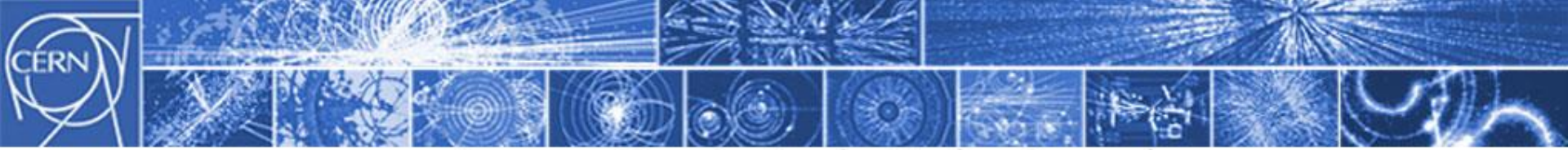


CERN

European Organization for Nuclear Research
Organisation Européenne pour la Recherche Nucléaire

Accelerating Science with OpenStack

Jan van Eldik
Jan.van.Eldik@cern.ch

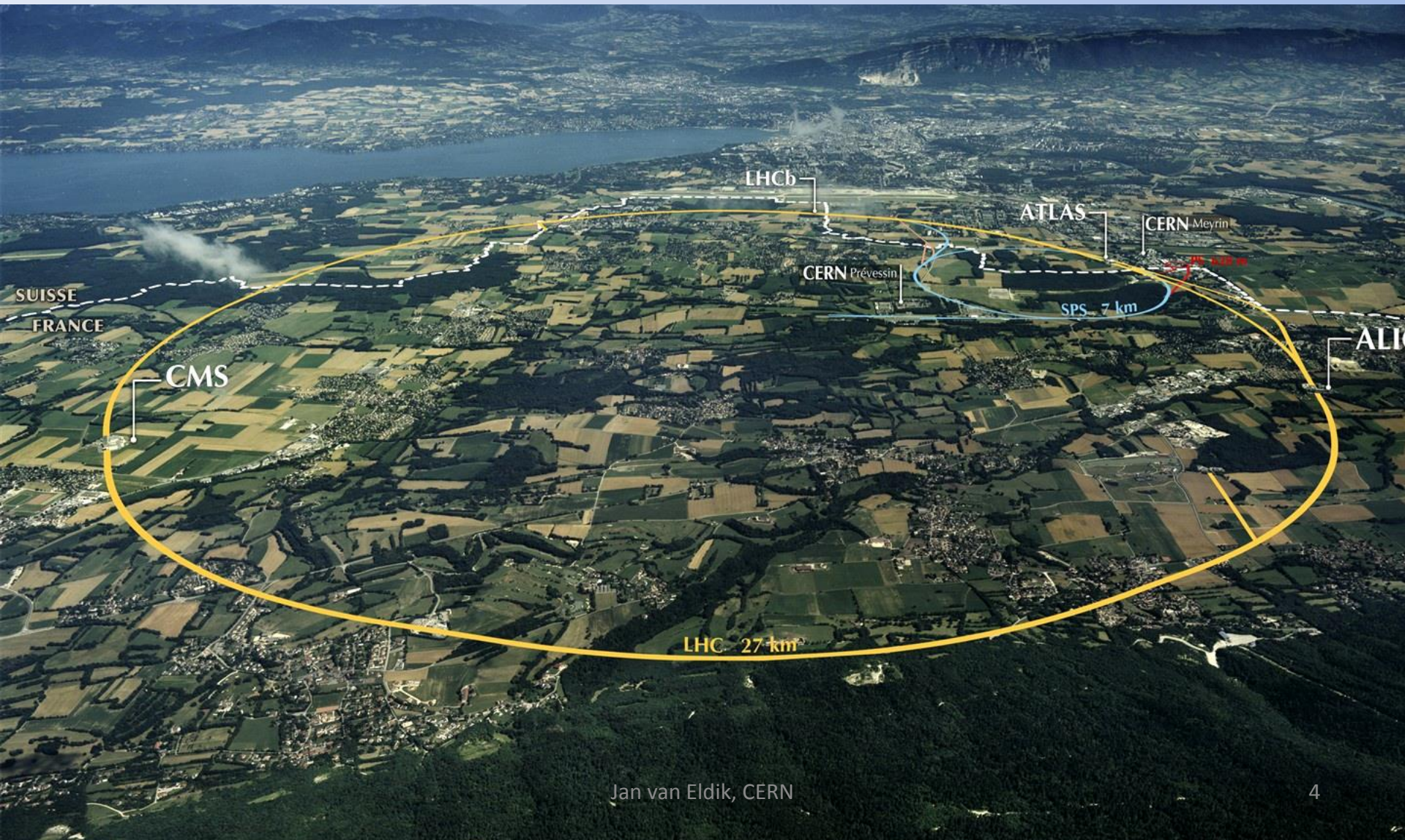
openstack™
CLOUD SOFTWARE

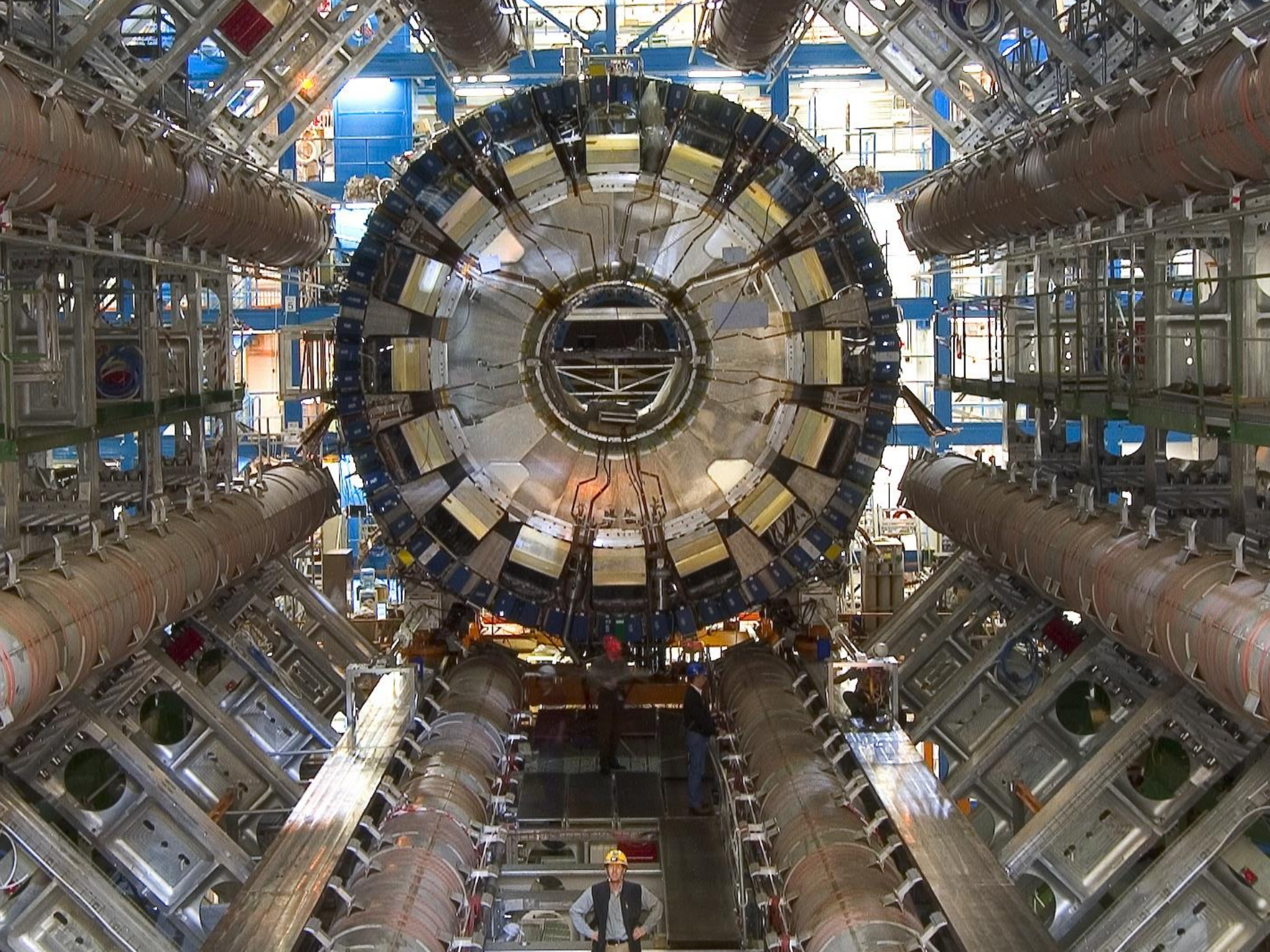
CERN

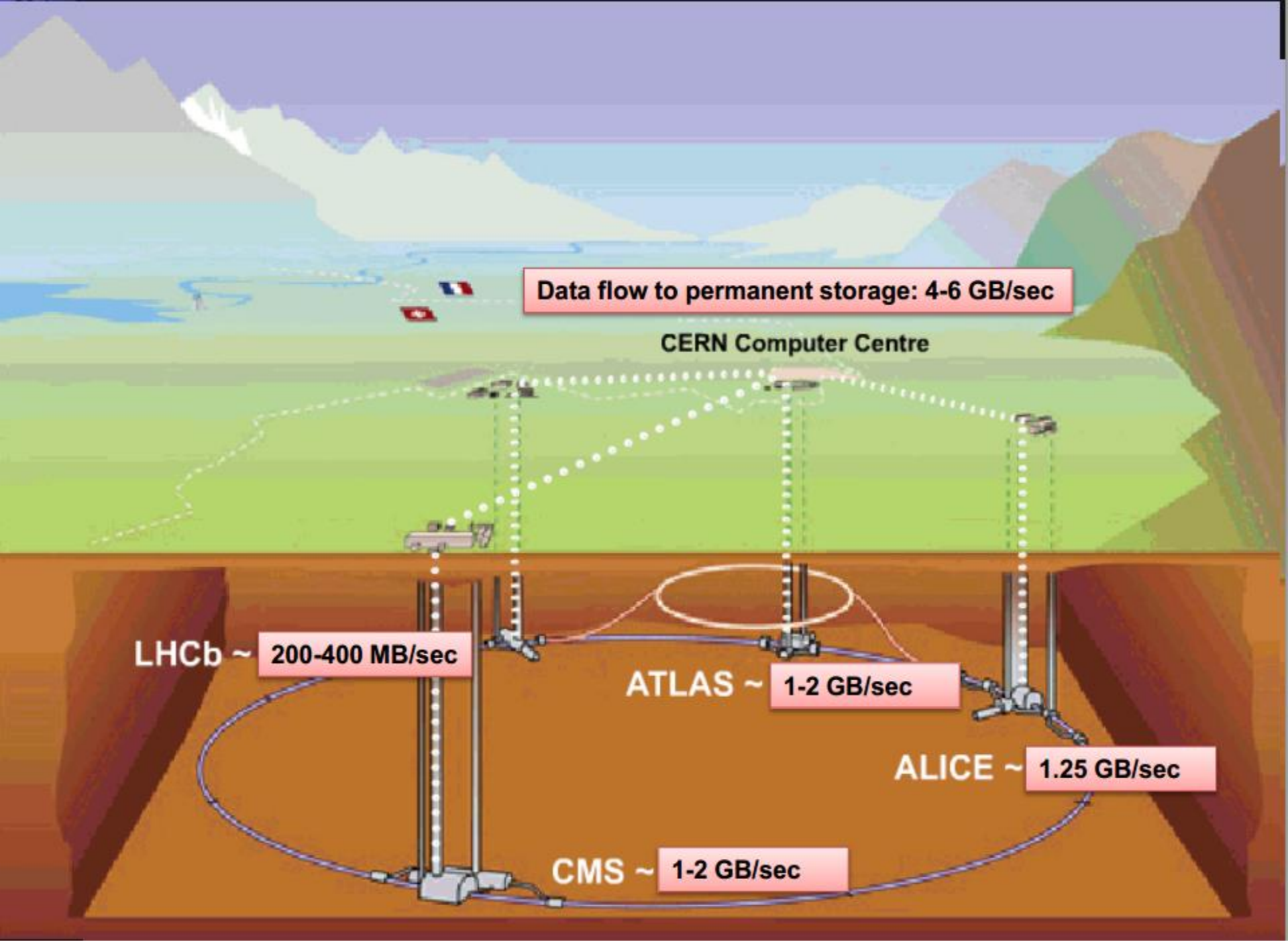
What is CERN ?

