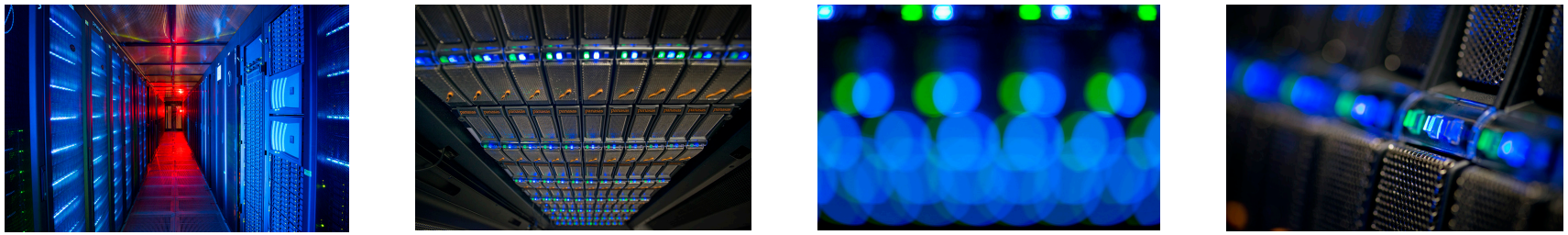


JASMIN (STFC/Stephen Kill)

Experiences and challenges in the development of the JASMIN cloud service for the environmental science community

ECMWF Visualisation in Meteorology Week, 28 September 2015

Philip Kershaw, CEDA Technical Manager

Victoria Bennett, Jonathan Churchill, Martin Juckes, Bryan Lawrence,

Cristina del Cano Novales, Sam Pepler, Matt Pritchard, Matt Pryor, Ag Stephens

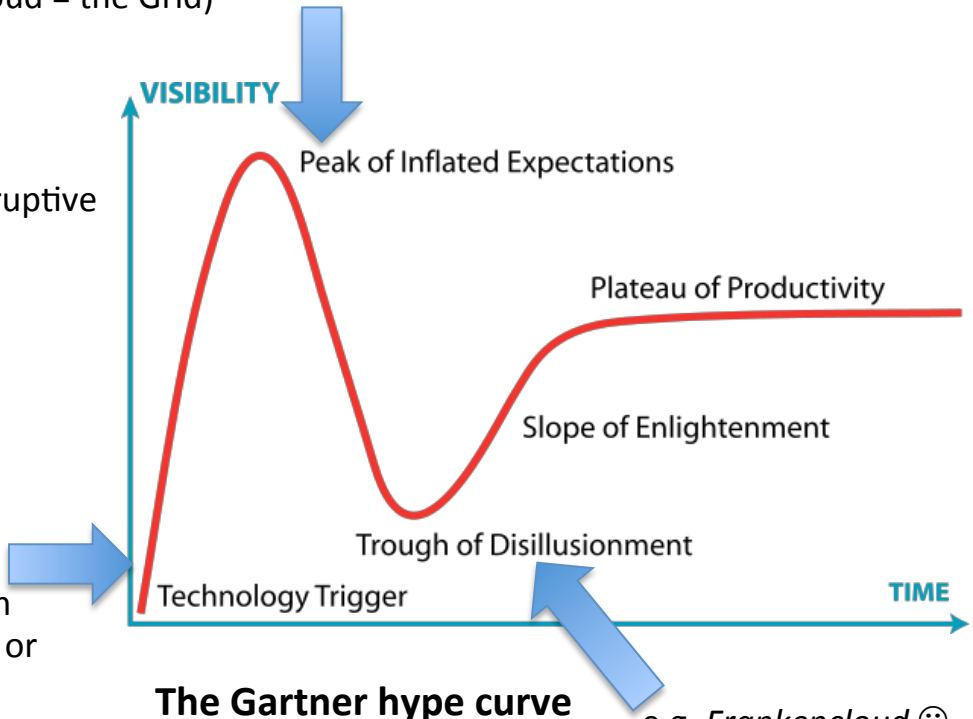
Introduction

- What is cloud?
- How can it be applied for science applications
  - Touch on relationship with HPC, Grid
- Practical experience building a community cloud for JASMIN
- Example application: IPython Notebook
- Challenges and next steps

