Warehouse-Scale Computing

Ozalp Babaoglu

Warehouse-Scale Computing

- What is a warehouse-scale computer?
- Warehouse-scale computers (WSCs) form the foundation of most internet services that we use today: searching, social networking, navigating, video sharing, online shopping, email, and in general, cloud computing
- They are also the basis for high-performance computing (HPC) for doing basic sciences
- The name was coined by Barroso and Hölzle of Google
  - Barroso, Luiz André, Urs Hölzle, and Parthasarathy Ranganathan. "The Datacenter as a Computer: Designing Warehouse-Scale Machines", Third Edition, Morgan & Claypool Publishers series *Synthesis Lectures on Computer Architecture* Lecture #24 (2018)

Warehouse-Scale Computer: Exterior

Warehouse-Scale Computer: Interior Cloud

Warehouse-Scale Computer: Interior HPC

Course Outline

- Examine the technical challenges in the design and operation of WSC
- How are WSC built and programmed?
- What makes them cost-effective and attractive for businesses?
- Why have they become so ubiquitous?
- Who are the major players that operate them?
- How do they store and process huge collections of information?
- How can they be made secure?
- How can they be made fault tolerant?
- How can their energy consumption be contained to practical levels?
- Which software technologies are most appropriate for their efficient operation and management?

Course Outline

- Hardware ***infrastructures*** for WSC
  - compute and storage elements
  - networking fabric
  - energy efficiency and heat dissipation
  - failures, redundancy, availability
- WSC software ***technologies***
  - web services
  - virtualization
  - containerization
  - scaling and fault tolerance
  - database

Course Outline

- WSC and cloud computing
  - Technological and business opportunities offered by cloud computing
  - Cloud economics
  - Cloud ***deployment*** models
  - Cloud ***computing*** models
  - Potential risks of cloud computing and challenges for cloud adoption
  - Modern cloud computing landscape
  - Cloud ***dependability***, cloud ***outages***, cloud ***forensics***

Administrative Info

- Lecture Schedule: 6 – 9 March 11.00 — 14.00
- Slides of lectures and pdf of the textbook are available from the course web site: https://www.cs.unibo.it/babaoglu/wsc/
- If you are taking the course for credit, you will have to make arrangements with me for grading options, which can be any one of written report on a subject that is relevant to the course and that is of interest to you or an oral presentation of a research paper

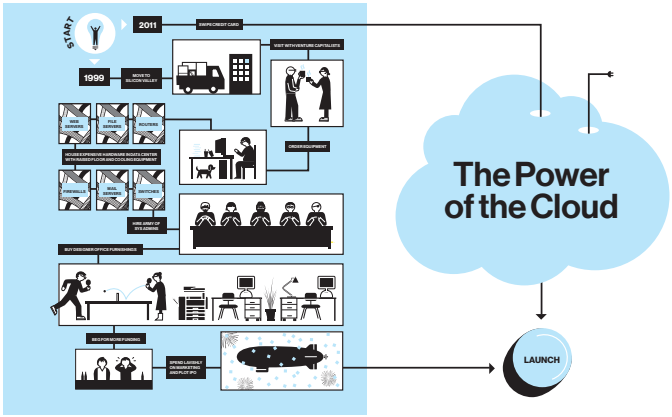
Tech Company Startup Scenario circa 1990s

- You want to start a business around a ***killer app*** that you have developed
- Move to Silicon Valley
- Pitch your idea to venture capitalists and secure funding
- Rent (expensive) office space
- Order (expensive) computing, storage and networking hardware
- Build ***on-premise*** infrastructure by hiring (expensive) system administrators to install, configure, monitor and maintain the hardware
- Advertise
- Launch your product, dream of *IPO*
- Buy more (expensive) hardware because of increasing demand
- Keep hardware and software up to date, worry about outages and security breaches

Tech Company Startup Scenario Today

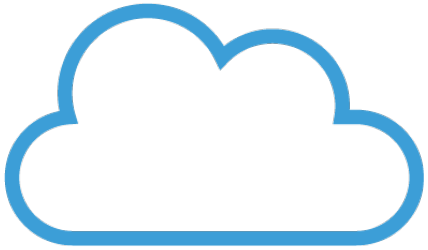
- Come up with the idea for a killer app and develop it
- Buy resources from a ***public cloud service provider*** from the comfort of your living room
- Deploy your app on the cloud servers
- Advertise
- Launch your product, dream of *IPO*
- Sit back and let the cloud service provider worry about ***software updates***, ***hardware maintenance***, ***resource provisioning*** to meet changing demand, ***security measures*** and ***monitoring***

Why Cloud Computing?



Why the “Cloud”?

- Historically, the “cloud” symbol was used to indicate anything that is “non local”, typically residing somewhere in the Internet
- “In the cloud”



What is Cloud Computing?

- Clouds are the ***utilities*** for computing, just like conventional utilities for water, gas and electricity
- ***Cloud computing*** is a remote virtual pool of on-demand shared resources offering compute, storage, database and network services that can be rapidly deployed at scale
- ***Cloud computing*** is ***on-demand delivery*** of IT resources ***over the Internet*** with ***pay-as-you-go pricing***
- ***Cloud computing*** allows ***hardware*** to be treated (licensed, installed, configured, initialized, sized) just like ***software***

What is Cloud Computing?

The illusion of ***infinite*** computing resources available ***on demand***, thereby eliminating the need for users to plan far ahead for provisioning

The elimination of an ***up-front*** commitment, thereby allowing companies to start small and increase hardware resources only when there is an increase in their needs

The ability to ***allocate*** and ***release*** resources ***as needed*** on a ***short-term*** and ***pay as you go*** basis

A Paradigm Shift

- Cloud computing represents a ***paradigm shift*** from ***local*** to ***network-centric computing*** and ***network-centric content***
- In this new paradigm, users relinquish control of their ***data*** and ***code*** to ***Cloud Service Providers***
- Cloud computing has ***reshaped*** the business technology landscape more than any other force in recent years
- It has transformed the information technology industry by making ***software central as a service*** and by shaping the way ***hardware*** is designed and purchased

Attributes of Cloud Computing

- Principal attributes of cloud computing:
  - on-demand, self-service
  - scalability (both scale-up and scale-out)
  - rapid elasticity
  - shared infrastructure and resource pooling
  - metered service for pay-as-you-go
  - high availability (redundancy, check-point/restart)
  - high security

