Universidade de Aveiro
2016
Departamento de Eletrónica, Telecomunicações e Informática

Manuel António de Matos Pereira

Nas nuvens ou fora delas, eis a questão
Clouding or not clouding, that is the question

Universidade de Aveiro
2016
Departamento de Eletrónica, Telecomunicações e Informática

Manuel António de Matos Pereira

Nas nuvens ou fora delas, eis a questão
Clouding or not clouding, that is the question

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Sistemas de Informação, realizada sob a orientação científica do Professor Doutor Aníbal Manuel de Oliveira Duarte, Professor Catedrático do Departamento de Eletrónica Telecomunicações e Informática da Universidade de Aveiro

Ao meu filho Santiago...

O Júri

Presidente	Professor Doutor Joaquim Arnaldo Carvalho Martins Professor Catedrático da Universidade de Aveiro
Vogal – Arguente Principal	Professor Doutor Henrique Manuel Dinis dos Santos Professor Associado C/ Agregação da Universidade do Minho – Escola de Engenharia
Vogal - Orientador	Professor Doutor Aníbal Manuel de Oliveira Duarte Professor Catedrático da Universidade de Aveiro

Agradecimentos

Em primeiro lugar agradeço ao Professor Doutor Aníbal Manuel de Oliveira Duarte por ter aceitado este desafio, pela sua orientação, inesgotável disponibilidade, partilha do seu vasto conhecimento académico e da sua enorme experiência. Pelas suas opiniões e críticas, colaboração no solucionar de dúvidas e problemas que foram surgindo ao longo da realização deste trabalho que só foi possível com a sua preciosa ajuda.

Um agradecimento especial aos meus pais, Manuel Domingues Pereira e Laura Neves de Matos que sempre me apoiaram incondicionalmente.

Por fim, agradeço a todos os meus familiares, amigos, colegas de estudo e de trabalho que ao longo dos anos me moldaram o modo de ser, pensar e atuar.

Palavras-chave **computação cloud; planeamento de capacidade; desempenho; suporte a decições; Interoperabilidade; modelação de infraestruturas;**

Resumo

O proposito desta dissertação é contribuir no sentido de uma melhor compreensão sobre a decisão de ir ou não ir para uma solução na cloud quando uma organização é confrontada com a necessidade de criar ou expandir um sistema de informação.
Isto é feito recorrendo à identificação de factores técnicos e económicos que devem ser tomados em conta quando planeamos uma nova solução e desenvolver um framework para ajudar os decisores.

Os seguintes aspetos são considerados:

- Definição de um modelo de referência genérico para funcionalidades de um Sistemas de Informação.
- Identificação de algumas métricas básicas para caracterizar performance e custos de Sistemas de Informação.
- Analise e caracterização de Sistemas de Informação on-premises:
  - Arquiteturas
  - Elementos de custo
  - Questões de Performance
- Analise e caracterização de Sistemas de Informação Cloud:
  - Topologias
  - Estruturas de custo
  - Questões de Performance
- Estabelecimento de framework de comparação para a cloud versus on-premises
- Casos de uso comparando soluções na cloud e on-premises;
- Produção de guidelines (focadas no caso das clouds publicas)

Para ilustrar o procedimento, são usados dois business cases, ambos com duas abordagens: uma dedicada aos Profissionais de IT (abordagem técnica), outra aos Gestores/Decisores (abordagem económica).

Keywords

cloud computing; capacity planning; performance; decision support; Interoperability; infrastructure modeling;

Abstract

The purpose of this dissertation is to contribute towards a better understanding about the decision to go or not to go for cloud solutions when an organization is confronted with the need to create or enlarge an information system.
This is done resorting to the identification of technical and economic factors that must be taken into account when planning a new solution and developing a framework to help decision makers.

The following aspects are considered:

- Definition of a generic reference model for Information systems functionalities.
- Identification of some basic metrics characterizing information systems performance.
- Analysis and characterization of on-premisis information systems:
  - Architectures
  - Cost elements
  - Performance issues
- Analysis and characetrization of cloud information systems.
  - Typology
  - Cost structures
  - Performance issues
- Establishment of a comparison framework for cloud versus on-premises solutions as possible instances of information systems.
- Use cases comparing cloud and on-premises solutions.
- Production of guidelines (focus on public cloud case)

To illustrate the procedure, two business cases are used, both with two approaches: one dedicated to IT Professionals (Technical approach), other to Managers/Decision Makers (Economic approach).

Index

Figure Index

Table Index

Acronyms

AD	Active Directory
ARPANET	Advanced Research Projects Agency Network
ARM	originally Acorn RISC Machine, later Advanced RISC Machine
ARR	Application Request Routing
BI	Business Intelligence
BPAAS	Business Process As A Service
CAPEX	Capital Expenditure
CDN	Content Delivery Network
CPU	Central Processing Unit
DC	Domain Controller (AD)
DMZ	Demilitarized Zone
DNS	Domain Name System
DoS	Denial of Service
DRM	Digital Right Management
DSL	Digital Subscriber Line
DVR	Digital Video Recorder
EIL	Enterprise Integration Layer
EPG	Electronic Program Guide
HD	High Definition
HFC	Hybrid Fiber Coax
HTTP	Hyper Text Transfer Protocol
HTTPS	HHTP Secure
IIS	Internet Information Server
IAAS	Infrastructure As A Service
IMAP	Internet Message Access Protocol
IMMS	Internet Media Management System
IRR	Internal Rate of Return
KPI	Key Performance Indicator
LAN	Local Area Network
LDA	Local Delivery Agent

MDA	Mail Delivery Agent
MSFC	Microsoft Failover Cluster
MTA	Mail Transfer Agent
MUA	Mail User Agent
NAS	Network Attached Storage
NFR	Non-Functional Requirements
NIC	Network Interface Card
NLB	Network Load Balancing
NPV	Net Present Value
OLA	Operation Level Agreement
OLTP	On-Line Transaction Processing
OPEX	Operational Expenditure
OS	Operating System
OU	Organizational Unit
ROI	Return On Investment
PAAS	Platform As A Service
POP	Post Office Protocol
PoP	Point of Presence
SAAS	Software As A Service
SAN	Storage Area Network
SAS	Serial Attached SCSI
SD	Standard Definition
SDN	Software-defined networking
SLA	Service Level Agreement
STB	Set Top Box
SSL	Secure Sockets Layer
SVOD	Subscription Video On Demand
TCO	Total Cost of Ownership
TVOD	Transactional Video On Demand
VCR	Video Tape Recorder
VLAN	Virtual LAN
VOD	Video On Demand
WCF	Windows Communication Framework
WSE	Web Services Extensions

1. Introduction

Nowadays, information systems are present in almost all activity sectors, including health, education, entertainment, commerce, manufacturing, banking, insurance, tourism, transportation, and many others. Their impact is twofold:

- They are used as a mean of reinventing and making more effective the conventional activities and procedures of many sectors;
- They make possible to deal with new complexities, constant transformations, rapidly evolving business models and ultra large scale operations – which, themselves, are also, to a high extent, consequences of the increasing usage of information and communication technologies.

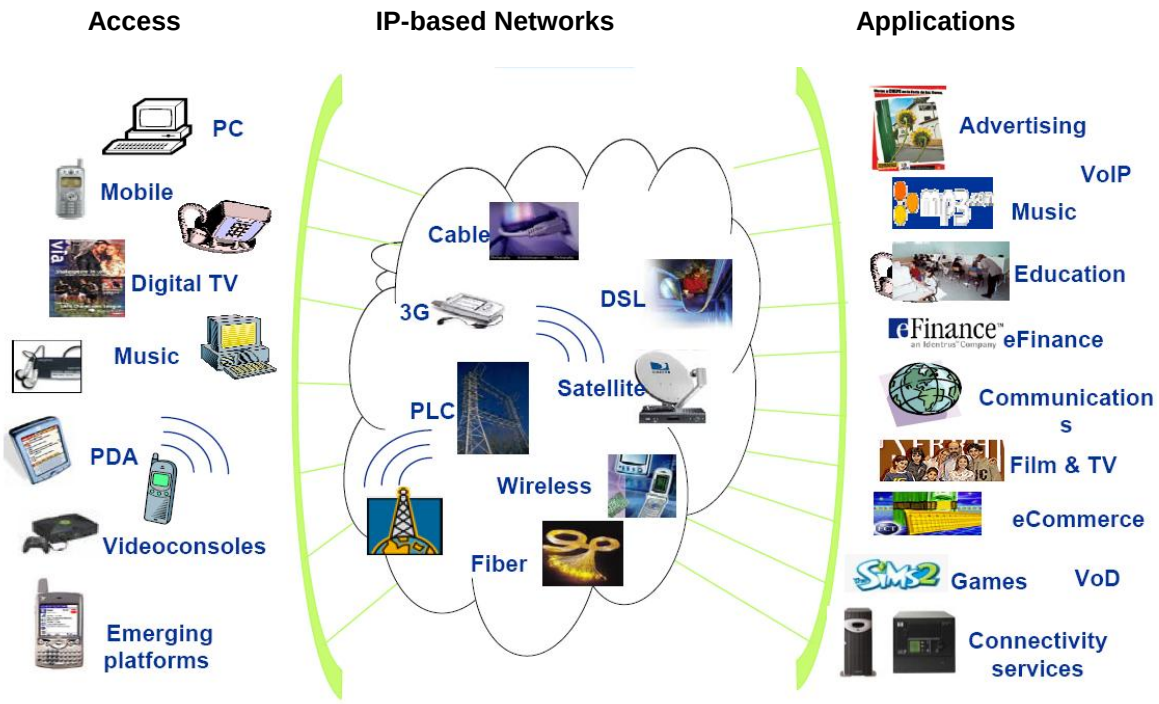


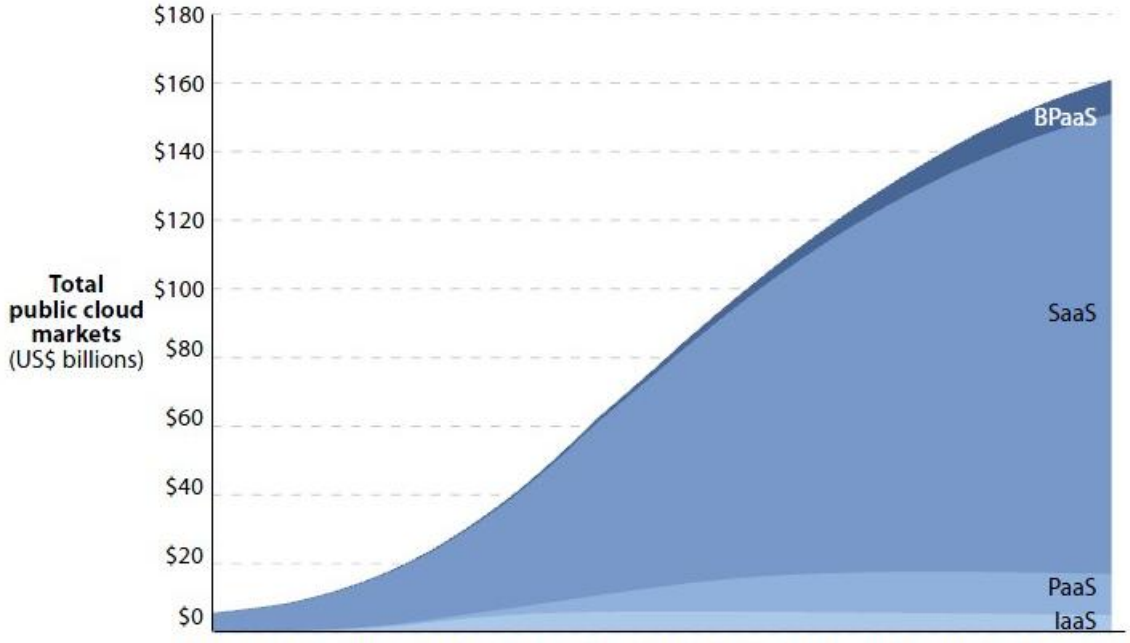
Figure 1 – New Environment [1]

Nowadays, one of the most visible forms of information systems is the Internet. Its evolution has made it available to almost everyone, anytime, anywhere and on any device. This new technological paradigm provides opportunities to expand and even create new business models supported by information and communication technologies. One of this new service models is based on cloud computing, that requires business and IT architects to understand the cloud computing paradigm, its benefits and limitations.

Motivation

Cloud Computing adoption continues accelerating in the organizations. According to an IDG, Enterprise Cloud Computing Study 2014 [2]:

1. 69% of enterprises have at least one application or a portion of their computing infrastructure in the cloud
2. Investment in cloud computing have increased 19% since 2012, with the average investment of large-scale enterprises (+1,000 employees) reaching $3.33M in 2014. Mid- and smaller scale enterprises with less than 1,000 employees spent $400K in 2015 on cloud solutions and technologies.
3. Speed of Deployment, Total Cost of Ownership (TCO) and replacing on-premises legacy technology are the most common reasons.
4. The top three application areas organizations are currently migrating or using are Email/Messaging, Collaboration/Conference Solutions and Customer Relationship Management (CRM) solutions.
5. The biggest challenges to implementing a Cloud Strategy are Security, Integration and Information Governance.
6. 24% of enterprise's IT budgets in 2016 are already allocated to cloud solutions, with the largest percentage allocated to SaaS-based applications (49%) followed by IaaS (28%), PaaS (18%) and other (5%).



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
BPaaS ($)	0.15	0.23	0.35	0.53	0.80	1.26	1.95	2.93	4.28	6.00	7.66	9.08	10.02
SaaS ($)	5.56	8.09	13.40	21.21	33.09	47.22	63.19	78.43	92.75	105.49	116.39	125.52	132.57
PaaS ($)	0.05	0.12	0.31	0.82	2.08	4.38	7.39	9.80	11.26	11.94	12.15	12.10	11.91
IaaS ($)	0.06	0.24	1.02	2.94	4.99	5.75	5.89	5.82	5.65	5.45	5.23	5.01	4.78

Figure 2 – Forecast: Global Public Cloud Market Size [3]

To finish depicting the above panorama it must be said that, currently, there is a generalized belief that cloud technologies can provide cost savings, rapid provision and scalability with minimum management effort to organizations[4] .

Objectives

The purpose of this dissertation is to illustrate possible evaluation models to support decision when IT professional and Managers / Decision Makers face the need of new Information System functionality.

Starting with information system characteristics, storage, computing and network state-of-the-art, to general and performance requirements. What to take in account when planning new or extend present Information Systems.

Passing by Cloud definition, essential characteristics, Service and Deployment Models from NIST, the most common accepted.

Characterize Workloads, and define a quantitative model identifying some technological and financial measures/metrics to use in comparison.

Supported on two business cases, apply the quantitative model and compare the information system solution on-premises with the several cloud providers, technological and financial comparisons and conclusions.

2. Information Systems Characterization and Requirements

An Information System exists to provide one or a set of functionalities to its users.

Regardless of these final functionalities that are specific to each application, two sets of requirements are transversal to all solutions: General (or non-functional) requirements and capacity / performance requirements.

In general, an information system performs one or a combination of the following operations:

1. Storage
2. Compute or
3. Network.

2.1 General Requirements

Availability. The degree of uptime for the solution, taking into account contention probabilities, which includes an indication of response time to problems and incidents, planning and maintenance schedules and impacts, and business continuity capability.

Performance. The extent to which the solution is assured to deliver a level of output.

Elasticity. The configurability and expandability of the solution, including the ability to adjust consumed service capacities up or down, and the scale or limitations of the capacity changes.

Manageability. The degree of automation and control available for managing the solution.

Recoverability. The solution's recovery point and recovery time objectives.

Interoperability. The degree to which services can interact with other services and infrastructure. Interoperability is described from two perspectives: (1) portability–the serial process of moving a system from one cloud environment to another, and (2) interconnectability–the parallel process in which two co-existing environments communicate and interact.

Security and privacy. Describes the attributes that indicate the effectiveness of the controls on access to services and data protection, and the physical facilities from which the services are provided. These attributes should provide an indication of physical protection, logical protection, controls and monitoring measures in place, compliance to country and corporate requirements, compliance with regulatory and statutory laws and obligations, and remediation processes.

Configurability. Describes the features and functions of the services, including the available basic services, the available standard options to add to the base services, the available customizable options and features to add to the base services, the ability to develop custom features, and the planned roadmap of functions and features for the service.

Long-distance migration. "Long distance" is defined as greater than 20 km of conductor between disparate data centers (or cloud provider sites). Inter-site latency is assumed to be at least 10 ms or worse. The characteristics of long-distance migration include cross-provider migration, cost-sensitive migration, open standards compliance, and live and at-rest migration parameters.

2.2 Data Centers Requirements

Before starting comparing on-premises with Public clouds, it's important to look at some of the investments needed to host properly on-premises. It's the only way to make a fair comparison, based on the ANSI TIA 942 2005 – Telecommunication Infrastructure Standard for Data Centers [5]:

1. Server Room
    a. Power
    b. Standby Power (UPS and/or Generator)
    c. HVAC (heating, ventilation, and air conditioning)
    d. Uninterruptible Power Supply
    e. Fire and Water infiltration Protection
    f. Access Control System
2. Networking
    a. Switches (Ethernet, Fiber Channel, Infiniband)
    b. Routers
    c. Hardware firewall
    d. Software firewall
    e. Network Intrusion Detection System (IDS) - optional
    f. Network Intrusion Protection System (IPS) - optional
3. Servers
    a. Server Rack
    b. Server Chassis/Enclosure
    c. Servers
4. Storage
    a. SAN
    b. NAS
    c. VTL/TL (Backup)
5. Licensing
    a. Hardware Management (iLO, iRMC, iDRAC, Storage services, etc.)
    b. Virtualization (VMware, Hyper-V, KVM, OVM (XEN), etc.)
    c. Operating System (Windows, Linux, etc.)
    d. DBMS (MSSQL, ORACLE, MONGODB, POSGRESSQL, etc.)
    e. Others
6. Staff specialize in the several infrastructure components:
    a. Datacenter infrastructure
    b. Server Systems
    c. Network Systems
    d. Storage Systems
    e. Operating Systems (also needed in Cloud model depending on the chosen service model)

2.3 Three-tier architecture

In software engineering, multitier architecture (often referred to as n-tier architecture) is a client–server architecture in which **presentation**, **application processing**, and **data management** functions are physically separated. The most widespread use of multitier architecture is the three-tier architecture[6].

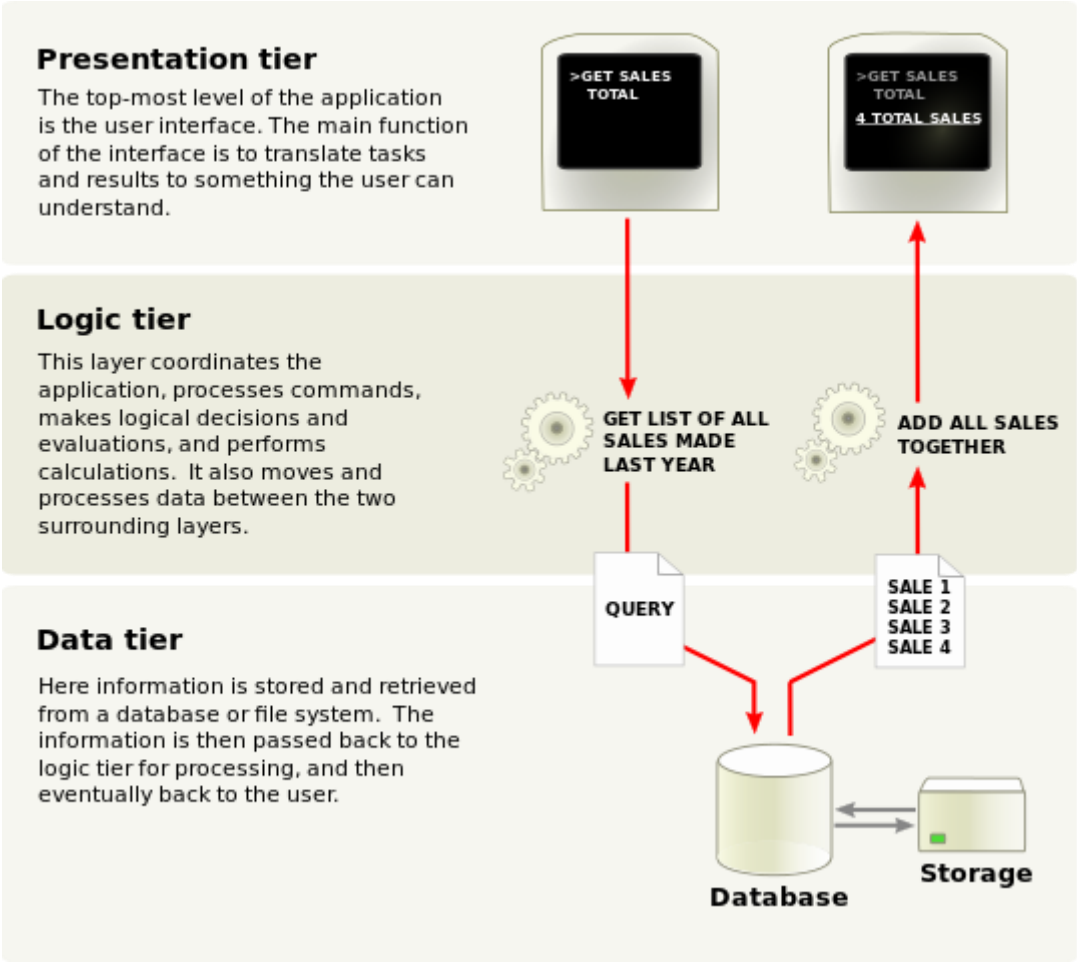


Figure 3 – Overview of a three-tier application [6]

2.4 Performance Requirements

The following set of quantitative performance metrics was selected since they cover the most important aspects of the dimensioning of an information system:

1. Storage (total users, capacity);
2. Compute (response time, performance);
3. Network (number of simultaneous users, bandwidth)

They will be the base of computing capacity for this work and as we can verify in more detail in the next sub-topics they are all converging.

2.5 On-premises Systems Characterization

2.5.1 Storage Area Networks (SAN) State-of-the-Art

Even though global file systems are enabling software defined scale-out NAS, and virtualization is enabling software defined block storage, most of today's storage is still located in either a traditional SAN or a NAS storage environment in the data center. Storage area networks (SAN) and network attached storage (NAS) have been around for decades now, and they still work very well, but they are now becoming converged together within new storage solutions.
A storage area network (SAN) is still the most efficient solution for high performance structured data such as databases, email, and other high transaction input/output (I/O) applications or high throughput block I/O[7].

2.5.1.1 Tiering

Storage tiering, and more specifically automated storage tiering, has become a baseline element. Even so, significant differentiation gives storage managers a delightful choice when it comes to evaluating competing solutions. This differentiation is particularly important to those organizations that seek best-of-breed products, where tiered data storage is a significant requirement.

