Ciência de Dados em Larga Escala

Inês Dutra and Zafeiris Kokkinogenis

DCC-FCUP
room 1.31
ines@dcc.fc.up.pt
zafeiris.kokkinogenis@gmail.com

23/24

Parallel and distributed platforms

- High Performance Computing (HPC)
- High Throughput Computing (HTC)
- Grid Computing
- Cloud Computing

Parallel and distributed platforms

HPC

- processors dedicated to a specific task or application
- objective: accelerate sequential tasks
- usually clusters of machines with homogeneous software and hardware

Parallel and distributed platforms

- "Milipeia" is a cluster at the University of Coimbra
  - 130 compute nodes
    Sun Fire X4100
  - 520 cores opteron@2.2 GHz
    2 GB p/core, GigE
  - 1 login node and 1
    management node
  - 6 TB main storage
  - 1.6 Tflop/s sustained performance
  - CentOS, GNU, Pathscale and Intel compilers, Cernlib, several numerical libraries

Parallel and distributed platforms

ENTREVISTAS / INOVAÇÃO

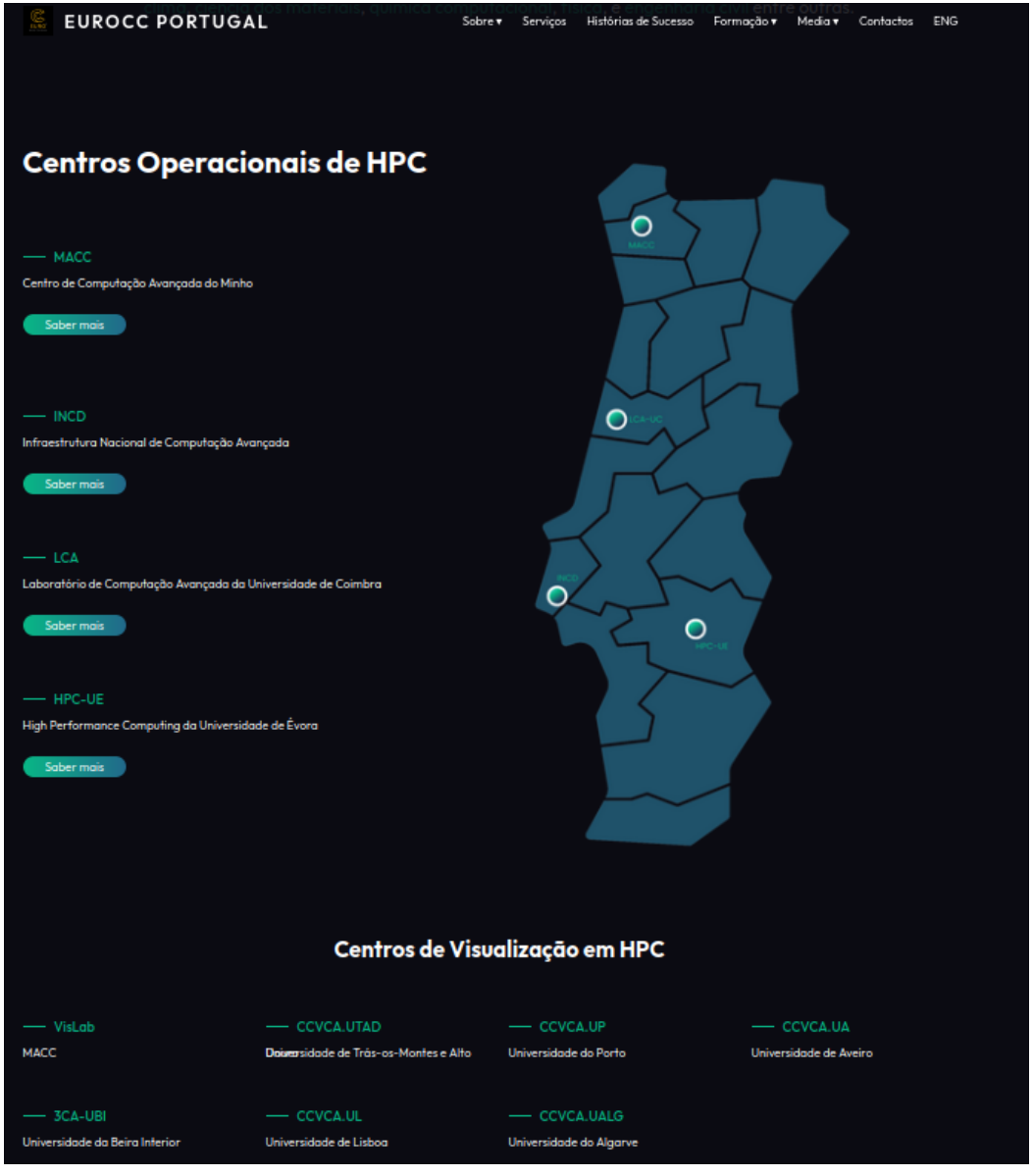
Deucalion – o supercomputador que servirá a estratégia nacional da computação avançada