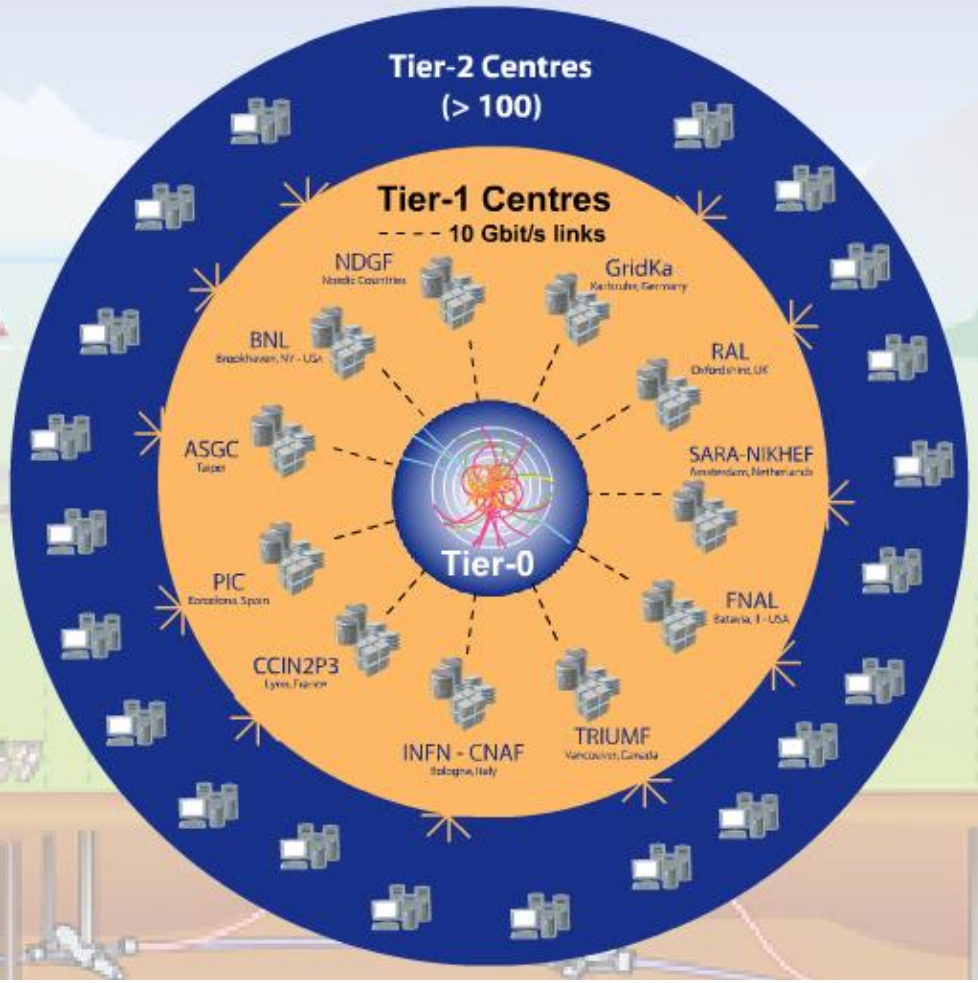
- Conseil Européen pour la Recherche Nucléaire – aka European Laboratory for Particle Physics
- Founded in 1954 with an international treaty
- Between Geneva and the Jura mountains, straddling the Swiss-French border
- Our business is fundamental physics , what is the universe made of and how does it work

