クラウドコンピューティング基礎論

第４回

創造情報•小林克志

ikob@acm.org

Outline

1. Administravia

2. Homework review

3. Scale-up / Scale-out

    1. CPU

    2. Storage

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

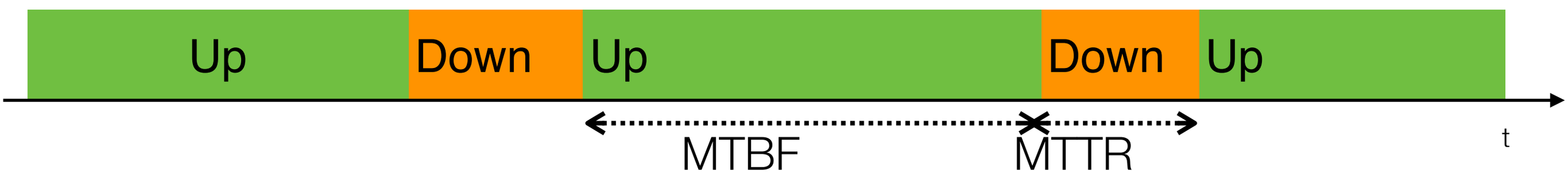
Indices of reliability

1. Classic definition, for predictions.
   A = MTBF / (MTBF + MTTR)
   Mean Time Between Failure (MTBF),Mean Time To Recover(MTTR)

2. Actual, to evaluate services, such for 24hr 7days service.
   A = Uptime / (Uptime + Downtime)

3. In fact, on Service Level Agreement (SLA) evaluation.
   A = Uptime / (Uptime + Unplanned Downtime)

   • excluding planned down time



Today’s Assignment

- MS provides Azure Cloud Services, an IaaS, which offers Service Level Agreement (SLA) as 99.95% uptime. This SLA is applied when customers deploy two or more role instances in different fault and upgrade domains.

  Your Web service deployed on MS Azure is required 99.95% availability as well as Cloud Services. Therefore, the service configuration is a redundant that the applications run on different domains’ instances.
  Each application program is expected service down for maintenance, such as, to apply security patch, at least one per month.

- How long downtime is allowed at the monthly maintenance without any critical impact(*) on the availability ?

(*) Discussion about “critical impact” is welcome.

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

本日の課題

- MS は Azure Cloud Services を 99.95% の可用性を SLA として提供している。この SLA の適用要件として、複数のドメインで 2つ以上のインスタンスの稼働が求められている。

- Azure 上に展開した Web サービスでは Cloud service と同等の可用性が求められている。
  したがって、2 つのアプリケーションプログラムを異なるドメインのインスタンスで稼働させる冗長構成とした。
  それぞれのプログラムではセキュリティパッチなどメンテナンスのため月間 1 回の停止が見込まれる。

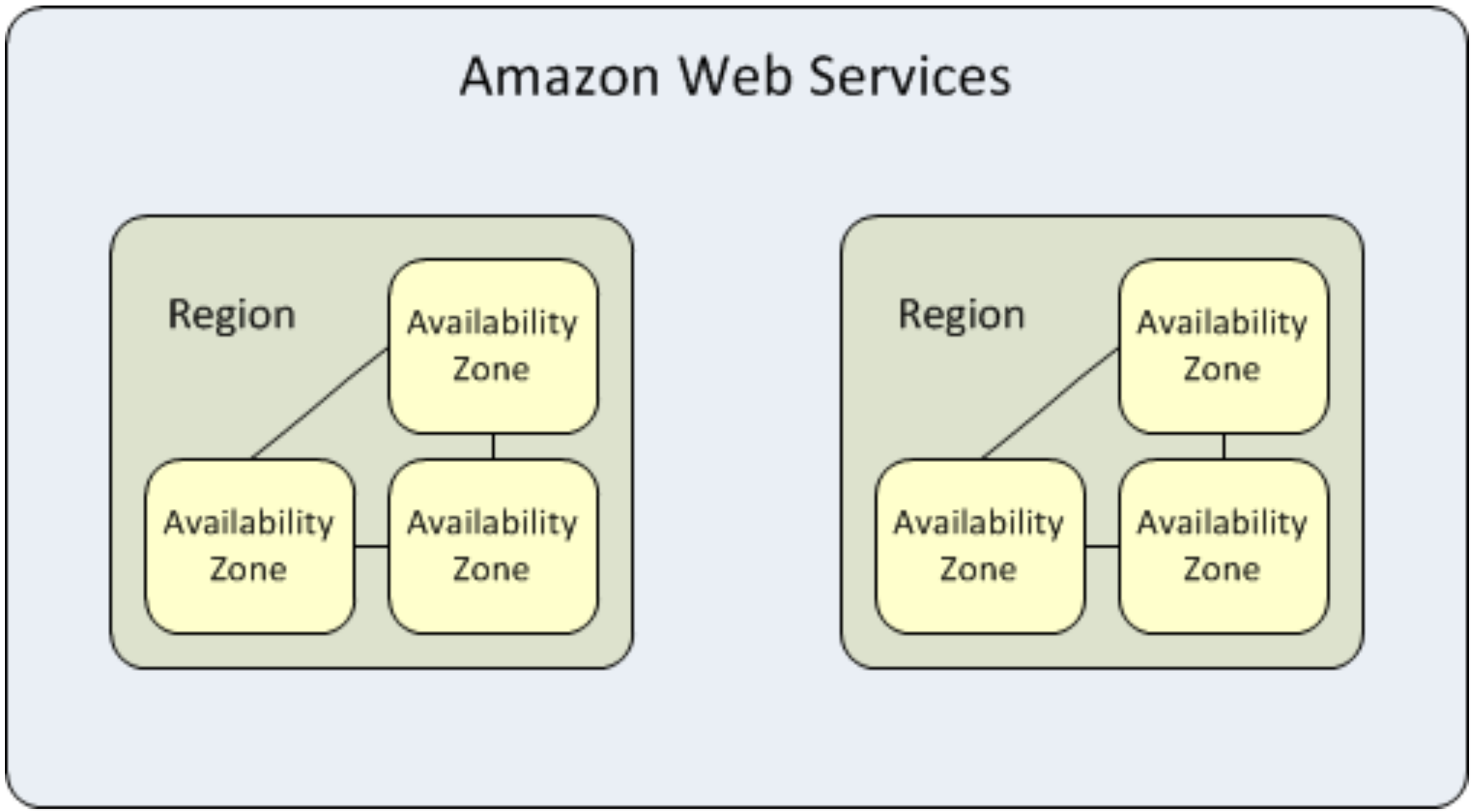
- 可用性に深刻な影響を与えない月例メンテナンスに許される停止時間は？

  (*) 「深刻な影響」の考察があってもよい。

- 講義 Web フォームから記入すること。

Wikipedia より

AWS region and zone



Amazon, “Amazon Elastic Compute Cloud: User Guide for Linux”, March, 2015.

- How long downtime is allowed at the monthly maintenance without critical impact on the availability ?

- Strictly speaking, the reliability of the Web system is always lower than that of its infrastructure due to series connection.