by admin | Published 24/02/2021

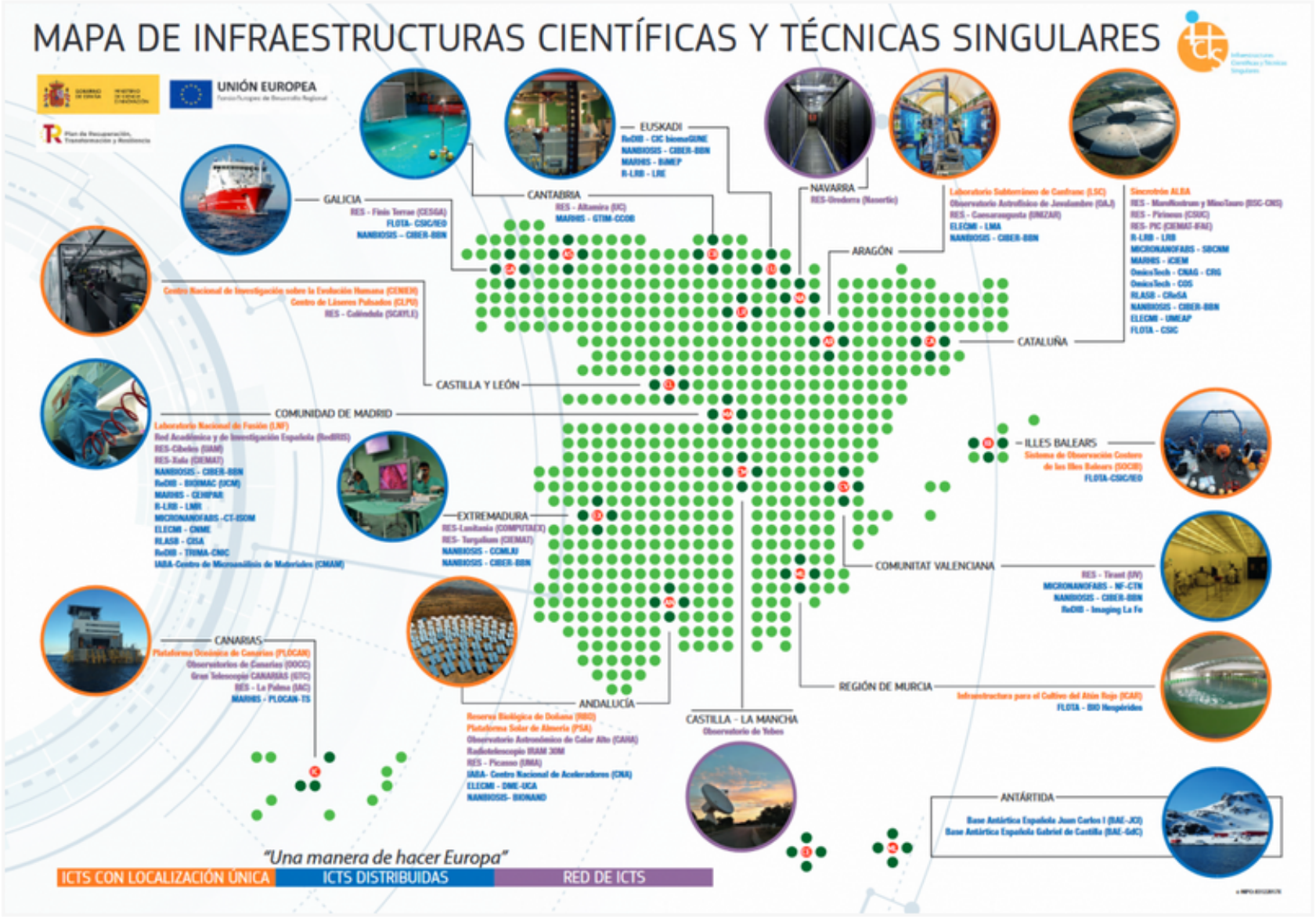
Em entrevista ao Guimarães Agora!, o docente da EEUM Rui Oliveira, diretor do Minho Advance Computing Center, infraestrutura nacional de computação avançada operada pela Universidade do Minho, dá-nos uma perspetiva do que vai fazer, quando e como o supercomputador enquadrado numa estratégia nacional e europeia, reforçando a sua competitividade no mundo.

Leia a entrevista aqui

Parallel and distributed platforms



Parallel and distributed platforms



Parallel and distributed platforms

TOP10 System - November 2023

R_{max} and R_{peak} values are in PFlop/s. For more details about other fields, check the TOP500 description.

R_{peak} values are calculated using the advertised clock rate of the CPU. For the efficiency of the systems you should take into account the Turbo CPU clock rate where it applies.