The Large Hadron Collider






Data flow to permanent storage: 4-6 GB/sec
CERN Computer Centre
LHCb ~ 200-400 MB/sec
ATLAS ~ 1-2 GB/sec
ALICE ~ 1.25 GB/sec
CMS ~ 1-2 GB/sec



Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~200 centres):
- Simulation
- End-user analysis

- Data is recorded at CERN and Tier-1s and analysed in the Worldwide LHC Computing Grid
- In a normal day, the grid provides 100,000 CPU days executing over 2 million jobs

The CERN Data Centre in Numbers

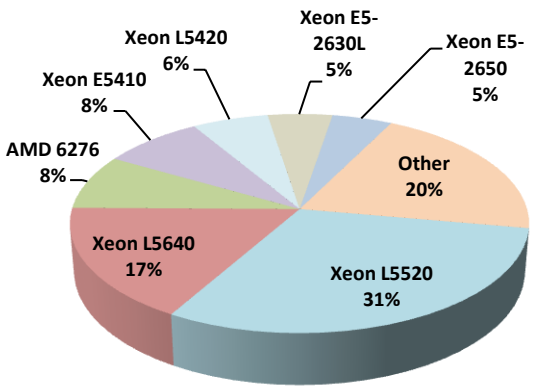
– Hardware installation & retirement
  - ~7,000 hardware movements/year; ~1800 disk failures/year

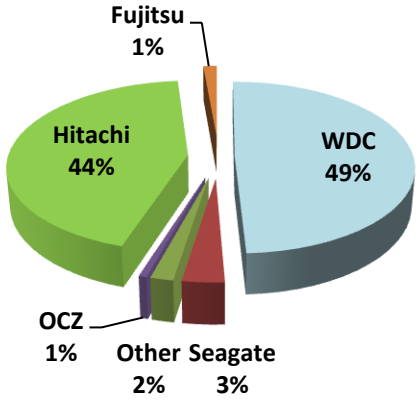
Racks	1127
Servers	10,070
Processors	17,259
Cores	88,414
HEPSpec06	744,277

Disks	79,505
Raw disk capacity (TiB)	124,660
Memory modules	63,326
Memory capacity (TiB)	*298*
RAID controllers	3,091

Tape Drives	120
Tape Cartridges	52000
Tape slots	66000
Data on Tape (PiB)	75

High Speed Routers	29
Ethernet Switches	874
10 Gbps/100Gbps ports	1396/74
Switching Capacity	6 Tbps
1 Gbps ports	27984
10 Gbps ports	5664





IT Power Consumption	2392 KW
Total Power Consumption	3929 KW

Not too long ago….

- Around 10k servers
  - Dedicated compute, dedicated disk server, dedicated service nodes
  - Majority Scientific Linux (RHEL5/6 clone)
  - Mostly running on real hardware
  - Last couple of years, we've consolidated some of the service nodes onto Microsoft HyperV
  - Various other virtualisation projects around
- Many diverse applications ("clusters")
  - Managed by different teams (CERN IT + experiment groups) ….
  - … using our own management toolset
    - Quattor / CDB configuration tool
    - Lemon computer monitoring

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**

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
- with 200Gbit/s network to CERN

Time to change strategy

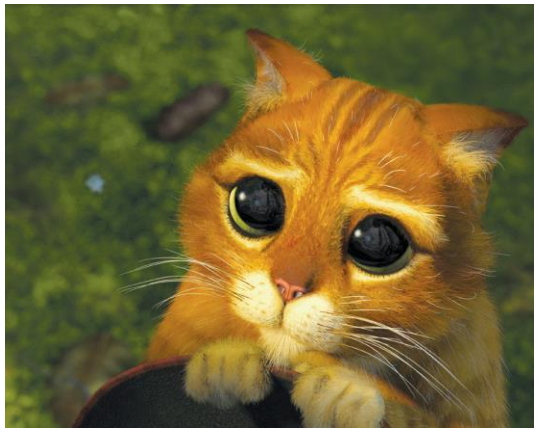
- Rationale
  - Need to manage twice the servers as today
  - No increase in staff numbers
  - Tools becoming increasingly brittle and will not scale as-is
- Approach
  - CERN is no longer a special case for compute
  - Adopt an open source tool chain model
  - Our engineers rapidly iterate
    - Evaluate solutions in the problem domain
    - Identify functional gaps and challenge them
    - Select first choice but be prepared to change in future
  - Contribute new functionality back to the community

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
- Improve responsiveness
  - Self-Service with coffee break response time

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 vm00042.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
  - Looking to use Cinder with Ceph backing store
- Live migration to maximise availability
  - KVM live migration using Ceph

Current Status of OpenStack at CERN

- Grizzly-release, based on RDO packages
  - Excellent experience with the Fedora Cloud SIG and RDO teams
  - Cloud-init for contextualisation, oz for images (Linux and Windows)
- Components
  - Current focus is on Nova with KVM and Hyper-V
  - Glance (Ceph backend), Keystone (Active Directory), Horizon, Ceilometer
  - WIP: Cinder, with Ceph and NetAPP backends
  - Stackforge Puppet components to configure OpenStack
- Production service since Summer 2013
  - Today: 2 Nova cells with 950 hypervisors
  - 3700 VMs integrated with CERN infrastructure
  - Additional Nova Cells for 3000 servers being set up

Project | Admin

System Panel

- Overview
- Instances
- Volumes
- Flavors
- Images
- Projects
- Users
- System Info