The cloud and the hype

Hype, ignorance, fear, misunderstanding and conflation with what has gone before (e.g. cloud = the Grid)

An enabling and disruptive technology



The Gartner hype curve

Nb. Technology driven development vs. user or science driven

- Different communities and application domains have reached different places on the curve
- Cloud for science a fast evolution
  - Magellan Report 2011

http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Magellan_Final_Report.pdf

e.g. *Frankencloud* ☺

http://www.entrepreneur.com/article/247140

Understand in order to exploit: a cloud definition

"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction." – **NIST SP800-145**

5 essential characteristics

- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured service

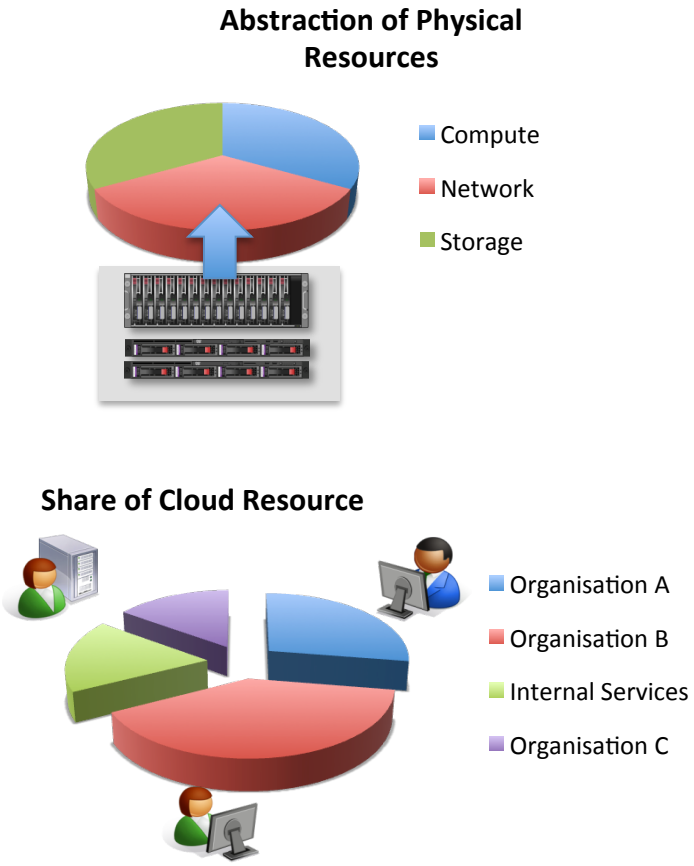
3 service models

- IaaS (Infrastructure as a Service)
- PaaS (Platform as a Service)
- SaaS (Software as a Service)

