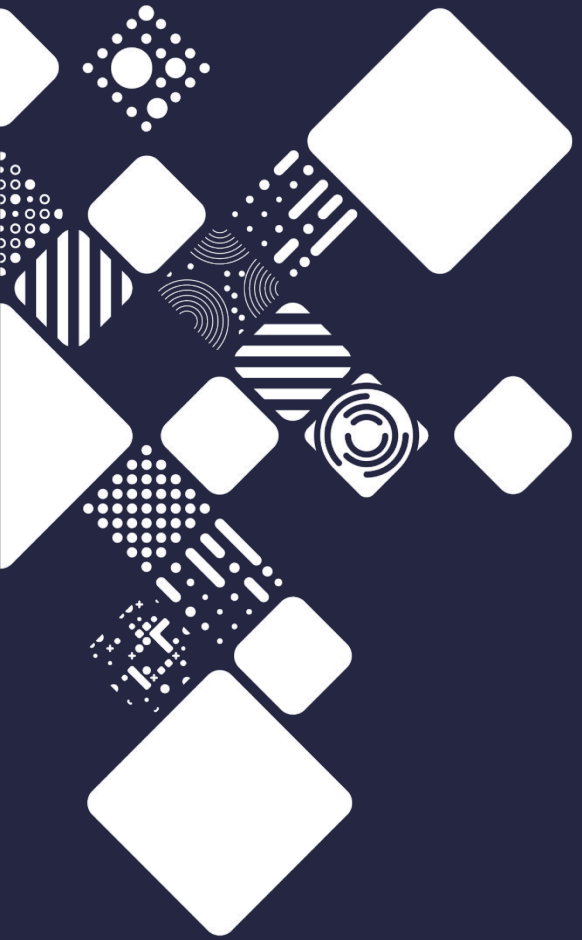


Supporting DNA sequencing at scale with OpenStack

Dave Holland

wellcome sanger institute

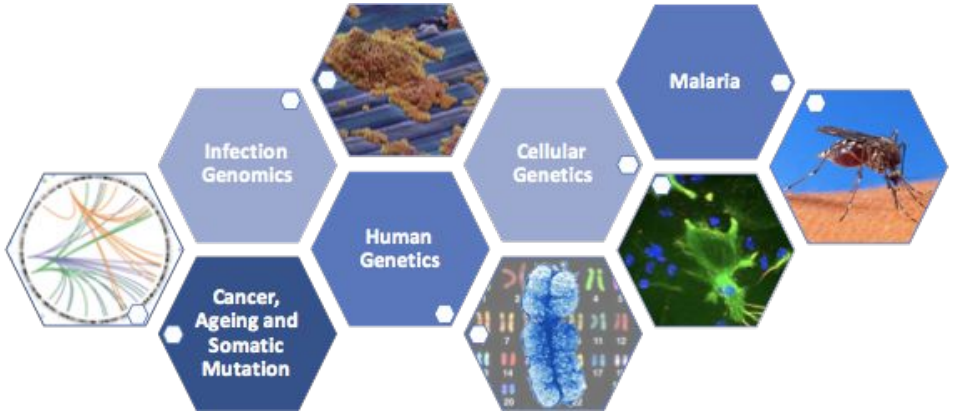
What I'll talk about

- The Sanger Institute's work
- Motivations for using OpenStack
- Our journey
- Our projects that are using OpenStack
- Next steps

wellcome sanger institute

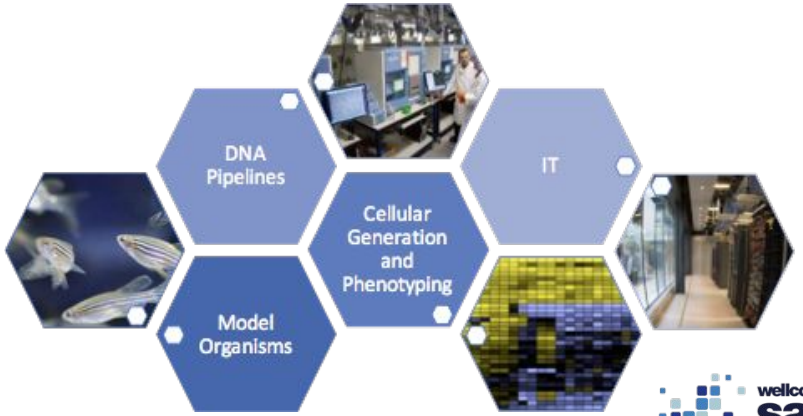
wellcome sanger institute

Sanger Science

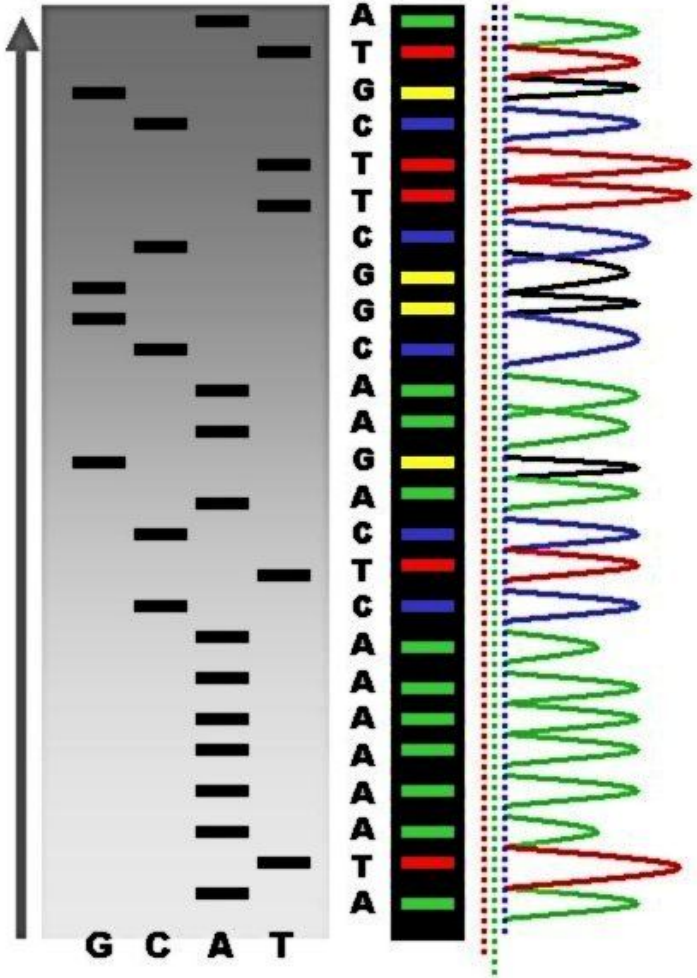


Scientific Research Programmes

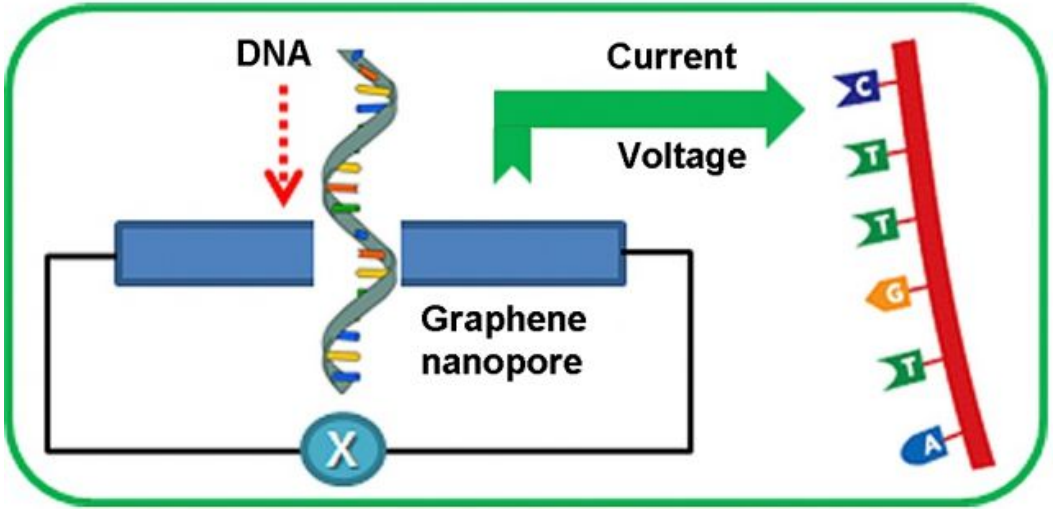
Core Facilities



wellcome sanger institute



DNA sequencing (the 30-second summary)



wellcome sanger institute

Then...