"All tiering offerings have certain things in common. First, and at a minimum, the array hosts multiple physical media types, usually including solid-state drives, high-performance disk (either Fibre Channel or SAS) and high-capacity disk, with plenty of permutations of those basic components. Second, systems include software that embodies rules and methods for moving data from one physical tier or media type to another. Even though these features are common at a base functional level, there's enormous variation in the way they're implemented.

A key technical driver for tiering adoption has been solid-state storage or solid-state drives (SSDs). Early tiering efforts around Tier 1 (Fibre Channel), Tier 2 (SAS) and Tier 3 (SATA) failed because organizations couldn't accurately provision for hot versus cold data. Thus, many tiered arrays remained 80% Tier 1 to ensure adequate performance. The marginal cost savings of the remaining 20% didn't justify the added complexity and effort. SSD has been a game-changer in that it delivers huge IOPS performance gains in a very small footprint.
At this point, nearly all storage vendors agree that best-practice architectures include a small percentage of solid-state storage accompanied by high-capacity hard disk drives (HDDs), resulting in far fewer spindles.
For the purposes of this dissertation, we'll draw a distinction between SSD and flash cache, though the technology is essentially the same. SSD can be thought of a distinct Tier 0, available for application provisioning as with any other storage media. Flash cache is general purpose in nature, enhancing the entire array. Most vendors support both types, and the majority also supports a "hybrid pool" in which LUNs may consist of both SSD and various types of HDDs.

What is state-of-the-art storage tiering?

1. Tiered storage strategies must encompass at least three drive types, including solid-state storage.
2. Flash memory is an integral part of the offering.
3. Sophisticated algorithms identify "hot" data and move it automatically to the appropriate tier.
4. Storage arrays can simultaneously be optimized for cost and performance.
5. Optimization decisions are largely automated to minimize administrative intervention."[8]

In addition to auto-tiering the current storage systems have some other features, of which I consider worth mention:

1. Data Deduplication
2. Thin Provisioning and
3. End-user Services (e.g.: FTP, CIFS or NFS)

2.5.2 Computing Systems State-of-the-Art

Independently if we are choosing a Rackmount Server or a Blade System Enclosure system, the computing technology is basically the same (e.g.: CPU, RAM and HDD). Blade Systems provide better space and energy efficiency, but are only cost-effective for +6 servers.
Today's typical enterprise server offer is using Xeon CPUs up to 24 cores each and with speeds up to 3.7Ghz per core. Summed with up to 6TB DDR 4 (2400mhz) memory supported by 128GB modules.
All the same to refer a trend in the last years that's the ARM based servers (microservers), for some specific workloads (e.g.: web servers) working in parallel can have great performance and low power consumption.

2.5.2.1 Microservers

"For certain kinds of high-volume, low-compute-power workload — such as web page serving, search engine query execution, or parallelized data processing tasks — a new species of server, the microserver, may occupy less data center space and consume less power than the traditional Xeon- or Opteron-based enterprise server.

HP is among the leading vendors in the emerging microserver space with its Project Moonshot. The first Moonshot server, announced in November 2011, was an ARM-based SoC platform called Redstone, which HP installed as a demonstration system in its Discovery Lab. Based on an existing ProLiant SL6500 chassis with server trays that can hold a mixture of compute or storage cartridges, the Redstone Development Server Platform delivers four times the density of the traditional ProLiant server (accommodating up to 72 compute nodes per tray, or 288 per 4U chassis) while consuming a tenth of the power, according to HP."[9]

2.5.3 Network Systems State-of-the-Art

There are at least three different types of networks in a datacenter, Ethernet, Fibre Channel and Infiniband (IB). For different proposes and with different speeds: Ethernet (up to 100Gbits/s) for server-server and client-to-server communications; Fibre Channel (up to 32Gbits/s) and InfiniBand (up to 50Gbits/s) for server-storage connections.

2.5.3.1 The software-defined data center

“Server and storage virtualization are mature concepts, but for maximum data center efficiency and flexibility, the third major IT component — networking — arguably needs to undergo a similar process of decoupling the control layer from the physical hardware. Software-Defined Networking (SDN) is a new field, and the major networking players are still working out their responses to it — (EMC-owned) VMware's purchase of SDN specialist Nicira in July 2012 was a key move in this respect.

The standard-setting and promotion body for SDN is the Open Networking Foundation (ONF) and its major vehicle is the OpenFlow protocol. In April 2012, Google — a founding ONF member — disclosed details of a large-scale OpenFlow implementation on its datacenter backbone network, which carries the search giant's internal traffic (Google's other, internet-facing, backbone network carries user traffic).

Google's SDN experience included faster deployment time, the ability to emulate the backbone network in software for testing, easy upgradability, significant improvements in network utilization (close to 100 percent, compared to an industry average of 30-40%) and a generally high degree of stability. Echoing its strategy with servers, Google built its own network switches for this project — something that, if widely adopted, will make vendors of traditional data center networking equipment nervous.”[9]

2.6 Problem Statement

Nowadays, in face of the need for a new information system functionality, planners are faced with two fundamentally different types of choices:

a) Build it as a physical, on-premises solution?
    a. Owned premises?
    b. Collocation?
    c. …
b) Build it, as a virtual, in the cloud solution?
    a. Public cloud?
    b. Private cloud?
    c. …

The following diagram, illustrates the type of questions and possible answers associated with the problem:

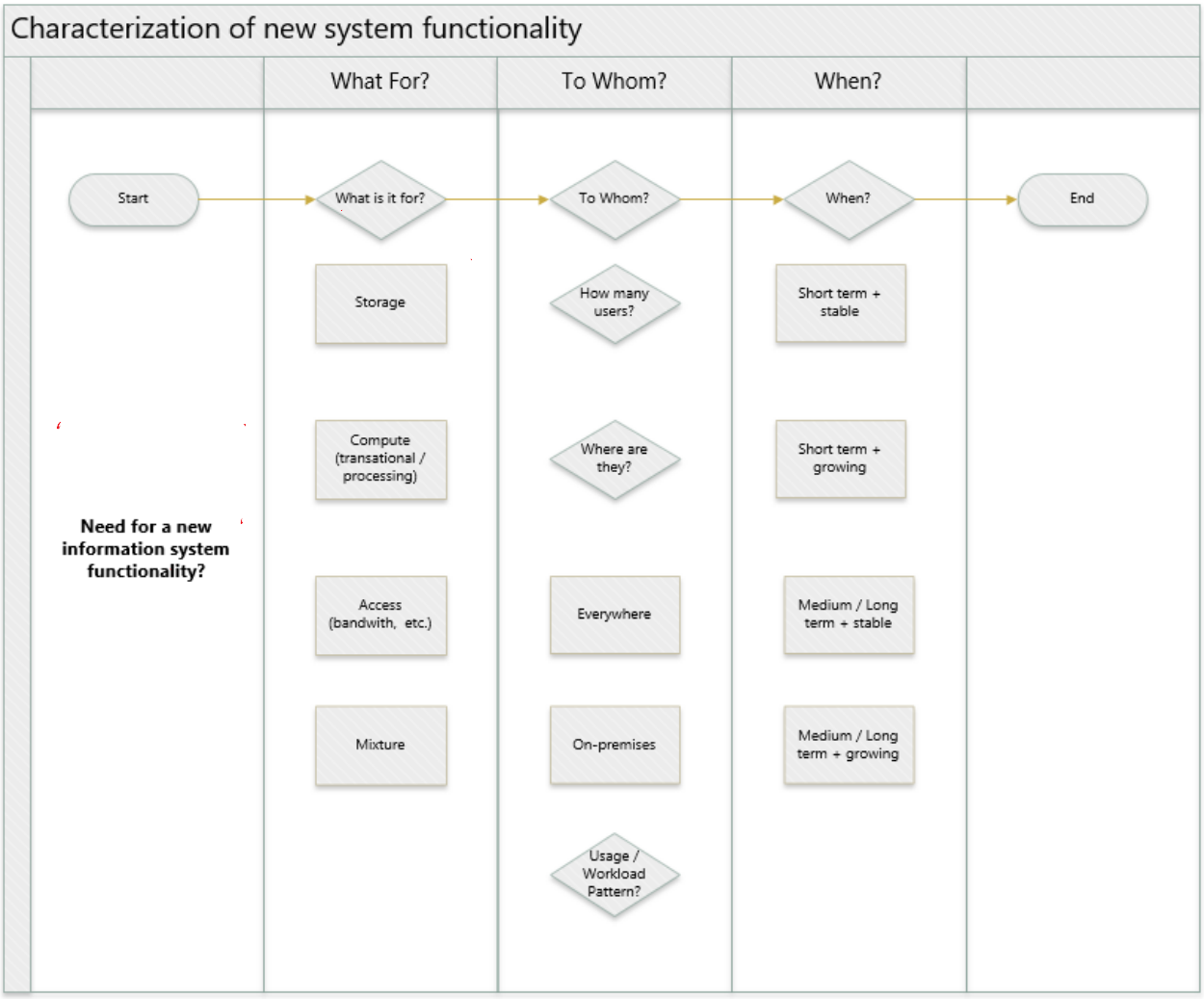


Figure 4 – Decision tree associated with the process of establishing a new information system functionality

Let's use some typical answers to the decision tree to create a template.

Template

1. **What is the major functionality**: Storage; Computing, Network or Mixture
2. **Number of Users**: up to (express in thousands or millions)
3. **User location**: on-premises; internet; Mixture
4. **Workload cargo**: low; medium; high.
5. **Workload Pattern**: stable; on and off; growing fast; unpredictable bursting; predictable bursting.
6. **When it's needed**: Short-term; Medium-term; Long-term.
7. **Duration (life cycle):** up to 1 year; 2 years; 3 years; 4 years; +4 years; unknown

2.7 Information Systems Planning

Planning of an information system involves the following fundamental steps [10]:

- Setting-up of the organizational context in which the information system is going to be used.
- Requirements specification
- Identification of possible engineering solution(s)
- Economic and financial assessment of possible solutions.
- Decision about the specific engineering solutions business plans to be implemented.

A brief outlook is presented about each of these steps.

- Setting-up of the organizational context in which the information system is going to be used.
  - *What is the business and general ecosystem of the organization?*
  - *Why is the information system needed and what will be its role.*
  - *How does it relate to other systems in the organization?*
- Requirements specification of the information system:
  - *What will it be? (General System Perspective)*
    Among possible answers are the following:
    - A stand-alone system to be used in a single location.
    - A system with distributed access to be shared from different locations
    - etc
  - *What will it do? (Functional and Operational Requirements)*
    The main aspects are the following:
    - What will the system do?
    - Which information must it handle?
    - Which information flows will take place?
    - Where will information originate and where will it be used, processed, stored, etc.
    - How should it perform? (Basic requirements about the data / information that will be processed in terms of volume, access bandwidth, latency, jitter tolerance, etc.)
  - *Who will be the users, what will they do with it and with which usage patterns? (User Characterization)*
    The mains aspects include the following:
    - Educational level.
    - Possible preferences.
    - Technical expertise.
  - *Constraints:*
    This might include such aspects as the following:
    - How much can be spent on it and when? (Financial constraints)
    - When is it to be operational? (Time constraints)
    - How reliable must it be? (Reliability constraints)
    - How safe/secure must it be? (Safety/security constraints)
    - Which information usage limitations have to be considered (Regulatory constraints)
    - Which environmental limitations have to be considered? (Ecological constraints)

- Identification of possible engineering solution(s):
  - *Which technical solutions could, possibly, satisfy the above requirements?* (Solution design)
  - *How could they be deployed and how could they evolve over time?* (Planning and roll-out scenarios)
- Identification of required resources:
  - *Human, material, financial resources.*
- Economic and financial assessment of possible alternative solutions.
  - *How much costs each one of the identified solutions?*
    - How much goes into CAPEX and into OPEX?
    - How are costs spread over time?
    - Economic and financial indicators:
      - TCO, NPV, IRR, ROI, etc.
- Decision
  - *Once made the economic and financial assessment, prepare decisions about which technical and planning solutions to take.*

In this dissertation the idea is to contribute with a general approach that might later be adapted to specific use cases. For this reason the above steps are simplified into the following subset:

- Definition of a set of general requirements
- Definition of a set of quantitative performance metrics
- TCO of identified solutions, CAPEX and OPEX.

2.8 Workloads

In association with the identification of performance metrics it is also necessary to define different workload patterns, some examples: stable or growing fast.

Furthermore important is the characterization of the different types of systems: static and transactional.
Static typical example: Company Web Site (mainly to present information, web presence)
Transactional typical example: e-commerce site (I/O operations)

2.8.1 Workload Patterns

Workload Patterns or sometimes called Usage Pattern are essential to better understand the application and its requirements.

Extremely uncommon are the Applications that have a Stable Workload.

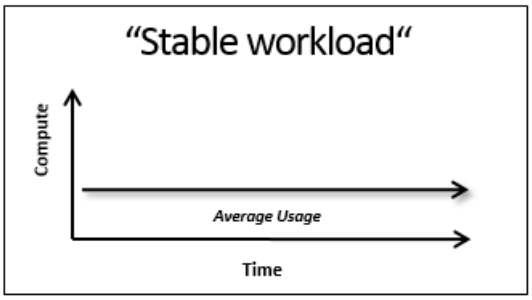


Figure 5 – Stable Workload Pattern [11]

Because of this, each of the other patterns showed in figure 8 utilizes key property of the cloud: the ability to scale up and down. Main difference with traditional IT here is in pay-per-use nature of the cloud: you don't have to pay for resources when you don't use them.

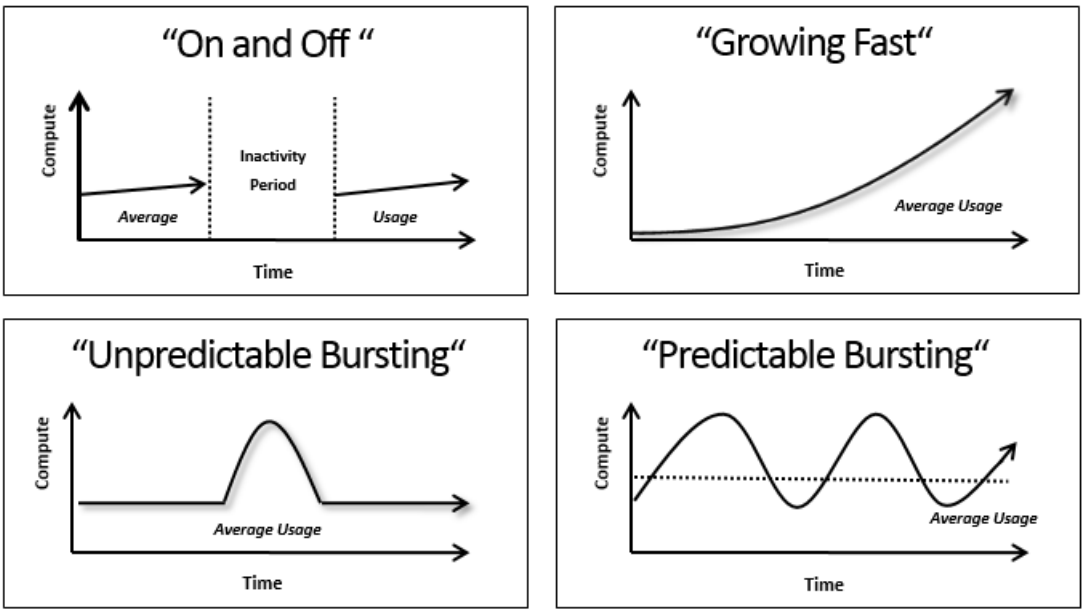


Figure 6 – Workload Patterns [11]

On and Off: Similar to Predictable Bursting, On and Off can have seasonal or time-bounded workloads where there are all or almost nothing processing requirements. Important enterprise workloads including those that are run monthly, quarterly and annually exhibit this type of behavior. (e.g.: Salary processing, etc.)

Unpredictable Bursting: Something happens that triggers heavy usage requirements so normally you would have had to scale design considerations to try and 'predict' what this requirement 'could be'. This may be one of the worst cases for which to plan: consider an emergency response system for which normal operation is fairly well understood, but then a hurricane hits and just when you most need the system, it gets overwhelmed.

Predictable Bursting: Think ticket system for a Cinema on weekend Nights or for UEFA Champions League Final. Most of the year or days of the week, demand is much less. Even though the additional load is expected, it is expensive to maintain this capacity because it is under-utilized when demand is lower.

Growing Fast: This is interesting in the case of smaller startup companies or groups in larger companies and/or can be associated with new development. How to plan, both during development and operation. Elasticity can be a huge opportunity/savings area.

2.9 Definition of the Quantitative Model

Based on "Basic Queuing Theory and Operational Analysis" topic from "Web Performance Measurement & Capacity Planning: Briefing Paper"[12] refers that Queuing network models (QN models) have been used successfully to analyze the performance of multi-device computer systems.

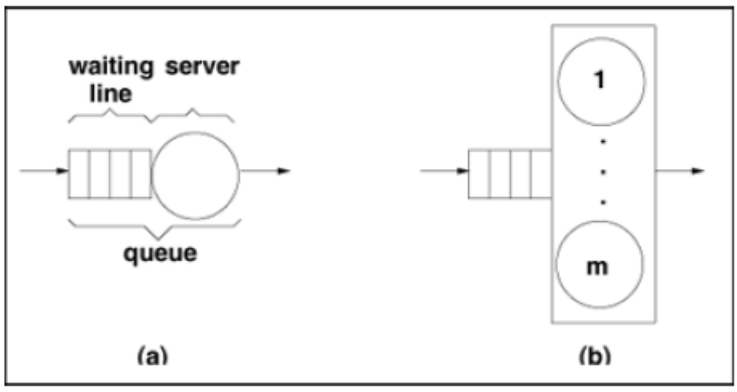


Figure 7 – (a) Single queue with one resource server (b) Single queue with m resource servers [13]

In information system, **queuing system** consists of one or more resources that provide some desired service to arriving customers. The system consists of the resources with a line/queue in which these customers must wait to be served. The queue must be able to hold zero or more customers. Note that oftentimes the term **queue** is used to refer to both the servers and the line together. After all, it doesn't make sense to wait in a queue for nothing, correct?
Such a system consists of the following dynamics, in order:

1. A customer arrives at the queuing system.
2. The customer waits in the queue for service by a server.
3. The customer receives service from a server.
4. The customer departs (ideally happily) from the queuing system.

Bringing this discussion back to IT systems, a customer is a packet, message, or frame, the queue is some buffer or block of memory in which the packet can be placed while it waits for processing, a server is the link, switch, or other resource on which the packet is waiting, and a departure is a packet that is transmitted by the server.

A queuing system can often be depicted by the following diagram:

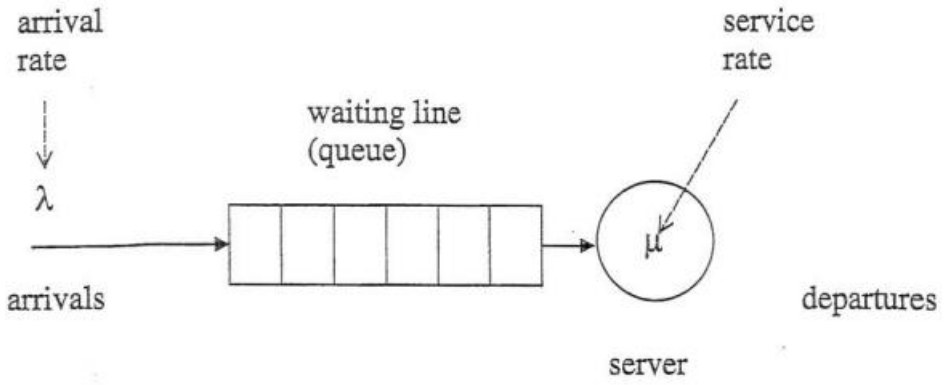


Figure 8 – Queuing system diagram [13]

2.9.1 Basic Performance

This chapter focuses on the quantitative aspects of QN models and introduces the input parameters and performance metrics that can be used, also presents the approach known as Operational Analysis [13] used to establish relationships among quantities based on measured or known data about computer systems.

Motivating problem: Suppose that during an observation period of 1 minute, a single resource (e.g., the CPU) is observed to be busy for 40 seconds. A total of 4000 transactions are observed to arrive to the system. The total number of observed completions is 4000 transactions (i.e., as many as arrivals occurred in the observation period).
What is the performance of the system (e.g., the mean service time per transaction, the utilization of the resource, the system throughput)?

Before solving this problem, some operational analysis notation is required for the measure data. Below is a partial list of such measured quantities:

- $\boldsymbol{T}$: length of time in the observation period
- $\boldsymbol{K}$: number of resources in the system
- $\boldsymbol{B_i}$: total busy time of resource i in the observation period T
- $\boldsymbol{A_i}$: total number of service requests (i.e., arrivals) to resource i in the observation period T
- $\boldsymbol{A_0}$: total number of request submitted to the system in the observation period T
- $\boldsymbol{C_i}$: total number of service completions from resource i in the observation period T
- $\boldsymbol{C_0}$: total number of requests completed by the system in the observation period T

Now, from these known measurable quantities, called operational variables, a set of quantities can be obtained. A partial list includes the following:

- $\boldsymbol{S_i}$: mean service time per completion at resource i; $S_i = \frac{B_i}{C_i}$
- $\boldsymbol{U_i}$: utilization of resource i; $U_i = \frac{B_i}{T}$
- $\boldsymbol{X_i}$: throughput (i.e., completions per unit time) of resource i; $X_i = \frac{C_i}{T}$
- λ_i : arrival rate (i.e., arrivals per unit time) at resource i; $\lambda i = \frac{A_i}{T}$
- $\boldsymbol{X_0}$: system throughput; $X_0 = \frac{C_0}{T}$
- $\boldsymbol{V_i}$: average number of visits (i.e., the visit count) per request to resource i; $V_i = \frac{C_i}{C_0}$

2.9.1.1 Example

Using the notation above, the motivation problem can be formally stated and solved in a straightforward manner using operational analysis. The measure quantities are:

T = 60 sec
K = 1 resource
B_i = 40 sec
$A_i = A_0$ = 4000 transactions
$C_i = C_0$ = 4000 transactions

Consequently, the derived quantities are:

$S_i = \frac{B_i}{C_i} = \frac{40}{4000} = \frac{1}{100}$ second per transaction

$$U_i = \frac{B_i}{T} = \frac{40}{60} = 66\%$$

$$\lambda_i = \frac{A_i}{T} = \frac{4000}{60} = 66{,}6\ tps$$

$$X_0 = \frac{C_0}{T} = \frac{4000}{60} = 66{,}6\ tps$$

2.9.2 Utilization Law

We have seen above, how the utilization of a resource is defined, dividing the numerator (Bi) and the denominator (T). If we divide both by the number of completions from resource I, Ci, during the observation interval:

$$U_i = \frac{B_i}{T} = \frac{\frac{B_i}{C_i}}{\frac{T}{C_i}}$$

The ratio Bi/Ci is simply the average time that the resource was busy for each completion from resource I, i.e., the average service time Si per visit to the resource. The ratio T/Ci is just the inverse of the resource throughput Xi. Thus, the relation known as the Utilization Law can be written as:

$$U_i = S_i \times X_i$$

If the number of completions from resource I during observation interval T is equal to the number of arrivals in that interval, i.e., if Ci = Ai, then Xi = λi the relationship given by the Utilization Law becomes:

$$U_i = S_i \times \lambda_i$$

2.9.2.1 Example

The bandwidth of a communication link is 48,000 bps and it is used to transmit 1500-byte packages that flow to the link at a rate of 3 packets/second. What is the utilization of the link?