4 deployment models

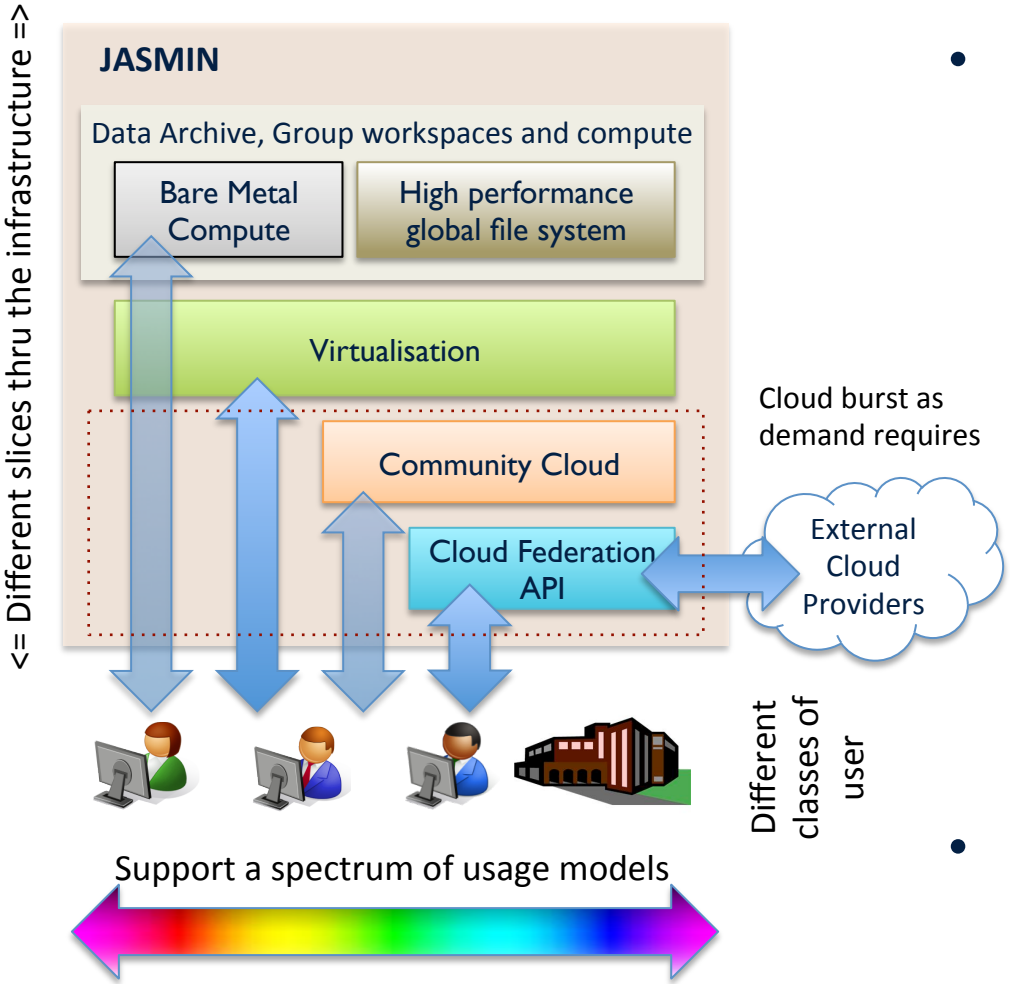
- Private cloud
- Community cloud
- Public cloud
- Hybrid cloud

How can Cloud help the Research Community?



- Address Big Data problem
  - Bring users to the data
  - Potential for near-limitless compute
- Long-tail research
  - Provide data analysis downstream of primary production in a way that is **customised** for researchers e.g. virtualised desktops
- Resource pooling
  - Divide up resources easily amongst different tenant research groups

JASMIN and Cloud



- A big data analysis facility for the environmental sciences
  - 16 Petabytes high-performance disk
  - 4000 computing cores
    - (HPC, Virtualisation)
  - High-performance network design optimised for i/o throughput
  - Virtualisation
  - Cloud - 200 socket Vmware vCloud licence (100 servers, 1600 cores)
- A *combination* of capabilities deliver what is needed

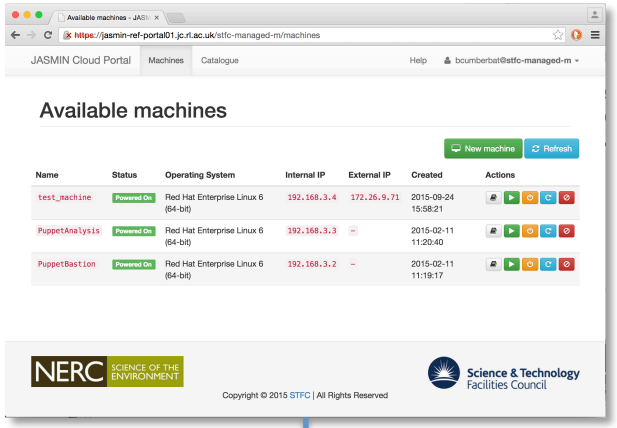
CEMS
Climate, Environment & Monitoring from Space

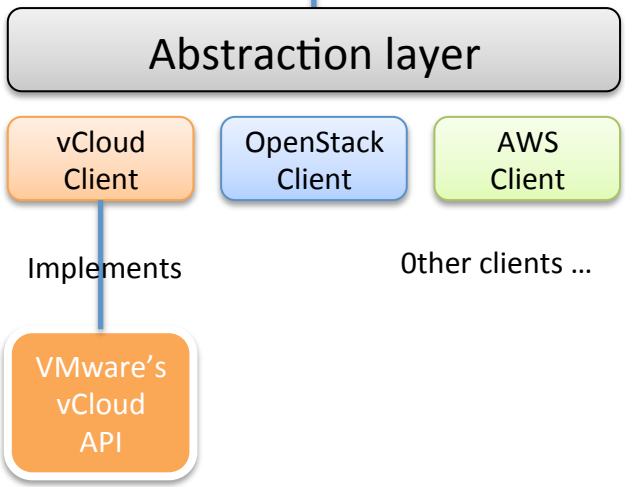
CEMS and JASMIN I: first steps with Cloud

