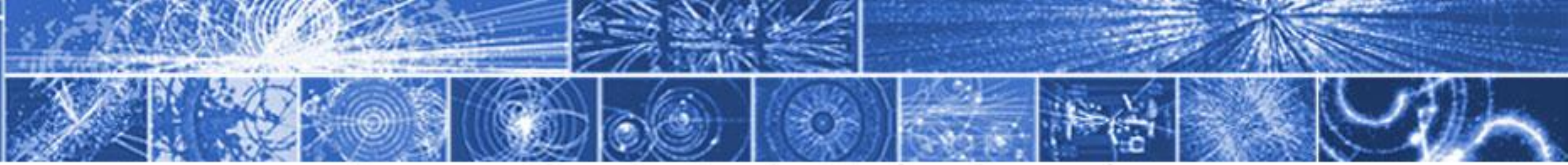


CERN

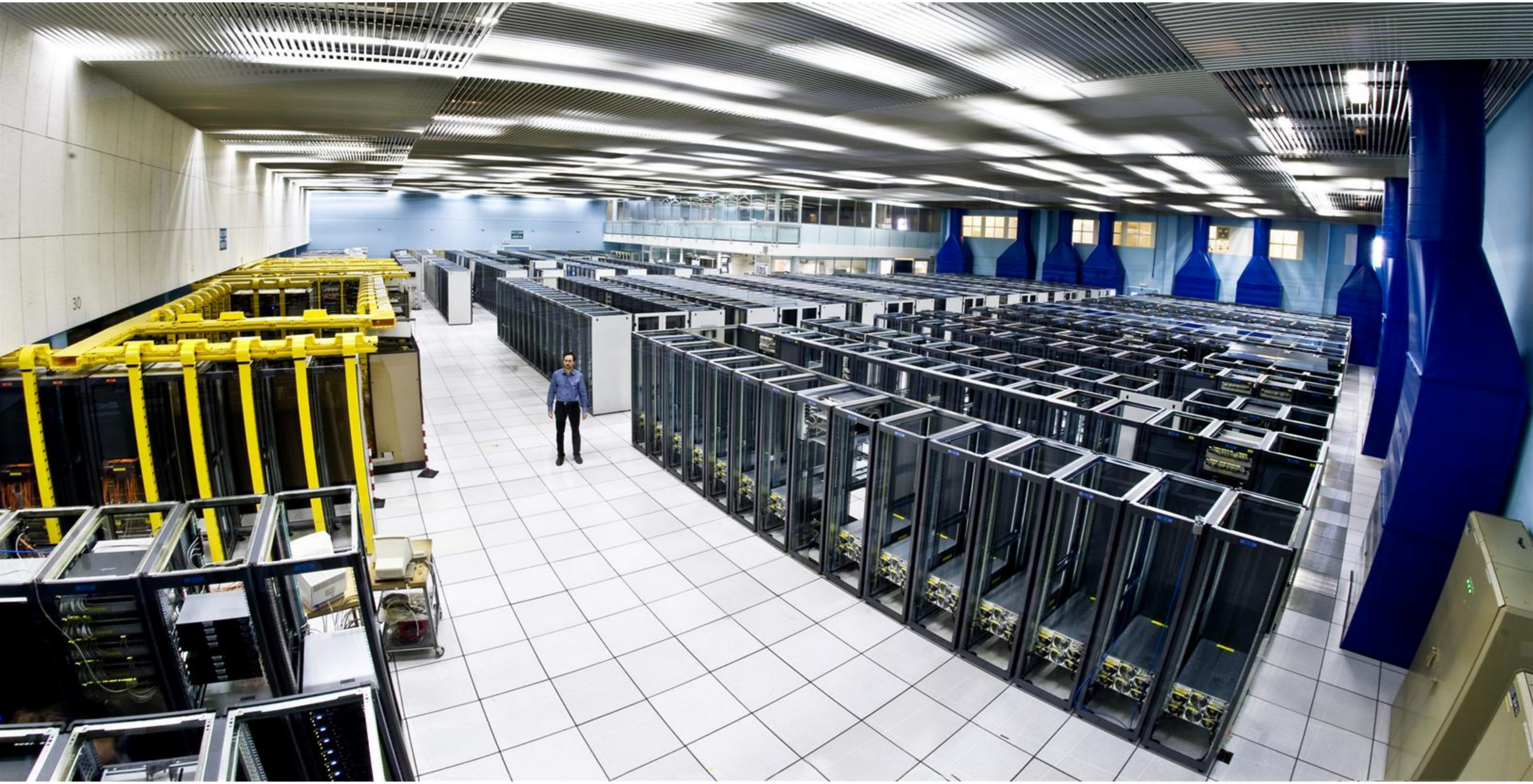
European Organization for Nuclear Research
Organisation Européenne pour la Recherche Nucléaire

CERN Agile Infrastructure

Tim Bell
Tim.Bell@cern.ch

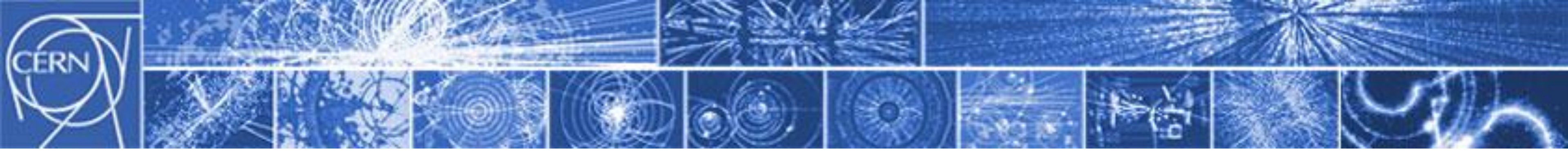
ATLAS Distributed Computing Tier-1/Tier-2/Tier-3 Jamboree
10th December 2012

The CERN Meyrin Data Centre



New data centre to expand capacity

- Data centre in Geneva at the limit of electrical capacity at 3.5MW
- New centre chosen in Budapest, Hungary
- Additional 2.7MW of usable power
- Hands off facility
- Deploying from 2013 with 200Gbit/s network to CERN



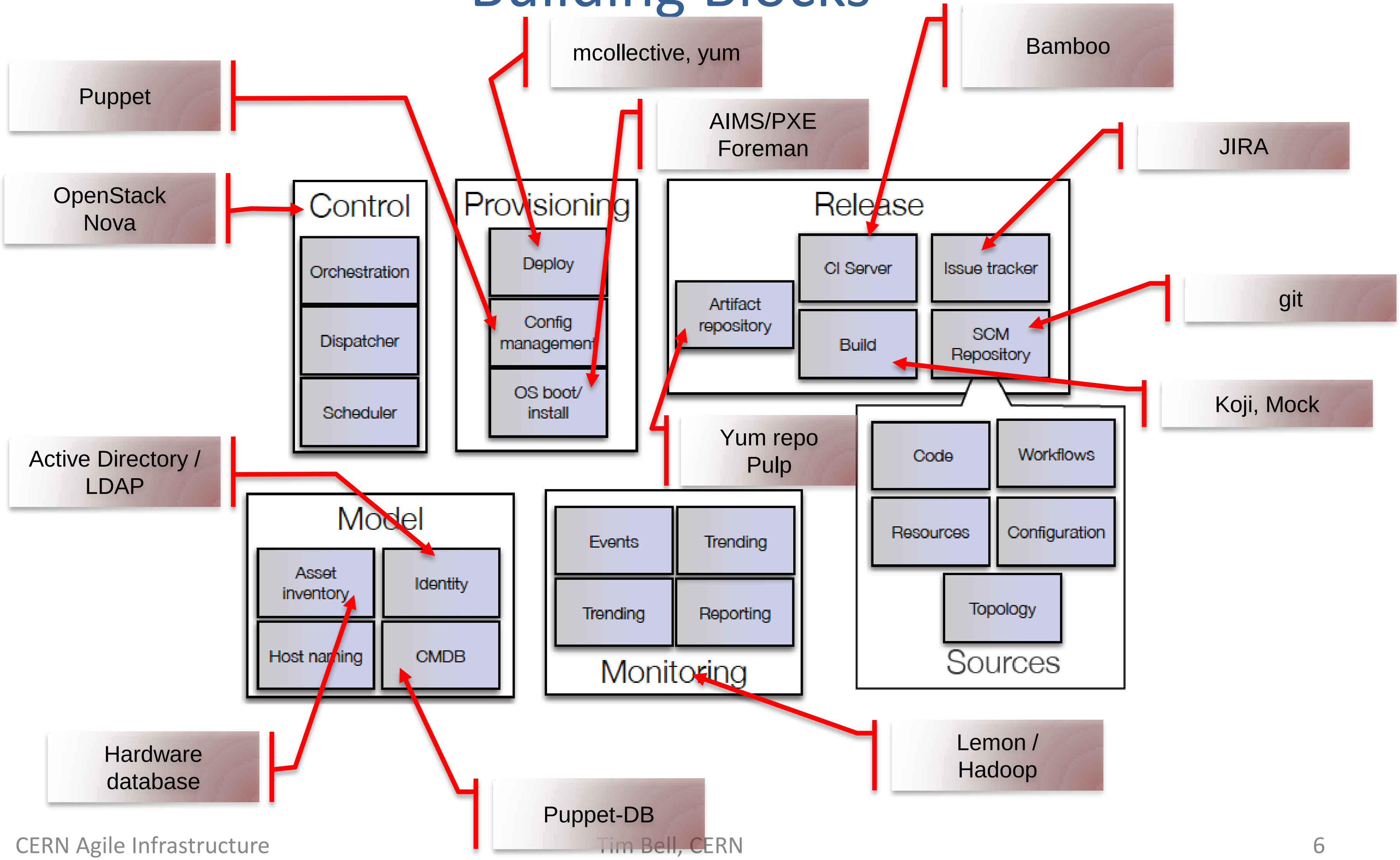
Time to change strategy

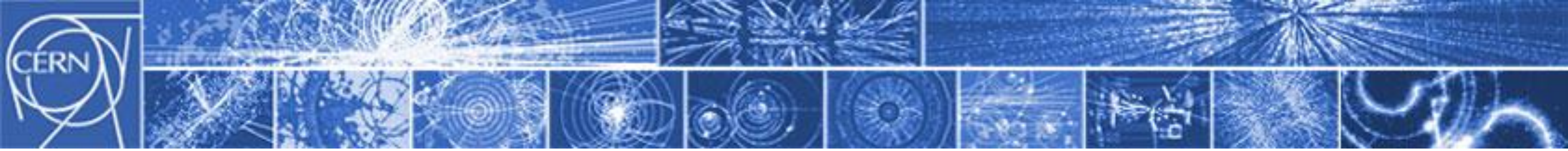
- Rationale
  - Need to manage twice the servers as today
  - No increase in staff numbers
  - Tools becoming increasingly brittle and will not scale as-is
- Approach
  - CERN is no longer a special case for compute
  - Adopt a tool chain model using existing open source tools
  - If we have special requirements, challenge them again and again
  - If useful, make generic and contribute back to the community
  - Address urgent needs first in
    - Configuration management
    - Monitoring
    - Infrastructure as a Service