Attributes of Cloud Computing: On-demand Resourcing

- No ***up-front*** investments
- Consume resources only when you need them (***on-demand***) and without requiring any intervention by the cloud service provider (***self-service***)
- Particularly important when demand is ***not*** known in advance

Attributes of Cloud Computing: Scalability

- Services can be scaled to keep up with increasing demand
  - *Scale-up*: increase the ***capacity*** of resources allocated to the service (also know as ***vertical scalability***) — ***faster*** servers, ***more*** memory
  - *Scale-out*: increase the ***number*** of resources allocated to the service (also know as ***horizontal scalability***) — ***more*** servers to exploit parallelism

Attributes of Cloud Computing: Rapid Elasticity

- *Elasticity* is the ability to ***add*** or ***remove*** resources to better match the demands of a dynamic workload
- *Elasticity* = *Scale-out* + *Scale-in*
- *Rapid elasticity* is the ability to add or remove resources at a ***fine grain*** (one server at a time) and with a small lead time (***minutes, seconds*** or ***milliseconds***) rather than ***weeks***

Attributes of Cloud Computing: Shared Infrastructure

- Resources are ***pooled*** to provide a ***shared infrastructure*** to facilitate elasticity for a large number of users
- Software technologies such as ***virtualization*** are necessary to maintain the ***isolation*** among users

What is Cloud Computing: Economy of Scale

- Unprecedented ***economies of scale*** are possible by operating extremely large infrastructures
- Exploit ***volume discounts*** for hardware, software, real estate, energy, personnel
- Fixed costs can be ***amortized*** over large number of users

What is Cloud Computing: Growth and Global Reach

- Scalability and elasticity are conducive to growth
- Users automatically inherit the global presence of the cloud service provider to get closer to customers

Attributes of Cloud Computing: Metered Service for Pay-as-you-go

- Fine-grained metering (minutes, seconds) admits paying for resources only for the period that they are actually used
- "***Pay-for-what-you-eat***" model
- "***Cost associativity***" — "1,000 server hours" of credit can be spent either as "1,000 servers for 1 hour" or "1 server for 1,000 hours"
- For "embarrassingly parallel" tasks, the first choice is a much better choice

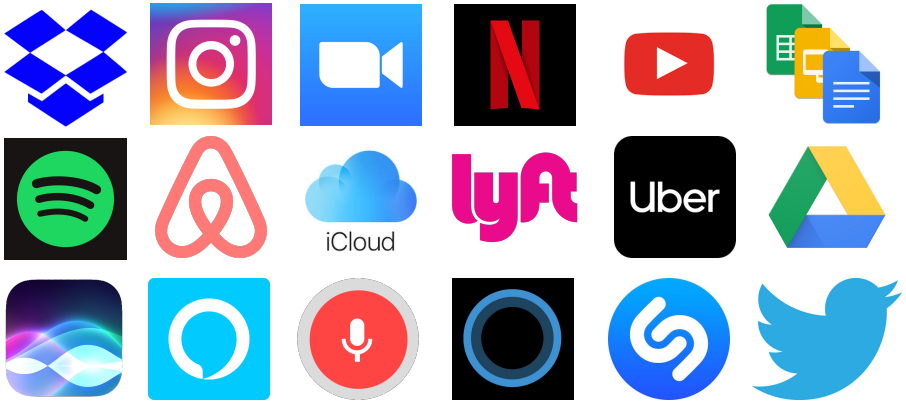
Attributes of Cloud Computing: High Availability

- Fault tolerance techniques such as ***redundancy*** and ***check-point/restart*** can be used to build cloud services that are highly available
- ***Geographic distribution*** of cloud service provider's infrastructure increases ***failure independence*** to reduce probability of total outages

Attributes of Cloud Computing: High Security

- The cloud service provider can invest heavily in ***physical security*** of its facilities
- It can also invest heavily in ***automated tools*** and ***personnel*** to monitor and defend its infrastructure from ***cyber-attacks***
- There are no guarantees and much rests in our ***trust*** towards the cloud service provider

Notable Cloud-Based Applications



Notable Cloud Customers

Notable Cloud Service Providers

Why the Cloud?

- Ability to rapidly deploy applications and services
- Scale to meet peak demands
- Increase revenue, efficiency, and reduce cost

Why the Cloud?

- Cloud computing is what the business world calls a “no-brainer”
- It helps cut costs, offers enhanced security and stability and gives companies greater flexibility
- According to one estimate, by 2023 50% of all business workloads are expected to run in some form of cloud environment

Why the Cloud?

- By adopting cloud computing, a business is transformed from a ***capital expenditure*** (CapEx) to ***operating expenditure*** (OpEx) model
- Allows an organization to focus on its ***core business*** instead of IT operations
- Business built on ***standardized***, ***up-to-date*** and consolidated IT
- Transfers ***investment***, ***risk*** and ***human resources*** to the cloud provider
- Offers high levels of ***availability*** and high levels of ***security***
- Allows a business to become instantly ***global*** and more ***flexible***

Cloud Provider World-Wide Locations

Why Now?

- Unprecedented ***economies of scale*** through large data centers
- Pervasive ***broadband Internet***
- Fast ***x86 virtualization*** ("x86" refers to the instruction set architecture of the Intel 8086 processor that became an industry standard)
- Wide availability of ***open-source software*** systems (Linux, Apache)

Cloud Computing in a Pandemic

- The COVID-19 health crisis further accelerated the shift to cloud
- To rapidly enable remote workers at scale, many businesses turned to cloud-based solutions
- At the height of the pandemic, approximately 60% of the global workforce was operating in a work-from-home environment
- It is estimated that 30% of the workforce will go remote permanently in the "new normal" following COVID-19 — nearly doubling pre-pandemic work-from-home numbers

Challenges for Cloud Adoption

- Vendor lock-in
- Security, confidentiality and accountability
- Data transfer bottlenecks
- Performance unpredictability
- Service outages
- Cross-border compliance
- Bugs in large distributed systems
- Large-scale systems are affected by phenomena characteristic to ***complex systems*** — "unintended consequences" or "emergence"

Challenges for Cloud Adoption: Vendor lock-in

- Once a customer is hooked to one cloud service provider, it may be difficult to move to another
- Opportunities:
  - ***Standardization*** of cloud interfaces will simplify moving to other providers
  - *Multi-clouds*

Challenges for Cloud Adoption: Security, confidentiality and accountability

- Sensitive, private data that is handed over to the provider may be ***revealed*** to unauthorized parties, data may be ***lost*** or ***corrupted*** and it may be difficult or impossible to hold the provider ***accountable***
- Opportunities:
  - *Encryption* technologies
  - *Firewalls*
  - *Virtual Private Networks*
  - *Data replication*

