Wisdom is not a crop harvested by age.

The Cambridge Research Computing Service

Its all about the data – well High Performance Data Analytics

Sean McGuire

IDC Oxford Sept 2016

From EDSAC to Openstack

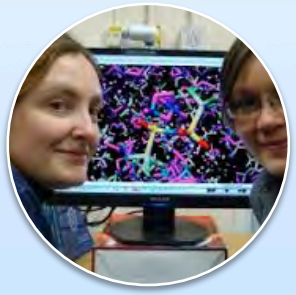
Cambridge has had an in house research computing development and service for the last 77 years

EDSAC 1949

Darwin 1996

Bio-Medical Cloud 2016

Focus area's



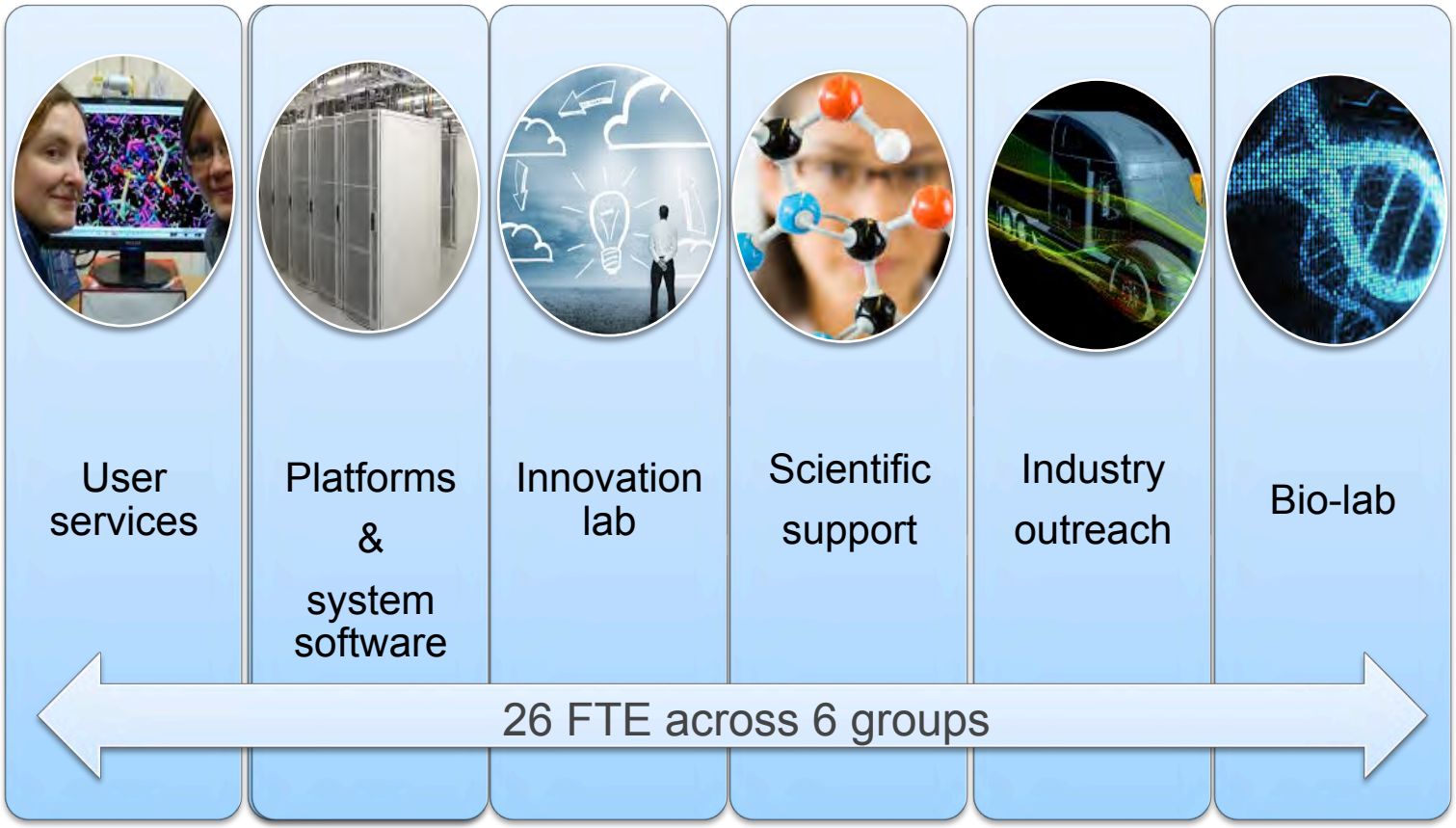
Research support

Industrial outreach

HPC & data solution development

Driving discovery, impact & innovation

Service structure



Cambridge as an HPC market

- University annual turnover is £1.2B
- Annual R&D budget £370M
- One of the world top research intensive Universities
- Cambridge is also a major technology centre
  - 1535 technology companies in surrounding science parks
  - £12B annual revenue
  - 53000 staff
- We have a mandate to provide HPC services to the University and the wider Cambridge technology cluster

Research outputs & growth

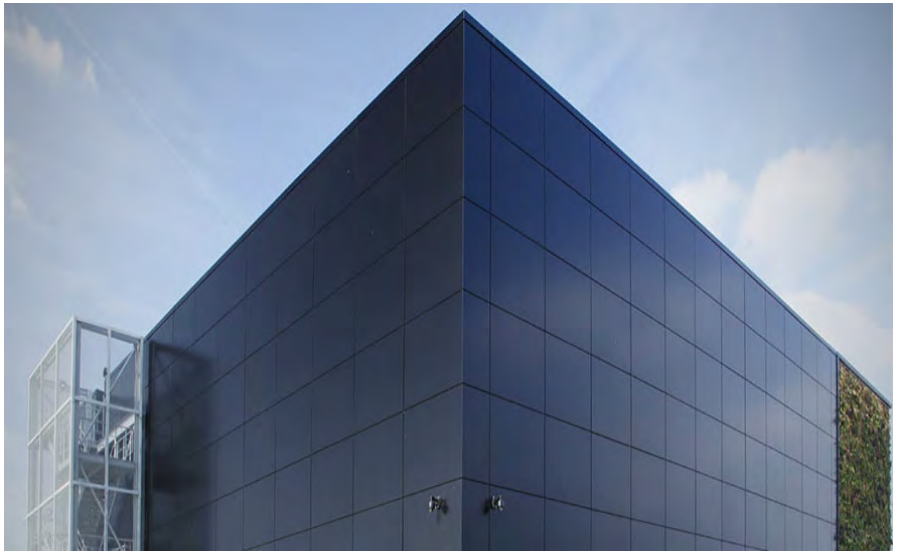
- **665** active accounts across **197** research groups from 42 Departments
- **500** publications a year

- Supporting a Cambridge grant portfolio of **£120M**
- **10%** of the grant income in Physical Sciences, Clinical Medicine and Technology

- New user growth at **38%** CAGR
- System load is growing annually at **40%**
- User support tickets are growing at **40%** a year

New energy efficient data centre

- £15M purpose built, 4 halls
- High levels of resilience,
- 280 Cabinets, 3000Kw IT Load
- 24/7 DC staff
- ISO2007 security standards



- HPC
- 100 racks -2000 Kw IT load
- evaporative cooling, water cooled back door, best in class PUE 1.1

Current HPC platforms

- 900 Dell Servers - 450 TF sustained DP performance
  - 600 node (9600 core) full non blocking Mellanox FDR IB 2,6 GHz sandy bridge (185 TF) – entered at 93 in TOP500
  - 128 node 256 card NVIDIA K20 GPU cluster 250 TF full non blocking dual rail Mellanox FDR connect IB – entered at No.2 in green 500
  - Various development, test, POC, inc ARM and standalone customer systems
- 6 PB of high performance Lustre parallel file system
  - Various generations Dell MDxxxx hardware
  - In house design, implementation and support
  - Provides unparalleled price performance

Why OpenStack in research computing

make computing, data, applications and workflows more accessible, flexible and secure.

