University of Bologna

Dipartimento di Informatica – Scienza e Ingegneria (DISI)

Engineering Bologna Campus

Class of

Infrastructures for Cloud Computing and Big Data M

Class Starting…

Basics, Objectives, and initial Models

Antonio Corradi

Academic year 2018/2019

CLASS MAIN GOAL

The course aims at delivering a novel vision of systems (mainly **distributed)** and at building a **deep, formal, practical, and meditated experience** of their **operations**

We are immersed into those systems, personally, socially, and as part of organizations

We are interested in a **system viewpoint**, i.e., **what is behind those systems,** and their **behavior** and **impact, both** from the **user perspective** but **more important** with **the point of view** of the **implementers and designers**

In particular we focus on the **experience of operations** rather than in static planning and configuration

we **aim at the entire life cycle operations**

COURSE TARGETS

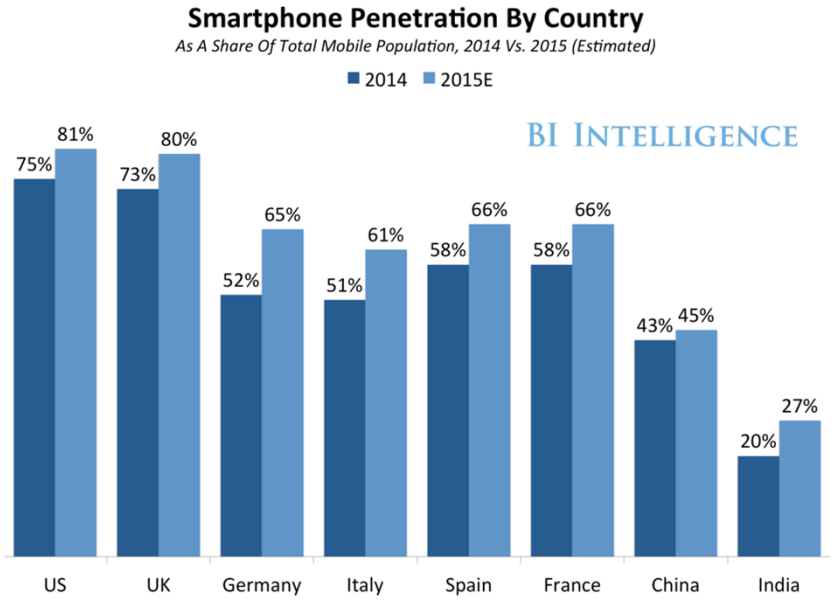
There are many Distributed Systems you use in your everyday experience

Private Personal PC

Private Smartphone

Corporate PC

Corporate Smartphone /Tablet



In Italy, we have a large number of cellular phones, but not so many smartphones, and also a very deep and large usage of them

Also other (Cloud) remote resources are used

COURSE TARGETS

Distributed Systems pervasively available

Within companies / organizations used in work day experience to support any business aspect
But also at private user level

Personal machines and local servers

Internal Electronic Data Processing (EDP) data center

Outsourced resources Cloud

In general, companies have a *conservative attitude* toward ICT resources, but have also consolidated usage of *not on-premises resources*

COURSE TARGETS

Large global corporations to provide **Cloud services (Amazon, Google, IBM, PAs,…)**

- **Organization of internal architecture to provide Cloud services with needed Quality of Service**
- **Cloud Data Center Organization**
- **Interaction with other Data Centers and Cloud**
- **Intra and inter Cloud**

In general, one **Cloud provider** has several local data centers and keep them as a **central bone**, but has to maintain *external available resources* and *extra-organization agreement* for special dedicated situations

CLOUD is a REVOLUTION…

Cloud is a buzzword to be used in advertising and it is sometimes depicted as a revolution

The are many books about Cloud as a revolutionary technology

In general terms, there is not such a *solution of continuity* both under an **organization** and a **technical perspective**

CLOUDS are CHEAPER… and WINNING…

Range in size from "edge" facilities to **megascale**

Scale economies

Approximate **costs for a small size center** (1K servers) and a **larger**, 50K server center

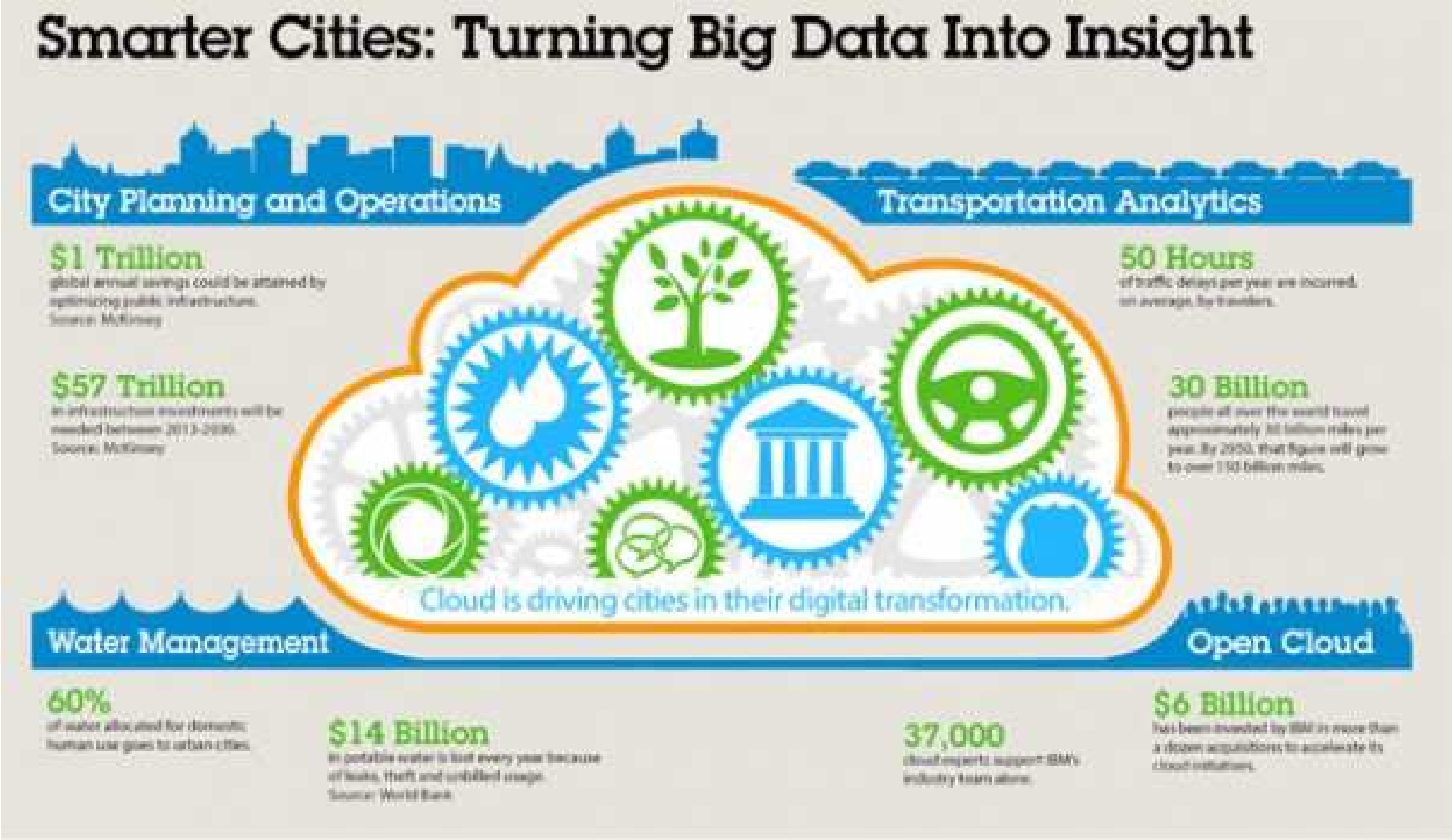
Technology	Cost in small-sized Data Center	Cost in Large Data Center	Cloud Advantage
Network	$95 per Mbps/ month	$13 per Mbps/ month	7.1
Storage	$2.20 per GB/ month	$0.40 per GB/ month	5.7
Administration	~140 servers/ Administrator	>1000 Servers/ Administrator	7.1