capillary sequencer
~2Mb / day

...and now

Illumina NovaSeq 6000

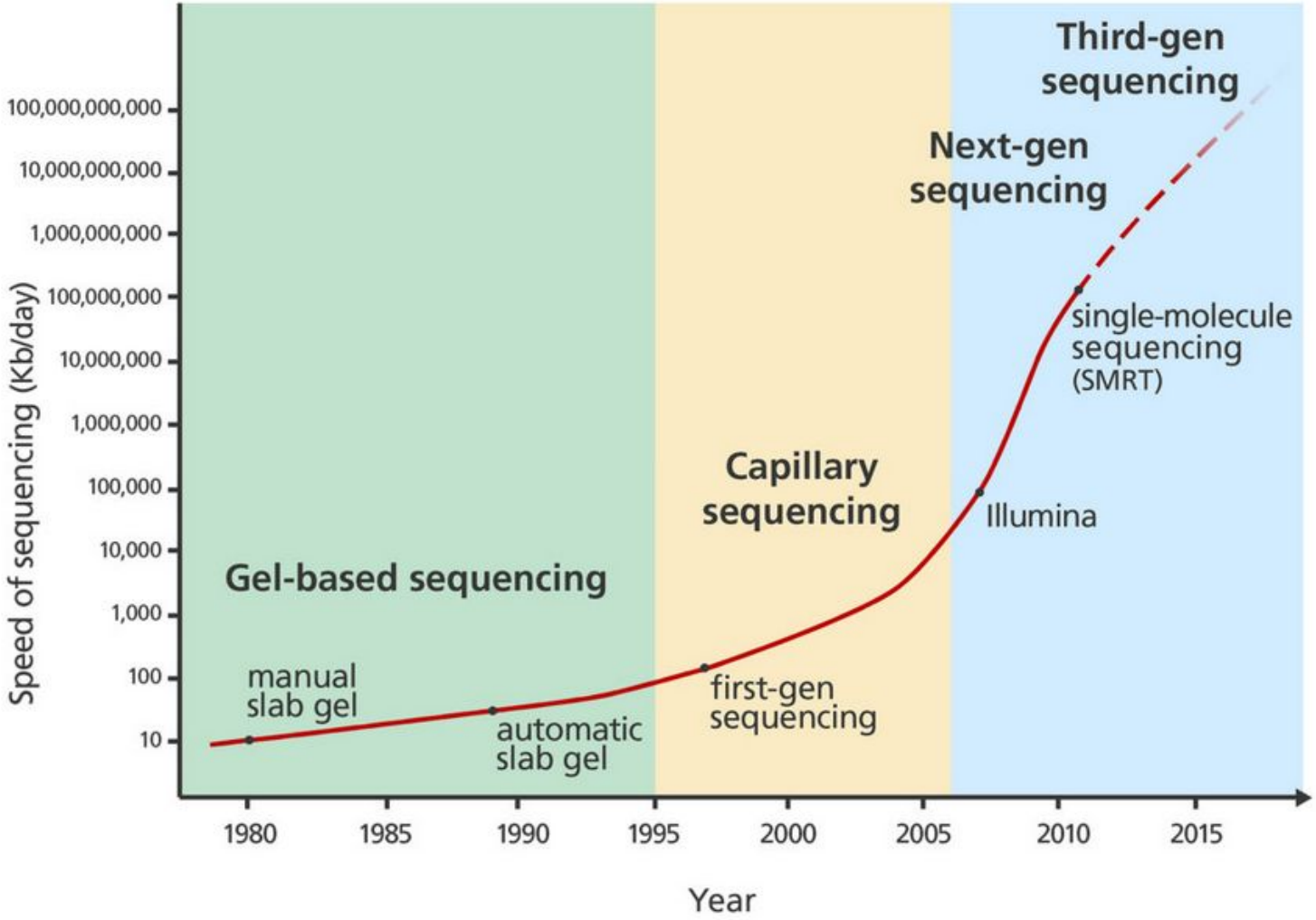
~3Tb per sequencer / day

(2-day cycle)

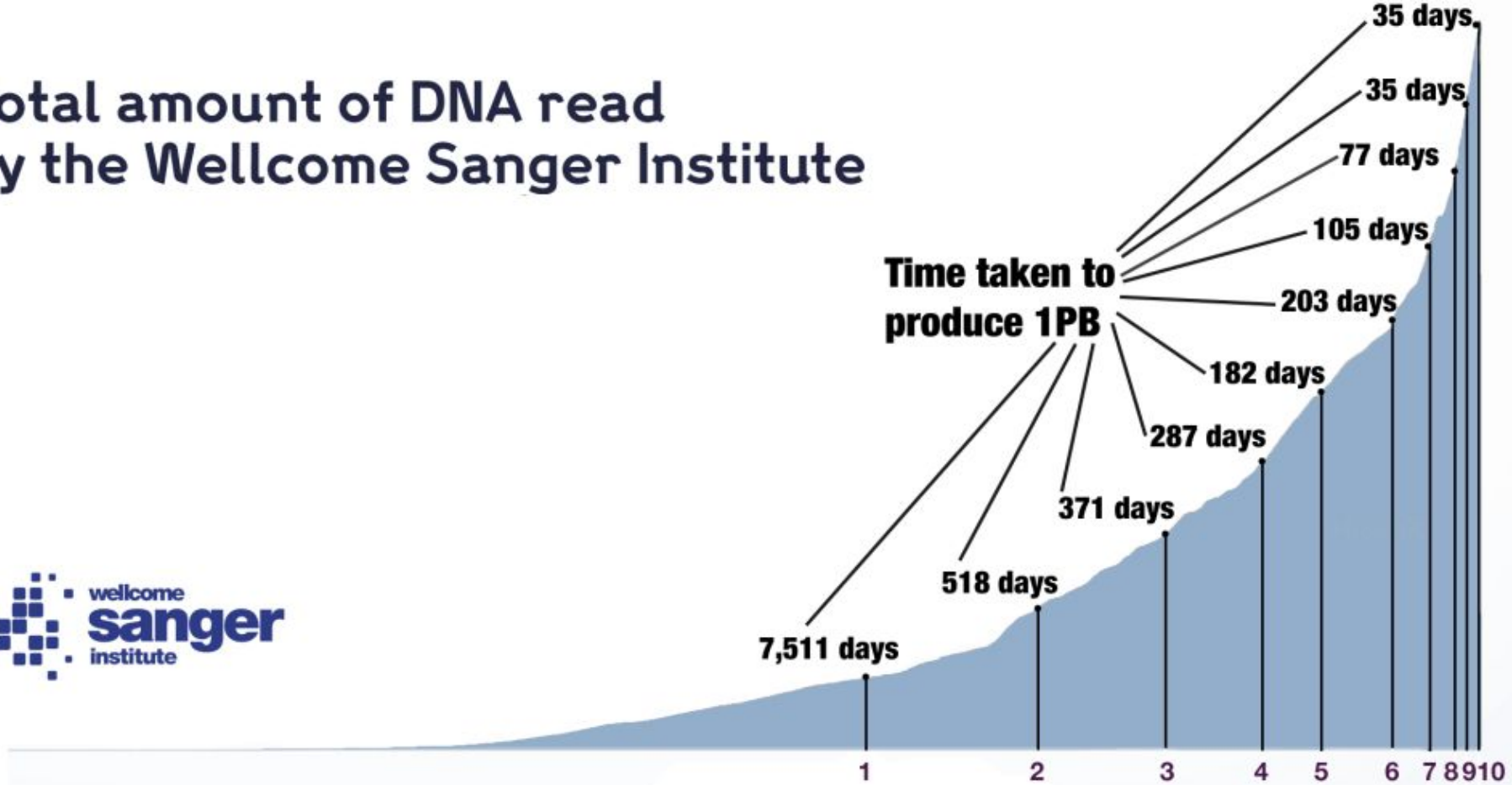
wellcome sanger institute

Oxford Nanopore
MinION and SmidgION
up to 20Gb / 48 hours

wellcome sanger institute


Third-gen sequencing
Next-gen sequencing
Capillary sequencing
Gel-based sequencing
single-molecule sequencing (SMRT)
Illumina
first-gen sequencing
automatic slab gel
manual slab gel
Speed of sequencing (Kb/day)
100,000,000,000
10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1980
1985
1990
1995
2000
2005
2010
2015
Year

