DNA Sequence Bioinformatics Analysis with the Galaxy Platform

University of São Paulo, Brazil
28 July - 1 August 2014

Dave Clements
Johns Hopkins University

Robson Francisco de Souza
University of São Paulo

José Ernesto Belizário

University of São Paulo

Course logistics

- Classes: Monday to Friday, 9:00 - 17:00
- Lunch Break: 12:00 - 14:00
- Local: Blue Auditorium at ICB IV building
- We will have free coffee inside the room
- Please, make sure you signed our attendance control list by the end of the morning and afternoon sessions
- Official certificates will be provided by ICB for present at a minimum 70% of the classes

Course infrastructure

- Two Wi-Fi networks were setup by CEFAP for this course
  - CEFAP Galaxy 2
  - CEFAP Galaxy 5
  - Password: @galaxy1
- Nuvem USP accounts were setup for most participants
- All slides and other course material will be available at

  https://wiki.galaxyproject.org/Events/SaoPaulo2014

Mailing-list

- A Google Groups mailing list / discussion group was setup for all participants
- All non-personal questions related to the course must be directed to the mailing list, not to the instructors personal e-mail
- All participants are strongly encouraged to initiate discussions both in the classroom and through the mailing list
- To add subscribe another e-mail to the list, send a request to

  galaxy-cefap-2014+subscribe@gmail.com

- If you subscribed a Google account, you can access the mailing list website and archive at

  https://groups.google.com/d/forum/galaxy-cefap-2014

Eating

- Closest cafeteria/restaurant: ICB-I
  - Self-service, charged by weight
- Alternative restaurant: FEA
- Other restaurants inside the campus
- Off-campus: many options

Nuvem USP

and a brief introduction to cloud computing

Robson Francisco de Souza

2014

What is “cloud computing”?

The computer industry started with big systems…

“There is no reason anyone would want a computer in their home”

Ken Olsen, 1977, president and founder of Digital Equipment Corporation (DEC)
arguing against the PC

Data center and mainframes

IBM PC

Apple and others

Personal computers

What is “cloud computing”?

Personal computing paradigm

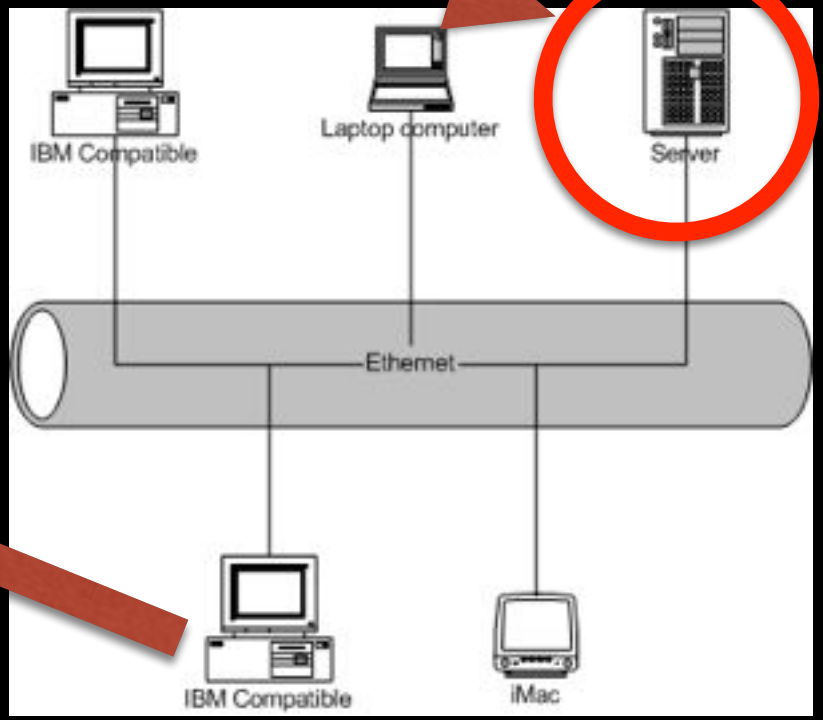
- All the tools and all data are stored and processed on the user’s computer
- My documents, my images, my music, my anything… right here!