Prepare the move to the clouds

- Improve operational efficiency
  - Machine ordering, reception and testing
  - Hardware interventions with long running programs
  - Multiple operating system demand
- Improve resource efficiency
  - Exploit idle resources, especially waiting for disk and tape I/O
  - Highly variable load such as interactive or build machines
- Enable cloud architectures
  - Gradual migration to cloud interfaces and workflows
  - Support autoscaling such as Webcast
- Improve responsiveness
  - Self-Service with coffee break response time

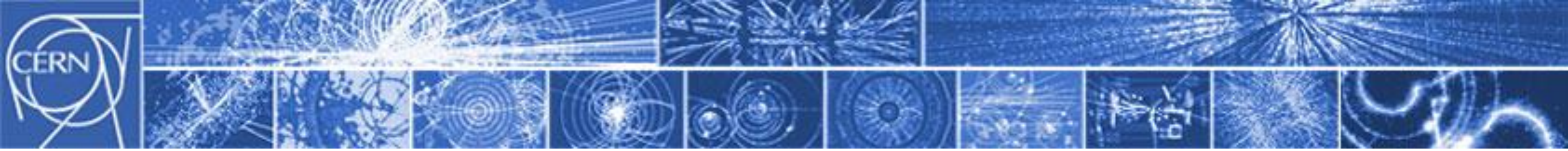
Building Blocks



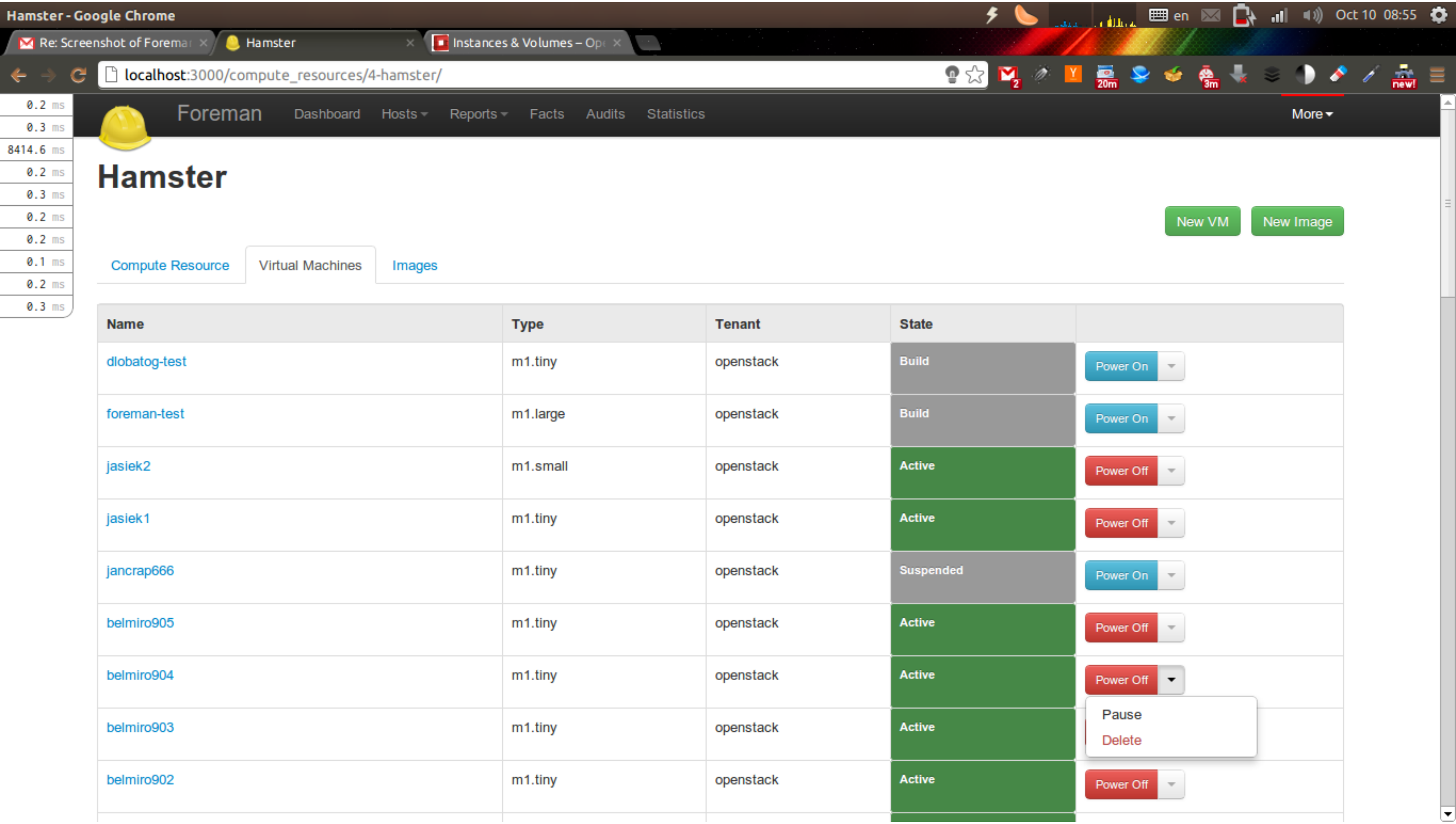


Configuration Management

- Using Puppet to replace Quattor for configuring machines
  - Active Community with many off-the-shelf configurations
  - Proven scalability such as Zynga with over 100,000 nodes
  - ATLAS online/offline, BNL, IN2P3 … are already using it
  - Publishing our modules to http://github.com/cernops such as CVMFS, VOMS, BDII if other sites are interested
- The Foreman provides a GUI
  - View configuration reports, errors, statistics, ….
  - Create new machines dynamically in a few minutes
  - Customise machines by adding modules or setting parameters
- Other tools
  - Git for recipe management
  - Mcollective for running commands across multiple machines



Foreman to manage Puppetized VM



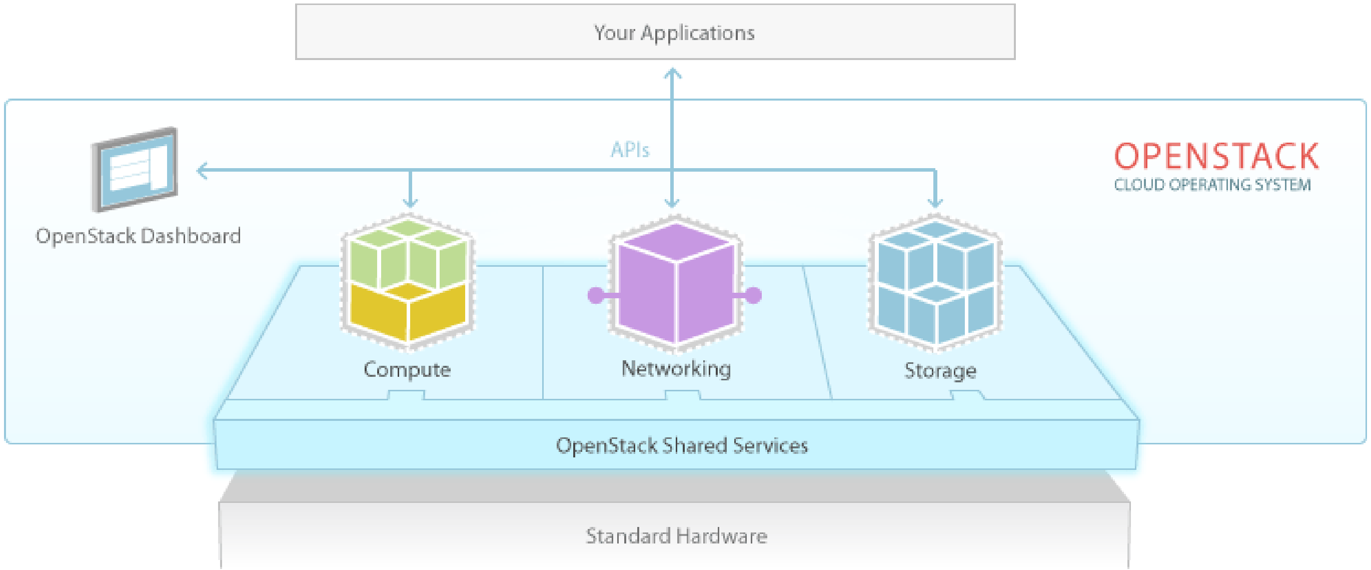
```
0.2 ms
0.3 ms
8414.6 ms
0.2 ms
0.3 ms
0.2 ms
0.2 ms
0.1 ms
0.2 ms
0.3 ms
```

Hamster

Name	Type	Tenant	State	
dlobatog-test	m1.tiny	openstack	Build	Power On
foreman-test	m1.large	openstack	Build	Power On
jasiek2	m1.small	openstack	Active	Power Off
jasiek1	m1.tiny	openstack	Active	Power Off
jancrap666	m1.tiny	openstack	Suspended	Power On
belmiro905	m1.tiny	openstack	Active	Power Off
belmiro904	m1.tiny	openstack	Active	Power Off
belmiro903	m1.tiny	openstack	Active	
belmiro902	m1.tiny	openstack	Active	Power Off

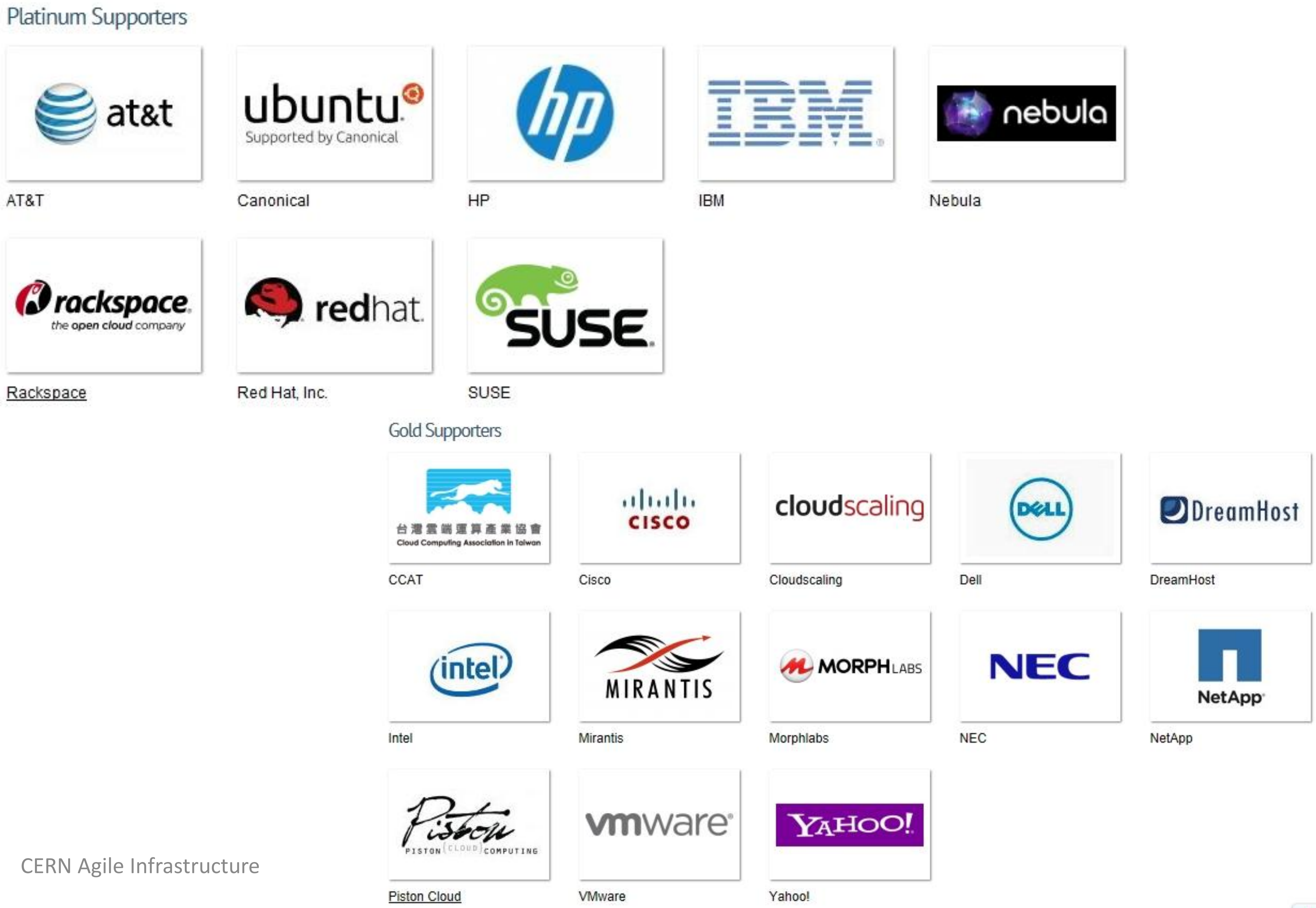
What is OpenStack ?

