Virtualization & Cloud Activities

Dr Charles LOOMIS (LAL)

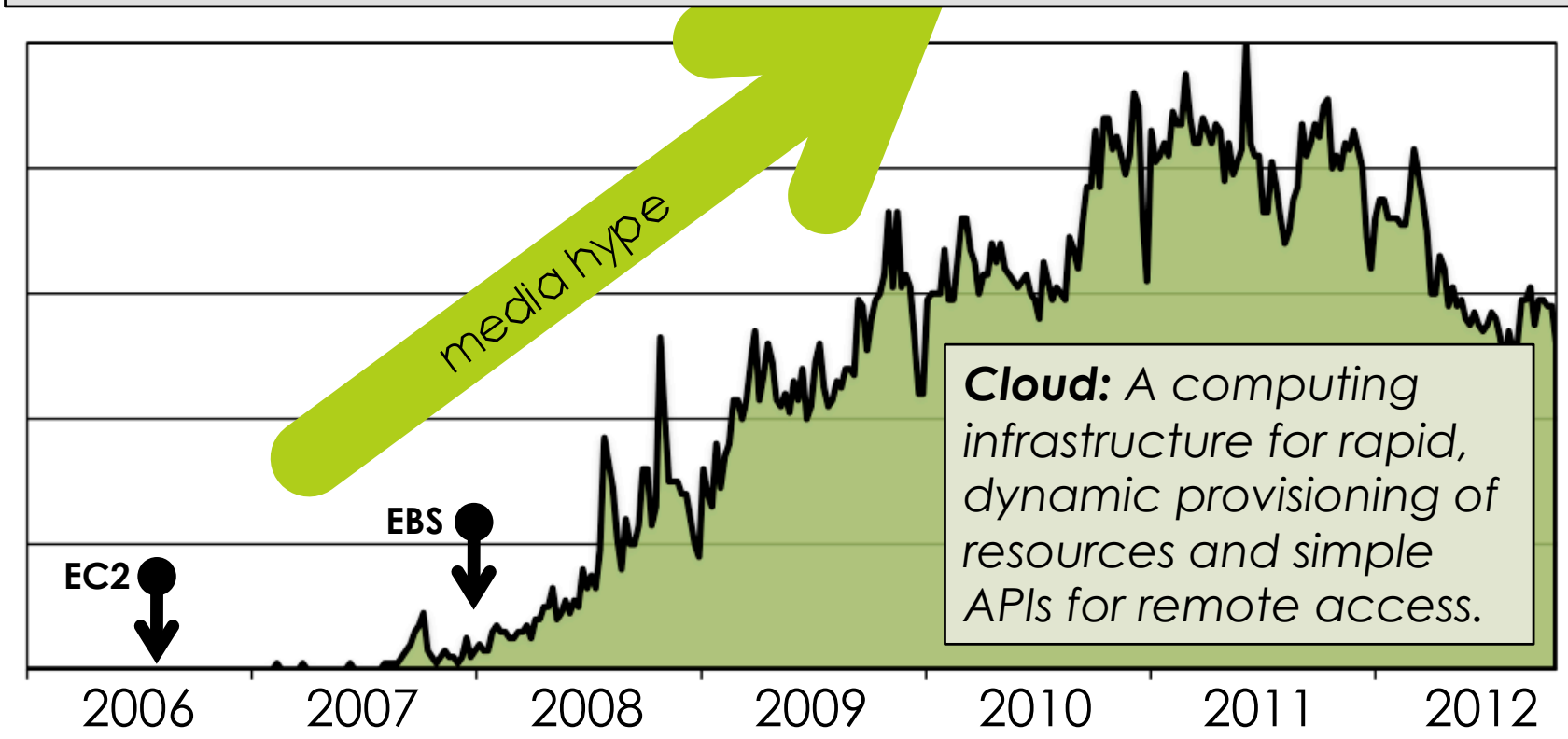
18 June 2014

Contents

- Introduction: cloud technology
- Survey of activities within IN2P3
- StratusLab: IaaS cloud distribution
- VirtualData: shared scientific computing services
- National and European Initiatives
- Conclusions