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	4,742,808	585.34	1,059.33	24,687
3	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Microsoft Azure United States	1,123,200	561.20	846.84	
4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899

Parallel and distributed platforms

Tianhe-1A

- supercomputer in China
- 14,336 Xeon X5670 processors
- 7,168 Nvidia Tesla M2050
- consumes 4.04 megawatts (MW) of electricity
- cost to power and cool the system can be significant:
  - e.g. 4 MW at \$0.10/kWh $\rightarrow$ \$400 an hour or about \$3.5 million per year

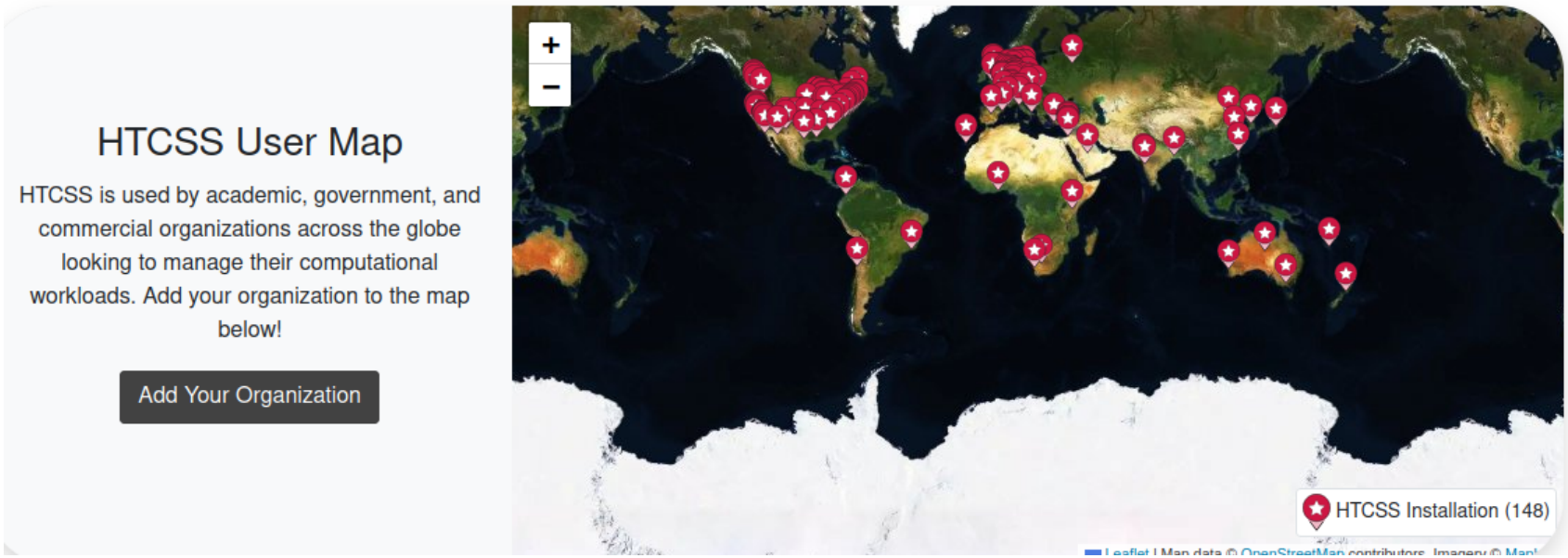
Parallel and distributed platforms

HTC

- opportunistic model
- processors are “scavenged” when idle
- potentially heterogeneous machines
- objective: increase throughput (number of jobs per time unit)

Parallel and distributed platforms

HTCondor



Parallel and distributed platforms

Grid Computing

- HPC or HTC model
- processors are “scavenged” when idle or allocated by a batch scheduler
- potentially heterogeneous machines
- objective: increase throughput or accelerate sequential execution
- usually bare metal (no virtualization)