Allow a wide range of services to be delivered from a single framework

Make research computing easier to use, easier to share, decreasing the time to science and increasing innovation and research outcomes

OpenStack activities @ Cambridge

- Scientific-WG https://wiki.openstack.org/wiki/Scientific_working_group
- Development and deployment of Openstack in the broad research computing environment with a particular focus on bio-medical computing
- Proof of concept work for OpenStack for use within the SKA, contracted by Astrophysics under the direction of Prof Paul Alexander

- The work being undertaken by the Research Computing Service is an active collaboration with the following partners:-
  - Dell
  - Intel
  - Redhat
  - Mellanox
  - Nexenta
  - StackHPC
  - Emerging scientific OpenStack community

OpenStack use cases within research computing

1) Research computing & storage infrastructure as a service (IaaS)

- *Provides central IT infrastructure, for departmental local IT staff to construct persistent IT services. Local IT staff build and support platforms*
- *Enfranchises local IT staff, drives efficiencies, increases agility*

2) Research computing & storage platform as a service (PaaS)

- *Provides central IT platforms, for research group or departmental local IT staff to build persistent IT services. Local IT staff build and support platforms*
- *Enfranchises research staff and local IT staff drives efficiencies increases agility*

3) Application development cloud (IaaS, PaaS)

- *Bare metal or virtualised provision of dynamic resources for platform or application development and testing. Uptake by local IT staff and research staff*
- *Drives innovation, in terms of application development, cloud computing utilisation, infrastructure*

OpenStack use cases within research computing

4) Research computing as a service (PaaS, SaaS)

- *Allowing easy access to virtualised or containerised compute infrastructure, long tail science, new user communities, science portals, can include application software*
- *Easy access, collaboration, sharing methods, workflows data greatly democratising research computing driving discovery*

5) HPC as a service (SaaS)

- HPC cluster, with workload scheduler, MPI workload, parallel efficiency, either shared tenant pool or individual, PVC or not
- *multitenant security models, user defined HPC environment, run anywhere Hybrid HPC, customisable, flexible repeatable HPC*

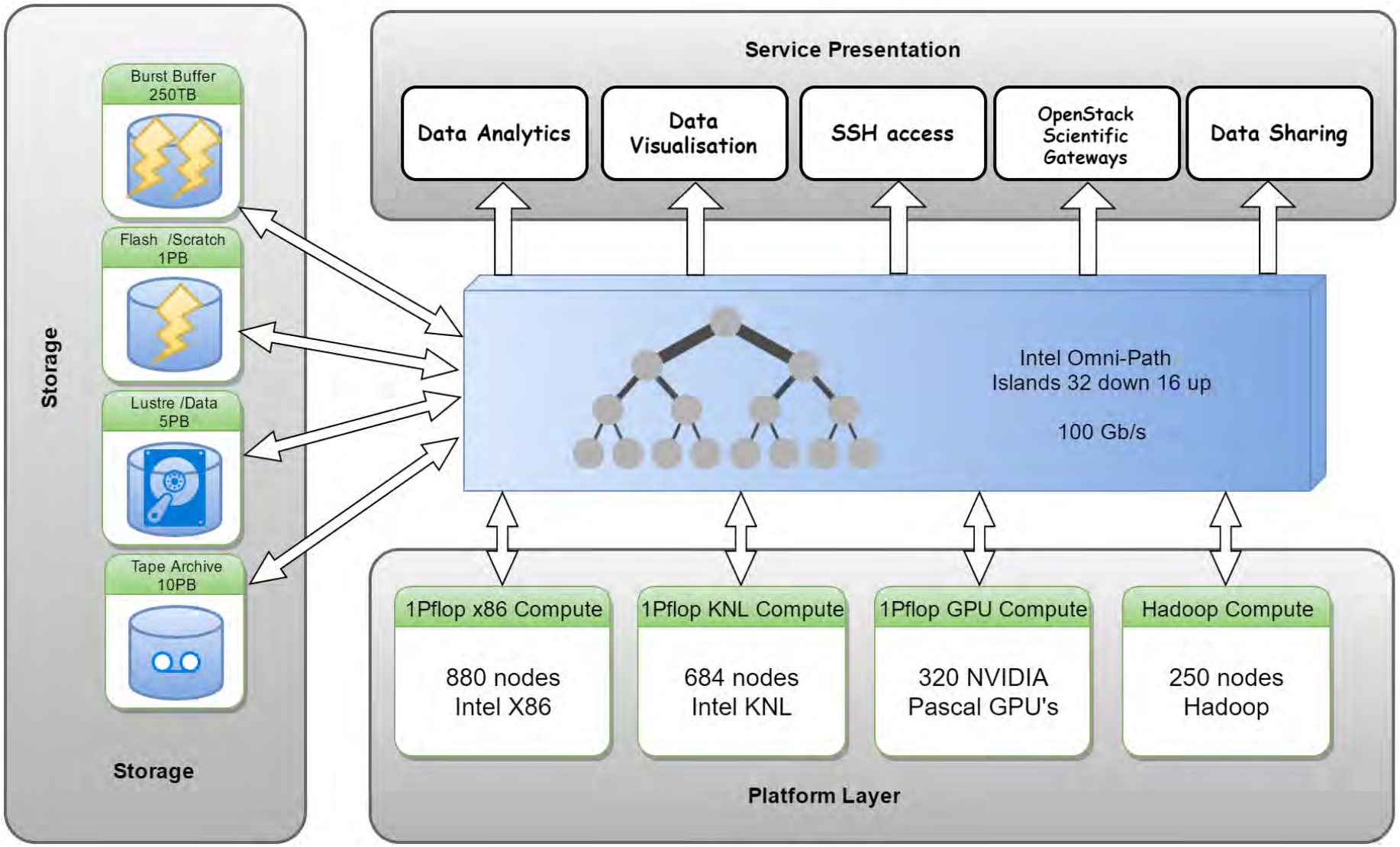
6) HPDA as a service (SaaS)