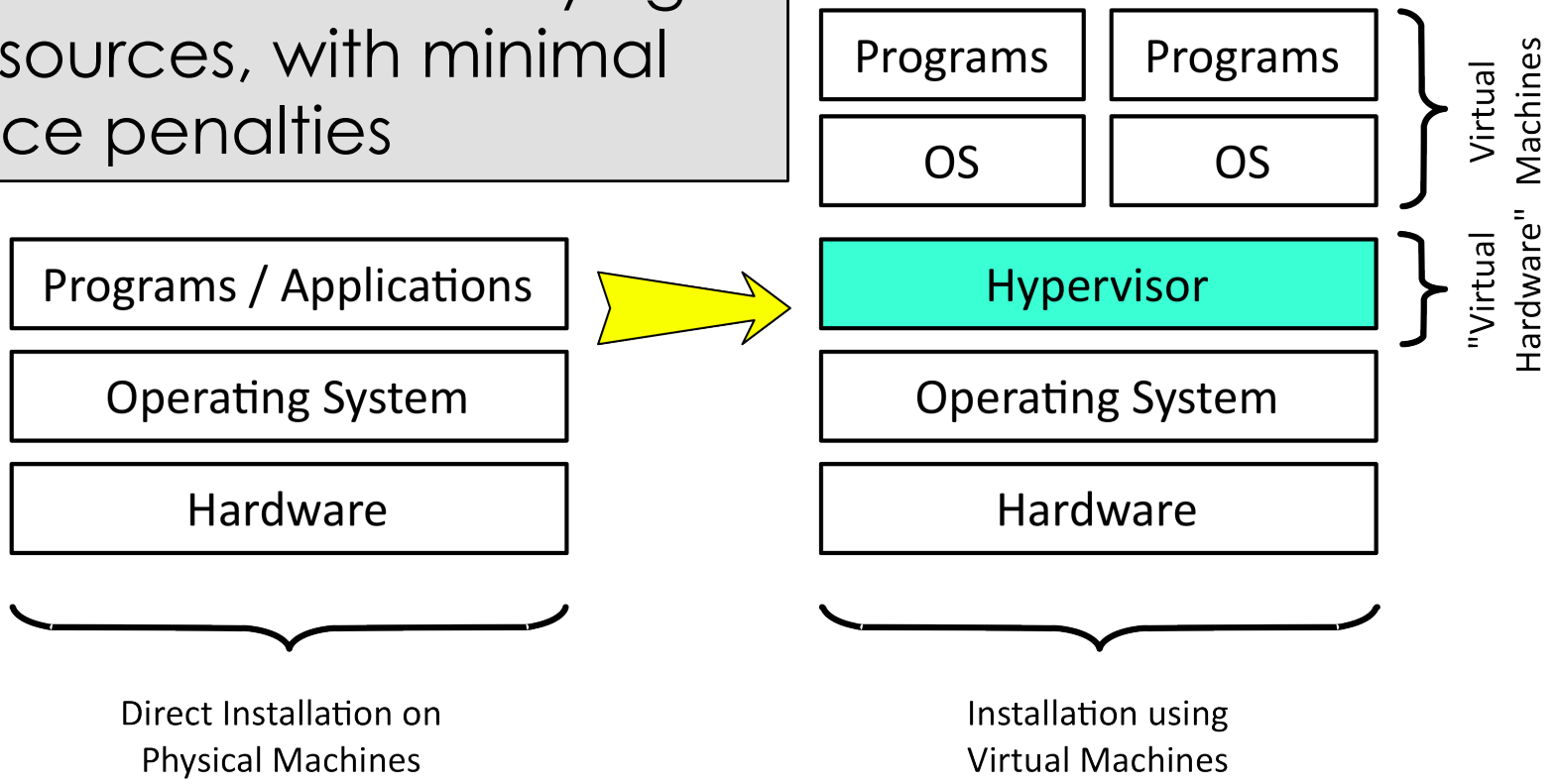
Cloud Hype

Is cloud technology useful for scientific computing?



Virtualization

Logical separation of services and applications from the underlying physical resources, with minimal performance penalties



Web Service Interfaces

SOAP

- Prioritizes easier service implementation by developers
- **Complex specifications, limited interoperability**
- Complex tooling
- RPC architectures only

REST, XMLRPC, …

- **Focus on easy access by service clients**
- Embraces HTTP protocol semantics
- Universal support from programming languages
- RPC and ROA possible

Simple access to, and integration of different services.