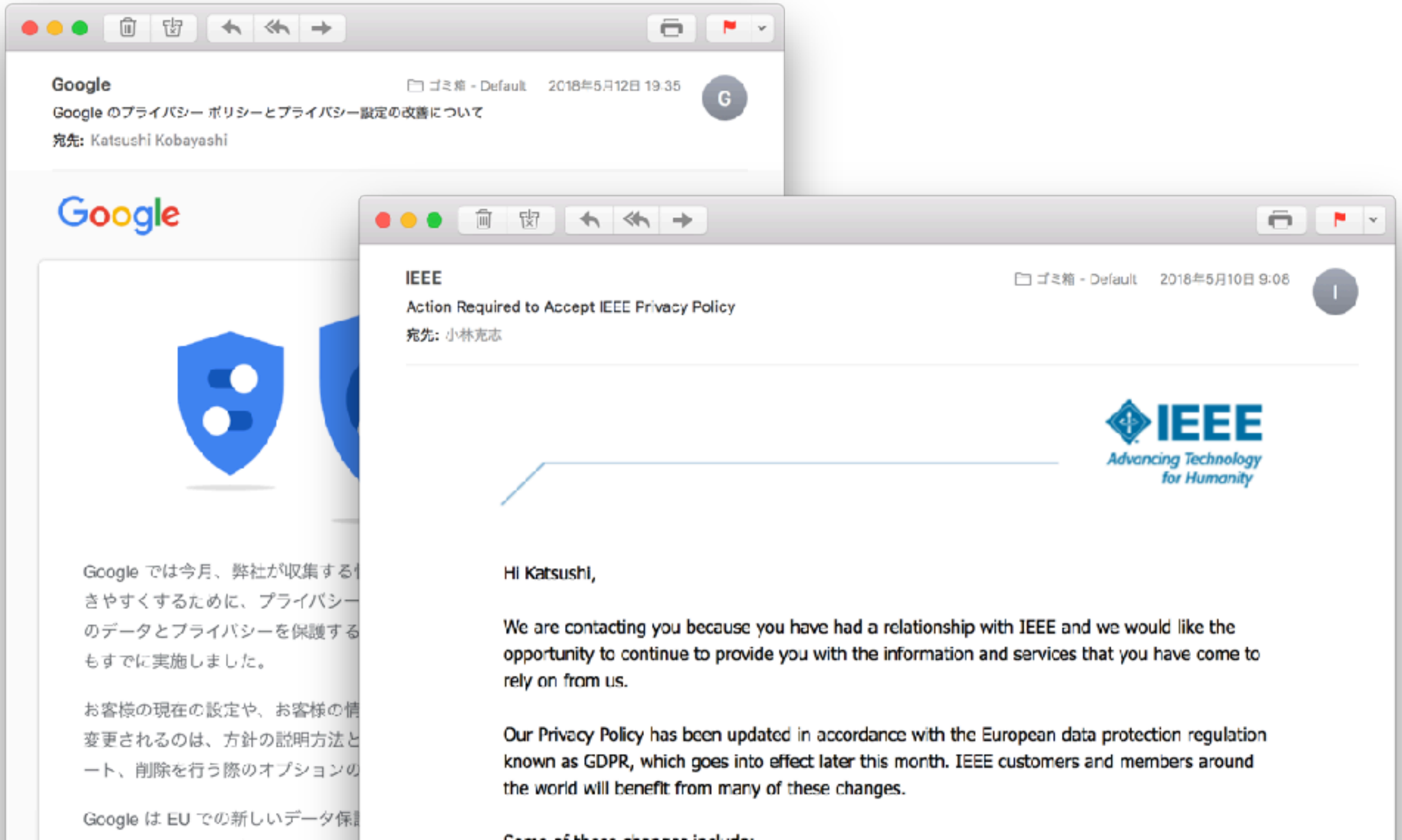
    - How much critical is allowed ?
      Is there alternative approaches ?

- How much reliability is expected each Azure domain ?

    - If allowed down-time of each domain, 15.8hrs/month, is the maintenance window, how much reliability is degraded ?

- If allowed down-time of entire system (22min/month) is the maintenance window, how much reliability is degraded ?

Have you got e-mails regarding to update privacy policy ?



GDPR

○ 「データ保護指令」に基づく各国法に代わり、2018年5月25日からは「一般データ保護規則」（GDPR: General Data Protection Regulation）がEU加盟国（及びEEA協定に基づきEU法の適用を受けるアイスランド、リヒテンシュタイン、ノルウェー）に直接適用される。

【事業者の義務の例】

	GDPR	個人情報保護法
センシティブデータ	取扱い禁止	取得と提供には本人の事前同意が必要
アクセス権	全ての個人データが対象	6ヶ月以上保有の個人データのみ対象
データポータビリティの権利	認められる	開示請求権あり
データの取扱いの記録義務	全ての取扱いが対象	第三者提供時のみ対象
データ漏えい時の監督当局への通知義務	リスクをもたらす可能性が高い場合には72時間以内に通知する義務	委員会告示等に従い報告する努力義務 ただし、時間制限の規定なし
データ保護オフィサー	次の場合に任命義務あり ●定期的かつ体系的な大規模監視を必要とする場合 ●大規模のセンシティブデータを処理する場合	任命義務なし ただし、従業者の監督義務や安全管理措置を講じる義務あり

【EU域外の事業者にも適用される可能性：域外適用】

✓ EU域内の個人に向けた商品/サービスの提供
✓ EU域内の個人の行動監視（追跡）

※言語・通貨・消費者への言及等の事情によりEUに対する商品/サービスの提供の意図が明白か否かが基準

に伴う個人データの取扱いに対しては、EU域外所在の事業者についてもGDPRが適用される(当該EU域外所在事業者は、ＥＵ域内に拠点をもつ代理人を指定しなければならない)

【違反時の制裁金】

✓ 最大2,000万ユーロまたは全世界年間売上高の４%の制裁金

GDPR : EU General Data Protection Regulation, https://www.kantei.go.jp/jp/singi/it2/senmon/dai14/siryou2-2.pdf

EU General Data Protection Regulation (GDPR)

- Apply all institutes including not only companies, but also NPO, universities.
  - The policy updates are actions for the regulation.

- Cross border data transfer is a critical topic for non-EU origin companies.
  - The costs of GDPR compliance are expensive, if the origin region is not certified “adequate level of protection” by EU.
  - “Japanese government” is likely optimistic, which EU is expected to certificate Japan, soon.