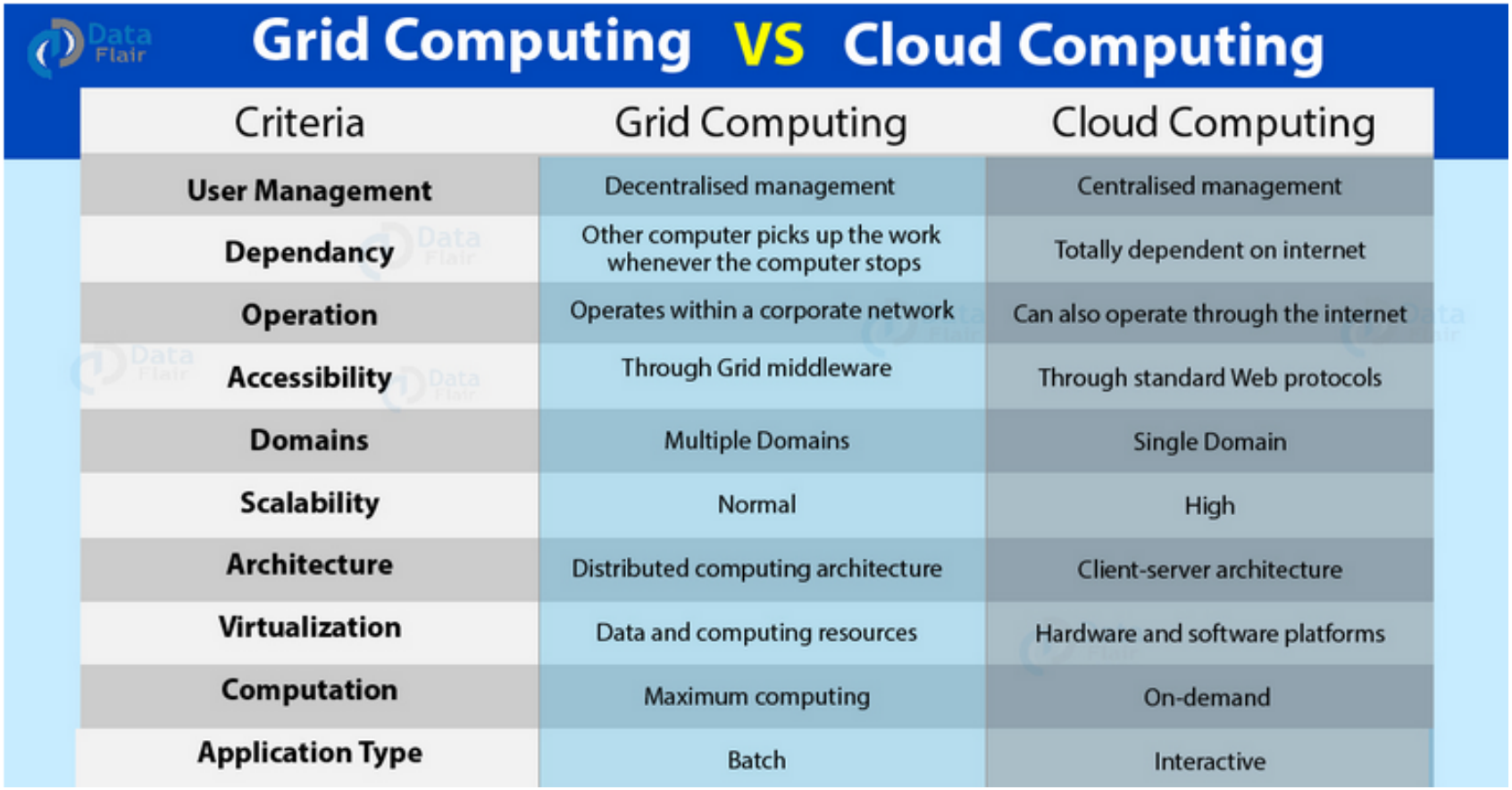
Parallel and distributed platforms

Cloud Computing

- Usually HPC-oriented, but not always
- Based on Virtual Machines (VM), on-demand
- Various infrastructure models

Parallel and distributed platforms

Grids x Clouds



Grid Computing VS Cloud Computing

Criteria	Grid Computing	Cloud Computing
User Management	Decentralised management	Centralised management
Dependancy	Other computer picks up the work whenever the computer stops	Totally dependent on internet
Operation	Operates within a corporate network	Can also operate through the internet
Accessibility	Through Grid middleware	Through standard Web protocols
Domains	Multiple Domains	Single Domain
Scalability	Normal	High
Architecture	Distributed computing architecture	Client-server architecture
Virtualization	Data and computing resources	Hardware and software platforms
Computation	Maximum computing	On-demand
Application Type	Batch	Interactive

Cloud Computing

"*Cloud computing* is an information technology (IT) paradigm that enables ubiquitous access to shared pools of configurable system resources and higher-level services that can be rapidly provisioned with minimal management effort, often over the Internet. Cloud computing relies on sharing of resources to achieve coherence and economies of scale, similar to a public utility."