Starting by identifying the operational variables provided or that can be obtained from the measured data:
The link is a resource (K=1)
The throughput of resource Xi is 3 packages/second

What is the average transmission time?

Each packet has 1,500 bytes x 8 bits = 12,000 bits/packet
It takes 12,000 bits / 56,000 bits/sec = 0,214 sec to transmit a packet over this link.
Therefore, Si = 0,214 sec/packet.

Using the Utilization Law, we compute the utilization of the link as:

$$S_i \times X_i = 0.214 \times 3 = 0{,}642 = 64.2\%$$

2.9.3 Service Demand Law

The Service Demand, denoted as Di, is defined as the total average of time spent by a typical request of a given type obtaining service from resource i. Throughout its existence, a request may visit several devices, possibly multiple times. However, for any given request, its service demand is the sum of all service times during all visits to a given resource. When considering various requests using the same resource, the service demand at the resource is computed as the average, for all requests, of the sum of the service times at that resource.
By definition, service demand does not include queuing time since it is the sum of service times. If different requests have very different service times, using a multiclass model is more appropriate. In this case, define Di,r, as the service demand of requests of class r at resource i.

To demonstrate the concept of service demand, consider that six transactions perform 3 I/Os on a disk. The service time in msec, for each I/O and transaction is given on table 2. The last line shows the sum of the service times over all I/Os for each transaction. The average of this sum is 36.8 msec. This is the service demand on this disk due to the workload generated by the six transactions.

	Transaction No.					
I/O No.	**1**	**2**	**3**	**4**	**5**	**6**
1	11	14	13	11	12	14
2	13	13	12	10	13	13
3	11	14	11	11	12	13
Sum	35	41	36	32	37	40

Table 1 – Service times in msec for six requests

By multiplying the utilization Ui of a resource by the measurement interval T one obtains the total time the resource was busy. If this time is divided by the total number of completed requests, C0, the average amount of time that the resource was busy serving each request is derived. This is precisely the service demand. So:

$$D_i = \frac{U_i \times T}{C_0} = \frac{U_i}{\frac{C_0}{T}} = \frac{U_i}{X_0}$$

This relationship is called the Service Demand Law, which can also be written as Di = Vi × Si, by definition of the service demand. The above equation indicates that service demand can be computed directly from the device utilization and system throughput.

2.9.3.1 Example

A web server is monitored for 5 minutes and it's CPU is observed to be busy 80% of the monitoring period. The web server log reveals that 36,000 requests are processed it hat interval. What is the CPU service demand of requests to the web server?

The observation period T is 300 (= 5 x 60) seconds.
The web server throughput, X0, is equal to the number of completed requests C0 divided by the observation interval; $X0 = \frac{36000}{300} = 120\ requests/sec$
The CPU utilization is Ucpu = 0.8. Thus, the service demand at the CPU is:

$$D_{cpu} = \frac{U_{cpu}}{X_0} = \frac{0.8}{120} = 0{,}066\ seconds/request$$

2.9.4 The Forced Flow Law

There is a simple way to relate the throughput of resource I, Xi, to the system throughput, X0. From the data from previous database server example, assume that every transaction that completes performs an average of two I/Os. What is the throughput of that disk in I/Os per second?

Since 3.8 transactions complete per second (i.e., the system throughput, X0) and each one performs two I/Os on average on disk 1, the throughput os disk 1 is 7.6 (= 2.0 x 3.8) I/Os per second. In other words, the throughput of a resource (Xi) is equal to the average number of visits (Vi) made by a request to that resource multiplied by the system throughput (X0). This relation is called the Forced Flow Law:

$$X_i = V_i \times X_0$$

The multiclass version of the Forced Flow Law is $Xi,r = Vi,r \times X0,r$

2.9.4.1 Example

What is the average number of I/Os on each disk in the previous database server example?
The value of Vi for each disk i, according to the Forced Flow Law, can be obtained as $\frac{Xi}{X0}$. The database server throughput is 3.8 tps and the throughput for each disk in I/Os per second is 32. Thus, $V_i = \frac{X_i}{X_0} = \frac{32}{3.8} = 8.4$ visits to disk per database transaction.

2.9.5 Little's Law

Let us consider the following figure, which describes a generic queuing system.

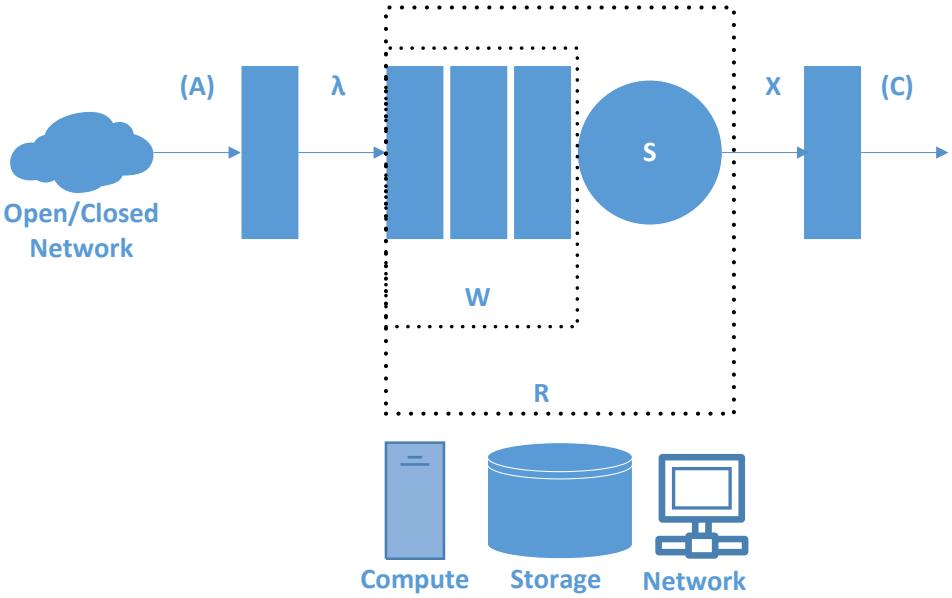


Figure 9 – What's in a queue? [14]

A - Arrival Count

λ - Arrival Rate (A/T)

W - Time spent in Queue

R - Residence Time (W+S)

S - Service Time

X - System Throughput (C/T)

C - Completed tasks count

T – Time Interval

Little's Law states the Following:

$$R = \lambda W$$

Where

R = Residence Time (average number of customers in the system, W+S)

λ = Arrival rate (average arrival rate)

W = Time spent in Queue (average time that a customer spends in the system)

The average waiting time and the average number of items waiting for a service in a system are important measurements for a manager. Little's Law relates these two metrics via the average rate of arrivals to the system.

2.9.5.1 Example

Assuming customers arrive at the rate of 20 per minute, λ=20, and they stay in the system an average of 25 seconds, W=25s, in the system. This means that the average number of costumers in the system at any time is 5, R=5.

$$R = 20 \times 0.25 = 5$$

2.10 Quantitative Model in Practice

Four basic formulas are proposed to calculate most of these measurements:

1. **Basic Performance**

Little's Law says that in a stable process there is only three characteristics that govern that process. If one characteristic changes, the others will change too. The three characteristics are:

Arrival Rate, **Time Spent in Queue** and **System Throughput**

We can summarize Little's Law by saying:

$\lambda = X\,W$ (Arrival Rate = System Throughput * Time Spent in Queue) Or

$W = \frac{\lambda}{X}$ (Time Spent in Queue = Arrival Rate / System Throughput)

Example:

E-commerce website as an average Arrival Rate of 17.5 clients, the throughput is 600 clients/minute or 10 clients/second.

Calculating Time Spent in Queue with Little's Law:

$W = \frac{R}{X} = \frac{17.5}{10} = 1.75$ Seconds (average time a client as to wait in the queue at the website)

2. **Utilization,** measures the fraction of time that the resources are busy

Utilization Law: $\mathrm{U_i} = \mathrm{X_i} \times \mathrm{S_i} = \lambda_{\mathrm{i}} \times \mathrm{S_i}$

Where X is the average throughput of a queue per unit of time and S is the average service time of a request

Example:

E-commerce website:

$U_i = 10 \times 1.75 = 17.5\,\%$

3. **Forced Flow,** measures average throughput of a resource and the percentage of utilization

- Forced Flow Law: $X_i = V_i \times X_0$

Example:

Transactions on a database server averaged 4.5 I/O operations, and if during a one hour period, 7,200 transactions were executed

$$X_{server} = \frac{7200}{3600} = 2\ tps\ (transactions\ per\ second)$$

4. **Service Demand**, sum of all service times for a request at the resource and is related to the system throughput and utilization

- Service Demand Law: $D = V_i \times S_i = \frac{X_i}{X_0} \times \frac{U_i}{X_i} = \frac{U_i}{X_0}$

Example:

UNIX system was monitored for 15 minutes and it was observed that the CPU was 80% busy during the monitoring period, and the number of HTTP requests counted in the log was 40,000

$U_{cpu} = 80\%$

$$X_{server} = \frac{40000}{15 \times 60} = 44.4\ requests\ per\ second$$

$$D_{cpu} = V_{cpu} \times S_{cpu} = \frac{U_{cpu}}{X_{server}} = \frac{0.80}{44.4} = 0.018\ seconds\ as\ the\ CPU\ demand\ for\ HTTP\ request$$

2.11 Metrics

Based on open data center alliance usage (ODCA): Standard Units of Measure For IaaS Rev 1.1[15], there are two base measures, quantitative and qualitative.

2.11.1 Quantitative Metrics

For IaaS, we begin with quantitative units for the three major components that the cloud provider needs to describe:

1. Storage
2. Compute (incorporating CPU and memory)
3. Network

For storage, measurement units must allow comparison of capacity, performance, and quality. Capacity can be measured in terabytes (TB). Performance can be provided in IOPS per TB.

For compute, there must be a consistent benchmark that is useful for comparison across a wide range of cloud subscriber needs. ODCA propose SPECvirt_sc2010 from www.SPEC.org. This benchmark covers three principal performance areas meaningful to many cloud subscribers: Web hosting (user interface intensive), Java hosting (compute intensive) and mail (database/transaction intensive). To represent memory needs, ODCA suggest use of a default gigabytes-per-SPECvirt ratio and descriptions of double and quadruple memory density above this level.

For networks, measurement units must allow comparison of bandwidth, performance, and quality. Bandwidth can be represented in gigabits per second (gb/s). Performance can be quantified in latency/jitter/throughput per minute. Quality in networks, as in storage, is rated by level: Bronze, Silver, Gold, and Platinum.

2.11.2 Storage Quantitative Metrics/Tiers

Based on the Five Tier Storage Model [16] and on my experience, the technology was associated with the respective tiers.

Tier 0: SSD SLC, 2,5", RAID5

Business need: Extremely time sensitive, high value, volatile information needs to be captured, analyzed and presented at the highest possible speed. The primary example is currency trading. Note that this is a special-case situation not found in most business environments. **Solution**: Only storage with the highest, subsecond response speeds is good enough in the currency trading environment, where a single trade can make or lose more than the cost of the entire storage system. The usual solution is solid state storage, although new high speed disk technologies may compete in this space.

Tier 1: FC or SAS 6Gb 15k rpm, 2,5", RAID1

Business need: Transactional data requires fast, 100% accurate writes and reads either to support customers or meet the requirements of high-speed applications. One common example is online retail. Numerous surveys have shown that even relatively short delays in response to customer actions can result in lost sales, making high performance storage essential.
Solution: Generally latest-generation, high-speed disk systems are used. These systems carry a premium price, but this cost is justified because slower performance systems would directly impact the business. However, even as disk becomes faster, solid state storage prices are decreasing and availability is increasing. As this trend continues solid state "drives" will find their way into the Tier 1 systems of increasing numbers of organizations.

Tier 2: FC or SAS 6Gb 15k rpm, 2,5", RAID5

Business need: This tier supports many major business applications from email to ERP. It must securely store the majority of active business data, where subsecond response is not a

requirement but reasonably fast response still is needed. Email systems, which generate large amounts of data, are a prime example. While users can tolerate a slightly slower response times that is required for transactional systems, they are quickly frustrated by consistently slow response.

Solution: Tier 2 technology is always a balance between cost and performance. The latest entrant in this tier is XIV, now part of IBM, which offers large storage volumes and good-enough performance for Tier 2 at a very low price. The one catch is that to accomplish that, XIV systems come in two standard sizes. Multiple systems can be chained together to handle larger amounts of data, but the size minimum can lock out the lower end of the SMEs.

Tier 3: SATA 6Gb 7,2k rpm, 2,5", RAID 6

Business need: As data ages, reads drop off rapidly. However, that data often is still used for trend analysis and complex decision support. For instance, financial data needs to be kept accessible at least until the end of the fiscal/tax year. However, it does not need to stay on more expensive Tier 1 and Tier 2 systems. Similarly emails more than a couple of weeks old are seldom accessed, but the business may still find it desirable to keep them on easily accessible systems.

Solution: Tier 3 technologies can have two different characteristics. Much of the data in Tier 3 is really semi-active. MAID technology is a good choice for that data. However, it also handles data that supports decision support analysis. Businesses that do a lot of complex analysis of historical business data might consider a storage system designed to support complex queries such as the Sybase IQ series of column-based systems designed specifically to support complex analysis as a Tier 3B solution for that data.

Tier 4: VTL/TAPE

Business need: Compliance requirements today are driving a tremendous explosion in storage for historical data. In the United States, for instance, state and federal civil courts often require companies to produce large amounts of historic emails, sometimes going back years, in civil torts. Often businesses rely on backup tapes to recover this data. However, backup tape procedures were never designed to preserve data going back several years. Tapes are lost, reused, or may have deteriorated. The technologies to read them may no longer be available. J.P. Morgan and other companies have learned the inadequacies of using backup tapes to archive old data to their cost in some highly publicized court cases. Backup tapes have another problem when used for archiving. They contain a snapshot of the entire corporate data population across all applications at a particular moment. This means that when an Information Technology Office is responding to a court request, for instance, for emails between specific people between specific dates that may span several years, it has to resurrect and search a large number of tapes containing extraneous data such as corporate financials and HR records for individual emails and reconstruct them into a chronological file.

Businesses need a better system for long-term storage of historical data. This is as much an issue of procedure as technology. Rather than backup tapes, it needs tapes or other media containing archives of a specific data type – for instance all corporate email week by week. These need to be formally archived in an organized fashion with full records of exactly what is on each tape or other medium, where it is, what technology is needed to recover it, and when it needs to be migrated to a new physical medium to ensure that data integrity is maintained. Migration procedures need to be established to ensure that each tape is replaced before it reaches its end of life, and old tape technologies that are being replaced in the data center need to be preserved in the archive to read old tapes.

Solution: Tier 4 will contain very large amounts of data, but on the other hand no one expects this data to be instantly available. Courts, for instance, routinely give organizations two-to-four weeks to produce documents in discovery. For this reason, tape is by far the most cost-effective physical medium for much of this data. If the old data is needed either to respond to a court discovery or to support internal analysis, typically users can tolerate the time needed to mount the relevant archive tapes. Removable disks may also be considered. However, they tend to be more expensive and delicate than tape.

	Disks and Technology	IOPS	Latency	Throughput	Reference Cost
StorageT0	SSD SLC, 2,5', RAID5	25k	2ms	1.800MB/s	20.000 $
StorageT1	SAS 6Gb, 15krpm, 2,5', RAID1	5k	5ms	900MB/s	6.000 $
StorageT2	SAS 6Gb, 15krpm, 2,5', RAID5	1k	15ms	400MB/s	2.000 $
StorageT3	SATA(NSAS 6Gb) 7,2krpm,	1k	20ms	100MB/s	1.000 $
StorageT4	VTL / TAPE	-	-	-	<200 $

Table 2 – Average Tier characteristics (on-premises)

2.11.3 Computing Quantitative Measures

In a very simple analysis, compute quantitative measures are based on two elements: the CPU and the Memory:

The number of CPU and/ or cores times it's clock speed:

$CPU = \text{number of cores x speed in mhz}$

The total amount of memory times it's clock speed:

$Memory = total\ memoy\ in\ GB\ x\ speed\ in\ mhz$

2.11.4 Network Quantitative Measures

Just as the size of the water pipe to your house determines how much water can flow into your home, the capacity of the network pipe into your system is called "bandwidth" and determines how much data can flow.

However as we know, bandwidth is not equal to performance in Mbps, if it was so, 1000 Mbits NIC's always had a throughput of 1000 Mbits per second. Performance depends on the performance of all interconnected devices, and most of the times on the computing/Disk performance of the systems. So network performance is dependent on bandwidth, interconnected devices and systems performance.

2.12 Qualitative Measures: For Service Assurance Level

This usage model does not define how the cloud provider manages the infrastructure; instead it focuses on how the cloud subscriber wants to consume infrastructure. Therefore, we can define levels similar to CMMI/COBIT (levels 1–5) or using ITIL processes.

A framework of four levels of service assurance differentiation–Bronze, Silver, Gold, and Platinum–is identified. Each of these levels stands by itself and can be applied to various industry sectors and IT environments.

The attributes of the service levels, expressed in NFR (non-functional requirements) terms, are:

- **Availability**. The degree of uptime for the solution, taking into account contention probabilities, which includes an indication of response time to problems and incidents, planning and maintenance schedules and impacts, and business continuity capability.
- **Performance**. The extent to which the solution is assured to deliver a level of output.
- **Elasticity**. The configurability and expandability of the solution, including the ability to adjust consumed service capacities up or down, and the scale or limitations of the capacity changes.
- **Manageability**. The degree of automation and control available for managing the solution.
- **Recoverability**. The solution's recovery point and recovery time objectives.
- **Interoperability**. The degree to which services can interact with other services and infrastructure in multiple clouds. Interoperability is described from two perspectives: (1) portability–the serial process of moving a system from one cloud environment to another, and (2) interconnectability–the parallel process in which two co-existing environments communicate and interact.
- **Security and privacy**. Describes the attributes that indicate the effectiveness of a cloud provider's controls on access to services and data protection, and the physical facilities from which the services are provided. These attributes should provide an indication of physical protection, logical protection, controls and monitoring measures in place, compliance to country and corporate requirements, compliance with regulatory and statutory laws and obligations, and remediation processes.
- **Configurability**. Describes the features and functions of the services, including the available basic services, the available standard options to add to the base services, the available customizable options and features to add to the base services, the ability to develop custom features for the cloud subscriber, and the planned roadmap of functions and features for the service.
- **Long-distance migration**. "Long distance" is defined as greater than 20 km of conductor between disparate data centers (cloud provider sites). Inter-site latency is assumed to be at least 10 ms or worse. The characteristics of long-distance migration include cross-provider migration, cost-sensitive migration, open standards compliance, and live and at-rest migration parameters.

2.13 Capacity Planning Process

According with the article “How to Do Capacity Planning” [17], the basic steps to develop a capacity plan are:

1. Determine Service Level Requirements
    a. Define workloads
    b. Determine the unit of work
    c. Identify Service Levels for each workload

2. Analyze current system capacity
    a. Measure service levels and compare to objectives
    b. Measure overall resource usage
    c. Measure resource usage by workload
    d. Identify components of response time

3. Plan for the Future
    a. Determine future processing requirements
    b. Plan future system configuration

To accomplish the base of this capacity plan we must understand Workloads, unit of work and Service Level Agreements.

A system has several individual processes running, system processes, application processes and others processes. From the application processes we must identify those who are supporting our service.
Example: if our service is a e-mail service, the processes used by that service define the workload.

$$\text{Workload} = \sum \text{App processes for Service (CPU + RAM + Disk + Network)}$$

Then we must define the unit of work, for a mail service the unit of work may be the transaction send or receive an e-mail. And when talking about performance we accomplish these work using resources like CPUs, disk I/O and network. Measuring the utilization of these resources is important for capacity planning, but not relevant for determining the unit of work.

$$\text{Unit of Work} = \sum \text{Transaction (CPU + RAM + Disk + Network)}$$

The next step is to establish a Service Level Agreement, in the case of a e-mail service we might establish requirements regarding the number of transactions within a given period of time or we might that each request be processes within a certain time limit. SLA's include more than these capacity planning metrics, in the next chapter we will see other metrics like availability and reliability.

$$SLA = number\ transaction\ per\ second\ or\ MAX\ time\ for\ a\ transaction\ in\ ms$$

2.14 QoS in IT Systems

According with the book Performance by Design: Computer Capacity Planning by Example[13] a IT system has the following QoS attributes: response time, throughput, availability, reliability, security, scalability, and extensibility.

Response Time

Description: The time it takes a system to react to a human request.

Example: the time it takes for a page to appear in your browser with the results of a search.

Measure: usually measured in seconds.

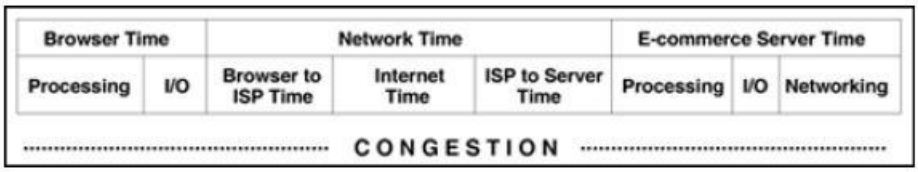


Figure 10 – Breakdown of response time example [13]

Client Time		Network Time			Server Time			
Processing	I/O	Client to ISP Time	Internet Time	ISP to Server Time	Processing	I/O	Networking	Storage

Table 3 – Breakdown of response time (adaptation)

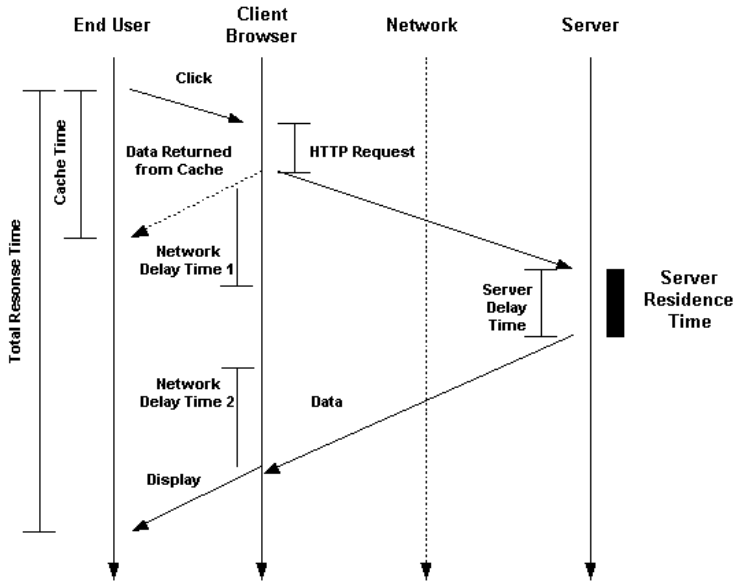


Figure 11 – Anatomy of a Web Transaction [13]

$$\text{Response Time} = \sum \text{Browser Time} + \text{Network Time} + \text{Server Time}$$

Throughput

Description: The rate at which requests are completed from a computer system.