Total amount of DNA read by the Wellcome Sanger Institute



wellcome sanger institute

Cumulative total of Petabases of DNA sequenced by the Wellcome Sanger Institute

UK Biobank - Vanguard

Whole genome sequencing will 'transform the research landscape for a wide range of diseases'

UK Biobank enters exciting new phase of genetics research

News article 6 Apr 2018

biobank uk

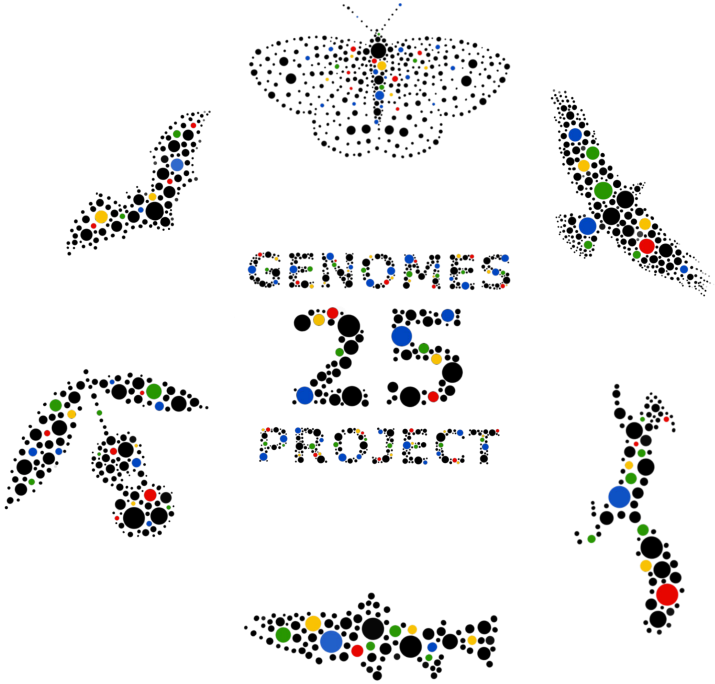
Improving the health of future generations

UK Biobank is working with the Wellcome Sanger Institute to sequence the full genomes of 50,000 UK Biobank volunteer participants

50,000 whole-genome sequences over 18 months

wellcome sanger institute

25 Genomes Project



25 previously unsequenced UK organisms:

- giant hogweed
- red squirrel
- golden eagle
- European robin
- ...


Darwin Tree of Life
Project (UK)

EARTH BIOGENOME PROJECT

ABOUT EBP GOALS WORK + PROGRESS

MEDIA + PUBLICATIONS EVENTS CONTACT

A MOONSHOT FOR BIOLOGY

Sequence the DNA of all life on Earth in 10 years

Motivations for using OpenStack

wellcome sanger institute

The Sanger Institute
Traditional HPC Environment

LSF 9

~10,000 cores in main compute farm

~10,000 cores across smaller project-specific farms

~15PB Lustre high-performance storage

Limited security - “isolation” is only POSIX file permissions

Limited flexibility - no root access, incompatible software dependencies

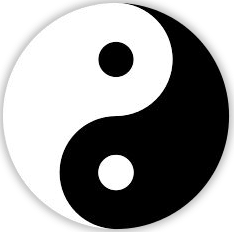
Pipelines and stacks are complex - scientific reproducibility is hard

wellcome sanger institute

HPC and cloud computing are complementary

Traditional HPC

- Highest possible performance
- A mature and centrally managed compute platform
- High-performance Lustre filesystems for data intensive analysis



Cloud compute

- Full segregation of projects ensures data security
- Developers no longer tied to a single stack
- Reproducibility through containers / images and infrastructure-as-code

But there's a catch or two...

- Large number of traditional/legacy pipelines
  - They require a performant shared POSIX filesystem, while cloud workloads support object stores
- We do not always have the source code or expertise to migrate
- We need multi-gigabyte per second performance
- The tenant will have root
  - and could impersonate any user, but Lustre trusts the client's identity assertions, just like NFSv3
- The solution must be simple for the tenant and administrator

wellcome sanger institute

Our journey

wellcome sanger institute

Step by step

2015: sysadmin training and experiments

2016: trial system offered to scientists (Kilo)

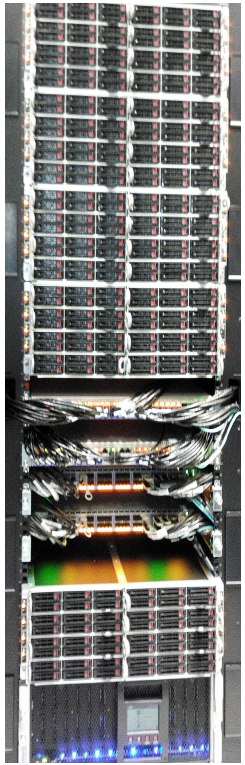
2017: production Flexible Compute Environment (Liberty)

2018: move to Pike; added hypervisors and storage

2019: move to Queens; Ceph upgrade

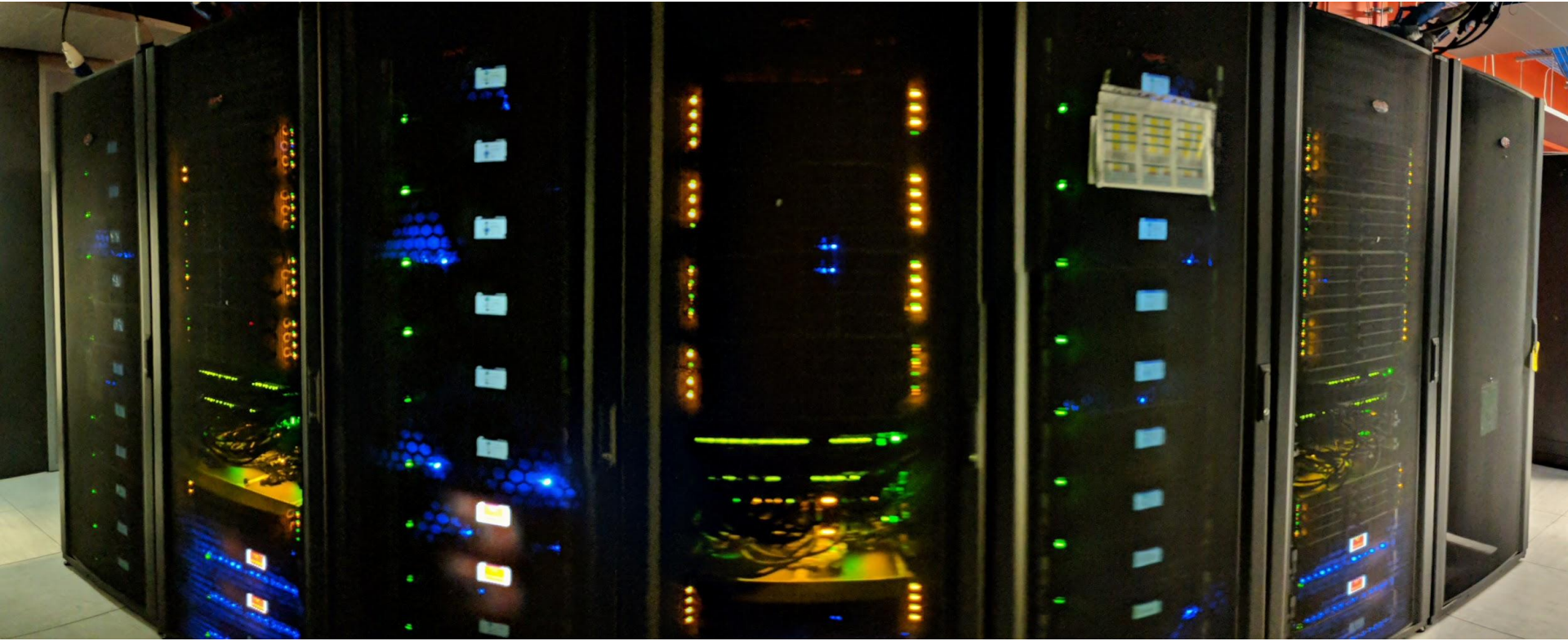
wellcome sanger institute

Production hardware



- 107 compute nodes (Supermicro) each with:
  - 512GB RAM, 2x 25GbE, 2x Intel E52690v4 (14 cores @ 2.6GHz)
- 6 control nodes (Supermicro)
  - 256GB RAM, 2x 100GbE, NVMe
- Total 53 TB RAM, 2996 cores, 5992 with hyperthreading
- 9 Ceph nodes running “Jewel”
  - 60x 6TB SATA, 2x 2TB NVMe, 2x 100GbE
- 1PB usable, 3-way replicated
- Arista 7060CX switches, spine/leaf

