Big data and HPC on-demand: Large-scale genome analysis on Helix Nebula – the Science Cloud

Rupert Lueck
Head of IT Services, EMBL Heidelberg

Helix Nebula General Assembly
ESA/ESRIN Frascati, 16 January 2013

EMBL

HELIX NEBULA
THESCIENCECLOUD

EMBL: European Molecular Biology Laboratory

- Intergovernmental Research Organization
- Supported by 20 Member States (+1 associated:)
- One of the world‘s foremost life science institutions
- EIROforum member
- 1500 staff
  >70 nationalities

The Five Branches of EMBL

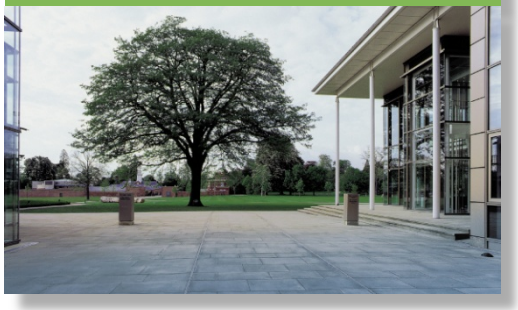
Heidelberg

Basic Molecular Biology Research
Main Lab / Headquarters

Hamburg

Structural Biology
DESY

Hinxton



European Bioinformatics Institute (EBI)
Sanger Centre

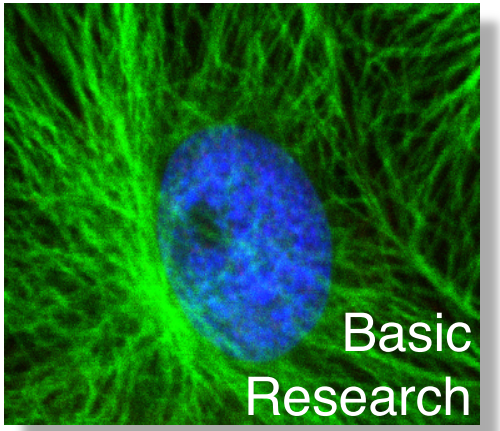
Grenoble

Structural Biology
ILL, ESRF, IBS, UVHCI

Monterotondo

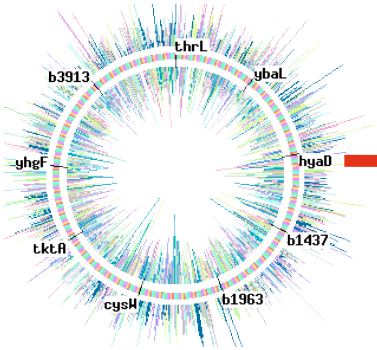
Mousebiology
CNR, EMMA

EMBL's Missions

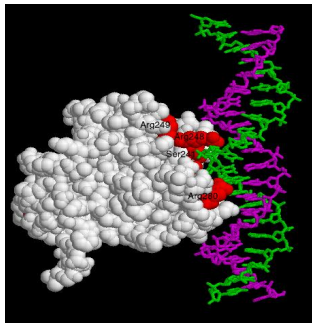


EMBL

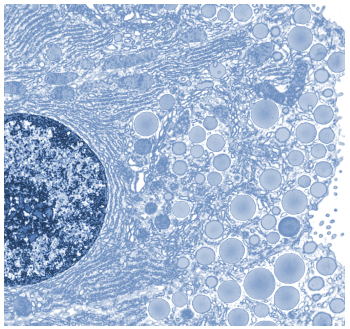
Systems Biology: From Molecules to Organisms