Challenges for Cloud Adoption: Data transfer bottlenecks

- Data-intensive applications may become impractical in the cloud
- Transferring 1TB of data on a 1Mbps network takes 8,000,000 seconds or about 10 days
- At 1Gbps, this time reduces to 8,000 seconds, or slightly more than 2 hours
- Opportunities:
  - Use a ***courier service*** (FedEx) to send hard disks instead of sending data over the network
  - Invest in high-speed networks
  - Edge-computing, 5G

Challenges for Cloud Adoption: Performance Unpredictability

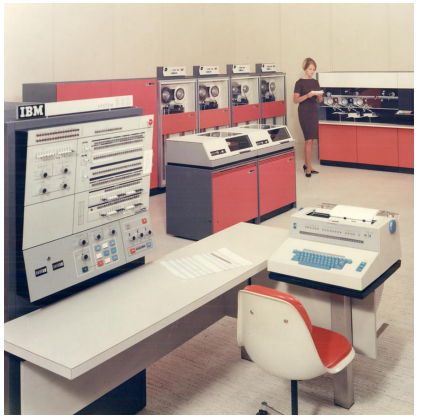
- In a shared infrastructure, performance of individual applications may become unpredictable due to interference from other applications
- Opportunities:
  - Better ***physical isolation*** — over provisioning
  - Better ***virtual isolation*** — virtualization

Challenges for Cloud Adoption: Service Outages

- What happens when the service provider cannot deliver?
- While rare, outages of a provider's entire infrastructure are not unheard of
  - Power outages in the electrical grid
  - Lighting strikes
  - Flooding after tropical storms
- Opportunities:
  - Stipulate stringent ***Service Level Agreements*** (SLAs) with the provider
  - ***Geo replication*** — deploy service on multiple "availability zones" of cloud provider

History of Cloud Computing

- Early "time sharing" systems with large mainframe computers housed in machine rooms



History of Cloud Computing

- Users run programs installed on the mainframe through "dumb" terminals
- Text only, 80 columns by 24 lines
- All data stored and computed on the mainframe

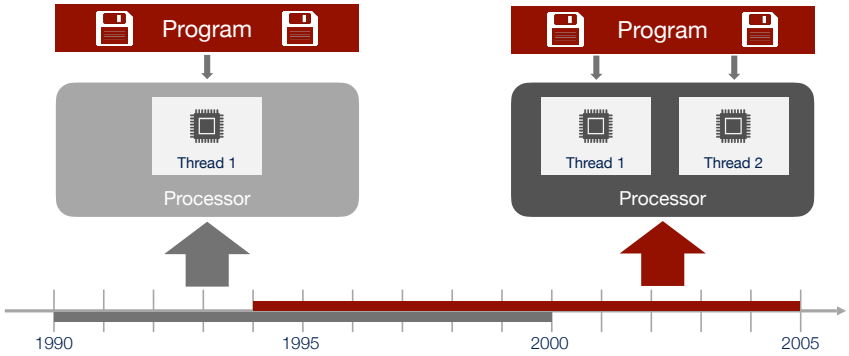
History of Cloud Computing

- With advent of workstations, move to a "client-server" model
- ***Clients*** now run on ***workstations*** (managing the user interface, other graphics) and access ***servers*** running on the mainframe that do the heavy computing

History of Cloud Computing

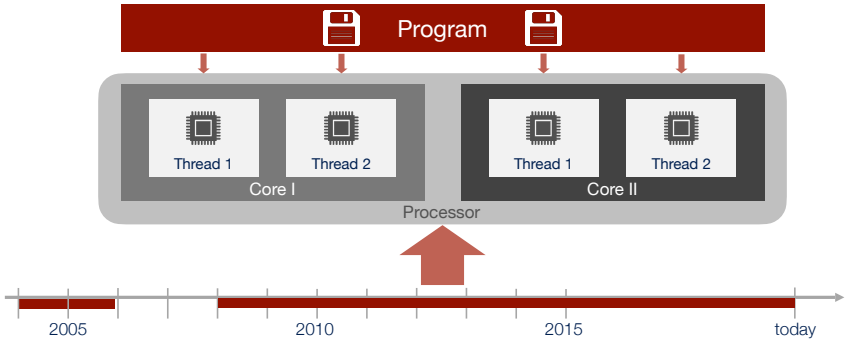
- An ***Operating System*** (Windows, Unix, Linux) creates the illusion that each user alone is using the mainframe
- Resources (CPU, memory, disk, network bandwidth) are ***multiplexed*** among the different programs running on behalf of users
- On the server, application-level parallelism exploited through "threads"

Server Side Evolution: Single to Multi-threaded Execution



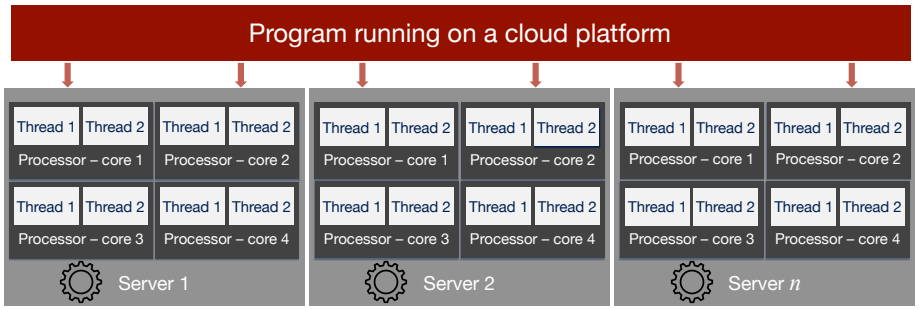
- Multiple ***threads*** on the same processor run ***concurrently*** (not in parallel)

Server Side Evolution: Multi-threaded to Multi-core Execution



- Multiple threads on multiple cores run ***in parallel***

Server Side Evolution: Multi-threaded, Multi-core, Multi-server



- Massive ***parallelism*** with ***multiple servers*** (processors), ***multiple cores*** and ***multiple threads***

WSC vs HPC

- "HPC clusters" are collections of independent computers that are connected together using standard LANs and off-the-shelf switches
- In a way, WSCs are just larger, more regular clusters — the regularity is necessary when there are tens of thousands of servers that must be maintained
- Yet, WSCs are also different from HPC clusters
  - HPC clusters generally have much faster processors and much faster networking
  - HPC applications are generally much more interdependent and communicate more frequently
  - HPC emphasizes thread-level or data-level parallelism, emphasizing latency for a single task rather than bandwidth for independent, request-level tasks
  - HPC clusters generally have long-running tasks that have high server utilization, whereas WSC utilization is typically less than 50%

