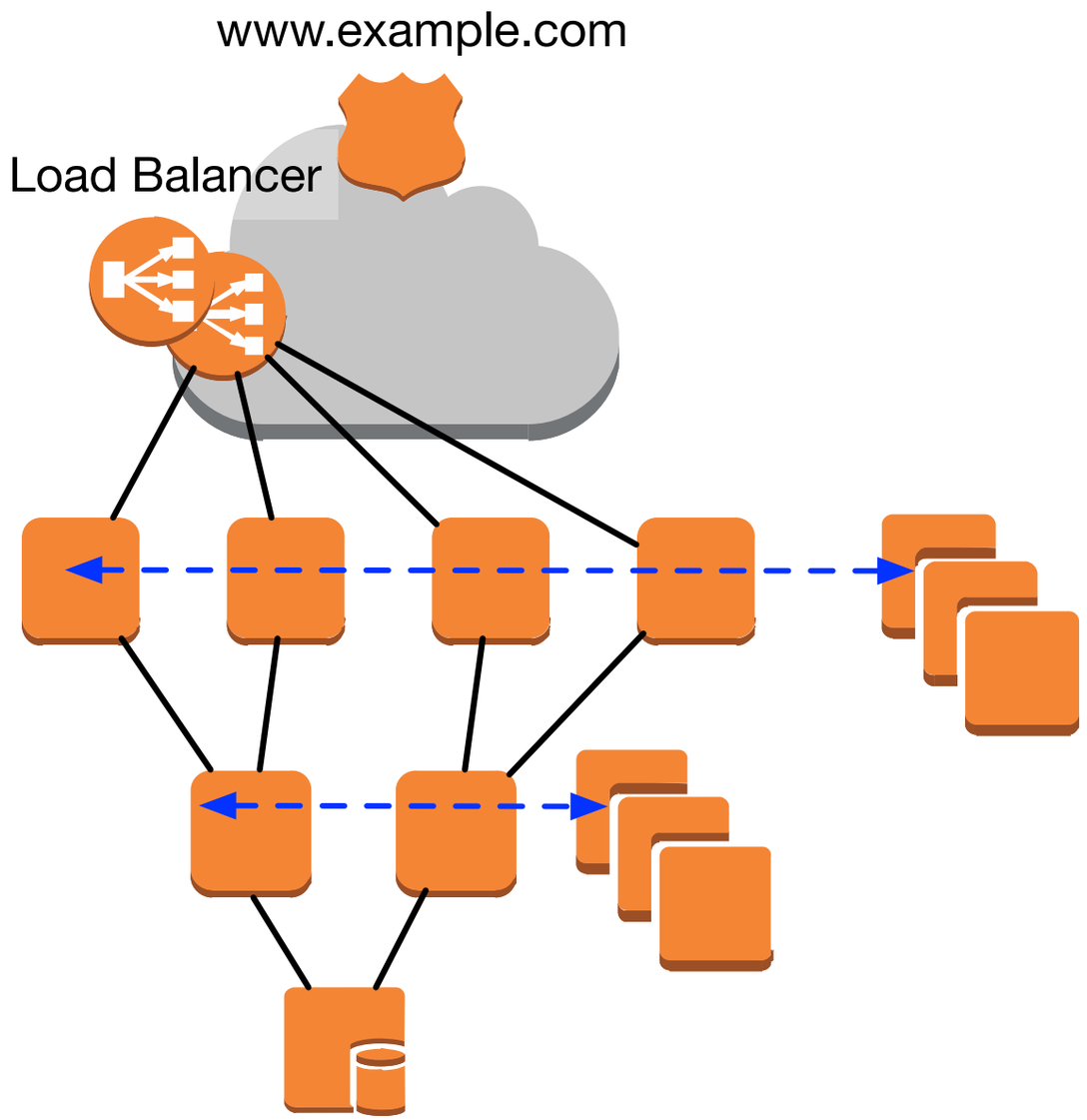
- Availability: 99.99% uptime availability SLA for each of the data and control plane operations.
- Throughput: 99.99% of requests complete successfully
- Latency: 99.99% of <10 ms latencies at the 99th percentile
- Consistency: 100% of read requests will meet the consistency guarantee for the consistency level requested by you.

Outline

- Administravia
- Quiz and homework review
- IEEE 802.3 LAN aka. Ethernet
  - Frame Format
  - Switch
  - Spanning Tree (Loop avoidance)
  - mapping service

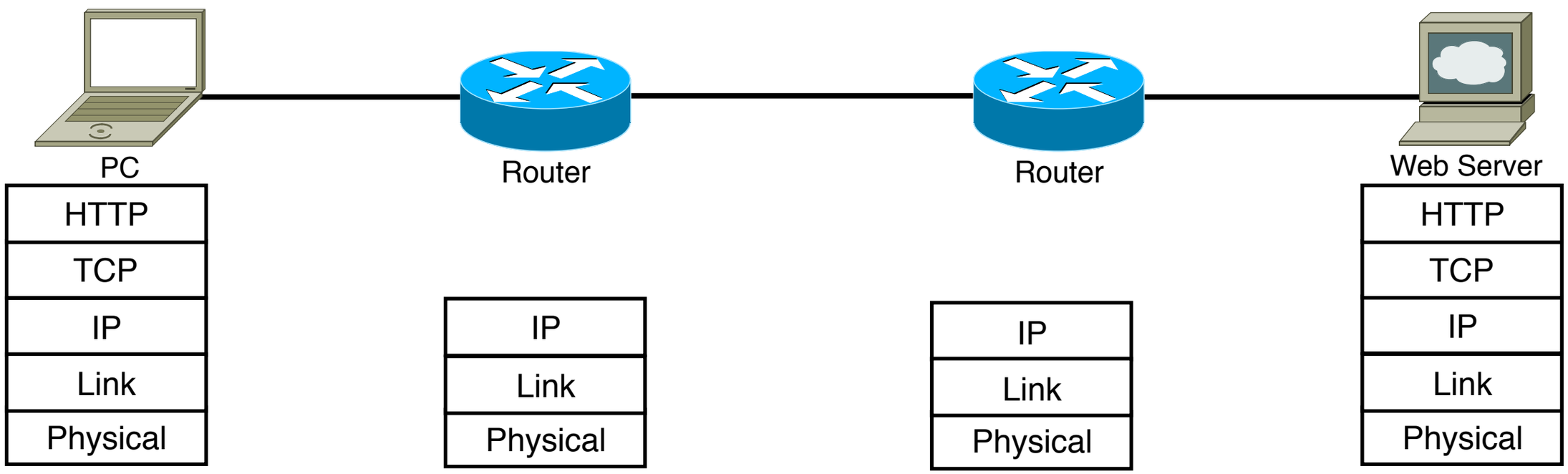
3-tier Web hosting architecture

- Goal: Improve scalability and reliability within affordable costs
- Pros:
  - Scale-out servers at Web and Application-tier
  - Isolate Internet / DB access.
- Cons:
  - Complicated software development
  - Bottleneck at backend-DB, or datastore.