Each data center is
11.5 times
the size of a football field

Data from a slide by Roger Barga, Head of Cloud Computing, Microsoft

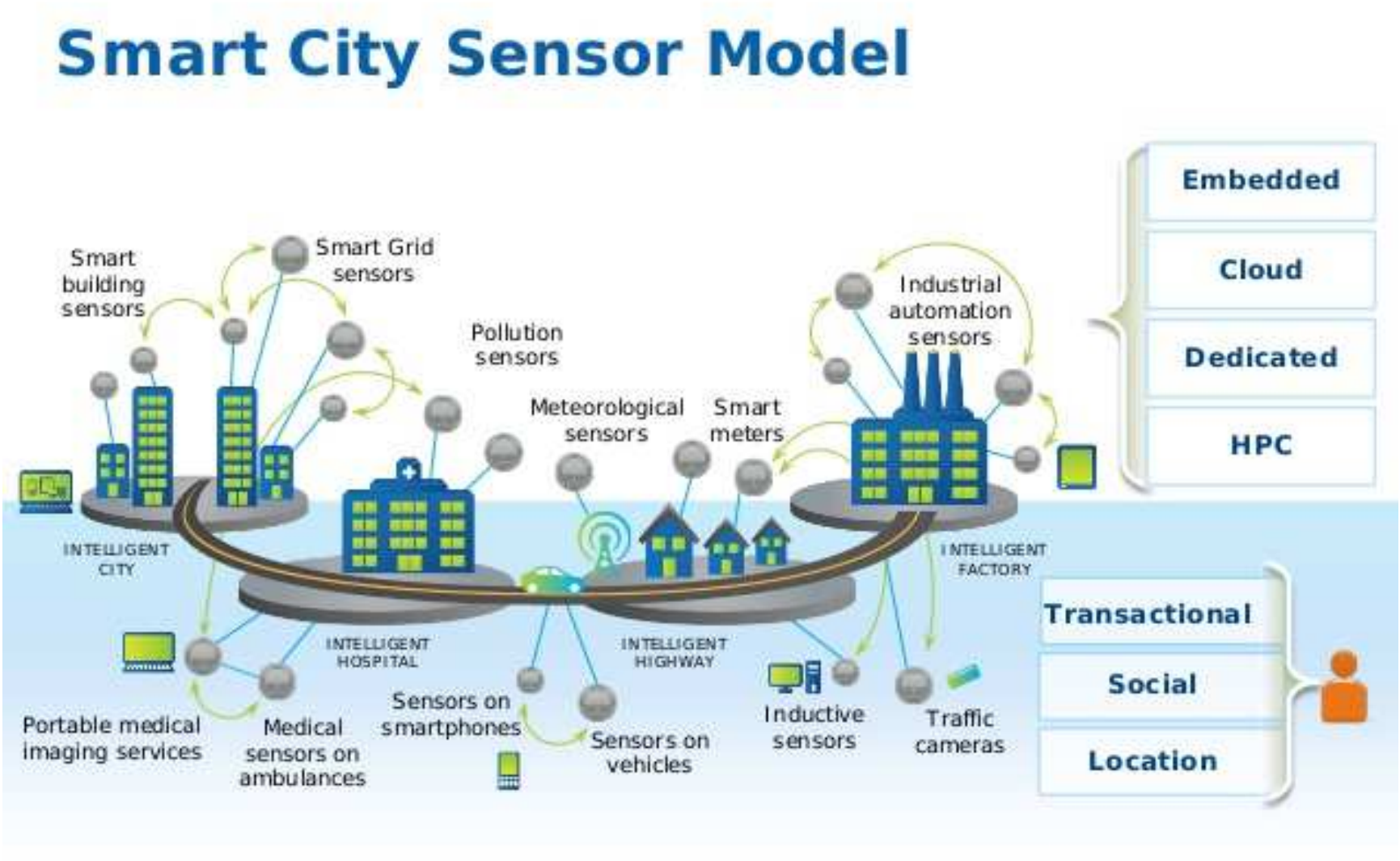
CLOUD AND BIG DATA

Smart cities and different services



SMART CITIES FOR SENSING

Smart cities and sensing data

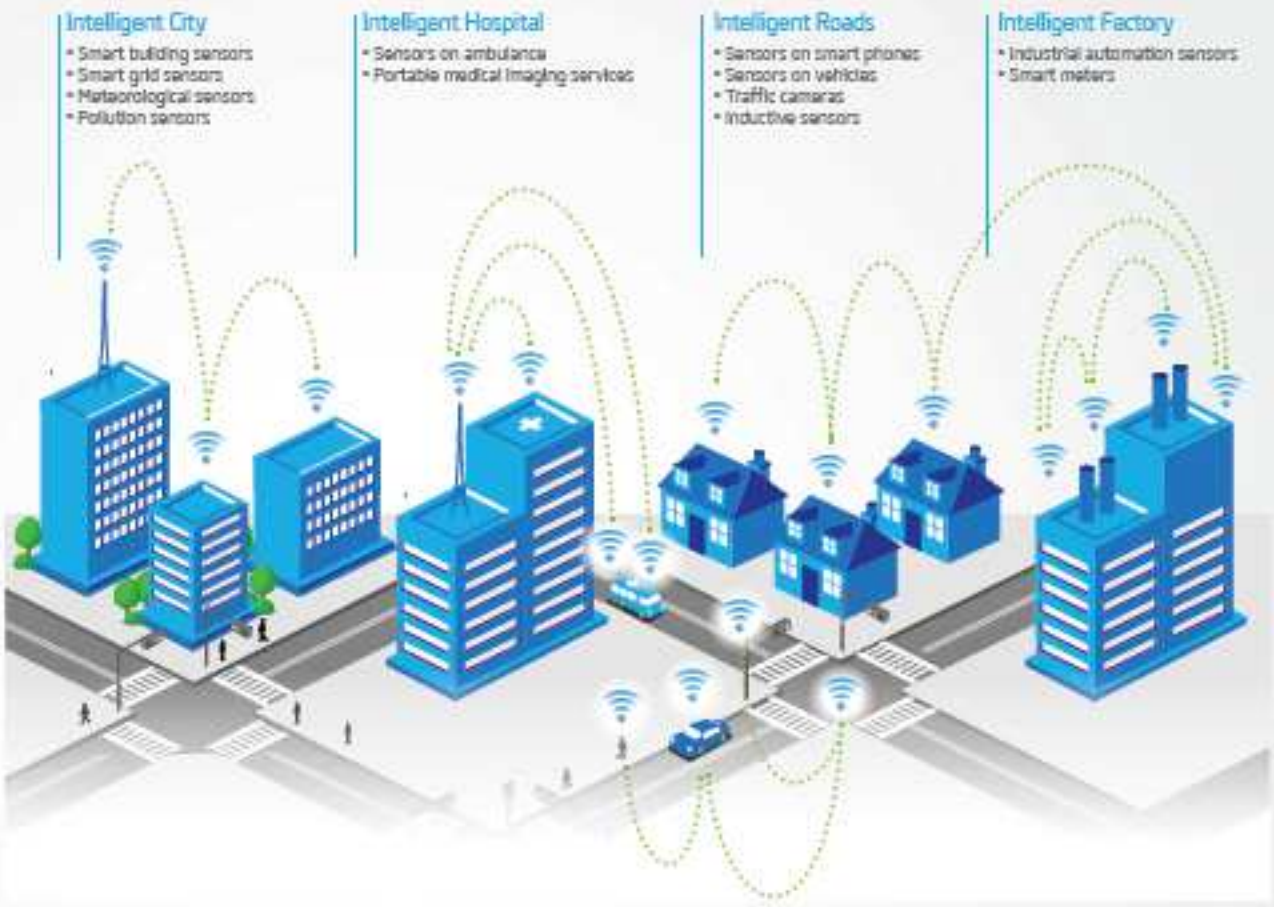


SMART CITIES FOR BIG DATA

Smart cities produce many data of many different kind

Big Data in Context: Smart City Example

In addition to the transactional, social, and location data generated by people, device sensors generate in real time some of the fastest-growing big data. Processing and analytics can be applied to these valuable data sources via provisioned embedded, cloud, or dedicated IT infrastructure and storage and high-performance computing solutions.