Outline

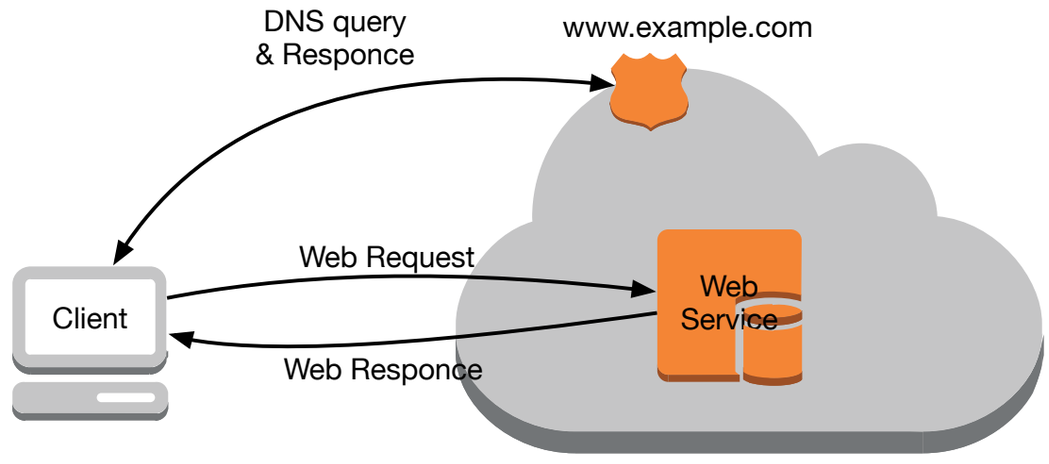
1. Administravia

2. Homework review

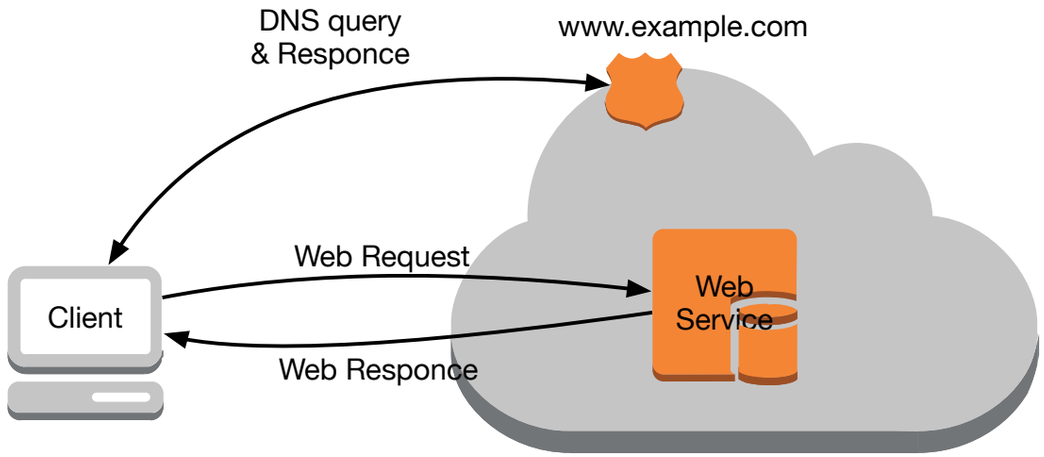
3. Scale-up / Scale-out

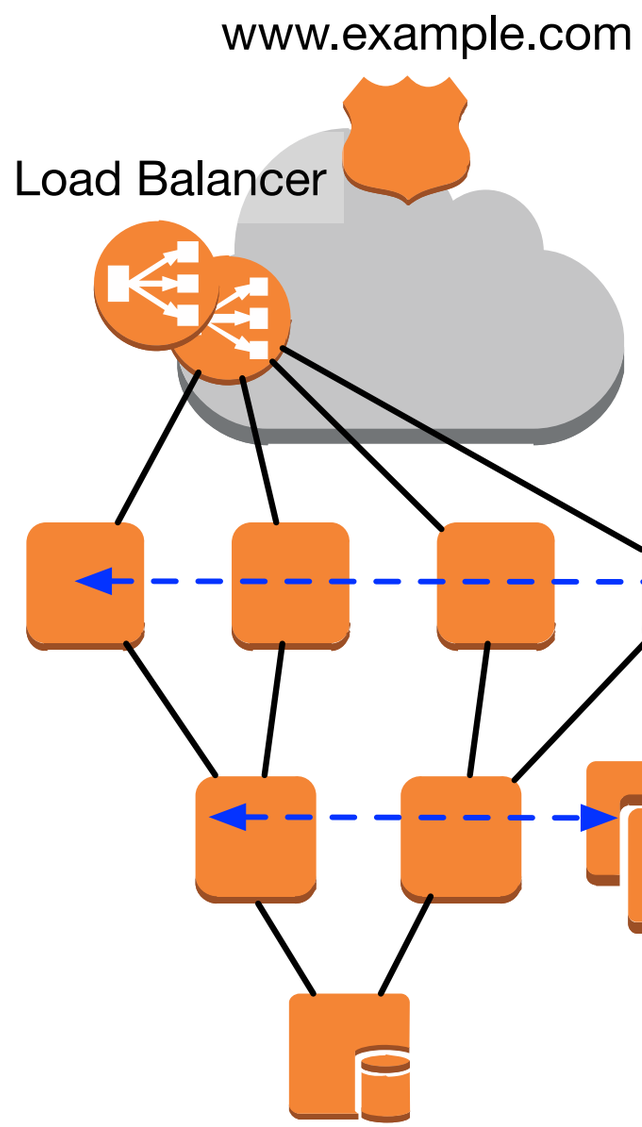
Web service

- Is more complicated than the simplest model
  - Static contents delivery with a single Web server is out of date.
- Provides dynamic contents to adapt customers with datastore.
- Required service scalability to support huge number of customers, 100M or more.
- Adopts multi-tier architecture
  - Three-tier : Web + Application + DB



3-tier Web architecture and scalability





Nature of Web services

- Bursty access patterns and traffic loads, but satisfaction should be kept
  - Service unavailable, e.g., 506 error, is critical, which regards as opportunity loss.
    - 2,68B USD on-line sales on Cyber Monday 2014 (Fundivo)
      - 7M VoD streams at Barack Obama’s victory speech in 2009 (Akamai)
- Legacy approaches are NOT practical :
  - To increase server performance
  - Reserve resources to meet peak-loads
    - unacceptable costs, e.g., HW, power

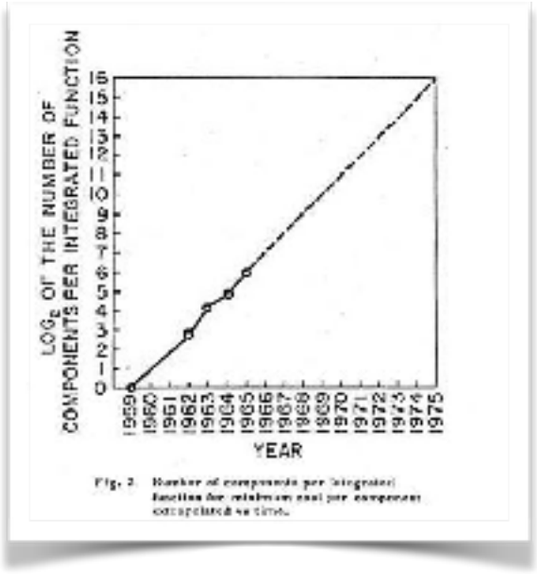
http://www.fundivo.com/stats/cyber-monday-statistics/
http://www.akamai.com/html/about/press/releases/2009/press_012009.html

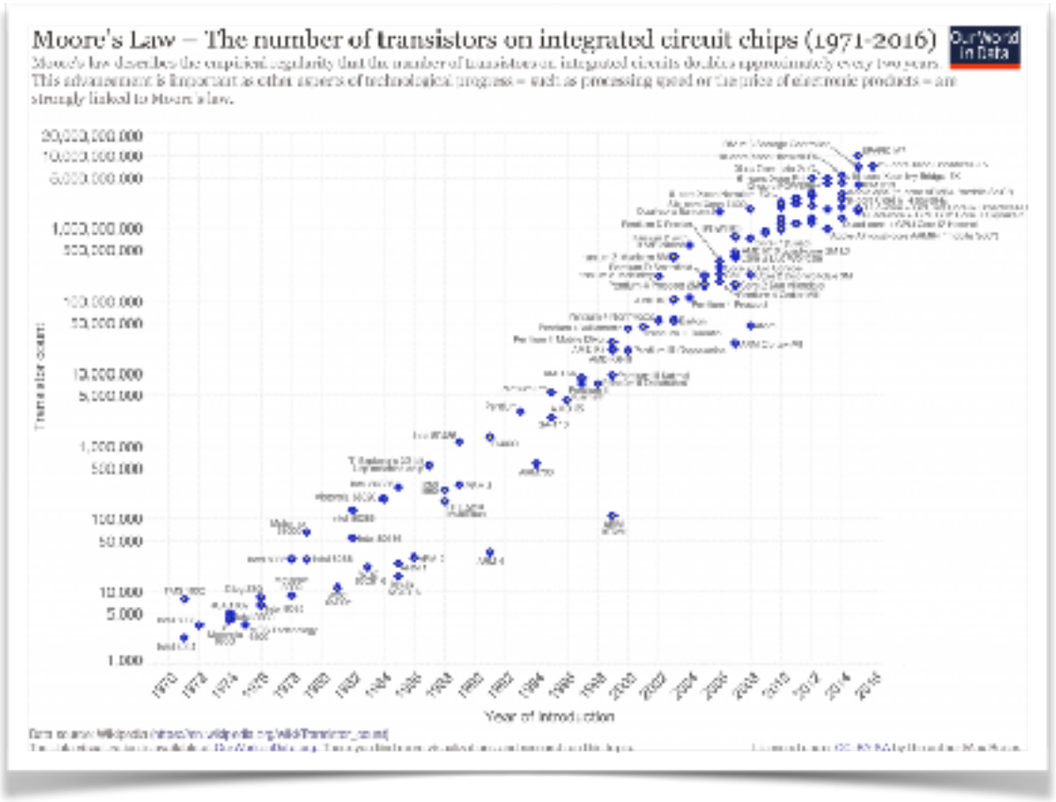
Scale up (vertical) and scale out (horizontal)

- Scale-up: To improve performance of components.
    - Clock-up CPU, BUS, Increase DRAM or Storage capacity.
        - Cost is NOT scale with performances. $$$ for higher performance system.
        - State of the art technologies cannot be employed by other than COTS*.
            - At the 32-nm node, a chip needs to sell about 30 to 40 million units to recoup the costs associated with it, Hsu (VP of Cadence) said. At the 20-nm node, the "breakeven" point jumps to between 60 and 100 million units, Hsu said. (EE Times)
- Scale-out : To increase number of components
    - Gradually improve within affordable cost.
    - Note: Most Web services bottlenecks are at I/O rather than CPU.