- Deployed a private cloud based on VMware *vCloud Director* in common JASMIN-CEMS environment
- Goal: self-service configurable VMs for scientific analysis next to the data
- But,
  - vCloud web portal too complicated for external users
  - It couldn't be locked down sufficiently for deployment alongside the data archive and sensitive services
  - Demand for processing with batch compute which didn't need the flexibility of cloud
- Solutions
  1) Build a custom cloud portal
  2) The *Inside-outside* project

1) Custom Cloud Portal

JASMIN Cloud Management Interface



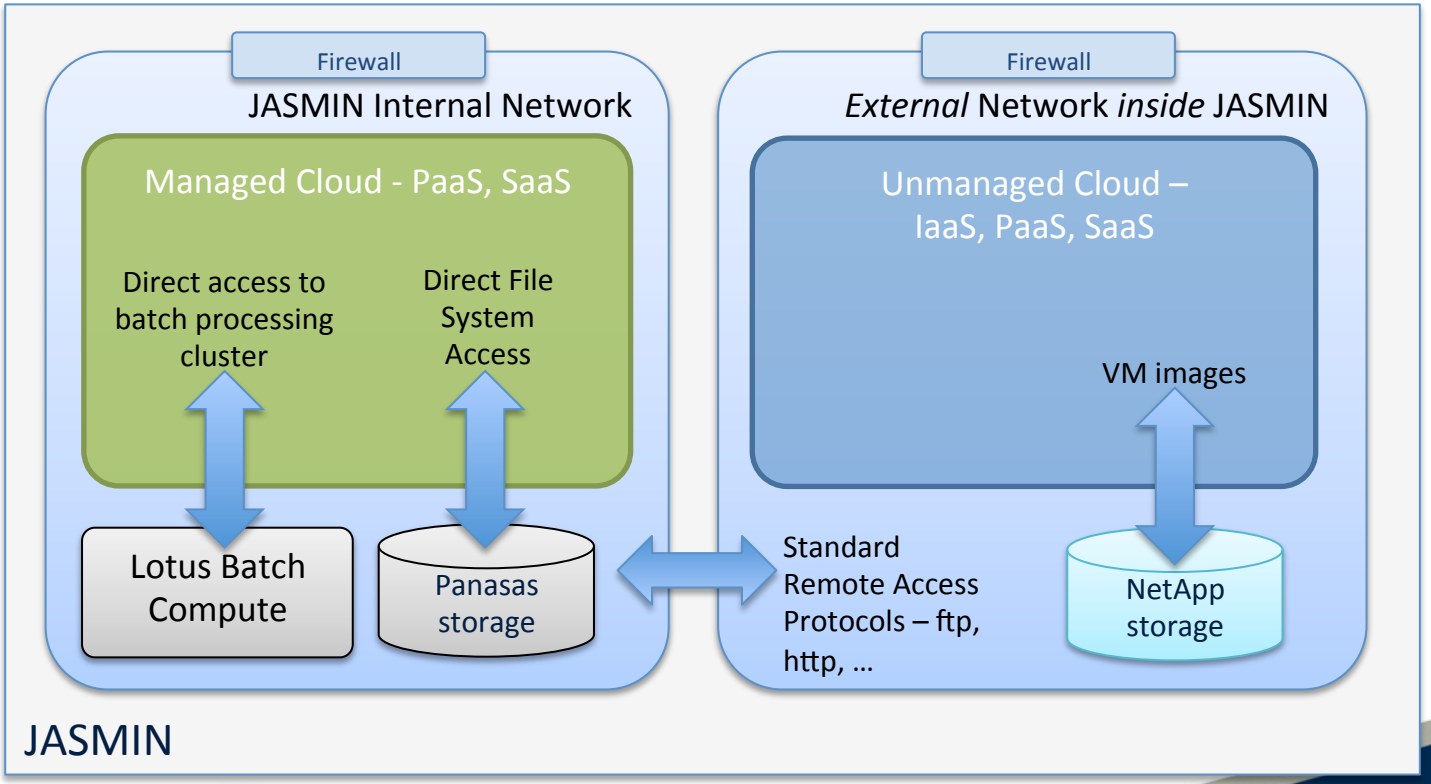


- Building on top of vCloud
- Keep it simple: provide just enough functionality to provision and configure VMs
- Right tools for right users: scientists, developers and administrators
- Abstraction from vCloud also provides a route to cloud federation / bursting
- Thin or thick client?