BIG DATA EXPECTATIONS

DIGITALIZATION …

Market and big data investments

6.3 billion of USD **2012**
48.3 billion of USD **2018**
expected 45% per year
not only public investments but also private ones

ICT industry market in 2020

5 trillion of USD **2020**
driven by platform for **Mobile broadband**, **Social** business, **Cloud** services, and **Big data** and **analytics**

European effort

Many initiatives also within **Horizon 2020**, also connected with **Open** and **Linked data** (Bologna Open data)
NESSI platform proposal on Big data

BIG DATA AND MORE

Information systems require a **quality-aware vision that can the organize the whole data lifecycle**

5 V's for new data processing

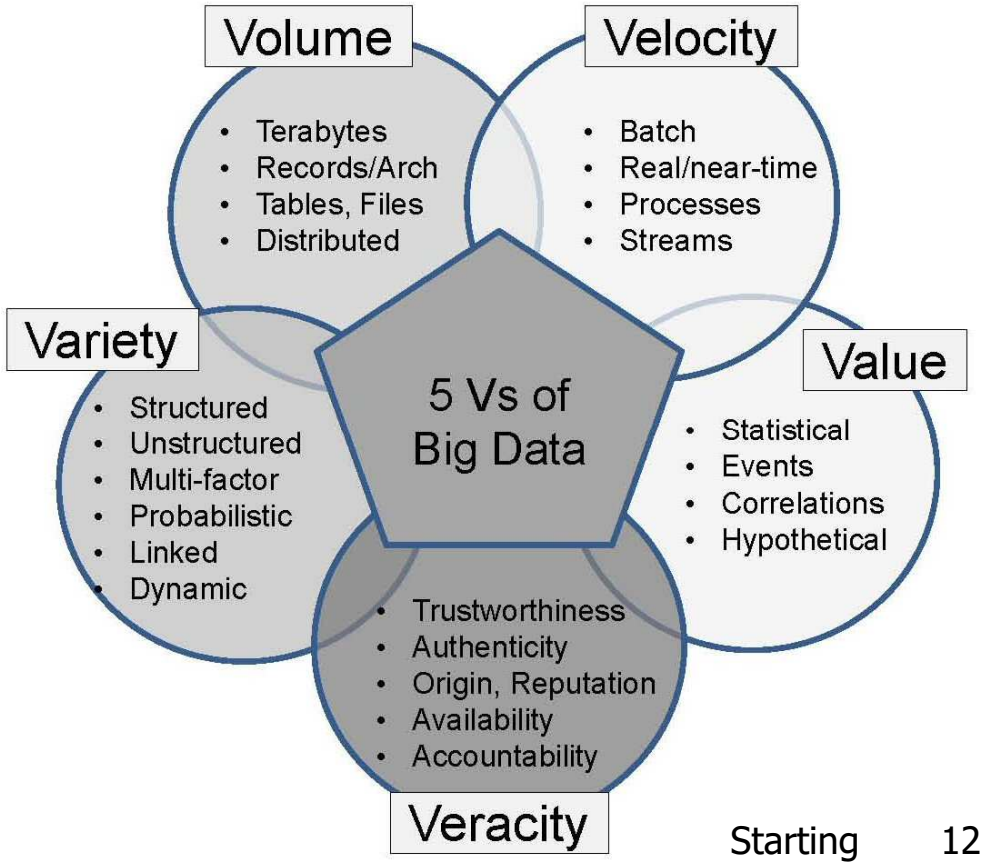
and

novel data treatment

- **Volume of Data**
- **Variety of Data**
- **Velocity**
- **Value**
- **Veracity**