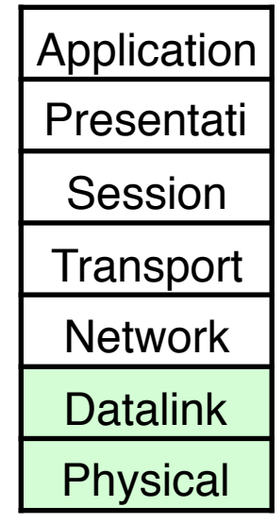


Internet Architecture

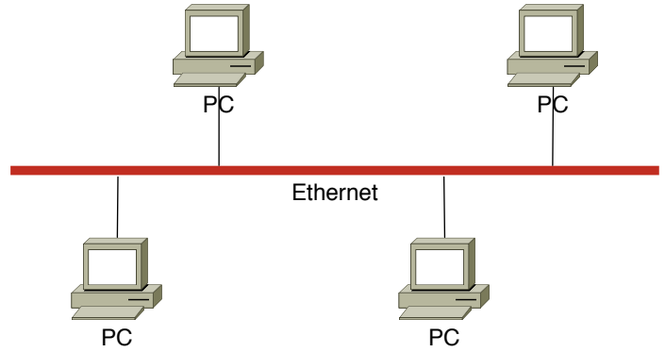
- End-to-End:Dumb networks and smart end systems.
  - Networks deliver packets in best-effort basis.
  - End systems take care other than it, e.g., reliability, sequence of data, error check, congestion avoidance.



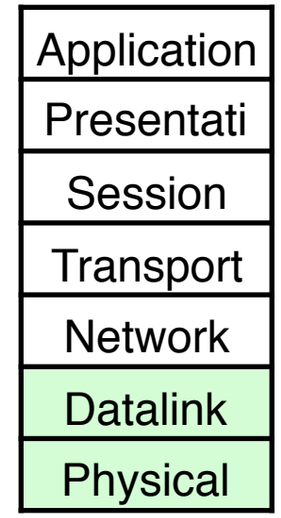
Ethernet (IEEE 802.3)



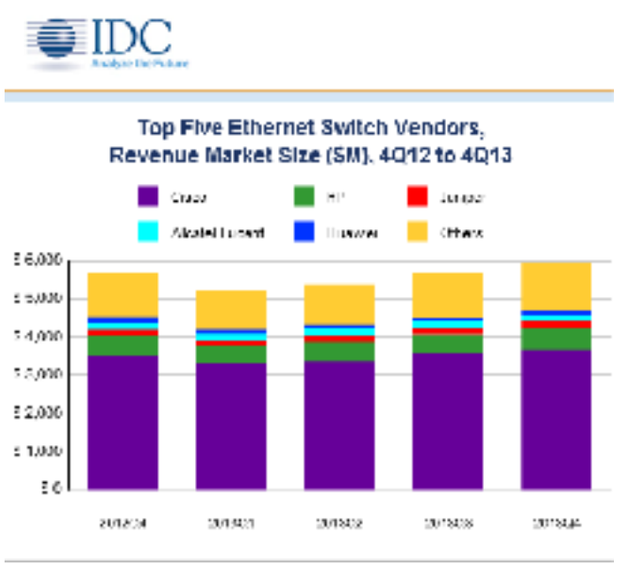
- Originally developed in 1970’ at XEROX PARC.
    - BUS type Local Area Network (LAN) protocol
      Broadcast (= Subnet) / Collision Domain
    - Career Sense Multiple Access with Collision Detection (CSMA/CD)
      designed for shared media, such as coaxial cable.
- IEEE 802 LAN/MAN Standards Committee
    - 802.1 LAN/MAN (802.1D Bridge, 802.1Q Virtual LAN (VLAN))
    - 802.3 Ethernet (Physical, MAC)
        - Coax : 10BASE5, 10BASE2
        - UTP(Unshielded Twisted Pair) : 1000BASE-TX, 10GBASE-T
        - SMF(Single Mode Fiber) : 10GBASE-LR, 100GBASE-LR4, LR10, ER4, 40GBASE-LR4
        - MMF(Multi Mode Fiber) : 1000BASE-SX, 100GBASE-SR10, 25GBASE-SR
        - Backplane : 10GBASE-KR
    - 802.11 Wireless LAN aka. WiFi



Ethernet in Today



- Composed of ethernet switch (SW) only, not of shared media
  - Shared media (CSMA/CD) support has dropped after 10GbE.
- Many SWs have L3 (IP) switching feature.
  - No significant difference between L2/L3 switching.
- L2/L3 (FDB/FIB) entries is limited and expensive.
  - ToR SW: 16k - 64k
  - Virtualization requires more entries.
- 6 billion USD revenue @ 4Q2013 (IDC), 3% growth
  - Server : 13 billion USD, - 4.1%



Today’s quiz

1. Tell your Laptop’s IPv4 address.

2. Tell your Laptop’s wireless MAC address.

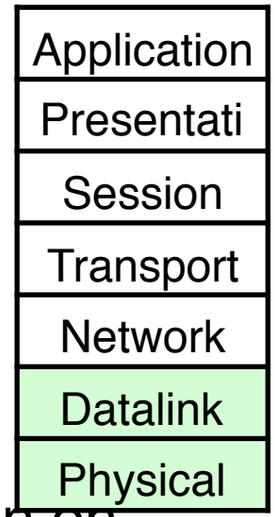
3. Tell the vendor of your wireless MAC.

- You can find them with:
  ifconfig for MacOS
  ipconfig for Windows
  ip address for Linux

- The list of assigned OUI is :
  http://standards-oui.ieee.org/oui/oui.txt

- Submit your answers in Japanese or in English via the course web.

Ethernet Frame Format



- While two types of encapsulation format, DIX and LLC-SNAP, DIX is common on IP
  - Jumbo frame, > 1,514 bytes frame size, is frequently used, but brings the risk of trouble
    - Must configure same frame size within the domain by manual operation.
  - Append additional field in case of 802.1Q VLAN tagging.

Ethernet2/DIX

8 Octets	1	6	6	2	46-1500 (or more)	4
preamble	SFD	DST addr.	SRC addr.	Type	data	FCS

Ethernet2/DIX w 802.1Q VLAN

8 Octets	1	6	6	2	2	2	46-1500 (or more)	4
preamble	SFD	DST addr.	SRC addr.	Type 0x8100	TCI	Type	data	FCS

3 bits	1	12
Priority	CFI	VLAN ID

Ethernet Addressing

- 6 octets MAC (Media Access Control) address format to identify node interface
    - Upper 3 octets : OUI (Organization Unique ID)
      Lower 3 octets : NIC (Network Interface Controller Specific)
      E.g., MAC address of NIC is “00:1b:21:a6:67:cd”, “00:1b:21” means NIC vended by Intel Corp.