What is “cloud computing”?

- Local area network: the return of the server
- My documents, my images, my music, my anything… right here but shared!

At least a subset of your data is now here!

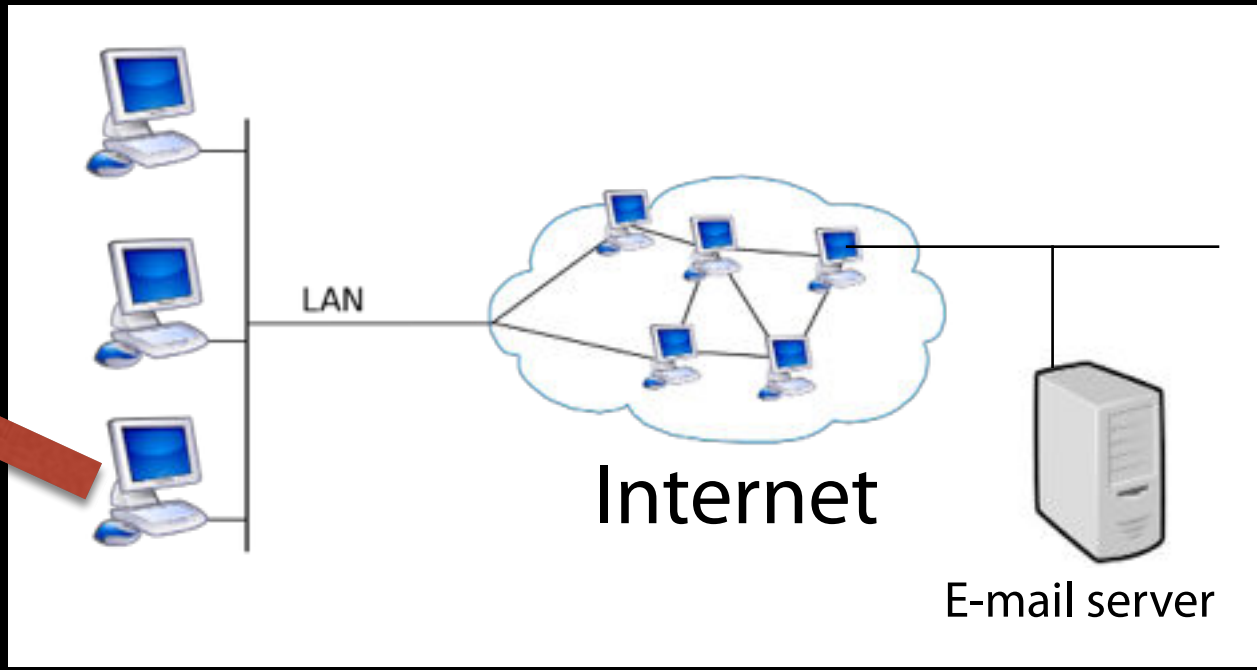
Your good old PC is still the same + a network card



What is "cloud computing"?

- Enters the Internet …
- My e-mail comes from outside… and I don't know or care how that happens: things are getting a little "cloudy"
- The e-mail server's location could be unknown as well

