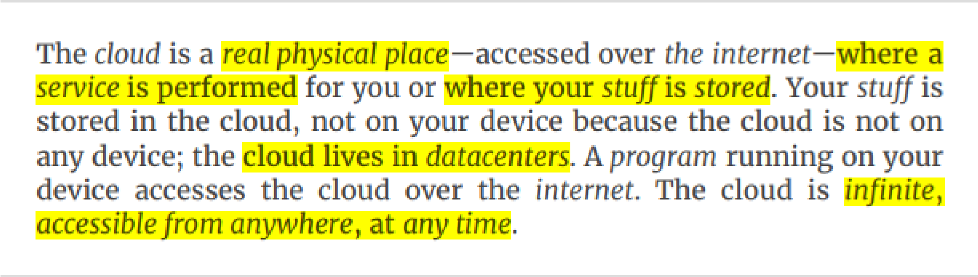
Explain cloud like I’m 10

- **Racks**: way of stuffing as many computers together as humanly possible.
- **Cloud computing** (just a service): accessing computer services over the internet.
- **Cloud** (~network of computers): big building with a lot of computers inside.
    - Cloud services run in the cloud.
    - Cloud lives in datacenters
    - Cloud sells computer as a service -> rent computer over internet.
    - The term ‘cloud’ comes from the symbol used to represent a network when drawing diagrams.



The *cloud* is a *real physical place*—accessed over *the internet*—where a *service* is performed for you or where your *stuff* is stored. Your *stuff* is stored in the cloud, not on your device because the cloud is not on any device; the cloud lives in *datacenters*. A *program* running on your device accesses the cloud over the *internet*. The cloud is *infinite, accessible from anywhere*, at *any time*.

- 2 Kind of Cloud
    - Cloud provider: own datacenters, let customers rent computer to build services.
        - Responsible for maintaining and operating their computer.
    - Cloud services (Gmail, YouTube, Facebook): perform a job for you in a cloud.
        - Services: compute, memory, storage, network, bandwidth, database…
- Cloud Computing began with **EC2 (Elastic Compute Cloud) in 2007**.
    - Easy to use: anyone with credit card and internet connection can rent
    - Permissionless: Not having to ask for permission.
    - Only pay for what you use.
    - OpEx vs CapEx
    - ‘Cloud Native’ software must be able to deal with computers failing at any time.
    - Datacenters can fail at any time.
        - Cloud allows programs to failover to different datacenters -> increase reliability.

Cloud Computing Concepts, Technology & Architecture

Chap 3. Understanding Cloud Computing

- **Cloud***:* a distinct IT environment that is designed for the purpose of remotely provisioning scalable and measured IT resources
- **IT Resource:**
  - **Software-based**: virtual server or a custom software program.
  - **Hardware-based**: physical server, or a network device
- **On-Premise:** An IT resource that is hosted in a conventional IT enterprise within an organizational boundary (that does not specifically represent a cloud) is considered to be located on the premises of the IT enterprise, or *on-premise.*
- **Scaling:**
  - **Horizontal Scaling ('Scaling Out')**: the process of increasing the number of nodes and machines in the resource pool.
  - **Vertical Scaling ('Scaling Up'):** the process of increasing the power of an existing system, such as the CPU or RAM, to meet the rising demands.
- **Cloud Service:** any IT resource that is made remotely accessible via a cloud (cloud service can exsist as a simple Web-based software program).
- **Cloud Service Consumer:** temporary runtime role assumed by software program when it accesses a cloud service.

Chap 4. Fundamental Concepts and Models

- **Cloud Delivery Models**
  - **Infrastructure-as-a-Service (IaaS**): self-contained IT environment comprised of infrastructure-centric IT resources (hardware, network, connectivity, operating systems, and other "raw" IT resources).
  - **Platform-as-a-Service (PaaS):** "ready-to-use" environment typically comprised of already deployed and configured IT resources.
  - **Software-as-a-Service (SaaS):** software program positioned as a shared cloud service and made available as a "product."
- **Cloud Deployment Models –** specific type of cloud environment (distinguished by ownership, size, access)
  - **Public Cloud:** publicly accessible cloud environment owned by a third-party cloud provider.
  - **Community Cloud:** similar to public cloud except that its access is limited to a specific community of cloud consumers.
  - **Private Cloud:** owned by a single organization. The same organization is technically both the cloud consumer and cloud provider.
  - **Hybrid Cloud:** comprised of 2 or more different cloud deployment models.