Overview

Logged in as: svckey Settings Help

Select a month to query its usage:

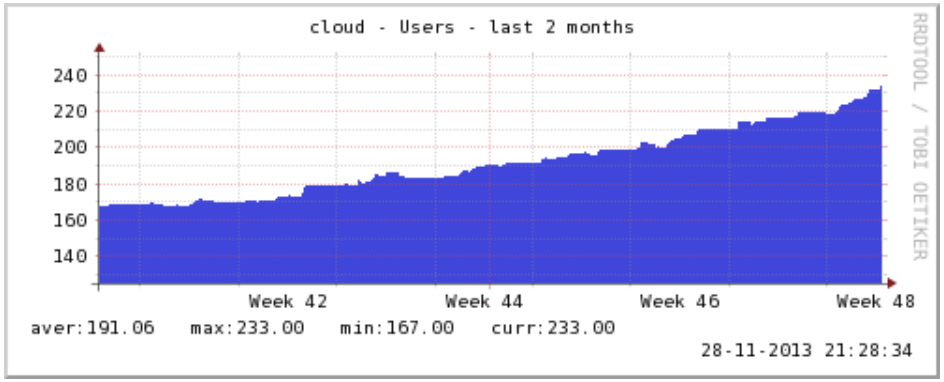
November 2013 Submit

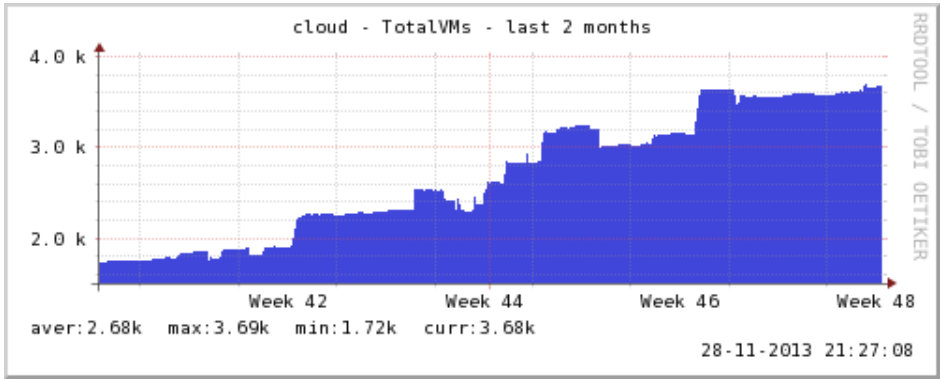
Active Instances: 3769 **Active RAM:** 38TB **This Month's VCPU-Hours:** 2243565.59 **This Month's GB-Hours:** 323587796.31

Usage Summary

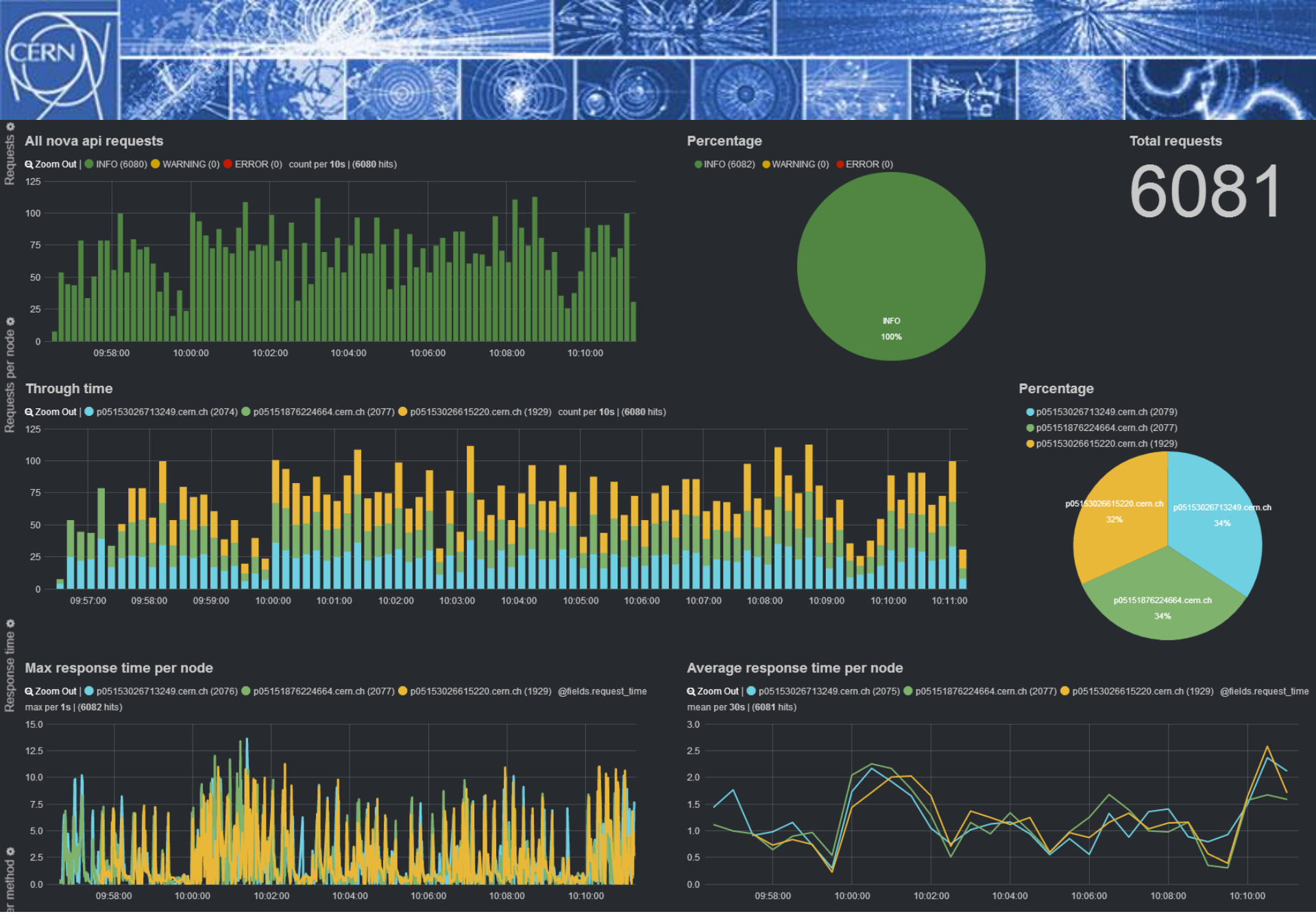
Project Name	VCPUs	Disk	RAM	VCPU Hours
IT Batch - Wigner	8680	249550	16TB	723386.97
IT Batch	4320	121820	8TB	235843.51
IT Batch - shared	1480	42550	2TB	123619.01
ATLAS Cloud Test	993	19860	1TB	79515.95
IT Plus	880	25300	1TB	74139.33
IT Monitoring	283	5660	566GB	63023.68
IT Dashboard	230	8860	458GB	50069.15
NA61 Data production	200	0	100GB	133642.18
LHCb Cloud Workers	137	2950	274GB	80964.44
PH LCGAA	136	4610	269GB	17407.40
IT Configuration Management Services	129	2790	256GB	27766.94
IT Agile CI	106	2120	212GB	28837.11
IT LFC	76	1520	152GB	12361.95