Enabling Technologies for Clouds

- Hardware technologies
  - Commodity computing units (servers)
  - Commodity storage elements (both ***magnetic*** and ***solid state***)
  - High-speed local-area networking (Gigabit Ethernet, fibre optic)
  - Broadband wide-area networking (Internet)
- Software technologies
  - Computing — Virtualization, containerization
  - Performance, reliability — Scale-out, replication, check-point/restart
  - Storage — RAID, NAS, SAN
  - Database — NoSQL, MapReduce

Server Farms, Data Centers

- Hardware technologies that enable cloud computing have given rise to ***Warehouse-Scale-Computers*** (WSC)
- WSC are built as ***server farms*** located in ***data centers***
- Data centers can be as large as 20,000 m^2 (2 hectares) and house as many as 100,000 servers mounted on standard 19-inch racks that are 73,5 inches high

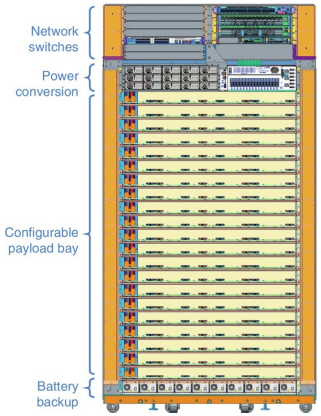
Data Center — Rackmount Server

1U (Rack unit) = 44.45mm is a unit of measure for the height of rack-mountable equipment such as servers

1U Rackmount Server - Dual Xeon E5 - 4 x 2.5-inch Hot-swap Bays	
122D10L	1221D10L-UV5C
Chassis	1U Rackmount Chassis (22-inch deep, 4 x 2.5-inch hot-swap drive bays)
Motherboard	Supermicro® X10DRL-i Motherboard
Processor	2 x Intel Xeon E5-2630v4 (10-Core, 2.20GHz, 25MB Cache) - 20-Cores / 40 Threads total
Memory	64GB (4 x 16GB) - DDR4 ECC Registered Memory
OS Drive	2TB SSD - Endurance: 1,200 TBW -- Samsung 860 EVO
	No Riser Selected (Riser required to use expansion slots)
10GbE LAN Card	No 10 GbE Network Adapter Selected
Rack Rails	Standard Rack Mounting Kit and Rails (Included)
Compatibility	FreeBSD
OS	FreeBSD 12.0 - QA install, no media, community support
Installation	No Windows installation required (Arbitrary test OS will remain for QA)
Warranty	1 Year Limited Warranty, Return to Depot, Parts & Labor
Notes	
	Price as Configured: $4,198.97

Data Center — Google Rack

- Servers are typically mounted in racks that are 19 inches wide and 42U high (2 m × 1.2 m × 0.5 m)



Data Center — Server Farms

- Inside a Data Center https://youtu.be/XZmGGAbHqa0?t=126

Client Side Evolution

- Workstations evolved into "thin clients" running on systems with limited resources (CPU, memory, network bandwidth)
  - 1990s — "X terminal"
  - 2011 — Google's Chrome OS on inexpensive laptops
  - Today — Mobile devices on our laps or in our pockets

History of Cloud Computing

- Cloud computing as we know it today is intimately tied to ***Amazon***
  - 1994 — Amazon founded as an online book retailer
  - 1998 — Business expanded to music and videos (CDs, DVDs)
  - 2002 — ***Amazon Web Services*** (AWS) started to provide data on web site popularity, Internet traffic patterns and statistics
  - 2006 — Amazon launches ***Elastic Compute Cloud*** (EC2) and ***Simple Storage Service*** (S3) to monetize under-utilized resources in its e-commerce IT infrastructure
- 2010 — ***Microsoft Azure*** launched
- 2011 — ***Google Cloud Platform*** (GCP) launched

Cloud Economics

Cloud Economics: Provider's Point of View

- Large data centers offer unprecedented economies of scale for the cloud service provider
- It may be 5-7 times more economical to operate a ***very large*** data center (~100,000 servers) when compared to operating a ***medium size*** data center (~10,000 servers)

Resource	Cost in Medium Size Data Center	Cost in Very Large Data Center	Ratio
Network	$95 / Mbps / month	$13 / Mbps / month	7.1x
Storage	$2.20 / GB / month	$0.40 / GB / month	5.7x
Administration	≈140 servers/admin	>1,000 servers/admin	7.1x

Cloud Economics: Provider's Point of View

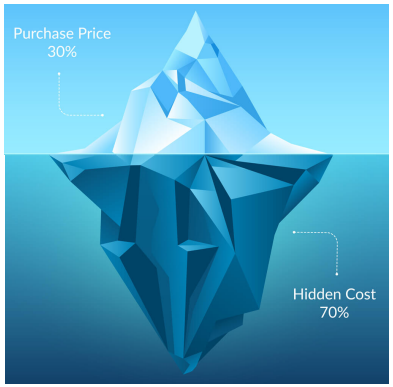
- From the cloud provider's point of view, ***economies of scale*** argue for ***bigger and bigger*** data centers
- Aside from capital investment considerations for real estate, facilities, equipment and personnel, there are limits imposed by ***power consumption*** and ***heat dissipation***

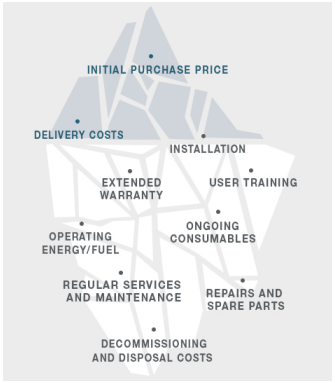
Cloud Economics: Total Cost of Ownership

- In comparing the economics of ***on-premise IT*** to ***cloud computing***, we need to consider ***Total Cost of Ownership*** (TCO)
- TCO includes not only the ***upfront purchasing cost*** of hardware and software, it also includes costs for ***design*** and ***deployment***, ***ongoing infrastructure*** (software maintenance, upgrades, hardware replacement every three years), ***ongoing ops*** (hiring, training, certifying personnel to operate, administer, monitor infrastructure)
- *TCO* = *CapEx* + *OpEx*