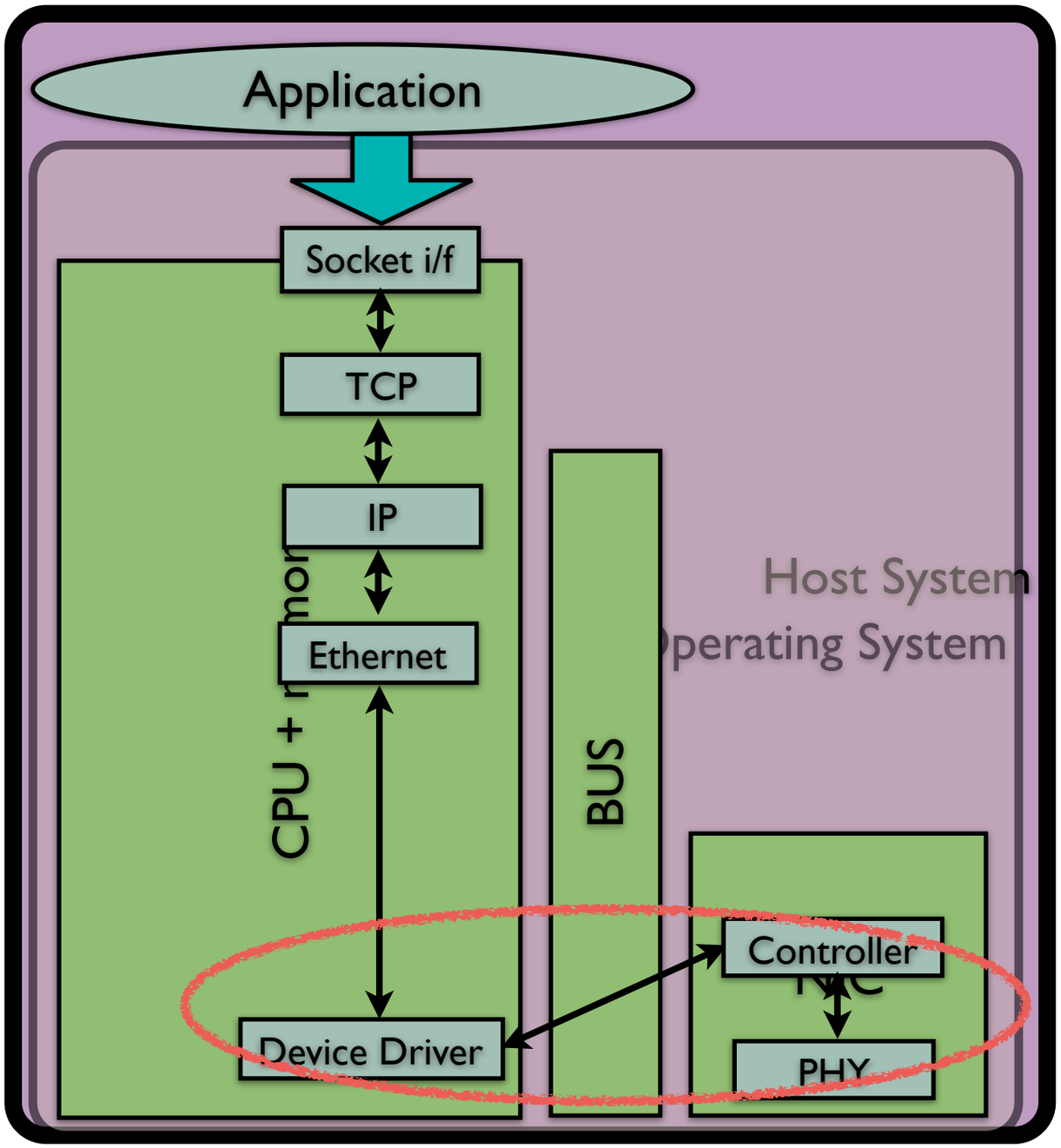
- Special addresses:
  Broadcast : FF:FF:FF:FF:FF
  (Site) Local : 0x02 bit of first octet is ON
  Multicast : 0x01 bit is ON

- Note: IP multicast address is mapped to specific MAC address

8 Octets	1	6	6	46 - 1500 or more	4
preamble	SFD	DST addr.	SRC addr.	Header + Data	FCS

Ethernet NIC and Host

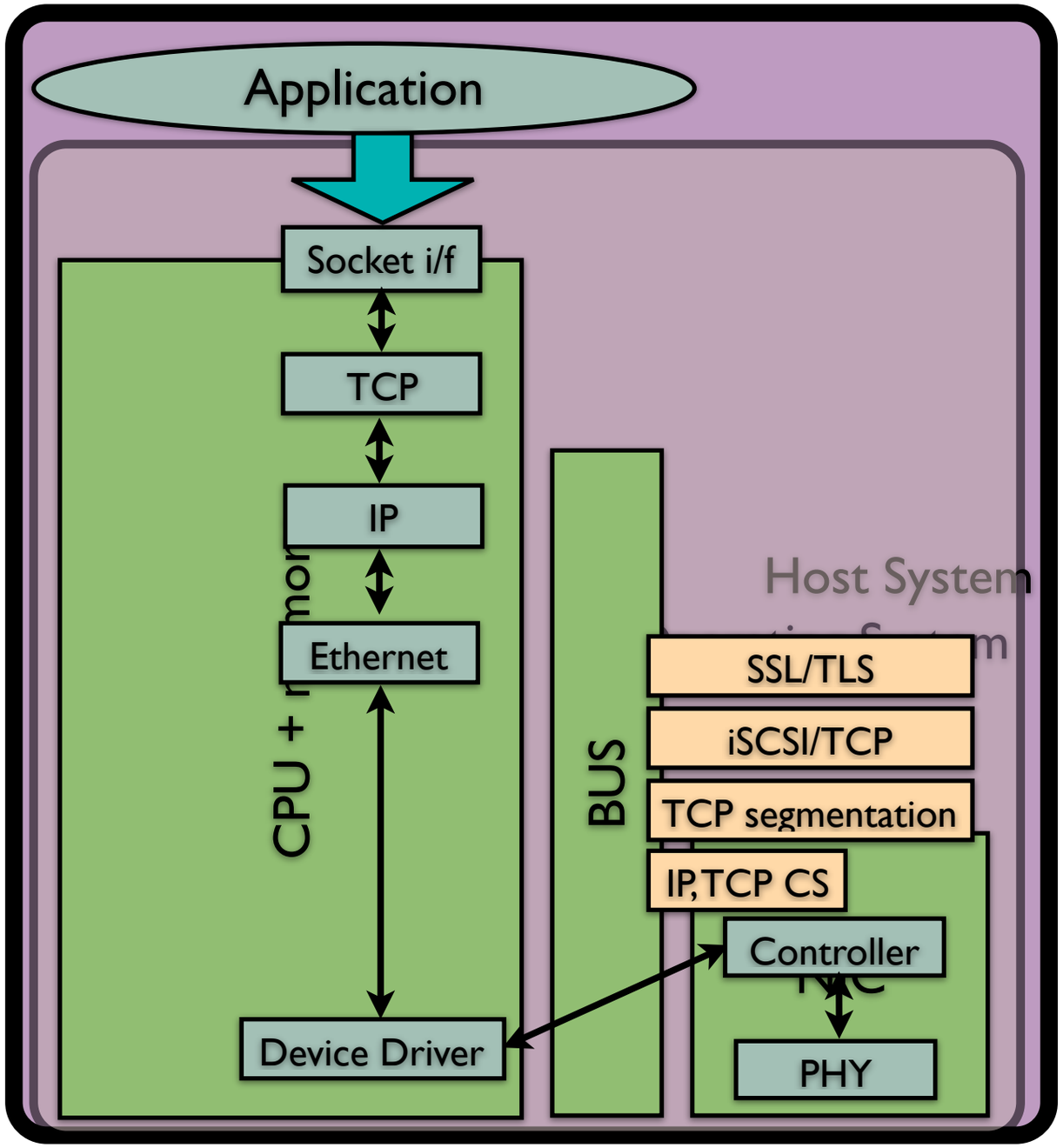
- NIC copy ethernet frame data between host memory and interface queue using DMA
  - OS aware completion status by NIC interrupts or by polling.
- Tx:
  - Scatter - Gather DMA eliminate memory copy in case of fragmented frame, e.g., BSD mbuf
- Rx:
  - Only receive specific destination(s) and broadcast (all 1’s).
    - Own MAC address and registered multicast destinations.
  - All broadcast must be processed