OpenStack production service since August 2013





Cell	Nodes	Cores	VMs
Geneva	655	17776	2619
Wigner	291	9312	1104
Total	**946**	**27088**	**3723**


All nova api requests
Zoom Out | INFO (6080) WARNING (0) ERROR (0) count per 10s | (6080 hits)
125
100
75
50
25
0
09:58:00
10:00:00
10:02:00
10:04:00
10:06:00
10:08:00
10:10:00
Requests
Percentage
INFO (6082) WARNING (0) ERROR (0)
INFO
100%
Total requests
6081
Through time
Zoom Out | p05153026713249.cern.ch (2074) p05151876224664.cern.ch (2077) p05153026615220.cern.ch (1929) count per 10s | (6080 hits)
125
100
75
50
25
0
09:57:00
09:58:00
09:59:00
10:00:00
10:01:00
10:02:00
10:03:00
10:04:00
10:05:00
10:06:00
10:07:00
10:08:00
10:09:00
10:10:00
10:11:00
Requests per node
Percentage
p05153026713249.cern.ch (2079)
p05151876224664.cern.ch (2077)
p05153026615220.cern.ch (1929)
p05153026615220.cern.ch
32%
p05153026713249.cern.ch
34%
p05151876224664.cern.ch
34%
Max response time per node
Zoom Out | p05153026713249.cern.ch (2076) p05151876224664.cern.ch (2077) p05153026615220.cern.ch (1929) @fields.request_time
max per 1s | (6082 hits)
15.0
12.5
10.0
7.5
5.0
2.5
0.0
09:58:00
10:00:00
10:02:00
10:04:00
10:06:00
10:08:00
10:10:00
Response time
Average response time per node
Zoom Out | p05153026713249.cern.ch (2075) p05151876224664.cern.ch (2077) p05153026615220.cern.ch (1929) @fields.request_time
mean per 30s | (6081 hits)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
09:58:00
10:00:00
10:02:00
10:04:00
10:06:00
10:08:00
10:10:00
per method

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

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 to Folsom release
- Now in use in our production instance
  - Map project roles (admins, members) to groups
  - Documentation in the OpenStack wiki

What about Hyper-V?

- We have used Hyper-V/System Centre for our server consolidation activities
  - 3400 VMs (2000 Linux, 1400 Windows)
  - 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, …

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
- ATLAS and CMS have deployed OpenStack on their farm
  - Simulation (low I/O, high CPU)
  - Analysis (high I/O, high CPU, high network)

ATLAS Sim@P1	1,200 Servers	28,800 cores (HT)
CMS OOOO cloud	1,300 Servers	13,000 cores
CERN IT Grizzly cloud	946 Servers	27,778 cores (HT)

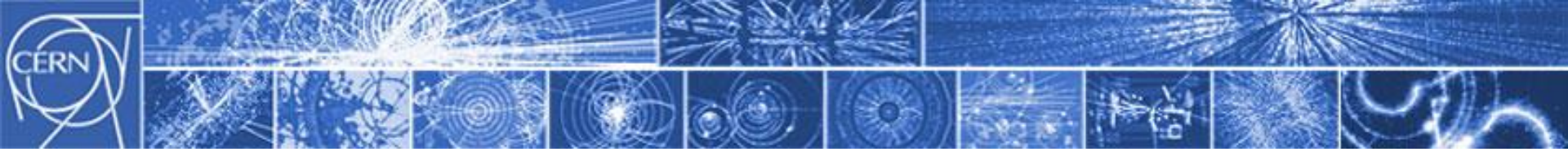
Upcoming challenges

- Upgrade to Havana
- Exploit new functionality
  - Deploy Cinder, with Ceph and NetAPP backends
  - Kerberos, X.509 user certificate authentication
  - OpenStack Heat
  - Replace nova-network by Neutron
  - Federated clouds
  - Keystone Domains to devolve administration
  - Bare metal for non-virtualised use cases such as high I/O servers
  - Load Balancing as a service