https://en.wikipedia.org/wiki/Cloud_computing

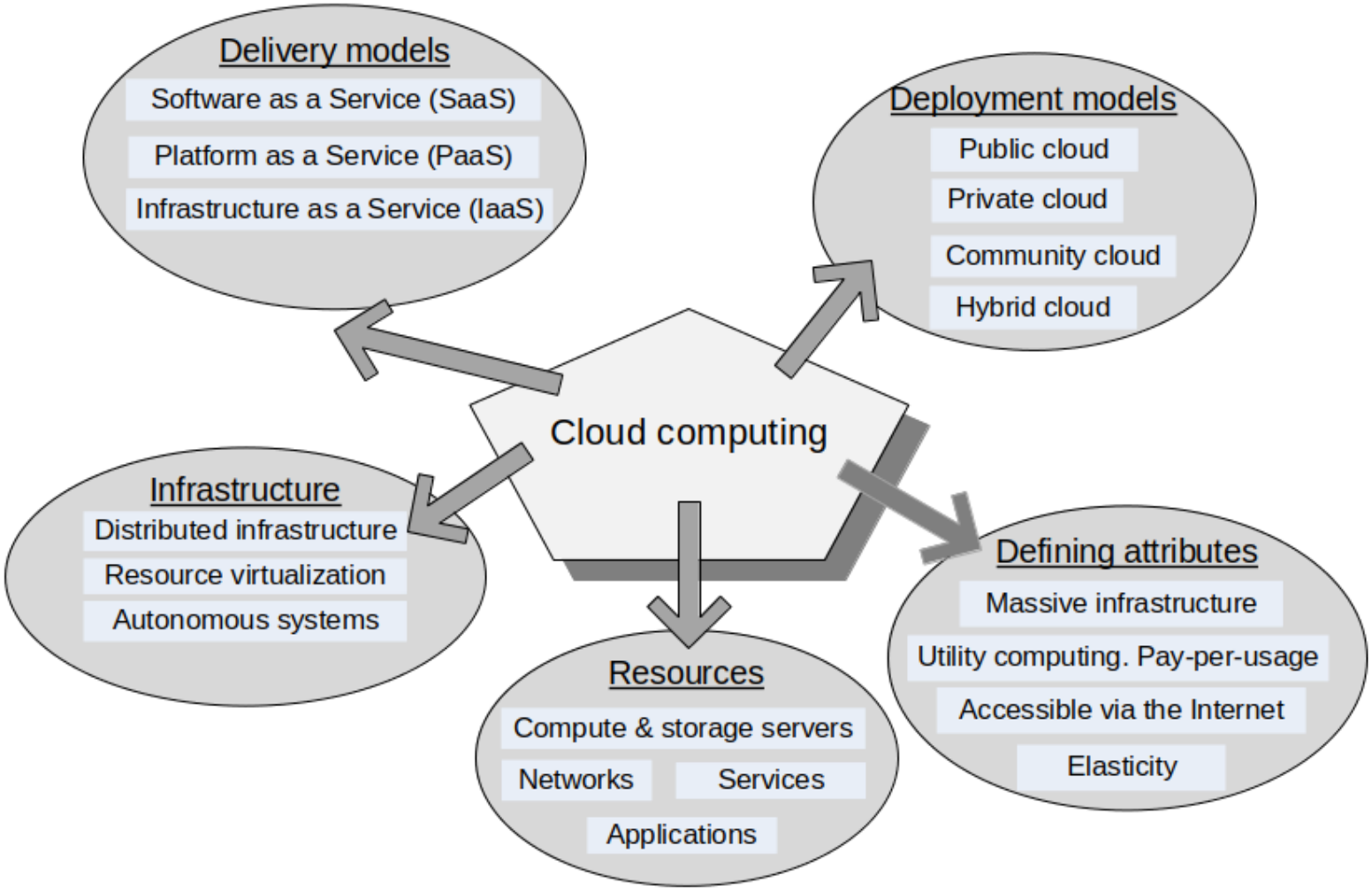
Parallel and distributed platforms

"Simply put, cloud computing is the delivery of computing services – servers, storage, databases, networking, software, analytics and more – over the Internet ("the cloud"). Companies offering these computing services are called cloud providers and typically charge for cloud computing services based on usage, similar to how you're billed for gas or electricity at home."

https://azure.microsoft.com/en-gb/overview/what-is-cloud-computing/

Cloud Computing

Models, Resources, Attributes



Marinescu's cloud computing book

Cloud Computing: early models

Rationale: information and data processing can be done more efficiently on large farms of computing and storage systems accessible via the Internet. Early models:

- Grid computing – initiated by American National Labs in the early 1990s; targeted primarily at scientific computing.
  "Grid computing is the collection of computer resources from multiple locations to reach a common goal. The grid can be thought of as a distributed system with non-interactive workloads that involve a large number of files." from Wikipedia
- Utility computing – initiated in 2005-2006 by IT companies and targeted at enterprise computing.
  "Utility computing is a service provisioning model in which a service provider makes computing resources and infrastructure management available to the customer as needed, and charges them for specific usage rather than a flat rate." from Wikipedia

Cloud Computing characteristics

"Cloud Computing offers on-demand, scalable and elastic computing (and storage services). The resources used for these services can be metered and users are charged only for the resources used." from Marinescu's book
Shared Resources and Resource Management:

- Cloud uses a shared pool of resources
- Uses Internet technology to offer scalable and elastic services
- The term "elastic computing" refers to the ability of dynamically and on-demand acquiring computing resources and supporting a variable workload
- Resources are metered and users are charged accordingly.
- It is more cost-effective due to resource-multiplexing. Lower costs for the cloud service provider are passed to the cloud users.