* Commercial Off-The-Shelf : Products available in the commercial marketplace.
http://www.eetimes.com/document.asp?doc_id=1260470

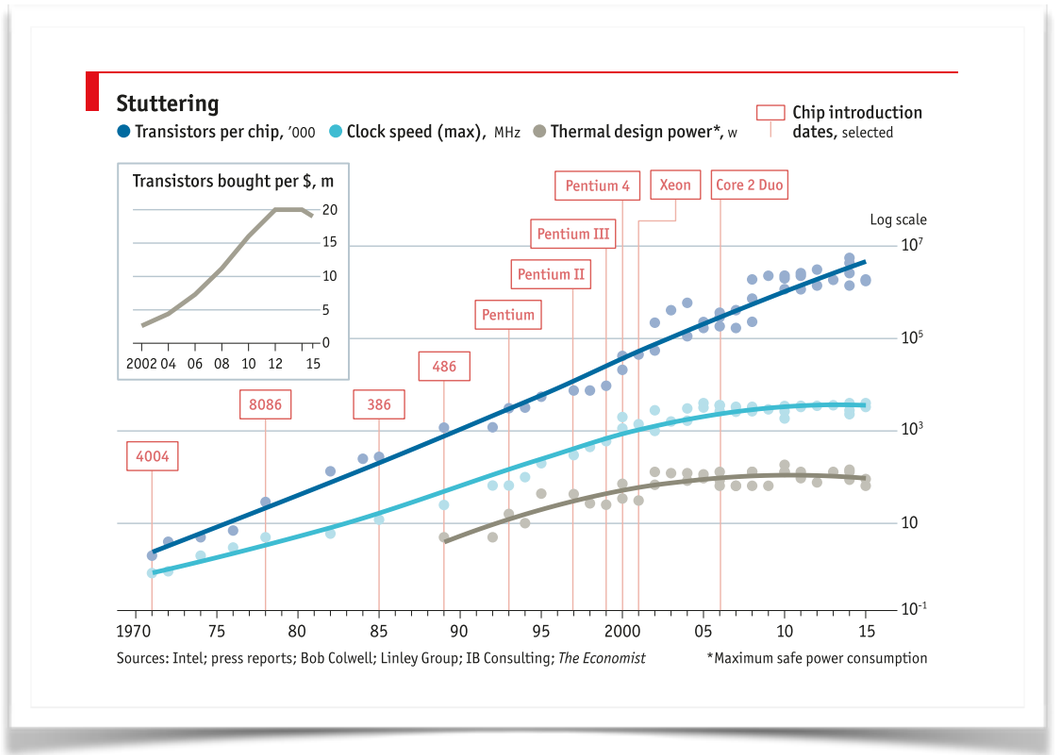
Moore’s law





- A doubling every year in the number of components per integrated circuit.

- Not covers CPU speed, such as clock rates, single threads performance.



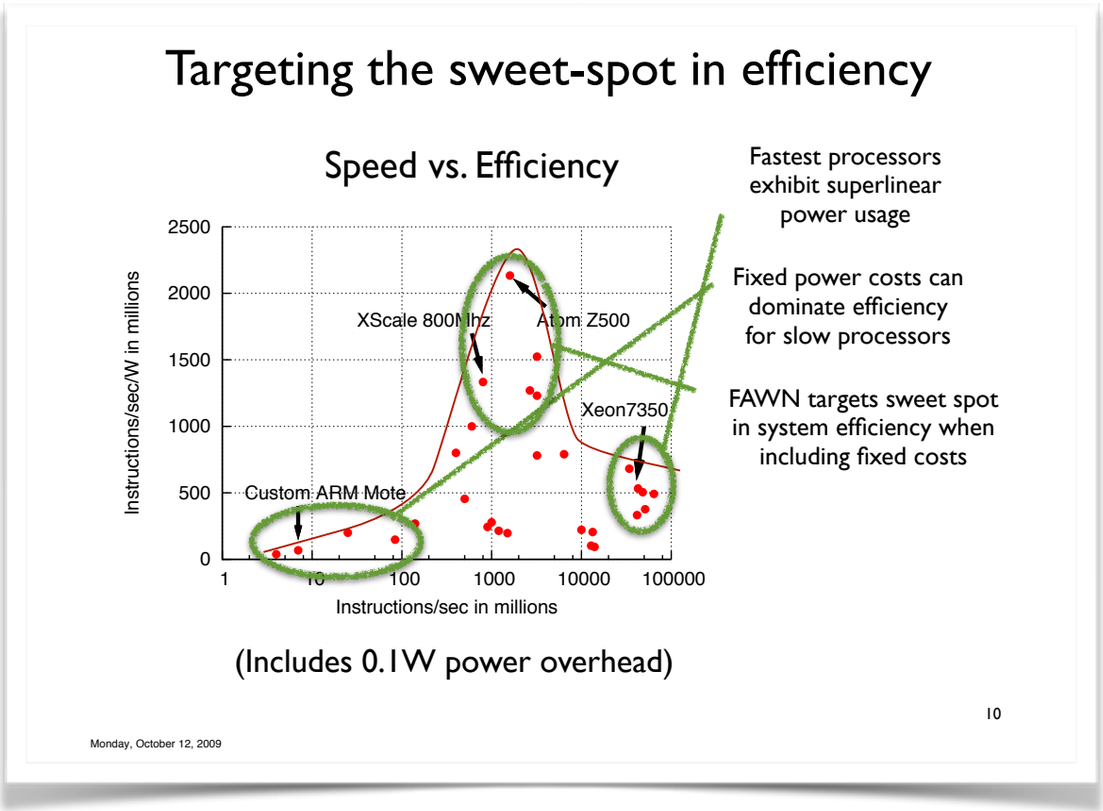
http://www.intel.com/pressroom/kits/events/moores_law_40th/index.htm?iid=tech_mooreslaw+body_presskit

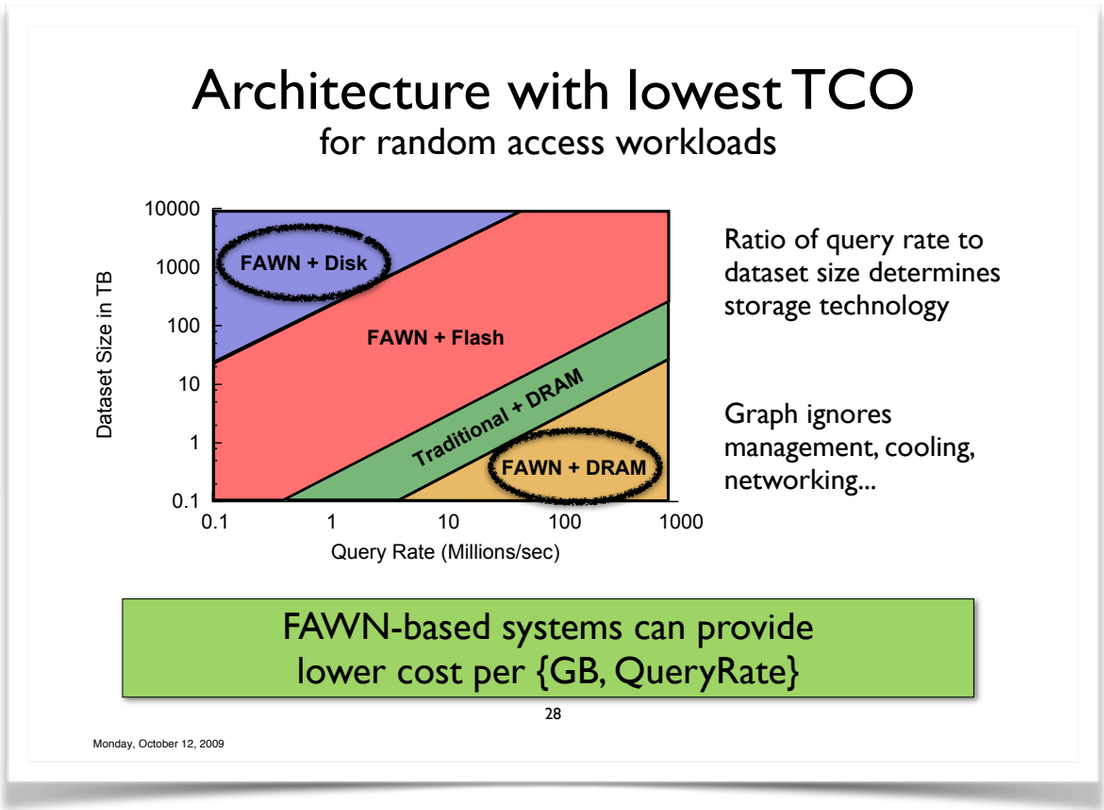
https://ourworldindata.org/technological-progress/#note-2

http://www.economist.com/technology-quarterly/2016-03-12/after-moores-law

CMU-FAWN(Fast Arrays of Wimpy Nodes)

- Embedded CPU cluster system presented better performance than server CPU, in terms of power consumption and of I/O.
    - SeaMicro (AMD) and HP released 512 cores and 90 cores clusters, respectively.