wellcome sanger institute



wellcome sanger institute

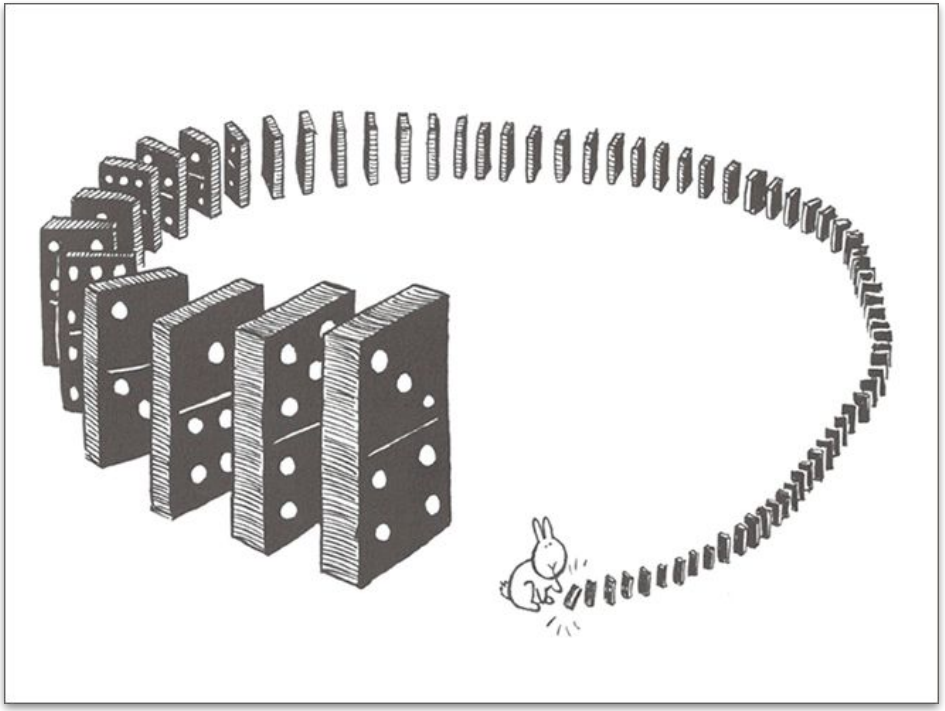
Challenges

Liberty (RHOSP8) - “Delta”

- default file descriptor limits for Glance, Cinder, RabbitMQ etc way too low
- NIC hardware acceleration works fine with the correct kernel/driver (thanks to Mellanox)
- races when scheduling many instances; mitigate by directing nova_osapi to a “primary” server
- faulty fibres caused NIC flaps and RabbitMQ problems

wellcome sanger institute

RabbitMQ



Andy Riley's "Bunny Suicides"

Liberty (RHOSP8) - “Delta”

- not enough hypervisor memory reserved, neutron-openvswitch-agent fails to allocate memory
- various package upgrades (Glance, Neutron, Dnsmasq etc) for particular bugs encountered or security fixes
- work with Arista on portchannel/LACP instability

wellcome sanger institute

Ceph

- standalone Ceph to support multiple OpenStack deployments
- generally robust in the face of spontaneous machine hangs and rack failures
- OSD start-up race due to single shared lock
- niggles with ceph-ansible

Ansible for customisation

- scheduler tweaks (stack not spread, CPU/RAM overcommit)
- hypervisor tweaks (instance root disk on Ceph or hypervisor)
- enable SSL for Horizon and API
- change syslog destination
- add “MOTD” to Horizon login page
- change session timeouts
- register systems with Red Hat
- and more…

- but deployer’s Puppet overwrites some of these

Monitoring

- evolved from “bare minimum” to customer and engineer views, scorecard, availability report
- custom Nagios scripts, active and passive checks
- Grafana for metrics
- rsyslog and ELK for logging

wellcome sanger institute

HPC Services

farm3 LSF	scratch114	Zeta VM creation	Zeta controllers
farm4 LSF	scratch115	S3 interface	Zeta compute
cgp LSF	scratch116	Cinder (volume)	Ceph nodes
pcs5 LSF	scratch117	Glance (image)	CloudForms VMs
sf2 LSF	scratch118	Neutron (network)	Zeta undercloud
vr LSF	scratch119	Horizon (web UI)	CVX servers
		CloudForms	
globus			Eta networkers
ELK stack	farm3 login nodes	Eta VM creation	Eta controllers
UKB seq-stor	farm4 login nodes	Eta web UI	Eta compute
UKB lus13	iRODS		Eta undercloud

- OK
- Warning Threshold Breached
- Critical Threshold Breached
- Acknowledged/Maintenance
- No Data Available

Host	Service	Status	Duration	Attempt	Last Check	Status Information
zeta.internal.sanger.ac.uk	Cinder API Health	Ok	3d 15h 48m 50s	1/5	09/11/2018 12:21:55	HTTP OK: Status line output matched "200" - 754 bytes in 0.196 second response time
	Glance-API and Glance-Registry API Health	Ok	3d 15h 49m 30s	1/5	09/11/2018 12:22:47	HTTP OK: Status line output matched "200" - 22609 bytes in 0.244 second response time
	HAProxy Health	Ok	3d 15h 51m 31s	1/5	09/11/2018 12:20:57	HAPROXY OK - nova_placement (Active: 3/3) glance_api (Active: 3/3) heat_cfn (Active: 3/3) gnocchi (Active: 3/3) nova_novncproxy (Active: 3/3) nova_metadata (Active: 3/3) redis (Active: 1/3) keystone_admin (Active: 3/3) panko (Active: 1/3) heat_api (Activ
	Heat API Health	Ok	3d 15h 50m 26s	1/5	09/11/2018 12:24:11	HTTP OK: Status line output matched "200" - 223 bytes in 0.274 second response time
	Horizon Web GUI Health	Ok	3d 15h 50m 35s	1/5	09/11/2018 12:21:55	HTTP OK: HTTP/1.1 200 OK - 9690 bytes in 0.036 second response time
	Instance Creation	Ok	3d 3h 3m 58s	1/1	09/11/2018 12:20:43	OK success - test duration 40.253437289 seconds
	Keystone-Public-SSL API Health	Ok	3d 15h 52m 3s	1/5	09/11/2018 12:21:23	HTTP OK: Status line output matched "200" - 513 bytes in 0.010 second response time
	Neutron agent status	Ok	3d 15h 50m 36s	1/1	09/11/2018 12:25:03	OK
	Neutron API Health	Ok	3d 15h 48m 7s	1/5	09/11/2018 12:24:09	HTTP OK: Status line output matched "200" - 265 bytes in 0.008 second response time
	Nova service status	Ok	3d 15h 50m 36s	1/1	09/11/2018 12:25:03	OK
	OpenStack database backup status	Ok	10d 2h 52m 12s	1/5	09/11/2018 12:24:35	OK latest backup overcloud_db_zeta-ctrl0_20181109_010001.tar.bz2 is 11 hours old
	Ping	Ok	3d 15h 52m 46s	1/5	09/11/2018 12:24:54	PING OK - Packet loss = 0%, RTA = 0.16 ms
	SSH	Ok	3d 15h 54m 27s	1/5	09/11/2018 12:21:16	SSH OK - OpenSSH_7.4 (protocol 2.0)
	SSL Certificate	Ok	3d 15h 52m 53s	1/5	09/11/2018 12:24:32	OK - Certificate '*.internal.sanger.ac.uk' will expire on Sat 18 Jan 2020 11:11:00 GMT.

Metrics

- collectd + libvirtd + Python graphitesend
- custom scripts to aggregate CPU use for efficiency “score”
- need to scale up number of carbon-cache processes
- Grafana dashboards

wellcome sanger institute

Grafana - Zeta overview

Secure | https://metrics.internal.sanger.ac.uk/dashboard/db/zeta-overview?refresh=5m&orgId=1

Zeta overview

Zoom Out | now-3h to now-1m | Refresh every 5m

Total hypervisor load average

2230.6

VCPUs in use.