- OpenStack is a cloud operating system that controls large pools of compute, storage, and networking resources throughout a datacentre, all managed through a dashboard that gives administrators control while empowering their users to provision resources through a web interface



Principles

- Open Source
  - Apache 2.0 license, NO ‘enterprise’ version
- Open Design
  - Open Design Summit, anyone is able to define core architecture
- Open Development
  - Anyone can involve development process via Launchpad & Github
- Open Community
  - OpenStack Foundation in 2012, Now 190+ companies, 3000+ developers, 6000+ members

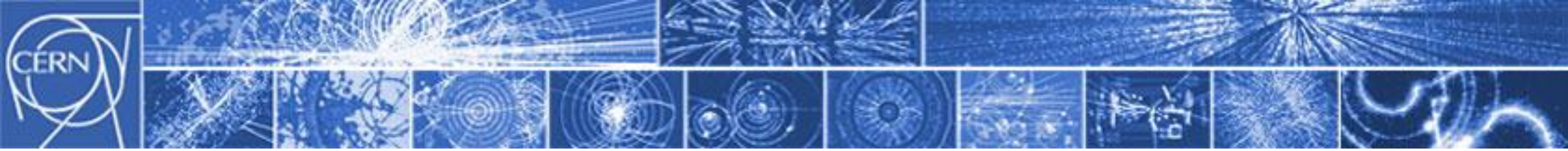


Platinum Supporters

- AT&T
- Canonical
- HP
- IBM
- Nebula
- Rackspace
- Red Hat, Inc.
- SUSE

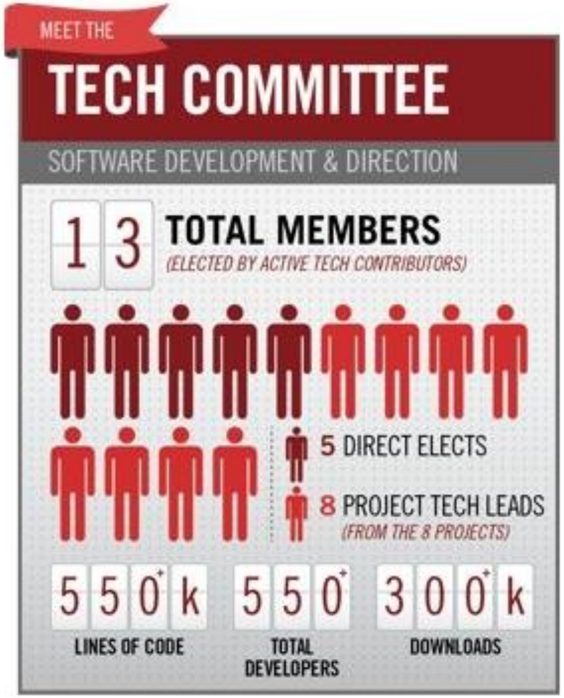
Gold Supporters

- CCAT
- Cisco
- Cloudscaling
- Dell
- DreamHost
- Intel
- Mirantis
- Morphlabs
- NEC
- NetApp
- Piston Cloud
- VMware
- Yahoo!



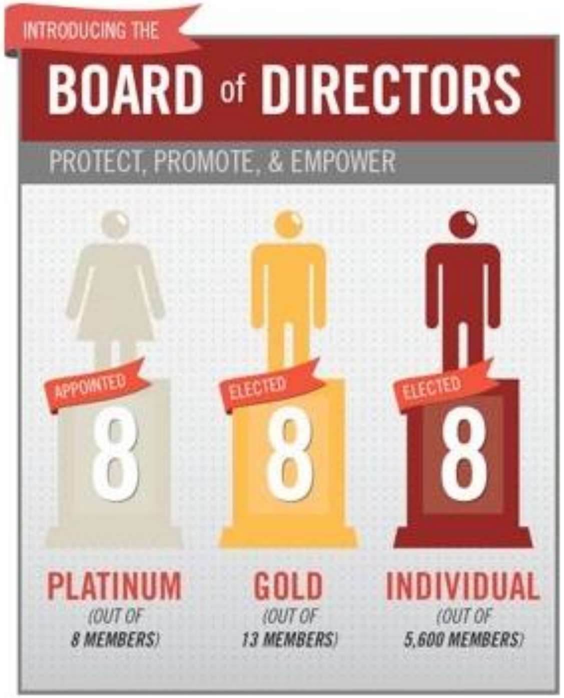
Foundation Governance

Technical Committee



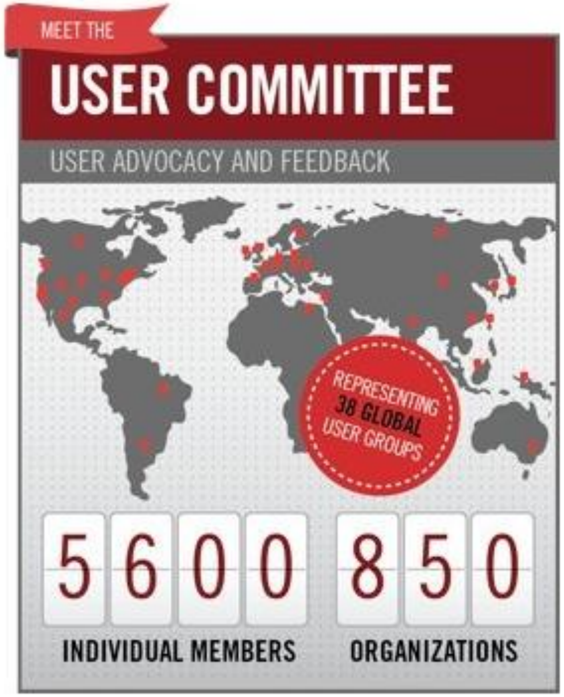
OpenStack Technical Committee – an evolution of the Project Policy Board – will define and steward the technical direction of OpenStack software development and includes elected Project Technical Leads for each of the core software projects.

Board of Directors

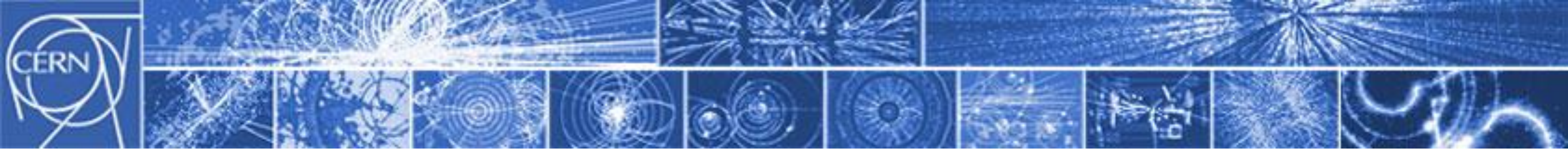


The Board of Directors provides strategic and financial oversight of Foundation resources and staff. Alan Clark, Director of Industry Initiatives, Emerging Standards and Open Source at SUSE, was elected Chairman of the Board, and Lew Tucker, Vice President and CTO of Cloud Computing at Cisco, was elected Vice Chairman of the Board.

User Committee



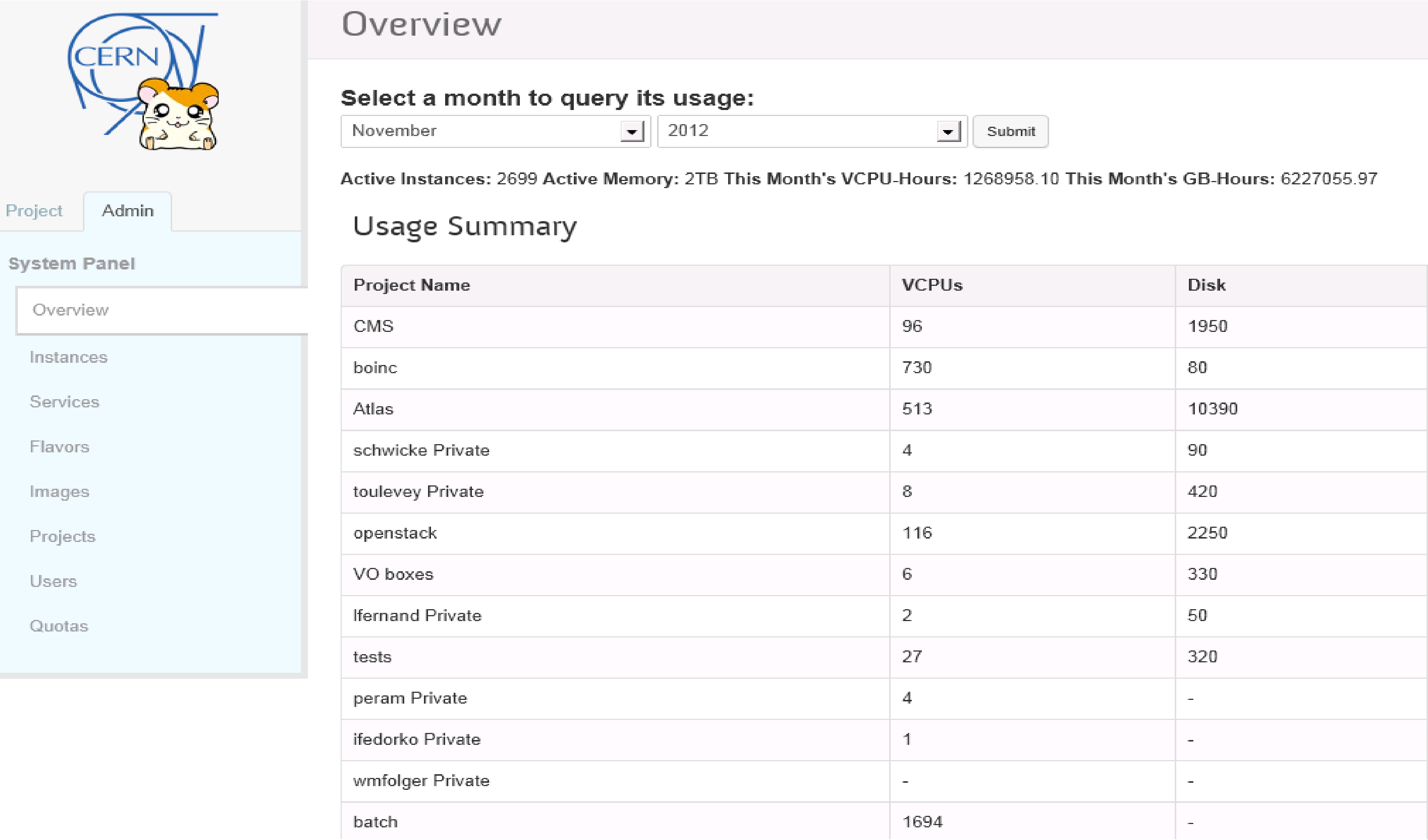
Tim Bell, Operating Systems and Infrastructure Services Group Leader at CERN, was appointed by the Board of Directors to help establish a new User Committee, created to represent a broad set of enterprise, academic and service provider users with the Technical Committee and Board of Directors.