Ethernet NIC (cont'd)

- On modern NIC, a lot off-load features are implemented into NIC, such as, checksum, segmentation, TCP, iSCSI, SSL/TLS

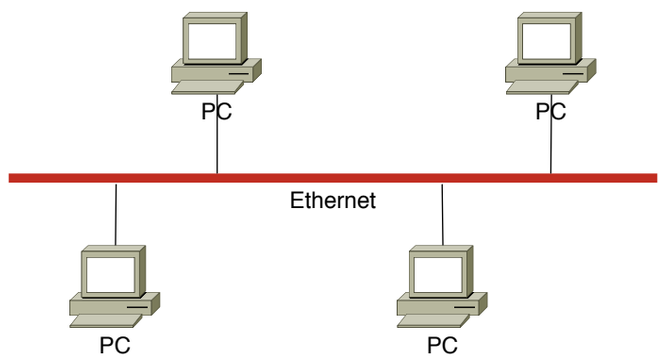
- ~~Layer-violated~~ Cross-layer features bring complicate protocol stack codes.



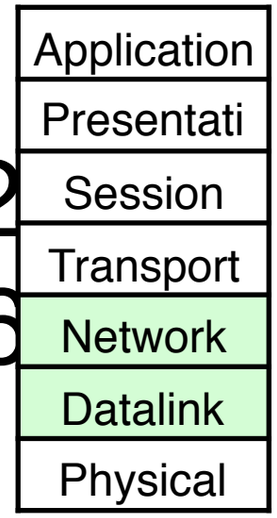
Ethernet Broadcast

Application
Presentati
Session
Transport
Network
Datalink
Physical

- Pros:
  - Plug-and-play : IP ARP, L3 Dynamic Host Configuration Protocol (DHCP)
  - Flexible topology : Can extend network by just adding SW devices, which meets scale-out approaches
- Cons:
  - Processing costs of broadcast frames are critical, in case of large # of end terminals.
  - Potential of broadcast storm :
    Avoids with Spanning Tree Protocol, Network Design



Address Resolution Protocol (ARP: RFC82
Neighbor Discovery Protocol (ND: RFC486

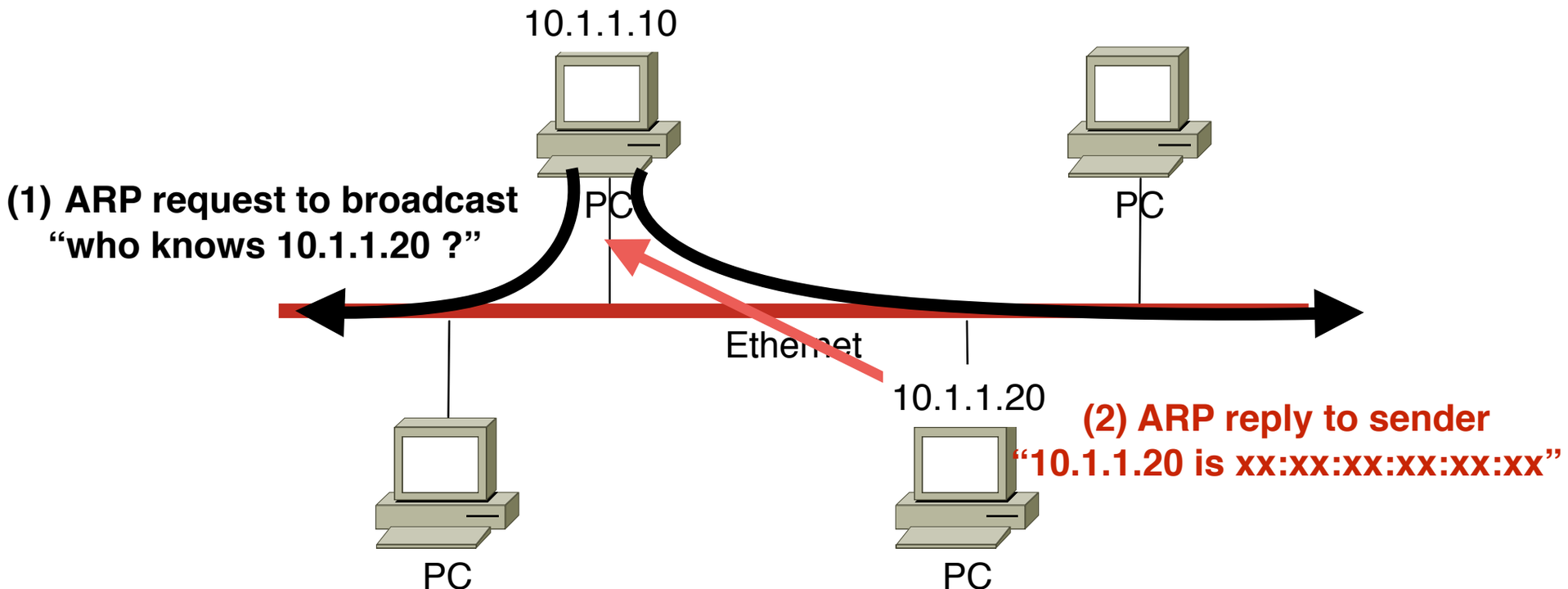


- Even if a host knows the destination IP,
  the host does not know the destination MAC address.