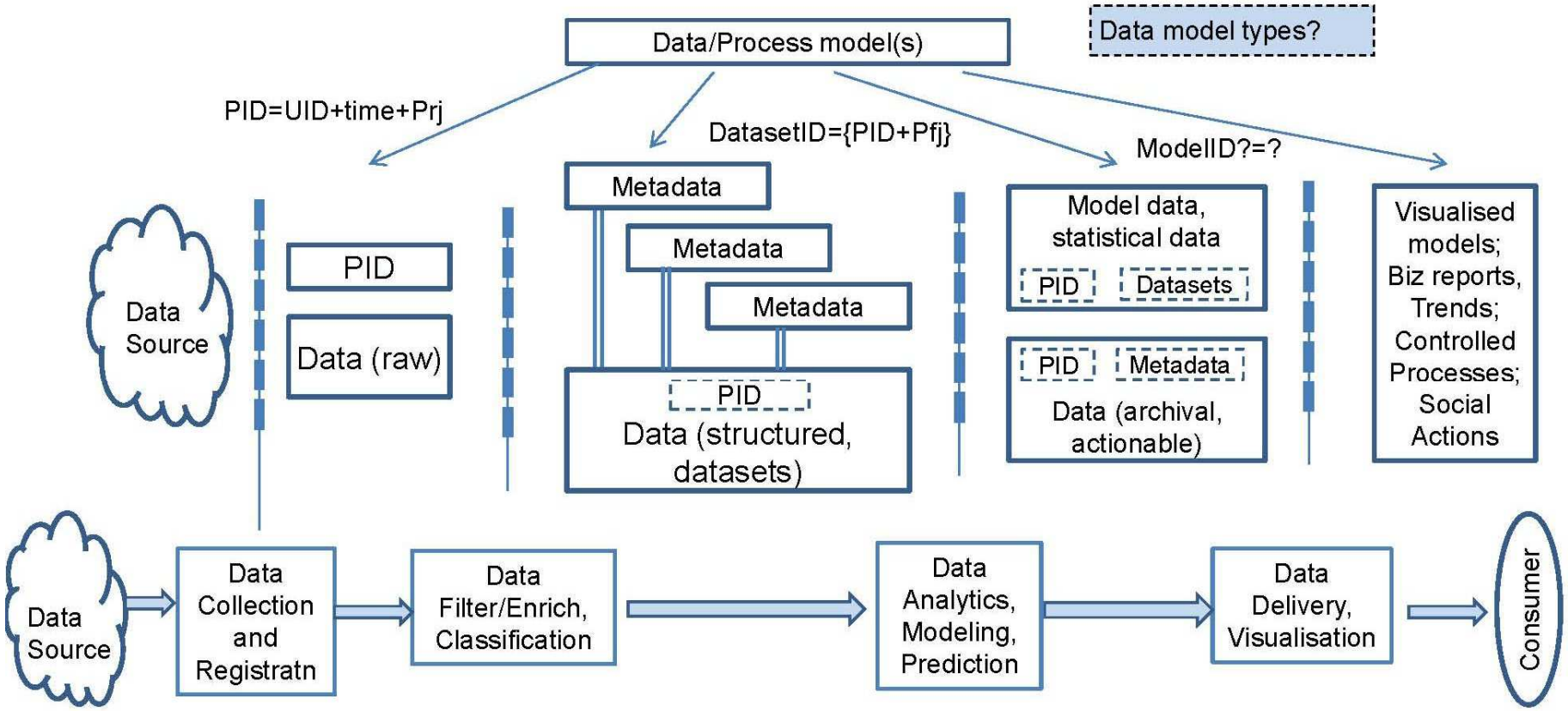
6 V's also Data Dynamicity

- **Variability**



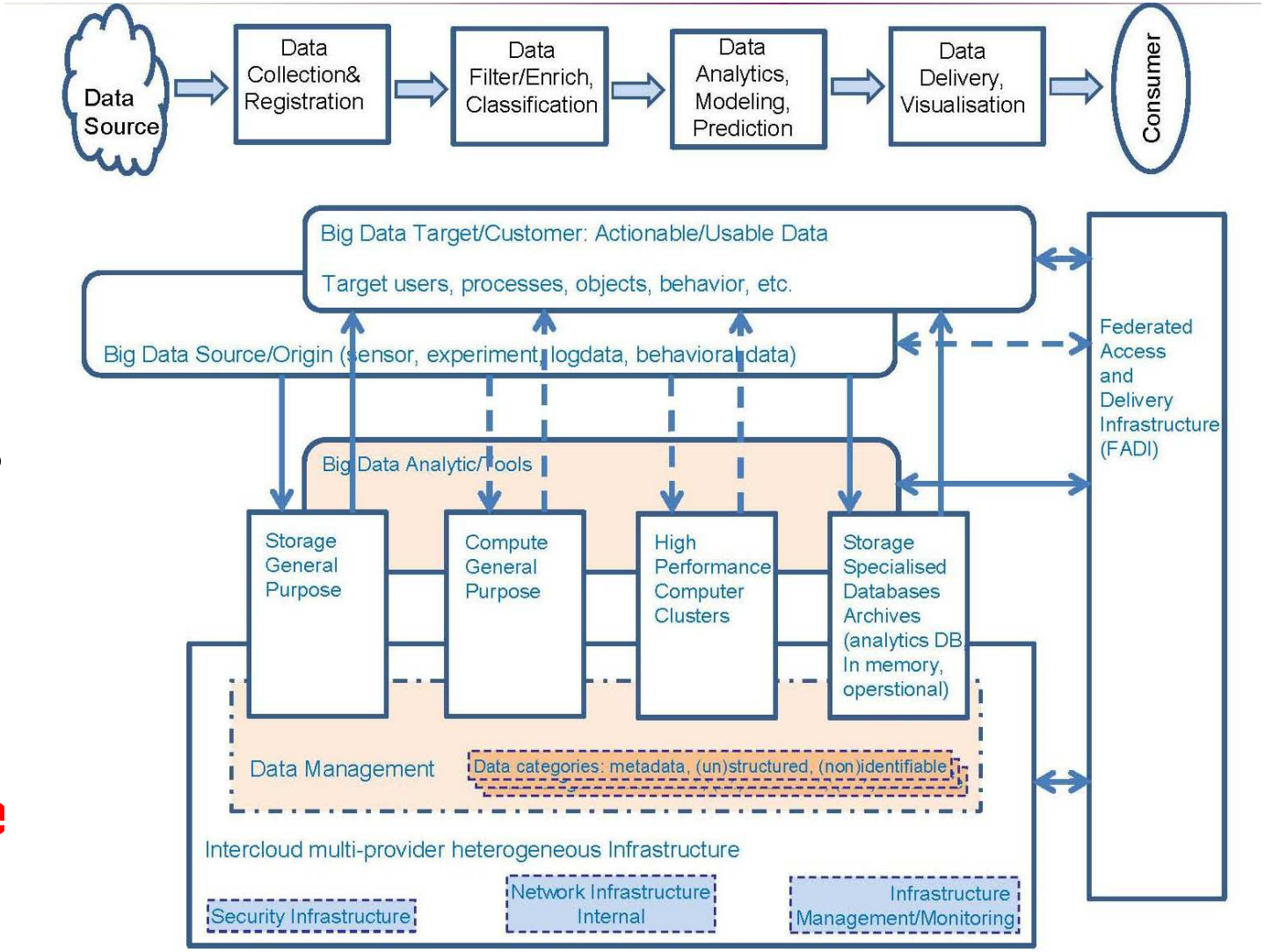
DATA TRANSFORMATION MODEL

The main workflow is to move data from source to sink via a pipeline easy to map and describe



INTEGRATED DATA ARCHITECTURE

Novel Information System organization require **new architectures with novel design principles based on quality-aware services**



TYPICAL SERVICE ENVIRONMENTS

While there are **many application areas that can offer complete scenarios** where you can find all the **topics** and the **solutions** we are interested in this class, we can **focus attention to one specific area**

The **smart city** topic is very **hot** and **pursued** in **several senses**

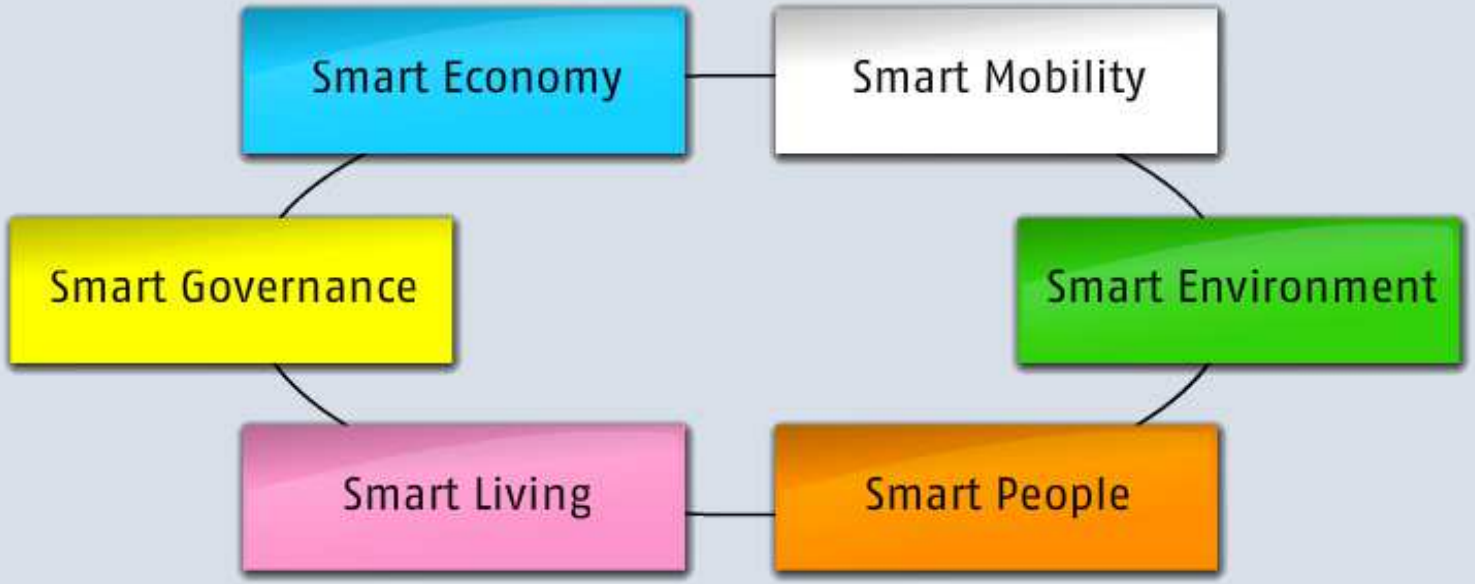
- It is a goal of public administrations and EU policy financing
- It is a area that can contain many (open) data and sets
- It is an area where streams of data can be harvested
- It is an area where citizen can move around and require services also in a localized way

The smart city contains many **data** but also include, require, and can manage many **IT resources**

Smart City:
smart perspectives

The smart city model

A Smart City is a city well performing in 6 characteristics, built on the 'smart' combination of endowments and activities of self-decisive, independent and aware citizens.



SMART CITY SCENARIO

In a **smart city**, we may consider and appoint **attention** to some **specific behaviors that produce a big data system in interaction with other ones (in the complexity stemming from global interaction)**

- **Group** of replicated resources and interacting components
- **Co-creation** of new contents such as videos, pictures, etc.
- **Collection** of big data
- **Harvesting** of open data
- **Management** of resources and people information
- **Public services**
- **Specific workflow** for communities

We can also **focus on** some **locality** to work with and test and experience **a smaller-size isolated system**

REQUIREMENT FOR SERVICES

In distributed systems, while the service must be correctly provided, it **is a compulsory goal** the **Quality of Service (QoS)**, in the sense of **provisioning with some parameters** and **respecting requirements**

The QoS has many **different meanings**, because it is a **quality indicator**

- It can stress **response time**, **security**, **correctness**, **availability**, **confidence**, **user satisfaction,** …

QoS goals (conflicting?) in the **Old** and the **New World**

- **Old world:** typically, main goals **reliability** and **enforced consistency**

New world: scalability and availability matters **most of all**

Focus on **extremely rapid response times:** Amazon estimates that each millisecond of delay has a measurable impact on sales!