Chap 5. Cloud-Enabling Technology

- **Data Center Technology**

Data center (specialized IT infrastructure) used to house centralized IT resources (servers, databases, networking, telecomunications devices, and software systems).

  - **Virtualization**
    - **Physical layer:** facility infrastructure (computing/networking systems and equipments)
    - **Virtualization layer:** operational and management tools that abstract physical computing/networking resources as virtualized components.
      - ➔ **Easier to allocate, operate, release, monitor and control.**
  - **Computing hardware**
    - rackmounted server arrays and multi-core CPU architectures
    - specialized high-capacity network hardware: content-aware routing, LAN and SAN fabrics, and NAS gateways.
- **Virtualization Technology**
  - **Virulization:** process converting a physical IT resource into a virtual one.
  - **Server virtualization:** process of abstracting IT hardware into virtual servers using virtualization software
- **Web Technology**
  - **Web applications – 3-tier model**
    - **Presentation layer:** the user-interface
    - **Application layer:** implementation of application logic
    - **Data layer:** persistent data stores
- **Multitenant Technology**
  - Enable multiple tentnants to access the same application logic simultanously
- **Service Technology**
  - **Web services – industry standarded:**
    - **Web Service Description Language (WSDL)**
    - **XML Schema Definition Language (XML Schema)**
    - **Simple Object Access Protocol** (SOAP messages= header + body)
    - **Universal Description, Discovery, and Integration (UDDI)**
  - **REST services- 6 design constraints:**
    - **Client-Server**
    - **Stateless**
    - **Cache**
    - **Interface/Uniform Contract**
    - **Layered System**
    - **Code-On-Demand**

- **Service Agents:** Event-driven programs designed to intercept messages at runtime.
    - Active service agents: making changes to the message content or changes to the message path.
    - Passive: do not change message content, capture certain part of its content, for logging, monitoring, or reporting purposes.
- **Service Middleware –** 2 common types of middleware platform:
    - **Enterprise service bus (ESB):** service brokerage, routing, message queuing
    - **Orchestration platform:** host and execute workflow logic that drives the runtime composition of services.

Chap 7. Cloud Infrastructure Mechanisms

7.1 Logical Network Perimeter establishes a virtual network boundary that can encompass and isolate a group of cloud-based IT resources.

- Virtual Firewall:

- Similar to traditional hardware but operates as a software instance
- Actively filter network traffic entering/exiting isolated network.

-> Protect isolated network from unauthorized access, malicious activities

- Virtual Network (acquired through VLANs**):** IT resource isolates the network environment within the data center.

-> Help with traffic management

7.2 Virtual Server: a form of virtualization sofware that emulates a physical server.

- Used by cloud providers to provide individual virtual server instances (same physical server) with multiple cloud consumers.

- **VIM** coordinates the physical servers in relation to the creation of virtual server instances.