Excess Computing Capacity

Amazon

- Dimensioned to handle Christmas rush
- Idle machines/resources other times of year
- Monetize investment in these services
- *Allowed resources to be offered at excellent prices*

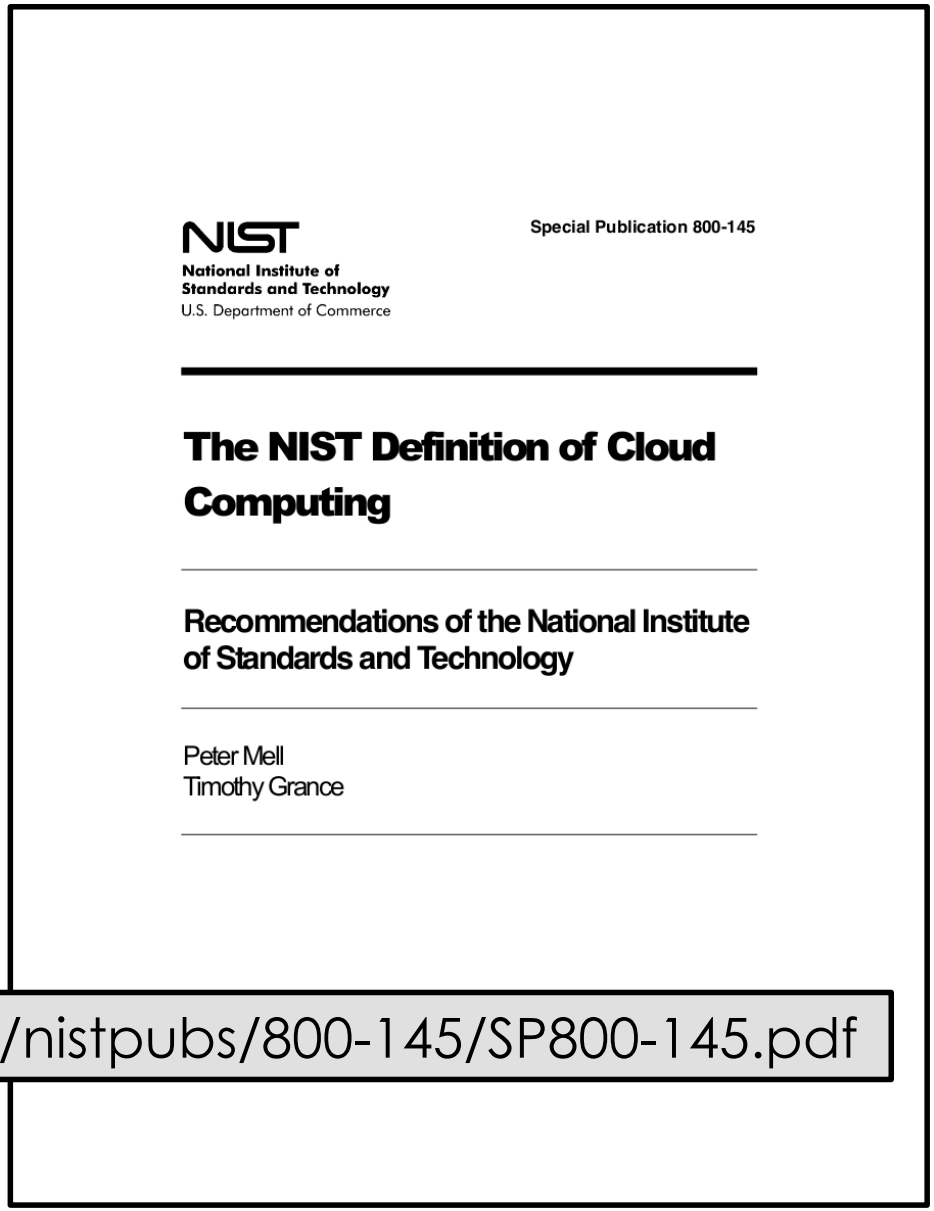
Data Centers

- Moved from monetizing existing investment to profit center
- Now Amazon and others have dedicated centers for the cloud!

Economies of scale of larger data centers offer potential cost savings.

What is a Cloud?

- NIST: Best Definitions
  - Essential characteristics
  - Service models
  - Deployment models
- Just 2 pages of text!


NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Special Publication 800-145