Your good old PC + network card



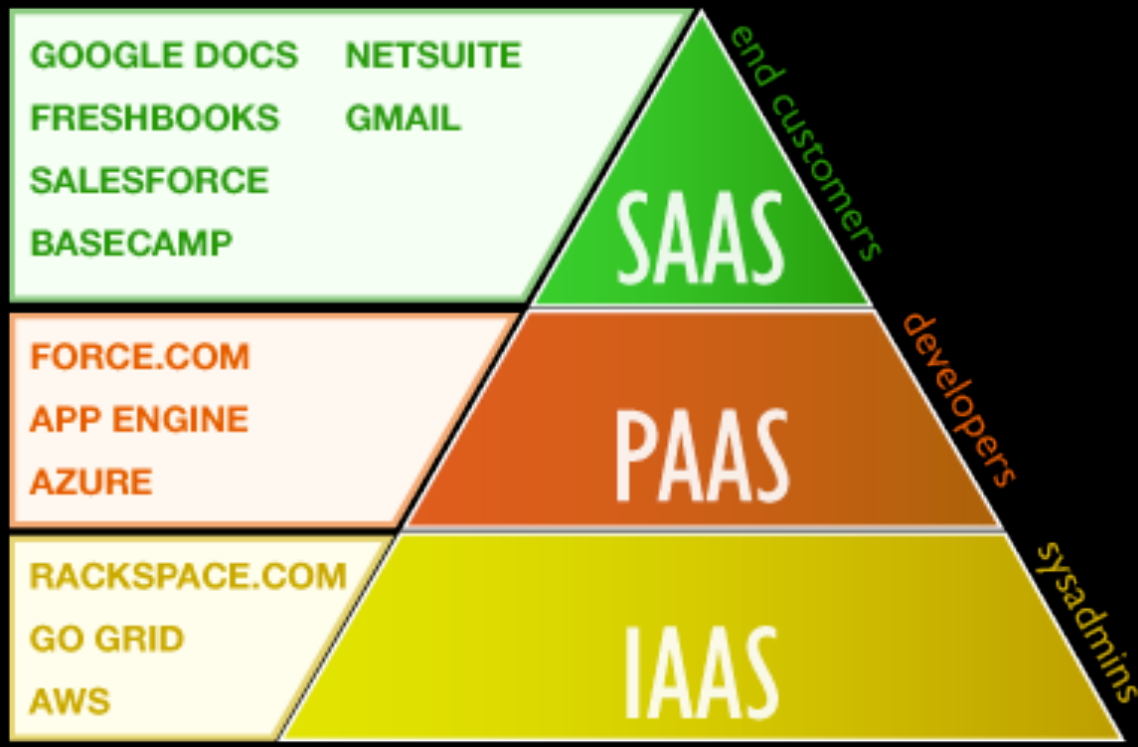
What is “cloud computing”?

- Enters the Web… and, sometime later, web applications
- One example is webmail. Using such a service
  - You no longer needed to install e-mail reading software on your computer
  - You can read your messages anywhere where
    - there is an Internet connection and
    - you can run a web browser
- With web applications, the service provider (GMail, Yahoo, etc) takes care of your data ⇨ software as a service (SaaS)

What is “cloud computing”?

- Software as a Service (SaaS)
  - Probably the first type cloud business model
  - Characteristics
    - No need for installation or upgrades
    - Data (user configuration, application data) may be kept in the server
    - Accessibility: anywhere
    - Usually serves multiple users

Approaches to Cloud Computing



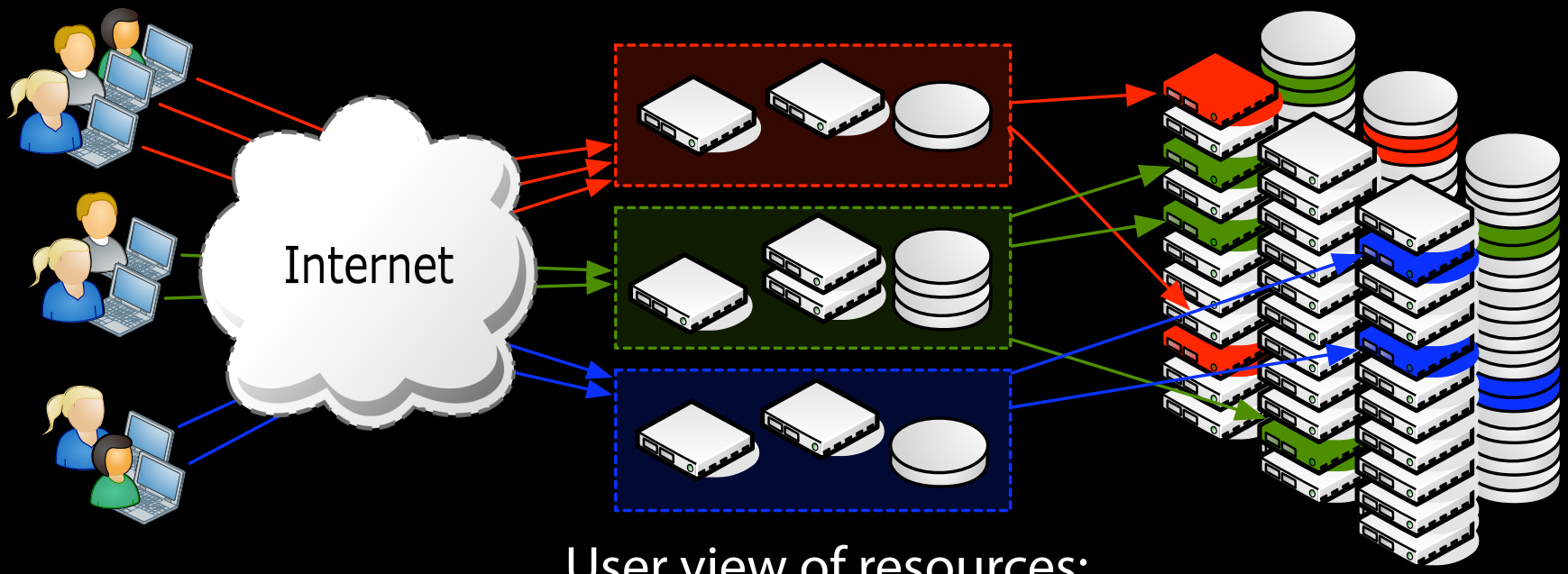
- Software as a Service
- Platform as a Service
- Infrastructure as a Service