- ARP/ND is used to Resolve of MAC address from IP

1. Send ARP/ND request incl. own/target IP address using broadcast/multicast.
   - IPv4 : Dest = `ff:ff:ff:ff:ff:ff` (Broadcast addr.)
     IPv6 : Dest = `33:33:FF:22:22:22` (Multicast addr. assigned to ND purpose)

2. Target IP host replies ARP/ND response to requested node.



Outline

- Administravia
- Quiz and homework review
- IEEE 802.3 LAN aka. Ethernet
  - Frame Format
  - Switch
  - Spanning Tree (Loop avoidance)
  - mapping service

Ethernet Switch (Bridge)

Application
Presentati
Session
Transport
Network
Datalink
Physical

- Bridge / Switch ?
  - Bridges (port = 2) split collision domains
  - First SW product (port > 3) released by Kalpana in 1989.
    - VLAN provides more flexibility in network topology
- 802.1D specifies how Ethernet Bridge works as:
  - FDB (Filtering Database)
    - Learn MAC / port mapping.
    - Forward frames according the mapping.
  - STP (Spanning Tree Protocol)
    - Avoid loop topology.
- Cheap management efforts and flexible network topology

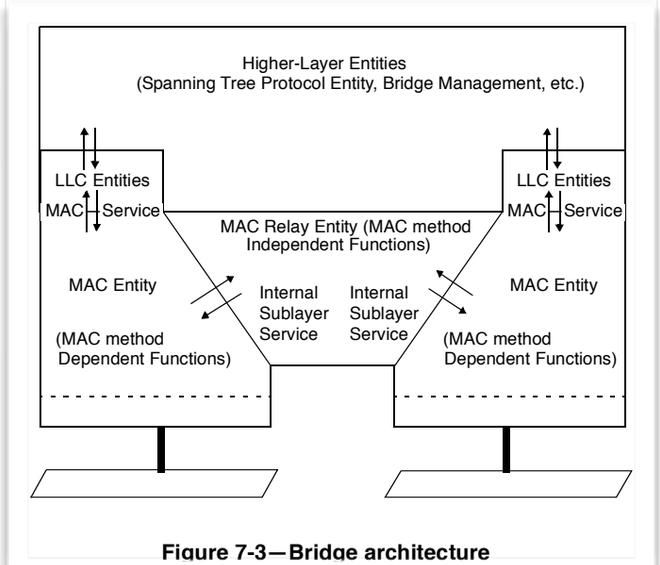


Figure 7-3—Bridge architecture

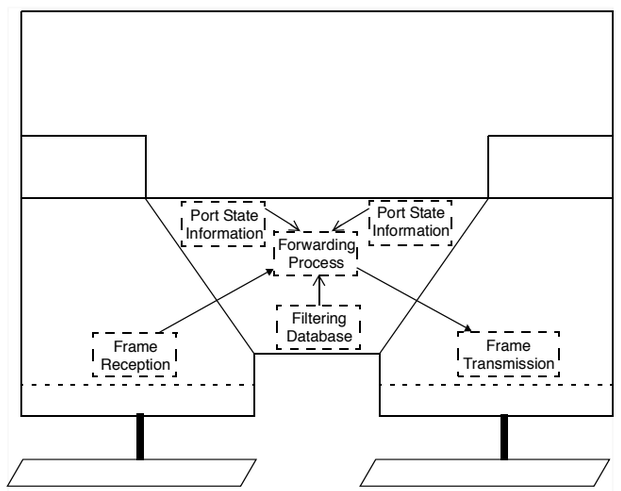
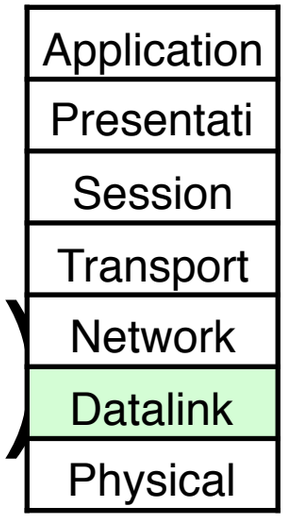


Figure 7-4—Relaying MAC frames

Ethernet Switch (Learning and Forwarding)



- When receiving a frame:

1. Create/update src. addr. and rx port map entry in FDB

2. If the frame is unicast, and FDB has dst. addr. entry, forward frame only to the mapped port.

3. Otherwise, forward it to all ports other than the receiving port.

4. FDB entries expires, if no update until aging time.

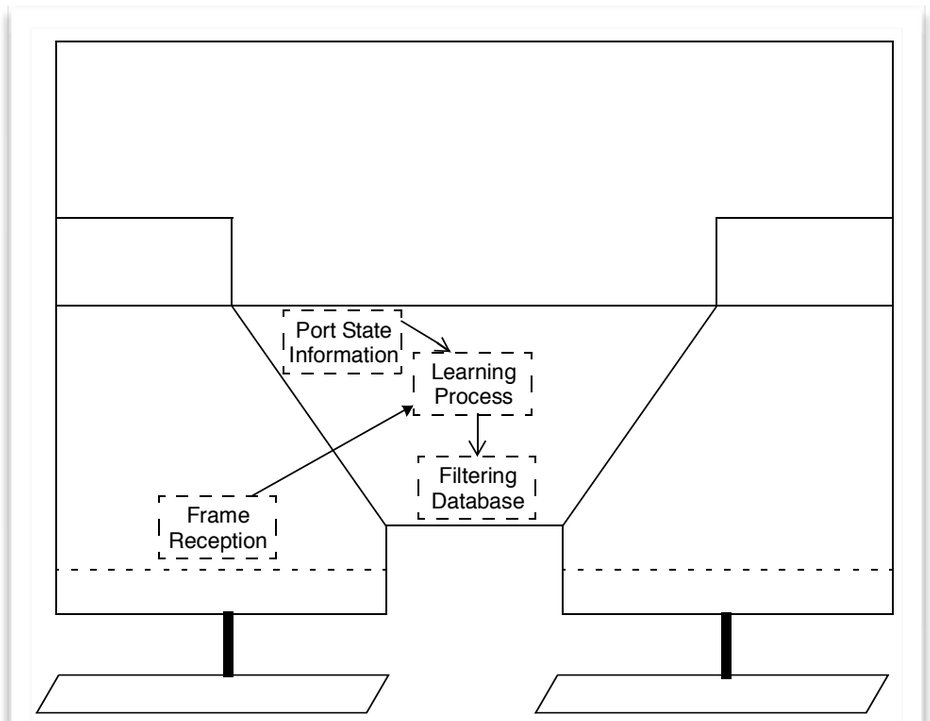
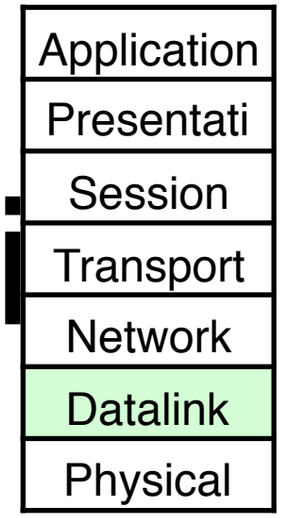


Figure 7-5—Observation of network traffic

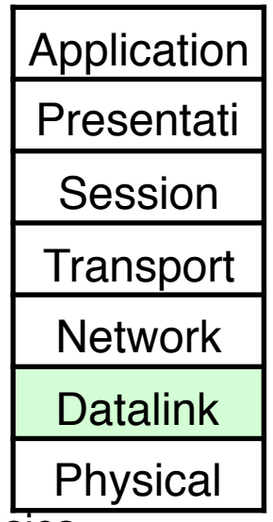
Impact of Broadcast Traffi



- Affect all nodes.
  - Nodes must process all broadcast packets.
    - Processing cost is not small in case of too many broadcasts.
  - 300,000/s ARP req./reply @ 50,000 node
    - CPU Load 27%@Linux, 100%@WInXP with 3GHz Xeon
- Optimize broadcast domain, or number of end terminals.
  - L2: Virtual LAN (VLAN) , L3: IP domain

Elmeleegy, Khaled, and Alan L. Cox. "Etherproxy: Scaling ethernet by suppressing broadcast traffic." *INFOCOM 2009, IEEE*. IEEE, 2009.