The NIST Definition of Cloud Computing

Recommendations of the National Institute of Standards and Technology

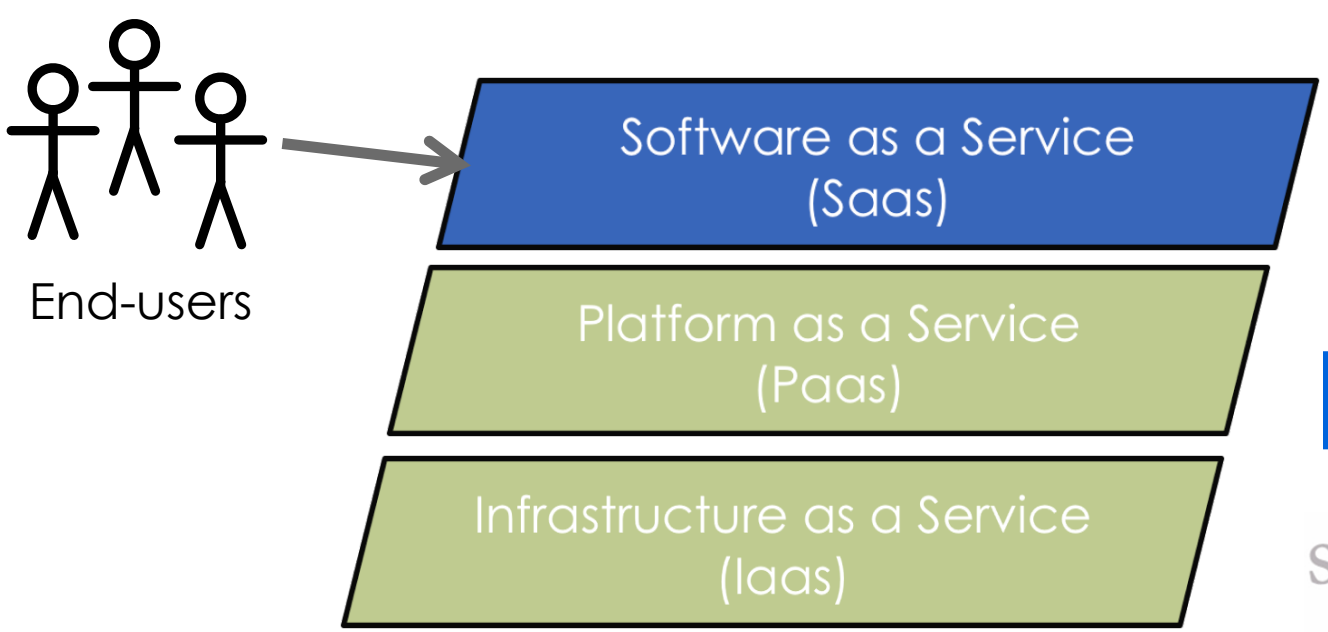
Peter Mell
Timothy Grance

http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

Service Models

What functionality does a cloud provide to users?

Software as a Service



Doodle®

salesforce.com
Success On Demand.

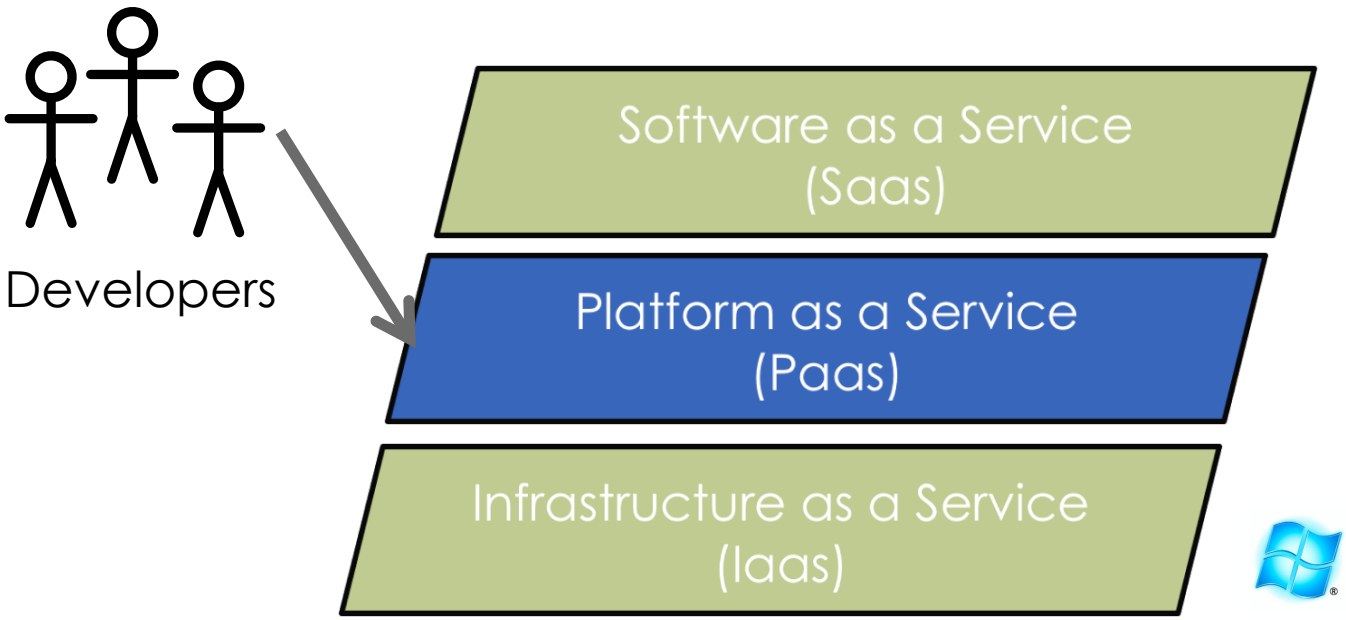
Advantages

- Simple use: web interface; no software installation
- Very accessible: laptop, smartphone, …