Based on a slide by Enis Afgan (Johns Hopkins University)
illustration from FiberTown

http://blog.fibertown.com/2011/07/27/who-else-wants-to-understand-the-3-types-of-cloud-computing/

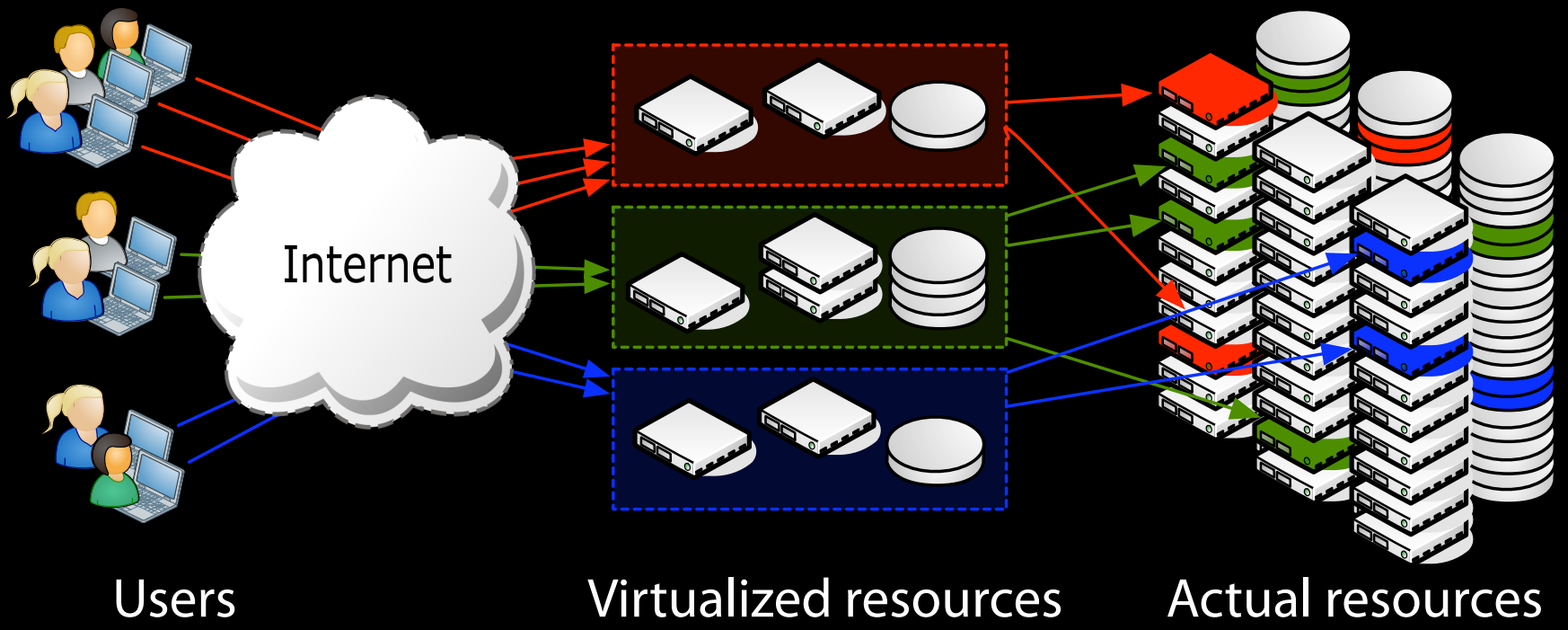
IaaS: Infrastructure as a Service

- The objective is to turn hardware into a service
  - Provide computing power (CPU time), storage and/or network
  - Users request the amount of resources they want: pay as you go
  - Scalability: no technical limitations on resources (self-service model)
- Ideal for anyone who can't spend time/money managing computing systems, like small labs and individual researchers