6950

Memory used

36.9 TiB

Instances

553

Overloaded

1.5

Cluster efficiency

32.1

m1.3xlarge

0

m1.2xlarge

0

m1.xlarge

5

m1.large

12

VCPUs created per tenant

VCPUs

8 K, 6 K, 4 K, 2 K, 0

13:00, 13:30, 14:00, 14:30, 15:00, 15:30

aker_prod Current: 30 aker_test Current: 102 casm_general Current: 152
casm_wr Current: 16 cellgeni Current: 238 cgp_farm Current: 446 cgp_prostate Current: 45
cgp_wr Current: 38 cloudforms_cancer Current: 4 cloudforms_npg Current: 62
corebio Current: 1 cosmic Current: 81 genserv_ci Current: 38 hgi Current: 108
hgi_ci Current: 32 hgi_dev Current: 56 hgiarvados Current: 2.74 K humgen Current: 4
jjnaristadev Current: 2 mo_dev Current: 63 nagios Current: 1 npg Current: 171
npg_dev Current: 47 npg_esa Current: 572 pathogen Current: 4
sequencescape_test Current: 14 signal_ci Current: 13 ssg_cgp_irods_41_dev Current: 80

Memory allocated per tenant

38 TiB, 29 TiB, 19 TiB, 10 TiB, 0 MiB

13:00, 13:30, 14:00, 14:30, 15:00, 15:30

aker_prod Current: 31 GiB aker_test Current: 107 GiB casm_general Current: 715 GiB
casm_wr Current: 76 GiB cellgeni Current: 1.8 TiB cgp_farm Current: 861 GiB
cgp_prostate Current: 378 GiB cgp_wr Current: 201 GiB cloudforms_cancer Current: 34 GiB
cloudforms_npg Current: 65 GiB corebio Current: 8 GiB cosmic Current: 254 GiB
genserv_ci Current: 84 GiB hgi Current: 157 GiB hgi_ci Current: 269 GiB
hgi_dev Current: 59 GiB hgiarvados Current: 17.5 TiB humgen Current: 34 GiB
jjnaristadev Current: 2 GiB mo_dev Current: 294 GiB nagios Current: 8 GiB
npg Current: 1.3 TiB npg_dev Current: 204 GiB npg_esa Current: 4.9 TiB

Grafana - Zeta overview | Grafana - Zeta project report

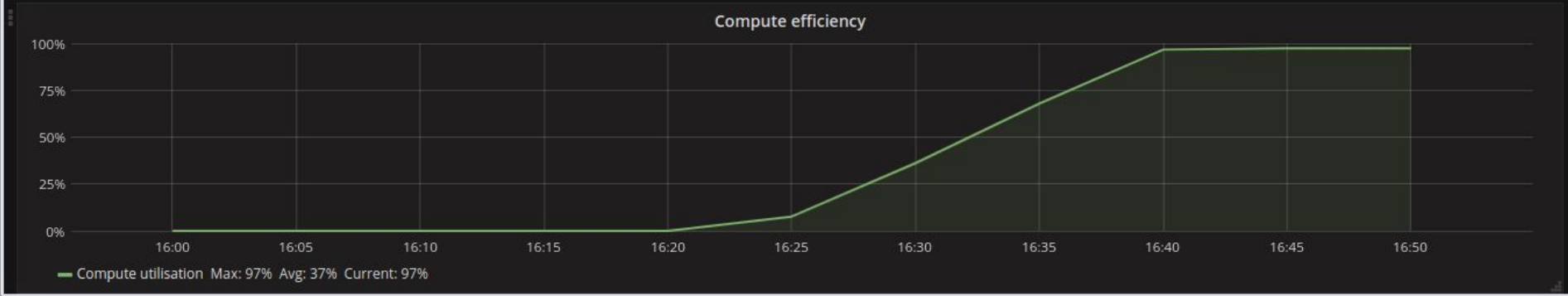
https://metrics.internal.sanger.ac.uk/dashboard/db/zeta-project-report?refresh=5m&orgId=1&var-project=ssg_isg&from=now-1h&...

Apps | Bookmarks | Send Link | Calendar – dh3@sa | Stuff | VMware | Google | Safari | NetApp | sciaas | Other bookmarks

Zeta project report | Zoom Out | now-1h to now-1m | Refresh every 5m

project | ssg_isg

Compute efficiency	VCPU	Memory used	Block storage	Space used by glance
97.3%	2218	24.6 TiB	1.709 GiB	16.0 GiB



Metrics | Arista Telemetry

https://arista-cvp.internal.sanger.ac.uk/telemetry/metrics?from=1541069338267&intfMetrics=INTERFACE_OUT_BITRATE~INTERFACE...

Apps | Bookmarks | Send Link | Calendar – dh3@sa | Stuff | VMware | Google | Safari | NetApp | sciaas | Other bookmarks

ARISTA | Events | Devices | Metrics | CloudTracer

Metric type: Interfaces

Viewing 4 metrics for 1 interface

Metric view: Graphs grouped by interface

Metrics

Metric or group

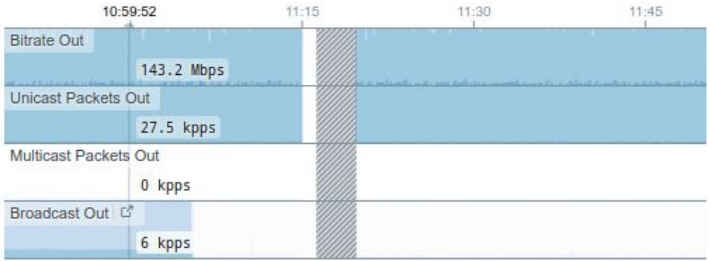
- [ ] Multicast Packets In
- [ ] Broadcast In
- [x] Tx Traffic Rate
  - [x] Bitrate Out
  - [x] Unicast Packets Out
  - [x] Multicast Packets Out
  - [x] Broadcast Out

Interfaces

Device or interface

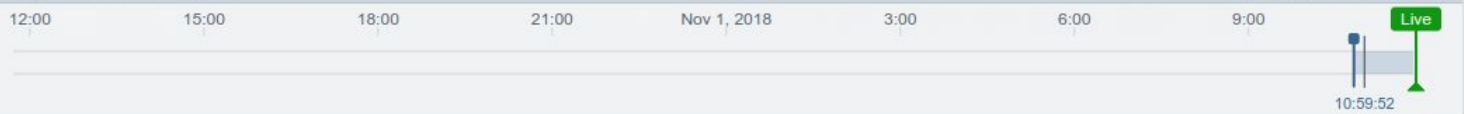
- [ ] Ethernet16/3
- [ ] Ethernet16/4
- [x] Ethernet17/1
- [ ] Ethernet17/2
- [ ] Ethernet17/3
- [ ] Ethernet17/4
- [ ] Ethernet18/1

Ethernet17/1 on leaf-3-1



Nov 1, 2018 10:48:58 – Now

Show Last: 1h 30m 5m 30s



Upgrades and expansions

wellcome sanger institute

OpenStack growth

- entire system physically relocated in July 2018 due to power requirements
- 99 compute nodes added during Q4 2018
- newer hardware generation
  - different CPU/RAM ratio ⇒ new flavours
  - AVX512

wellcome sanger institute

wellcome sanger institute

Queens (RHOSP13) - “Eta”

- current production system
- many of our original customisations and fixes no longer necessary
- sideways-upgrade was time-consuming but gave reassurances of stability and known process

wellcome sanger institute

Queens (RHOSP13) - “Eta”