Disadvantages

- Data questions: access, ownership, reliability, etc.
- Integration (mash-up) is often difficult

Platform as a Service



Google App Engine

GridGain CLOUD COMPUTING

Windows Azure

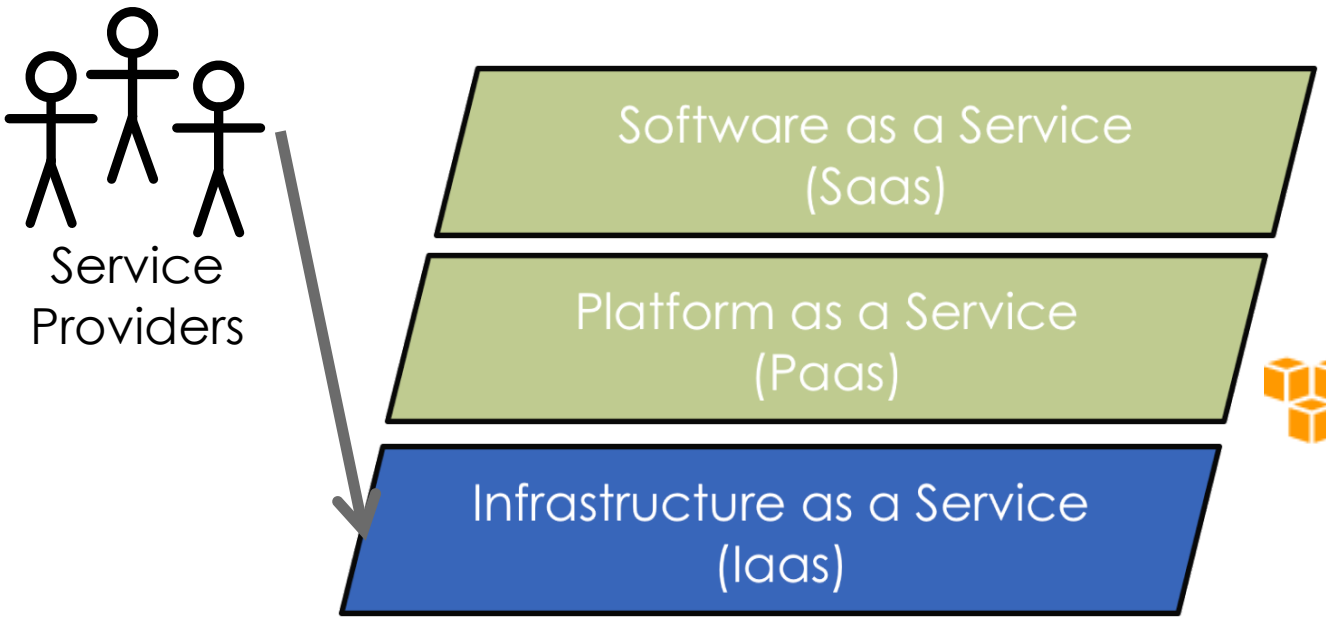
Advantages

- Programmers can forget about low-level plumbing
- Load balancing; automatic failover, etc.

Disadvantages

- Restricted number of languages
- Applications not portable between different providers

Infrastructure as a Service



ElasticHosts
Flexible servers in the cloud

GOGRID
A ServePath Company

amazon web services™

flexiant™
utility computing on demand

Advantages

- Full control over machine; access as "root"
- Customized environment, down to operating system

Disadvantages

- Multiple APIs
- No std. machine format
- VM creation is time-consuming and difficult

Deployment Models

Who are the targeted users?