Cloud Computing advantages

Data Storage

- Data is stored in the cloud, in certain cases closer to the site where it is used
- Data appears to the users as if stored in a location-independent manner.
- The data storage strategy can increase reliability, as well as security, and can lower communication costs.

Management:

- The maintenance and security are operated by service providers
- The service providers can operate more efficiently due to specialisation and centralisation

Cloud Computing advantages

- Resources, such as CPU cycles, storage, network bandwidth, are shared
- When multiple applications share a system, their peak demands for resources are not synchronised thus, multiplexing leads to a higher resource utilization
- Resources can be aggregated to support data-intensive applications
- Data sharing facilitates collaborative activities. Many applications require multiple types of analysis of shared data sets and multiple decisions carried out by groups scattered around the globe

Cloud Computing advantages

- Eliminates the initial investment costs for a private computing infrastructure and the maintenance and operation costs.
- Cost reduction: concentration of resources creates the opportunity to pay as you go for computing.
- Elasticity: the ability to accommodate workloads with very large peak-to-average ratios.
- User convenience: virtualization allows users to operate in familiar environments rather than in idiosyncratic ones.

Types of Clouds

- Public Cloud - the infrastructure is made available to the general public or a large industry group and is owned by the organization selling cloud services.
- Private Cloud – the infrastructure is operated solely for an organization.
- Hybrid Cloud - composition of two or more Clouds (public, private, or community) as unique entities but bound by a standardised technology that enables data and application portability.
- Other types: e.g., Community/Federated Cloud - the infrastructure is shared by several organizations and supports a community that has shared concerns.

Why Clouds?

- It is in a better position to exploit recent advances in software, networking, storage, and processor technologies promoted by the same companies who provide Cloud services.
- Economical reasons: It is used for enterprise computing; its adoption by industrial organizations, financial institutions, government, and so on has a huge impact on the economy.
- Infrastructures Management reasons:
  - A single Cloud consists of a mostly homogeneous (now more heterogeneous) set of hardware and software resources.
  - The resources are in a single administrative domain (AD). Security, resource management, fault-tolerance, and quality of service are less challenging than in a heterogeneous environment with resources in multiple ADs.