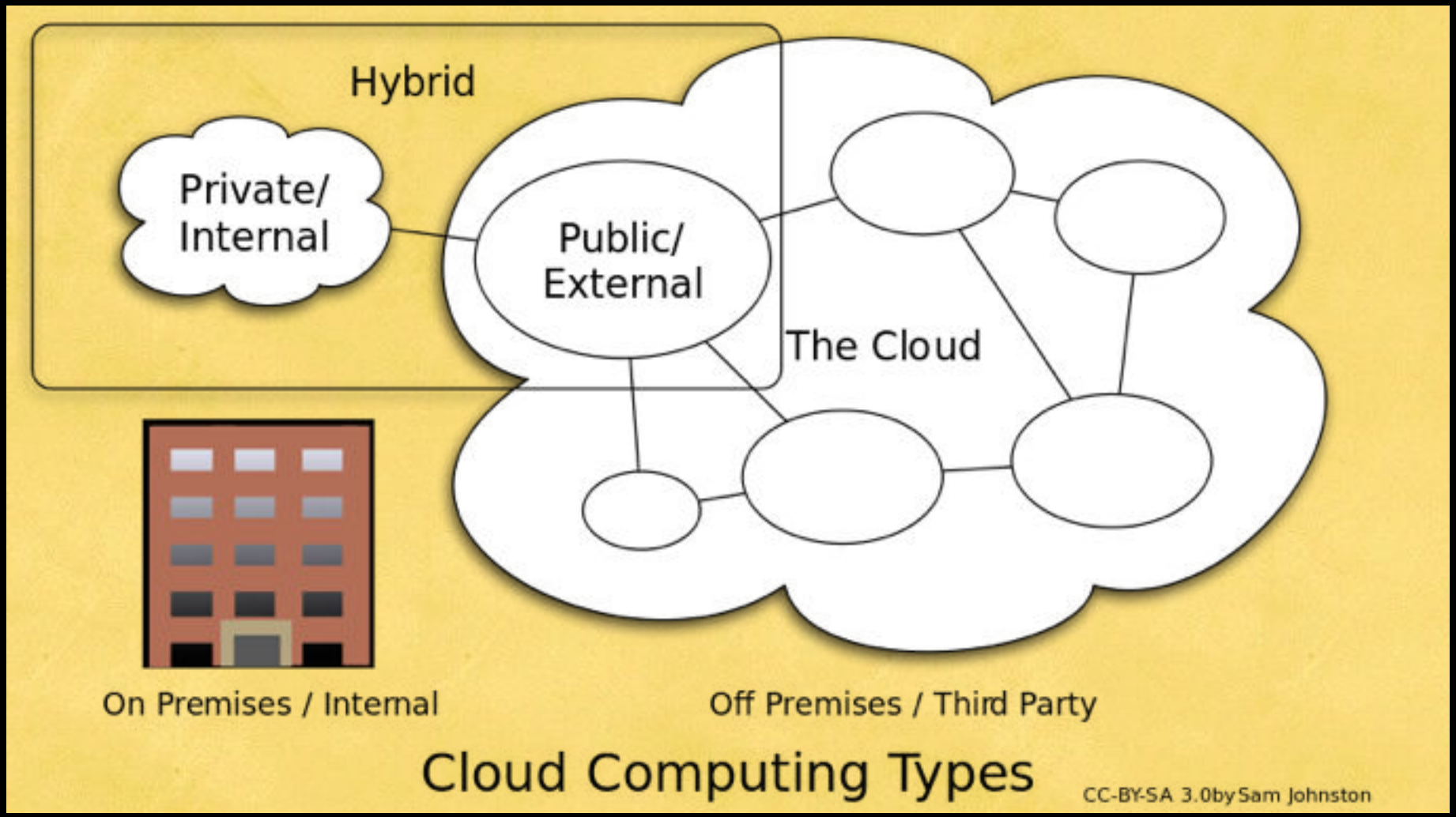


IaaS: Infrastructure as a Service

- How to turn hardware into a service?
  1. Buy physical hardware for users and connect to the internet
  2. Virtualization: allocate computer resources dynamically via software
- Quick demonstration: a virtual machine running Windows 7 on Mac OS X



Cloud Computing



What is “cloud computing”?