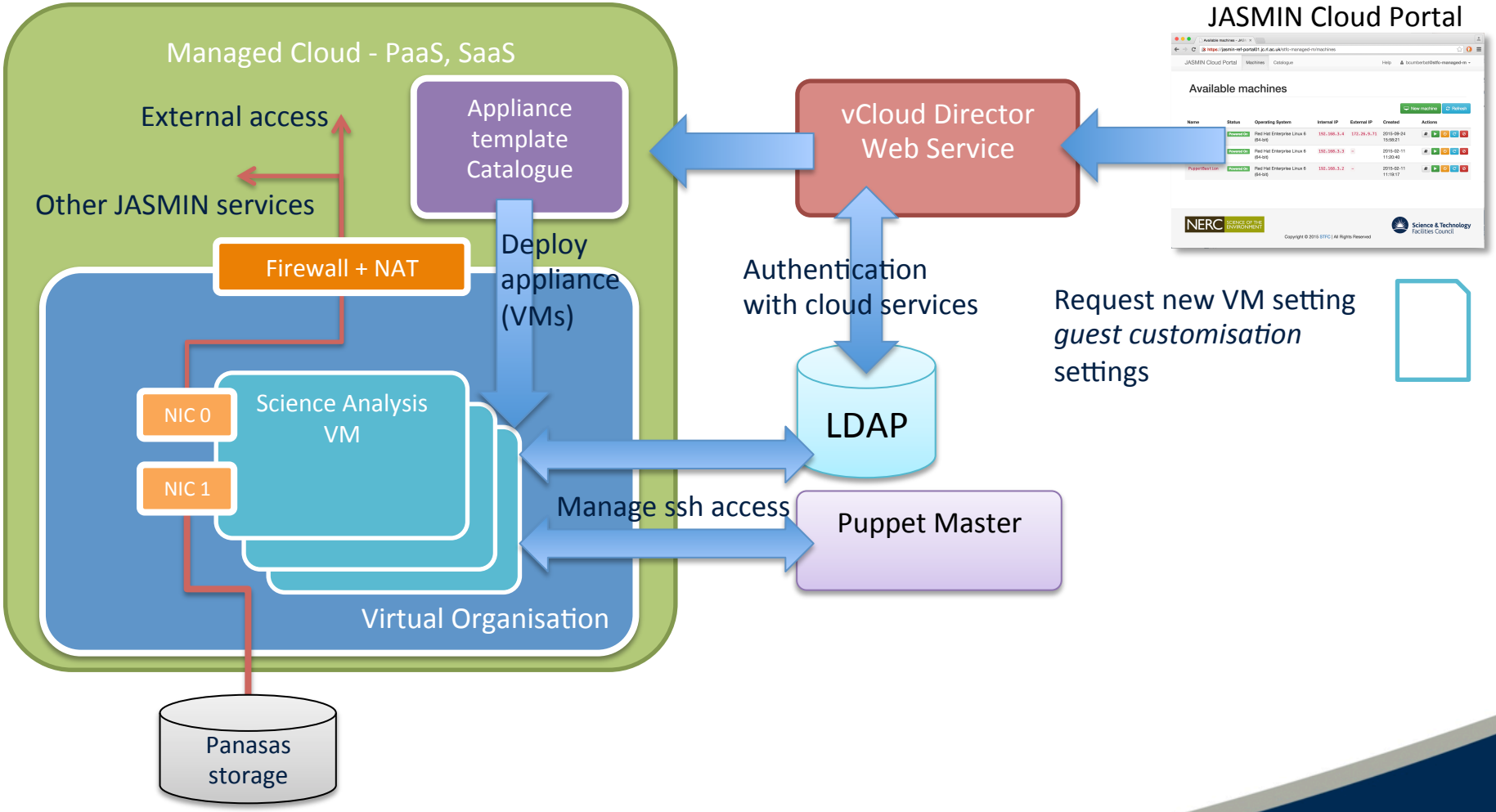
2) The *Inside-Outside* Project

Create,

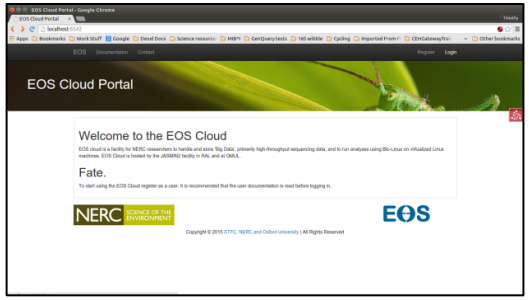
- an isolated part of the network **inside** JASMIN
- that would give users all the freedom they would have **outside** but
- the benefit of good bandwidth to the data archive.



Tenancy management and the VM deployment workflow



Pooling RAM with EOS Cloud

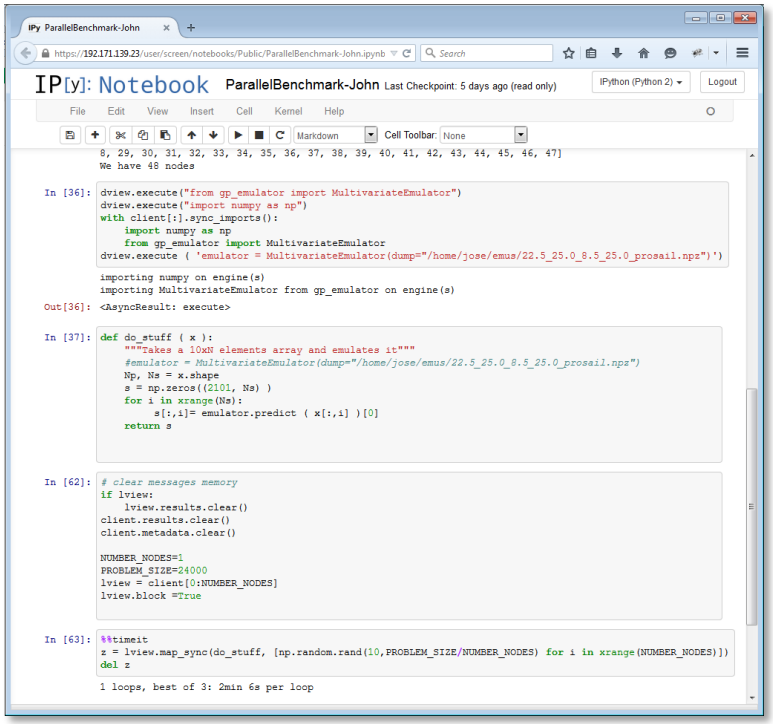


Cloud BioLinux

- Desktop as a Service for environmental genomics hosted on JASMIN Cloud
- Exploits key characteristics of cloud: pooling and elasticity
- The problem: bioinformatics apps are memory hungry
- Solution: using virtualisation seamlessly share access to a ‘fat’ node with 512 GB RAM in the tenancy
- A token system allows users to boost their VM to use the additional memory for a metered period