Deployment Models: Private

- Single administrative domain, limited number of users
- Informal resource allocation via ‘hallway conversations’
- E.g. running cloud for site services, managed by group of system administrators

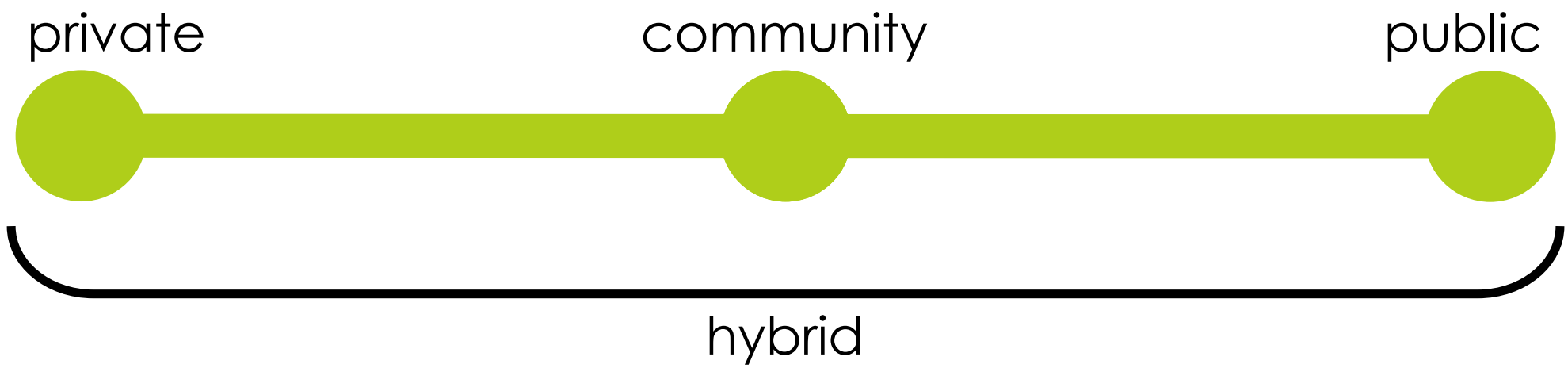
Deployment Models: Public

- People outside of cloud provider's institute; aimed at the general public
- Resource allocation by payment
- E.g. Amazon Web Services (EC2, S3, …)

Deployment Models: Community

- Different institutes but with common interests and procedures
- ‘Horse trading’ used to allocate resources between individuals and groups
- E.g. bioinformatics community in France

Deployment Models: Hybrid



- Combination of other deployment models
- Used for redundancy and/or peaks in demand
- "Sky" computing, brokered federation, bursting, …

Clouds for Scientific Computing

Are there benefits in **deploying** a cloud?

Maximize “Profits”

- Increase occupancy through broader user base
- Easily consolidate services on fewer machines
- More easily manage diversity of computing environments

Simplified Maintenance

- Hardware: services become independent of underlying physical machine
- Cloud services: single set of services for managing access to computing resources
- Scientific platforms: become separate layer deployed, controlled and managed by domain experts

Clouds for Scientific Computing

Are there benefits in **using** a cloud?

Customized Environment

- Operating systems suited to your application
- Your applications preinstalled and preconfigured
- CPU, memory, and swap sized for your needs

Dynamic Provisioning

- New storage and compute resources in minutes (or less)
- Resources freed just as quickly to facilitate sharing
- Create temporary platforms for variable workloads

Flexible Service Deployment

- Deploy user-level, network-accessible services
- Create domain-specific analysis platforms
- Redundant, scalable services easily deployed

Agile Scientific Computing

- Faster analyses
  - Quick deployment of preconfigured virtual machines
  - Lower learning curves via customized VMs/services
- Fewer failures
  - Less interference between processes on machine
  - Avoiding misconfigurations and incomplete OS installations

Virtualization & Cloud Expertise

What are IN2P3 laboratories currently doing with cloud?

Survey

- Recently asked IN2P3 labs about their use of:
  - Virtualization
  - Commercial cloud services
  - Scientific cloud infrastructures
- Responses
  - 15 from total of 21 IN2P3 institutes

Virtualization

- **All institutes** use virtualization technologies.

Virtualization Tech.	
KVM	7
Xen	4
XenServer	2
ESXi/vSphere	8
HyperV	4
ProxMox	3
VirtualBox	1
TOTAL	**29**

Most laboratories use more than one virtualization solution!

Size of Infrastructures	
physical machines	2—30
virtual machines	2—80

Use of Platforms		
Laboratory Svcs.	10 of 15	DHCP, Web, LDAP, AD, Print Servers, DNS, NFS, mail, …
Scientific Svcs.	4 of 15	Mokka, CMS, AGATA, ALTO, Fermi, LSST

Commercial Cloud Services

- None of the institutes use these services and
- None have plans to do so in the future!