Examples: DB transaction per second; HTTP requests/sec; Disk I/Os per second

Measure: per second

$$\text{Throughput} = \text{minimum}\,[\text{servercapacity}, \text{offeredworkload}]$$

Availability

Description: The fraction of time that a system is up and available to its customers.

Examples: 99.99% availability per year

Measure: percentage

$$\text{Availability} = (1 - 0.9999)\ \text{X 30 days 24h / day X 60min / hr} = 4{,}32\ \text{minutes}$$

Reliability

Description: the probability that it functions properly and continuously over a fixed period of time.

Example: 0,1% request errors (with very few error rates the reliability tends to the availability)

Measure: percentage

Security

Security is a combination of three basic attributes:

Confidentiality: only authorized individuals are allowed access to the relevant information.

Data Integrity: information cannot be modified by unauthorized users.

Non-repudiation: senders of a message are prevented from denying having sent the message.

Scalability

A system is said to be scalable if its performance does not degrade significantly as the number of users, or equivalently, the load on the system increases.

Extensibility

Extensibility is the property of a system to easily evolve to cope with new functional and performance requirements.

2.15 Capacity Planning Engineering Approach

An engineering approach to accomplish capacity planning as suggested by Menascé in 1999 [13] with the following steps:

1. Define performance metrics
    a. Number of simultaneous users (capacity)
    b. Response time (performance)
    c. Storage.
2. Characterize the workload
3. Measure performance
4. Analyze results
5. Develop cost and performance alternatives
6. Assess alternatives
7. Implement the best alternative

2.15.1 Scalable Architectures

On a traditional on-premises datacenter, one of the major concerns is the infrastructure ability to scale.

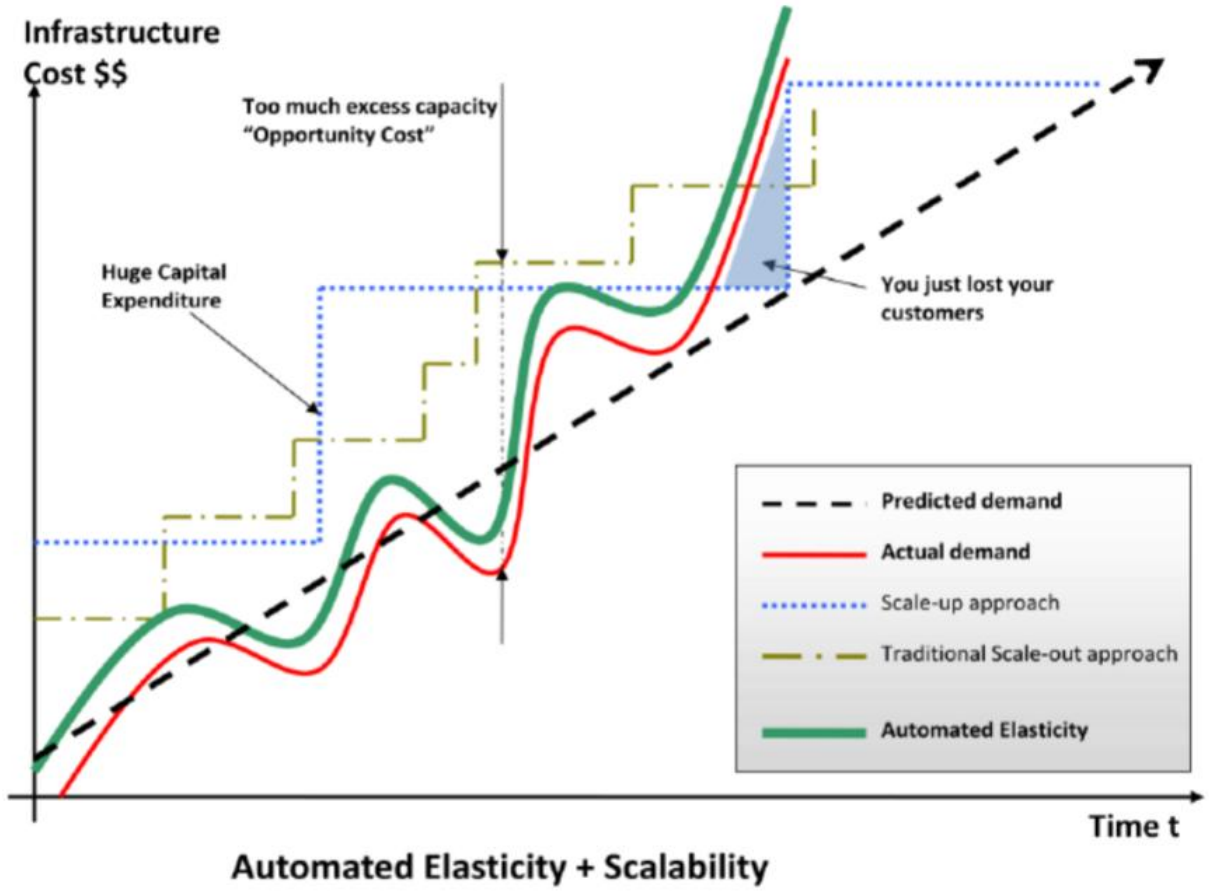


Figure 12 – Automated Elasticity + Scalability [4]

3. Telco-OTT Business Case State-of-the-Art

For better context of the main Business Case, the Contend Delivery Network and Content Delivery Techniques concepts where included in the following sub-topics.

3.1 Content Delivery Network

A content delivery network or content distribution network (CDN) is a globally distributed network of proxy servers deployed in multiple data centers. The goal of a CDN is to serve content to end-users with **high availability** and **high performance**. CDNs serve a large fraction of the Internet content today, including web objects (text, graphics and scripts), downloadable objects (media files, software, documents, etc.), applications (e-commerce, portals, etc.), **live streaming media**, **on-demand streaming media**, and social networks [18].

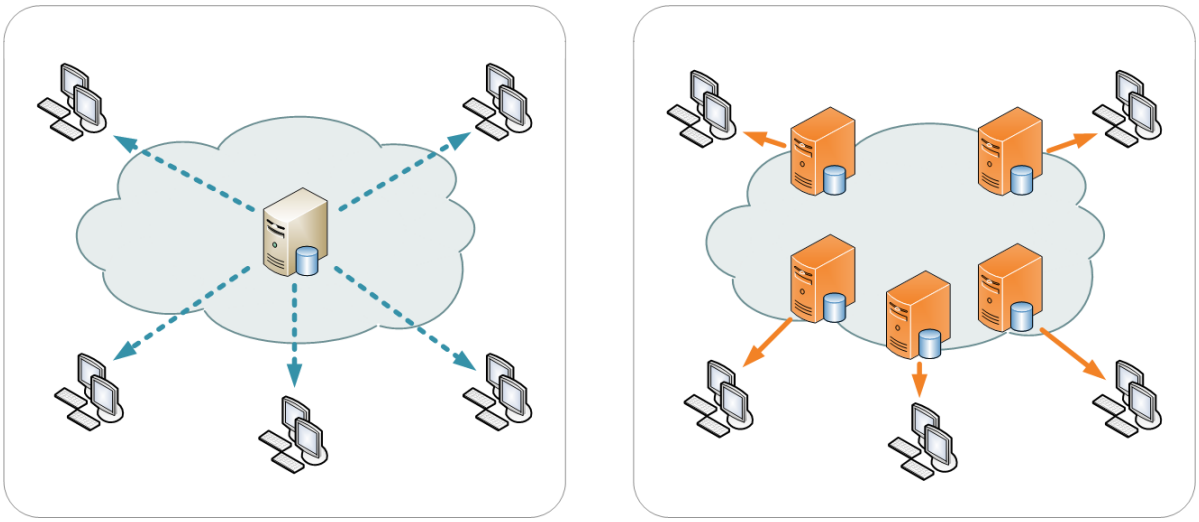


Figure 13 - (Left) Single server distribution (Right) CDN scheme of distribution [18]

Content delivery networks (CDN) are the transparent backbone of the Internet in charge of content delivery. Whether we know it or not, every one of us interacts with CDNs on a daily basis; when reading articles on news sites, shopping online, watching YouTube videos or read social media feeds.

No matter what you do, or what type of content you consume, chances are that you'll find CDNs behind every character of text, every image pixel and every movie frame that gets delivered to your PC and mobile browser.

To understand why CDNs are so widely used, you first need to recognize the issue they're designed to solve. Known as latency, it's the annoying delay that occurs from the moment you request to load a web page to the moment its content actually appears onscreen.

That delay interval is affected by a number of factors, many being specific to a given web page. In all cases however, the delay duration is impacted by the physical distance between you and that website's hosting server. A CDN's mission is to virtually shorten that physical distance, the goal being to improve site rendering speed and performance.

How a CDN Works

To minimize the distance between the visitors and your website's server, a CDN stores a cached version of its content in multiple geographical locations (a.k.a., points of presence, or PoPs). Each PoP contains a number of caching servers responsible for content delivery to visitors within its proximity.

In essence, CDN puts your content in many places at once, providing superior coverage to your users. For example, when someone in London accesses a US-hosted website, it is done through a local UK PoP. This is much quicker than having the visitor's requests, and your responses, travel the full width of the Atlantic and back[19].

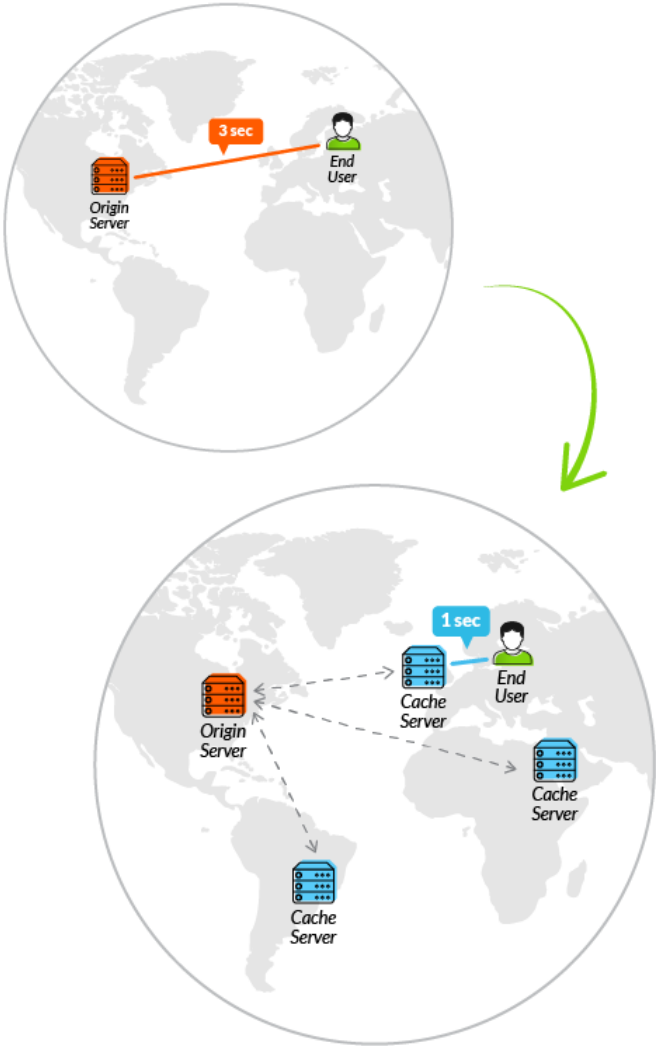


Figure 14 – CDN mission [19]

3.2 Content Delivery Techniques

Media delivery on the Web today uses three general delivery methods: traditional streaming, progressive download, and adaptive streaming. Refer to figure bellow for a simplistic stream workflow overview.

Figure 15 – How Streaming Video & Audio Work [20]

3.2.1 Traditional Streaming

RTSP (Real-Time Streaming Protocol) is a good example of a traditional streaming protocol. RTSP is defined as a stateful protocol, which means that from the first time a client connects to the streaming server until the time it disconnects from the streaming server, the server keeps track of the client's state. The client communicates its state to the server by issuing it commands such as PLAY, PAUSE or TEARDOWN (the first two are obvious; the last one is used to disconnect from the server and close the streaming session).

After a session between the client and the server has been established, the server begins sending the media as a steady stream of small packets (the format of these packets is known as RTP). The size of a typical RTP packet is 1452 bytes, which means that in a video stream encoded at 1 megabits per second (Mbps), each packet carries approximately 11 milliseconds of video. In RTSP the packets can be transmitted over either UDP or TCP transports—the latter is preferred when firewalls or proxies block UDP packets, but can also lead to increased latency (TCP packets are re-sent until received).[21]

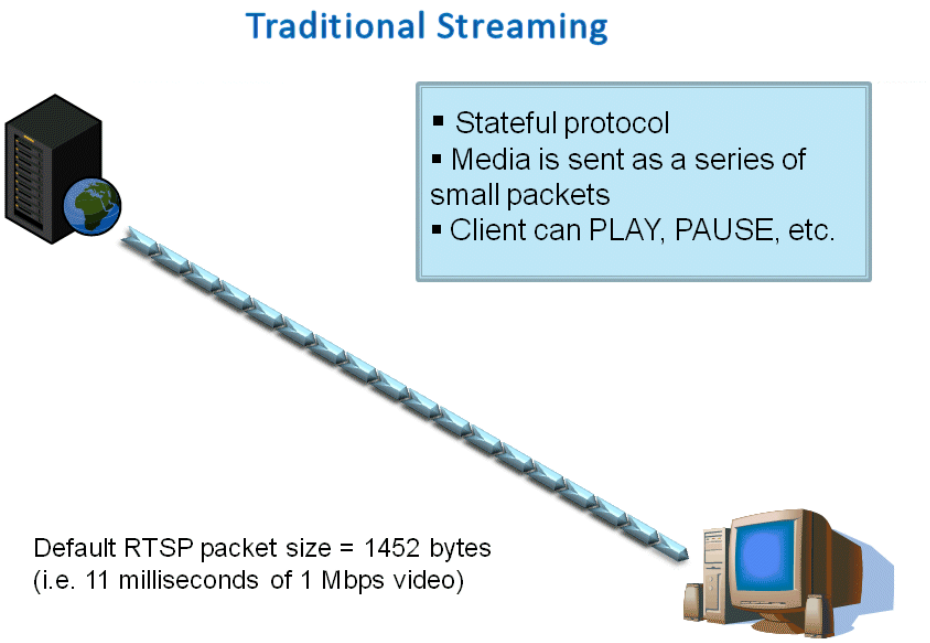


Figure 16 – RTSP is an example of a traditional streaming protocol [21]

HTTP, on the other hand, is known as a stateless protocol. If an HTTP client requests some data, the server responds by sending the data, but it won't remember the client or its state. Each HTTP request is handled as a completely standalone one-time session.

Windows Media Services supports streaming over both RTSP and HTTP. But if HTTP is a stateless protocol, how can it be used for streaming? Windows Media Services uses a modified version of HTTP officially known as MS-WMSP (known in Windows Media Services as the Windows Media HTTP Streaming Protocol, or more commonly just as Windows Media HTTP). MS-WMSP uses standard HTTP for transfer of data and messages but also maintains session states, effectively turning it into a streaming protocol like RTSP. Windows Media Services has also supported RTSP streaming since 2003 (in Windows Media Services 9 Series) over both UDP and TCP. Its implementation of the protocol is publicly documented as MS-RTSP.

The most important things to remember about traditional streaming protocols such as RTSP and Windows Media HTTP (MS-WMSP) are:

- The server sends the data packets to the client at a real-time rate only—that is, the bit rate at which the media is encoded. For example, a video encoded at 500 kilobits per second (kbps) is streamed to clients at approximately 500 kbps.
- The server only sends ahead enough data packets to fill the client buffer. The client buffer is typically between 1 and 10 seconds (Windows Media Player and Silverlight default buffer length is 5 seconds). This means that if you pause a streamed video and wait 10 minutes, still only approximately 5 seconds of video will have downloaded to the client in that time.

Other examples of traditional streaming protocols include Adobe Systems' proprietary Real Time Messaging Protocol (RTMP) and RealNetworks' RTSP over Real Data Transport (RDT) protocol. The Dynamic Streaming stream-switching feature in the Adobe® Flash® Platform is based on the RTMP protocol and is, therefore, considered a traditional streaming method—not adaptive streaming.[21]

3.2.2 Progressive Download

Another common form of media delivery on the Web today is progressive download, which is nothing more than a simple file download from an HTTP Web server. Progressive download is supported by most media players and platforms, including Adobe Flash, Silverlight, and Windows Media Player. The term "progressive" stems from the fact that most player clients allow the media file to be played back while the download is still in progress—before the entire file has been fully written to disk (typically to the Web browser cache). Clients that support the HTTP 1.1 specification can also seek to positions in the media file that haven't been downloaded yet by performing byte range requests to the Web server (assuming that it also supports HTTP 1.1).
Popular video sharing Web sites on the Web today, including YouTube, Vimeo and Quicktime, almost exclusively use progressive download.

Unlike streaming servers that rarely send more than 10 seconds of media data to the client at a time, HTTP Web servers keep the data flowing until the download is complete. If you pause a progressively downloaded video at the beginning of playback and then wait, the entire video will eventually have downloaded to your browser cache, allowing you to smoothly play the whole video without any “hiccups”. There is a downside to this behavior as well—if 30 seconds into a fully downloaded 10 minute video, you decide that you don't like it and quit the video, both you and your content provider have just wasted 9 minutes and 30 seconds worth of bandwidth. To try to mitigate this problem, IIS 7.0 provides a helpful extension called Bit Rate Throttling, which allows content providers to throttle the download bit rate in exactly the same way that a streaming server would to reduce costs.[21]

3.2.3 HTTP-Based Adaptive Streaming

Adaptive streaming is a hybrid delivery method that acts like streaming but is based on HTTP progressive download. It's an advanced concept that uses HTTP rather than a new protocol. Both IIS Smooth Streaming and Move Networks’ Adaptive Stream are examples of adaptive streaming. Even though the two technologies use different codec’s, formats, and encryption schemes, they both rely on HTTP as the transport protocol and perform the media download as a long series of very small progressive downloads, rather than one big progressive download.
In a typical adaptive streaming implementation, the video/audio source is cut into many short segments ("chunks") and encoded to the desired delivery format. Chunks are typically 2-to-4-seconds long. At the video codec level, this typically means that each chunk is cut along video

GOP (Group of Pictures) boundaries (each chunk starts with a key frame) and has no dependencies on past or future chunks/GOPs. This allows each chunk to later be decoded independently of other chunks.

The encoded chunks are hosted on a HTTP Web server. A client requests the chunks from the Web server in a linear fashion and downloads them using plain HTTP progressive download. As the chunks are downloaded to the client, the client plays back the sequence of chunks in linear order. Because the chunks are carefully encoded without any gaps or overlaps between them, the chunks play back as a seamless video.

The "adaptive" part of the solution comes into play when the video/audio source is encoded at multiple bit rates, generating multiple chunks of various sizes for each 2-to-4-seconds of video. The client can now choose between chunks of different sizes. Because Web servers usually deliver data as fast as network bandwidth allows them to, the client can easily estimate user bandwidth and decide to download larger or smaller chunks ahead of time. The size of the playback/download buffer is fully customizable.[21]

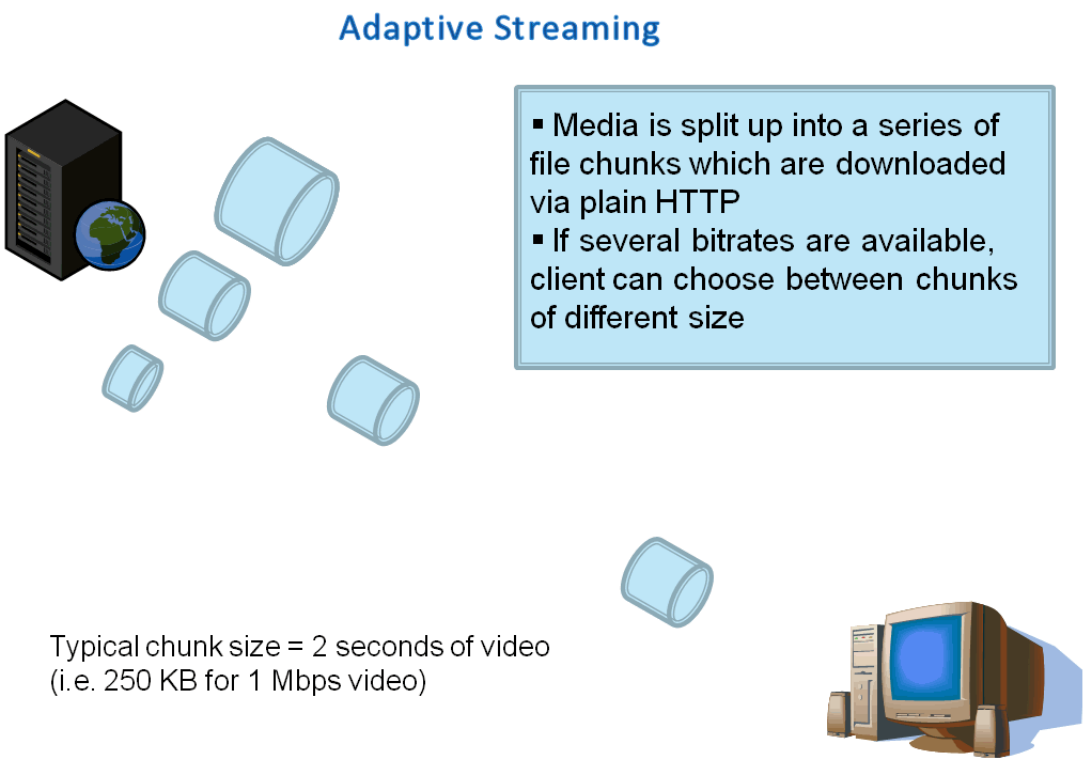


Figure 17 – Adaptive streaming is a hybrid media delivery method [21]

Adaptive streaming, like other forms of HTTP delivery, offers the following advantages over traditional streaming to the content distributor:

- It's cheaper to deploy because adaptive streaming can use generic HTTP caches/proxies and doesn't require specialized servers at each node.
- It offers better scalability and reach, reducing "last mile" issues because it can dynamically adapt to inferior network conditions as it gets closer to the user's home.
- It lets the audience adapt to the content, rather than requiring content providers to guess which bit rates are most likely to be accessible to their audience.