Ramping to 15,000 hypervisors with 100,000 VMs by end-2015

Conclusions

- OpenStack is in production at CERN
  - Work together with others on scaling improvements
- Community is key to shared success
  - Our problems are often resolved before we raise them
  - Packaging teams are producing reliable builds promptly
- CERN contributes **and** benefits
  - Thanks to everyone for their efforts and enthusiasm
  - Not just code but documentation, tests, blogs, …

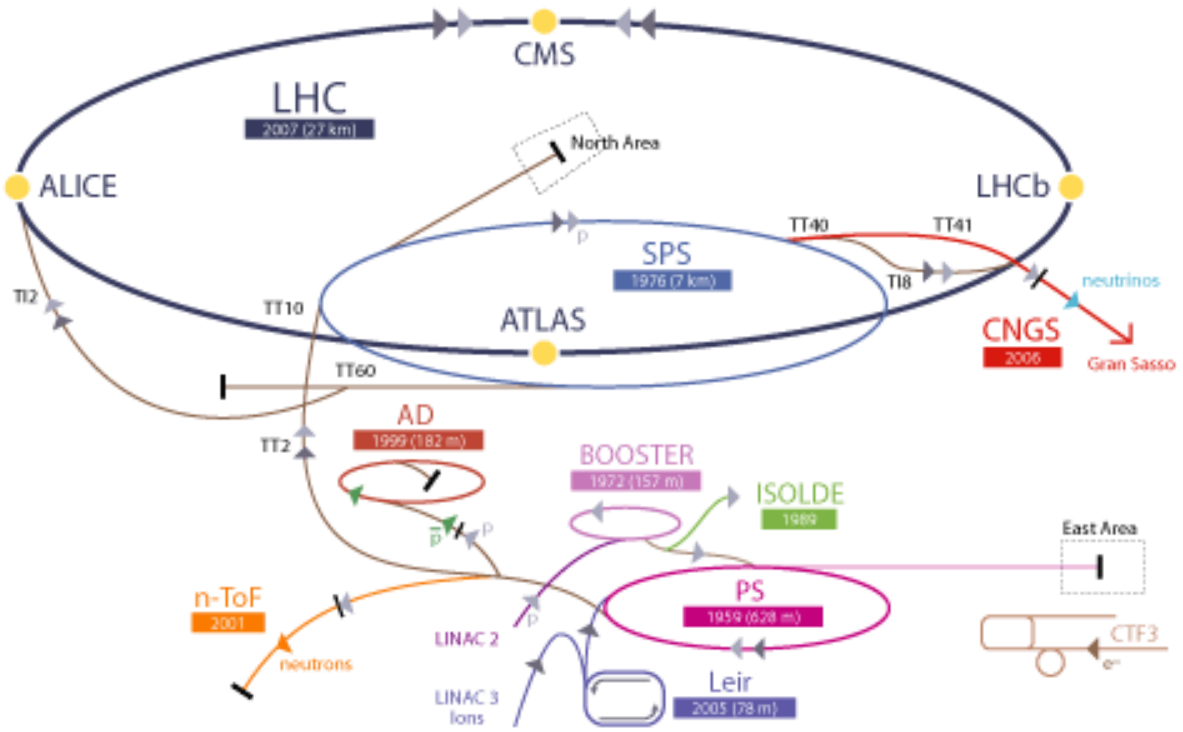


Backup Slides

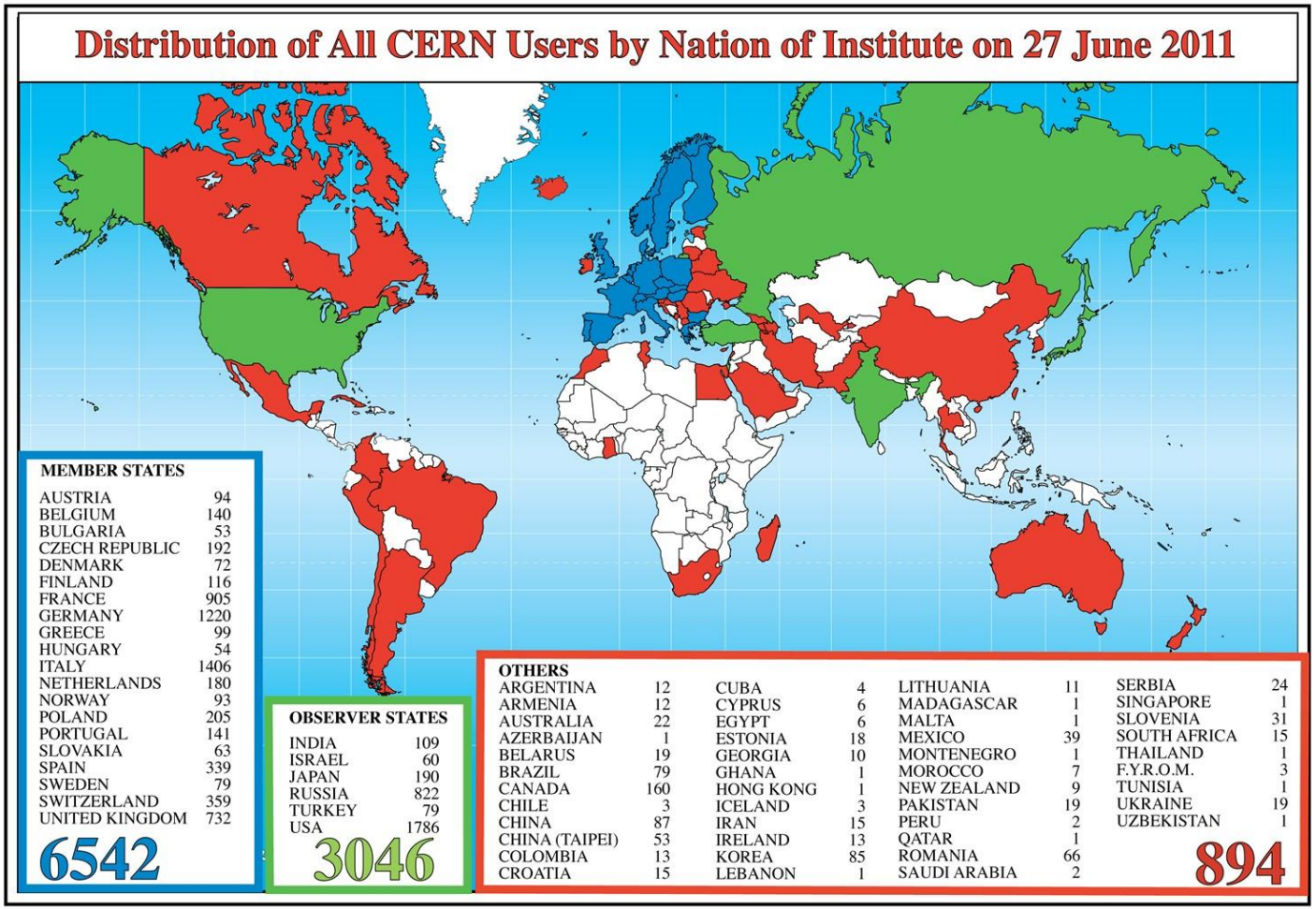
References