OpenStack Cloud Components

- Nova – Compute Layer like Amazon EC2
- Swift – Object Store like Amazon S3
- Quantum – Networking such as SDN or load balancing
- Horizon – Dashboard GUI
- Cinder – Block Storage like Amazon EBS
- Keystone – Identity
- Glance – Image management

- Each component has an API and is pluggable
- Other non-core projects interact with these components such as load balancers and usage tracking



Overview

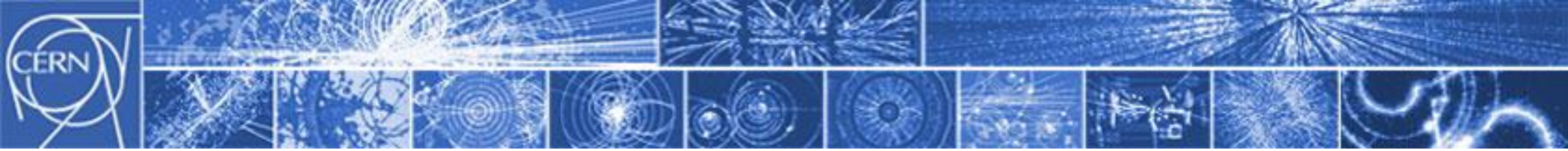
Select a month to query its usage:

November | 2012

Active Instances: 2699 **Active Memory:** 2TB **This Month's VCPU-Hours:** 1268958.10 **This Month's GB-Hours:** 6227055.97

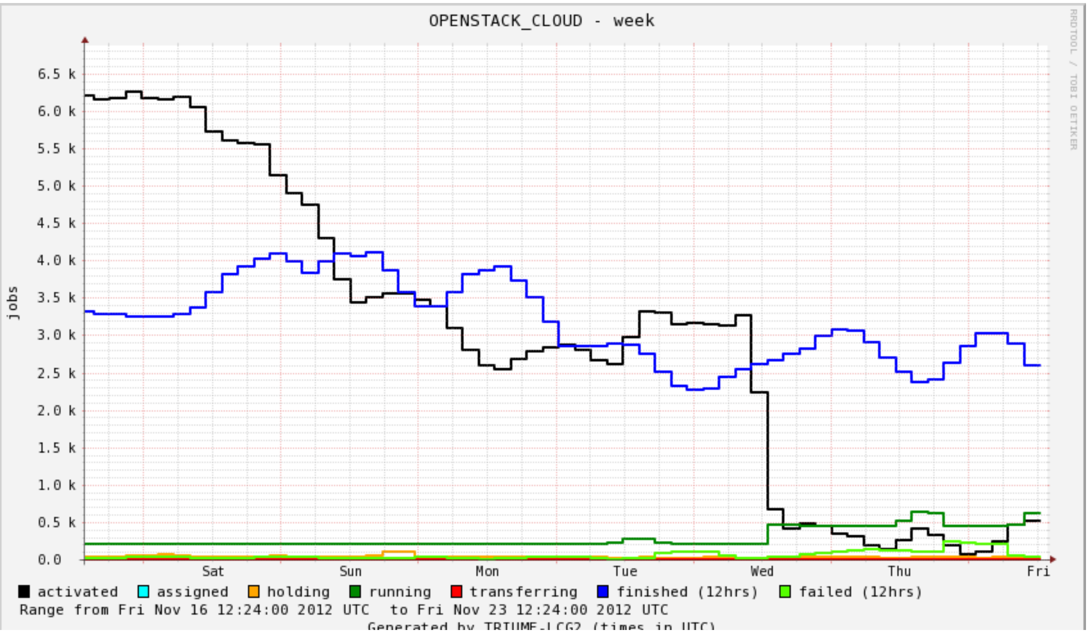
Usage Summary

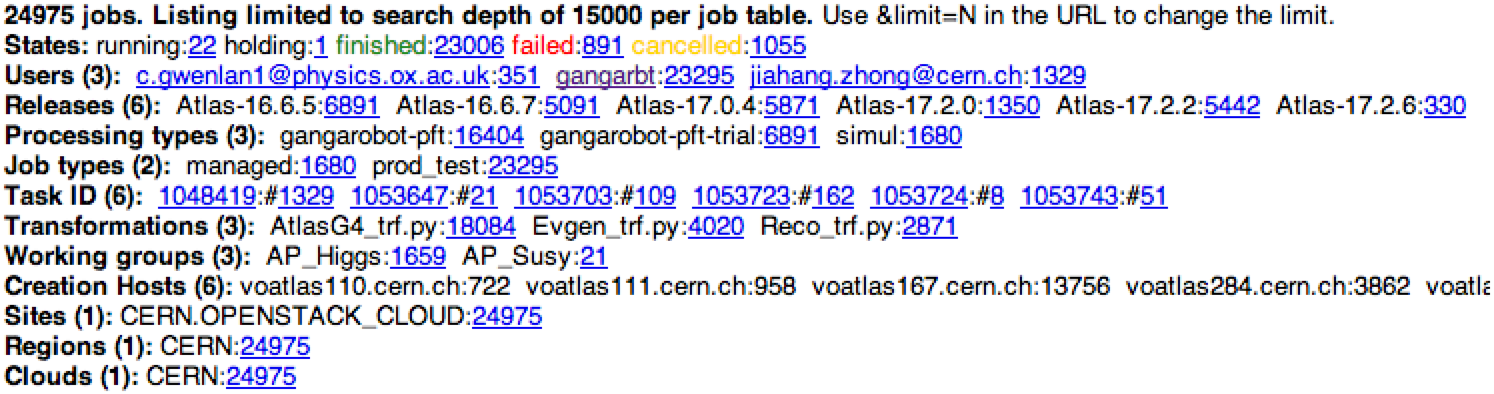
Project Name	VCPUs	Disk
CMS	96	1950
boinc	730	80
Atlas	513	10390
schwicke Private	4	90
toulevey Private	8	420
openstack	116	2250
VO boxes	6	330
lfernand Private	2	50
tests	27	320
peram Private	4	-
ifedorko Private	1	-
wmfolger Private	-	-
batch	1694	-



Current Status

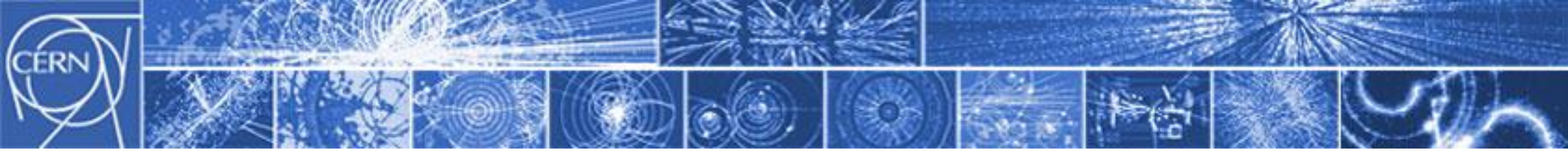
- Pre-production facility based on the OpenStack Essex release
  - Using KVM on Scientific Linux and Hyper-V on Windows
  - Around 200 hypervisors with over 2,700 VMs
  - Integrated into CERN user and group management systems
    - E-groups define who can administer and use a project
  - Host certificates, external connectivity, DNS names ready
  - Running multiple use cases
    - Batch services
    - Atlas workloads with IT/ES
    - Server consolidation tests for VOBoxes
- All is puppet managed
  - Puppetlabs provide an excellent set of manifests for configuration
  - More than 2,000 hosts managed by the central CERN instance





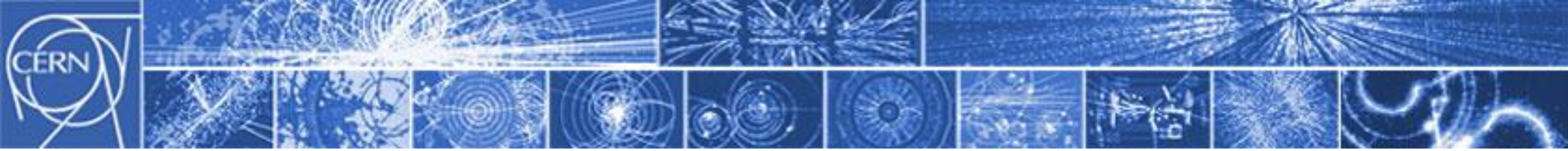
24975 jobs. Listing limited to search depth of 15000 per job table. Use &limit=N in the URL to change the limit.
States: running:22 holding:1 finished:23006 failed:891 cancelled:1055
Users (3): c.gwenlan1@physics.ox.ac.uk:351 gangarbt:23295 jiahang.zhong@cern.ch:1329
Releases (6): Atlas-16.6.5:6891 Atlas-16.6.7:5091 Atlas-17.0.4:5871 Atlas-17.2.0:1350 Atlas-17.2.2:5442 Atlas-17.2.6:330
Processing types (3): gangarobot-pft:16404 gangarobot-pft-trial:6891 simul:1680
Job types (2): managed:1680 prod_test:23295
Task ID (6): 1048419:#1329 1053647:#21 1053703:#109 1053723:#162 1053724:#8 1053743:#51
Transformations (3): AtlasG4_trf.py:18084 Evgen_trf.py:4020 Reco_trf.py:2871
Working groups (3): AP_Higgs:1659 AP_Susy:21
Creation Hosts (6): voatlas110.cern.ch:722 voatlas111.cern.ch:958 voatlas167.cern.ch:13756 voatlas284.cern.ch:3862 voatla
Sites (1): CERN.OPENSTACK_CLOUD:24975
Regions (1): CERN:24975
Clouds (1): CERN:24975