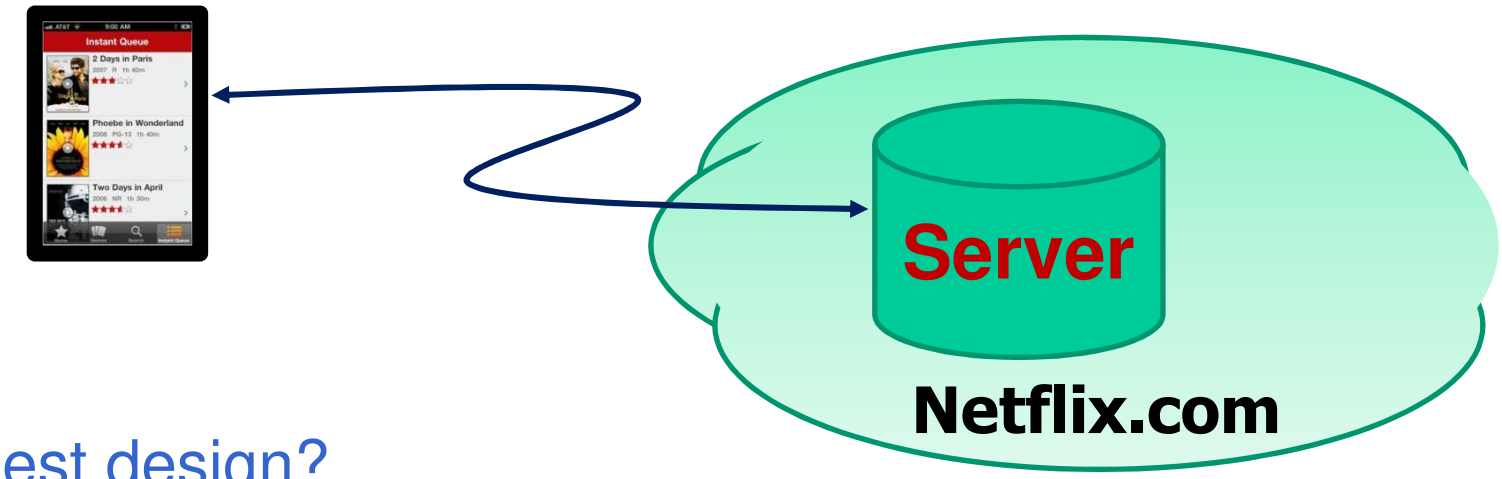
BEHIND THE WOODS: SUPPORT FOR…

To **provide QoS** distributed systems have to support some coverage of **properties** and **functions**

- **Replication**: usage of multiple copies of resources
- **Grouping**: keeping together different copies and behavior
- **Simplified delivery**: new tools and technologies to fasten development & deployment of complex applications
- **Automated management**: infrastructures taking care of management burden with minimal human intervention
- **Batch data processing:** storage/processing of massive amounts of data, such as for Google Web indexing
- **Streaming data**: dealing with information series coming from a set of grouped info, such as a video, sensors, etc.

AN EXAMPLE: NETFLIX

Personal service to **play movies** on demand



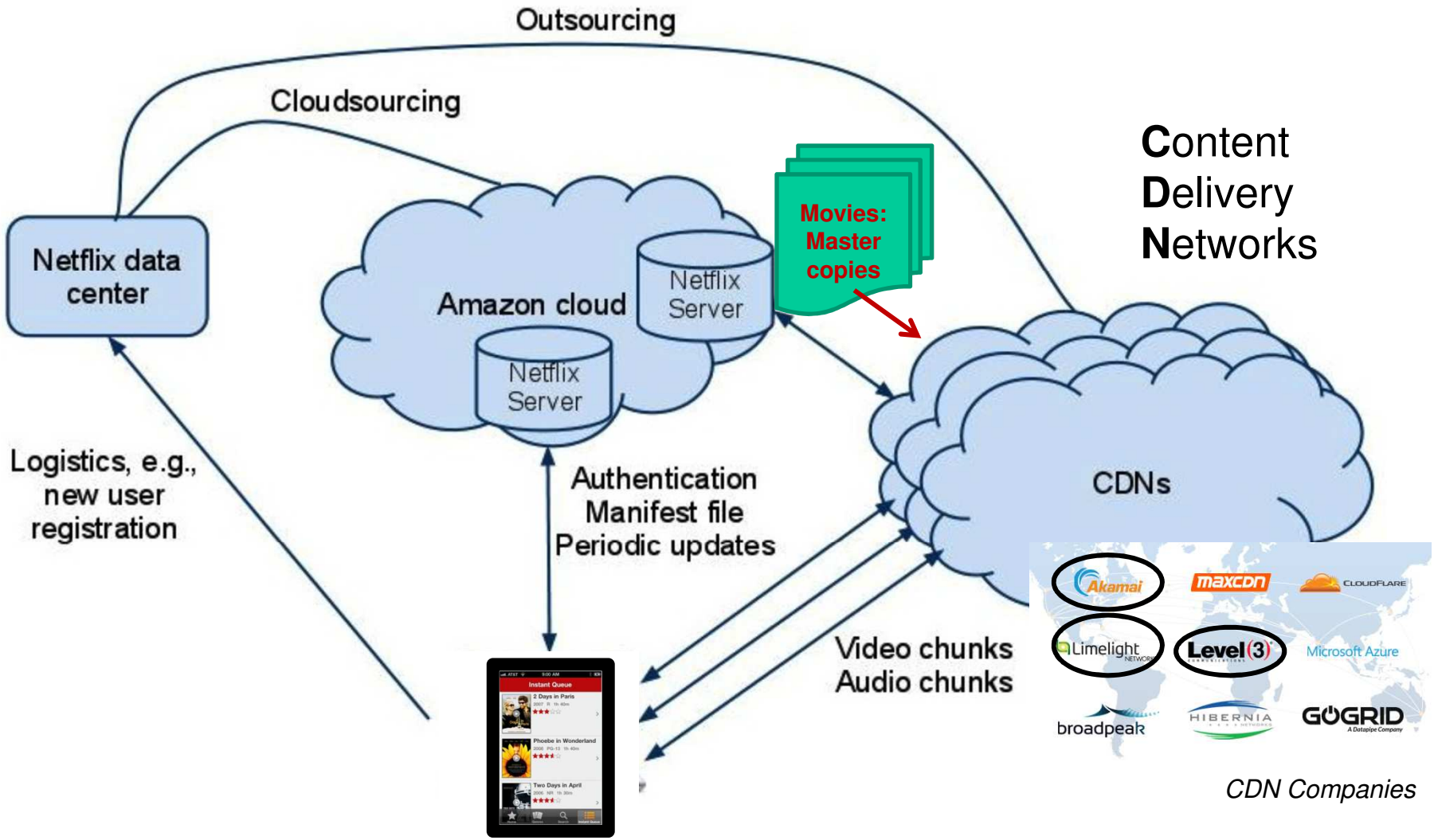
Simplest design?

Netflix owns the data center and content distribution infrastructure

BUT, in the reality….