It also offers the following benefits for the user:

- Fast start-up and seek times because start-up/seeking can be initiated on the lowest bit rate before moving to a higher bit rate.
- No buffering, no disconnects, no playback stutter (as long as the user meets the minimum bit rate requirement).
- Seamless bit rate switching based on network conditions and CPU capabilities.
- A generally consistent, smooth playback experience.

4. Cloud Computing

4.1 History

The fundamental concept of cloud computing originated in the 1950s, when corporations and learning institutes prioritized the efficiency of their large-scale mainframe computers, allowing multiple users both physical access to the computer from multiple terminals as well as shared central processing unit time.[22]

But the overarching concept of delivering computing resources through a global network is rooted in the sixties. The idea of a global network was introduced in the sixties by J.C.R. Licklider, who was responsible for enabling the development of ARPANET in 1969. His vision was for everyone on the globe to be interconnected and accessing programs and data at any site, from anywhere [23]. Sounds a lot like what we are calling cloud computing today.

Since the internet only started to offer significant bandwidth in the nineties, cloud computing for the masses has been something of a late developer.

One of the first milestones in cloud computing history was the arrival of Salesforce.com in 1999, which pioneered the concept of delivering enterprise applications via a simple website. The services firm paved the way for both specialist and mainstream software firms to deliver applications over the internet.

After the dot-com bubble burst in the early 2000s, companies such as e-commerce giant Amazon played a key role in the development of cloud computing.
“The development was Amazon Web Services in 2002, which provided a suite of cloud-based services including storage, computation and even human intelligence through the Amazon Mechanical Turk.
Then in 2006, Amazon launched its Elastic Compute cloud (EC2) as a commercial web service that allows small companies and individuals to rent computers on which to run their own computer applications. Amazon EC2/S3 was the first widely accessible cloud computing infrastructure service.”[23]

Another big milestone came in 2009, as Web 2.0 hit its stride, and Google and others started to offer browser-based enterprise applications, though services such as Google Apps.
The most important contribution to cloud computing has been the emergence of "Apps" from leading technology giants such as Google and Microsoft. When these companies deliver services in a way that is reliable and easy to consume, the effect to the industry as a whole is a wider general acceptance of online services.[23]

Other key factors that have enabled cloud computing to evolve include the maturing of virtualization technology, the development of universal high-speed bandwidth, and universal software interoperability standards.

The present availability of high-capacity networks and low-cost computers, together with the widespread adoption of virtualization, automation and service-oriented architecture, has led to the version of cloud computing we know today.

Origin of the term

The word "cloud" is commonly used in science to describe a large agglomeration of objects that visually appear from a distance as a cloud and describes any set of things whose details are not further inspected in a given context. Another explanation is that the old programs that drew network schematics surrounded the icons for servers with a circle, and a cluster of servers in a network diagram had several overlapping circles, which resembled a cloud. In analogy to the above usage, the word cloud was used as a metaphor for the Internet and a standardized cloud-like shape was used to denote a network on telephony schematics. Later it was used to depict the Internet in computer network diagrams. With this simplification, the implication is that the specifics of how the end points of a network are connected are not relevant for the purposes of understanding the diagram [24].

4.2 Generational Shift

We're all accustomed to the rapid and continuous evolution of technology but, from a broader scale, technology can be viewed as having passed through three distinct generational phases. Technology eras are bounded by three fundamental periods: the mainframe, client-server, and the Web.

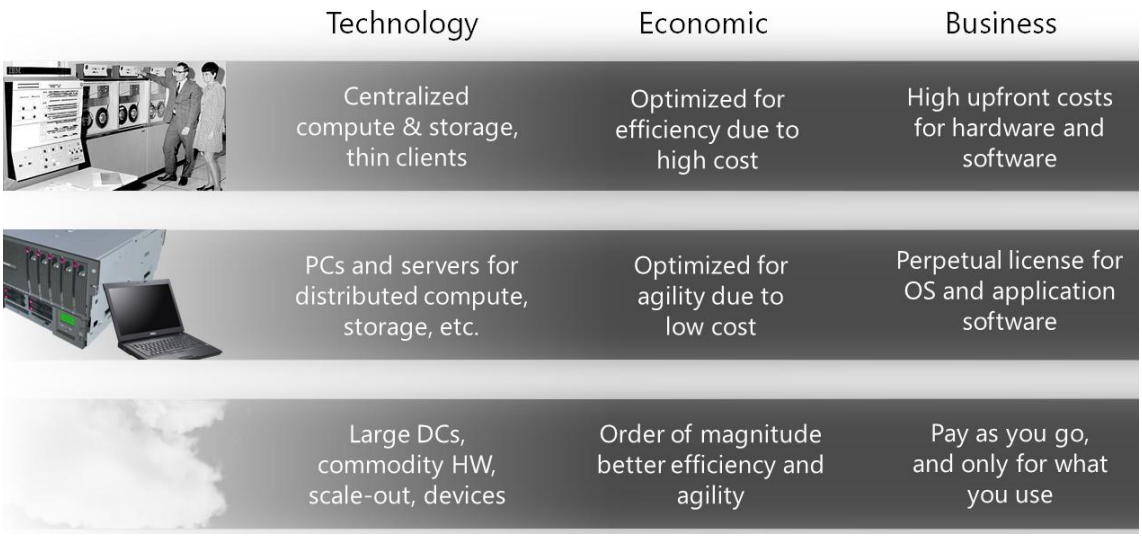


Figure 18 – The three fundamental periods [25]

Each shift can be summarized as a combination of the technology that's often the catalyst of the shift, the economic environment that makes the technology conducive.

4.3 Economies of Scale

Cloud economies of scale are stronger than commonly understood, of course, very few organizations with the wherewithal (technical and financial) to run public data-centers of the scale needed to reach optimum efficiency.
At massive scale, the cloud provider is able to negotiate steep discounts with hardware, power and bandwidth providers which further secures its advantages. When 100k+ machines are on order, hardware providers will tailor everything from their hardware designs to their support capabilities to win the business. Not only does the cloud provider require huge bandwidth today but it's in a position to secure long-term contracts on such bandwidth which provides strong financial incentives to the bandwidth provider.

The economics of scale continue. A public cloud provider is able to pool demand from a diverse customer base. Customers in different businesses require peak computing loads at different times. Customers across different industries are discordant with some industries experiencing peak demand in the Winter (e.g.: retailers) while others experience peak demand in the Summer (e.g.: music festivals). Customers in different time zones across the world are 'awake' at different times. The public cloud provider can dramatically increase the utilization of its hardware by multiplexing across a range of these customers and to a far greater extent than a provider to a single business.

Lastly, through extensive automation and the ability to run fewer applications for many customers, the labor costs in a data-center become a very small proportion of overall costs. In a contemporary on-premises data-center, an IT professional may be able to manage 100 machines. In a public cloud data-center, one person needs to support 1,000s of machines.

4.4 Definition

Although there are many definitions of cloud computing, the US National Institute of Standards and Technology (NIST) has published a working definition that has captured the commonly agreed aspects of cloud computing.

The NIST Definition of Cloud Computing:
“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 five essential characteristics, three service models, and four deployment models.” [26]

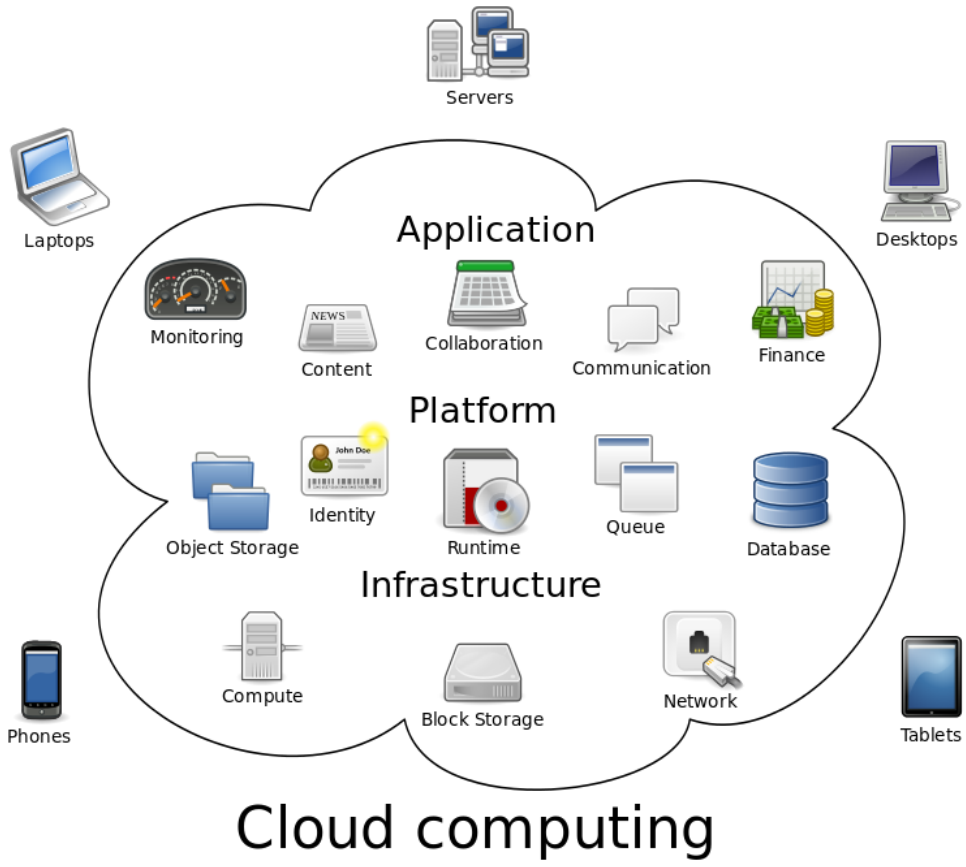


Figure 19 – Cloud Computing Metaphor [24]

4.5 Essential Characteristics

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.

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).

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.

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.

Measured service. Cloud systems automatically control and optimize resource use by leveraging a metering capability1 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.

4.6 Service Models

Though service-oriented architecture advocates "everything as a service" (with the acronyms EaaS or XaaS or simply aas), cloud-computing providers offer their "services" according to different models, of which the three standard models per NIST are Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

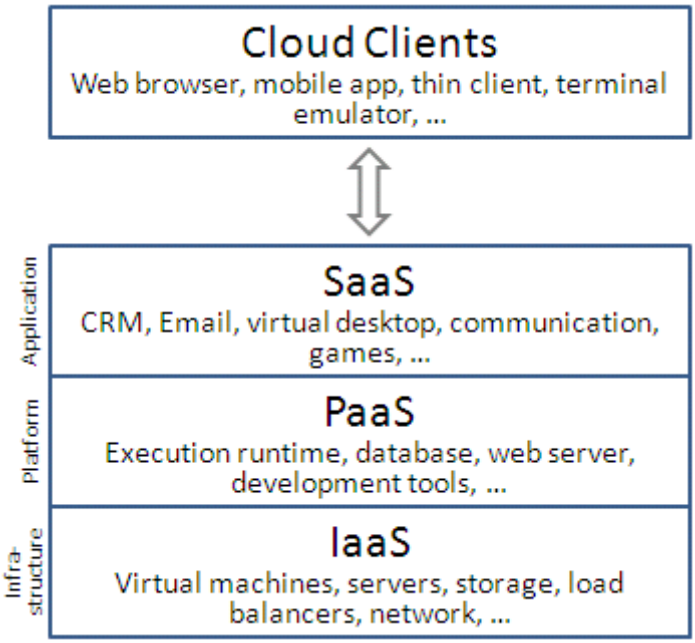


Figure 20 - Service Models arranged as layers in a stack [24]

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.

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.

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).

Colocation Service Model

This Service Model is more uncommon, and not every cloud providers include this offer.
A colocation centre (also spelled co-location, or colo) or "carrier hotel", is a type of data centre where equipment, space, and bandwidth are available for rental to retail customers. Colocation facilities provide space, power, cooling, and physical security for the server, storage, and networking equipment of other firms—and connect them to a variety of telecommunications and network service providers—with a minimum of cost and complexity[27].

Typically the rack or racks are physically secured in a cage, some other services are offered by the provider like “eyes” and “hands”. Where the operator from the colocation center can observe the systems and inform the client of any visible alarm from the system (eyes) and can perform a system reset or in some cases even a simple hardware replacement (e.g.: Disk) (hands).

4.7 Deployment Models

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.

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.

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.

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).

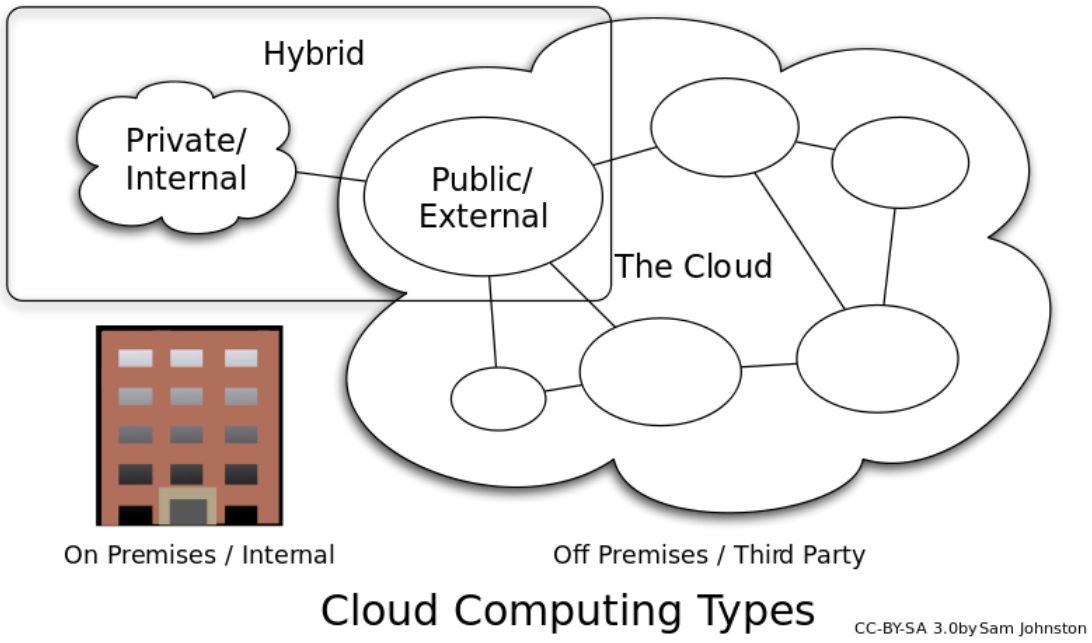


Figure 21 – Deployment Models [24]

4.8 Main Providers

According with RightScale 2015 state of the cloud report [28], based on % of survey respondents running Applications, these are the main providers:

1. Amazon AWS
2. Microsoft Azure IaaS
3. Rackspace Public Cloud
4. Microsoft Azure PaaS
5. Google AppEngine

For the purpose of our analysis we will compare the on-premises with the following cloud providers:

1. Amazon AWS
2. Microsoft Azure
3. Google Cloud Platform.

Amazon, Azure and Google to be analyzed as a full cloud platform and Rackspace was removed because its business model is mostly to provide Cloud Servers and Dedicated Servers mainly based on technology from the other cloud providers, making Rackspace more a broker then a pure Cloud Provider. However being more used then Google Cloud Platform.

4.8.1 Quick comparing Top 3 Cloud Providers

Cloud vendor competition is heading to peek, including a pricing war and new feature offerings in the battle for new clients. Officially launched in 2006, Amazon is the market leader with Azure following, while Google entered latter and it's trying to recover the gap.

According to the Cloudyn White Paper WHO MOVED MY CLOUD? Part II: What's behind the cloud vendors AWS, Azure and GCP[29] compares the three top cloud vendors, picture 1 states the differences between the three giants, in terms of speed, block storage, network, pricing & models.

	AWS	GCE	Azure
Features			
Instances: families and types	5 families 23 types	4 families 15 types	2 families 13 types
Load balancer	Elastic LB (internal) DNS service (internal) Auto-scaler (external)	Combined solution	Traffic management combined solution
I/O speeds	Second place	First place	Third place
Block storage	1 TB	Up to 10 TB	1 TB
Global network	Wide regional offering, most data centers and access points	Narrower regions, centers and access, faster per connection	Global datacenter infrastructure
Connection	Open Internet	Regional fiber network	Open Internet
Billing			
Pricing	Per hour – rounded up	Per minute – rounded up (minimum 10 minutes)	Per minute – rounded up commitments (pre-paid or monthly)
Models	On-demand, Reserved, Spot	On-demand, sustained-use	On-demand, short term commitments (pre-paid or monthly)

Figure 22 – AWS vs GCE vs Azure [29]

Concluding that companies preparing for cloud migration must do their homework to find the best route for them. Each vendor has its own strengths and companies must figure out which KPI's are most important to them and choose a vendor accordingly.

5. Business Cases

5.1 Assumptions

Comparing on-premises with cloud provider solution is not a direct approach, consequently some assumption had to be done to balance and make the comparison as fair as possible.

It was assumed a client already had a datacenter facility, there is no meaning in include the costs of building a proper datacenter. However, to balance the comparisons the costs of collocation were included in on-premises costs. Even knowing Colocation service already includes some basic energy and network service, but this duplicated cost was assumed because of is insignificant financial impact.

The on-premises equipments cost was based on the average costs of state-of-the-art enterprise solutions from the major manufactures based on my professional experience, with a lifetime period of 4+ years. The costs of Maintenance and Support are also included and adapted, some manufactures offer more and others less than the 4 years.

Labor costs were partially ignored, Cloud Providers tend to flag zero labor costs, however staff with skills like operating systems, storage, networks, etc. is still needed in IaaS deployment on the Public Cloud. Only on some simple SaaS, for instances the E-mail Business case we can ignore IT staff costs on a Public cloud.
Nevertheless, obviously some staff skills are only needed on-premises like Server Room skills on the “On-premises Data Center Requirements” topic and physical Rack, Server, Storage and Network assembling. Once more to balance the comparisons the cost of two IT professionals where included in on premises costs, based on the average US salary.

Based on the initial systems characterizations, two different business cases where chosen, one where the users reside on the internet, and other where the users reside mainly (and most of the system usage) on-premises.

The first and main business case compares on-premises IaaS to Cloud Providers IaaS solutions, in order to create a viable comparison, VM vs VM (or Instance VS Instance) was adopted. So, the same computing capacity (CPU and Memory) and Storage should be achieved in both scenarios (on-premises and Cloud Providers).
Some Operational Laws from “Defining of the Quantitative Model”, chapter 9, where applied, to confirm technology provider proposed capacity and predict future growth.
Technology provider specified the Hardware requirements to on-premises, that give the base computing capacity to compare with the several cloud providers. Because it is an IaaS solution, Applications Software costs where the same to both scenarios and so excluded from comparison. For the same reason, CDN costs were also excluded from the comparison.

The second business case compares on-premises hardware infrastructure to Cloud Providers SaaS solutions. In this second Business case, Cloud Providers set the mailbox capacity and features in the best solutions, focus mainly in E-mail Service. Some capacity and the TCO comparison is presented, however no purpose in comparing infrastructure and operational laws because we are comparing with a SaaS solution.

The costs comparison has four years duration:
On-premises (Hardware Acquisition + Colocation + Energy consumption + Internet Access) vs Public Cloud (Computing and Storage + Internet Access + Bandwidth consumption)

Because larger server configuration on some Cloud Providers where only available on the USA, we assumed the USA as default location and also the Dollar as our currency.

About Financial Analysis, as we do not have data on revenue estimates, it is not possible to obtain or calculate the NPV indicator, because Net Present Value needs the amount paid in each year and the receipts obtained with the investment. The IRR indicator can be calculated from the NPV, so as well is not possible. ROI is the indicator of return on investment = (revenues - costs) / costs; Therefore is not possible for the same cause.
So our Financial Analysis will be based on TCO, and to do this we need to know two simple concepts:

1. **Opex:** is an ongoing cost for running a product, business, or system.
2. **Capex**: is the capital expenditures realized in that year.

Total Cost of Ownership (TCO) is an important measure to the managers because it includes all the additional costs required to support and maintain for its full useful life and not only the initial cost on the purchase moment.
In the Information Technology industry, TCO has been used extensively in the acquisition of information technology hardware. TCO of hardware typically incorporated the estimated cost of out-of-warranty repairs; the cost of annual service agreements, and the prorated cost of additional hardware and software required to leverage the hardware. Today, many firms rent services from cloud providers, eliminating large investments in hardware and reducing the annual costs and the TCO incurred for running a firm's software applications.

So, in this case is important to estimate all the costs of the lifetime of the hardware and software that the manager must have into account in order to make the best decision according your benefits may impact in your core business. You cannot forget the possible hidden costs, for example, the labor costs, the training for users over the time, etc. The decision depends of the return of your investment and the benefits that you may have in your main business. It´s important to evaluate the profit gaining or the internal economical or social performance after make the investment and all the cost for maintain and support the solution during the lifetime.

5.2 Telco-OTT Business case

Let's start from some IPTV, OTT and Telco-OTT definitions and their relation.

IPTV Definition

"IPTV is defined as multimedia services such as television/video/audio/text/graphics/data delivered over IP based networks managed to provide the required level of quality of service and experience, security, interactivity and reliability."
Official definition approved by the International Telecommunication Union.

Deployed over copper, fiber, wireless, and HFC networks to deliver digital TV services over IP Networks, including real-time broadcast TV services.
Leverages IP technologies to deliver an enhanced TV experience in the context of Triple Play (video, voice, data) services.

OTT definition

In broadcasting, over-the-top content (OTT) is the delivery of audio, video, and other media over the Internet without the involvement of a multiple-system operator in the control or distribution of the content. The Internet provider may be aware of the contents of the Internet Protocol packets but is not responsible for, nor able to control, the viewing abilities, copyrights, and/or other redistribution of the content. This model contrasts with the purchasing or rental of video or audio content from an Internet service provider (ISP), such as pay television, video on demand or an IPTV video service. OTT refers to content from a third party that is delivered to an end-user, with the ISP simply transporting IP packets [30].

Telco-OTT definition

Telco-OTT (Over-The-Top) is where a telecommunications service provider delivers one or more services across an IP network. The IP networks is predominantly the public internet although sometimes telco-run cloud services delivered via a corporation's existing IP-VPN from another provider, as opposed to the carrier's own access network. It embraces a variety of telco services including communications (e.g. voice and messaging), content (e.g. TV and music) and cloud-based (e.g. compute and storage) offerings.

