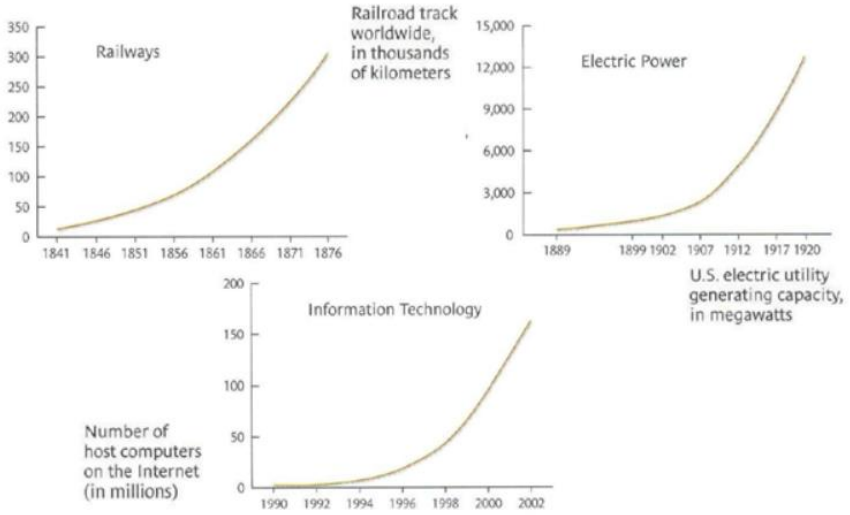
Understanding Cloud concepts

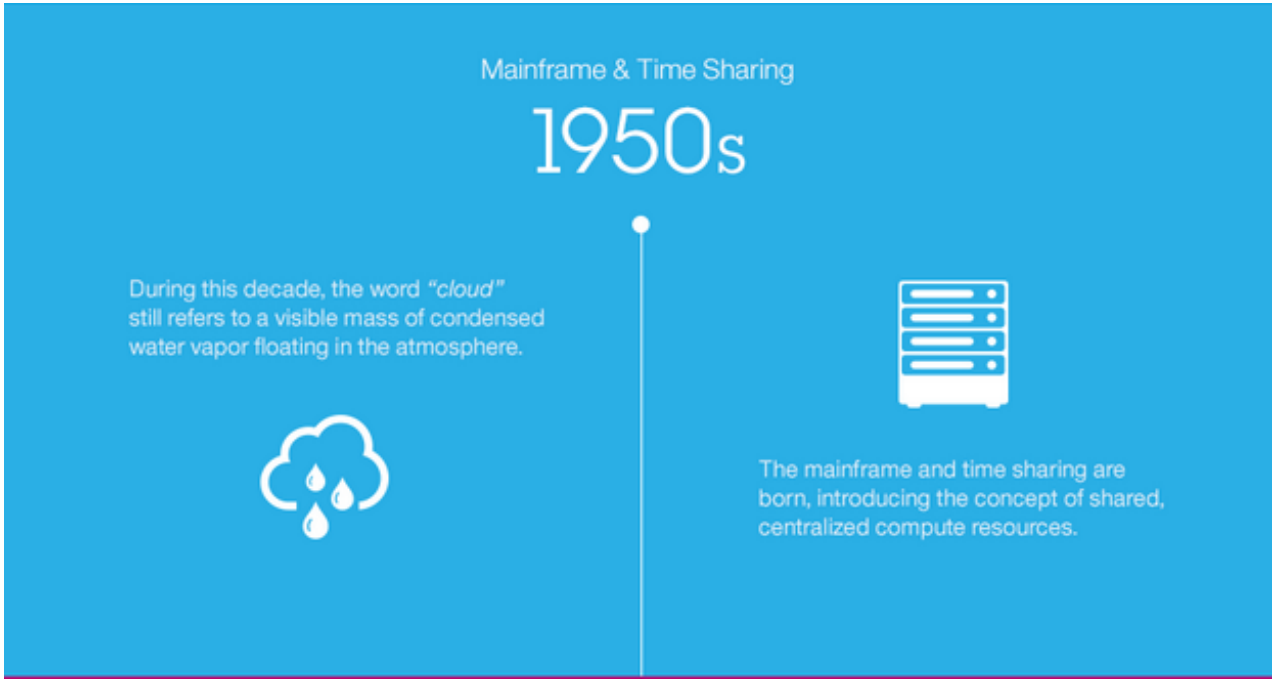
Same needs

- Face growing demand
- Creation of scale economies
- Improve production in an efficient way

Same effects

- Prices reduction
- Increase adoption of technology
- "Standardization" of the service




Mainframe & Time Sharing
1950s
During this decade, the word "cloud" still refers to a visible mass of condensed water vapor floating in the atmosphere.
The mainframe and time sharing are born, introducing the concept of shared, centralized compute resources.

ARPANET

1969

The first working prototype of ARPANET is launched

linking four geographically dispersed computers over what is now known as the Internet.

