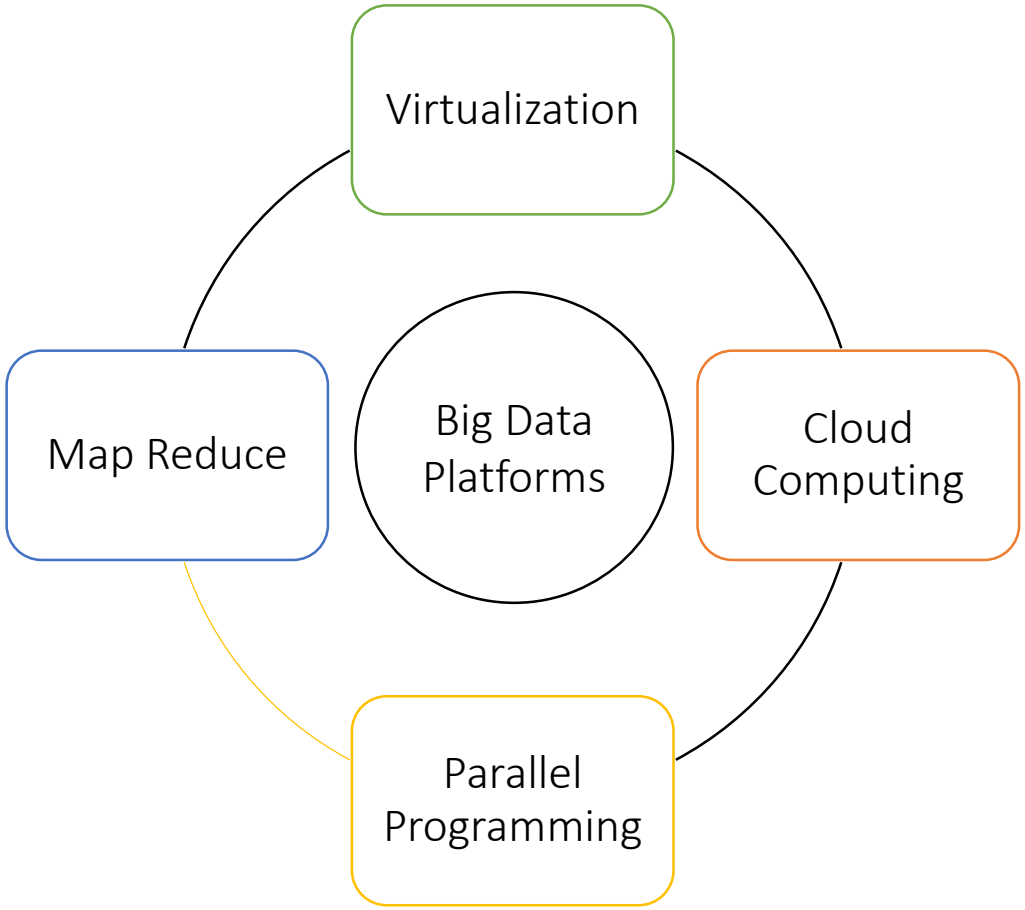
# Big Data Platform Elements

Edgard Luque

# Big Data Platform Elements



# Virtualization

# What is Virtualization?

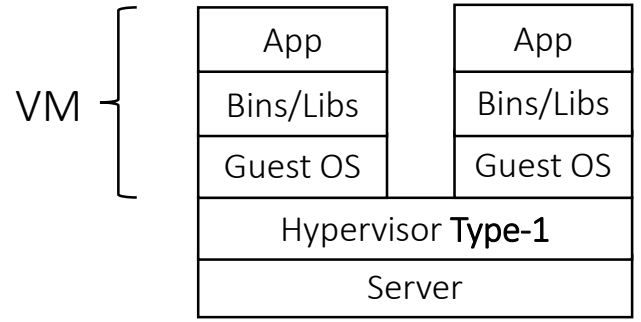
- *"Virtualization means that Applications can use a resource without any concern for where it resides, what the technical interface is, how it has been implemented, which platform it uses, and how much of it is available"*