Cloud Economics: Total Cost of Ownership

- "30-70" rule





Cloud Economics: Total Cost of Ownership

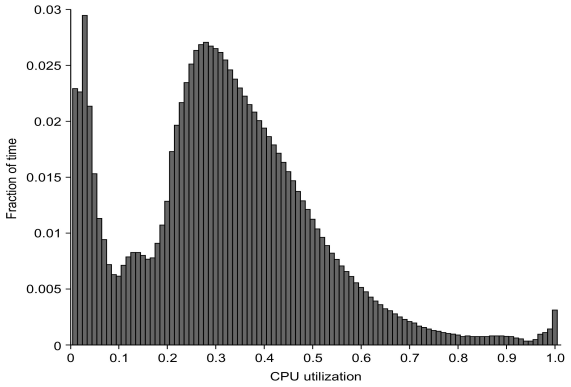
- Moving to the cloud, much of the *CapEx are eliminated* and much of the *OpEx are transformed* into *subscription* fees

Cloud Economics: Keeping Servers Busy

- For the cloud provider who has invested large sums of money in servers, the busier they are, the better
- The provider would like to maintain server utilization as close to 100% as possible, eliminating idle time

Cloud Economics: Keeping Servers Busy

- Average CPU utilization of more than 5,000 Google servers during a six-month period

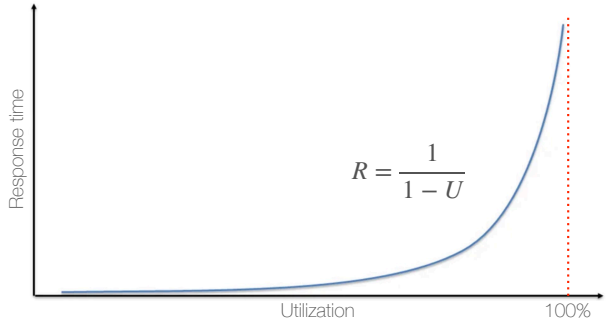


Cloud Economics: Keeping Servers Busy

- The cloud provider and cloud client have ***conflicting*** goals regarding server utilization
- While the provider wants to keep ***utilization high*** to maximize return on investment, the client wants to keep ***utilization low*** to minimize response times

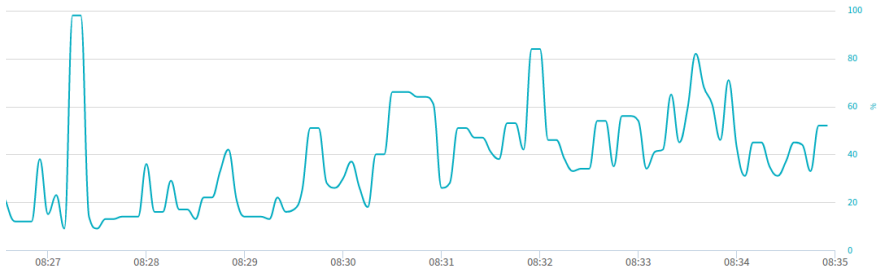
Cloud Economics: Keeping Servers Busy

- As the server utilization approaches 100%, the ***response time*** (delay) of the server grows without bound



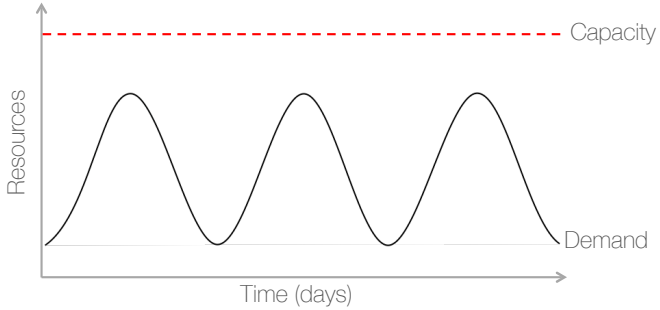
Cloud Economics: Keeping Servers Busy

- Server utilization is not only highly ***skewed***, it is also very ***dynamic***

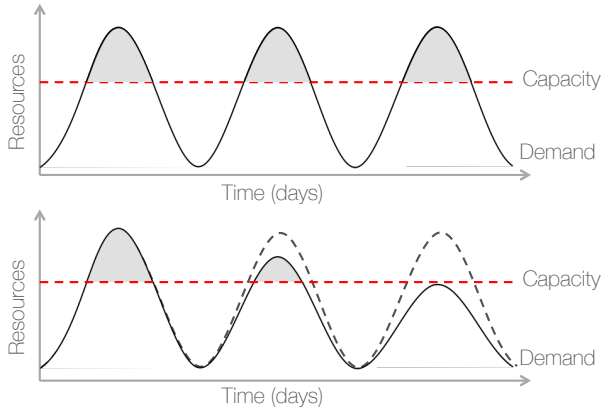


Cloud Economics: Resource Provisioning (Capacity Planning)

- Excessively high delays may result in ***Service Level Agreements*** being violated
- In order to maintain user satisfaction levels high (and respect ***Service Level Agreements***), resources (servers) have to be ***provisioned*** appropriately



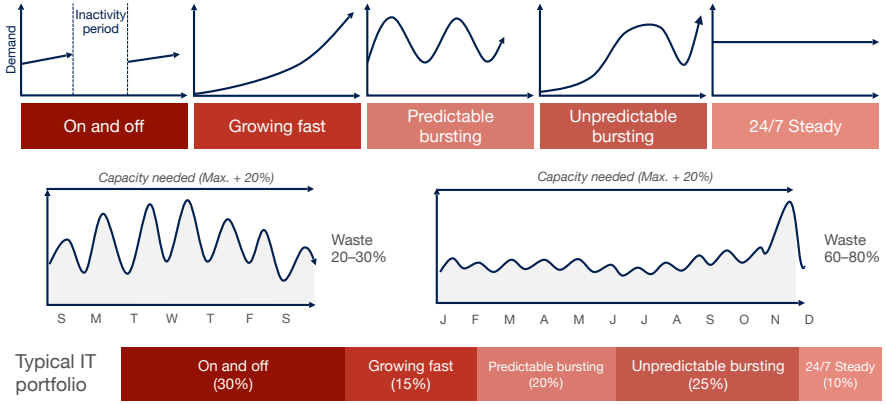
Cloud Economics: Risks of Under Provisioning