VLAN (Virtual LAN)

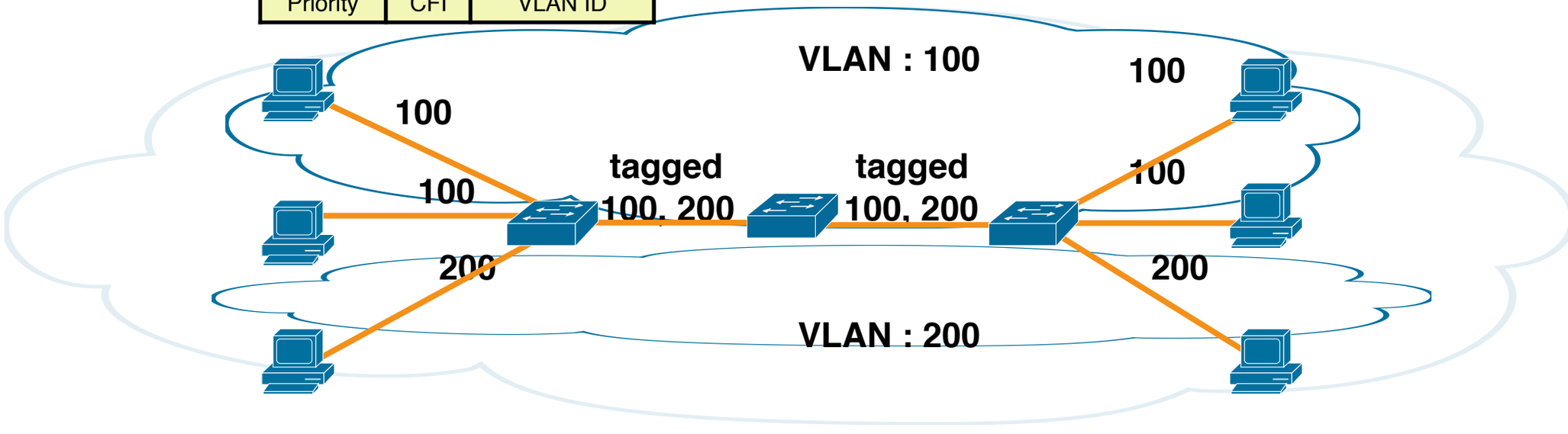


- Split network into more than one broadcast domains in order to mitigate broadcast traffic or to meet site policies.
- Two type of VLAN:
  - Tagged: Coexisting more than one VLANs using 802.1q VLAN ID (4,094 IDs)
  - Untagged: One VLAN for a physical media

Ethernet2/DIX w 802.1Q VLAN

8 Octets	1	6	6	2	2	2	46-1500 (or more)	4
preamble	SFD	DST addr.	SRC addr.	Type 0x8100	TCI	Type	data	FCS

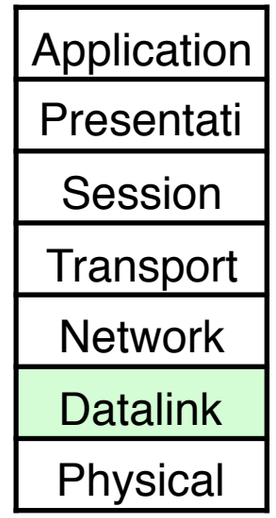
3 bits	1	12
Priority	CFI	VLAN ID



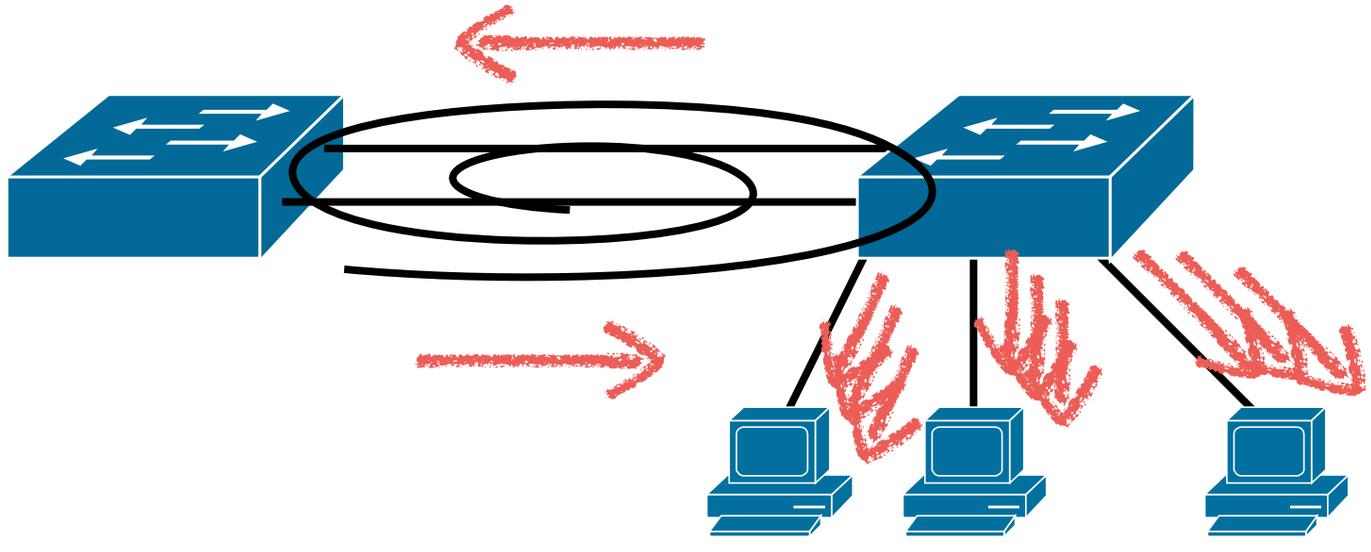
Outline

- Administravia
- Quiz and homework review
- IEEE 802.3 LAN aka. Ethernet
  - Frame Format
  - Switch
  - Spanning Tree (Loop avoidance)
  - Mapping service