Client-Server

Late 1970s

The term *"client-server"* comes into use

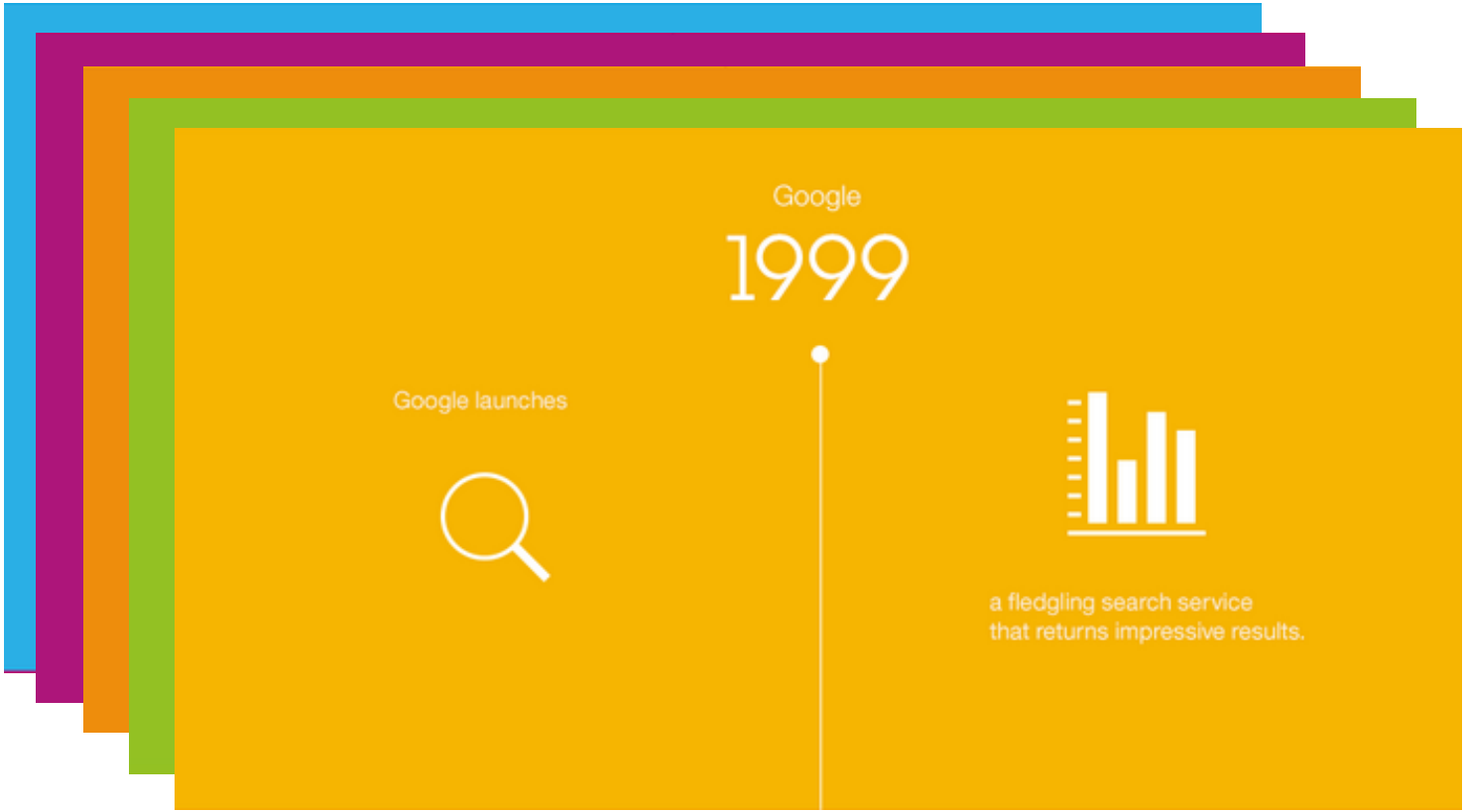
defining the compute model where clients access data and applications from a central server over a local area network.