- containerised overcloud services mean different ways of managing/customising services
- composable roles - RabbitMQ cluster separate from tenant network routers
- tenant networks can be VLAN via Arista ml2 plugin (but still VXLAN by default)
- enabled CPU overcommit, 8x, “o” flavours
- enabled jumbo frames

wellcome sanger institute

Ceph growth

- expansion: 1PB → 4.5PB usable
  - 9 servers → 51 servers
  - 540 OSDs → 3060 OSDs
- current use: 1.8PB as S3 objects, 200TB as Cinder volumes

wellcome sanger institute

Ceph upgrades

- in-place and automated (after extensive testing)
- version 10 (Jewel) to 12 (Luminous)
  - for bug fixes and features
  - completed
- storage format change: FileStore to BlueStore
  - for performance
  - underway

wellcome sanger institute

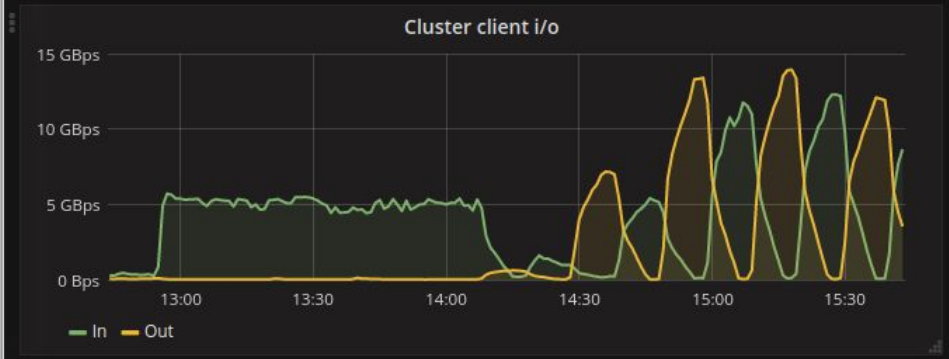
Grafana - Ceph status

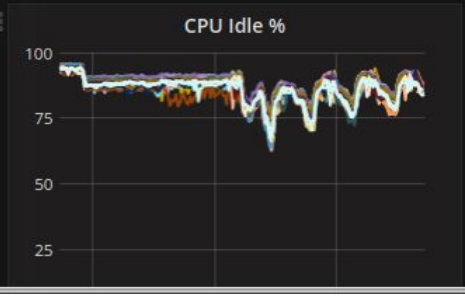
Secure | https://metrics.internal.sanger.ac.uk/dashboard/db/ceph-status-dh3-copy?refresh=5m&orgId=1&from=now-3h&to=now

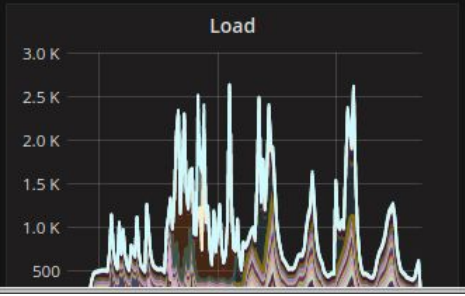
Apps | Bookmarks | Send Link | Exchange calendar | Stuff | VMware | HG | vc5red | Google | Safari | Keep | NetApp | Emacs keybindings | Other bookmarks

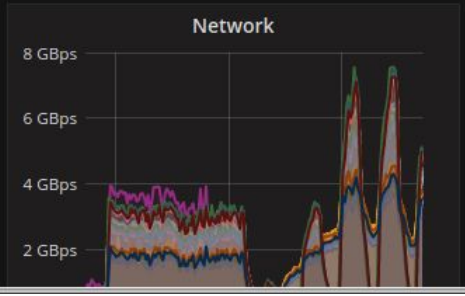
Ceph status dh3 copy | Zoom Out | Last 3 hours | Refresh every 5m

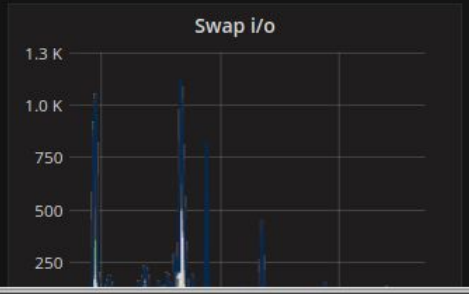
OSDs In	OSDs Up	Monitor quorum	% cluster space used
3060	3060	3	9











Enhancements

Ceph

- automated radosgw/S3 public bucket scan

```
From: ssg-isg@sanger.ac.uk
Date: Sun, 1 Apr 2018 00:01:26 +0100
To: dh3@sanger.ac.uk
Subject: S3 public-readable bucket warning

Hello dh3,

A recent review showed that you have public-readable rgw/S3 bucket(s):

   frobnitz public

This means that the contents are accessible by anyone on the Internet,
without authentication. If this is not your intention, please address
this URGENTLY, e.g. with both these commands:

  s3cmd setacl s3://bucketname --acl-private
  s3cmd setacl s3://bucketname --acl-private -r
```

Provider networks

- “Secure Lustre” - strong tenant isolation for a POSIX filesystem
- scientific instruments
- next iteration of LSF compute clusters
- niggles with security groups
- “a hammer for all screws”? e.g. faster S3 access

wellcome sanger institute

“But what’s it all for?”

Services and applications implemented on OpenStack

- Mutational Signatures
- HGI Arvados and Spark/Hail
- widespread test, dev, CI (Gitlab runners)
- dynamic expansion of LSF clusters
- Mattermost
- CellphoneDB - receptor/ligand database
- CloudForms
- ...

wellcome sanger institute

signal

Home Explore Analyse

Analyse > Jobs > test

test

Filter out kataegis mutations

Submitted

2018-3-27 at 11:53

Status

Complete

Reference build

GRCh37

Filename

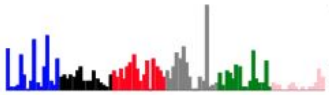
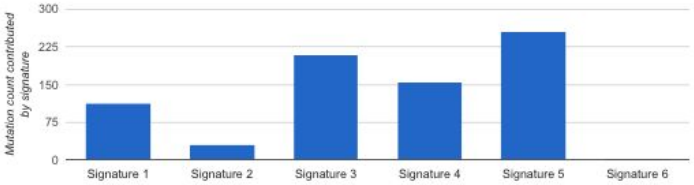
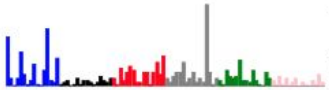
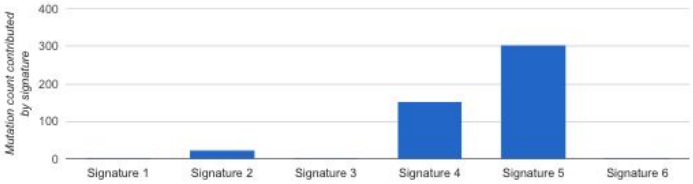
test.vcf.gz