CERN	http://public.web.cern.ch/public/
Scientific Linux	http://www.scientificlinux.org/
Worldwide LHC Computing Grid	http://lcg.web.cern.ch/lcg/ http://rtm.hep.ph.ic.ac.uk/
Jobs	http://cern.ch/jobs
Detailed Report on Agile Infrastructure	http://cern.ch/go/N8wp
HELiX Nebula	http://helix-nebula.eu/
EGI Cloud Taskforce	https://wiki.egi.eu/wiki/Fedcloud-tf

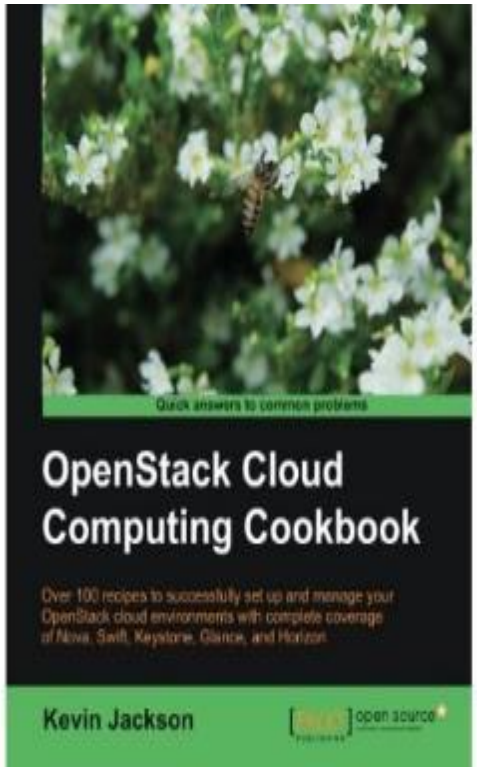
CERN Accelerator Complex

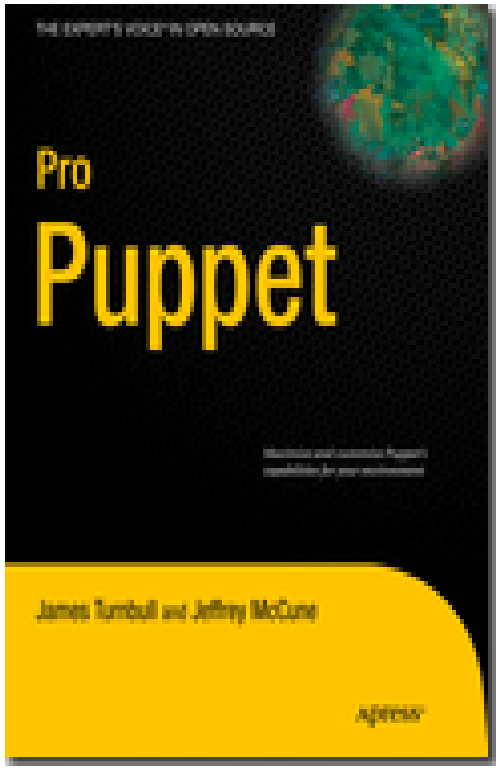


Community collaboration on an international scale



Training for Newcomers





Buy the book rather than guru mentoring

Job Opportunities