L2 switch and loop



- The most common and critical Ethernet trouble.
  - Broadcast storm caused by L2 switch loop
  - Call center service (aka 119) trouble at Tokyo Fire Department in Jun. 5, 2011
    - Somebody connected an ethernet cable at open SW port. Then, it caused broadcast storm



日本経済新聞

東京消防庁の119番不具合、職員の人為ミスか

2011/1/7 21:26 | 日本経済新聞　電子版

東京都内で５日、災害救急情報システムに不具合が生じ、約４時間半にわたり119番通報がつながりにくくなった問題で、東京消防庁は７日、ＬＡＮケーブルの誤接続が原因だったと発表した。同庁職員の人為的ミスとみられ、同日記者会見した松浦和夫警防参事は「都民に大変な迷惑をかけ、不安を与えてしまい、申し訳ない」と謝罪した。

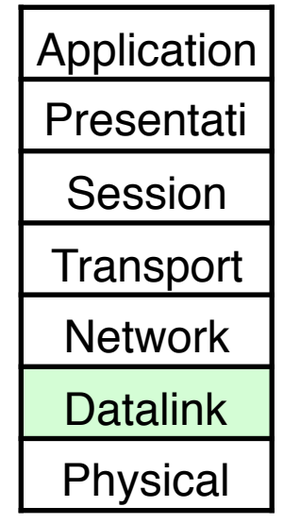
システム障害は千代田区の本庁舎内の災害救急情報センターで発生。同庁によると、ホストコンピューターと端末を結ぶ中継器にＬＡＮケーブルが誤接続されたため、過大なデータが流れて処理しきれなくなり、災害現場に応じて自動的に出動部隊を振り分ける機能が停止したという。

ＬＡＮケーブルは予備のもので、一方の端子だけ中継器に差し込まれていたが、職員が誤ってもう一方の端子を空いていた差し込み口につないだとみられる。誤接続の時期は不明で、職員も特定できていない。

同庁は再発防止策として、中継器の空いている差し込み口をふさいだり、ＬＡＮケーブルに「抜き差し厳禁」の印を付けたりしたほか、システムの改善も検討している。

障害は５日午前10時半ごろ発生。職員が手作業で出動可能な部隊を探し、出動命令を出したために時間がかかり、稲城市と島しょ部を除く都内全域で119番がつながりにくくなった。同庁は、人命に関わる影響は出ていないとしている。

Loop prevention and detection



- Build loop free topology :
  - Spanning Tree Protocol (STP:802.1D)
  - Rapid Spanning Tree (RSTP:802.1w)
  - Multiple Spanning Tree Protocol (MST:(802.1s)
  - Other vendor specific protocols (PVST, PVST+…)
- Loop detection (proprietary) :
  - When receiving probe frames (detecting loop), shutdown port / alert to operator.
- Storm control (proprietary):
  - When broadcast traffic exceed threshold, shutdown port.

802.1D Spanning Tree Protocol (STP)

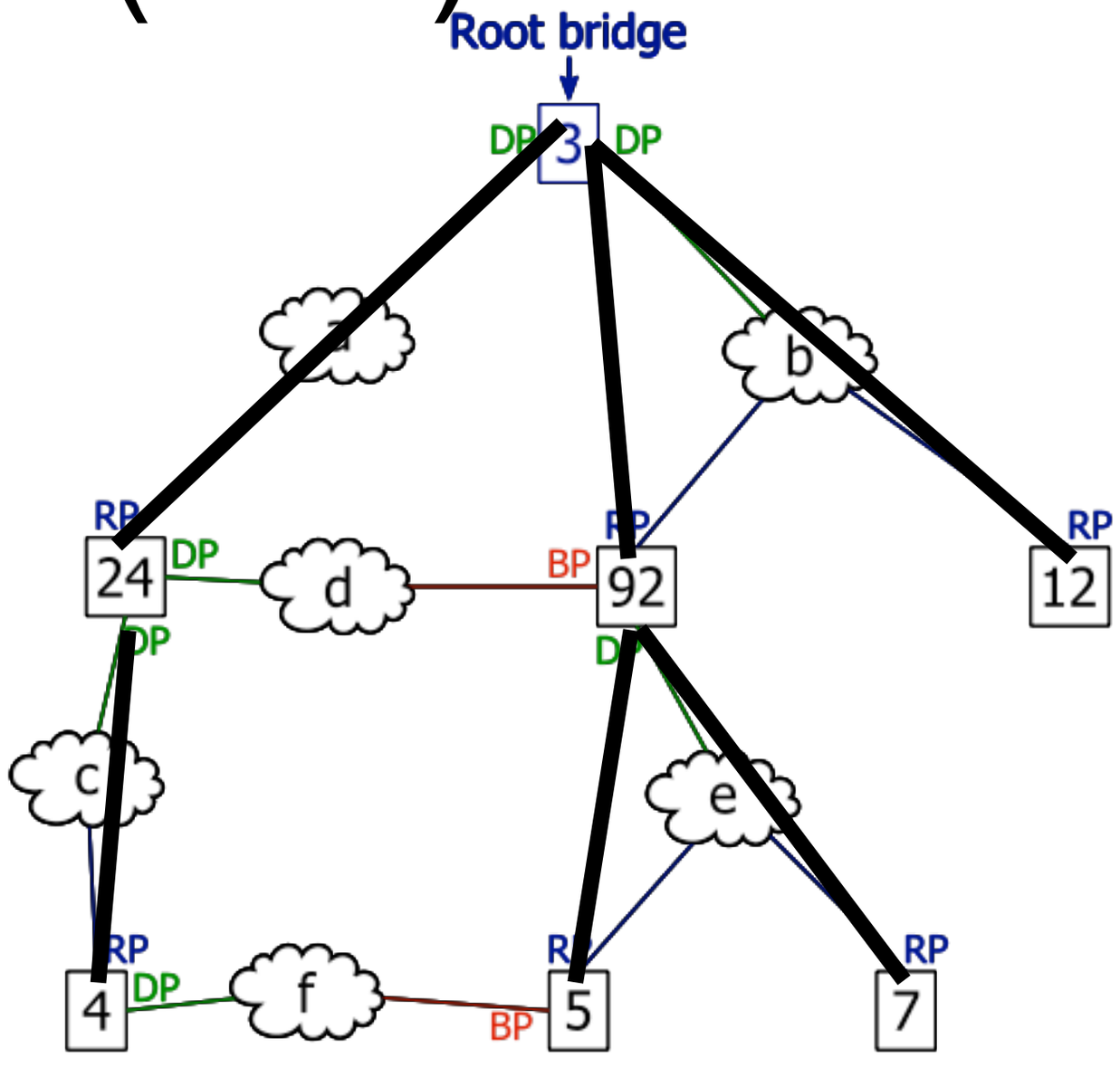
- Ensure loop free topology by closing redundant ports
  - Distributed algorithm exchanging BPDU (Bridge Protocol Data Unit) frames