Data center and mainframes

IBM PC

Apple and others

Personal computers

Personal computer

Internet / Web

Virtualization

Cloud infrastructure

Cloud or No Cloud?

Pros

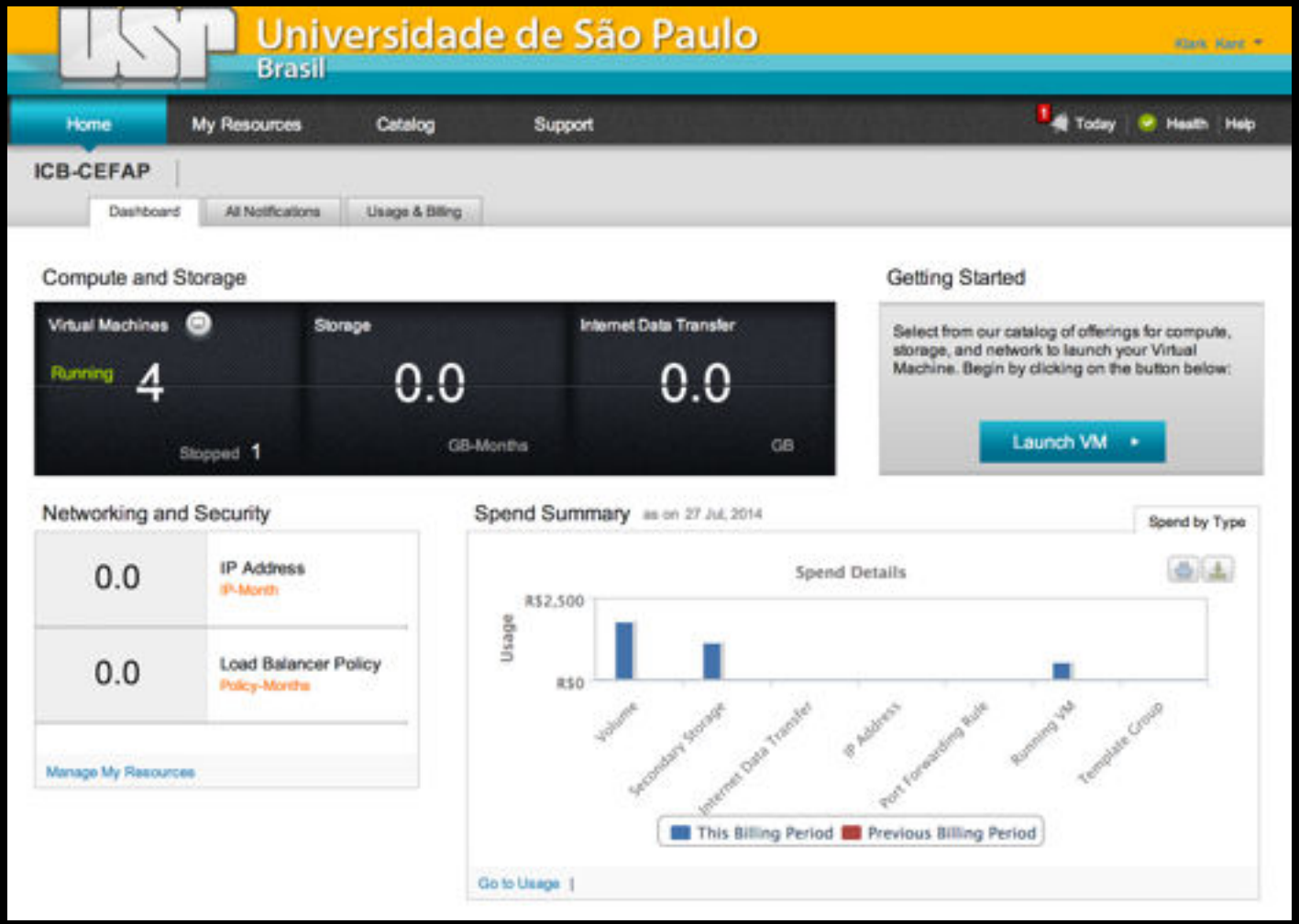
- Consumption based cost - cost reduction?
- Better utilization of resource
- Management done by cloud provider
- Faster deployment time
- Dynamic scalability