- Running on up to 450 cores with over 2,500 jobs completing in 12 hours



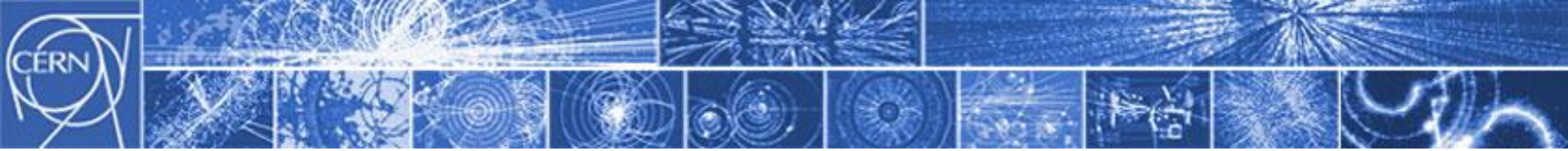
Current Status…

- Interfaces
  - GUI via the OpenStack dashboard
  - OpenStack and EC2 command line interfaces and APIs available
  - Users can be able to use standard images such as SLC6 or upload their own images
  - Different VM sizes are available (#cores,memory,disk) as on Amazon
- Production Preparation
  - High availability of OpenStack subsystems
  - Monitoring
- Ongoing extensions
  - Ceilometer to provide usage accounting by project
  - Boson for delegated quota management
  - Cinder for external block storage such as Gluster or NetApp



Future

- Plans
  - Production service in Q1 2013 in Meyrin data centre
  - Will be extended to Hungary at the start of 2013 with aim to run 90% virtualised services
  - Target is 100K to 300K VMs on 15K hypervisors by 2015
  - Explore federation opportunities with other clouds and share experiences/recipes
- Migration
  - Existing services such as Grid, lxbatch and VOBoxes will be increasingly run on virtual machines rather than physical boxes
  - Legacy tools such as Quattor to be replaced over the next 2 years
  - Usage of CERN Virtualisation Infrastructure (CVI) and Lxcloud test bed will be migrated to the CERN Private cloud

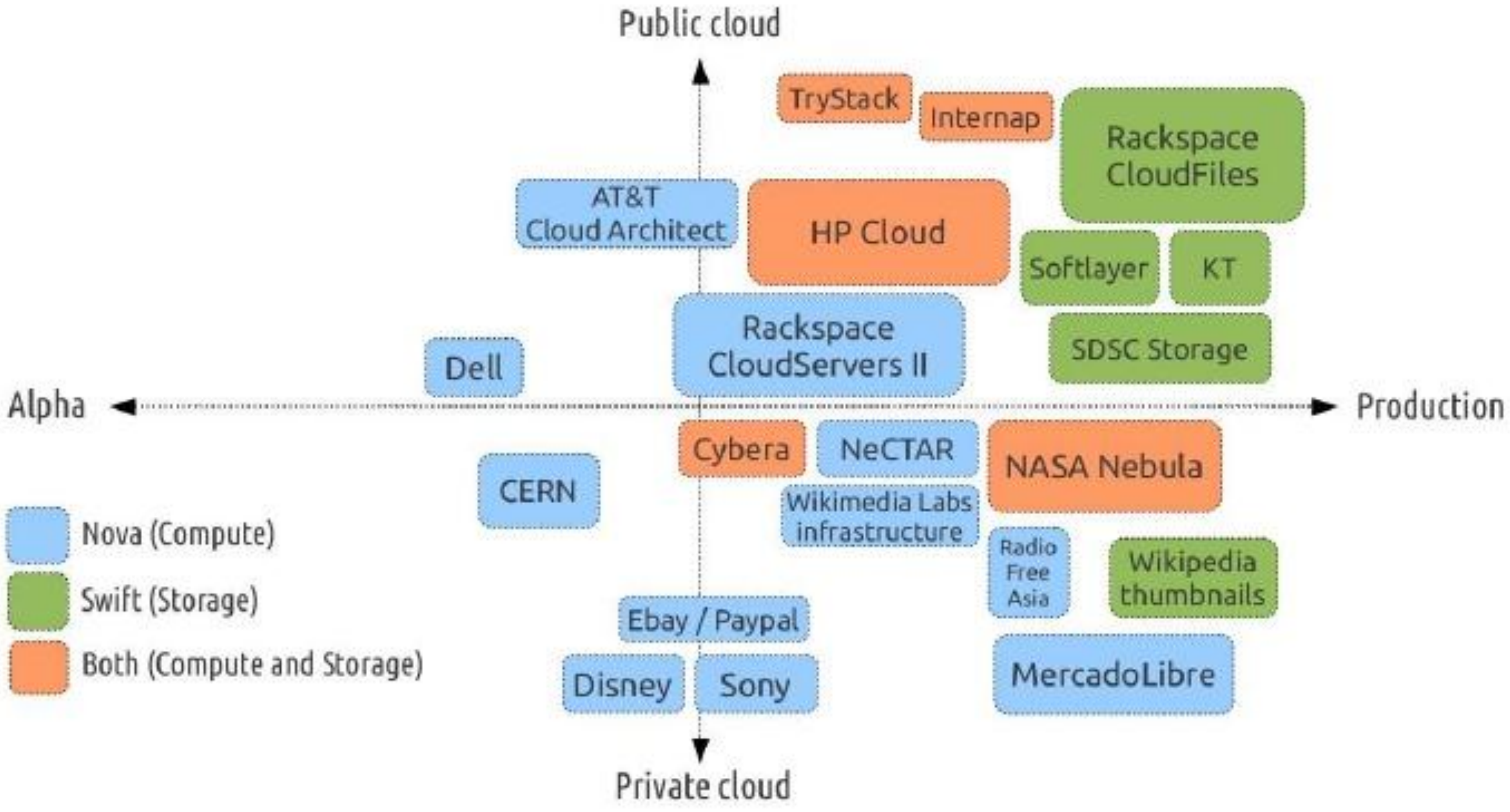


References

HEPiX Talk in Beijing	http://cern.ch/go/NF8f
OpenStack Summit San Diego 2012	http://www.youtube.com/user/OpenStackFoundation
OpenStack web site	http://openstack.org
CERN Puppet user group	http://cern.ch/go/vGR6
Paper presented at CHEP	http://cern.ch/go/N8wp
IT Technical Forum on Agile Infrastructure	http://cern.ch/go/FRW8
CERN Private Cloud user guide (Draft)	http://information-technology.web.cern.ch/book/cern-private-cloud-user-guide
Swiss OpenStack User group	http://www.openstack.org/blog/2012/12/1st-swiss-openstack-user-group/

Backup Slides

Known deployments



Public cloud

TryStack

Internap

Rackspace CloudFiles

AT&T Cloud Architect

HP Cloud

Softlayer

KT

Rackspace CloudServers II

Dell

SDSC Storage

Alpha

Production

Cybera

NeCTAR

NASA Nebula

CERN

Wikimedia Labs infrastructure

Radio Free Asia

Wikipedia thumbnails

Ebay / Paypal

Disney

Sony

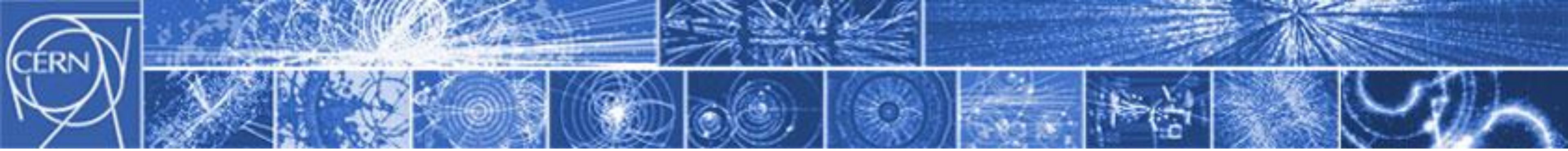
MercadoLibre

Private cloud

🟦 Nova (Compute)

🟩 Swift (Storage)

🟧 Both (Compute and Storage)



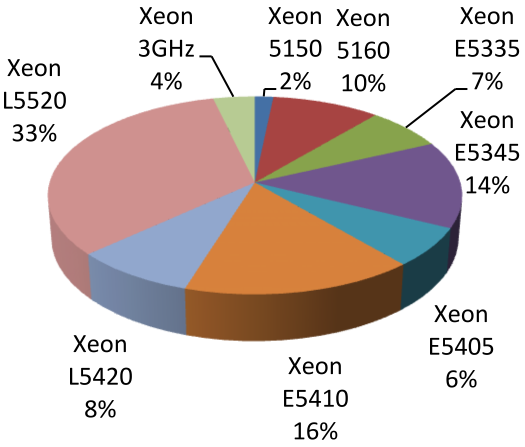
- Data Centre by Numbers
  - Hardware installation & retirement
    - ~7,000 hardware movements/year; ~1,800 disk failures/year

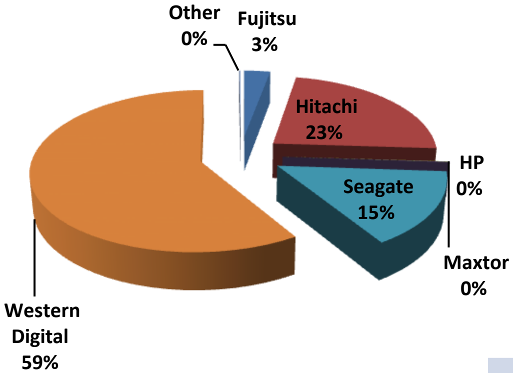
Racks	828
Servers	11,728
Processors	15,694
Cores	64,238
HEPSpec06	482,507

Disks	64,109
Raw disk capacity (TiB)	63,289
Memory modules	56,014
Memory capacity (TiB)	158
RAID controllers	3,749

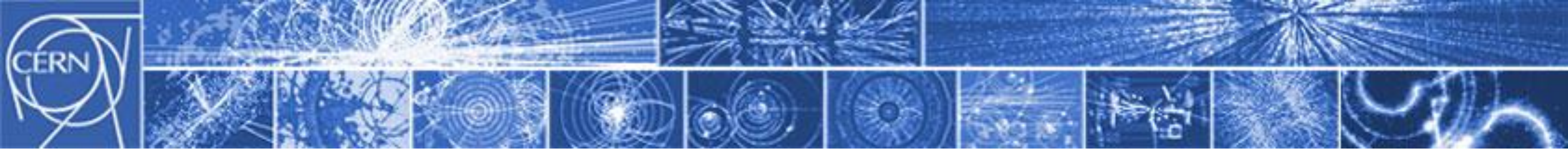
Tape Drives	160
Tape Cartridges	45,000
Tape slots	56,000
Tape Capacity (TiB)	73,000

High Speed Routers (640 Mbps → 2.4 Tbps)	24
Ethernet Switches	350
10 Gbps ports	2,000
Switching Capacity	4.8 Tbps
1 Gbps ports	16,939
10 Gbps ports	558