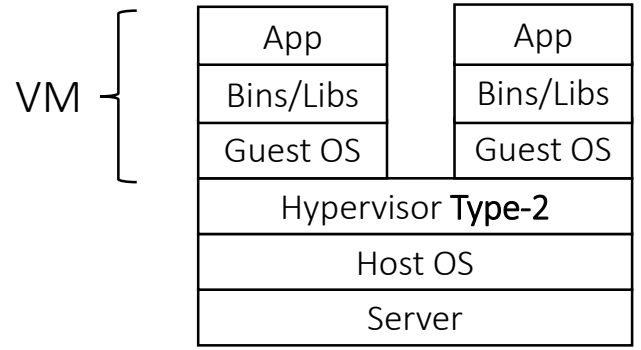
  ~Rick F. Van der Lans in Data Virtualization for Business Intelligence Systems

- We'll look at a few different types of virtualization:
  - Server Virtualization – can be HW-level or OS-level Virtualization
  - Storage Virtualization
  - Network Virtualization
  - Desktop Virtualization
  - Application Virtualization

# Server Virtualization: HW-level virtualization

- Ability to run multiple Virtual Machines (VMs or guests) on a single Physical Machine (host).
- Each Virtual Machine emulates the underlying physical hardware and has an Operating System (OS).
- Guest VMs are mostly completely isolated from each other.
- Each guest VM can run a different OS.
- Hypervisors (or Virtual Machine Monitors or VMMs) are used to create and run VMs. There are two types of hypervisors:
  - Type-1, Native or Bare-metal Hypervisors:
    - Run directly on the host's hardware.
    - Example: Hyper-V Hypervisor.
  - Type-2 or Hosted Hypervisors:
    - Run on the host's OS.
    - Example: VMware Player, VirtualBox
- Server Virtualization provides improved utilization, and scalability

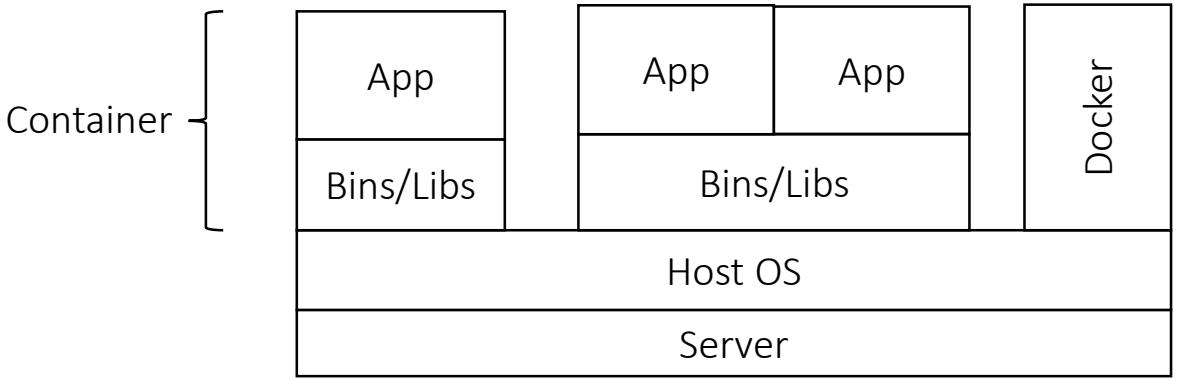



*Reference: https://en.wikipedia.org/wiki/Hypervisor*

# Server Virtualization: OS-level virtualization

- Ability to run multiple isolated Containers (user-space instances or guests) on a single Physical Machine (host).
- Containers do not emulate the underlying HW and don't have their own OS (they share the host OS). This lighter footprint allows hosts to support a higher density of guest Containers (as against guest VMs). But on the flip side raises Security concerns.
- Containers can also share binaries and libraries with other Containers.
- Each Container typically runs a single Application.
- Example: Docker




*Reference: https://en.wikipedia.org/wiki/Operating-system-level_virtualization*

# Review: Storage Definitions

- Block
  - A sequence of bytes.
  - Storage systems typically provide access to blocks.
  - The OS typically abstracts other logical views like files and records.