Stimulated by the availability of high performance fixed and mobile broadband networks as well as the rapid adoption of smartphones and tablet computers, telco-OTT is viewed by a selection of industry analysts and media commentators as the mechanism that mobile network operators need to employ in order to compete with the vast and growing range of over-the-top (OTT) services provided by non-telco companies.

Telco-OTT is a response to the fact that users will have multiple devices (smartphones, laptops or other connected devices such as TVs, games consoles) which almost inevitably will have

various different access providers (especially with the growth of public-access Wi-Fi). So to deliver consistent telco-branded services, at some points at least, they will need to be delivered over 3rd-party access [31].

Value Chain

Figure 23 – Value Chain

Major features

1. Switched Digital Broadcast Channels (SDB)
2. Video-on-Demand (VOD)
3. Digital Video Recorder (DVR, PVR)
4. Network based Personal Video Recorder (nPVR)
5. Interactive TV applications (iTV)
6. Electronic Program Guide (EPG)

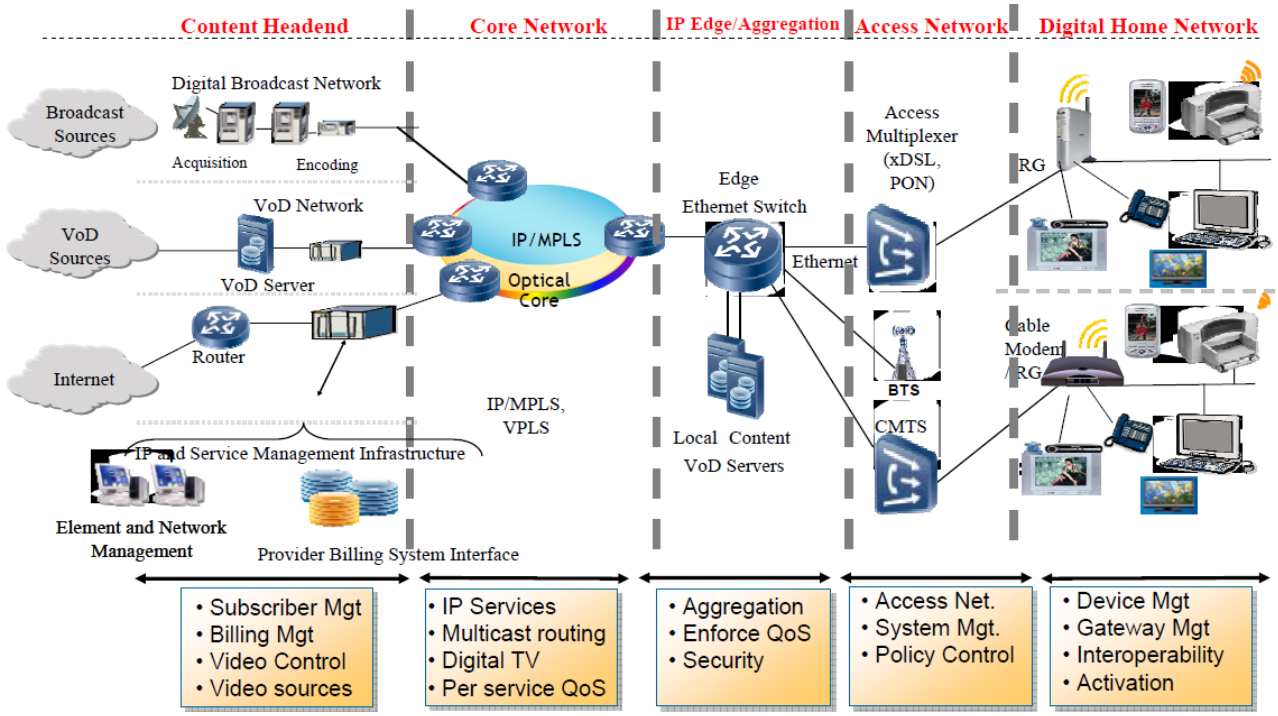


Figure 24 – The IPTV ecosystem [32]

Figure 25 – Mediaroom Platform Overview [33]

Usage Pattern

IPTV Usage pattern is predictable burst, with spikes of usage from late afternoon to first dawn hours and all weekend.

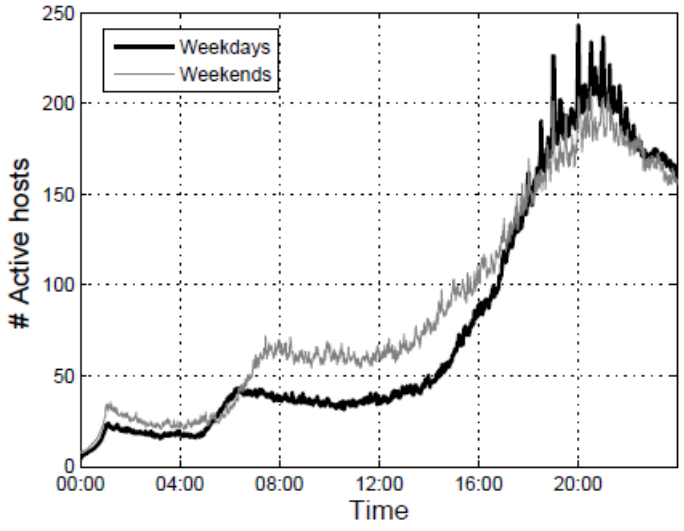


Figure 26 – Average number of active hosts [34]

Workloads

There are two typical usages evolving a user access to the service, first:

- *Online Experience* - web or App media streaming for PC, MAC and portable devices.
- *Interativo* – EPG (Electronic Programming Guide) for SmartTV.

Second, *IpTV Enterprise Service Framework, for* media catalogue, epg, broker and integration services.

Finally, *Back-end systems, for* accounting, auditing, management, etc.

The Media Stream Workflow

The media selection is based on a VOD catalog or EPG (for Live TV) that is presented to the user. A high level workflow of the entire process is shown in the following picture.

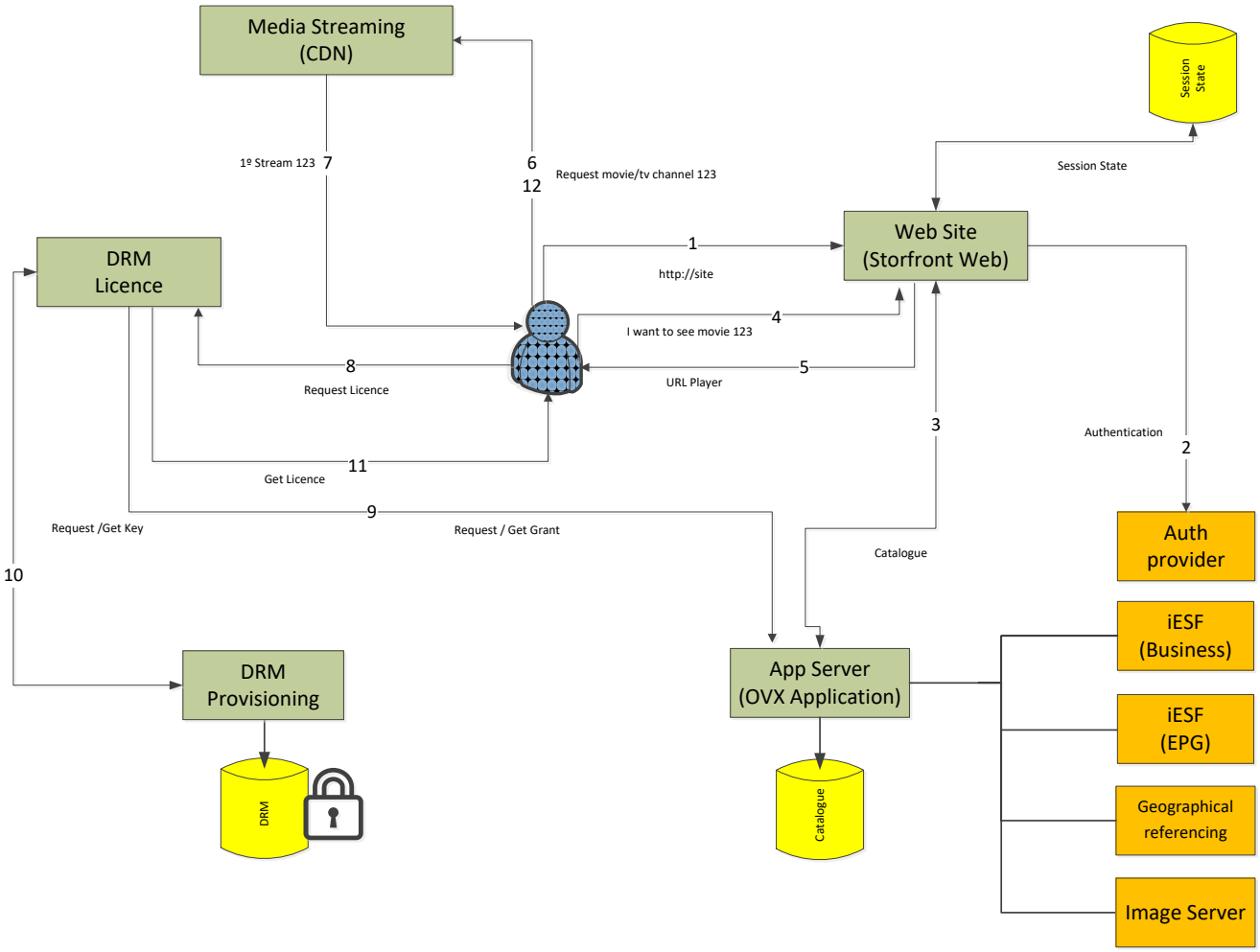


Figure 27 - Media streaming workflow

The workflow is centered on the end-user, and all communications between him and the platform are done through a web browser or Application that connect via http/https.

1. The user starts his contact with the platform via the website url (e.g. http://media.site.com). At the site the user identifies the content he wants to view.

The content provided is presented through a catalogue supplied by means of the OVX Application servers;

2. A form is presented to the user for him to log in to the platform. His credentials are validated by the Auth provider;
3. Once logged in, the presented catalog information can be dynamically updated in order to reflect specific information to the particular user;
4. Navigating through the catalog, the user identifies a content he wants to view;
5. The user's browser is then redirected to the location the media is maintained;
6. Through the redirection, a communication is established with the media streaming servers in the CDN;
7. As the content starts to be streamed, an url is identified to download a valid license to decrypt the stream;
8. Through the identified url by the streaming media server, a valid license for the media is requested to the DRM License servers;
9. The DRM License server validates (via OVX Application servers) that the user in question has a valid license to the content in question and if otherwise, carries out the request for issuance and billing for the requested license;
10. After a valid response of the OVX Application servers, the DRM License servers require a valid license to the user in question for the requested content;
11. The license is delivered to the user;
12. Through the license, the user has the ability to decode the requested content and start viewing it.

For the simplicity of this Business Case, the performance calculations will be based only on the Client to Edge/cache Server workflow (the system bottleneck), the following diagram will be considered:

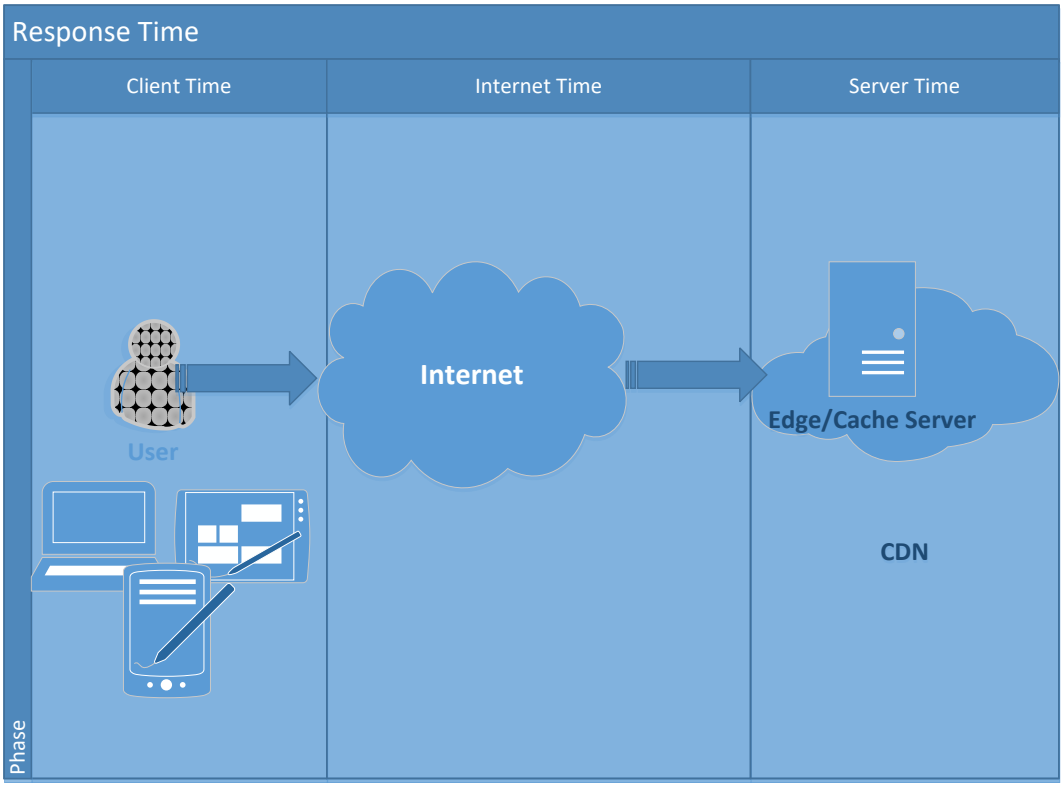


Figure 28 – Telco-OTT Response Time

By experiments, the average time to open the website or app and immediately ask for a live channel were from 10 seconds in 4G/ADSL networks to 20 seconds in poor 3G connection. All Edge/Cache Servers contain **presentation** (web or app layer) and **streams** for all offered channels.

A tracert running from client to Edge/cache Server showed the following response time:

Client Time: 10ms; Internet Time: 50ms; Server Time: 100ms

Response Time = client time + internet time + server time = 10 + 50 + 100 = 160ms

Based on the "Basic Performance" chapter 9.1, let us register some basic data measured for our performance calculations:

T = 60 sec (measured time interval)

K = 1 resource (calculations are made per Edge/cache Server)

B_i = 36 sec (60%)

$A_i = A_0$ = 400 transactions (Arrivals in T)

$C_i = C_0$ = 400 transactions (Completed in T)

Consequently, the derived quantities are:

$S_i = \frac{B_i}{C_i} = \frac{36}{400} = 0{,}09$ minutes or 5,4 second per transaction

$$U_i = \frac{B_i}{T} = \frac{36}{60} = 60\%$$

$$\lambda_i = \frac{A_i}{T} = \frac{400}{60} = 6{,}6\ tps$$

$$X_i = \frac{C_0}{T} = \frac{400}{60} = 6{,}6\ tps$$

Following, based on the "definition of the Quantitative Model" chapter 9 we will now use the four basic formulas previously presented in relation to performance calculation:

1. **Basic Performance**

Little's Law says that in a stable process there is only three characteristics that govern that process. If one characteristic changes, the others will change too. The three characteristics are:

Arrival Rate, **Time Spent in Queue** and **System Throughput**

We can summarize Little's Law by saying:

$\lambda = X\,W$ (Arrival Rate = System Throughput * Time Spent in Queue) Or

$W = \frac{\lambda}{X}$ (Time Spent in Queue = Arrival Rate / System Throughput)

Telco-OTT website as an average Arrival Rate of 100 clients, the throughput is 400 clients/minute. Calculating Time Spent in Queue with Little's Law:

$W = \frac{\lambda}{X} = \frac{100}{400} = 0{,}25$ Minutes or 15 seconds (average time a client as to wait in the queue at the website)

2. **Utilization,** measures the fraction of time that the resources are busy

- Utilization Law: $\mathrm{U_i} = \mathrm{X_i} \times \mathrm{S_i} = \lambda_{\mathrm{i}} \times \mathrm{S_i}$

Where X is the system throughput of a queue per unit of time and S is the average service time of a request

From basic performance calculations we already know the Server throughput is 6,6 transactions per second and the average service time is 0,09 minutes per transaction.

Starting by identifying the operational variables provided or that can be obtained from the measured data:

The Server is a resource (K=1)

The throughput of resource Xi is 6,6 transactions/second

The average service time of a request is 0,09 minutes

Using the Utilization Law, we compute the utilization of the link as:

$S_i \times X_i = 0.09 \times 6{,}6 = 0{,}6 = 60\%$

3. **Forced Flow,** measures average throughput of a resource and the percentage of utilization

- Forced Flow Law: $X_i = V_i \times X_0$

Transactions on server averaged 400 operations per minute:

$$X_{server} = \frac{400}{60} = 6{,}6\ tps\ (transactions\ per\ second)$$

Forced Flow Law gives us the number of visits to a resource (e.g.: CPU, RAM or HDD) in order to complete a transaction. To maintain the simplicity of the business case we will rather calculate the system throughput (X) based on the resource (server) throughput times the number of resources(N):

$$X = NX_{Server} = 12 \times 6{,}6 = 79{,}2\ tps$$

4. **Service Demand**, sum of all service times for a request at the resource and is related to the system throughput and utilization

- Service Demand Law: $D = V_i \times S_i = \frac{X_i}{X_0} \times \frac{U_i}{X_i} = \frac{U_i}{X_0}$

System was monitored for 60 minutes and it was observed that the CPU was 60% busy during the monitoring period, and the number of HTTP requests counted in the log was 24.000

$U_{cpu} = 60\%$

$X_{server} = \frac{24000}{3600} = 6.6\ requests\ per\ second$

$D_{cpu} = V_{cpu} \times S_{cpu} = \frac{U_{cpu}}{X_{server}} = \frac{0.60}{6.6} = 0.090\ seconds\ as\ the\ CPU\ demand\ for\ HTTP\ request$

By these calculations we know the system throughput is 4800 simultaneous users (12 times 400) and the average simultaneous users at Service at any time to be 1200 (12 times 100).
We have 100.000 active users every month and a maximum of 4800 users, we also can say the maximum system utilization rate is 4,8%.
Assuming utilization rate increases 5% a year we would need to duplicate the resources (Edge/cache servers) each year to maintain capacity. However could be continuously increased resource by resource knowing we will be adding 400 simultaneous user capacity.

Telco-OTT Scope

As previously indicated, for the simplicity of this Business Case the performance calculations where done only for the Edge/Cache Servers (OVX Presentation Services on table 6).
For informational purpose, the present architecture, assumes that an existing IPTV operator wants to add Telco-OTT functionality. The following components would be necessary and are addressed; however no performance calculations are done to them.

Infrastructure Services

- Authentication
- System Center (monitoring, configuration, backups, etc.)
- Remote Desktop /VDI
- Firewall

Telco-OTT (Over the Top)

- Online Experience
- ipTV Enterprise Service Framework;
- Interative;
- Relevant infra-structure services to support the above mentioned solutions.

Interface for Telco Operations Support Systems (OSS):

- Service and Channel Management
- System and Device Management
- Diagnostics
- Content meta-data for linear and on-demand

Interface for Telco Business Support Systems (BSS):

- Principal (devices, accounts, users, and subscriber groups) Management
- Rights Management
- Billing Management
- Offer Management

The Mediaroom (with Live TV Encoders and respective Back-end Systems) solution is out of the scope of the Business case.

Logical Architecture

The OTT will be comprised of several interconnected dedicated solutions. These solutions depend on each other to permit the end user a seamless IP based video experience, whether is for VOD or Live TV consumption.

The following picture illustrates the depicted solution in this platform.

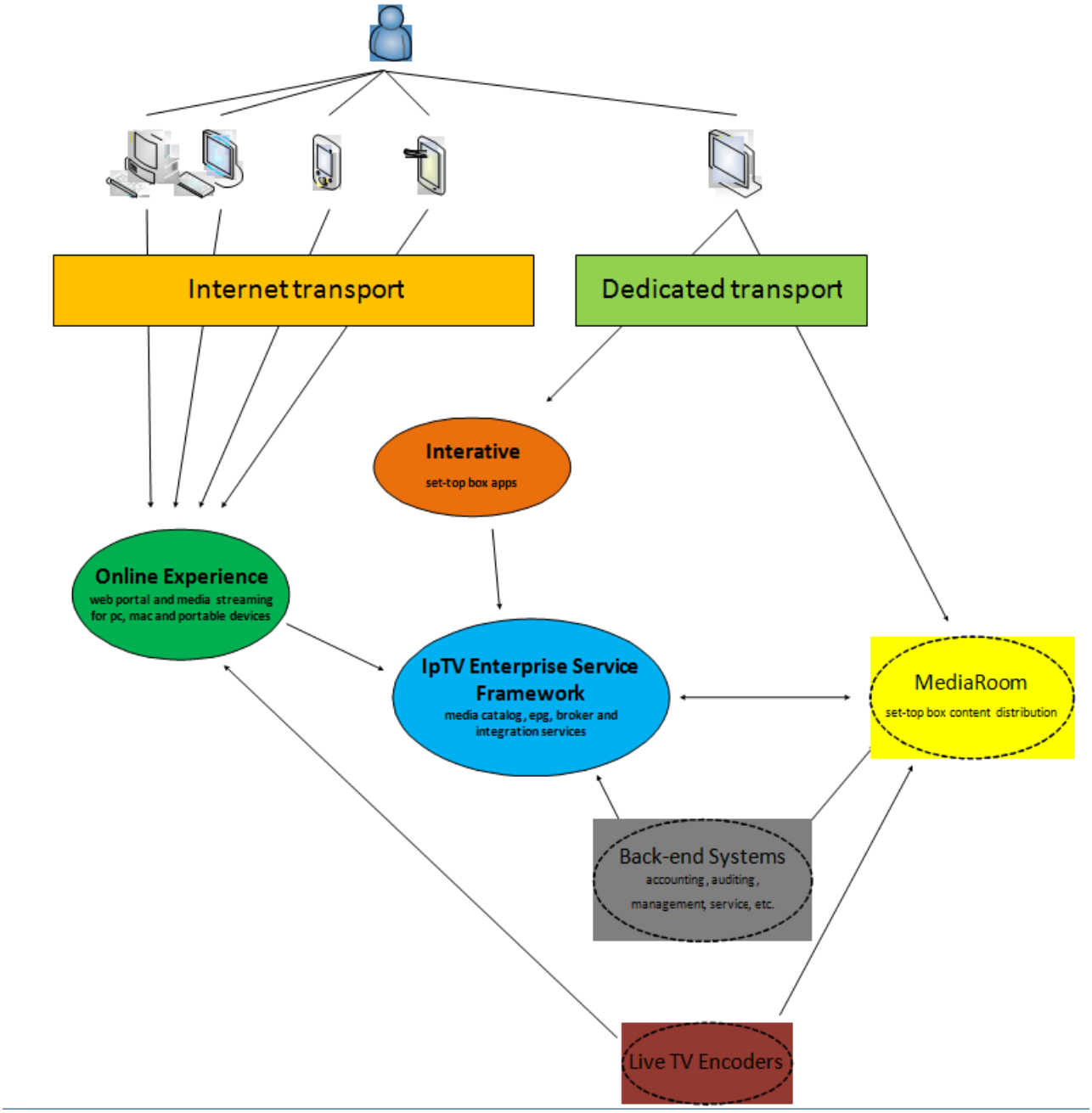


Figure 29 – Telco-OTT Solutions Framework

System Requirements / Characterization

1. What is the major functionality: Mixture
2. Number of Users: up to 100k
3. User location: internet
4. Workload cargo: high
5. Workload Pattern: unpredictable bursting
6. When it's needed: Medium-term
7. Duration (life cycle): +4 years

5.2.1 Technology Provider Requirements

Due to the complexity and dimension of the solution, the technology provider requirements where simplified. The physical infrastructure is able to support up to 100.000 monthly active users and 4800 simultaneous users from the 12 OVX Presentation service Servers (Edge/cache). The Edge/cache servers computing capacity and Network bandwidth would support 500 simultaneous users each, but a 20% reserve is used for Operating System and other services.