-> Uniform Implementation of the virtualization layer

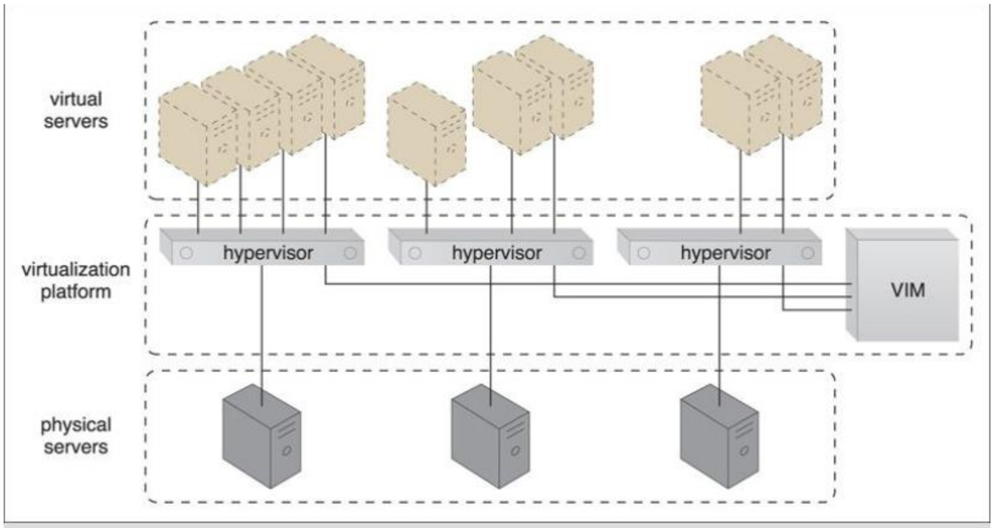
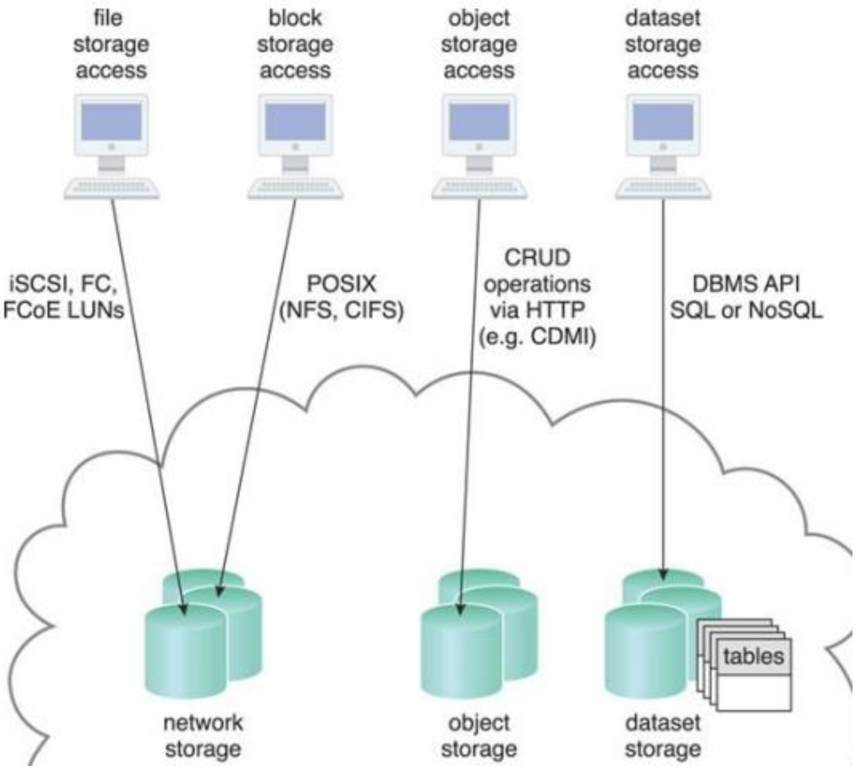


Figure 7.7. Virtual servers are created via the physical servers' hypervisors and a central VIM.

7.3 Cloud Storage Device

- **Cloud Storage Levels**
  - **Files**
  - **Blocks –** smallest unit of data that is still individually accessible.
  - **Datasets –** organized into table-based, delimited or record format.
  - **Objects –** data and metadata organized as Web-based resources.



- **Network Storage Interfaces**
- **Object Storage Interfaces**
- **Database Storage Interfaces**

7.4 **Cloud Usage Monitor:** software program responsible for collecting and processing IT resource usage data.

- **Monitoring Agent:** event-driven program to measure network traffic and message metrics.