- File System:
  - Controls how data is managed, stored and retrieved.
  - Without a file system, we would just have a large blob of data with no way to identify different connected pieces of information.
  - File systems are organized around groups of data called files, and groups of files called directories or folders.
  - Distributed files systems are files systems that are spread across multiple servers.

*Reference: Wikipedia*

# Review: Storage Definitions

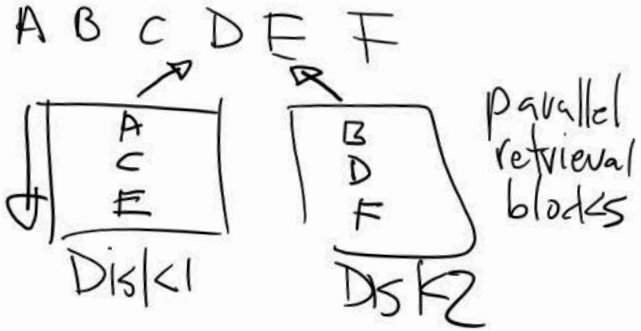
- Striping
  - Sequential blocks of data are stored on different physical storage devices in (typically) round-robin fashion.
  - Example: Disk1 <A, C, E>; Disk2 <B, D, F>
  - Striping is useful when requests for data are faster than a single storage device can deliver. Striping data across multiple storage devices allows for concurrent access to data thereby improving performance.

- Mirroring
  - Replication of data onto separate disks in real time.
  - Example: Disk1 <A, B, C>; Disk2 <A, B, C>
  - Improves data redundancy and reliability.

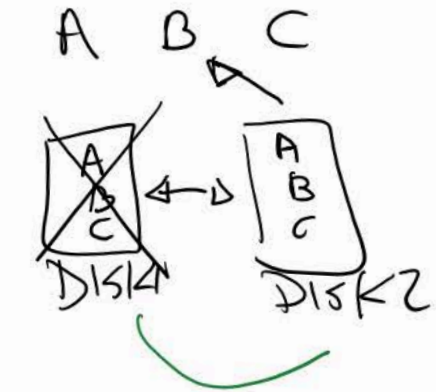
- Parity
  - When data on a crashed disk can be reconstructed using data on other disks (using the XOR operation)
  - Example: Disk1 <A:11010011>; Disk2 <B:10011001>; Disk3 <$P_{AB}$: 01001010>
    Essentially, $P_{AB}$ = A XOR B, so if one of the disk crashes,
    you can reconstruct using XOR operation between other two
  - Improves data redundancy