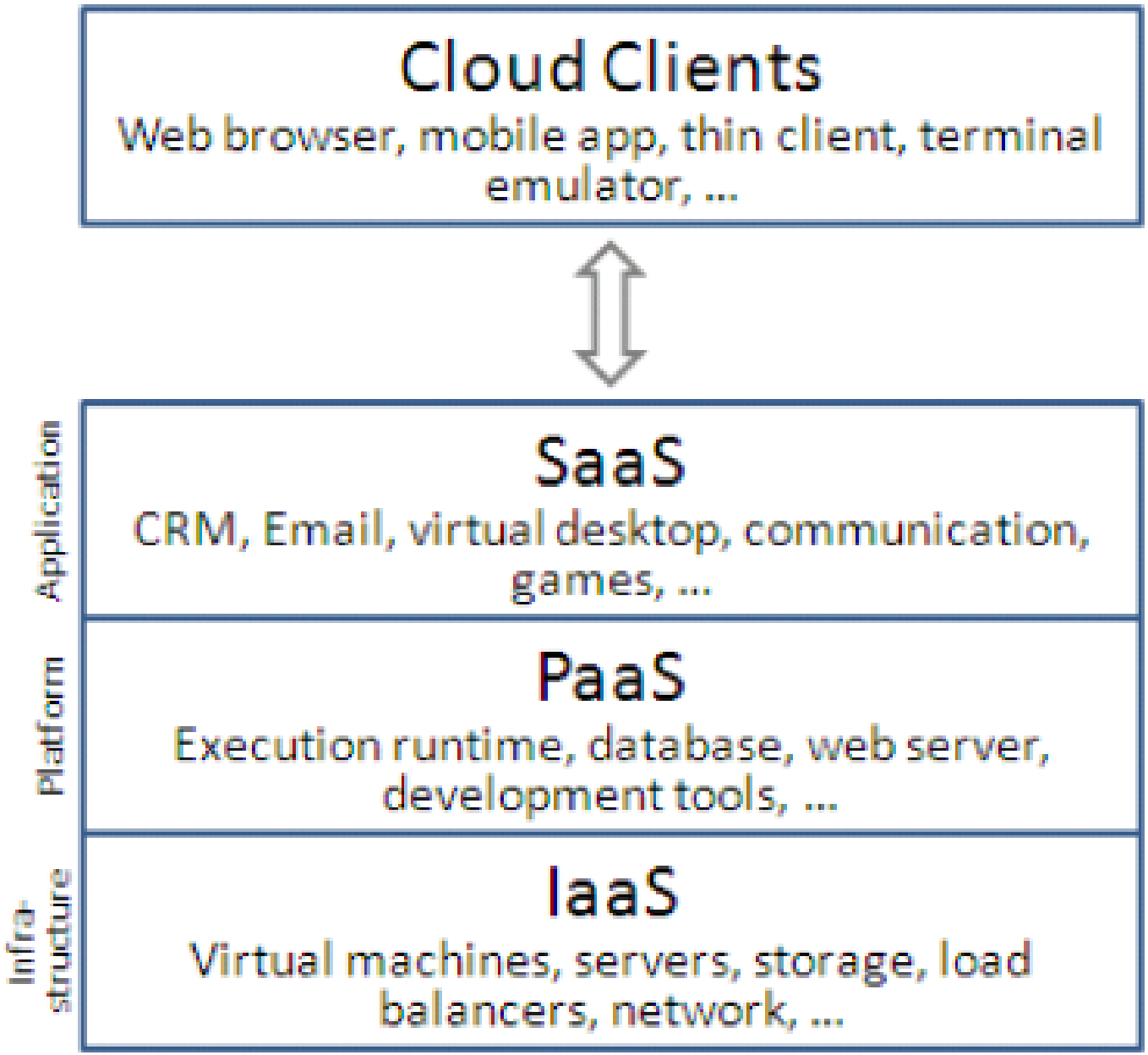
Challenges

- Availability of service: what happens when the service provider cannot deliver?
- Data confidentiality and auditability, a serious problem.
- Diversity of services, data organization, user interfaces available at different service providers limit user mobility; once a customer is hooked to one provider it is hard to move to another.
- Data transfer bottleneck; many applications are data-intensive.

Challenges

- Performance unpredictability, one of the consequences of resource sharing. How to use resource virtualization and performance isolation for QoS guarantees? How to support elasticity, the ability to scale up and down quickly?
- Resource management: It is a big challenge to manage different workloads running on large data centers. Are self-organization and self-management the solution?
- Security and confidentiality: major concern for sensitive applications, e.g., healthcare applications.

Cloud models



Cloud models

IaaS

- Infrastructure is compute resources, CPU, VMs, storage, etc
- The user is able to deploy and run arbitrary software, which can include operating systems and applications
- The user does not manage or control the underlying Cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of some networking components, e.g., host firewalls
- Services offered by this delivery model include: server hosting, storage, computing hardware, operating systems, virtual instances, load balancing, Internet access, and bandwidth provisioning.
- Example: Amazon EC2, GCP, Eucalyptus

Cloud models

PaaS

- Allows a cloud user to deploy consumer-created or acquired applications using programming languages and tools supported by the service provider.
- The user has control over the deployed applications and, possibly, application hosting environment configurations.
- The user does not manage or control the underlying Cloud infrastructure including network, servers, operating systems, or storage.
- Not particularly useful when:
  - The application must be portable.
  - Proprietary programming languages are used.
  - The hardware and software must be customised to improve the performance of the application.
  - Examples: Google App Engine, Windows Azure

Cloud models

SaaS

- Applications are supplied by the service provider.
- The user does not manage or control the underlying Cloud infrastructure or individual application capabilities.
- Services offered include: enterprise services such as: workflow management, communications, digital signature, customer relationship management (CRM), desktop software, financial management, geo-spatial, and search.
- Not suitable for real-time applications or for those where data is not allowed to be hosted externally.
- Examples: Gmail, Salesforce