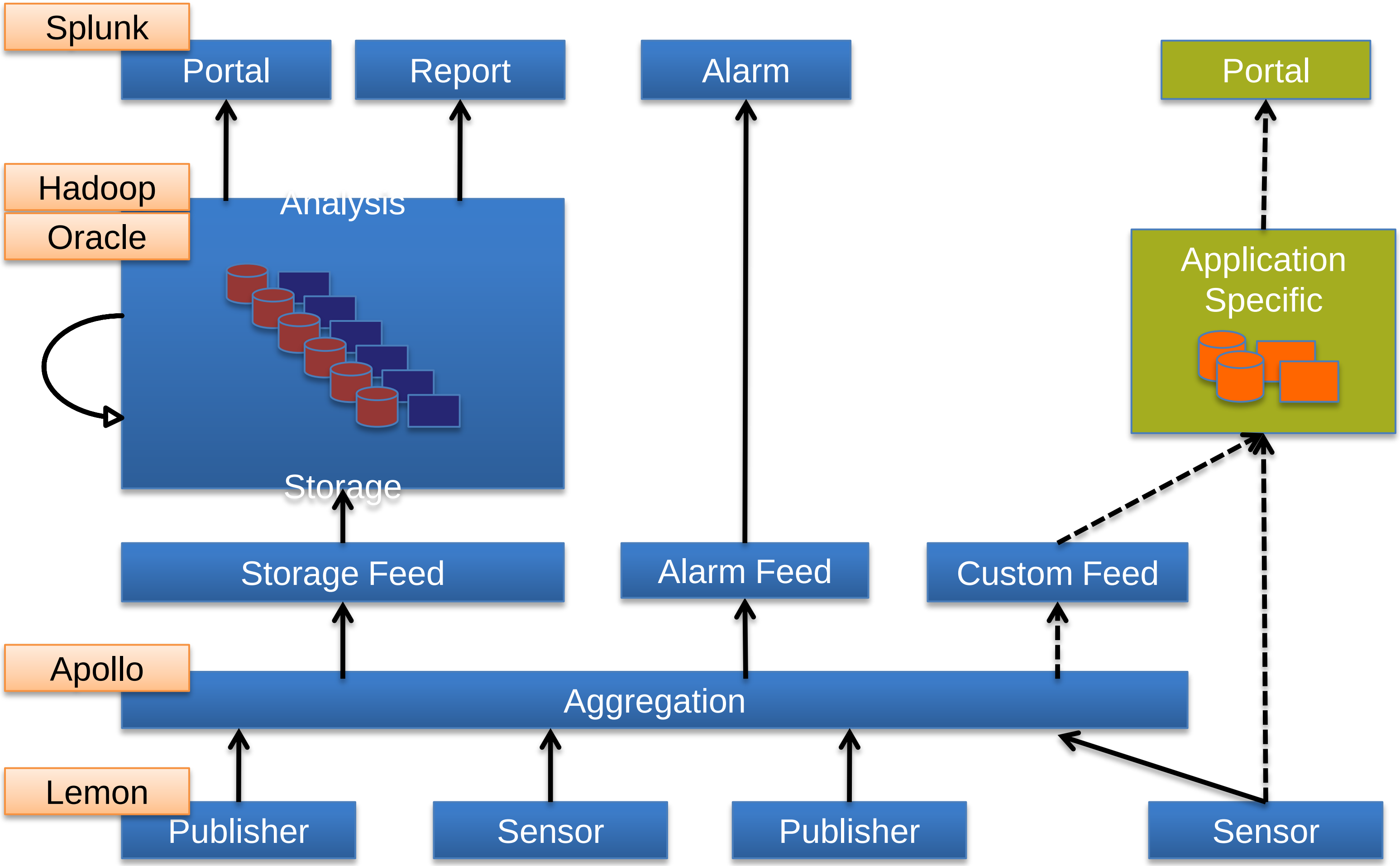
IT Power Consumption	2,456 KW
Total Power Consumption	3,890 KW



Public Procurement Purchase Model

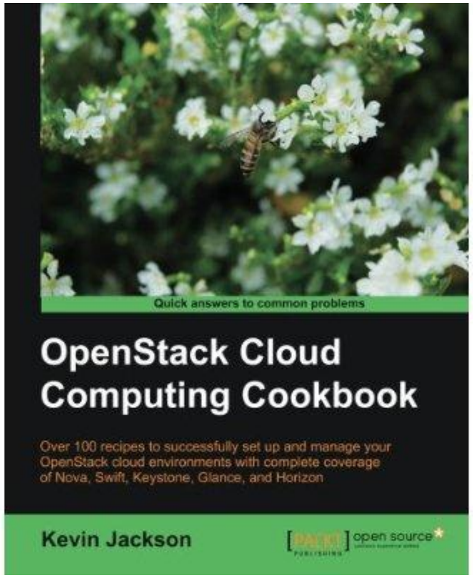
Step	Time (Days)	Elapsed (Days)
User expresses requirement		0
Market Survey prepared	15	15
Market Survey for possible vendors	30	45
Specifications prepared	15	60
Vendor responses	30	90
Test systems evaluated	30	120
Offers adjudicated	10	130
Finance committee	30	160
Hardware delivered	90	250
Burn in and acceptance	30 days typical 380 worst case	280
Total		**280+ Days**

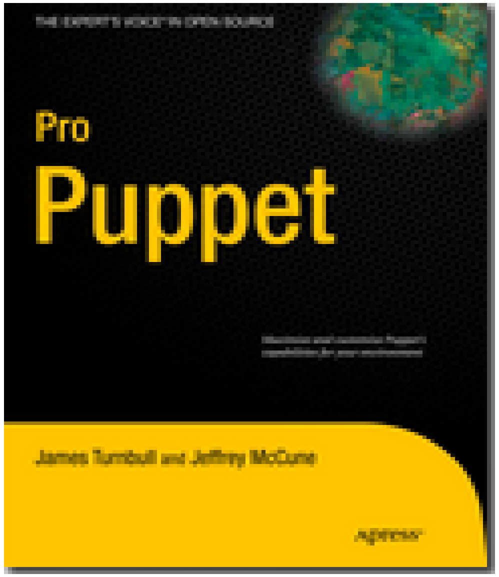
Monitoring


Splunk
Portal
Report
Alarm
Portal
Hadoop
Oracle
Analysis
Application Specific
Storage
Storage Feed
Alarm Feed
Custom Feed
Apollo
Aggregation
Lemon
Publisher
Sensor
Publisher
Sensor

Training and Support

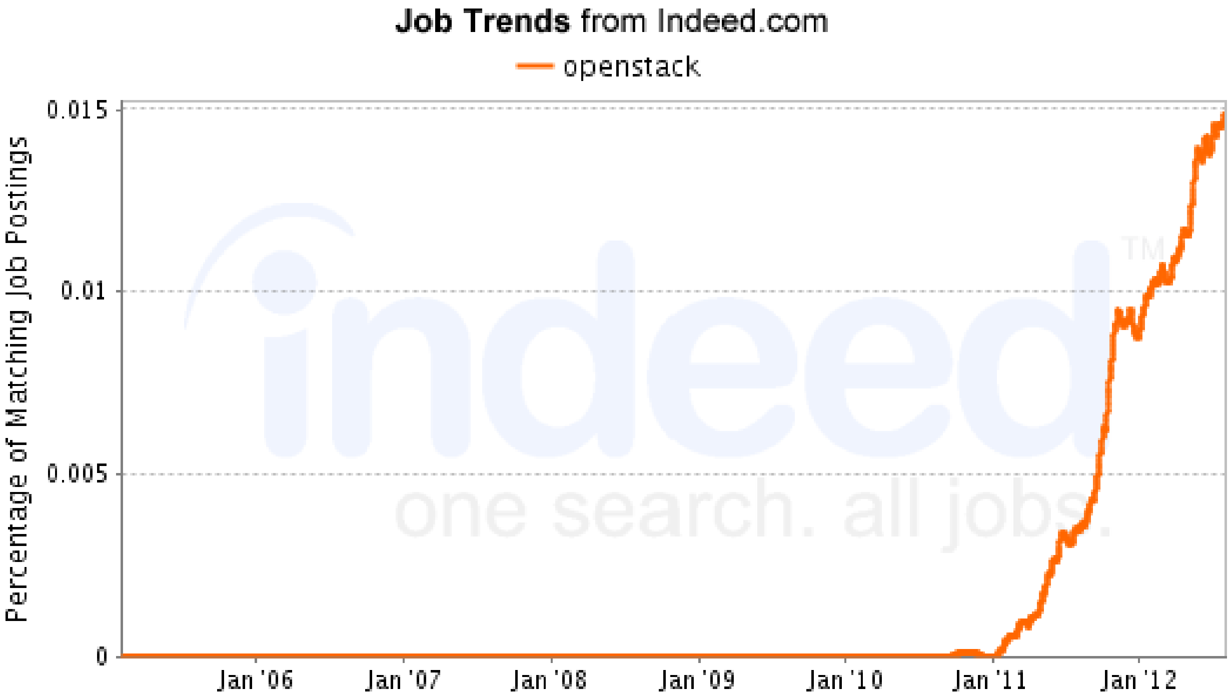
- Buy the book rather than guru mentoring
- Follow the mailing lists to learn
- Newcomers are rapidly productive (and often know more than us)
- Community and Enterprise support means we’re not on our own


Quick answers to common problems
OpenStack Cloud Computing Cookbook
Over 100 recipes to successfully set up and manage your OpenStack cloud environments with complete coverage of Nova, Swift, Keystone, Glance, and Horizon
Kevin Jackson
[PACKT] open source


THE EXPERT'S VOICE® IN OPEN SOURCE
Pro
Puppet
James Turnbull and Jeffrey McCune
Apress®