Netflix owns **neither** a data center **nor** a distribution infrastructure

NETFLIX: THE COMPLEX PICTURE

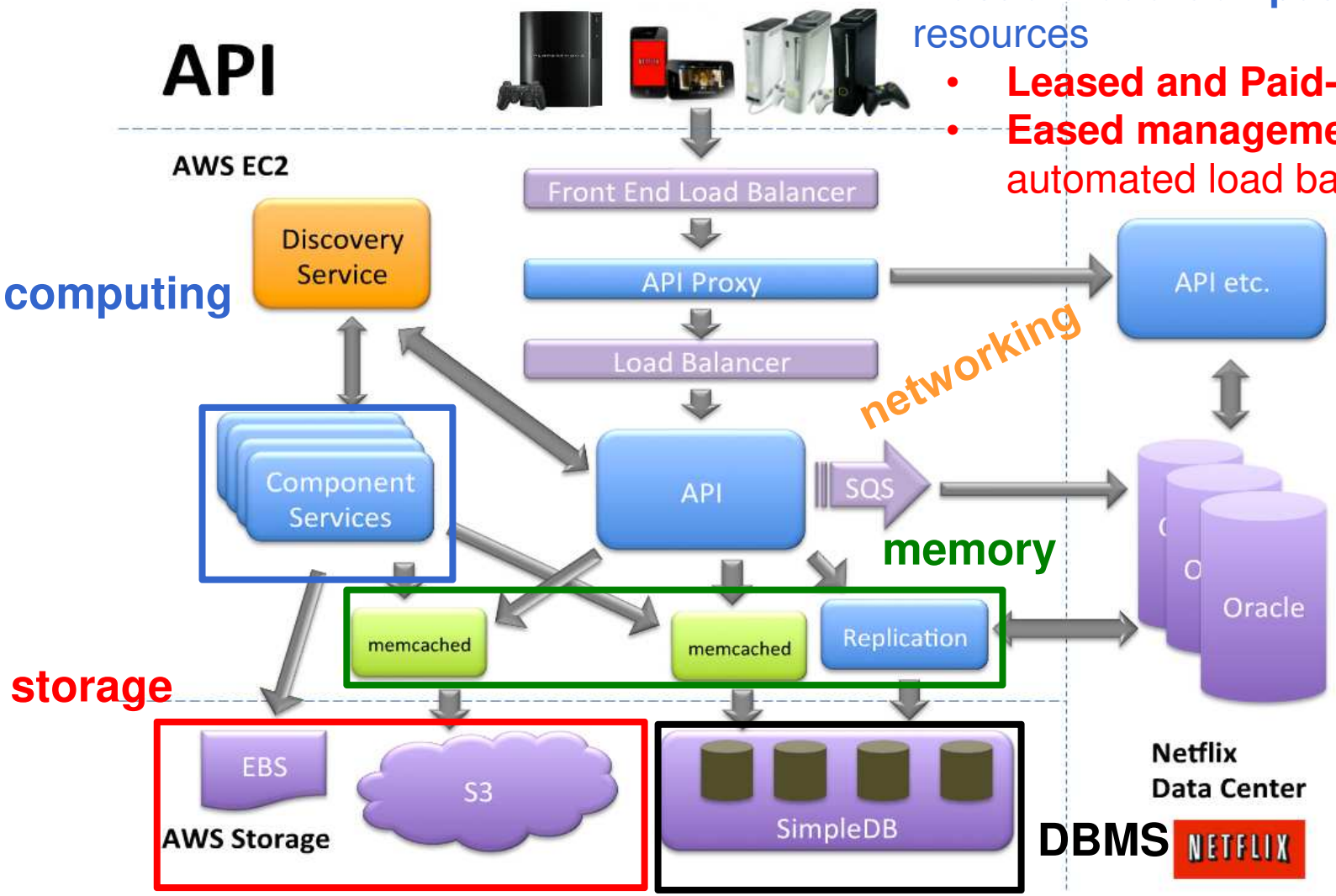


V.K. Adhikari *et al.*, "Unreeling Netflix: Understanding and Improving Multi-CDN Movie Delivery", *IEEE INFOCOM*, 2012.

NETFLIX & AWS EC2 in a NUTSHELL

Amazon Web Services (WS)
Elastic Cloud Computing (EC2)
resources

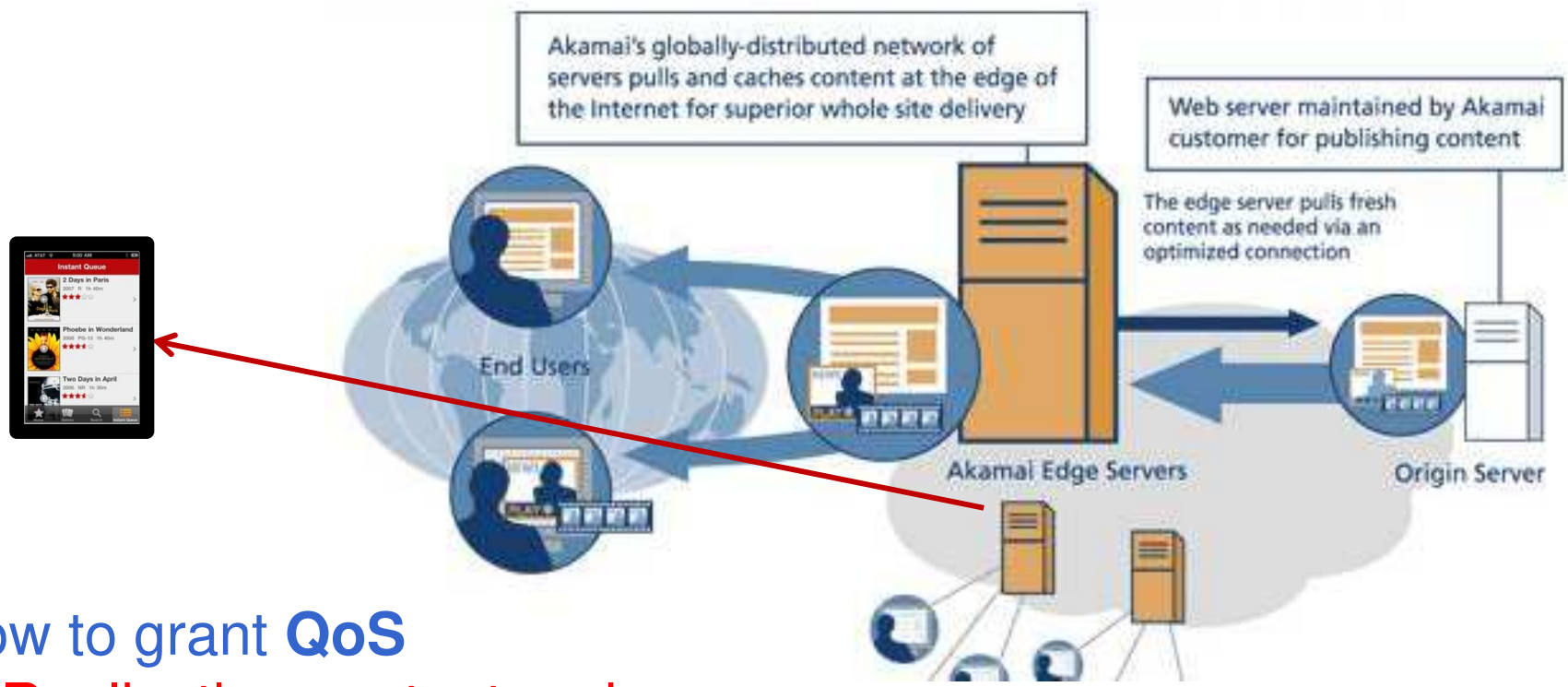
- **Leased and Paid-per-use**
- **Eased management** (e.g., automated load balancing)



NETFLIX & AKAMAI CDN in a NUTSHELL

Many resources

- Capillary **worldwide network**
- Externalized **infrastructure management**



How to grant **QoS**

- Replicating content and servers
- Low latency through identification of **nearby Edge Servers**

INDUSTRY 4.0

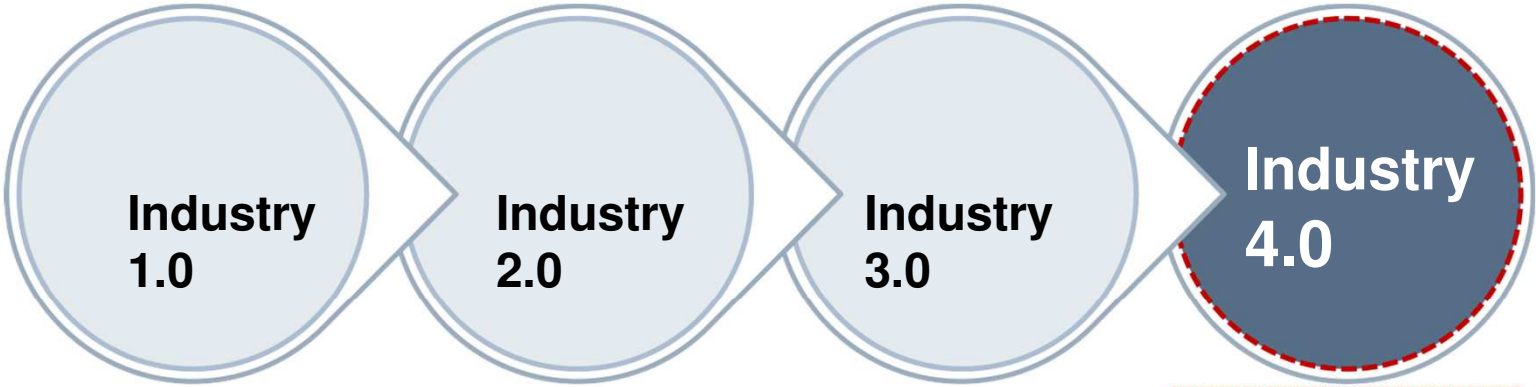
Industry 4.0 is a spreading trend toward an evolution of traditional **industrial processes**