FAWN: A fast array of wimpy nodes, David G. Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, and Vijay Vasudevan, SIGOPS ’09

But, mainframe is still a big business.

IBM

みずほ銀行、最新IBMメインフレーム「IBM z13」を採用

2015年4月28日

みずほ銀行、最新IBMメインフレーム「IBM z13」を採用

アプリケーションをプライベート・クラウド基盤に統合し約2割のコスト削減を見込む

日本IBMは、株式会社みずほ銀行（以下、みずほ銀行）がIBMの最新メインフレーム「IBM® z13」の採用を決定したことを発表します。最高レベルのセキュリティーと可用性を提供する「IBM z Systems」は現在、みずほ銀行のネットバンキング・サービスを支えるダイレクト・チャネル基盤および基幹業務である勘定系システム基盤として稼働しており、今回新たに海外勘定系システムの基盤としてその最新モデルである「IBM z13」が採用されました。みずほ銀行では、銀行業務のみならず信託や証券などの業務を一元的に支える共通ITプラットフォームの構築によるワン・ストップサービス実現を推進しています。最新の「IBM z Systems」のプライベート・クラウド基盤上にアプリケーションを統合・集約を進め、運用負荷を軽減することで、約2割のコスト削減とともに、顧客サービスのさらなる拡充を見込んでいます。新システムは2016年後半の稼働開始を予定しています。

Today’s quiz

- Tell an actual example of a systems that are designed scale-up and scale-out approaches other than computer system.

- Ex.)
  Scale-up: To switch larger lecture-room so that more student can be accepted.
  Scale-out: To switch distance learning using video on demand.

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

- Bonus points will be given for excellent jokes and stories.

本日のクイズ

- スケールアップ／スケールアウトについて（計算機以外の）具体例を示せ。

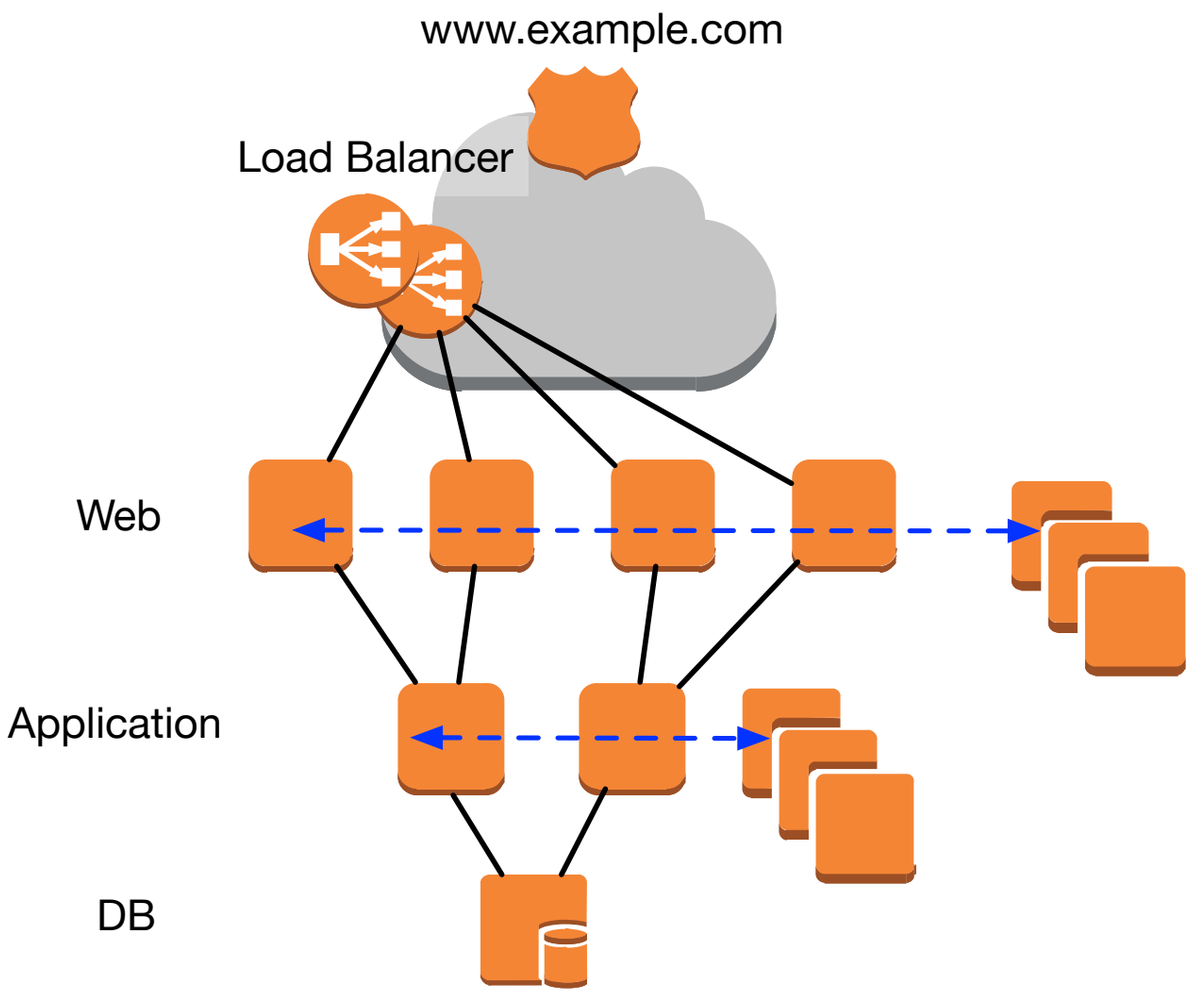
- 例）スケールアップ：多くの学生を受け入れるためにより大きな教室に変更する。
  スケールアウト：VoD を利用する遠隔学習に切り替える。

- 秀逸な小咄には加点する

- 講義 Web フォームから記入すること。

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.