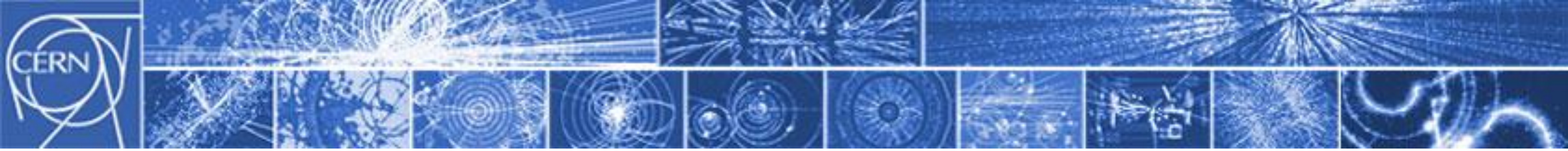
Staff Motivation

- Skills valuable outside of CERN when an engineer's contracts end

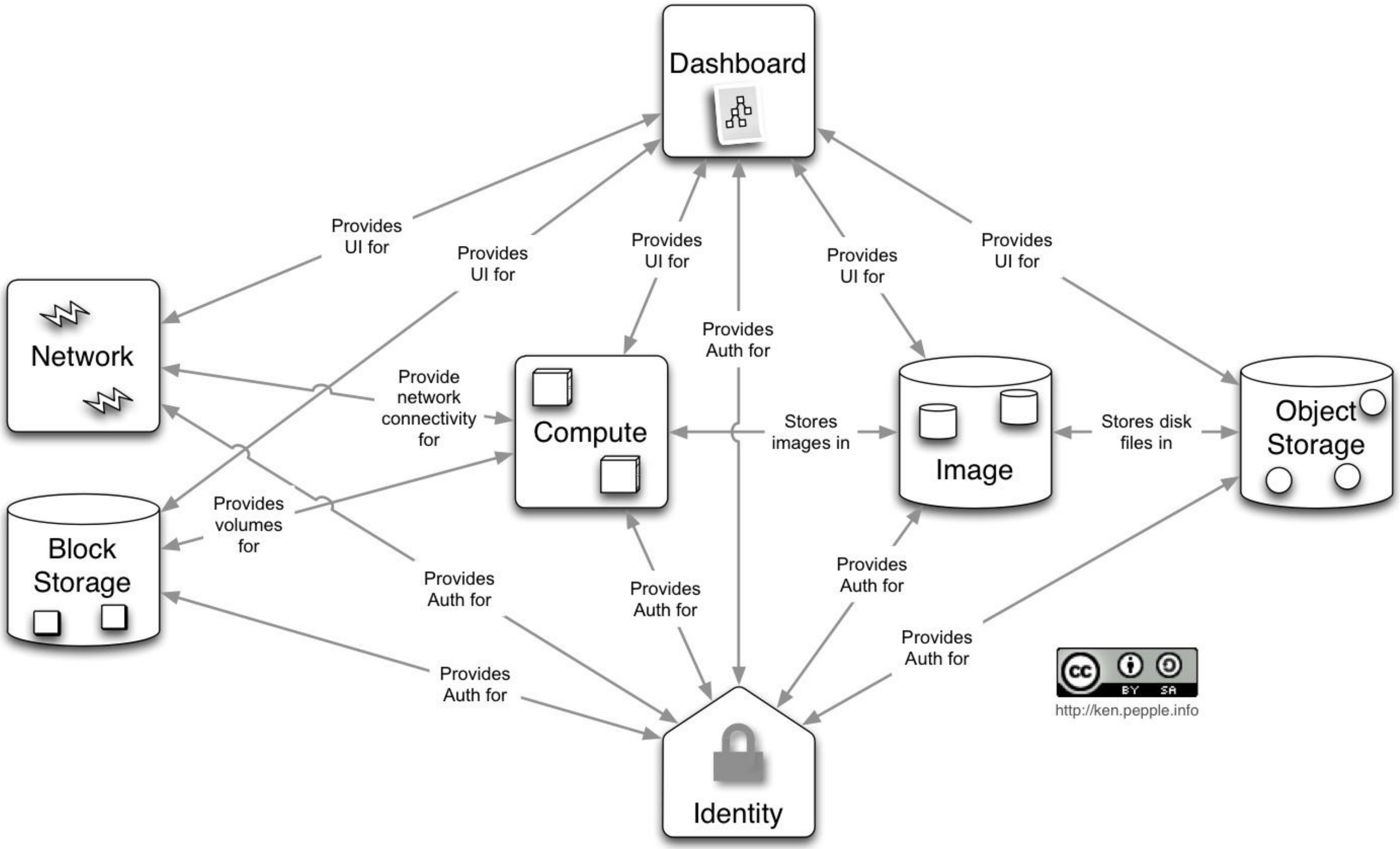


When communities combine…

- OpenStack’s many components and options make configuration complex out of the box
- Puppet forge module from PuppetLabs does our configuration
- The Foreman adds OpenStack provisioning for user kiosk to a configured machine in 15 minutes



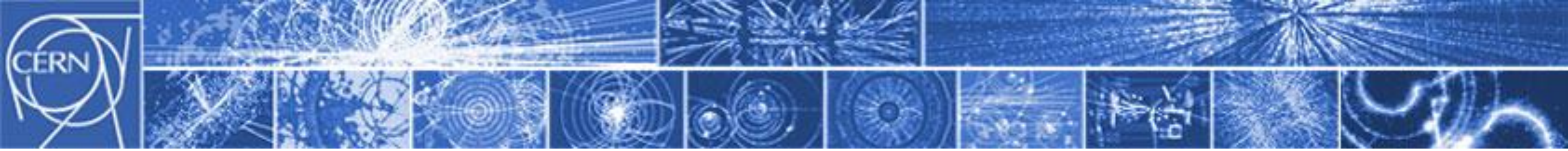
OpenStack Folsom



http://ken.pepple.info

OpenStack Board Structure

- "Individual Members" who participate on their own or as part of their paid employment. It’s free to join as an Individual Member and Individual Members have the right to run for, and vote for, a number of leadership positions.
- ”Platinum Members” are companies which make a significant strategic commitment to OpenStack in funding and resources. Platinum Members each appoint a representative to the Board of Directors
- ”Gold Members” are companies which provide funding and resources, but at a lower level than Platinum Members. Associate Members as a class elect representatives to the Board of Directors.

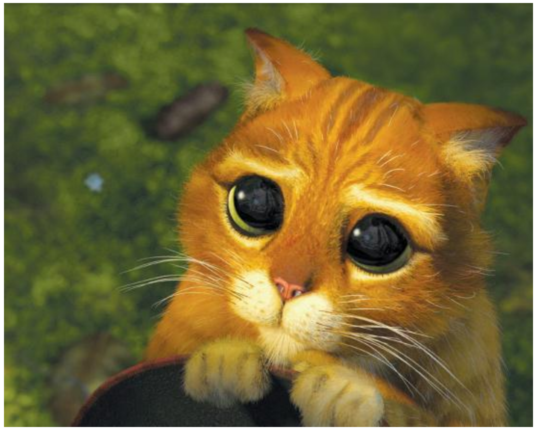


OpenStack Foundation

- The OpenStack Foundation is an independent body providing shared resources to help achieve the OpenStack Mission by Protecting, Empowering, and Promoting OpenStack software and the community around it, including users, developers and the entire ecosystem.
- Provides
  - Promotion of the OpenStack brand
  - Event management such as Summits, User Groups, …
  - Legal coverage of trademark, contributions
  - Continuous integration and Testing infrastructure
  - Developer tools to ease contribution

Service Model

cloudscaling



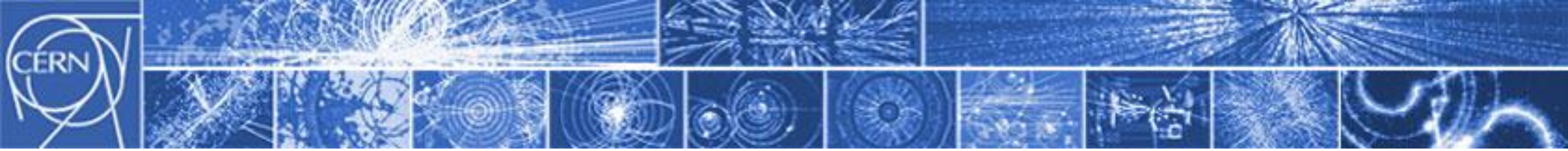
- Pets are given names like pussinboots.cern.ch
- They are unique, lovingly hand raised and cared for
- When they get ill, you nurse them back to health

- Cattle are given numbers like vm0042.cern.ch
- They are almost identical to other cattle
- When they get ill, you get another one

- Future application architectures should use Cattle but Pets with strong configuration management are viable and still needed

Supporting the Pets with OpenStack

- Network
  - Interfacing with legacy site DNS and IP management
  - Ensuring Kerberos identity before VM start
- Puppet
  - Ease use of configuration management tools with our users
  - Exploit mcollective for orchestration/delegation
- External Block Storage
  - Currently using nova-volume with Gluster backing store
- Live migration to maximise availability
  - KVM live migration using Gluster
  - KVM and Hyper-V block migration

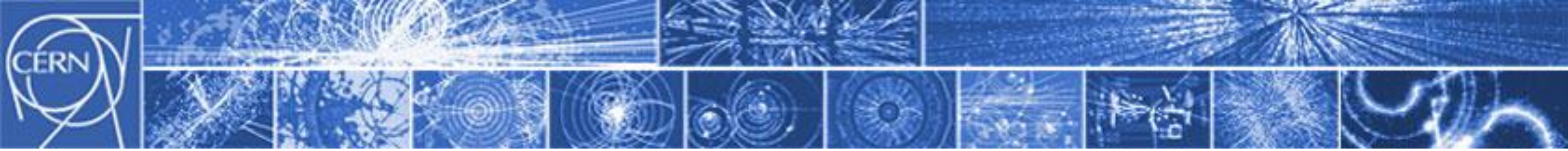


Active Directory Integration

- CERN’s Active Directory
  - Unified identity management across the site
  - 44,000 users
  - 29,000 groups
  - 200 arrivals/departures per month
- Full integration with Active Directory via LDAP
  - Uses the OpenLDAP backend with some particular configuration settings
  - Aim for minimal changes to Active Directory
  - 7 patches submitted around hard coded values and additional filtering
- Now in use in our pre-production instance
  - Map project roles (admins, members) to groups
  - Documentation in the OpenStack wiki

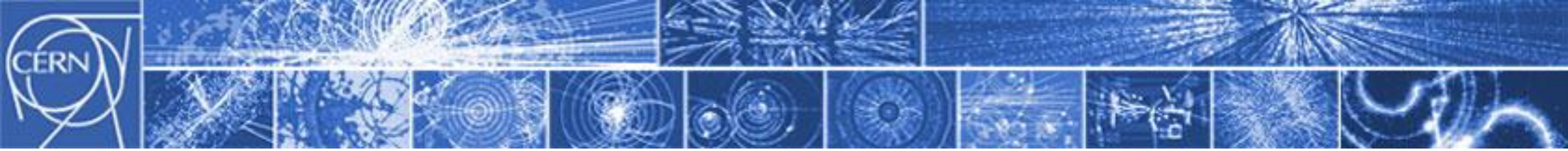
Welcome Back Hyper-V!

- We currently use Hyper-V/System Centre for our server consolidation activities
  - But need to scale to 100x current installation size
- Choice of hypervisors should be tactical
  - Performance
  - Compatibility/Support with integration components
  - Image migration from legacy environments
- CERN is working closely with the Hyper-V OpenStack team
  - Puppet to configure hypervisors on Windows
  - Most functions work well but further work on Console, Ceilometer, …



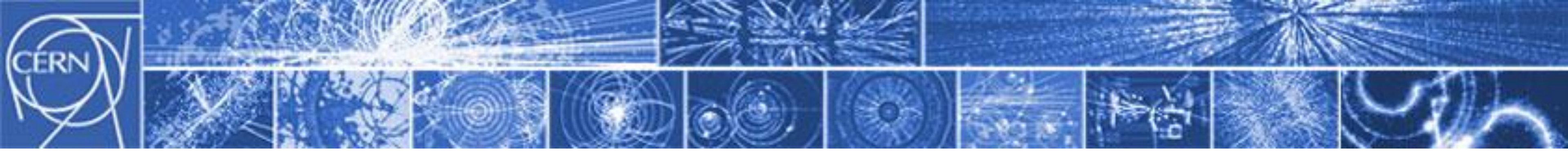
What are we missing (or haven’t found yet) ?