Cons

- Not a silver bullet
- Expensive for 24/7 use
- Offers scalability in terms of infrastructure, applications are still sequential
- The data transfer problem?
- Security?

IaaS at USP

http://nuvem.uspdigital.usp.br



IaaS at USP

- http://nuvem.uspdigital.usp.br
- Free private cloud: USP community only
  - 2 public IP addresses
  - 2 VLANs (virtual local area networks)
  - 100 virtual machines
  - 1000 backups / snapshots
  - 20 templates and/or ISOs
  - Windows server

IaaS at USP

Description	vCPU	Memory (vRAM)
Mini	1	512 MB
Web	1	1 GB
Standard	2	4 GB
Advanced	4	8 GB
High performance	8	16 GB
High performance with extra memory	8	32 GB

- Pre-installed OS on virtual hard disk
  - Linux: 20 GB
  - Windows: 100 GB
- Additional volumes of up to 2 TB

IaaS at USP

Physical infrastructure and current usage

- 576 servers
- 10.752 processing cores (CPUs)
- 368.640 graphical processing units (GPUs)
- 260 Terabytes of RAM memory
- 13 Petabytes of disk storage
- More than 2000 network interfaces (10 Gbps)
- 104 accounts
- 4700 VMs created (~30% capacity)

Data from Cyrano Rizzo (24/10/2013)
http://iptv.usp.br/portal/video.action?idItem=19268&idVideoVersion=15411

IaaS at USP

What does it offer for you?

- Elasticity: grow/shrink as you need
- Self-serving model
- Better connectivity

Need more computing power?

Laboratory of Advanced Scientific Computation (LCCA)

- Up to 10 dedicated vCPUs per VM
- 128 GB RAM

Perspectives

- Site-to-site VPN
- vGPU (CUDA)
- HPC

Data from Cyrano Rizzo (24/10/2013)
http://iptv.usp.br/portal/video.action?idItem=19268&idVideoVersion=15411

Galaxy on Nuvem USP

- For this course, Dave and Enis implemented the Galaxy VM models we will use!!!!
- It will require minimal computational expertise
- Hardware infrastructure is available for free for the USP community and this course's guests
- No need to perform any software installation!!!!
- Deploy a Galaxy instance in minutes!

Galaxy on the Cloud

- Enable execution of Galaxy on cloud infrastructures
  - Labs do not have to house compute resources
  - Support variable analysis data volume
  - Web-based Galaxy instantiation
- Goal is to keep Galaxy use unchanged but deliver flexibility and job performance improvement

Questions
&
Comments

Instructions for our practice:

http://goo.gl/nikgiU

Not using NuvemUSP? Try your own cluster in AWS.

Complete instructions available at http://usegalaxy.org/cloud