- System architecture optimised for large storage and Hadoop, allow virtual I/O to integrate large data sets, virtualised Hadoop, data federation
- Bring users to the data, can use own environment, portable, flexible. Hadoop is customisable, grow on demand, tenant separation.

Peta – A National HPC system for Petascale data intensive simulation and analytics

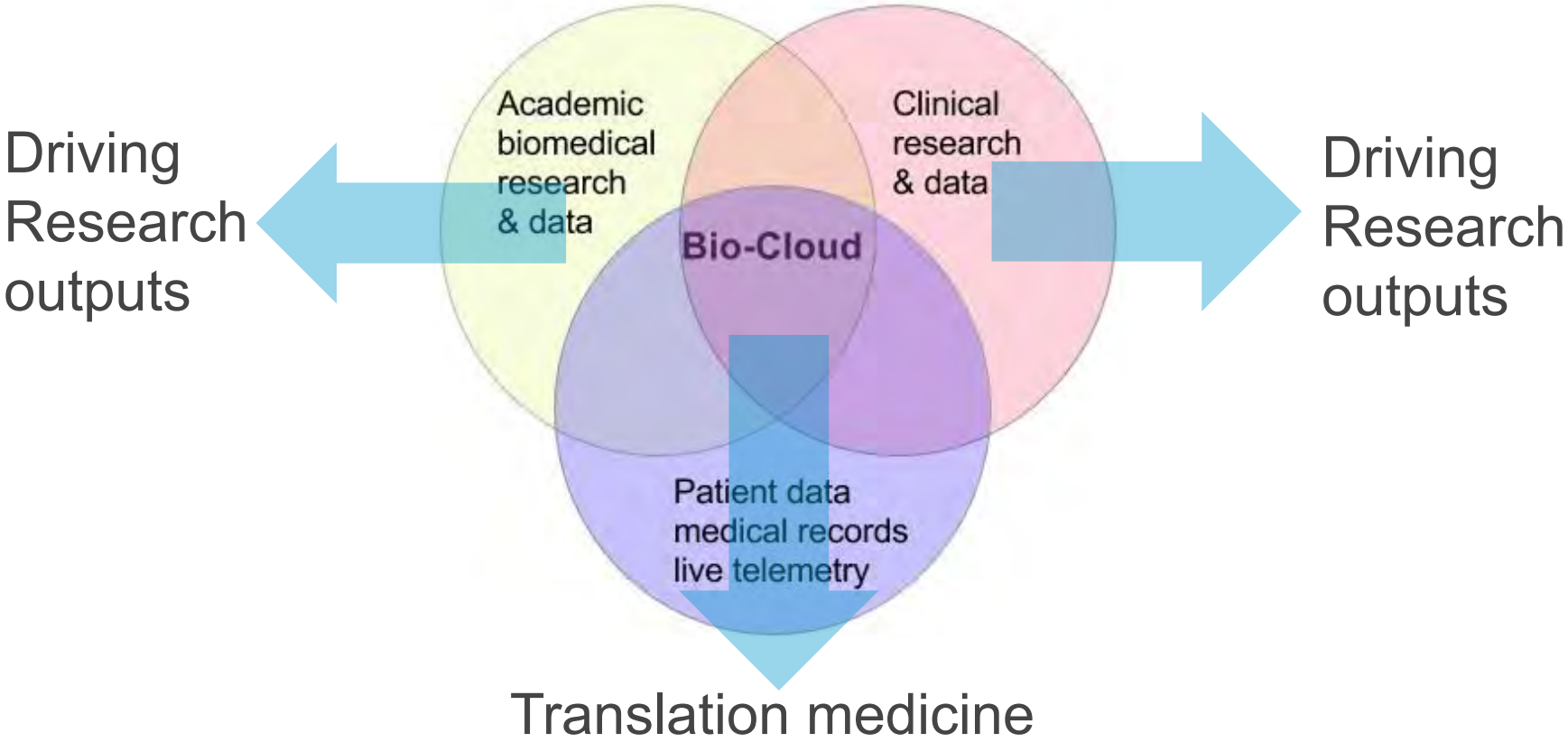
- Large £9M system investment, very cost effective due to
  - active supply chain management;
  - in house system design, system integration and solution support
- Probably, the largest UK academic HPC resource when deployed, mid 2017
- Heterogeneous architecture with single central file system across all elements
- Particular architectural focus on high throughput I/O and data analytics at scale
- Only system of its kind in the UK
- Procurement completed end 2016
- Service launch June 2017
- System software Openstack