Cloud models

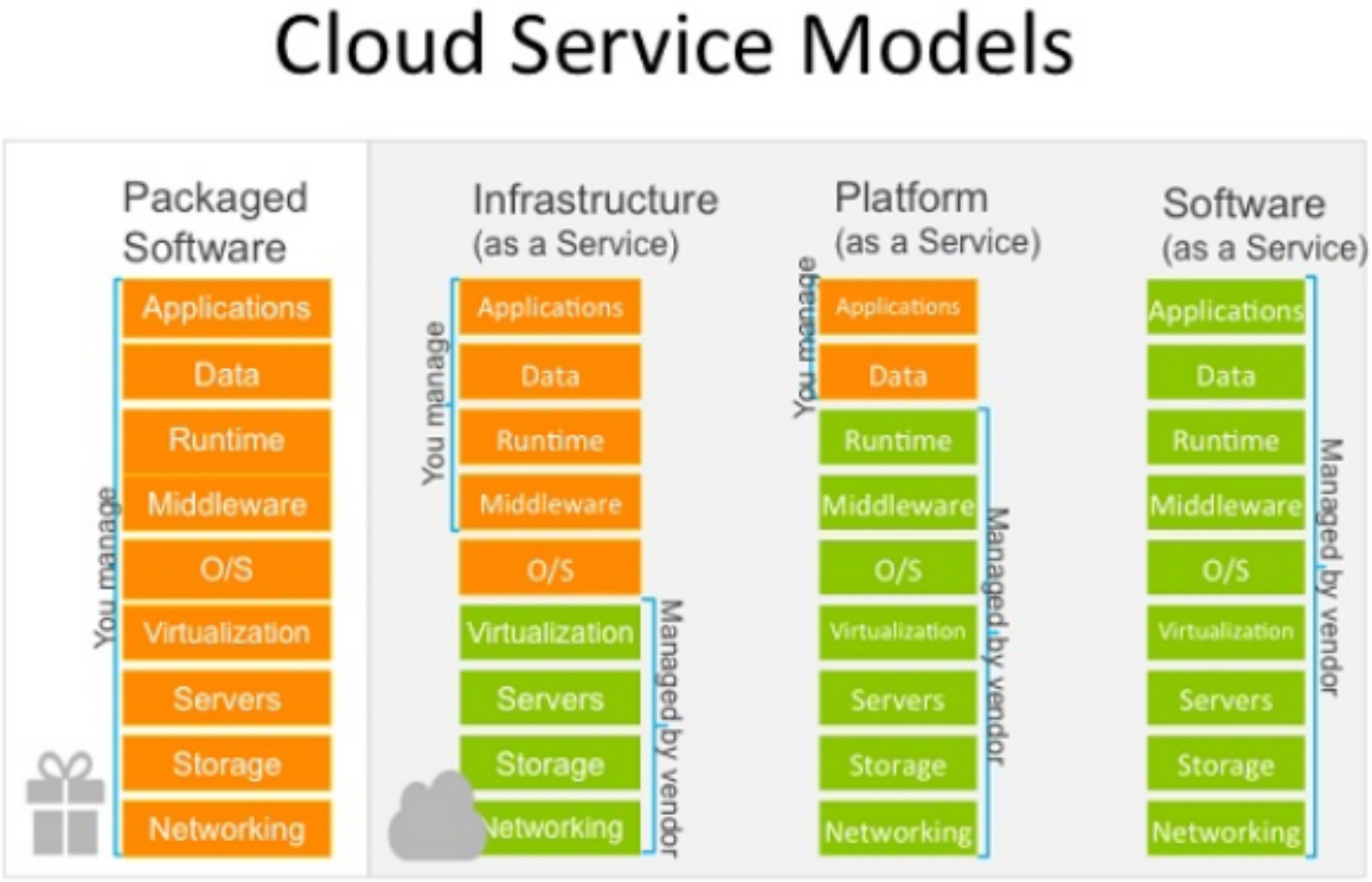


Figure 1. **Source: Microsoft Azure**

Cloud activities

Service management and provisioning including:

- Virtualization.
- Service provisioning.
- Call center.
- Operations management.
- Systems management.
- QoS management.
- Billing and accounting, asset management.
- SLA management.
- Technical support and backups

Cloud activities

Security management including:

- ID and authentication.
- Certification and accreditation.
- Intrusion prevention.
- Intrusion detection.
- Virus protection.
- Cryptography.
- Physical security, incident response.
- Access control, audit and trails, and firewalls.

Cloud activities

Customer services such as:

- Customer assistance and on-line help.
- Subscriptions.
- Business intelligence.
- Reporting.
- Customer preferences.
- Personalization.

Integration services including:

- Data management.
- Development

Ethical issues

- Paradigm shift with implications on computing ethics:
  - The control is relinquished to third party services.
  - Data is stored on multiple sites administered by several organizations.
  - Multiple services interoperate across the network.
- Implications:
  - Unauthorised access.
  - Data corruption.
  - Infrastructure failure, and service unavailability

De-perimeterization

- Systems can span the boundaries of multiple organisations and cross the security borders.
- The complex structure of Cloud services can make it difficult to determine who is responsible in case something undesirable happens.
- Identity fraud and theft are made possible by the unauthorised access to personal data in circulation and by new forms of dissemination through social networks $\rightarrow$ they could also pose a danger to Cloud resources.

Privacy issues

- Cloud service providers have already collected petabytes of sensitive personal information stored in data centers around the world. The acceptance of Cloud Computing therefore will be determined by privacy issues addressed by these companies and the countries where the data centers are located.
- Privacy is affected by cultural differences; some cultures favour privacy, others emphasize community. This leads to an ambivalent attitude towards privacy in the Internet which is a global system.

Cloud Computing

Back to on-premises?

CLOUD COMPUTING

By David Linthicum, Contributor, InfoWorld | FEB 9, 2024 2:00 AM PST

Why companies are leaving the cloud

Cloud is a good fit for modern applications, but most enterprise workloads aren't exactly modern. Security problems and unmet expectations are sending companies packing.

Don't look now, but 25% of organizations surveyed in the United Kingdom have already moved half or more of their cloud-based workloads back to on-premises infrastructures. This is according to a recent study by Citrix, a Cloud Software Group business unit.

The survey questioned 350 IT leaders on their current approaches to cloud computing. The survey also showed that 93% of respondents had been involved with a cloud repatriation project in the past three years. That is a lot of repatriation. Why?

Cost, not cloud

Security issues and high project expectations were reported as the top