1. Elect “root” bridge having minimum bridge ID (Priority + Bridge’s MAC)

2. Non-root bridges select Root Port (RP) which is most least cost to root.

3. Determine most least cost port in each segment as Designated Port (DP)

4. Close ports other than RP, DP as Blocking Port (BP)



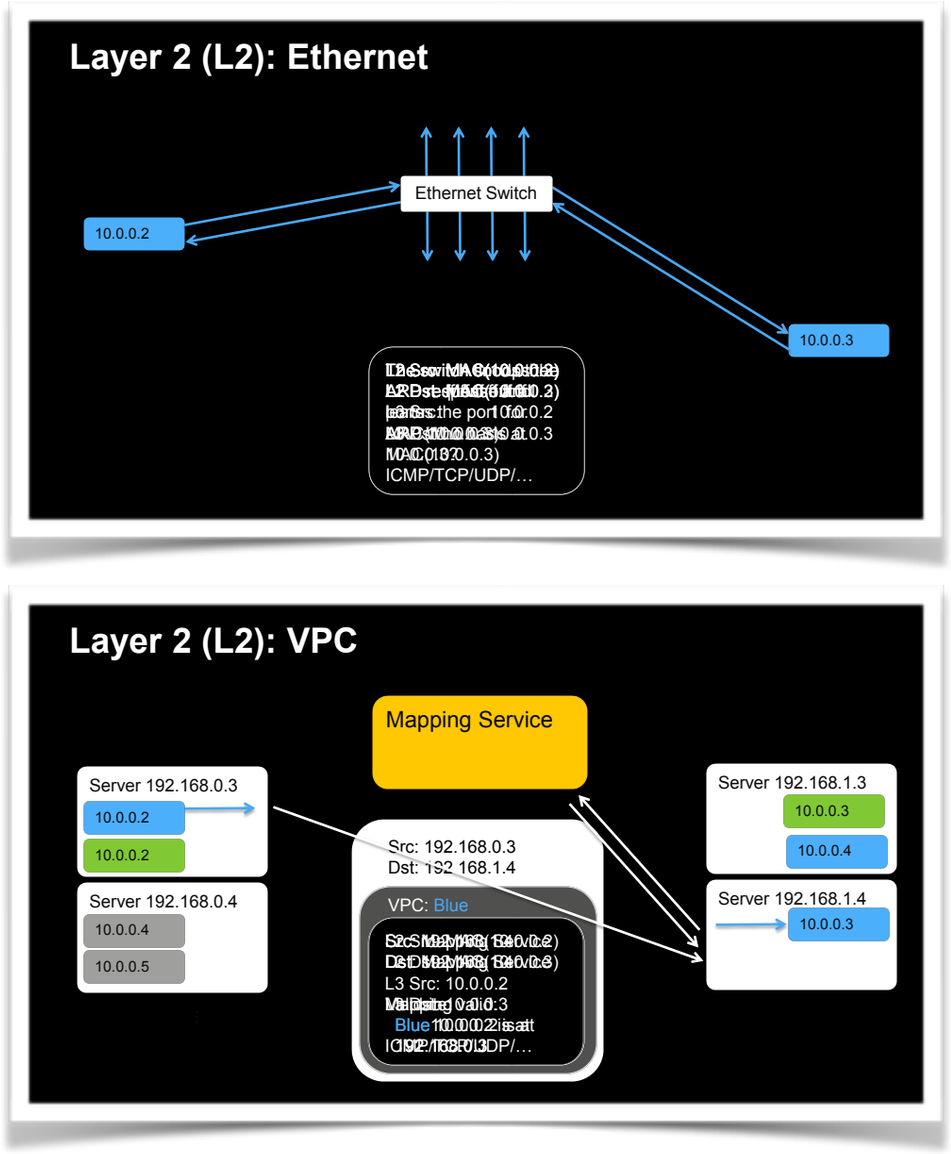
http://en.wikipedia.org/wiki/Spanning_Tree_Protocol より

Perlman, Radia. "An algorithm for distributed computation of a spanningtree in an extended LAN." *ACM SIGCOMM Computer Communication Review*. Vol. 15. No. 4. ACM, 1985.

Mapping Service in Production Cloud

Application
Presentati
Session
Transport
Network
Datalink
Physical

- In fact, broadcast / multicast traffic is not allowed in production cloud, such as, AWS.
  - Such traffic makes critical impacts on large networks.
  - Mapping service is used instead of broadcast ARP / multicast ND. But, the service is hidden such from OS, software.



Steve Mueller, “Another Day, Another Billion Packets”, AWS re:Invent 2016

Today’s assignment

- Read Microsoft Azure Cosmos DB technical document. Choose one consistency level other than “Strong” among five levels provided by Cosmos DB.

- Tell an application compatible with your chosen consistency level. Why ?

- Submit your answers in Japanese or in English via the course web.

本日の課題

- Microsoft Azure Cosmos DB の技術文書を読む。Cosmos DB が提供する5 つの整合性モデルのうち “Strong” 以外から一つを選択せよ。
- 選んだ整合性レベルに適合するアプリケーションを示せ。理由は？

- 講義 web から日本語か英語で回答すること。

- AWS service overview: http://aws.amazon.com/products/?nc2=h_ls
  or Wikipedia : http://en.wikipedia.org/wiki/Amazon_Web_Services

Today’s quiz

1. Tell your Laptop’s IPv4 address.

2. Tell your Laptop’s wireless MAC address.

3. Tell the vendor of your wireless MAC.

- You can find them with:
  ifconfig for MacOS
  ipconfig for Windows
  ip address for Linux

- The list of assigned OUI is :
  http://standards-oui.ieee.org/oui/oui.txt

- Submit your answers in Japanese or in English via the course web.

For hands-on exercise : Install two softwares

1.Wireshark : A packet capture and analyzer

- Just install package from http://www.wireshark.org

2.NS2 : Network simulator.

Three options are there:

A. Docker
   - Install docker software and take container:
     - https://github.com/ekiourk/docker-ns2
     - X-window configuration depends on OS.
B. Native application. If you are using Linux, use this option.
   - Install ns-allinone-2.35 from source because NS-2 package in some distribution may not work.
   - X-window, perl, gnuplot are also required.
C. Virtual Machine (VM)
   - Install Hypervisor Software.
     - Oracle VirtualBoX is free.
       vmware or others are also welcome.
   - Linux VM image with NS2 software will be available from the course Web.

演習に向けて：２つのソフトウェアをインストールする

1.パケットアナライザ wireshark

- http://www.wireshark.org を参考にインストールすること。

2.ネットワークシミュレータ NS2

A. Docker

- Docker をインストールし、コンテナを使用する。
- https://github.com/ekiourk/docker-ns2
- X-window の設定は OS に依存する。

B. VM で動作

- ハイパーバイザの導入
  - Oracle VirtualBoX であれば無償
    vmware 他でもかまわない
- NS2 付きの Linux 仮想マシンイメージを講義ページで配布する

C. Native

- Linux を利用している場合は、この方法を使う。
- ns-allinone-2.35 を install すること。
  - X-window, perl, gnuplot なども必要となる。