Industry 4.0 (**I4.0**) has multiple meanings

- connects / merges **production with ICT**
- merges **customer data with machine data**
- goes **M2M**: Machines communicate with machines
- components and machines autonomously manage **production in a flexible, efficient, and resource-saving manner**

INDUSTRY 4.0

Industry 4.0 is in the trends of the **industrial revolutions**



- **1760s–1900**
- Use of steam and mechanically-driven production facilities

- **1900–1970s**
- Electric power-driven mass production based on division of labor

- **1970s to date**
- Extensive use of controls, IT, and electronics for an **automated and high productivity** environment

- **Future** & **Smart:** based on integration of virtual and physical production systems

DEFINITION OF INDUSTRY 4.0

INDUSTRIE 4.0 represents the **coming fourth industrial revolution** on the way to an **Internet of Things**, **Data** and **Services**

"The information-intensive transformation of manufacturing in a connected environment of data, people, processes, services, systems and production assets with the generation, leverage and utilization of actionable information as a way and means to realize the smart factory and new manufacturing ecosystems"

Smart industry or "INDUSTRIE 4.0" refers to the **technological evolution from embedded systems to cyber-physical systems**…

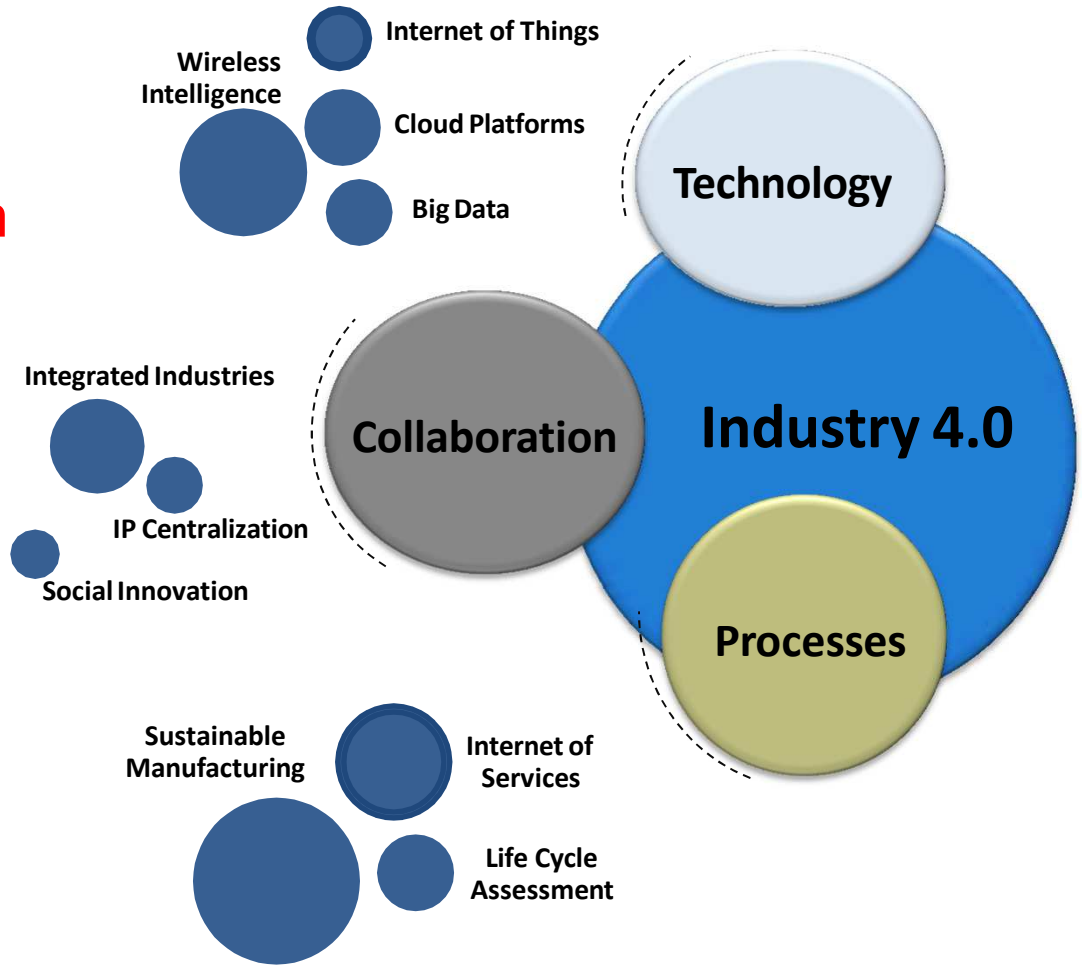
Decentralized intelligence helps create **intelligent object networking and independent process management**, with the **interaction of the real and virtual worlds** representing a crucial new aspect of the manufacturing and production process

Source: Frost & Sullivan

INDUSTRY 4.0 ENVIRONMENT

Industry 4.0 is in the sense of **product innovation in manufacturing**
as an effort
in three areas

- **Technology**
- **Collaboration**
- **Processes**



Source: Frost & Sullivan

COURSE CORE

The complexity of applications asks for ready-to-use off-the-shelf solutions

The answer toward a better usage is "Middleware"

We can give **a first definition**

Middleware is a **set of tools and components** already available for the best **system performance** mainly under the **user required perspective**

A middleware can make **available ready-to-use applications** if a user needs a new functions with no user intervention

A middleware can also **simplify the development of new applications** if the functions are not already available

A middleware can also follow **life cycle to adapt the system** to new requirements and trends

MIDDLEWARE

From the very **complex and differentiated user scenarios,** it is difficult to define **one middleware,** but **many different ones are available and suitable**

We speak of different middlewares for different usage

Different meaning for usage & for adoption and suitable for different environments

1. **personal usage** (for a private user)
2. **company usage** (for internal organization)
3. **global data center usage** (for large data center provider & cloud provider usage)

PRIVATE USERS

A first case is a **middleware to support the needs and requirements of a single user** that typically

- **Has several private machines** (traditional PC and also several smartphones)
- **Works on private data and applications** (typically configured and loaded but also *apps*)
- **Has to access to remote resources** (either company based or globally available on Cloud)

Examples of needed support services/functions:

- **Usage of personal Apps and communication tools**, such as **email, Whatsapp, Telegram**, …
- **Transparent synchronization** of tools across devices, such as in **Skype (for chat), Dropbox (file system)**, and many other services
- **Transparent reliability** through data replication, such as personal storage for backups in Amazon S3
- **Access** through UI and remote visual desktop

ORGANIZATION INTERNAL SUPPORT

A second case is a **middleware to support the needs and requirements** of either **a private or public organization with specific goals to provide services to internal users**

- **Has several user machines and applications** (traditional PC, mobile & small group resources, …)
- **Works on company server in local data center** (typically servers and their resources)
- **Has to access to remote resources** (either on other companies or on global Cloud)

Examples of needed support services/functions:

- **Transparent services**: replication/group synch, load balancing, naming, accountability,
- **Non Transparent Project management** and **support tools**: service monitoring, decision systems, …
- **Management** of **service delivery & used resources** (computing, storage, network, …): both via CLI and visual UI

CLOUD PROVIDERS

A third case is a **middleware to support the needs and requirements of a (general-purpose) data center typically available in Cloud**

- **Has several IT resources** (large quantities of servers in groups, large data servers and storage, more special purpose IT resources, …)
- **Offers services to several client organizations** (typically bare services, and more articulated ones)
- **Has to honor accepted contracts** (not only locally, but also coordinating with provider in need)

Examples of needed support services/functions:

- Management & monitoring of physical infrastructure & of support functions to enable sharing of resources
- Advanced physical resource management to grant: agreed quality levels, isolation (security & performance), …

THIS CLASS ISSUES

The course aims at elaborating on the knowledge of distributed systems for the whole life cycle operation, for the aspects related the execution

- **Operations in the entire life cycle**
- **System management**
- **Quality of service (QoS)**
- **Variations during the life cycle**
- **Recovery and tuning**