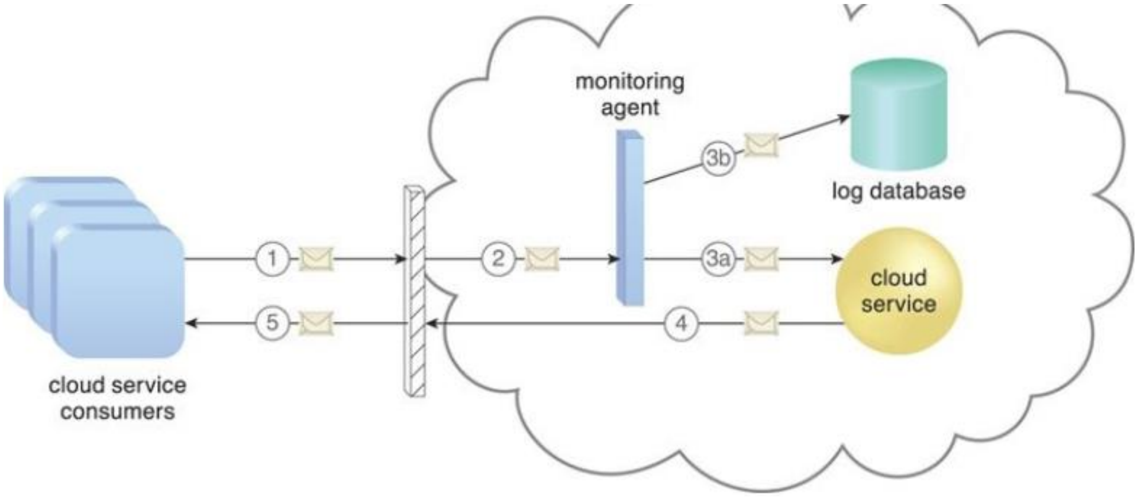


Figure 7.12. A cloud service consumer sends a request message to a cloud service (1). The monitoring agent intercepts the message to collect relevant usage data (2) before allowing it to continue to the cloud service (3a). The monitoring agent stores the collected usage data in a log database (3b). The cloud service replies with a response message (4) that is sent back to the cloud

- **Resource Agent:** a processing module that collects usage data by having event-driven interactions with specialized resource software.
  - Monitor a virtual server and detect an increase in usage.
- **Polling Agent:** a processing module that collects cloud service usage data by polling IT resources.

7.5 **Resource Replication:** creation of multiple instances of the same IT resource

-> Enhance IT resource's availability and performance

7.6 Ready-Made Environment: a component of the PaaS cloud delivery model

- Ready-made environments include pre-installed IT resources, such as databases, middleware, development tools, and governance tools

Chap 8. Specialized Cloud Mechanisms

8.1 Automated Scaling Listener: service agent that monitors and tracks communications between cloud service consumers and cloud services for **dynamic scaling** purposes.

- Workloads: volume of cloud consumer-generated requests or back-end processing demands triggered by types of requests

- **Types of responses** to workload fluctuation conditions:

    - **Auto-scaling**: Automatically scale or based on parameters previously defined by the cloud consumer.
    - **Automatic notification** of the cloud consumer: when workloads exceed current thresholds or fall below allocated resources (fig 8.1)
      -> Cloud consumer can choose to adjust its current IT resource allocation.

8.2 Load Balancer (runtime agent)

- **Objectives**: optimize IT resource usage, avoiding overloads, and maximizing throughput

- **Specialized runtime workload distribution functions:**

    - **Asymmetric Distribution** – larger workloads are issued to IT resources with higher processing capacities
    - **Workload Prioritization** – workloads are scheduled, queued, discarded, and distributed workloads according to their priority levels
    - **Content-Aware Distribution** – requests are distributed to different IT resources as dictated by the request content

- The load balancer mechanisms can exist as a:

    - multi-layer network switch
    - dedicated hardware appliance
    - dedicated software-based system (common in server operating systems)
    - service agent (usually controlled by cloud management software)

8.3 SLA (Service Level Agreement) Monitor: observe the runtime performance of cloud services to ensure that they are fulfilling the contractual QoS (Quality of Service) requirements (published in SLAs).

- Collected data are aggregated into SLA reporting metrics -> repair or failover cloud services

8.4 Pay-Per-Use Monitor: measures cloud-based IT resource usage by predefined pricing parameters and generates usage logs for fee calculations and billing purposes.