*Reference: Wikipedia*

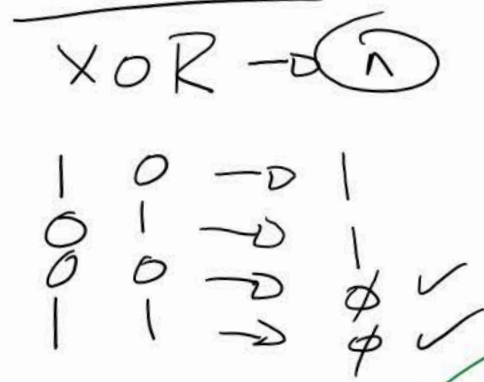
## Striping

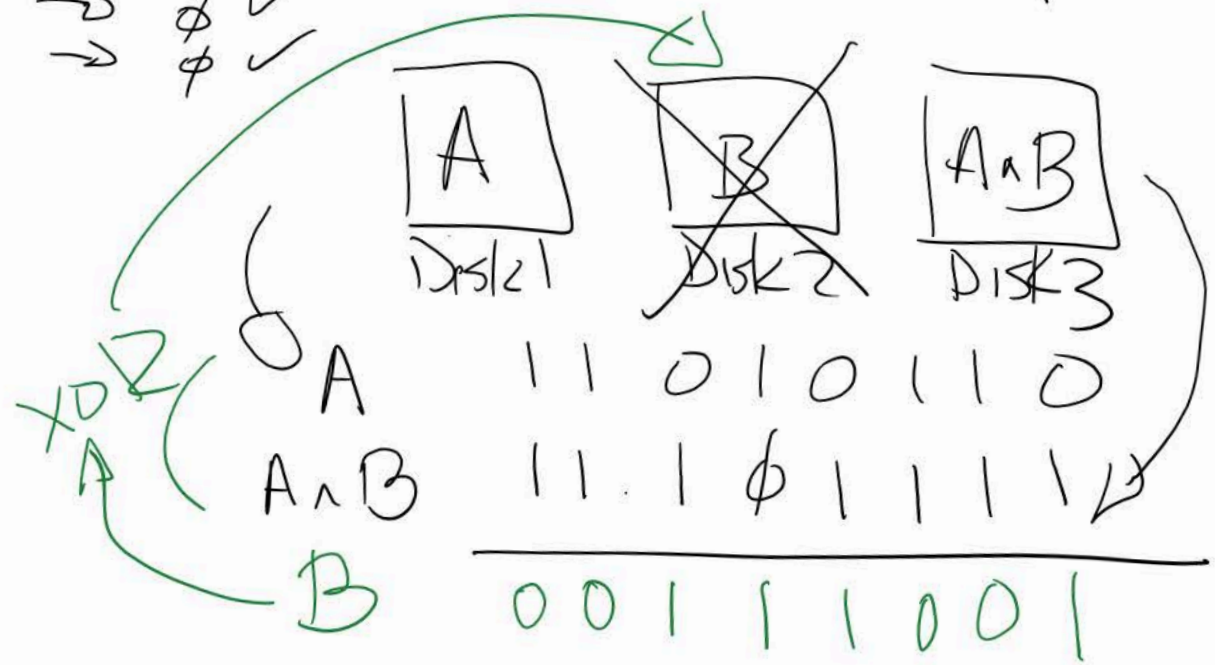



## Mirroring




## Parity



# Storage Virtualization

- Data is abstracted into what appears to be a single storage unit, while the physical storage actually spans multiple heterogeneous devices and often locations
- Storage Virtualization provides location independence, improved utilization, performance, reliability and availability
- Example: RAID (redundant array of independent/inexpensive disks)

*Reference: https://en.wikipedia.org/wiki/RAID*

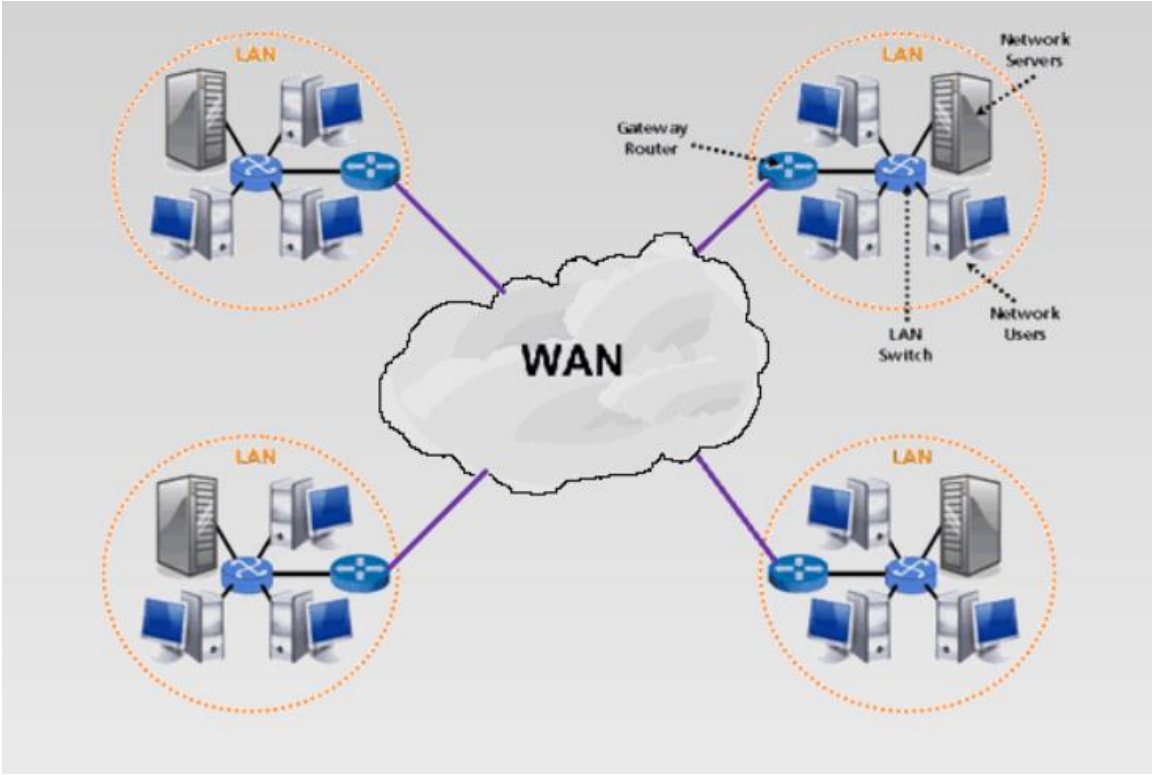
- Example: RAID (redundant array of independent/inexpensive disks)