and now a few examples…

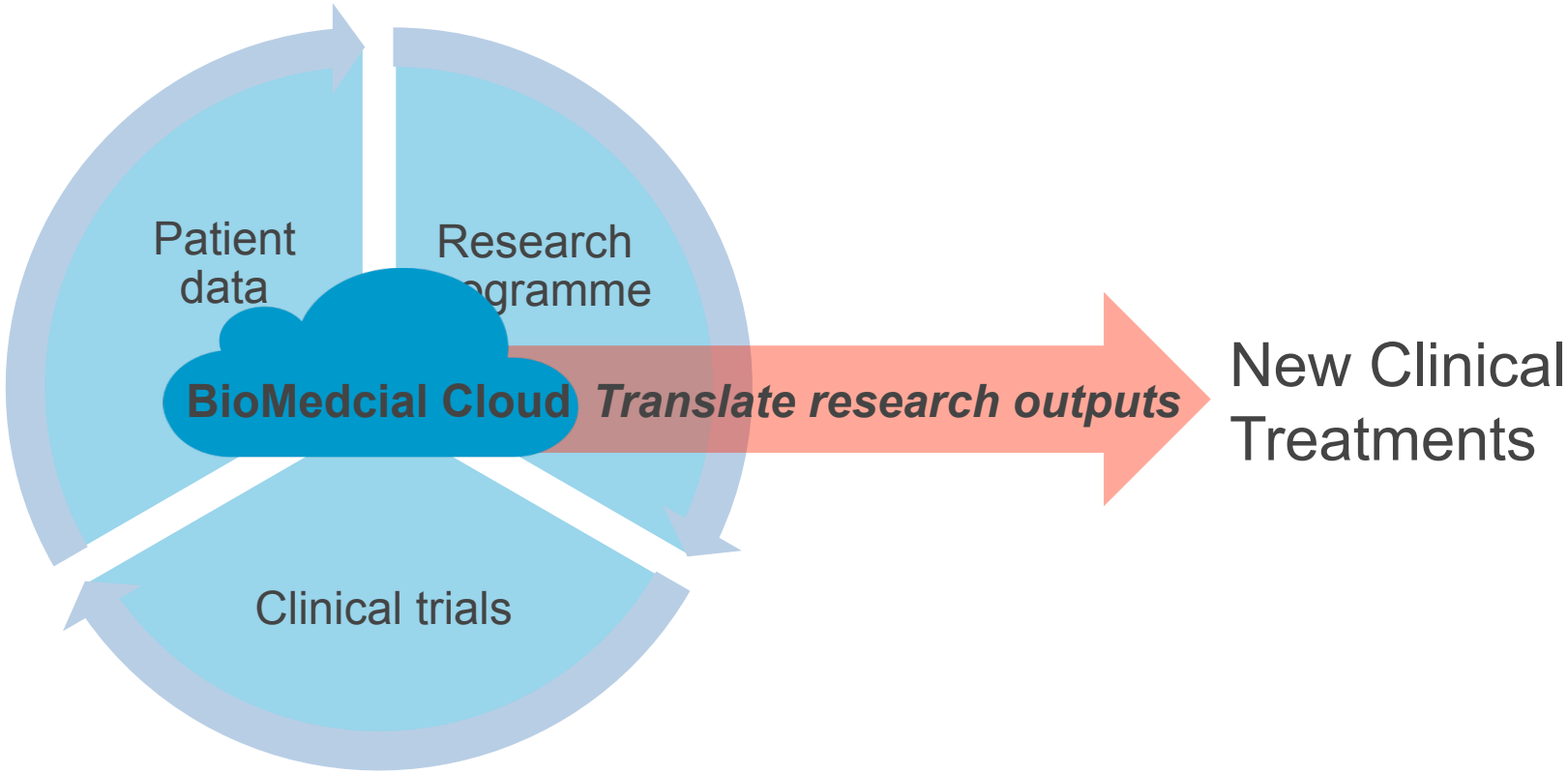

Storage
Burst Buffer
250TB
Flash /Scratch
1PB
Lustre /Data
5PB
Tape Archive
10PB
Storage
Service Presentation
Data Analytics
Data Visualisation
SSH access
OpenStack Scientific Gateways
Data Sharing
Intel Omni-Path
Islands 32 down 16 up
100 Gb/s
1Pflop x86 Compute
880 nodes
Intel X86
1Pflop KNL Compute