Credit Tim Booth, NERC CEH

Example Cloud-hosted Application: IPython Notebook

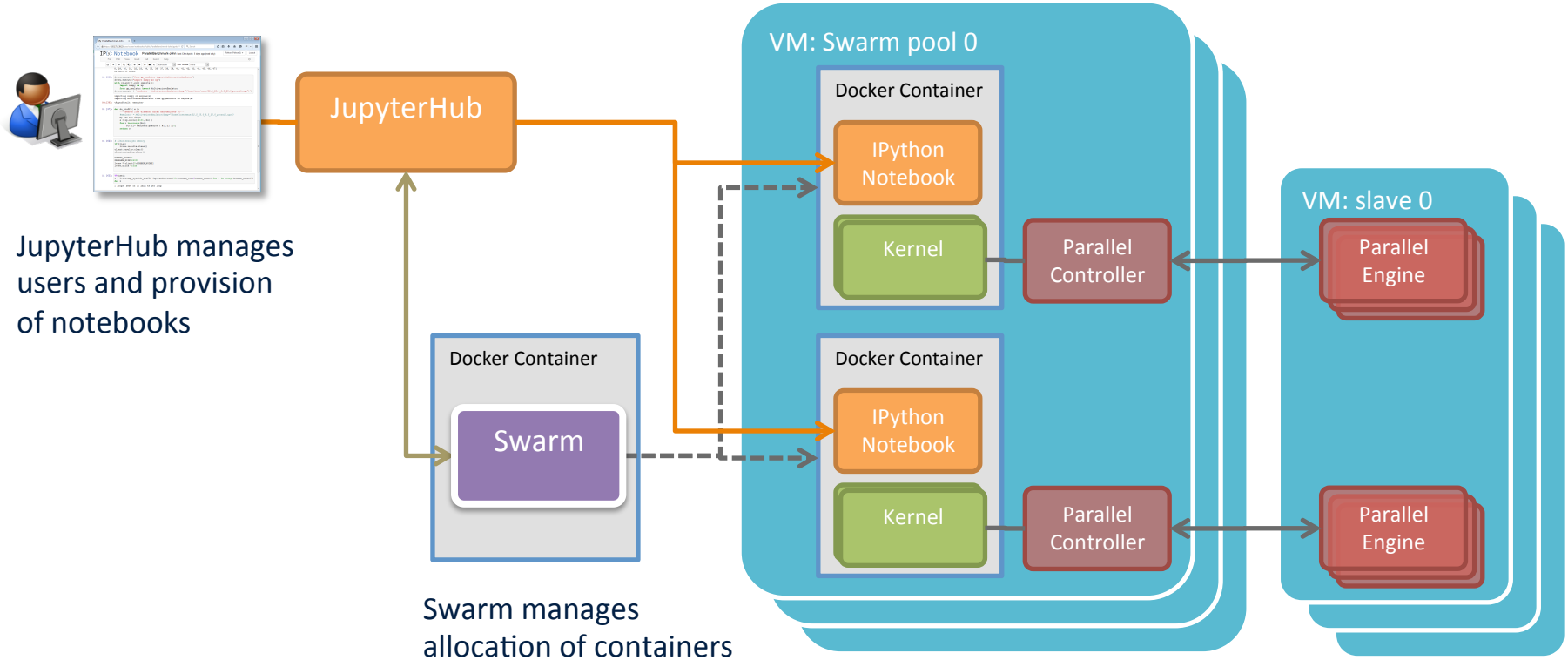


- Provides Python kernels accessible via a web browser
- Sessions can be saved and shared
- Trivial access to parallel processing capabilities – IPython.parallel
- New JupyterHub allows multi-user and notebook management

- Opportunity to open a middle ground in the application space between
  - batch compute / command line access: powerful but hard
  - web portals: easy to use but less flexible

- OPTIRAD: ESA-funded project, land surface data assimilation algorithms via notebooks on JASMIN cloud

JupyterHub, Swarm and Containers



JupyterHub manages users and provision of notebooks

Swarm manages allocation of containers for notebooks

Notebooks and kernels in containers

Nodes for parallel Processing

Conclusions

- Experiences from project delivery
  - Importance of key skills
  - Cross-cutting team spanning, developers, sys admins, DevOps
  - Effective linking of hardware deployment, cloud middleware and application layer development
  - Documented repeatable processes for operations
- Futures
  - Challenge: how to bridge together different types of resource and service seamlessly whilst preserving performance and user segregation
  - Effect and influence of new technologies: containers, object stores

Further information

- JASMIN and CEDA:
  - http://jasmin.ac.uk/
  - http://www.ceda.ac.uk
- JASMIN paper (Sept 2013)
  - http://home.badc.rl.ac.uk/lawrence/static/2013/10/14/LawEA13_Jasmin.pdf
  - Cloud paper to follow soon
- @PhilipJKershaw