Scientific Clouds

- 5/15 have one or more cloud infrastructures
- 4/15 plan will have within 12 months

Deployment Model	
Private	4
Community	1
Public	1

Cloud Software	
OpenStack	2
StratusLab	3

Size of Infrastructures	
physical machines	10—22
virtual machines	2—60

Use of Platforms		
Laboratory Svcs.	8 of 13	Grid UI, Web, LDAP, Rhodecode, Print Servers
Scientific Svcs.	3 of 13	Fermi, Local collaboration

IN2P3 Expertise

- Significant, production use of virtualization and cloud
- Expected doubling of number of institutes with clouds

Local Initiatives

What are we doing at LAL and UPSud?

StratusLab

- StratusLab: software distribution for creating IaaS clouds
  - Darn Simple Cloud: focus on simple installation and use
  - Standard cloud services: compute, storage, network
  - Advanced appliance sharing services and tools
  - Created through EU project; now open source collaboration
- Software evolution
  - Implementation of new scalable architecture
  - Adoption of the CIMI interface as native API
  - Improving tools for cloud administrators

stratuslab

StratusLab Clouds at LAL

- Private cloud infrastructure
  - Used for range of laboratory services
  - Also used for tests of new deployments

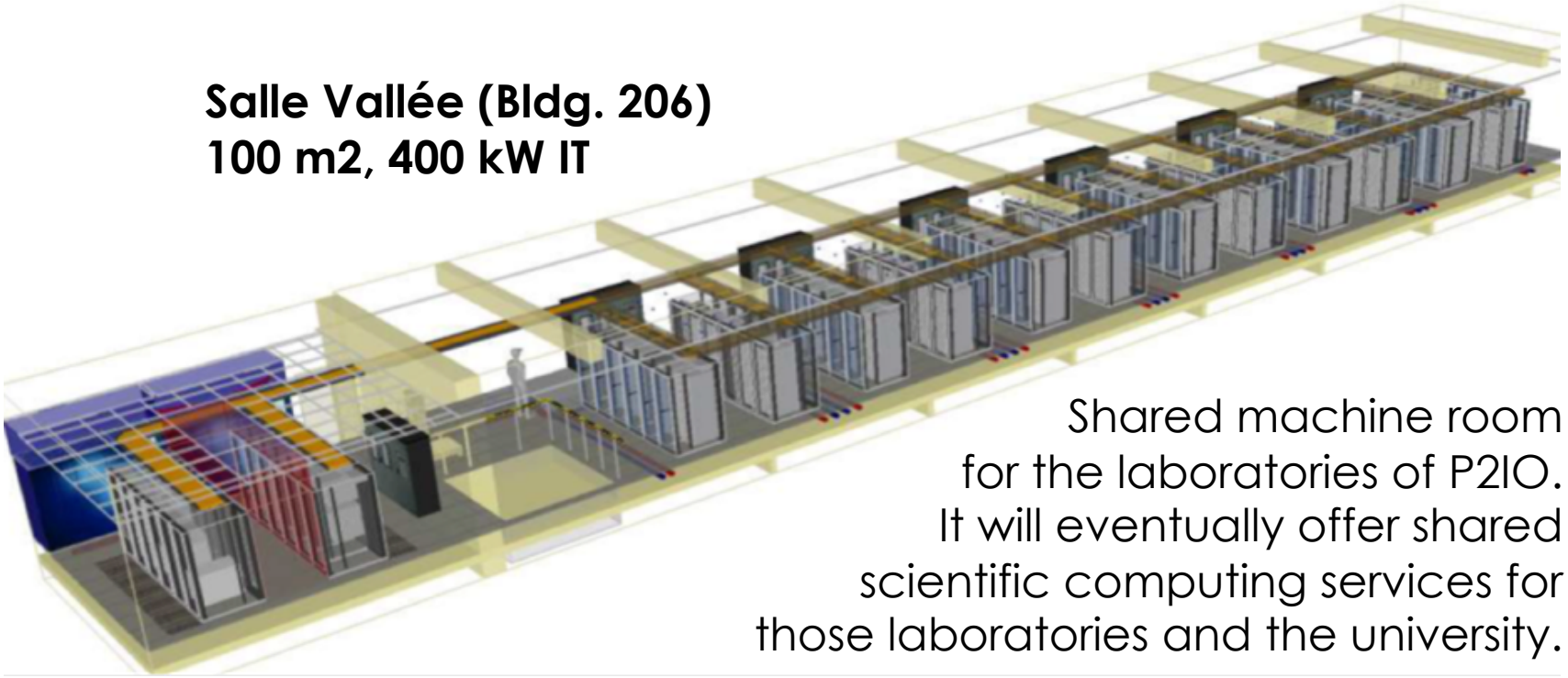
Private Cloud	
machines	2
CPU cores	96
Memory (GiB)	72
Storage (TiB)	300

- Public cloud infrastructure
  - Evolved from demonstration cloud for EU project
  - In continuous production since the end of 2010

Public Cloud	
machines	16
CPU cores	440
Memory (GiB)	772
Storage (TiB)	15

VirtualData

Salle Vallée (Bldg. 206)
100 m2, 400 kW IT



Shared machine room
for the laboratories of P2IO.
It will eventually offer shared
scientific computing services for
those laboratories and the university.