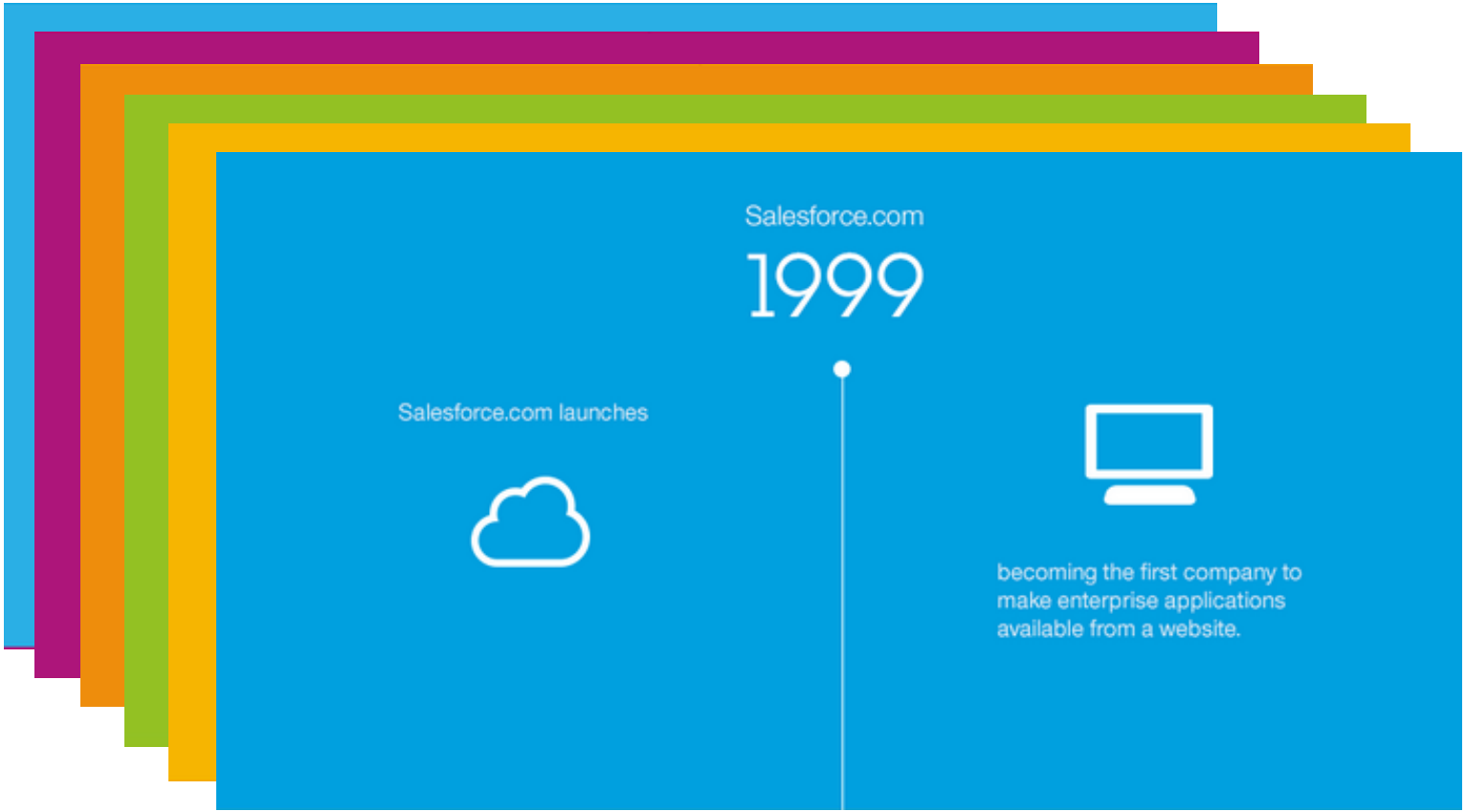
Pictures of Clouds

1995

Pictures of clouds start showing up in network diagrams

denoting anything too complicated for non-technical people to understand.


Google
1999
Google launches
a fledgling search service
that returns impressive results.

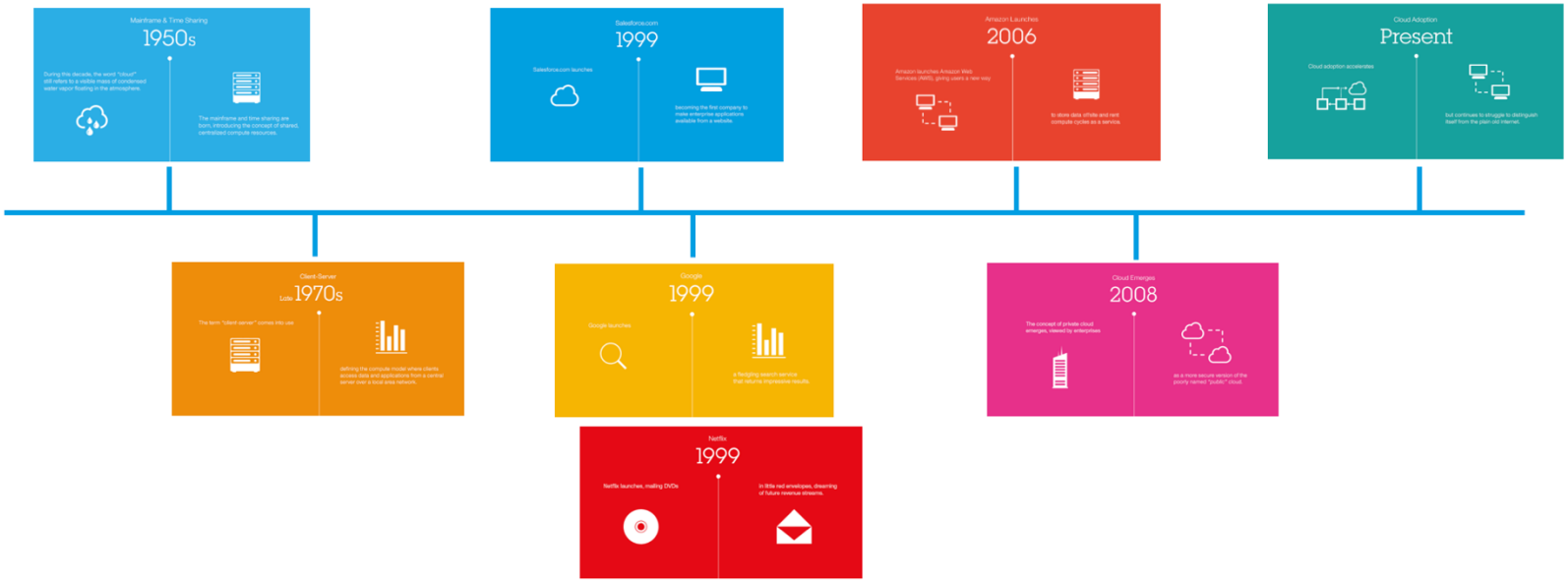

Salesforce.com
1999
Salesforce.com launches
becoming the first company to
make enterprise applications
available from a website.