- Some typical monitoring variables:

    - request/response message quantity
    - transmitted data volume
    - bandwidth consumption

8.5 Audit Monitor: collect audit tracking data for networks and IT resources in support of regulatory and contractual obligations

8.6 Failover System: automatically switch over to a redundant or standby IT resource instance whenever the currently active IT resource becomes unavailable

- Commonly used for mission-critical programs (like financial system) and reusable services (preventing a single service failure from impacting all dependent applications)

- **2 Basic configurations**:

- **Active-Active**: Redundant implementations of the IT resource actively serve the workload synchronously. When a failure is detected, the failed instance is removed from the load balancing scheduler. Whichever IT resource remains operational takes over the processing
- **Active-Passive:** A standby or inactive implementation is activated to take over the processing from the IT resource, and the corresponding workload is redirected to the instance taking over the operation

8.7 Hypervisor (software): generate virtual server instances of 1 physical server

- A hypervisor is limited to 1 physical server -> only create virtual images of that server

- VIM provides features for administering multiple hypervisors

8.8 Resource Cluster: combine multiple (geographically diverse) IT resource instances into a single IT resource (cluster) -> increase computing capacity, load balancing, and availability.

- Common **resource cluster types** include:

- **Server Cluster**
- **Database Cluster**
- **Large Dataset Cluster**

8.9 Multi-Device Broker: gateways contain the **mapping logic** to

transform data exchanges between a cloud service and different types of cloud

service consumer devices

8.10 State Management Database: storage device that is used to temporarily

persist state data for software programs.

- As an alternative to caching state data in memory, software programs can off-load state data to the database in order to reduce the amount of runtime memory

- State management databases are commonly used by cloud services, especially those involved in long-running runtime activities.

Chap 9. Cloud Management Mechanisms

9.1 Remote Administration System: provide tools and UI for external cloud resource administrators to configure and administer cloud-based IT resources.

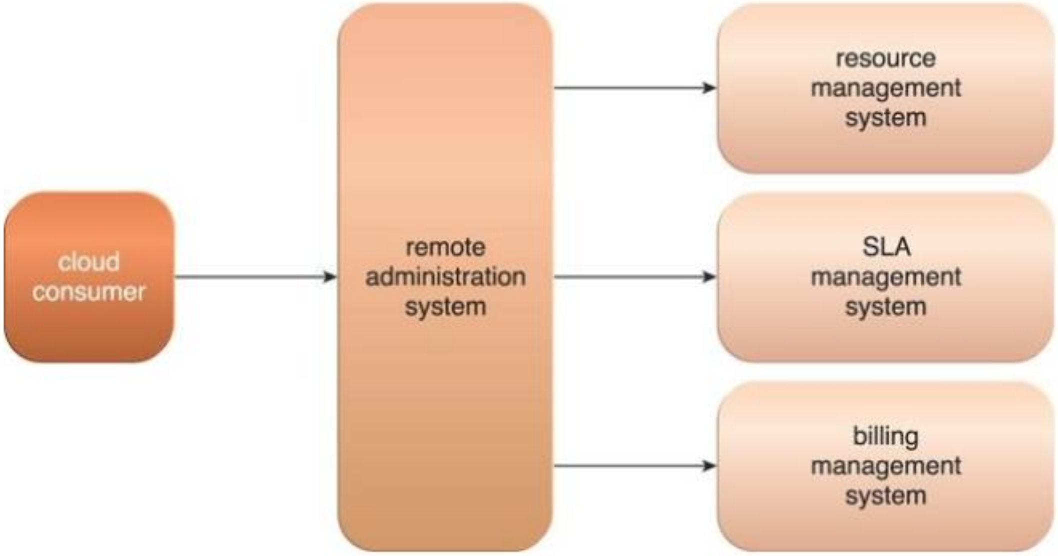

cloud consumer
remote administration system
resource management system
SLA management system
billing management system

Figure 9.2. The remote administration system abstracts underlying management

- Cloud provider uses tool and APIs (provided by remote adminstration) to develop and customize online portal for cloud consumers -> provide administrative controls.

- **2 primary types of portals:**

- **Usage and Administration Portal -** General Purpose Portal:
  - centralizes management controls to different cloud-based IT resources.
  - provide IT resource usage reports.
- **Self-Service Portal** – Shopping Portal that allows cloud consumers to search an up-to-date list of cloud services and IT resources that are available from a cloud provider (usually for lease). The cloud consumer submits its chosen items to the cloud provider for provisioning.