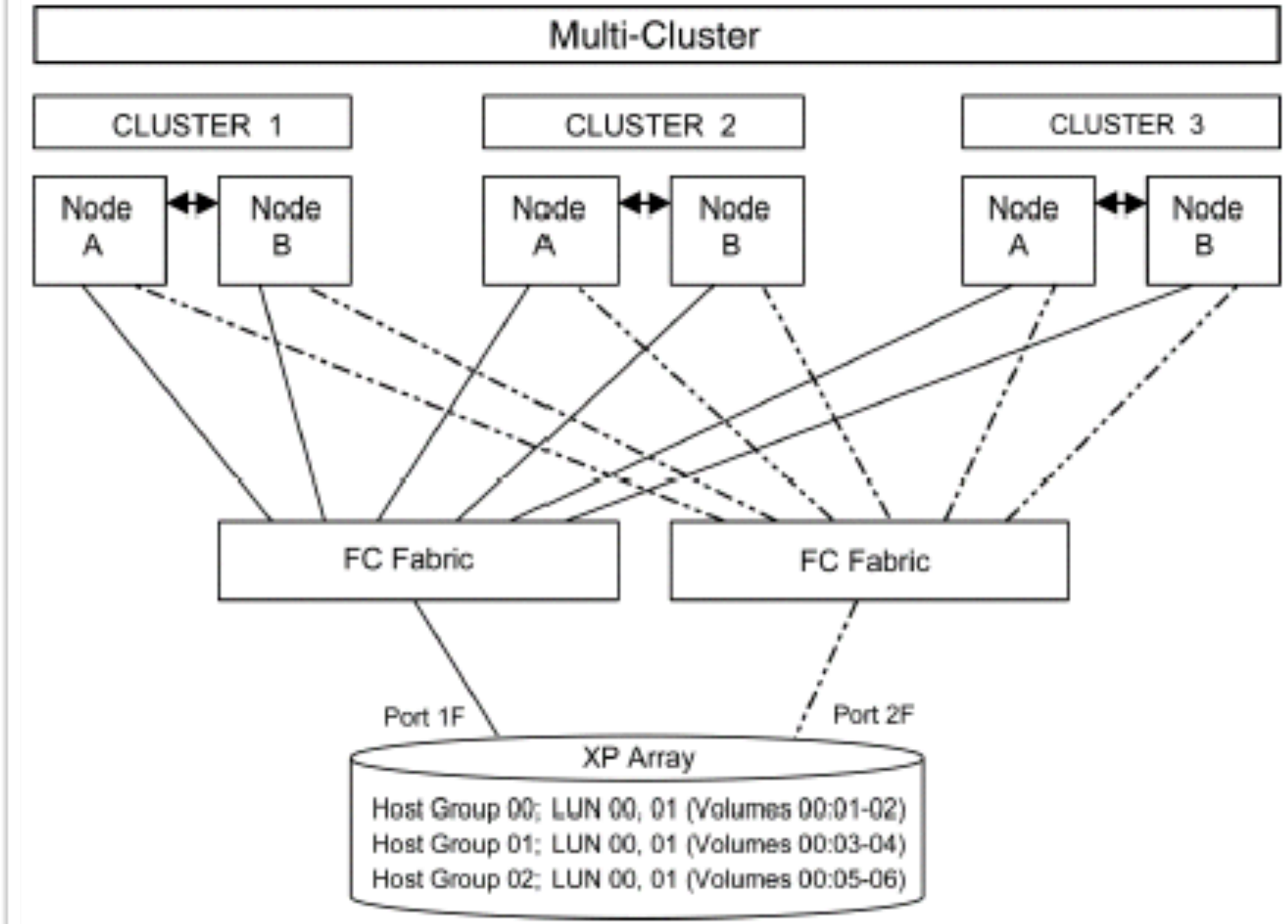
Scale up approaches in Storage

- Storage : Disk array, Storage Area Network (SAN)
    - Throughput : Improved by more # of disk headers, but IF bandwidth and disk controller may be next bottleneck.
    - Reliability : by redundant disks
    - Flexibility (SAN): volumes shared by 1+ servers with partitioning
- Server : High Availability (HA) Cluster.
    - $$$: Proprietary HA, e.g., HP, Oracle (formerly Sun)

Hard Disk Drive (HDD) & Disk Array

HGST Deskstar T7K500
http://www.hgst.com より

DELL PowerVault MD1220
http://www.dell.com より


Multi-Cluster
CLUSTER 1
CLUSTER 2
CLUSTER 3
Node A
Node B
Node A
Node B
Node A
Node B
FC Fabric
FC Fabric
Port 1F
Port 2F
XP Array
Host Group 00; LUN 00, 01 (Volumes 00:01-02)
Host Group 01; LUN 00, 01 (Volumes 00:03-04)
Host Group 02; LUN 00, 01 (Volumes 00:05-06)

Case of storage trouble

- More than three month trouble on Sakura's IaaS storage.
- Start service without enough experiences and evaluations because Sakura depended upon vendor.
- Sakura finally resolved it by changing boxes and tunings.

i.	性能限界におけるストレージ装置のテストを十分に行うことができなかった。設計仕様に基づいた最大収容数に相当するサーバを準備することができず、予想に基づいて生成した負荷により性能確認をしたため、実運用時に発生した問題に迅速かつ正確に対処することができなかった
ii.	クローン、スナップショットの作成所要時間が、共有ファイルシステム数の増大により遅延することが判明した
iii.	InfiniBand ポートにおいてパケットロスが観測され、期待していた性能が発揮されなかった。ファームウェア・アップデートにより解消されたが、対応完了まで 3 カ月を要した
iv.	監視・運用に必要なツール類が、ファイル数の増大、アクセスの増大に伴い利用できなくなった。このためストレージの状態を正確に把握することができなくなり、運用上重大な支障をきたすようになった
v.	ストレージ装置の仕様と動作について弊社エンジニアが全容を把握することができず、発生した障害に対して十分な対応を実施することができなかった。メーカーとの綿密な連携により対処を急いだが、対処のための調査と確認に長い日数がかかってしまう結果となった

さくらのクラウド・ストレージに関する報告書、2012年6月

Object storage

- Manage data as objects with data and metadata.
- Consist of distributed data- and metadata nodes.
    - Meet scale-out approach.
- E.g., Amazon S3, Google Filesystem, Openstack SWIFT.

- Block storage : Data as blocks within sector and tracks.
    - Included computing instances as boot disk.
    - Snapshot used for software distribution like VM images. Also for template images for auto-scaling.
- File storage : Data as file hierarchy.
    - Some are implemented on top of object storage.

(Distributed) Object storage: Google File System

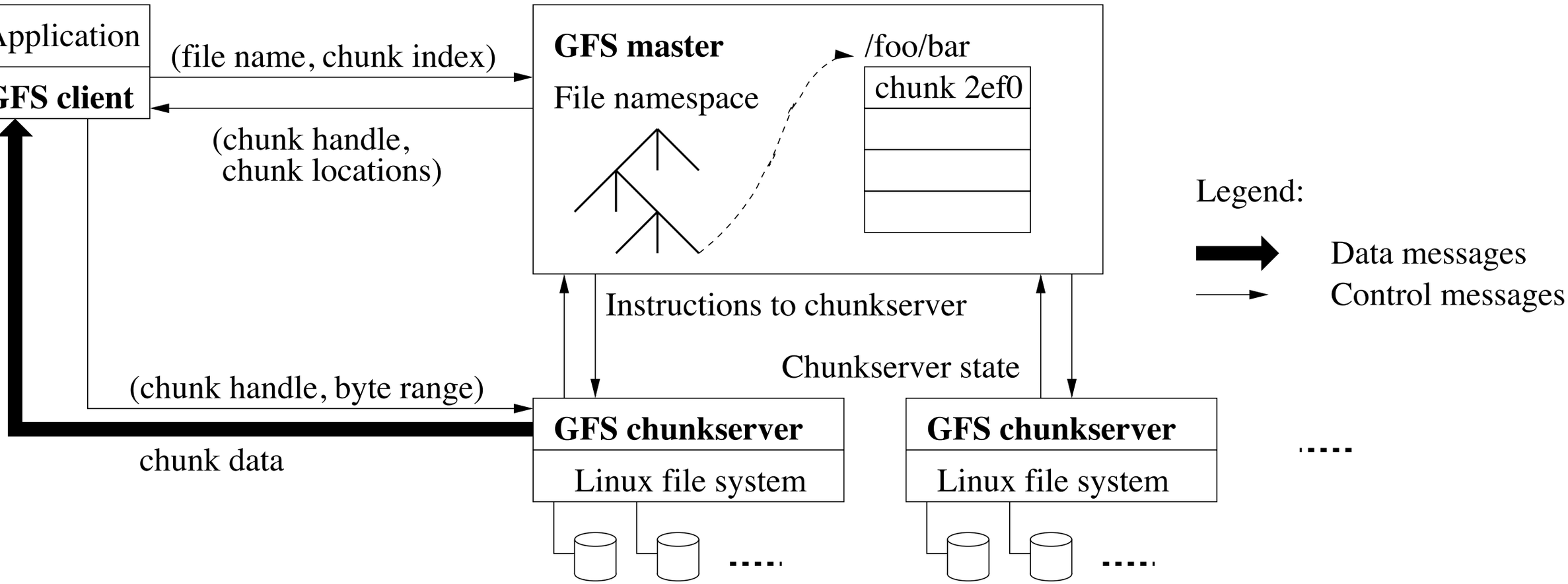
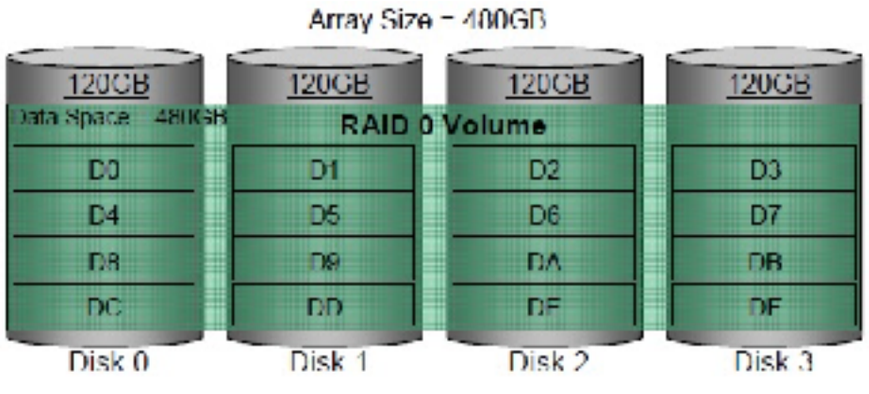