Mutation type

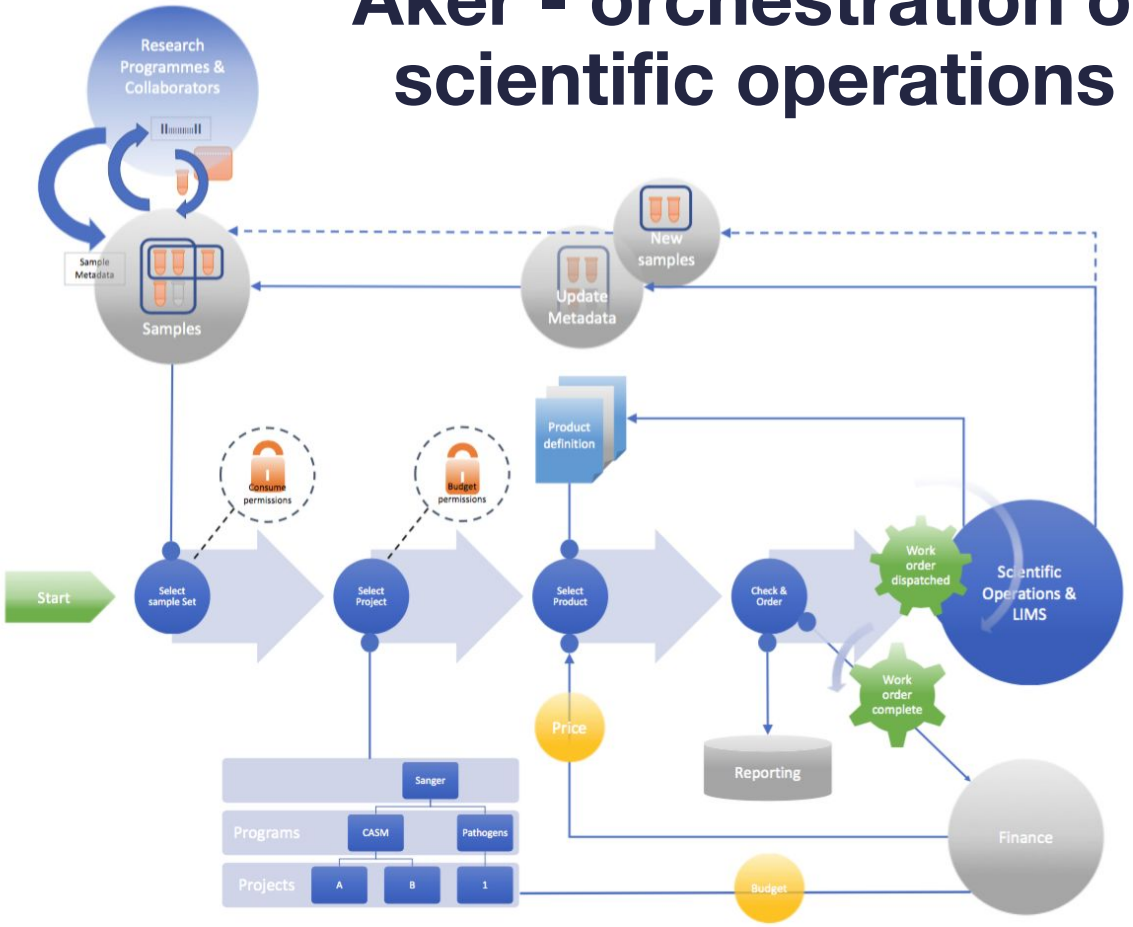
SUBSTITUTION

Signature source type

CANCER

Sample	Substitution catalog	Signature contributions
NORMAL	Sample has no substitutions.	
TUMOUR MORE		
MSK0.4_NORMAL	Sample has no substitutions.	
MSK0.4_TUMOUR MORE		

Aker - orchestration of scientific operations

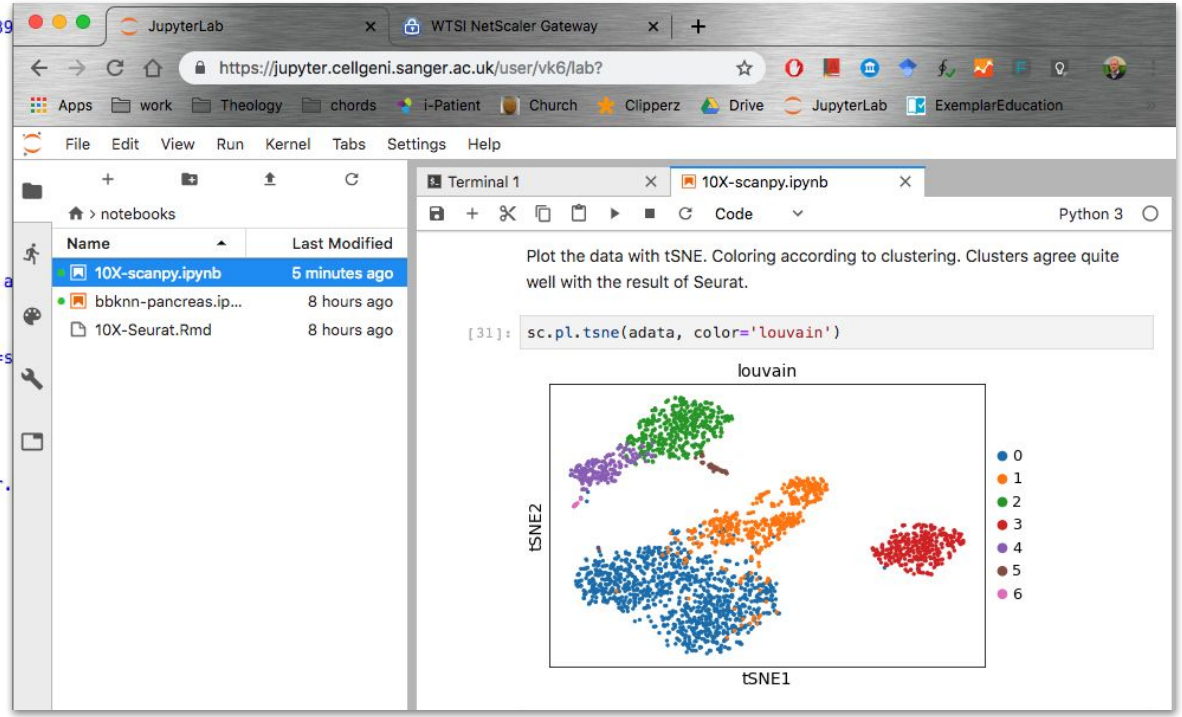


! jupyter-config.yaml ×

You, a day ago | 2 authors (Anton Khodak and others)

```yaml
proxy:
    secretToken:
    c65b057a6ba96ee5562f4ea7d78f949d3539
auth:
  admin:
    users:
      - ak27
      - vk6
      - svd
  type: ldap
  ldap:
    server:
      address: ldap-ro.internal.sanger.a
    dn:
      templates:
        - 'uid={username},ou=people,dc=s
ingress:
    enabled: false
    hosts:
     - jupyter.cellgeni.sanger.ac.uk
     - jupyter.cellgeni.internal.sanger.

singleuser:
  defaultUrl: "/lab"
  memory:
    limit: 20G
    guarantee: 16G
  cpu:
    limit: 4
    guarantee: 2
  image:
    name: quay.io/cellgeni/cellgeni-jupyter
    tag: v0.2.8
  lifecycleHooks:
    postStart:
      exec:
        command: ["bash", "/poststart.sh"]
```



UK Biobank - Vanguard

- 50,000 whole genome sequences in 18 months
- £30M MRC funding
- 4.5PB of sequence
- started August 2018; 50% done
- completion expected in October 2019

wellcome sanger institute

Compare and contrast

First draft of the human genome (Nature, 15 February 2001)

- took 10 years
- cost $2.7billion

HPCwire, 12 November 2018

Best Use of HPC in the Cloud

Readers' Choice: The Wellcome Sanger Institute using a private **OpenStack** cloud enhances IT environment necessary to sequence and assemble 100 complete human genomes per day.

wellcome sanger institute

Why use OpenStack?

In our developers' words:

"We pursued the VMware route [but found that] the level of access we would have on the VMware machines is too burdensome"

"Quick turnaround"

"Scaling"

wellcome sanger institute

Education and training

- Hashicorp products
- Ceph (for sysadmins)
- bespoke end-user training (OCF)

OpenStack - OpenStack - SSG Confluence - Mozilla Firefox

File Edit View History Bookmarks Tools Help

OpenStack - OpenStack - SSG

https://ssg-confluence.internal.sanger.ac.uk/display/OPENSTACK

Search

Confluence Spaces People Calendars Create

OpenStack

PAGE TREE

- Adding the Sanger CA
- Building images with op
- Cloudforms
- Commonly encountere
- Creating DNS entries u
- Customising your instai
- Distributed applications
- Docker training - 2016/
- Flavours
- Getting started
- Guidance for running D
- How do I get started wi
- Instances affected by o
- iRODS 4.1.10 images -
- Loadbalancer as a serv

Pages

Edit Save for later Watching Share

OpenStack

Created by Luke Burling, last modified by Peter Clapham on 24 Sep, 2018

This is a collection of end-user documentation for the Sanger OpenStack systems, a.k.a. the Flexible Compute Environment.

New to OpenStack? Start here...

Mile-high overview: How do I get started with Flexible Compute

If you want to use the web interface: read Welcome to OpenStack

For information on the command-line: Using the Openstack CLI

Requesting a new project group or an account on the flexible compute platform

Flexible Compute Capacity Management

OpenStack object store (S3-alike) quickstart

Work-In-Progress: Internal Cloud Best Practice Advice

How can I tell if should I be running my application or software on the FCE ?

But what tools are available to me ?

How can I request a new feature ?

Having trouble?