684 nodes
Intel KNL
1Pflop GPU Compute
320 NVIDIA
Pascal GPU's
Hadoop Compute
250 nodes
Hadoop
Platform Layer

Bio-Medical-Cloud

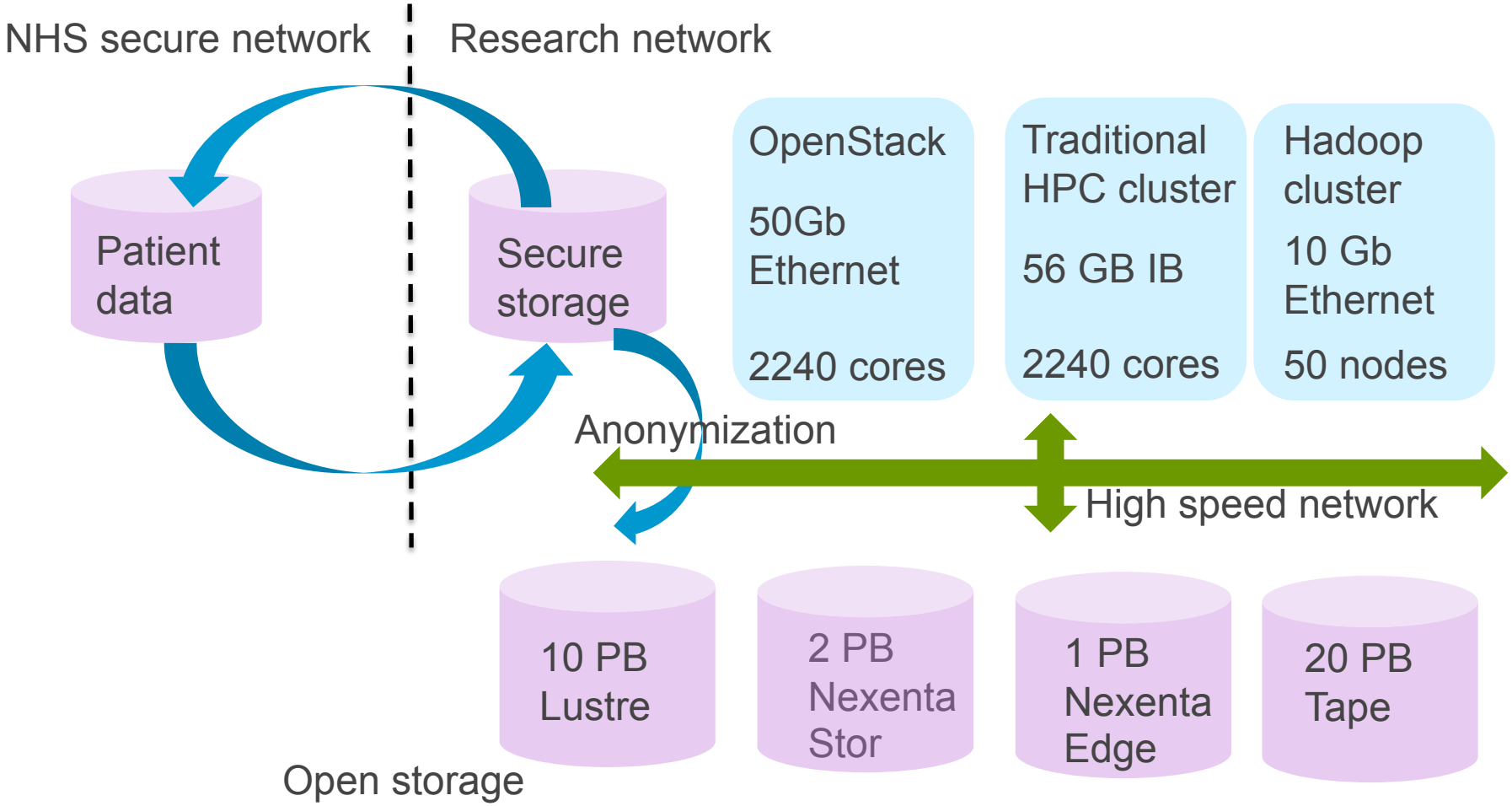
- One central Bio-Medical compute & data capability for both biological and medical research at the University and Hospital linking research staff, research data and patient data