Shared Storage

Technology provider requires a total of 100 TB in shared storage, including backup space for one month retention (similar to backup service offered on the Public Clouds).

	Size (usable)	Suggested RAID
SQL Hosts	12	RAID 1
Hyper-V Hosts	20	RAID 5
Short-term Retention Backup	52	RAID 5
Total	**102 TB**	

Table 4 – Shared Storage RAID sizes

Technology, RAID and Disk configurations:
Tier 1: SAS 6Gb 15k rpm, 2,5”, 12xRAID1 24x 1TB disks)
Tier 2: SAS 6Gb 15k rpm, 2,5”, 12xRAID5 (48x 1TB disks)
Tier 2: SAS 6Gb 15k rpm, 2,5”, 18xRAID5 (72x 1TB disks)

Computing

For this scenario we will need 32 physical servers with 32 cores, 256GB RAM and 300 GB local hard disk each.

	# Hosts	vCPUs	RAM
SQL Hosts	12	384	3072
Hyper-V Hosts	20	640	5120
Total		**1024**	**8192**

Table 5 – Total Computing Capacity for 1.5M users

The SQL Server engines will be supported by 12 physical hosts, forming 4 clusters of 3 nodes each. The remaining 20 physical hosts form a hyper-v cluster supporting the 120 VM’s (more info see next table – Servers per role).

	Role	# VMs (or Physical Servers)
Infrastructure Services (17 servers)	AD	2
	SQL	3 (Physical SQL Cluster)
	SMTP	2
	PKI	2
	System Center	5
	Remote Desktop /VDI	2
	Firewall	2
Online Experience (66 Servers)	**OVX Presentation Services**	**12**
	DRM License Server	6
	OVX Application Services	18
	OVX License GW Service	4
	VOD Ingest Server (HPC)	6
	VOD Origin Server	2
	Live Ingest	2
	Live Origin	2
	RMS – load Balancer	2
	RMS – App Server	4
	WMS – Origin Server	2
	WMS – Edge Server	4
	OVX database Services	3 (Physical SQL Cluster)
ipTV Enterprise Service Framework (22 Servers)	FrontEnd Services	2
	BizServices	5
	Distributed Cache	5
	Messaging and Integration Services	4
	Database Services	3 (Physical SQL Cluster)
Interativo (29 Servers)	FrontEnd Services	20
	Image Server	2
	Database Services	3 (Physical SQL Cluster)

Table 6 – Number of Servers per Roles

Physical Server Profile	Processor: 2x16 Cores @ 2.3GHz Memory: 256 GB RAM @ 1333Mhz NIC: 8 x @ 1 Gbit Other: 2 x Fiber Channel SAN Connector (HBA) @ 16 Gbit Local Disks: 2 x 300 GB SCSI – RAID 1 Logical Volumes: • C (100 GB RAID1 – OS); • D (Available free space – Backups and other operations).

Table 7 – Physical Server Profile

Virtual Machine Profile	4vCPU; 32GB RAM; 2x 125GB VHD

Table 8 – Virtual Machine Profile

By measure, the presented Virtual Machine profile supports 400 simultaneous streams, plus some gap for operating system and other apps. Each stream consumes 50MB of RAM and 250MB on the cache disk (the second 125GB disk), multiplied by 400 (its peak), the system will consume 20.000MB or 20GB and 100.000MB or 100GB of cache disk.

Network (CDN)

CDN detail for informational propose, in both scenarios (on-premises or cloud) CDN is required:

- Content Delivery Network is required, up to 4.800 simultaneous users.
- Based on a 2 Mbits per user stream, a total bandwidth of 12 Gbits is needed.
- Each Edge/cache Server supports up to 400 users
- A total of 12 Edge /Cache Servers are needed (divided in 2 locations, 6 each)

5.2.2 On-premises Energy Costs + Colocation

Based on article “Choosing Between Room, Row, and Rack-based Cooling for Data Centers” [35] on-premises Rack Energy Costs per month based on 80% utilization (is the most accepted percentage): Rack Density: 12kW per rack and Cost of Energy: $0.15/kWh

Month= 24h x 30 days = 720h

Total Month Costs per Rack: 720h x $0.15 = $108 per Rack

Total Year Costs: 108 x 12 = $1296 per rack, $2592 for the 2 racks

Total Energy Costs: $2592 x 4 years = $ 10.368

Based on colocationamerica.com [36] collocation provider offer, costs of Full Rack Colocation (42U) are $ 999 per year, so for 2 Racks and 4 years:
Year= $ 999 x 4 years= $ 3996

Total Colocation Costs: $ 3996 x 2 Racks = $ 7.992

5.2.3 Labor

The labor costs were based on the average wages on the United States in 2015, which is an average annual wages of $ 58.714.[37]
Total direct costs for 2 IT Professionals in 4 years:

Total Labor Costs: $ 58.714 x 2 IT Pros = $ 117.428 x 4 years = $ 469.712

5.2.4 On-premises Internet Access

Offering large debts presupposes a dedicated access between the customer premises and the provider network and then a contracted debt. In this scenario contention rates do not apply because they run right into the core of the operator, then (for the Internet) lead with the best effort Internet policies.

The price of dedicated access FE (Fast Ethernet) or GE (Gigabit Ethernet) varies with the customer's location. Admitting that the location is a Great Urban Center will have the following average reference prices:

Description	Price per Month	Price per Year	Total (4 years)
FE Access + 100Mbps Internet Connectivity	$ 350	$ 4.200	$ 16.800
GE Access + 1Gbps Internet Connectivity	**$ 800**	**$ 9.600**	**$ 38.400**
GE Access + 10Gbps Internet Connectivity	$ 1.100	$ 132.000	$ 528.000

Table 9 – on-premises Internet Access Costs

5.2.5 On-premises Computing Capacity costs

The on-premises IaaS scenario chosen to support the required multiple solutions/Applications is totally redundant, multi-vendor and the more common characteristics where chosen to create a private cloud. Consider next table for General description and costs:

Description	Unitary Price	Quantity	Total
Rack cabinet 42U height	$ 2.000,00	2	$ 4.000,00
Blade Enclosure with 16 Blades	$ 225.000,00	2	$ 450.000,00
SAN Storage with 50TB	$ 50.000,00	2	$ 100.000,00
Datacenter LAN Switches 10Gbits	$ 30.000,00	2	$ 60.000,00
Datacenter SAN Switches 8GBits	$ 5.000,00	2	$ 10.000,00
Backup System with 500TB LTO6	$ 50.000,00	1	$ 50.000,00
Virtualization Software (per CPU)	$ 2.000,00	64	$ 128.000,00
			$ 802.000,00

Table 10 – on-premises IaaS infrastructure

Server details

CPU	RAM	HDD
2x E5-2698V3 2.3GHz 16-core	256GB DDR4	2x 300GB SCSI HDD (RAID1)

Table 11 – Blade Server details

Computing capacity was calculated from the sum of the 32 blades most common CPU used in 2015, more suitable for server virtualization and cloud infrastructure. Choosing Blade Systems reduces rack space and energy consumption.

On-premises total computing capacity (from the 32 Blades):

Cores	RAM	Local usable Storage
768	8.192GB	9,6TB

Table 12 – on-premises total computing capacity

Storage details

Storage technology, Tier 2: SAS 6Gb 15k rpm, 2,5", RAID5 was chosen because it's still the most common for the general production environments.

On-premises total dedicated storage capacity (from the 2 Storage Systems):

SAN Storage
100TB (Tier 2)

Table 13 – on-premises total SAN storage capacity

5.2.6 Closest AWS Amazon Solution

Computing Capacity

Based on the on-premises configuration, we specify on the AWS TCO Calculator [38] based on the Physical server capacity. Only a little adjustment was made, because the maximum option of cores was 8 cores, so instead of 2 CPUs with 16 cores each like on-premises, it was chosen 4 CPUs of 8 cores each. This way we have the same 32 cores per server.

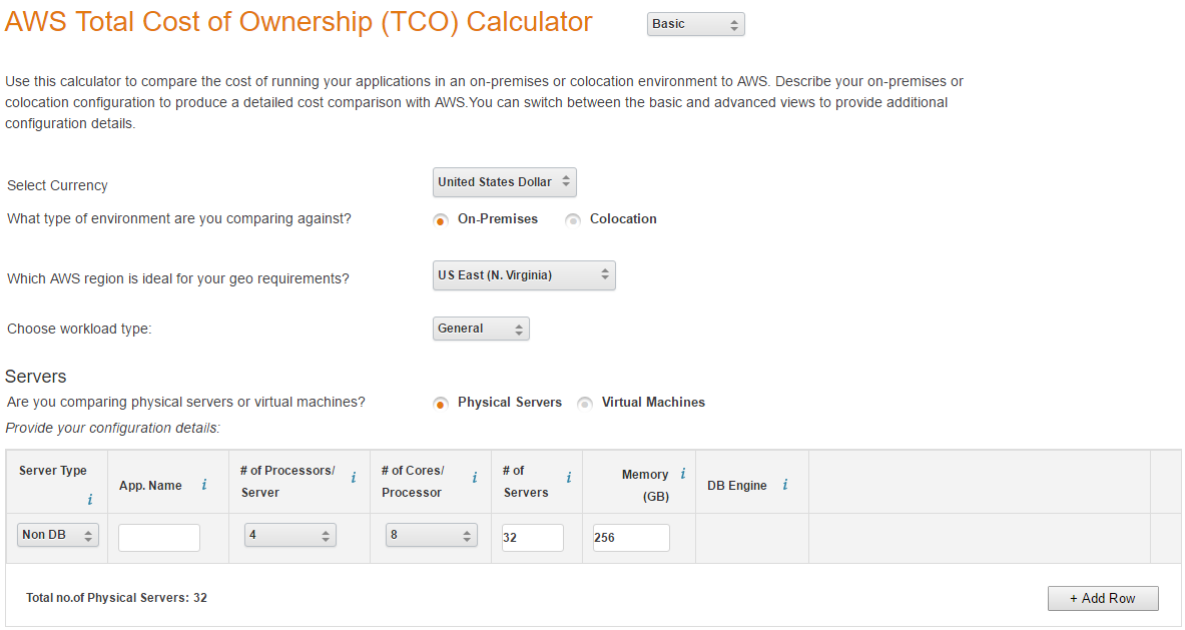


Figure 30 – AWS TCO Calculator (Servers) [38]

CPU	RAM	HDD
32-core	244GB	0GB

Table 14 – Instance details (r3.8xlarge)

AWS Amazon total computing capacity (from 32 instances):

Cores	RAM	Local usable Storage
1024	**7808GB**	**0TB**

Table 15 – AWS Amazon total computing capacity

Shared Storage

When choosing Storage we have three options and the following explanation:

"Type of on-premises storage used- storage area network (SAN), which provides block level data storage, network-attached storage (NAS), which provides file level data storage, and object storage, which manages data as objects. On-premises SAN and NAS systems are mapped to Amazon EBS, while object storage is mapped to Amazon S3."

We have no information in what kind of technology/Tier is supporting our storage capacity or the share ratio of that physical storage.

Figure 31 – AWS TCO Calculator (Storage)[38]

SAN Storage
100TB

Table 16 – AWS Amazon total SAN storage capacity

Closest Computing Capacity on AWS Amazon		
4 Yr. Total Cost of Ownership		
Server	$	1.186.013,00
Storage	$	131.800,00
Total	**$**	**1.317.813,00**

Table 17 - Closest Computing Capacity Costs on AWS Amazon

Network Consumption

Data Transfer IN To Amazon EC2 From	per GB	Total per Month	Total (4 Years)
Internet	$ 0	$ 0	$ 0
Another AWS Region (from any AWS Service)	$ 0	$ 0	$ 0
Amazon S3, Amazon Glacier, Amazon DynamoDB, Amazon SES, Amazon SQS, or Amazon SimpleDB in the same AWS Region	$ 0	$ 0	$ 0
Amazon EC2, Amazon RDS, Amazon Redshift and Amazon ElastiCache instances or Elastic Network Interfaces in the same Availability Zone			
Using a private IP address	$ 0	$ 0	$ 0
Using a public or Elastic IP address	$ 0.01	$ 100*	$ 4.800
Amazon EC2, Amazon RDS, Amazon Redshift and Amazon ElastiCache instances or Elastic Network Interfaces in another Availability Zone or peered VPC in the same AWS Region	$ 0.01	$ 100*	$ 4.800

Table 18 – AWS Amazon Data Transfer Costs IN

Data Transfer OUT	per GB	Total per Month*	Total (4 years)
First 1 GB / month	$0.00	$ 0	$ 0
Up to 10 TB / month	**$0.09**	**$ 900**	**$ 43.200**
Next 40 TB / month	$0.085	$ 3.400	$ 163.200
Next 100 TB / month	$0.07	$ 7.000	$ 336.000
Next 350 TB / month	$0.05	$ 17.500	$ 840.000

Table 19 – Cost of Data Transfer OUT from Amazon to Internet

*Assuming 10TB /month scenario.

5.2.7 Closest Microsoft Azure Solution

In the Microsoft Azure Pricing Calculator[39] to get close to similar on-premises configuration, the total of instances was doubled (from 32 to 64), because the largest instance (A11) only has 112GB of RAM, very far from the 256GB of RAM needed.

Computing Capacity

CPU	RAM	HDD
16-core	112GB	384GB

Table 20 – Instance details (A11)

Microsoft Azure total computing capacity (from 64 instances):

Cores	RAM	Local usable Storage
1024	7168	24,5TB

Table 21 – Microsoft Azure total computing capacity

Shared Storage

SAN Storage
100TB

Table 22 – Microsoft Azure Block Blob storage capacity

Closest Computing Capacity on Microsoft Azure

Estimated monthly cost converted to 4 Yr.			
	Azure	**Month**	**Total**
Server	$ 111.707,14	48	$ 5.361.942,66
Storage	$ 2.396,57	48	$ 115.035,36
Total			**$ 5.476.978,02**

Table 23 - Closest Computing Capacity Costs on Microsoft Azure

Network Consumption

Data Transfer OUT	Total per Month	Total (4 years)
10 TB / month	**$ 890,45**	**$ 42.741,6**

Table 24 – Cost of Data Transfer OUT from Azure to Internet

5.2.8 Closest Google Cloud Platform

In the Google Cloud Platform Pricing Calculator [40] to get close to similar on-premises configuration, the largest instance (n1-highmem-32) only has 208GB of RAM, choosing other instance to double instances like used on Microsoft Azure scenario increases costs extremely, so assumed the lost of 1536GB of RAM compared to the on-premises capacity.

Computing Capacity

CPU	RAM	HDD
32-core	208GB	375GB

Table 25 – Instance details (n1-highmem-32)

Google Cloud Platform total computing capacity (from 32 instances):

Cores	RAM	Local usable Storage
1024	6656	12TB

Table 26 – Google Cloud Platform total computing capacity

Shared Storage

Google Cloud Platform Standard storage total capacity:

SAN Storage
100TB

Table 27 – Google Cloud Platform Standard storage capacity

Closest Computing Capacity Costs on Google Cloud Platform

Estimated monthly cost converted to 4 Yr.			
	Google	Month	Total
Server	$ 35.593,63	48	$ 1.708.494,24
Storage	$ 2.662,40	48	$ 127.795,20
Total			**$ 1.836.289,44**

Table 28 - Closest Computing Capacity Costs on Google Cloud Platform

Network Consumption

Data Transfer OUT	Total per Month	Total (4 years)
10 TB / month	**$ 1136,64**	**$ 54.558,72**

Table 29 – Cost of Data Transfer OUT from GCP to Internet

5.2.9 Resume Comparison

Computing

	On-premises	Amazon	Microsoft	Google
vCPUs	1024	1024	1024	1024
RAM in GB	8192	7808	7168	6656
Local Disk in TB	9,6	0	24,5	12
Shared Disk in TB	100	100	100	100

Table 30 – Computing resume comparison

Total Costs

4 Years	On-Premises	Amazon	Microsoft	Google
Energy	10.368	0	0	0
Colocation	7.992	0	0	0
Labor	469.712	0	0	0
Computing	802.000	1.317.813,00	5.476.978,02	1.836.289,44
Network	38.400	43.200	42.741,6	54.558,72
Total	**$ 1.328.472**	**$ 1.361.013**	**$ 5.519.719,62**	**$1.890.848,16**

Table 31 – Total Costs comparison

5.2.10 Financial Analysis

On the Cloud Providers all expenditures are OPEX. On the on-premises option the costs are divided in CAPEX and OPEX, which require a large upfront investment (CAPEX):

CAPEX	Quantity	unitary price (year/equipment)	Total
Rack cabinet 42U height	2	$ 2.000,00	$ 4.000,00
Blade Enclosure with 16 Blades	2	$ 225.000,00	$ 450.000,00
SAN Storage with 50TB	2	$ 50.000,00	$ 100.000,00
Datacenter LAN Switches 10Gbits	2	$ 30.000,00	$ 60.000,00
Datacenter SAN Switches 8GBits	2	$ 5.000,00	$ 10.000,00
Backup System with 500TB LTO6	1	$ 50.000,00	$ 50.000,00
Virtualization Software (per CPU)	64	$ 2.000,00	$ 128.000,00
			$ 802.000,00

Table 32 – on-premises Telco-OTT CAPEX costs

And the following operational expenditures, OPEX:

OPEX	Quantity		unitary price (year/equipment)		4 Years
Energy Costs		$	2.592,00	$	10.368,00
Colocation		$	1.998,00	$	7.992,00
Internet GE		$	9.600,00	$	38.400,00
Labor (2 IT Pro.)	2	$	117.428,00	$	469.712,00
				$	**526.472,00**

Table 33 – on-premises Telco-OTT OPEX costs

5.3 E-mail Business Case

E-mail Definition

Electronic mail is a method of exchanging digital messages between users, started with computer devices, today used in almost every connected device.

Messages are exchanged between hosts using the Simple Mail Transfer Protocol with software programs called mail transfer agents (MTAs); and delivered to a mail store by programs called mail delivery agents (MDAs, also sometimes called local delivery agents, LDAs). Accepting a message obliges an MTA to deliver it, and when a message cannot be delivered, that MTA must send a bounce message back to the sender, indicating the problem.

Users can retrieve their messages from servers using standard protocols such as POP or IMAP, or, as is more likely in a large corporate environment, with a proprietary protocol specific to Novell Groupwise, Lotus Notes or Microsoft Exchange Servers. Programs used by users for retrieving, reading, and managing email are called mail user agents (MUAs).

Mail can be stored on the client, on the server side, or in both places. Standard formats for mailboxes include Maildir and mbox. Several prominent email clients use their own proprietary format and require conversion software to transfer email between them. Server-side storage is often in a proprietary format but since access is through a standard protocol such as IMAP, moving email from one server to another can be done with any MUA supporting the protocol.

Many current email users do not run MTA, MDA or MUA programs themselves, but use a web-based email platform, such as Gmail, Hotmail, or Yahoo! Mail, that performs the same tasks. Such webmail interfaces allow users to access their mail with any standard web browser, from any computer, rather than relying on an email client [41].

Major features

1. E-mail Service (Send/Receive Mail)
2. Task Management
3. Calendar Application (Agenda)
4. Contacts Manager (Address / Contacts List)
5. Distribuition Lists
6. Auto Reply (Out-of-office and other client Rules)

High Level Architecture

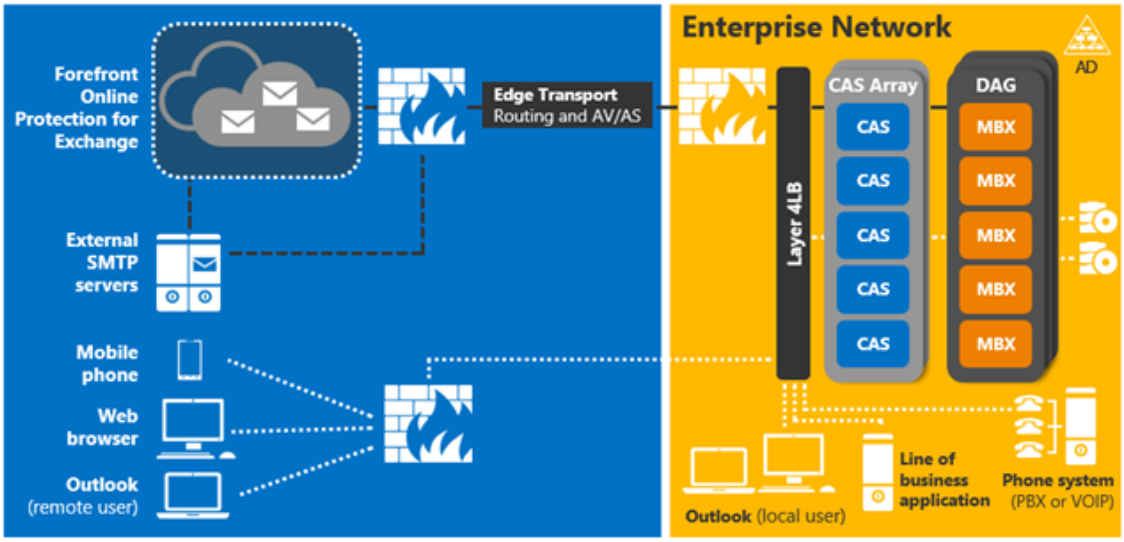


Figure 32 – Exchange e-mail system architecture [42]

System Requirements / Characterization

1. What is the major functionality: Storage
2. Number of Users: up to 10k
3. User location: Mixture
4. Workload cargo: medium
5. Workload Pattern: on and off
6. When it's needed: Short-term
7. Duration (life cycle): +4 years

5.3.1 Technology Provider Requirements

Storage

In order to provide up to 50GB mailbox per user to a total of 5k users, a total on-premises Storage of 250TB Tier 2 is required. 50GB x 5.000 users = 250.000GB

Work hours average up to 4.000 simultaneous users (80%), however peaks reaching the 5.000 (100) are possible and acceptable.