Amazon Launches

2006

Amazon launches Amazon Web Services (AWS), giving users a new way

to store data offsite and rent compute cycles as a service.



https://www.ibm.com/blogs/cloud-computing/2015/04/05/a-brief-history-of-cloud-1950-to-present-day/

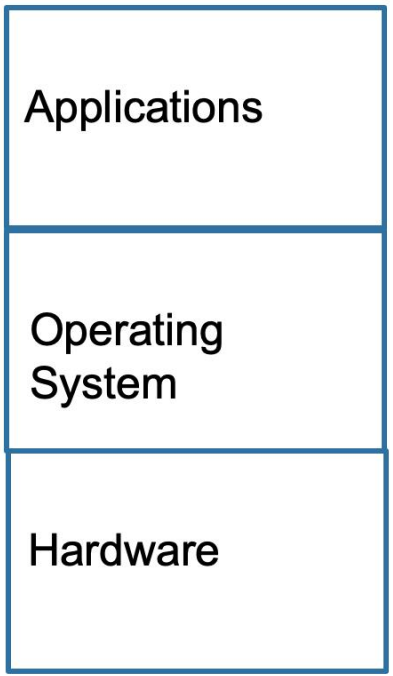
Applications

Operating System

Hardware

- Microsoft Excel
- Windows 7
- Laptop

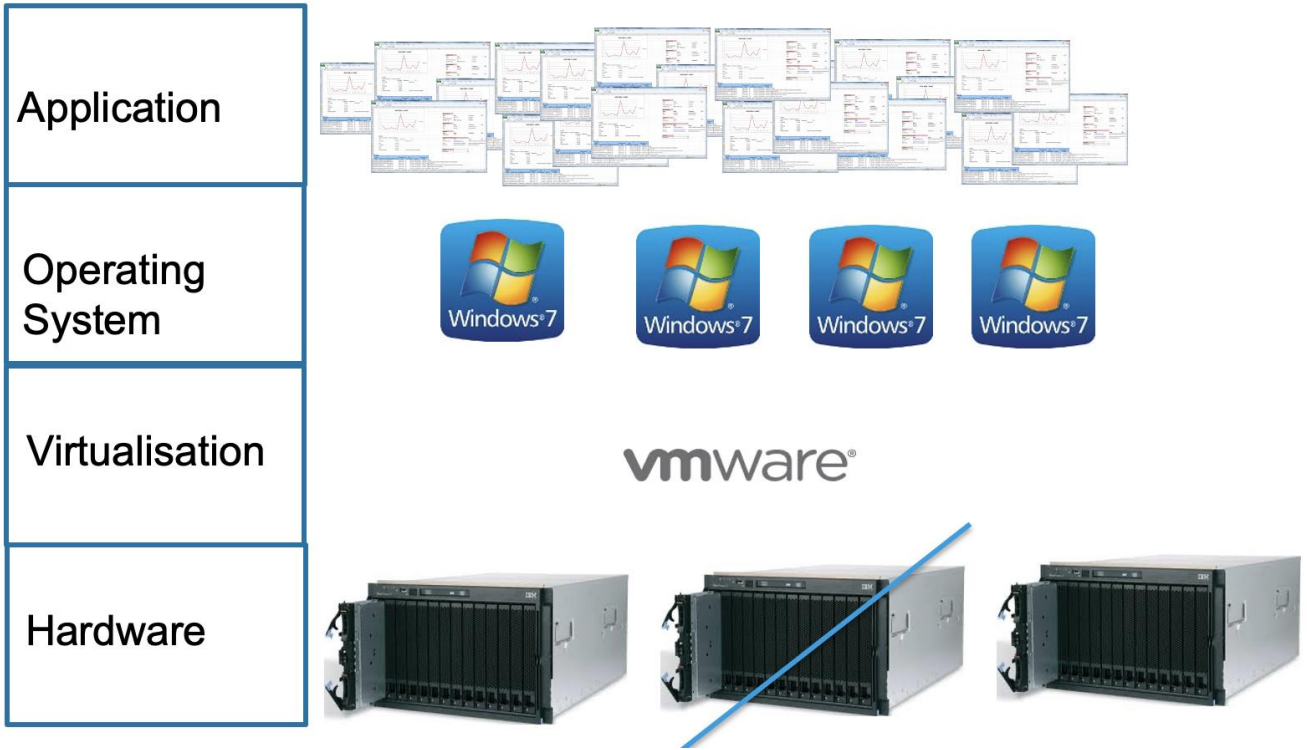
- Your laptop is a computer, CPU, RAM
- You can run more applications on a stack of laptops than you can a single laptop
- A Blade center is a "stack of laptops" managed as one single computer or “Server”