- Best practice for
  - Monitoring and KPIs as part of core functionality
  - Guest disaster recovery
  - Migration between versions of OpenStack
- Roles within multi-user projects
  - VM owner allowed to manage their own resources (start/stop/delete)
  - Project admins allowed to manage all resources
  - Other members should not have high rights over other members VMs
- Global quota management for non-elastic private cloud
  - Manage resource prioritisation and allocation centrally
  - Capacity management / utilisation for planning



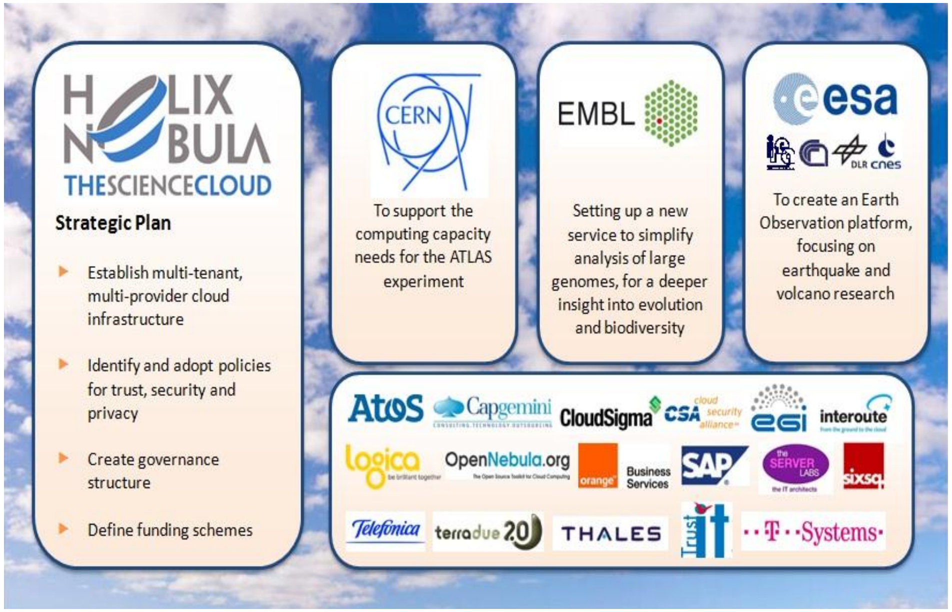
Opportunistic Clouds in online experiment farms

- The CERN experiments have farms of 1000s of Linux servers close to the detectors to filter the 1PByte/s down to 6GByte/s to be recorded to tape
- When the accelerator is not running, these machines are currently idle
  - Accelerator has regular maintenance slots of several days
  - Long Shutdown due from March 2013-November 2014
- One of the experiments are deploying OpenStack on their farm
  - Simulation (low I/O, high CPU)
  - Analysis (high I/O, high CPU, high network)



Federated European Clouds

- Two significant European projects around Federated Clouds
  - European Grid Initiative Federated Cloud as a federation of grid sites providing IaaS
  - HELiX Nebula European Union funded project to create a scientific cloud based on commercial providers



HELIX NEBULA THESCIENCECLOUD

Strategic Plan

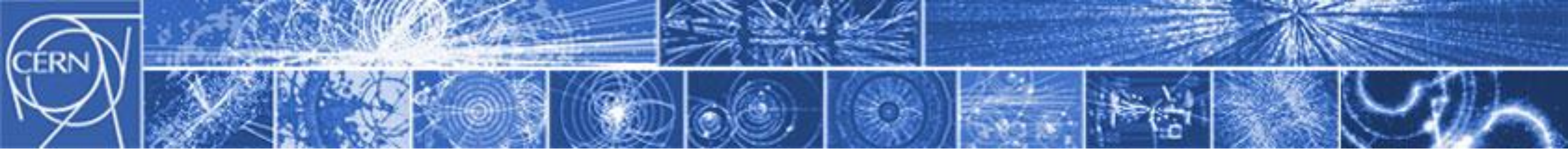
- Establish multi-tenant, multi-provider cloud infrastructure
- Identify and adopt policies for trust, security and privacy
- Create governance structure
- Define funding schemes

CERN: To support the computing capacity needs for the ATLAS experiment

EMBL: Setting up a new service to simplify analysis of large genomes, for a deeper insight into evolution and biodiversity

esa: To create an Earth Observation platform, focusing on earthquake and volcano research

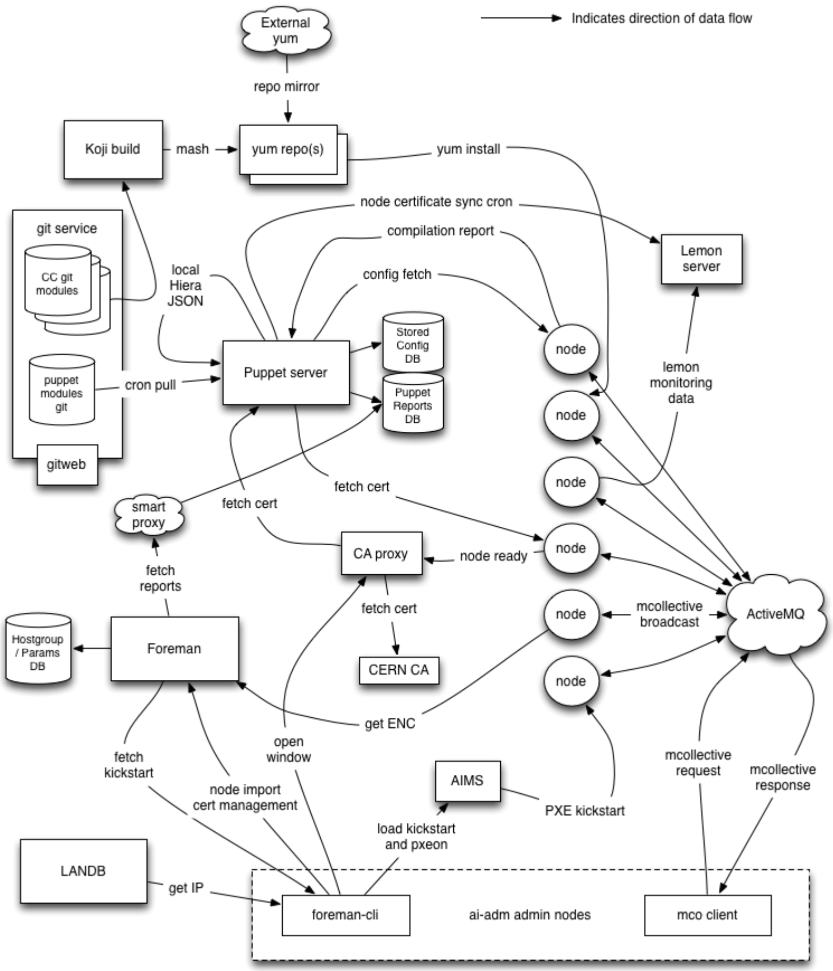
EGI Federated Cloud Sites			
CESGA	CESNET	INFN	SARA
Cyfronet	FZ Jülich	SZTAKI	IPHC
GRIF	GRNET	KTH	Oxford
GWDG	IGI	TCD	IN2P3
STFC			



Federated Cloud Commonalities

- Basic building blocks
  - Each site gives an IaaS endpoint with an API and common security policy
    - OCCI? CDMI ? Libcloud ? Jclouds ?
  - Image stores available across the sites
  - Federated identity management based on X.509 certificates
  - Consolidation of accounting information to validate pledges and usage
- Multiple cloud technologies
  - OpenStack
  - OpenNebula
  - Proprietary

New architecture data flows



Indicates direction of data flow

External yum

repo mirror

Koji build

mash

yum repo(s)

yum install

node certificate sync cron

compilation report

config fetch

Lemon server

git service

CC git modules

puppet modules git

gitweb

local Hiera JSON

cron pull

Puppet server

Stored Config DB

Puppet Reports DB

node

node

node

node

node

node

lemon monitoring data

smart proxy

fetch cert

fetch cert

CA proxy

node ready

fetch reports

fetch cert

CERN CA

mcollective broadcast

ActiveMQ

Hostgroup / Params DB

Foreman

get ENC

fetch kickstart

open window

node import cert management

AIMS

PXE kickstart

load kickstart and pxeon

mcollective request

mcollective response

LANDB

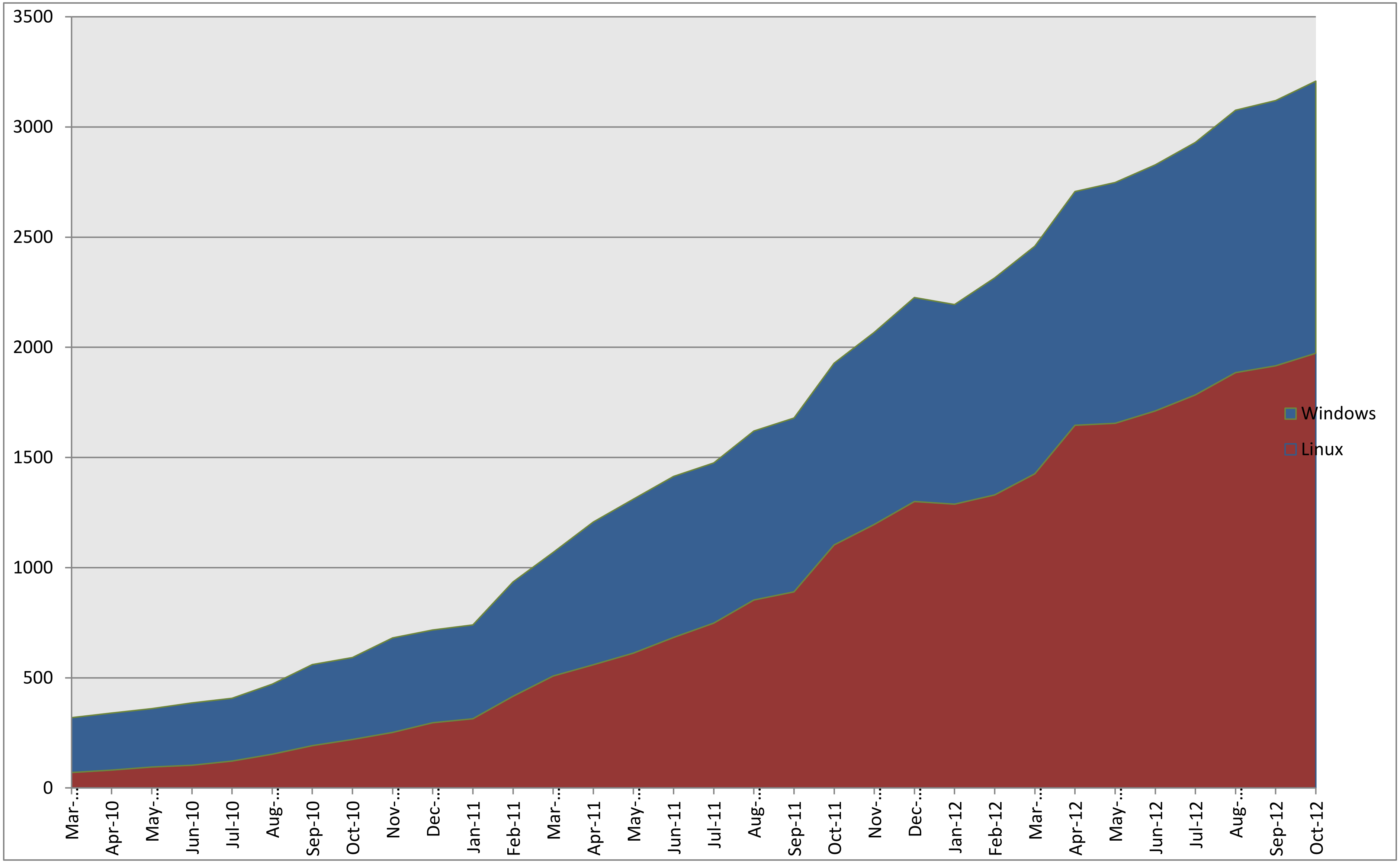
get IP

foreman-cli

ai-adm admin nodes

mco client

Virtualisation on SCVMM/Hyper-V


Mar-…
Apr-10
May-…
Jun-10
Jul-10
Aug-…
Sep-10
Oct-10
Nov-…
Dec-…
Jan-11
Feb-11
Mar-…
Apr-11
May-…
Jun-11
Jul-11
Aug-…
Sep-11
Oct-11
Nov-…
Dec-…
Jan-12
Feb-12
Mar-…
Apr-12
May-…
Jun-12
Jul-12
Aug-…
Sep-12
Oct-12
Windows
Linux