Figure 1: GFS Architecture

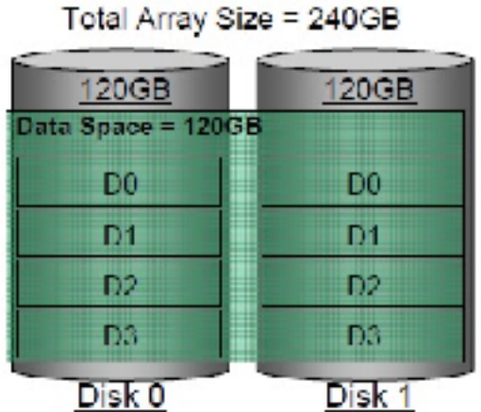
Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. "The Google file system." *ACM SIGOPS operating systems review*. Vol. 37. No. 5. ACM, 2003.

Redundant Arrays of Independent/ Inexpensive Disks (RAID)

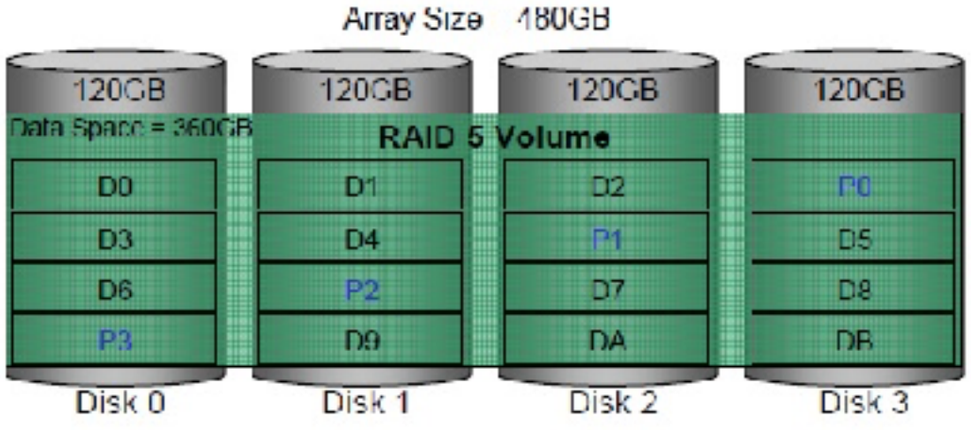
RAID 0 (striping, no redundancy)



RAID 1 (mirroring)

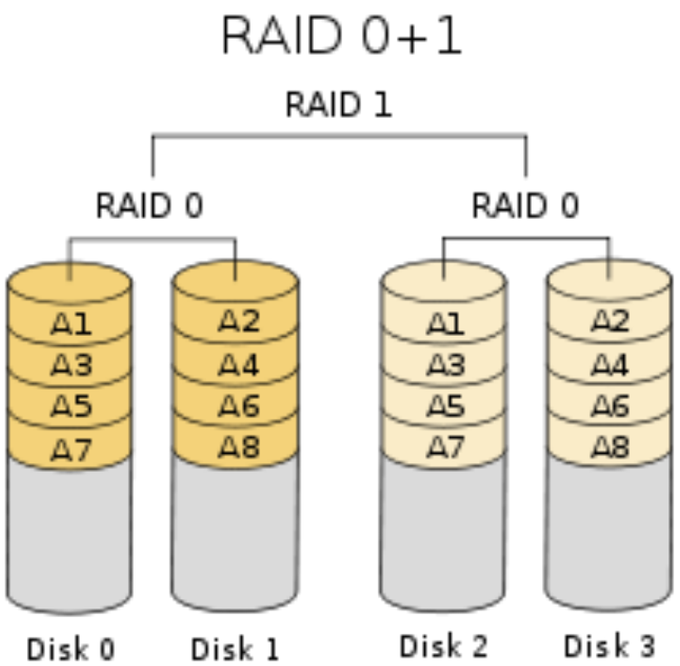


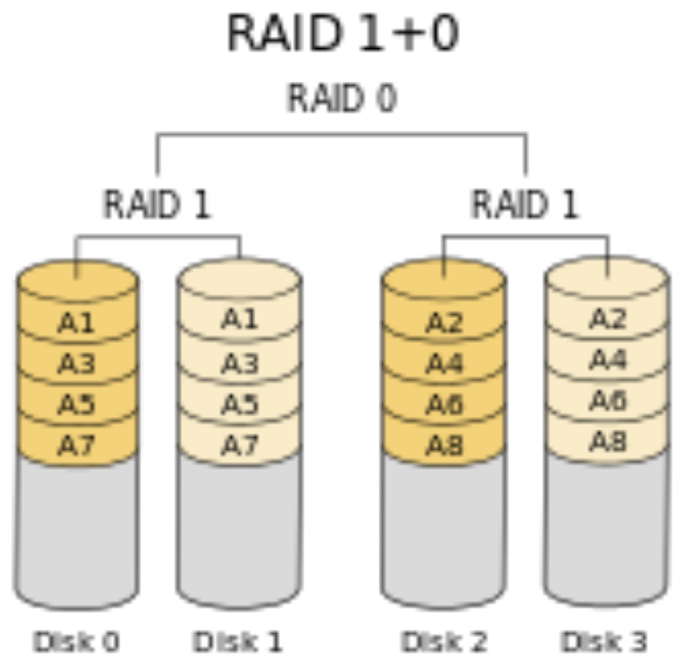
RAID 5 (striping with parity)



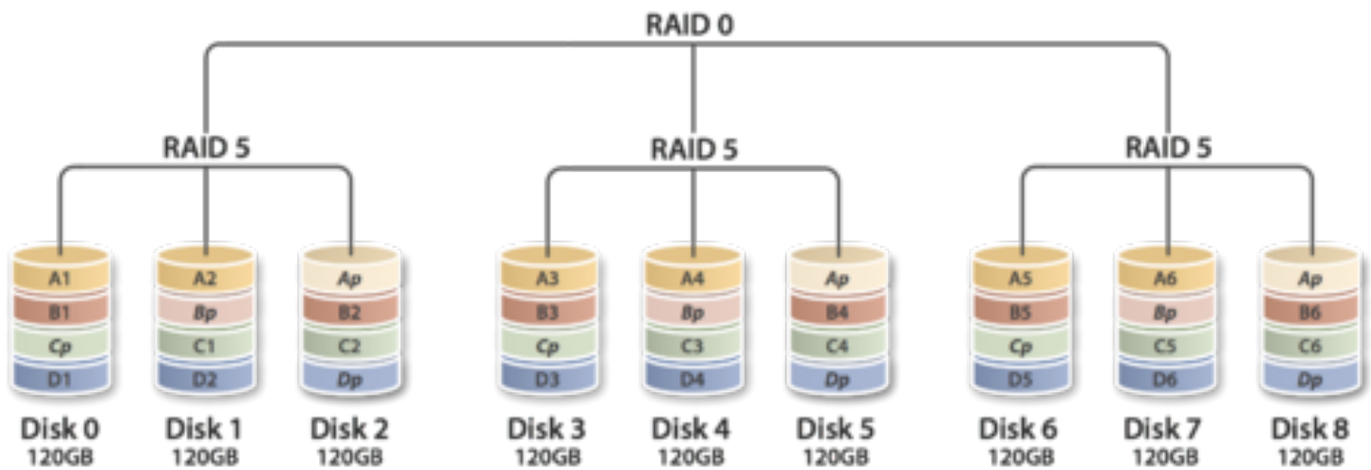
http://www.intel.com/support/chipsets/imsm/sb/CS-009337.htm