Computing

According to the technology provider, this topic represents on-premises total computing capacity requirements for 5K users:

Online Experience	Server Profile	# Servers
Mailbox Servers (MBX)	2 CPU; 32GB RAM	2
CAS Servers (CAS)	2 CPU; 32GB RAM	2
Edge Transport	2 CPU; 32GB RAM	2

Table 34 – E-mail Server Profile and # Servers

Network

Assuming a mixture user location, but assuming that users are mostly active during work-hours and on-premises, a commercial 100mbits internet connection is enough.

5.3.2 On-premises Energy Costs + colocation

Based on the article “Choosing Between Room, Row, and Rack-based Cooling for Data Centers”[35] on-premises Rack Energy Costs per month based on 80% utilization (is the most accepted percentage): Rack Density: 12kW per rack and Cost of Energy: $0.15/kWh

Month= 24h x 30 days = 720h
Total Month Costs per Rack: 720h x $0.15 = $108 per Rack
Total Year Costs: $ 108 x 12 = $1296 per rack
Total Costs: $ 1296 x 4 years = $ 5184

Based on colocationamerica.com [36] collocation offer, cost of Full Rack Colocation (42U) are $ 999 per year, so for 2 Racks and 4 years:
Year= 999 x 4 = $ 3996

5.3.3 Labor

The labor costs were based on the average wages on the United States in 2015, which is an average annual wages of $ 58.714[37].
Total direct costs for 2 IT Professionals in 4 years:

Total Labor Costs: $ 58.714 x 2 IT Pros = $ 117.428 x 4 years = $ 469.712

5.3.4 On-premises Computing Capacity cost

The on-premises scenario chosen to support the technology provider Applications is totally redundant. Consider table 1 for General description and costs:

Description	Unitary Price	Quantity	Total
Rack cabinet 42U height	$ 2.000,00	1	$ 2.000,00
Rackmount Server	$ 10.000,00	6	$ 60.000,00
SAN Storage with 125TB (Tier 2)	$ 100.000,00	2	$ 200.000,00
Datacenter LAN Switches 10Gbits	$ 30.000,00	2	$ 60.000,00
Datacenter SAN Switches 8GBits	$ 5.000,00	2	$ 10.000,00
Backup System with 500TB LTO6	$ 50.000,00	1	$ 50.000,00
Windows Server Software	$ 5.000,00	6	$ 30.000,00
Exchange Server Software	$ 5.000,00	6	$ 30.000,00
			$ 442.000,00

Table 35 – E-mail on-premises infrastructure

Server details

CPU	RAM	HDD
2x E5-2698V3, 2,3 GHz 16-core	32GB DDR4	2x 600GB SCSI HDD (RAID1)

Table 36 – Rackmount Server details

Computing capacity was calculated from the sum of the 6 server most common CPU used in 2016. On-premises total computing capacity (from the 6 rackmount servers):

Cores	RAM	Local usable Storage
144	192GB	3,6TB

Table 37 – on-premises total computing capacity

Storage details

Storage technology, Tier 2: SAS 6Gb 15k rpm, 2,5", RAID5 was chosen because it's still the most common for the general production environments. On-premises total dedicated storage capacity (from the 2 Storage Systems):

SAN Storage
250TB (Tier 2)

Table 38 – on-premises total SAN storage capacity

5.3.5 Amazon WorkMail e Amazon WorkDocs (SaaS)

For a SaaS scenario we don't need to be concerned about computing capacity or Network Consumptions. Typically, we pay for a service per user, this analysis was made based on public prices available, even so, we know for large companies the Providers make significant discounts.

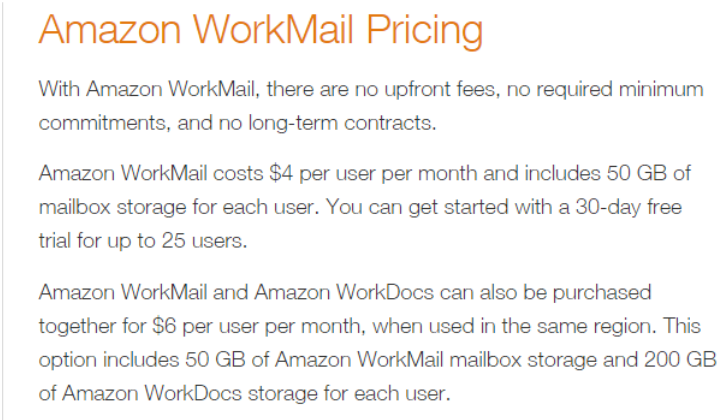


Figure 33 – Amazon WorkMail Pricing[43]

	Month	Year / 5k users	4 Years
Mailbox Size	50GB	-	-
SharedDocuments Size	200GB	-	-
Backup retention	30 days	-	-
Price	**$ 6**	**$360.000**	**$1.440.000**

Table 39 – Amazon WorkMail Solution

Information according to Amazon AWS WorkMail[43].

5.3.6 Office 365 Essentials (SaaS)

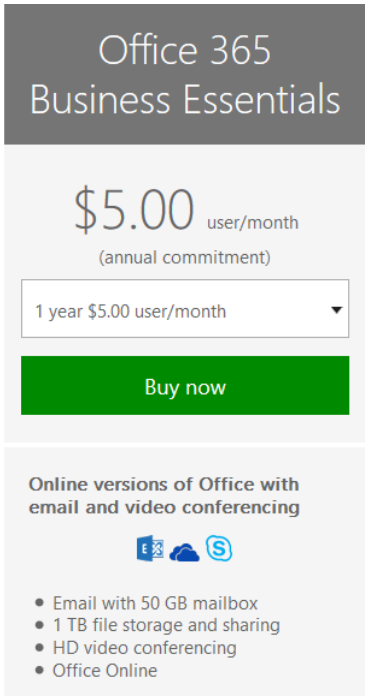


Figure 34 – Office 365 Business Essentials[44]

	Month	Year / 5k users	4 Years
Mailbox Size	50GB	-	-
SharedDocuments Size	1000GB	-	-
Backup retention	30 days	-	-
Price	**$ 5**	**$300.000**	**$1.200.000**

Table 40 – Office 365 solution

Information according to Microsoft Office 365 website[44].

5.3.7 G Suite Basic (SaaS)

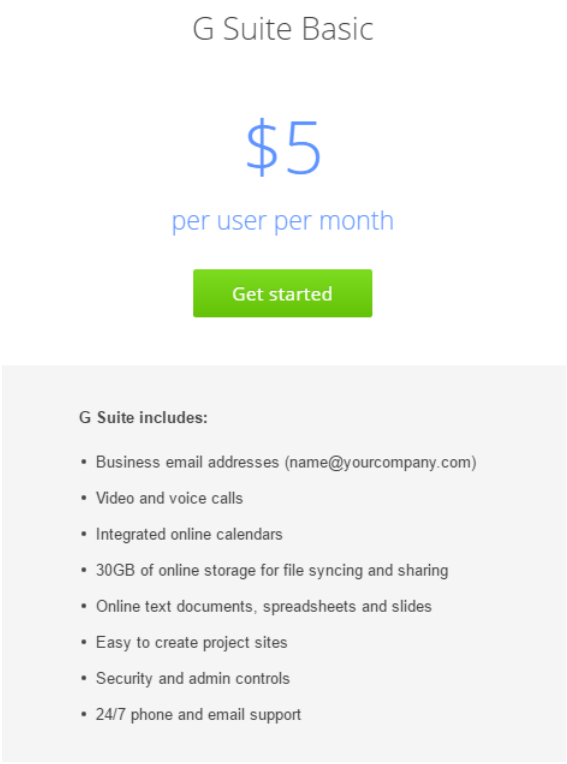


Figure 35 – G Suite Basic[45]

	Month	Year	4 Years
Mailbox Size	30GB	-	-
SharedDocuments Size	0GB (shared with mailbox)	-	-
Backup retention	30 days	-	-
Price	**$ 5**	**$300.000**	**$1.200.000**

Table 41 – G Suite Basic Solution

Information according to Google G Suite website[45].

5.3.8 Resume Comparison

	On-premises	Amazon	Microsoft	Google
Mailbox Size	50GB	50GB	50GB	30GB
SharedDocuments Size	0GB	200GB	1000GB	0GB
Backup retention	30 days	30 days	30 days	30 days
Price	**$ 442.000**	**$ 1.440.000**	**$ 1.200.000**	**$ 1.200.000**
Energy+Colocation	**$ 9.180**	**0**	**0**	**0**
Labor	**$ 469.712**	**0**	**0**	**0**
Total	**$ 920.892**	**$ 1.440.000**	**$ 1.200.000**	**$ 1.200.000**

Table 42 – E-mail solutions Resume Comparison

5.3.9 Financial Analysis

On the Cloud Providers option all expenditures are OPEX. On the on-premises option the costs are divided in CAPEX and OPEX, which require a large upfront investment (CAPEX):

CAPEX	Quantity	unitary price (year/equipment)	Total
Rack cabinet 42U height	1	$ 2.000,00	$ 2.000,00
Rackmount Server	6	$ 10.000,00	$ 60.000,00
SAN Storage with 125TB (Tier 2)	2	$ 100.000,00	$ 200.000,00
Datacenter LAN Switches 10Gbits	2	$ 30.000,00	$ 60.000,00
Datacenter SAN Switches 8GBits	2	$ 5.000,00	$ 10.000,00
Backup System with 500TB LTO6	1	$ 50.000,00	$ 50.000,00
Windows Server Software	6	$ 5.000,00	$ 30.000,00
Exchange Server Software	6	$ 5.000,00	$ 30.000,00
			$ 442.000,00

Table 43 – on-premises e-mail CAPEX costs

And the following operational expenditures, OPEX:

OPEX	Quantity	unitary price (year/equipment)	4 Years
Energy Costs		$ 1.296,00	$ 5.184,00
Colocation		$ 999,00	$ 3.996,00
Labor (2 IT Pro.)	2	$ 117.428,00	$ 469.712,00
			$ 478.892,00

Table 44 – on-premises e-mail OPEX costs

6. Conclusions and Future Work

Several conclusions may be achieved from this dissertation, the majority I will try to resume in this conclusion. However, every scenario must be carefully studied, and multiple requirements must be contemplated.

Computing Capacity it's cheaper upfront on the Cloud, but on the long run (+4 years) has the years past they will cost more compared to on-premises and On-Premises energy costs have little impact on total costs.

Network requirements and transactions are very important to help make your decision, if your information system users are on-premises; the Cloud is slower and costs more. If your Information System requires intensive uploads and downloads, we need to remember this consumption is charged by cloud providers. It may be the key factor for the best choice.

Public Cloud are typically a better solution for small a medium business, no physical infrastructure and specialized staff to manage that low level infra (Building, controls, energy, hardware, etc.). Public Cloud is sometimes better for some or parts of an Information System when your clients are on the internet or spread around /World.

As public cloud cost trend to become cheaper, I believe in a near future will be the first solution for many scenarios. By trends of energy and computing efficiency, the next challenge even for cloud providers is to predict computing capacity. To achieve this, average workload and growth must be understandable.

Understanding Information Systems Workloads and Operational Laws are the most important fraction for performance, required capacity and predicted growth. Application providers tend to specify hardware and software requirements for their systems, but calculating computing capacity and network requirements for hundreds or millions of users can be difficult. Workloads and Operational Laws are very useful for these decisions.

Depending on the Information System the best choice may be on-premises or Public Cloud, there is no "one size fits all". To the majority of organization the future is going to be Hybrid Clouds, adopting the best of both worlds, and putting the best suitable Information Systems on the Public Cloud and maintaining the others on-premises.

Planning an on-premises infrastructure solution has different concerns then a Cloud solution. Typically on-premises infrastructure is planned on the medium/long term (+4 years), concerning most of the times upfront the potential growth for that time period, because of this the capacity is usually over dimensioned for the first years.

Conversely, Cloud infrastructure provides elasticity and resources can quickly upscale or downscale as service requires. Cloud is based on the concept of shared resources, and because of this sometimes limiting the control and customization of the services, although some other times offering a complex service has a simple (shared) service.

Managers and IT Professionals usually tend to calculate only hardware price when comparing cloud cost with on-premise infrastructure. However, total cost of ownership (TCO) is much bigger than just a hardware price. Like showed on this work, we have to take in account not only hardware, software, energy and network, but also Hardware and Software support contracts, staff salary, training, operations, risk management and other things that cloud provider does for us (e.g.: SLA vs OLA financial impact).

Cloud Major Benefits

1. Public cloud tend to use newer technology (hardware and software).
2. Shift from capital expenses to operational expenses (pay as you go).
3. Public Clouds have greater economies of scale.
4. Elasticity, scale as needed and pay per use.
5. Reduced Implementation Time.
6. Lower maintenance Costs (hardware and software updates, maintenance or repairs).
7. Better Monitoring, Metrics and financial control.
8. Resiliency and Redundancy.
9. Better staff competence.
10. For Small businesses and Startup's (Easy to start up, easy to end).

Cloud Major Limitations

1. Applications readiness: not every app was design with cloud in mind.
2. Compliance: regulated industries like financial, government and healthcare organizations.
3. Customization: most financial organizations customize a product or platform to suit their unique requirements.
4. Security/Privacy: data is sensitive and a data breach would present a significant risk to the organization, example business-critical Intellectual Property.
5. Cost/ROI: migration challenges and economic issues make it better to leave certain apps on-premises for the biggest ROI. e.g.: if it costs 50k to provide 10k of value app.
6. Organization Capability: Is the organization cloud-ready? Workers skills and Organization processes.
7. Vendor Lock: Moving from one public cloud to another or back on-premises can represent huge challenges.
8. Large Data: with a pay per use model, large data processing and transfer are not cost-effective and performance ready to public cloud.
9. Control: with public cloud you don't have total control of IT infrastructure
10. Performance: technology is the same, but on-premises we can choose, dedicate and control what solutions or application run on what technology, giving this way more certain performance.

Future Work

Several quantitative/qualitative metrics and laws where investigated during this dissertation, only some of them where applied and to part (Edge/cache servers) of the business cases. It would be interesting to apply the others metrics and laws presented and also to investigate other ones. Furthermore, apply this metrics and laws to a larger scope then only part of the business cases.

It would also be interesting to have more financial data (e.g.: revenue) in order to calculate the other financial metrics.

Even on Cloud Providers like the ones studied on this dissertation the resources (financial, human, etc.) are not infinite as some users would think. Managers and Stake Holders ask for efficiency in all resources utilization. I believe the capacity of these organization professionals to understand Workloads and Operational Laws will be crucial for their success.

7. References

[1] J. K. Ponder, “https://www.itu.int,” April 2006. [Online]. Available: https://www.itu.int/.../2006/ponder-ngn-access-10-april-2006.ppt. [Acedido em 08 10 2016].

[2] I. E. MARKETING, “IDG Enterprise Cloud Computing Study 2014,” 2014. [Online]. Available: http://www.idgenterprise.com/resource/research/idg-enterprise-cloud-computing-study-2014/.

[3] L. Columbus, Forbes, 24 01 2015. [Online]. Available: http://www.forbes.com/sites/louiscolumbus/2015/01/24/roundup-of-cloud-computing-forecasts-and-market-estimates-2015/#7303d525740c. [Acedido em 08 10 2016].

[4] E. Gorelik, “Cloud Computing Models,” 2013. [Online]. Available: http://web.mit.edu/smadnick/www/wp/2013-01.pdf.

[5] A. N. S. I. (. /. T. I. A. (TIA), ANSI/TIA-942-A Telecommunications Infrastructure Standard for Data Centers, 2005.

[6] wikipedia, “Multitier Architecture,” 2016. [Online]. Available: https://en.wikipedia.org/wiki/Multitier_architecture.

[7] C. Poelker, “The state of storage in 2014,” computerworld, 1 12 2014. [Online]. Available: http://www.computerworld.com/article/2851912/storage-state-of-storage-in-2014.html. [Acedido em 17 11 2016].

[8] P. Goodwin, “Tiered data storage: State of the art,” TechTarget, 12 2012. [Online]. Available: http://searchstorage.techtarget.com/feature/Tiered-data-storage-State-of-the-art. [Acedido em 17 11 2016].

[9] C. McLellan, “The 21st Century Data Center: An overview,” ZDNET, 04 2013. [Online]. Available: http://www.zdnet.com/article/the-21st-century-data-center-an-overview/. [Acedido em 11 2016].

[10] A. Manuel de Oliveira Duarte, “Planning and Design of Information Systems - a Possible Roadmap,” University of Aveiro, Aveiro, 2015.

[11] V. Haydin, “Cloud Computing: Myths, Fears and Facts,” ELEKS LABS, 28 12 2012. [Online]. Available: http://elekslabs.com/2012/12/cloud-computing-myths-fears-and-facts.html. [Acedido em 13 10 2016].

[12] R. D. Woolley, “Web Performance Measurement & Capacity Planning: Briefing Paper,” Chief Information Officer’s Section, Salt Lake City, Utah 84114, January 2000.

[13] D. A. Menascé, A. V. Almeida and W. L. Dowdy, Performance by Design: Computer Capacity Planning by Example, Prentice Hall PTR, 2004.

[14] R. Campos, “Capacity Planning for Linux Systems,” 12 2011. [Online]. Available: http://pt.slideshare.net/xinu/capacity-planning-for-linux-systes. [Acedido em 11 2016].

[15] o. d. c. alliance, “Standard Units of Measure For IaaS Rev 1.1,” 2013. [Online]. Available: http://www.opendatacenteralliance.org/docs/Standard_Units_of_Measure_For_IaaS_Rev1.1.pdf. [Acedido em 04 02 2015].

[16] B. Latamore, “Five Tier Storage Model,” Wikibon, [Online]. Available: http://wikibon.org/wiki/v/Five_Tier_Storage_Model.

[17] J. R. &. J. Hill, How to Do Capacity Planning, TeamQuest Corporation, 2013.

[18] wikipedia, “Content delivery network,” wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Content_delivery_network. [Acedido em 26 10 2016].

[19] incapsula, “What is a CDN,” incapsula, [Online]. Available: https://www.incapsula.com/cdn-guide/what-is-cdn-how-it-works.html. [Acedido em 11 11 2016].

[20] T. V. WILSON, “How Streaming Video and Audio Work,” Howstuffworks, 12 10 2007. [Online]. Available: http://computer.howstuffworks.com/internet/basics/streaming-video-and-audio3.htm. [Acedido em 12 12 2016].

[21] A. Zambelli, Microsoft, 03 2009. [Online]. Available: https://www.iis.net/learn/media/on-demand-smooth-streaming/smooth-streaming-technical-overview. [Acedido em 11 2016].

[22] TechTarget, “The history of cloud computing and what's coming next: A CIO guide,” TechTarget, [Online]. Available: http://searchcio.techtarget.com/essentialguide/The-history-of-cloud-computing-and-whats-coming-next-A-CIO-guide. [Acedido em 05 11 2016].

[23] ComputerWeekly.com, “A history of cloud computing,” ComputerWeekly.com, [Online]. Available: http://www.computerweekly.com/feature/A-history-of-cloud-computing. [Acedido em 05 11 2016].

[24] “Cloud computing,” Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Cloud_computing. [Acedido em 24 08 2016].

[25] M. Nieuwpoort, Microsoft, 7 2 2011. [Online]. Available: http://www.slideshare.net/sparked/cloud-strategy-cloud-accelerate-workshop.

[26] P. Mell e T. Grance, “The NIST Definition of Cloud,” ADtranz, September 2011. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf. [Acedido em 4 2 2016].

[27] wikipedia, “Colocation Center,” [Online]. Available: https://en.wikipedia.org/wiki/Colocation_centre. [Acedido em 05 11 2016].

[28] RightScale, “http://www.rightscale.com/,” 2015. [Online]. Available: http://keystone-ms.com.au/uploads/Resources/RightScale-2015-State-of-the-Cloud-Report.pdf. [Acedido em 04 02 2015].

[29] CLOUDYN, WHO MOVED MY CLOUD? Part II: What’s behind the cloud vendors AWS, Azure and GCP?, 2015.

[30] wikipedia, “Over-the-top-content,” wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Over-the-top_content. [Acedido em 05 11 2016].

[31] wikipedia, “Telco-OTT,” wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Telco-OTT. [Acedido em 05 11 2016].

[32] S. Rahmanian, “IPTV Network Infrastructure,” 12 2008. [Online]. Available: http://www.cvt-dallas.org/IPTV-Dec08.pdf.

[33] Microsoft, “msdn_about_microsoft_digital_lifestyle,” 06 2007. [Online]. Available: http://download.microsoft.com/download/3/4/e/34ed7e22-516b-41f0-bd49-92cfc037b29d/msdn_about_microsoft_digital_lifestyle.pdf.

[34] G. Yu, T. Westholm, M. Kihl, I. Sedano, A. Aurelius, C. Lagerstedt e P. Ödling, “Analysis and characterization of IPTV user behavior,” 2009. [Online]. Available: http://portal.research.lu.se/portal/en/publications/analysis-and-characterization-of-iptv-user-behavior(8729fec4-ae66-417d-80bf-1a0fa768ec77).html.

[35] K. D. a. N. Rasmussen, “Choosing Between Room, Row, and,” p. 14, 2012.

[36] colocationamerica, “Full Rack Colocation,” colocationamerica, [Online]. Available: http://www.colocationamerica.com/colocation/full-rack.htm. [Acedido em 5 11 2016].

[37] Wikipedia, “Average Wage,” Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/List_of_countries_by_average_wage. [Acedido em 13 11 2016].

[38] Amazon, “AWS Total Cost of Ownership (TCO) Calculator,” Amazon, [Online]. Available: https://awstcocalculator.com/. [Acedido em 16 03 2016].

[39] Microsoft, “Pricing Calculator,” Microsoft, [Online]. Available: https://azure.microsoft.com/en-us/pricing/calculator/. [Acedido em 16 03 2016].

[40] Google, “Google Cloud Platform Pricing Calculator,” Google, [Online]. Available: https://cloud.google.com/products/calculator/. [Acedido em 16 03 2016].

[41] wikipeia, “Email,” [Online]. Available: https://en.wikipedia.org/wiki/Email. [Acedido em 11 2016].

[42] R. S. IV, “Exchange 2013 Server Role Architecture,” Microsoft, 23 01 2013. [Online]. Available: https://blogs.technet.microsoft.com/exchange/2013/01/23/exchange-2013-server-role-architecture/. [Acedido em 02 11 2016].

[43] Amazon, “Amazon Workmail,” Amazon, [Online]. Available: https://aws.amazon.com/workmail/. [Acedido em 26 10 2016].

[44] Microsoft, “Office 365 Business Essentials,” Microsoft, [Online]. Available: https://products.office.com/en-us/business/office-365-business-essentials. [Acedido em 26 10 2016].

[45] Google, “G Suite by Google Cloud,” Google, [Online]. Available: https://gsuite.google.com/intl/en/pricing.html. [Acedido em 26 10 2016].