Lost revenue

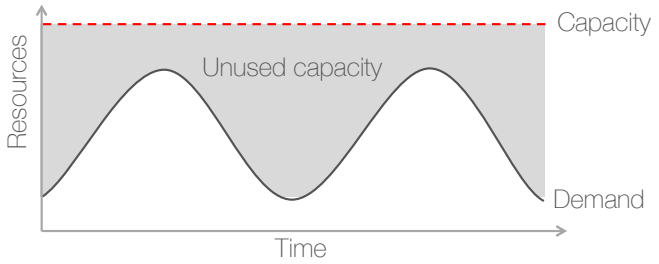
Lost customers

Cloud Economics: Demand Patterns



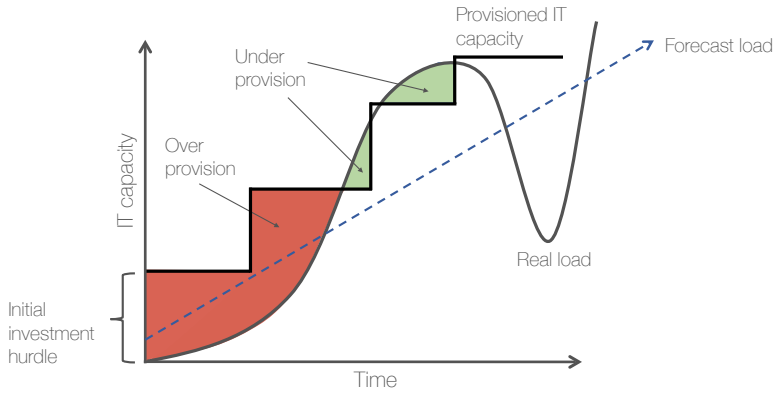
Cloud Economics: Static Over Provisioning

- In order to not lose revenue or customers, we need to over provision



- Static provisioning can be wasteful but may be necessary to meet *Service Level Agreements* with respect to response times

Cloud Economics: Pre-cloud Provisioning



Cloud Economics: Example

- Assume your service has a predictable demand where the peak requires ***500 servers at noon*** but only ***100 servers at midnight***
- Since the *average* utilization over a whole day is 300 servers, the actual cost per day is 300 × 24 = 7,200 server-hours
- But you must provision to the ***peak of 500 servers***, and you pay for 500 × 24 = 12,000 server-hours, a ***factor of 1.7 more***
- Therefore, as long as the ***pay-as-you-go cost per server-hour*** over 3 years (typical amortization time) ***is less than 1.7 times*** the cost of buying the server, ***cloud computing is cheaper***

Back to Amazon

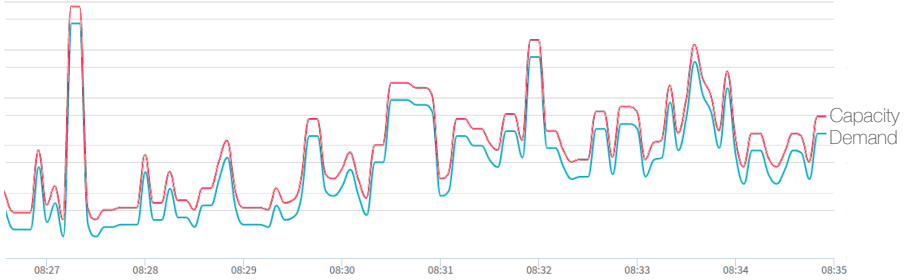
- For Amazon, static over provisioning was not only a solution for ***minimizing lost revenue*** and ***lost customers***, it also represented a new ***business opportunity***
- As its retail business grew, Amazon built bigger and bigger data centers to handle the increasing demand which resulted in huge amounts of ***aggregate unused capacity***
- Modern cloud computing was born when the unused capacity in large data centers was seen as a ***business opportunity*** that could be monetized and sold to customers as ***virtual servers***

Back to Amazon

- 2002 — Amazon launches ***Amazon web Services*** (AWS) as its retail computing infrastructure
- 2004 — ***Simple Queueing Service*** (SQS) launched as first public AWS service
- 2006 — AWS officially launched for public usage with SQS, ***Elastic Compute Cloud*** (EC2) and ***Simple Storage Service*** (S3)

Elastic Cloud Provisioning

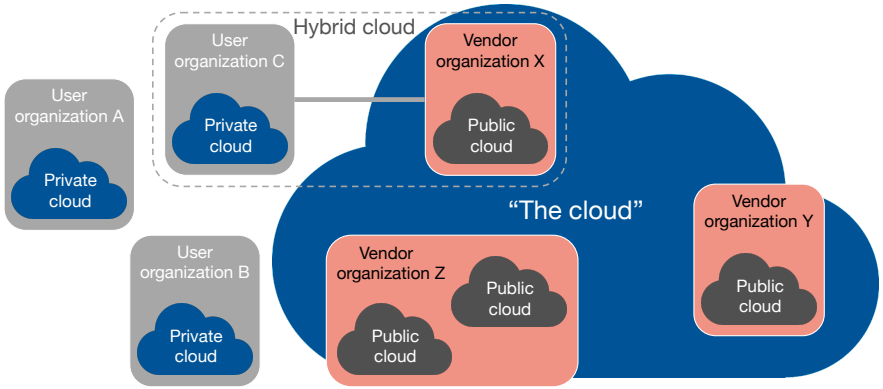
- The defining idea of cloud computing is ***Elastic Provisioning*** where the computing ***capacity*** closely matches the actual ***demand*** so as to minimize both ***wasted capacity*** and ***lost revenue***



Cloud Deployment Models

- Based on the identities of the ***provider*** and ***user*** roles
  - Private cloud
  - Public cloud
  - Hybrid cloud

Cloud Deployment Models



Private Cloud

- Provider and users belong to the ***same organization***
- Services built to be ***compatible*** with generic cloud interfaces (private or public)
- No risk of ***vendor lock-in***
- ***Security*** and ***data protection*** in the hands of the user
- Costs for hardware, space and administration similar to ***on-premise non-cloud*** architectures
- Private cloud may be suitable for those applications that have strict ***security*** or ***regulatory compliances*** for data and computations needs or where the ***migration costs*** are excessive