Less interest paid to

- Design phases
- Coding
- Preparation and static analysis

CLASS INTERESTS

- **Topics oriented toward the execution environment**
  - All the aspects are selected in the sense of their contribution toward a **better execution**
  - General topics are conjugated with the idea of their **presence and support for the execution part of the life cycle**, always the dominant in time
- **Individual experience**
  - **Capacity of reading technical papers**
  - **Skill to support going depth into a topic**
  - **Writing & Presentation on technical topics**
  - **Design a small project and solution sketch**

Distributed systems and Applications

Middlewares to support Distributed Systems

Where a suitable infrastructure (a middleware) handles and manages all system resources

Some interesting Middleware lines

- **Object** middleware **(CORBA,** COM, .NET**, …)**
- **Message exchange** middleware **(MOM)**
- **Overlay Networks, File systems, NoSQL support**
- **Cloud** systems & **middleware (OpenStack, CloudFoundry)**
- **Data processing** & **streaming** middleware **(Hadoop, SPARK)**

Middleware as a container of **support environment**

Some tools are common to all different kinds of middleware

CLOUD AS AN EVOLUTION

A **necessary and unavoidable step ahead**

Cloud Architectures and solutions

Possibility of **off-the-shelf solutions organized** around and with **Web-accessible resources** in **remote data centers**

- ready-to-use **Systems**
- easy **Systems**
- **pay-per-use Systems**
- transparent (or non) **Systems**
- flexible, extensible & elastic **Systems**
- reliable **Systems**
- secure **Systems**

PRE-REQUISITES...

- Skills on operations in different environments (previous lab presence is recommended)
- Skills on most significant models for distributed systems
  concurrency, processing, storage, …

LATERAL SKILLS

- Capacity of implementing and controlling real projects
- Capacity of exploring in an independent way
- Skills in project engineering
- Skills in English …

GOALS

Design of a service/application architecture
Execution and performance of the project

- **Analysis Capacities**
  - Understanding of **Principles** and **support environments** for general-purpose services and special-purpose ones
  - Understanding of **Projects** and **Solutions** at different levels: conceptual, architectural, at protocol level, algorithmic one, by using different technologies & components
- **Synthesis Capacities (see site)**
  - Speech based on some read **paper**, chosen & elaborated
  - Design of a ***chosen case study***
  - Presentation of a **written report** as a 'to-be-published' article

CLASS RESULT

The **final grading** stems from **an oral exam**

to ascertain the **knowledge** and **orientation** about the entire discipline, ranging on all topics, starting with the basics, going through the practical portions of middleware, and also with a possible follow-up on a chosen topic

You can also choose the project activities (for 4 credits), recommended for the Distributed System Computer Engineering path

Assignment of a project on a specific **subject assigned** and **done individually**

Project activity

Projects can deal with any topics of the class

- **Data Monitoring Aggregation for deployments OpenStack multi-region**
- **Monitoring and Scalability of CloudFoundry for PaaS**
- **Linked data and Semantic Data support for Storm real-time processing**
- **Storage Levels and Inputs in Apache Spark**
- **Load balancing in S4**
- **Enhancing networking in Openstack**
- **Multi-Cloud PaaS Services**
- **Infrastructures to support Blockchain**
- **…**

GRADING - WORKFLOW

The final score is via the oral exam
almaesami is the site for the enrollment

La **first step is the enrollment on the list and find the dates**
Scheduled days in almaesami and **oral exams for the class on dates:**

First exam	(Friday, **14th June 2019**)
Second exam	(Friday, **5th July 2019**)
Third exam	(Friday, **19th July 2019**)

And the oral …

La **first step (for the project activity) is the enrollment on the list and find the dates, give in the project, then the enrollment**
Scheduled days in almaesami and **oral exams for the class on dates:**

Giving in the two-part project (report & implemented project)

First exam	(Friday, **14th June 2019**)
Second exam	(Friday, **5th July 2019**)
Third exam	(Friday, **19th July 2019**)

And more oral exams…

CLASS WEB SITE

`https://middleware.unibo.it/ICCBD`

- Find there
  - Teaching contents (lessons, exercises)
  - Information & discussion exchange
  - Some project topic and area proposals
- The available lab
  - **LAB2** available non class schedule
  - Middleware tools there, also individual practice
    CORBA, OpenStack, Hadoop, SPARK, …
- Via Web
  - Many papers available
  - Some personal deepening hints

Hands-on Seminars

Planning of hands-on experience about some novel directions in relevant technologies

not within class hours

- Seminars to introduce company technology perspective
- Companies can give a picture of what is their experience and which technical roles and are significant for and with them

Importance of

- **Possibility of studying abroad / work experience**
- **Serious language skills (apart from technical)**

SOME MATERIALS and ITEMS

- **Class Slides** Available
  - on the web site of the class
  - at the copy center of the School
- **Some basic books**
  - G. Coulouris, J. Dollimore, T. Kindberg, "***Distributed Systems: Concepts and Design***", Addison-Wesley, (fifth edition) 2012.
  - A.S. Tanenbaum, M.v.Steen "***Distributed Systems: Principles and Paradigms***", Prentice-Hall, second edition 2006.
  - B. Forouzan, F. Mosharraf: "***Computer Networks, a top down approach***", McGrow-Hill, 2011.
  - M.L. Liu, "***Distributed Computing***", Addison-Wesley, 2003.

SOME (CLASSIC) REFERENCE BOOKS

- D.L. Galli, "***Distributed Operating Systems: Concepts and Practice***", Prentice-Hall, 2000.
- L. Peterson, B. Davie, "***Computer Networks, A Systems Approach***", Second edition, Morgan Kaufmann, 2000.
- V.K. Garg, "***Elements of Distributed Computing***", Wiley, 2002.
- J.F. Kurose, K.W. Ross, **"Computer Networking: a Top-Down Approach Featuring the Internet"**, McGraw-Hill, 2001).
- J. Siegel, "***CORBA 3: Fundamentals and Programming***", (second edition), OMG Press, Wiley, 2000.
- F. Halsall, "***Multimedia Communications***", Addison-Wesley, 2001.

SOME BOOKS ON LATEST TOPICS

- T. Erl *et al.*, "***Cloud computing : concepts, technology, & architecture***", Prentice Hall, 2013.
- B. Wilder, "***Cloud architecture patterns***", Beijing, 2013.
- A. T. Velte *et al.*, "***Cloud computing: a practical approach***", McGraw-Hill, 2010.
- J. Rhoton, "***Cloud computing explained***", Recursive Press, 2009.
- T. Fifield *et al.*, "***Openstack operations guide: set up and manage your OpenStack cloud***", O'Reilly, 2014.
- S. Holla, "***Orchestrating Docker***", Packt Publishing, 2015.
- O. Hane, "***Build your own PaaS with Docker***", Packt Publishing, 2015.

SOME BOOKS ON LATEST TOPICS

- T.D. Nadeau and K. Gray, "***SDN: software defined networks***", O'Reilly, 2013.
- L. Carlson, "***Programming for Paas***", O'Reilly, 2013.
- T. White, "***Hadoop: the definitive guide***", O'Reilly, 2012.
- E. Sammer, "***Hadoop operations***", O'Reilly, 2012.
- K. Rankin, "***DevOps troubleshooting***", Addison-Wesley, 2013.
- D. Sui *et al.*, "***Crowdsourcing geographic knowledge***", Springer, 2013.
- Z. Yan *et al.*, "***Semantics in mobile sensing***", Morgan & Claypool, 2014.
- R. Copeland, "***MongoDB applied design patterns***", O'Reilly, 2013.

Many sources – Internet apart

Please refer to articles on different topics in journals published by the two professional organization

ACM (**Association for Computing Machinery**) e

IEEE (**Institute of Electrical and Electronic Engineering**)

Groups **www.computer.org** **www.comsoc.org**

General magazine:

IEEE ***Computer***, *ACM* ***Communications***

IEEE **Internet Computing** e IEEE **Communications**

also **Distributed Systems OnLine** http://dsonline.computer.org

Depth into journals very specific and helpful

ACM **Computing Surveys** (ACM CS), ACM Transactions on...

IEEE Transactions on .... (IEEE Trans…, ACM Trans…)

IETF Request for Comments

You can see both from UNIBO sites and UNIBO students account