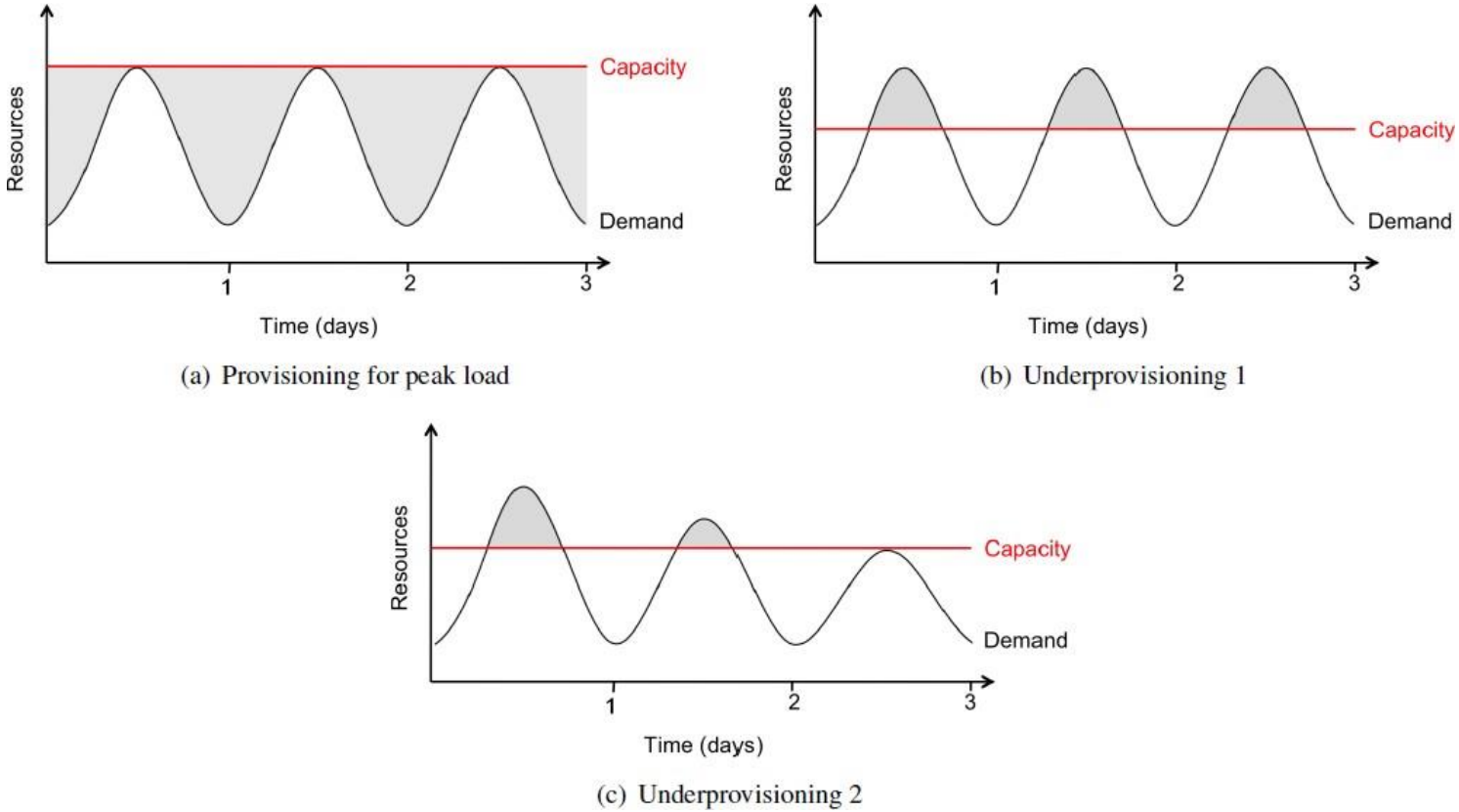
- Microsoft Excel
- Windows 7
- Blade Centre

HARDWARE UTILIZATION

- Imagine that each of these individual servers is being used at 20% of its maximum capacity.
- Therefore,80% of the server is not being used
- For example: CPU, memory, storage under utilised
- What is going to happen in case of a failure?

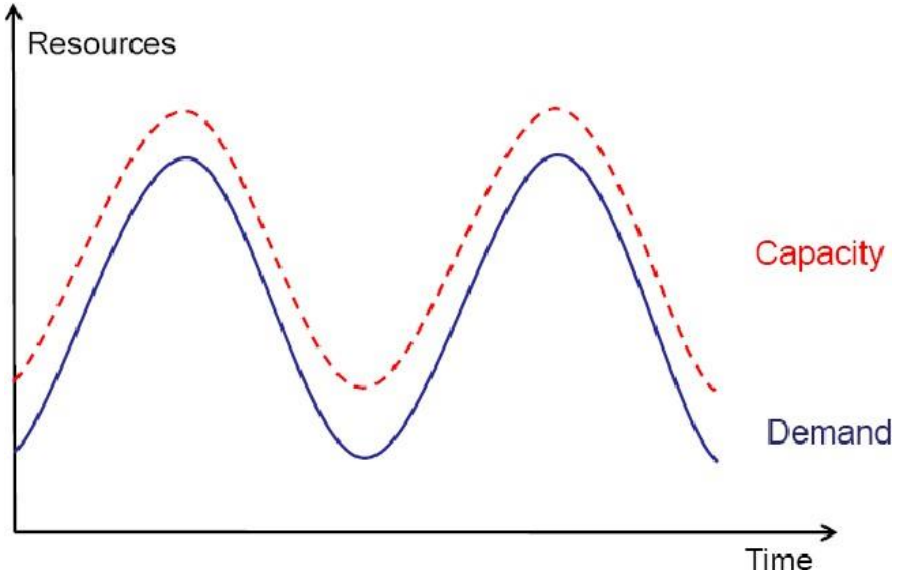

Application
Operating System
Virtualisation
Hardware
Windows 7
Windows 7
Windows 7
Windows 7
vmware