Public Cloud

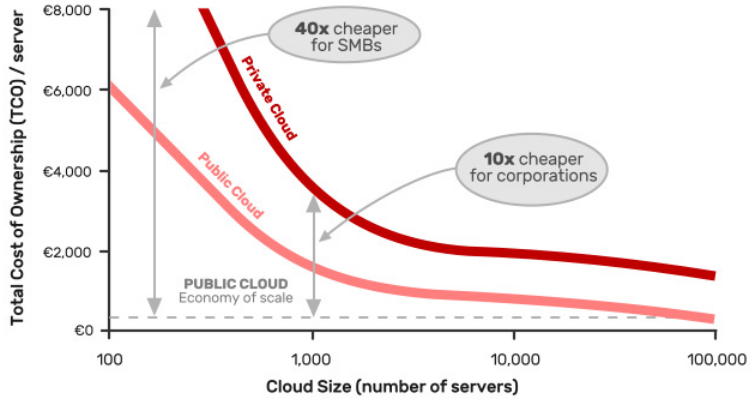
- Provider and users belong to ***different organizations***
- Providers pursue ***commercial interests***
- Users ***do not invest*** in procurement, operation and maintenance of hardware
- Possible risk of ***vendor lock-in***
- ***Security*** and ***data protection*** guarantees largely dependent on the cloud provider and may be determining factors

Hybrid Cloud

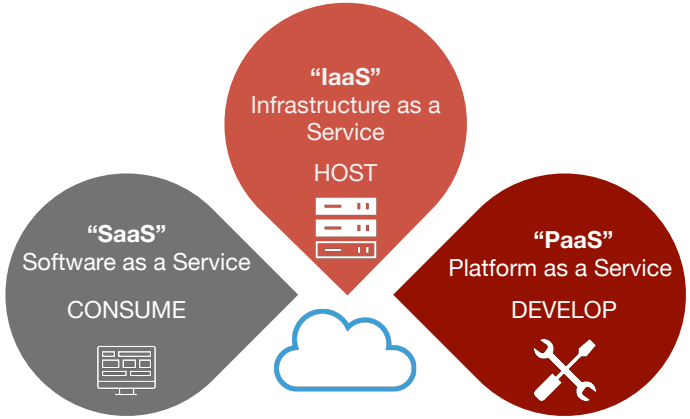
- Services from ***public and private clouds*** are used together within the ***same organization***
- Usage examples
  - Augment private cloud capacity with public clouds at times of ***peak demand***
  - Delegate certain functions such as ***data backup*** to public clouds
  - Leverage the ***security of private clouds*** together with the ***scale of public clouds***

Public vs. Private Clouds

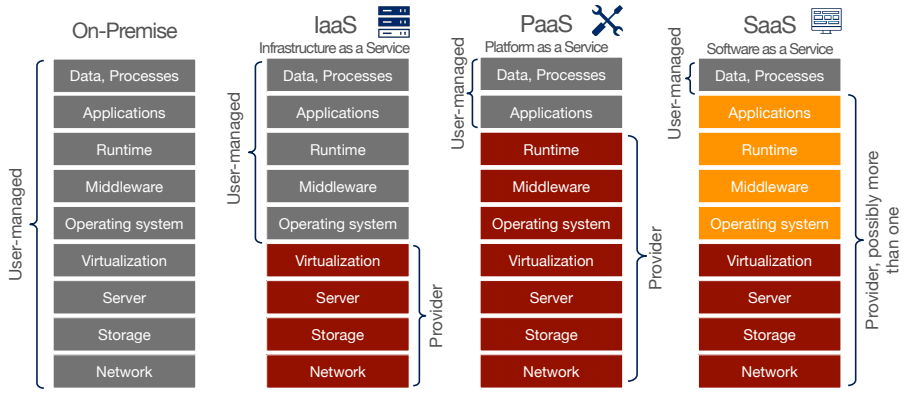
Public vs. Private Clouds



Cloud Computing Models



IaaS vs. PaaS vs. SaaS



Infrastructure as a Service

- *IaaS* offers the capability to host and provision ***computing***, ***storage***, ***networking*** and other fundamental computing resources
- Users are able to deploy and run arbitrary software, which can include ***operating systems*** and ***applications***
- The ***underlying cloud infrastructure*** is ***managed by the provider*** and the ***user has control*** over ***operating systems***, ***storage***, ***deployed applications***, and possibly ***limited control*** of some ***networking components*** (e.g., host firewalls)

Infrastructure as a Service

- Computing resources in the form of ***Virtual Machines*** with choices among many flavors of Windows server, Linux operating systems
- Functions include ***setup***, ***configuration***, ***load balancer***, ***backup***, ***site recovery***
- ***Advantages***: Highly flexible, highly scalable, highly available, cost effective, great way to future-proof your business
- ***Examples***: AWS Elastic Compute Cloud (EC2), Azure Virtual Machine, Google Compute Engine, Rackspace

Platform as a Service

- *PaaS* offers the capability to ***develop*** and ***deploy*** consumer-created or acquired applications
- Includes ***runtimes*** (java), ***databases*** (MySQL, Oracle), ***Web Servers*** (Apache), ***programming languages*** with ***libraries*** (PHP, Perl) supported by the provider
- The ***user does not manage or control*** the underlying cloud infrastructure including network, servers, operating systems, or storage
- The ***user has control*** over the deployed applications and, possibly, application hosting environment configurations

Platform as a Service

- *PaaS* allows the developer to focus on the creative side of ***application development***, as opposed to menial tasks such as installing development software, managing updates or security patches
- All brainpower and effort can be dedicated to ***creating***, ***testing***, and ***deploying***
- Application areas include ***multimedia***, ***big data analytics*** and ***software development*** when multiple developers collaborate
- *Examples*: AWS Elastic Beanstalk, Google App Engine, Apache Stratos, Force.com, IBM SmartCloud, Cloud Foundry

Software as a Service

- *SaaS* utilizes the Internet to ***deliver applications*** managed by the provider, usually on a subscription basis, to a large number of users
- Majority of SaaS applications run directly ***through a web browser***, which means they do not require any ***downloads*** or ***installations*** on the client side
- Users do not manage or control the underlying cloud infrastructure or the individual application capabilities, with the possible exception of user-specific ***application configuration settings***
- *SaaS* provides numerous advantages by greatly reducing the time and money spent on tedious tasks such as installing, managing, and upgrading software