| Popular RAID Types | Striping (provides excellent performance) | Mirroring (provides excellent redundancy) | Parity (provides good redundancy) | Minimum Number of Disks | Example (Disk – Blocks) | Comments |
|---|---|---|---|---|---|---|
| RAID 0 | Yes | No | No | 2 | Disk 1 -- A, C, E<br>Disk 2 -- B, D, F | Excellent Performance.<br>No Redundancy.<br>Do not use for critical applications. |
| RAID 1 | No | Yes | No | 2 | Disk 1 -- A, B, C<br>Disk 2 -- A, B, C | Good Performance.<br>Excellent Redundancy. |
| RAID 5 | Yes | No | Yes (Distributed Parity) | 3 | Disk 1 – A, C, $P_{EF}$<br>Disk 2 – B, $P_{CD}$, E<br>Disk 3 – $P_{AB}$, D, F | Good Performance.<br>Good Redundancy.<br>Most cost effective.<br>Fast Reads; Slow Writes. |
| RAID 10 | Yes | Yes | No | 4 | Disk 1 -- A, C, E<br>Disk 2 -- A, C, E<br>Disk 3 -- B, D, F<br>Disk 4 -- B, D, F | Excellent Performance.<br>Excellent Redundancy.<br>Great for mission critical applications.<br>Not as cost-effective as RAID 5. |

*Reference: https://en.wikipedia.org/wiki/RAID*

# Review: Network Definitions

- Local Area Network (LAN): A computer network with interconnected devices within a limited geographical area such as a house or building.
  - Switch: Connects devices together on a computer network
- Wide Area Network (WAN): A computer network that spans large geographical areas
  - Router: Carries traffic from one network to the other



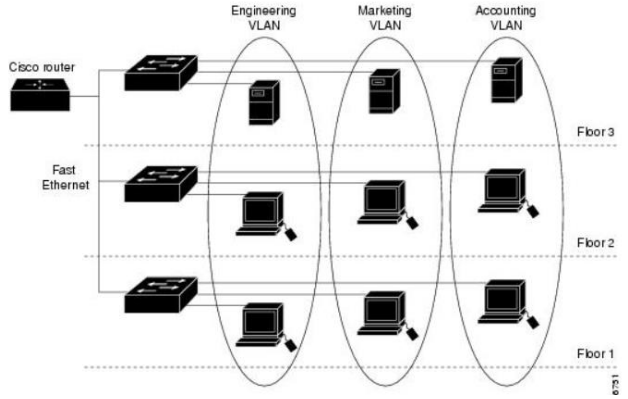

*Reference: Wikipedia*

*Image Source: http://netprivateer.com/lanwan.html*

# Network Virtualization

- Creation of logical, virtual networks that are decoupled from the (limitations of) underlying physical hardware.