Over- and under-provisioning



source: Above the Clouds: A Berkeley View of Cloud Computing (2009)

Elasticity



Definition

"Classical" definition: *The NIST* Definition of Cloud Computing* (2011)

> **Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.**

This cloud model is composed of 5 essential characteristics, 3 service models, and 4 deployment models.

(*) National Institute of Standards and Technology (U.S. Department of Commerce)

source: NIST Definition of Cloud Computing (2011)

Ecosystem of Cloud Computing

Three categories of players (people):

- **consumers of services**: end-users that use cloud services in their day-to-day business activities. They may have little understanding of where the service resides or how it is designed; they simply need the capabilities to get the job done
- **providers of services**: cloud providers offer a variety of functions ranging from infrastructure services to applications and tools
- **designers of services**: companies that build applications and tools. Often services are intended to work within a specific cloud ecosystem or can augment a packaged cloud application

Alert: these categories can mix-up! In fact, designers of services are often “consumers” from the viewpoint of the cloud providers

Five essential characteristics

1. **On-demand self-service**. A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
2. **Broad network access**. Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).
3. **Resource pooling**. The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a **sense of location independence** in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.
4. **Rapid elasticity**. Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.
5. **Measured service**. Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

source: NIST Definition of Cloud Computing (2011)

Service models

1. **Software as a Service (SaaS).** The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings. Examples: Microsoft 365, Google Apps
2. **Platform as a Service (PaaS).** The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment. Examples: Heroku, Google Apps Engine
3. **Infrastructure as a Service (IaaS).** The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls). Examples: Amazon EC2, Google Compute Engine

• • • [more on this topic in the next Unit…] • • •

source: NIST Definition of Cloud Computing (2011)

Deployment models

1. **Private cloud**. The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.
2. **Community cloud**. The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.
3. **Public cloud**. The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.
4. **Hybrid cloud**. The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

source: NIST Definition of Cloud Computing (2011)

Cloud advantages

***Designer of services* perspective:**

1. The appearance of infinite computing resources on demand.
2. The elimination of an up-front commitment by cloud users.
3. The ability to pay for use of computing resources on a short-term basis as needed.

***Provider of services* perspective:**

4. Economies of scale that significantly reduced cost due to many, very large data centers.
5. Simplifying operation and increasing utilization via resource virtualization.
6. Higher hardware utilization by multiplexing workloads from different organizations.

source: Above the Clouds: A Berkeley View of Cloud Computing (2009)

Virtualization

Computing resources can be virtualized.

- **Virtual Machine (VM)**: is an emulation of a computer system. A piece of software “pretends” to be hardware
- OS-level virtualization (a.k.a. **Container**): is an isolated user-space instance within an OS
- **Sandbox**: is a security mechanism for separating running program