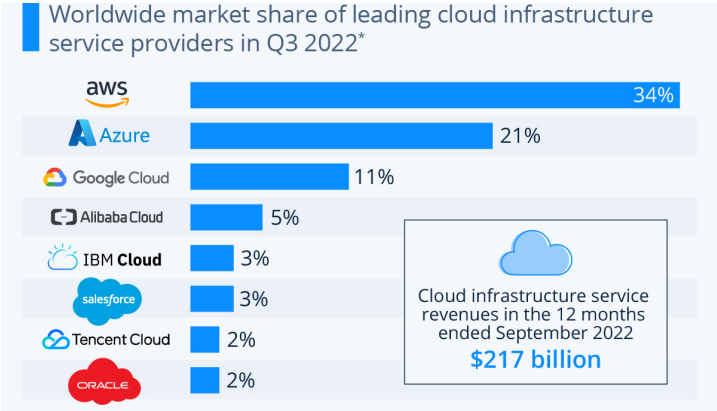
Software as a Service

- *SaaS* appropriate for
  - ***Startups or small companies*** that need to ***launch quickly*** without installing servers or software
  - ***Short-term projects*** that require ***quick, easy, and affordable collaboration***
  - Applications that ***aren't needed too often***, such as tax software
  - Applications that need ***both web and mobile access***
- *Examples*: Google G Suite, Dropbox, Office 365, Adobe Creative Cloud, TurboTax, Salesforce.com, SAP Concur

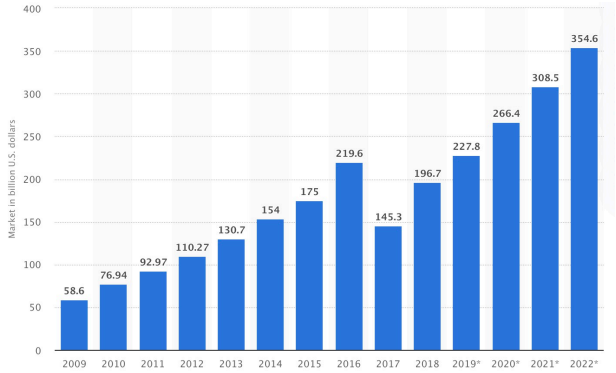
Public Cloud Computing Landscape 2023

- *IaaS* market dominated by the "big three" — ***AWS***, ***Microsoft Azure***, ***GCP***
- Together, they account for more than ***66% of the global cloud market*** with AWS taking the lion's share
- AWS revenue constituted ***14% of Amazon's total revenue*** of $110.8 billion in Q4 of 2021

Public Cloud Global Market Share

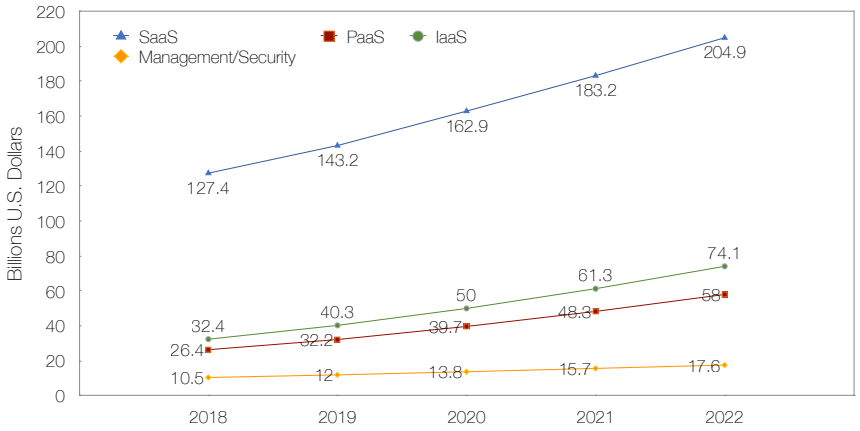


Global Public Cloud Market Revenues



- The cloud advertising segment was removed from public cloud services forecast segments starting from 2017

Public Cloud Revenue Trend



The "Big Three" Cloud Providers

	Amazon (AWS)			Microsoft (Azure)			Google (GCP)		
	2018	2019	2020e	2018	2019	2020e	2018	2019	2020e
Cloud Rev ($B)	$25.7	$34.9	$46.1	$10.0	$16.3	$23.6	$2.5	$4.3	$6.7
Cloud Rev Growth	47%	36%	32%	82%	62%	45%	135%	70%	55%
*Market Share	67%	63%	60%	26%	29%	31%	7%	8%	9%
CAPEX ($B)	$21.9	$26.5	$30.5	$11.6	$13.9	$14.9	$25.1	$26.8	$32.6
CAPEX growth	11%	21%	15%	43%	20%	7%	91%	7%	22%
Customers	Netflix, GE, Salesforce, Expedia, Adobe, Intuit, Kellogg's, Philips, BP			Walmart, Ford, NBC, Geico, T-Mobile, Daimler			Snap, Home Depot, Colgate, Disney, eBay, Spotify		
Other Key Metrics	76 Availability Zones within 22 geographic regions			Available in 140 countries and 54 geographic regions, with plans for 4 more			61 Availability Zones within 20 regions. Available in 200+ countries & territories.		

Cloud Pricing
Free Tiers

- *AWS*: 750 hours of Linux or Windows micro instances with 1GB of memory, 15GB of bandwidth, a load balancer, and access to a database, caching, and other tools for 12 months, as long as you don't exceed the limits
- *Microsoft Azure*: 750 hours of Linux or Windows machines with ample storage, SQL database, 15GB of bandwidth. Several other popular services are free for at least 12 months, and new customers also receive a $200 credit to try any other service for 30 days
- *Google Cloud Platform*: One month of a micro instance with 30GB of storage, plus a 12-month free trial with $300 credit to try any service. Limited access to many common tools is provided for free, always

Quick Pricing Comparison

	aws	Microsoft Azure	Google Compute Engine
CPUs	8	8	8
RAM	16GB	16GB	30GB
Storage	30GB free	128GB	$.02/GB per month
Bandwidth	10GB free	5GB free	$.12/GB per month
Price	$56.00/month	$150.45/month	$124.84/month

AWS EC2 Pricing Models

- *On-demand*: pay as you go without commitment. For unpredictable bursting demands
- *Reserved*: rent instances with one-time up-front payment for one or three years while receiving 35-75% discounts on the hourly charge. For steady demands
- *Spot*: instances offered with no SLA and risk interruption with two minutes notification if Amazon needs the capacity back. Spot instances are available at up to 90% discount compared to on-demand prices. For time-insensitive demands

AWS EC2 Pricing Models

- *Dedicated*: rent a physical EC2 server dedicated for your use. Dedicated Hosts can help reduce costs by allowing use of existing server-bound software licenses as well as address ***compliance and regulatory requirements*** for organizations that need to run their instances on dedicated servers instead of multi-tenant servers
- On-Demand, Reserved and Spot forms are also available with ***per-second billing*** (60 seconds minimum)

Cloud Pricing Calculators

- https://calculator.aws/#/
- https://cloud.google.com/products/calculator
- https://azure.microsoft.com/en-in/pricing/calculator/