Bio-Medical-Cloud & translational medicine

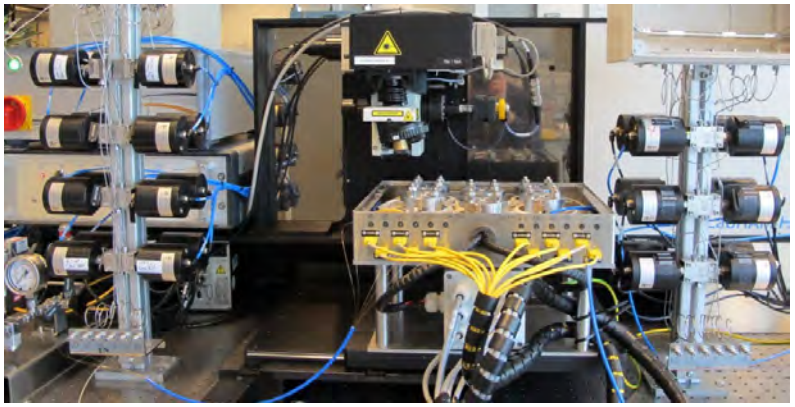


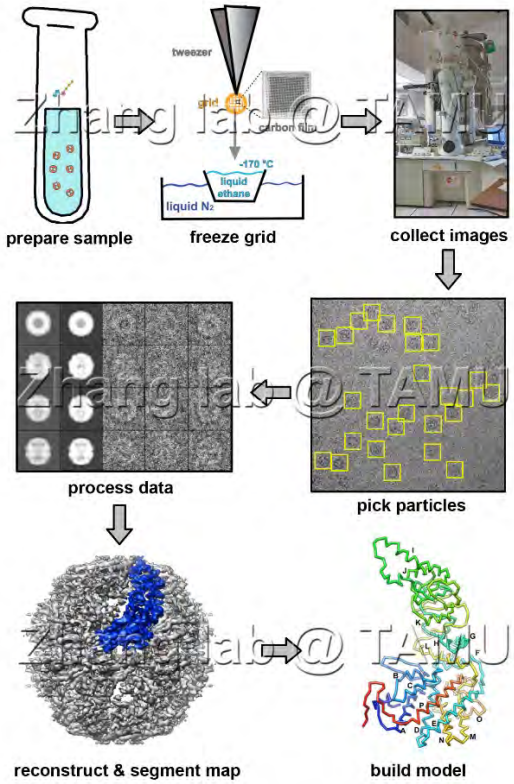
Bio-Medical-Cloud high level view



Medical imaging, microscopy and Cryo-EM

A revolution in biology and medicine from new instruments – but produce huge amounts of data and require large amounts of data processing





Predictive medical Informatics

- The Bio-Medical Cloud will underpin a new collaboration between the Addenbrooke's Hospital and the University to develop new predictive analytics systems for improving patient care

- Take electronic health records, live patient telemetry data, statistical model, run that through real time and produce patient interventions that improve health

Dr John Cromwell from University Hospital developed a new statistical model that takes patient medical records, live feeds from operating room runs areal time model in Statistica