- Example: VLAN, VPN
  - Virtual Local Area Network (VLAN)
    - Allows for grouping of hosts within a virtual LAN regardless of geographical location
    - Provides scalability, flexibility, simplified administration, and security
  - Virtual Private Network (VPN)
    - Securely extends a private network over a public network such as the internet
    - Users can remotely communicate with the private network as though they were directly connected to it with the same functionality, security and administrative policies
    - Provides flexibility, simplified administration, and security



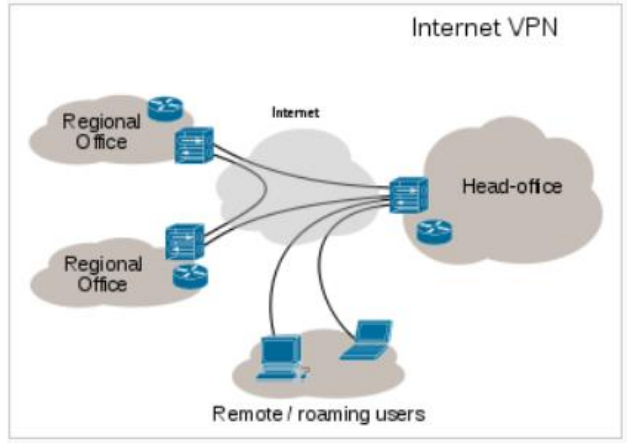




*Image Source: link*

*Image Source: https://en.wikipedia.org/wiki/Virtual_private_network*

# (Remote) Desktop Virtualization

- Enables access to applications on a remote OS using a virtual desktop.
- The remote OS carries the application and data, and only the display, keyboard, and mouse information are communicated with the local client device.
- Users (on the local client devices) must establish a session and be connected with the remote server to access the application.
- Makes installation, upgrades and management of applications easier for IT.

# (Remote) Desktop Virtualization (cont)

- Two kinds: RDS, VDI
  - Remote Desktop Services (RDS) aka Terminal Services
    - Provides remote desktop to multiple users on a Host OS
    - Provides users session-based isolation (session virtualization) - users share Host OS
    - Users have no admin privileges on the host OS
    - Can support higher user density
  - Virtual Desktop Infrastructure (VDI)
    - Provides remote desktop to multiple users on Guest OSs
    - Provides users VM-based isolation - each user gets a dedicated Guest OS
    - Users have admin privileges on the Guest OS
    - Support lower users density

# Application Virtualization

- Application Virtualization separates the Application from the OS, so Applications can be more easily deployed and delivered.
- The application is packaged and streamed from the server down the network to the client and, instead of being installed on the client device, is executed on the local device in a virtual bubble that is completely isolated from the client OS.
- Applications are streamed intelligently.
  - Only required parts are streamed as and when they are used.
  - Once the application has been streamed, it is cached on the client device so it doesn't have to be streamed every time a user uses it on the client. This also means the application can be used even when the client is not connected to the server.
  - When an application upgrade is available, the server copy is upgraded, and the upgrades are streamed down to the clients the next time the application is used on the client.
- Makes installation, upgrades and management of applications easier for IT.
- Examples: VMware ThinApp, Citrix XenApp and Microsoft App-V

*Reference: http://blogs.msdn.com/b/ianm/archive/2010/06/11/microsoft-virtual-desktop-101-making-sense-of-vdi-rds-app-v-med-v-and-desktop-virtualisation.aspx*

# Summary

- Four key elements make up big data platforms:
  - Virtualization, Cloud Computing, Parallel Programming, and MapReduce

- Virtualization:
  - "*Virtualization means that Applications can use a resource without any concern for where it resides, what the technical interface is, how it has been implemented, which platform it uses, and how much of it is available*"
  - Virtualization can occur at different levels of the stack: Server, Storage, Network, Desktop, and Application.