• • • [more on this topic next…] • • •

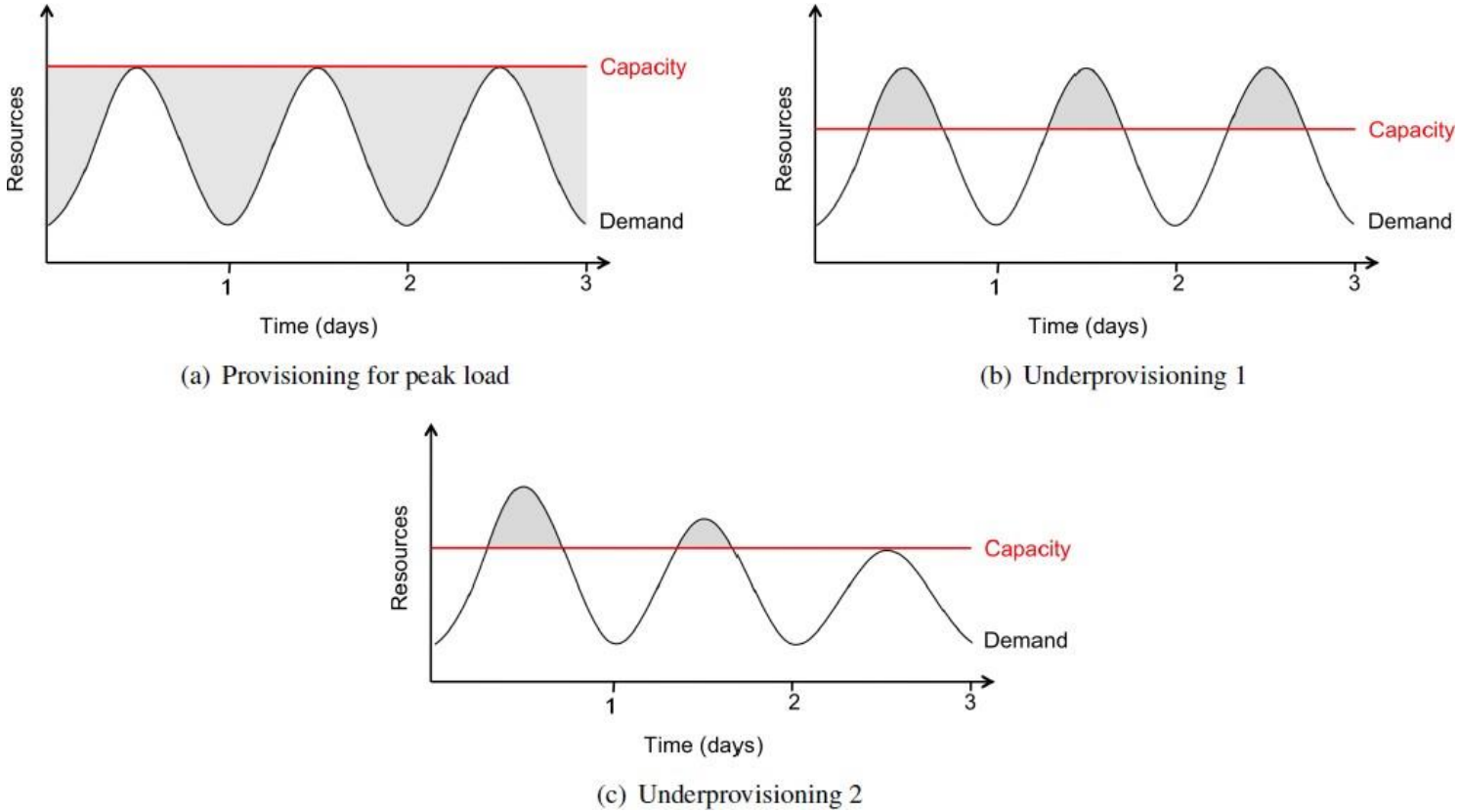
Self-service provisioning

Self-service **provisioning** is one of the most important capabilities of cloud computing.

With self-service, **cloud consumers can select and purchase cloud services, configure them, launch them into the cloud environment, and start using them within minutes (or perhaps even seconds)**.

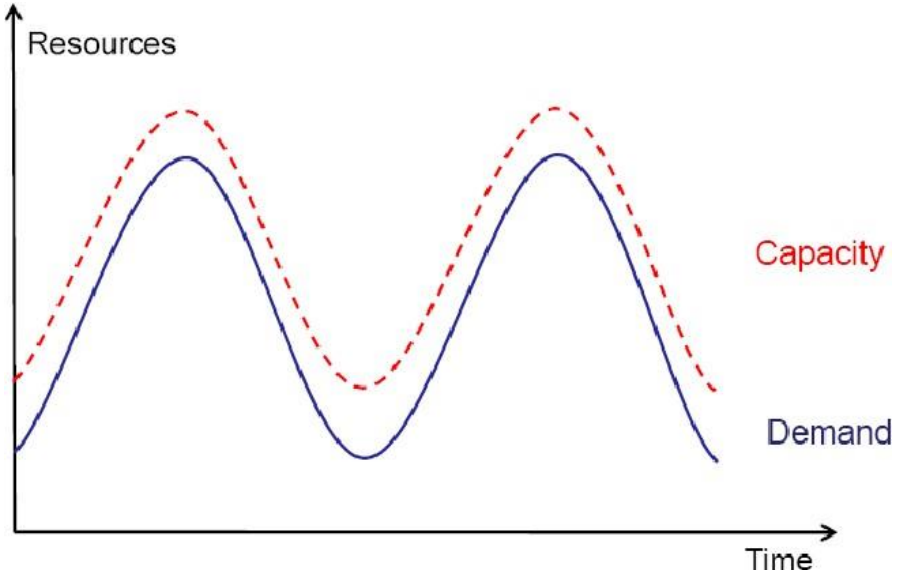
In the traditional data center model, that same consumer might have to file a request with IT operations for equipment or software, go through approval and payment processes, and then wait while IT procures the equipment, installs it, and configures it, and finally turns it over to the requesting consumer for use. The data center procurement process can take days, and often weeks.