Try this: Commonly encountered errors

More advanced topics

Tracking your resource use

Customising your instances

Sending email from OpenStack instances

Using affinity and anti-affinity groups

Loadbalancer as a service

Distributed applications: links and resources

Guidance for running Docker within flexible compute environments

The Sanger community

- user engagement
- Slack
- coffee mornings

The wider community

- OpenStack Summits
- OpenStack Days
- Scientific SIG
- openstack-discuss mailing list

The future

Upgrades

- planning to follow Red Hat LTS releases: Train/RHOSP16
- serious consideration of in-place “fast forward” upgrade
- sideways-upgrade is painful for users

wellcome sanger institute

Desirable features

Driven by user demand

- Manila - filesystem-as-a-service (on CephFS?)
- Sahara - Spark
- Ironic - bare-metal instances
- Murano - application catalogue
- Reaper - pre-emptible instances
- GPUs?

wellcome sanger institute

Containers

- Docker, Singularity, Kubernetes already in use
  - users need lots of sysadmin help
- high demand for managed service
- currently planning an evaluation

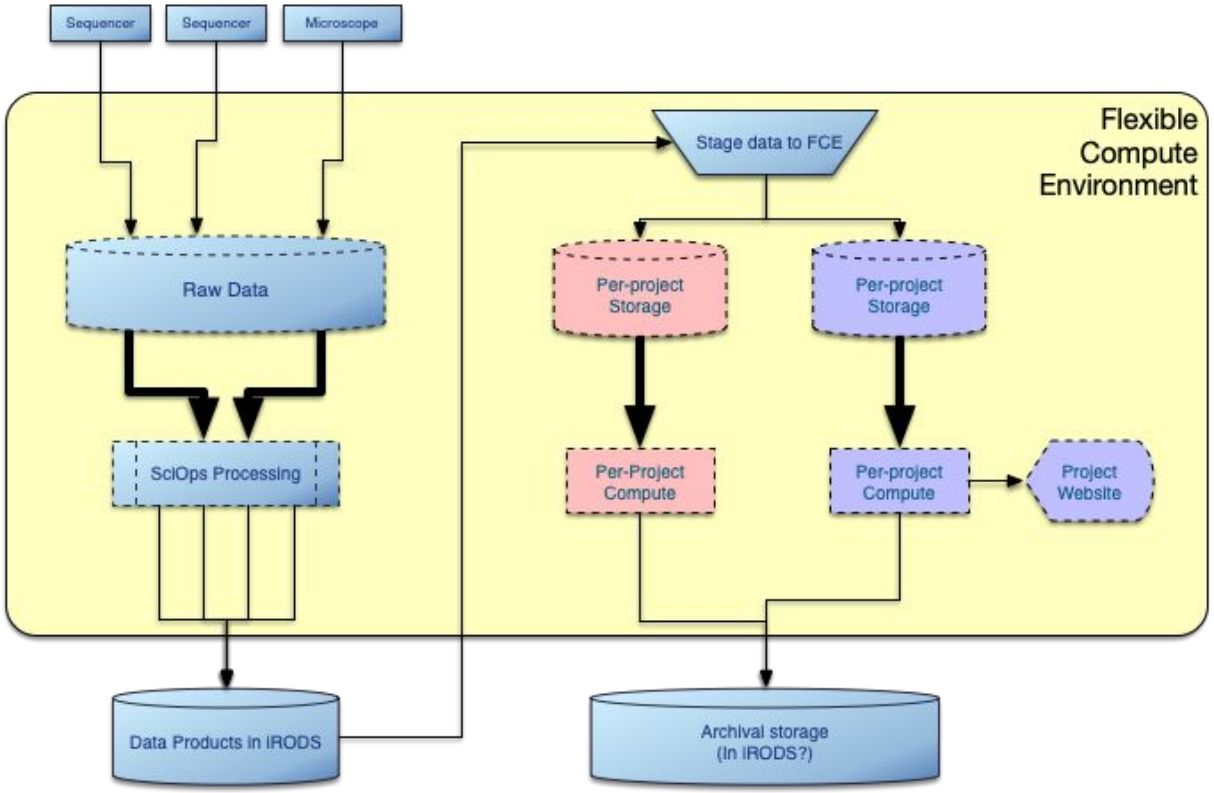
wellcome sanger institute

Offsite DR/BCP

- small OpenStack and Ceph clusters at JSDC
- currently investigating global load-balancing options
- open questions about user federation/identity management
- need to investigate Ceph replication

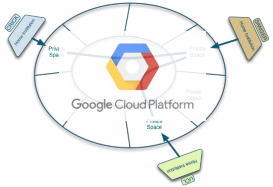
wellcome sanger institute

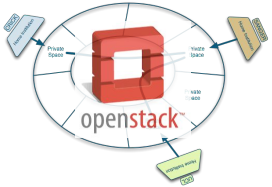
Data flow model



Evolution and federation

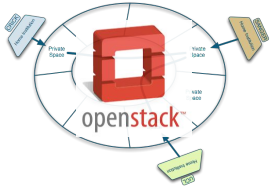
Public clouds



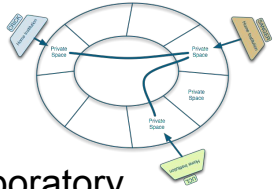


EBI
Embassy
Cloud

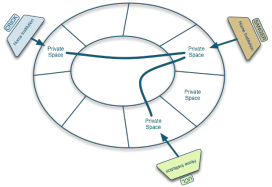
Private clouds



Sanger
Flexible
Compute



OICR
Collaboratory



GDC Cloud
(Genomic Data Commons)

wellcome
sanger
institute

Standardised pipelines

2013

2014

2015

2016

2017

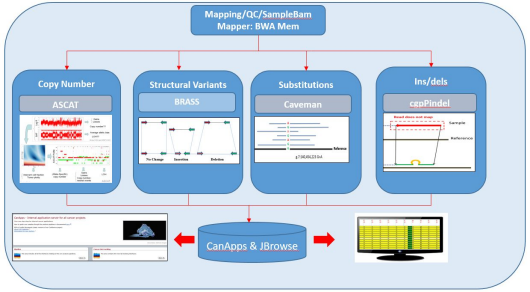
2018

ICGC/TCGA Project

- Analyse 2,800 WGS Cancers
- 3 Analysis Pipelines
- (Broad, Sanger & DKFZ)
- Need to use multiple Data Centres

Problem: analysis pipelines NOT portable

Software Go To Data:
Analysis Pipeline available in multiple Data Centres



Science Advancements

Large Collaborative Projects

IT Advancements

PAN PROSTATE CANCER GROUP

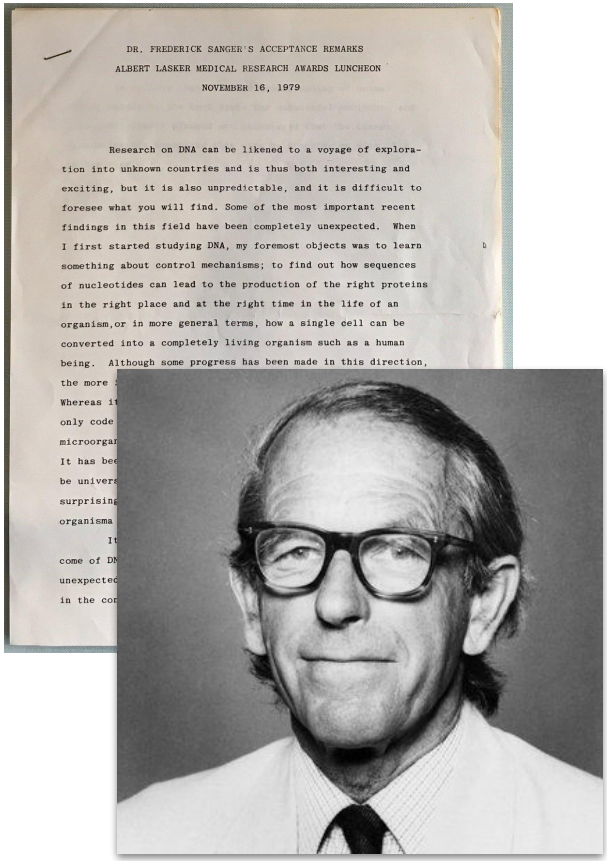
redhat

openstack

happy scientists

Future science

Conclusion



DR. FREDERICK SANGER'S ACCEPTANCE REMARKS
ALBERT LASKER MEDICAL RESEARCH AWARDS LUNCHEON
NOVEMBER 16, 1979

Research on DNA can be likened to a voyage of exploration into unknown countries and is thus both interesting and exciting, but it is also unpredictable, and it is difficult to foresee what you will find. Some of the most important recent findings in this field have been completely unexpected. When I first started studying DNA, my foremost objects was to learn something about control mechanisms; to find out how sequences of nucleotides can lead to the production of the right proteins in the right place and at the right time in the life of an organism,or in more general terms, how a single cell can be converted into a completely living organism such as a human being. Although some progress has been made in this direction,

"It is difficult to predict the further outcome of DNA research, but we may reasonably hope for further unexpected and exciting discoveries - some of which will help in the conquest of diseases." - Fred Sanger, Lasker Award acceptance in 1979

wellcome sanger institute

Dave Holland
dh3@sanger.ac.uk

https://hpc-news.sanger.ac.uk/

https://sangerinstitute.blog/

wellcome sanger institute