StratusLab public cloud infrastructure will probably evolve into the P2IO cloud service.

Cloud: Collaborative Platform

Local Users	
ABgenomica	An SME that provides appliances and services for bioinformatics
AppStat (LAL)	Development and analysis of machine learning algorithms
CSNSM	Use of the cloud to create temporary infrastructures for NARVAL tutorials
IGM	Scientific analyses for molecular genomics
Grid Observatory	Collection and analysis of traces of grid and cloud utilization
others…	

Further Afield

What links are there to national and international initiatives?

FranceGrilles

- Cloud technical group
  - Brings together administrators, developers, and users
  - Forum for exchanging information and expertise
  - Support others deploying and using cloud technologies
- Federated cloud Proof-of-Concept
  - Create a federated cloud demonstrator in France
  - Choices: grid certificates, CloudInit, Marketplace, …
  - Validate this with a real application
  - Ensure infrastructure works in FG operational environment

EGI

- EGI Federated Cloud Task Force
  - Investigate federation of cloud resources
  - Take advantage of existing operational infrastructure
  - Bring cloud resources into production infrastructure

- Technology choices
  - Standard API for all resources: OCCI
  - Federated identity: grid certificates
  - Appliances: CloudInit, AppDB

Helix Nebula

- Science cloud for Europe
  - Brings together large European cloud service providers
  - Central "marketplace" HNX to select and to use all service offerings
- Technical choices
  - No common API; use plugins to bridge different provider APIs
  - No federated identity; users have separate identities for each cloud
  - Broker provides unique interface to all clouds and also provides higher-level services for multi-cloud deployments

H2020 Projects

- Cyclone
  - Complete management of complex multi-cloud applications
  - Integration of advanced network features into cloud software distributions
  - End-to-end application security including communications, data storage and data access
- Others?
  - Helix Nebula project in e-infrastructures call?

Cloud Technology Challenges

What are the areas that need development?

Cloud & Big Data

- Legal constraints for some types of data
  - EC working with member states to harmonize rules
  - Protection guarantees provided by some public clouds
  - "National" clouds avoid cross-border constraints
- Reasonable network access
  - Large datasets require large bandwidth
  - Co-located computing only partial solution
  - Commercial world does not enjoy fast RENATER network

Cloud & Big Data

- Data curation
  - Long-term storage of data → hierarchical storage
  - Preservation of data for access by others → metadata also!
- Large volumes
  - Ability to store data affordably
  - Applications must expect errors in stored data
- Data access model(s)
  - Huge variety of different access models
  - Cassandra, Hadoop, MySQL, SRM, …

Appliance Management

- Understanding the contents of a given appliance
- Easily recreate appliances for software updates, etc.
- Securely share appliances between users and groups

- StratusLab Marketplace is a step in this direction

"Academic" Cloud Challenges

- Scheduling
  - Initial placement of machines
  - Migration of machines
- Good behavior: reducing hording of resources
  - Commercial providers use financial incentives
  - What can be done for academic environments?
- "Infinite" capacity
  - Amazon can be "infinite"; LAL (CNRS, …) cannot be
  - How to maintain illusion with finite resources?

Multiple Cloud Infrastructures

- Combining cloud infrastructures
  - Aggregation, brokering → multi-cloud deployments
  - Federation → identity management, standard APIs
  - Bridging private and community/public clouds
- Location is a feature!
  - Geographical service redundancy
  - Optimizing latencies based on client location
  - Simplified remote backups
- Cloud brokers like SlipStream or CompatibleOne make use of multiple cloud providers manageable.

Conclusions

Conclusions

- Extensive use and experience with virtualization for laboratory and scientific services.
- Significant (and growing) experience with cloud technologies and platforms.
- Cloud technologies and platforms provide a strong focus for multi-disciplinary collaboration.
- Making cloud technologies even more pervasive in scientific computing requires overcoming some interesting challenges.

Questions...