Over- and under-provisioning



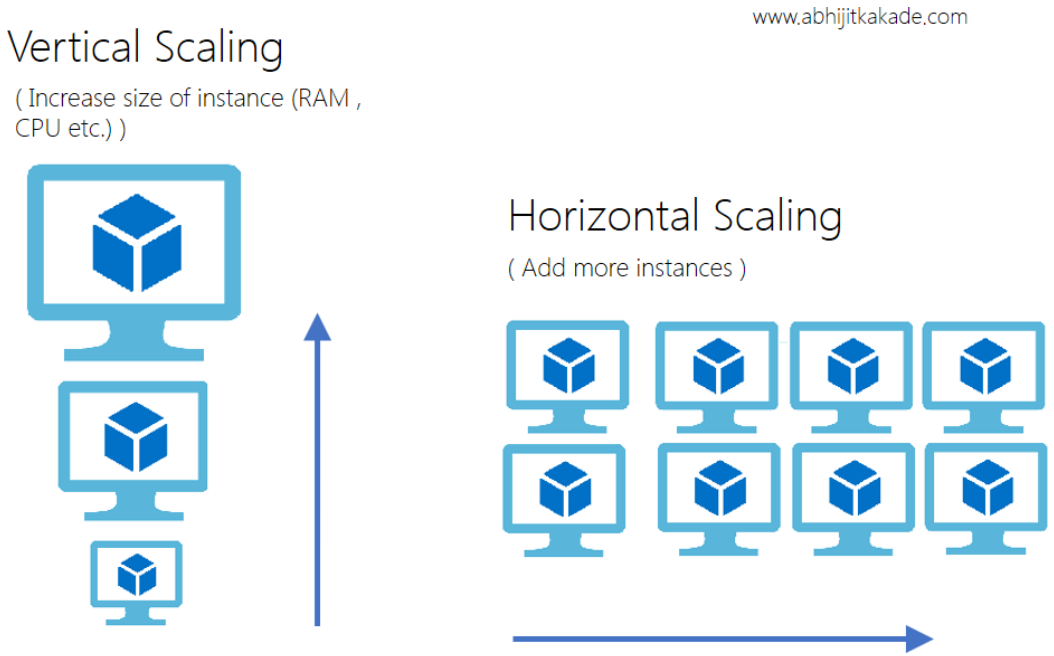
source: Above the Clouds: A Berkeley View of Cloud Computing (2009)

Elasticity



Scalability

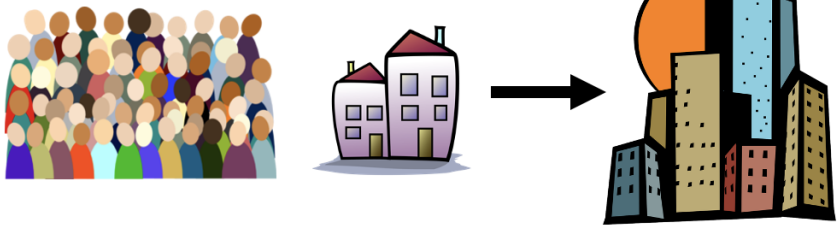
Scalability is the property of a system to handle a growing amount of work by adding resources to the system.



source: http://abhijitkakade.com/2019/04/horizontal-vs-vertical-scaling-azure-autoscaling/

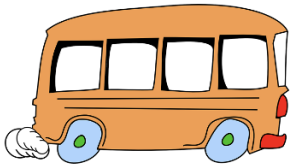
Scalability

32 People need to go from town to the Big City today



BUT

Your bus only holds 20 people

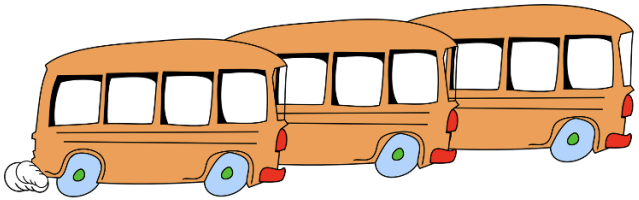


Vertical Scaling (1 bigger instance)

Double-decker bus holds 40 people

Horizontal Scaling (multiple instances)

More buses - 20 people each



Which solution is better?

Scalability

For Cloud-native point of view - multiple instances are preferred – Why?

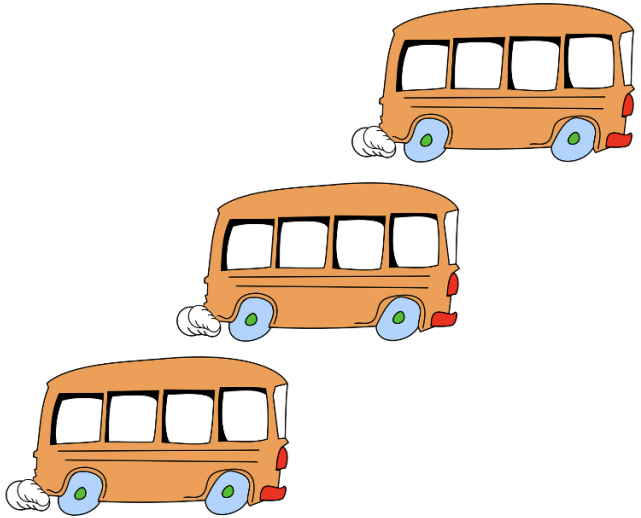
No single point of failure
- If a bus breaks down, you still have other buses
- Can activate another bus to carry the stranded passengers

Elastic scalability
- You can add or remove buses from service to meet the demand

Load distribution
- Passengers don't care which bus they go on, as long as they reach their destination

Flexibility
- Some buses can use different roads if there is a problem on the default route

Availability
- You can take down buses for maintenance, and still provide your service with other buses



Autoscaling

The amount of computational resources in a cloud environment (e.g., number of computing units – VMs – or the RAM of a computing unit) varies automatically based on the computational load.

Autoscaling ❼ “perfect” elasticity

Reading list

- P. M. Mell, T. Grance, NIST Definition of Cloud Computing (2011)
- M. Armbrust *et al.*, Above the Clouds: A Berkeley View of Cloud Computing (2009)

Questions, comments?