Cuts post operative infection rates by 58%

Large scale NGS sequencing & analytics

- We are jointly with Genomics England leading the developing of major open source gene analysis software stack called OpenCB, this software offers break through functionality and performance in large scale gene variant analysis.
- We will deploy the OpenCB on the Bio-Medical-cloud to drive genomics research at Cambridge, one such study is the Bridge study to sequence and analyse the genomes of 10,000 rare disease patients

SKA IT design Cambridge led (Astrophysics)

- Design work led by Prof Paul Alexander in Astrophysics

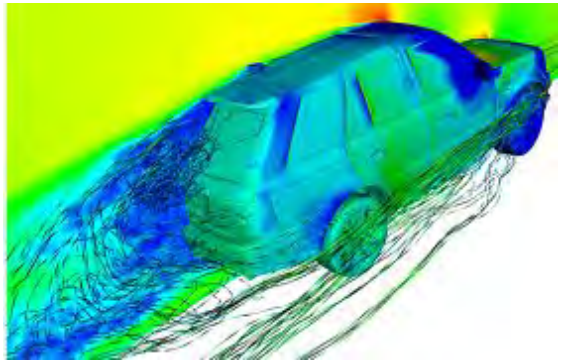
- RCS is contracted to help with
  - HPC compute design
  - HPC storage design
  - HPC operations

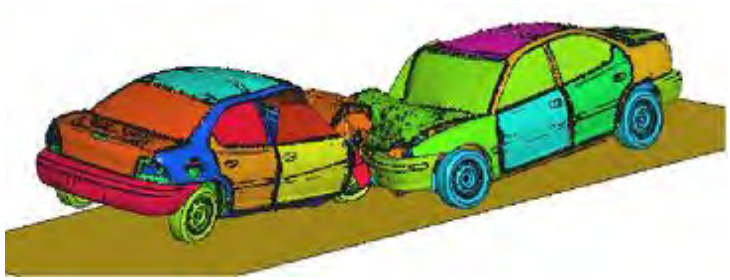
CHPC SA

- Long term strategic partnership
- Technology strategy, system design procurement, integration training and education
- Worked with CHPC on the last three national systems
- Strong pan African HPC training strand
- Strong SKA link-up

Jaguar Land Rover R&D project

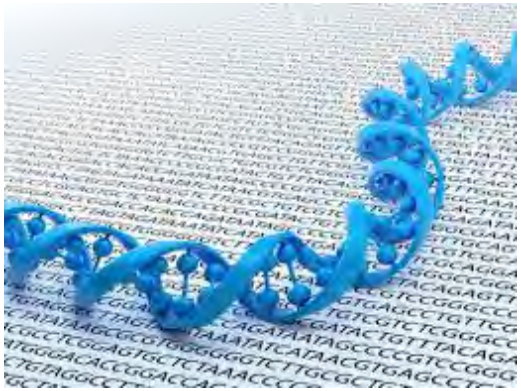
- 5 year research project with JLR
- Drive capability in simulation & data mining
- HPC design, implementation and operation best practice





UK 10K genome project

- RCS is providing all the data storage and computational recourse for this major UK genomics study
- The study involves the gene sequencing of 10K patients from the UK
- Data throughput problem requiring good reproducible I/O and porting of sequencing pipelines to shared cluster environment



Changes in our last 2 years

- Measure what we need to
  - Use those metrics to communicate our value
- Communication
  - User group
  - Governance
  - …increase our methods of listening
- Hiring Criteria
  - Skill sets
  - Where we advertise, what we advertise
  - From 5 to 33 heads

- …to create our Vision, Mission and Strategy

Trends

- Growth
  - HPDA
  - Data
  - Variety of silicon options – which are needed
  - New users from all 4 corners of…
- Our Research Software Engineers team++
- Simplify all we can for our users

Future focus & goals

- Life science – both next gen Omics analysis software development and translation medicine within Hospital environment
- Industrial outreach and industrial enablement in terms of HPC and data analytics usage = CORE Advantage
- National level HPC provision – on target to be the largest National Academic HPC facility in 2017
- Help drive convergence and Openstack and HPC – research computing clouds
- ***To continue to develop world leading research computing solutions that drive major scientific discovery, dramatically improve patient health outcomes and provide a positive impact on the UK Knowledge economy.***