Nested RAID configs





RAID 50



Wikipedia より

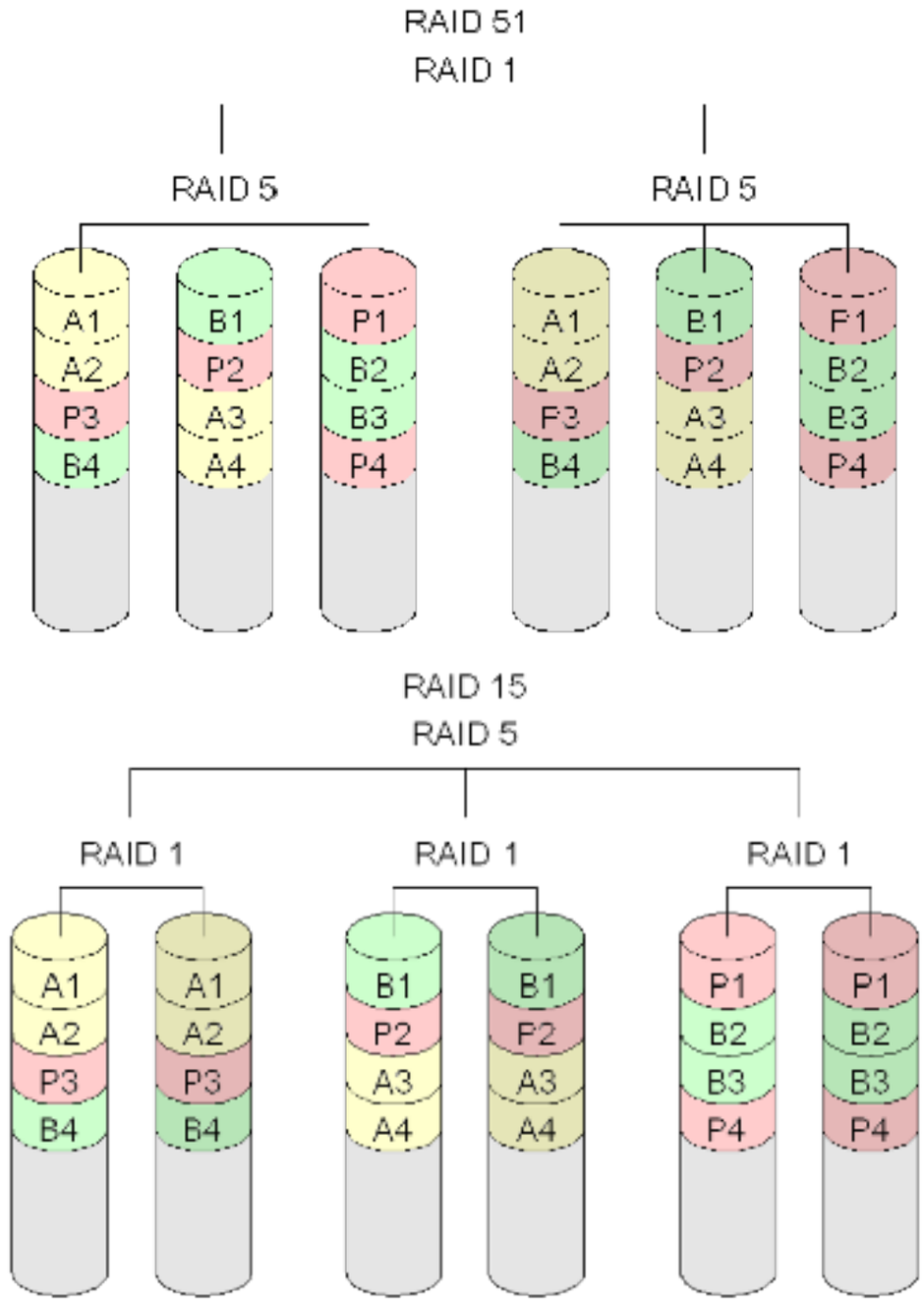
Today’s assignment

- Compare couple of nested RAID configurations
  (1) RAID-51, and (2) RAID-15.

- Discuss which configuration is better, and why ?

    - Consider the operation after HDD crash.

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



Wikipedia より

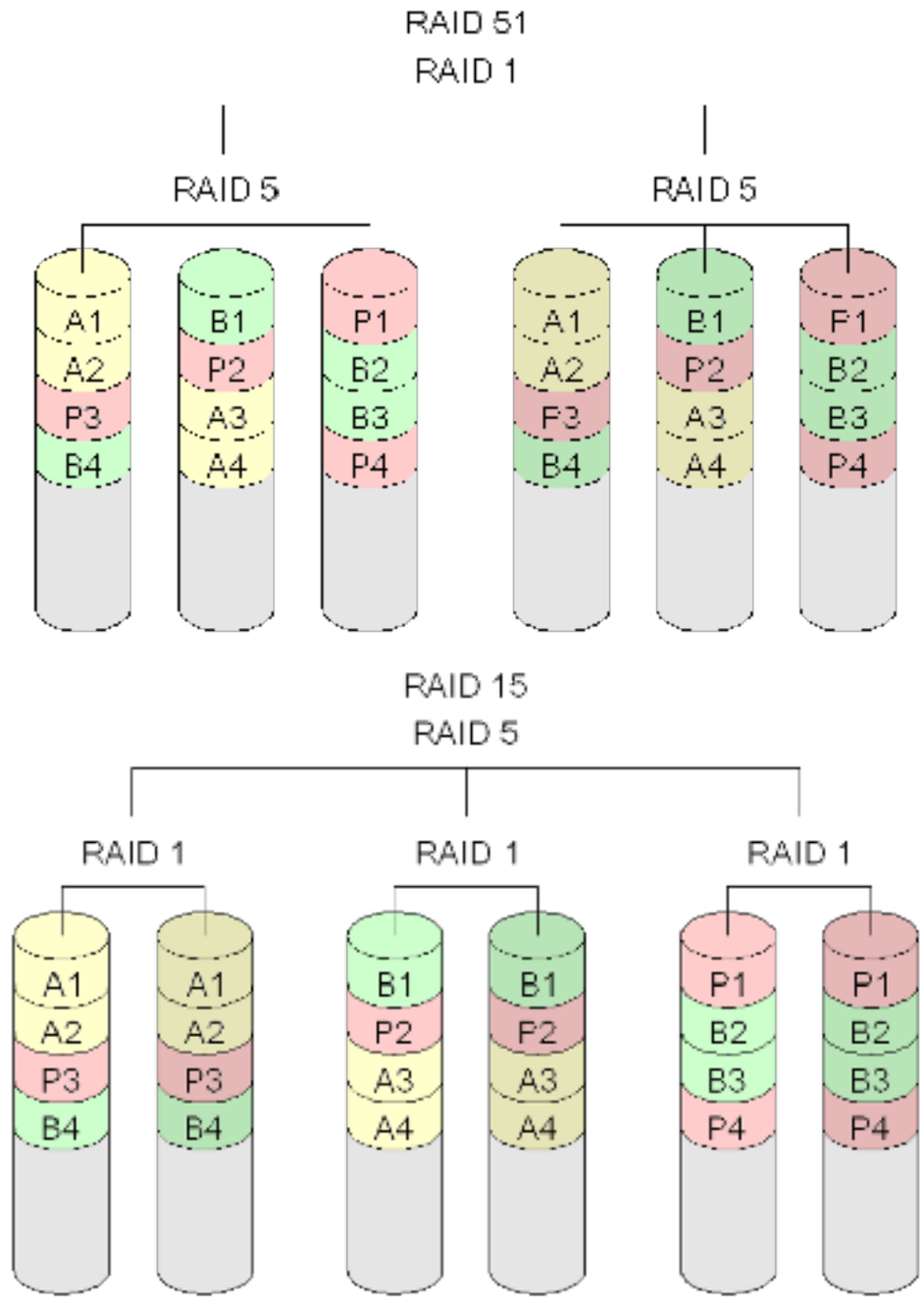
本日の課題

- 2 種類の nested RAID 構成、(1) RAID-51, and (2) RAID-15 を比較せよ。

- いずれの構成が好ましいか議論せよ？

  - ディスク障害後の運用について検討してみる。

- 講義 Web フォームから記入すること。



Wikipedia より