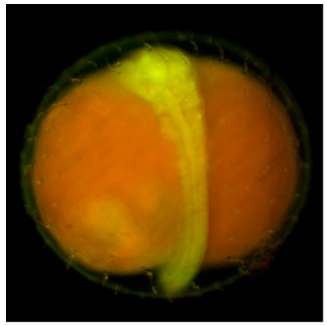
Genome



Protein/DNA

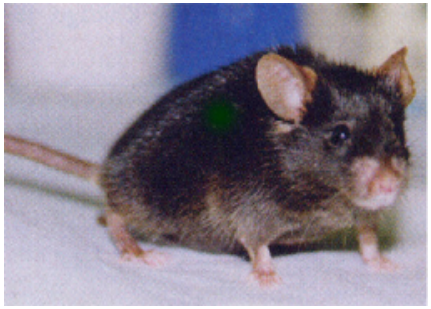


Cell



Embryo

Development



Organisms

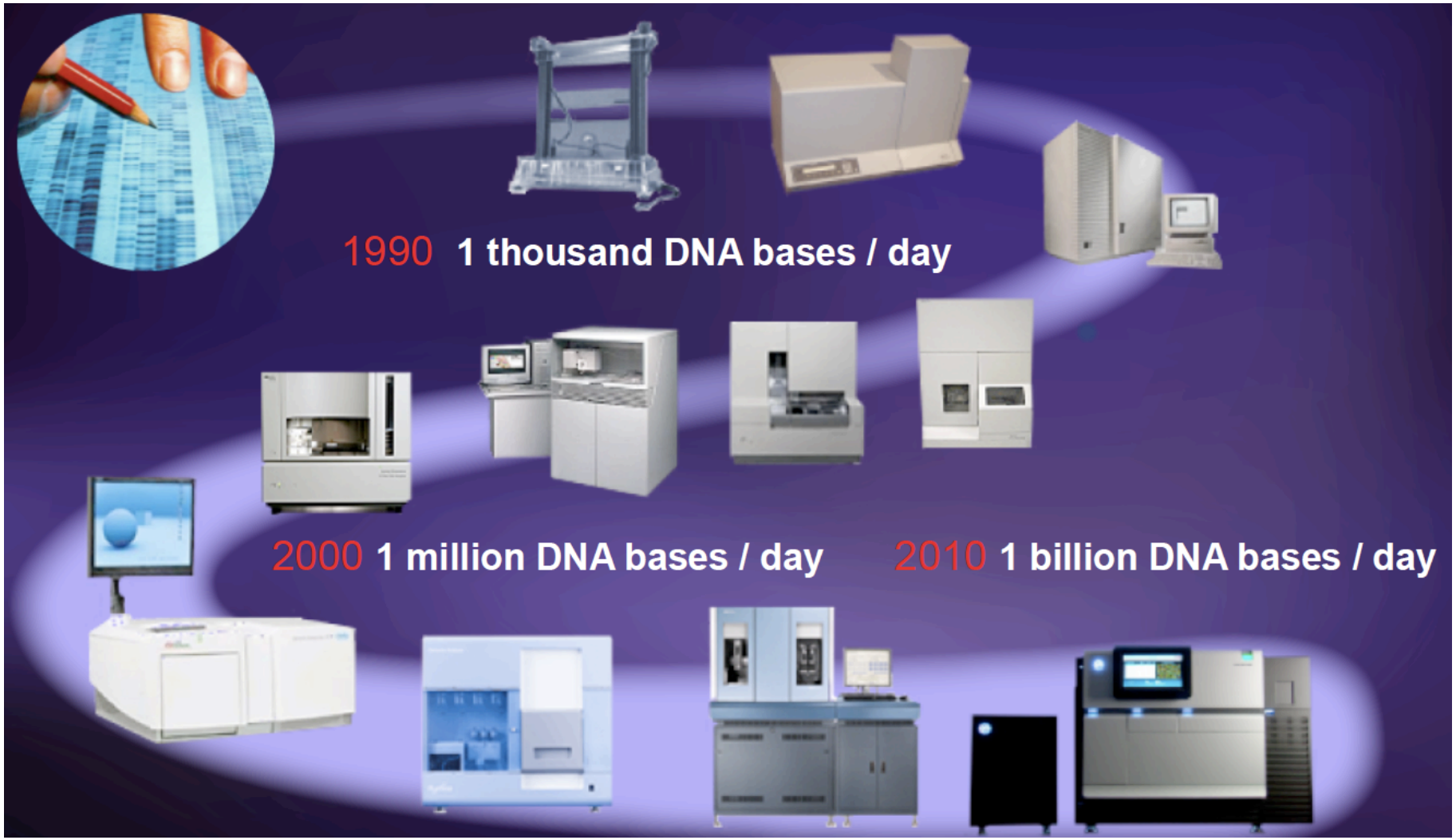
Complexity

Aging

Disease

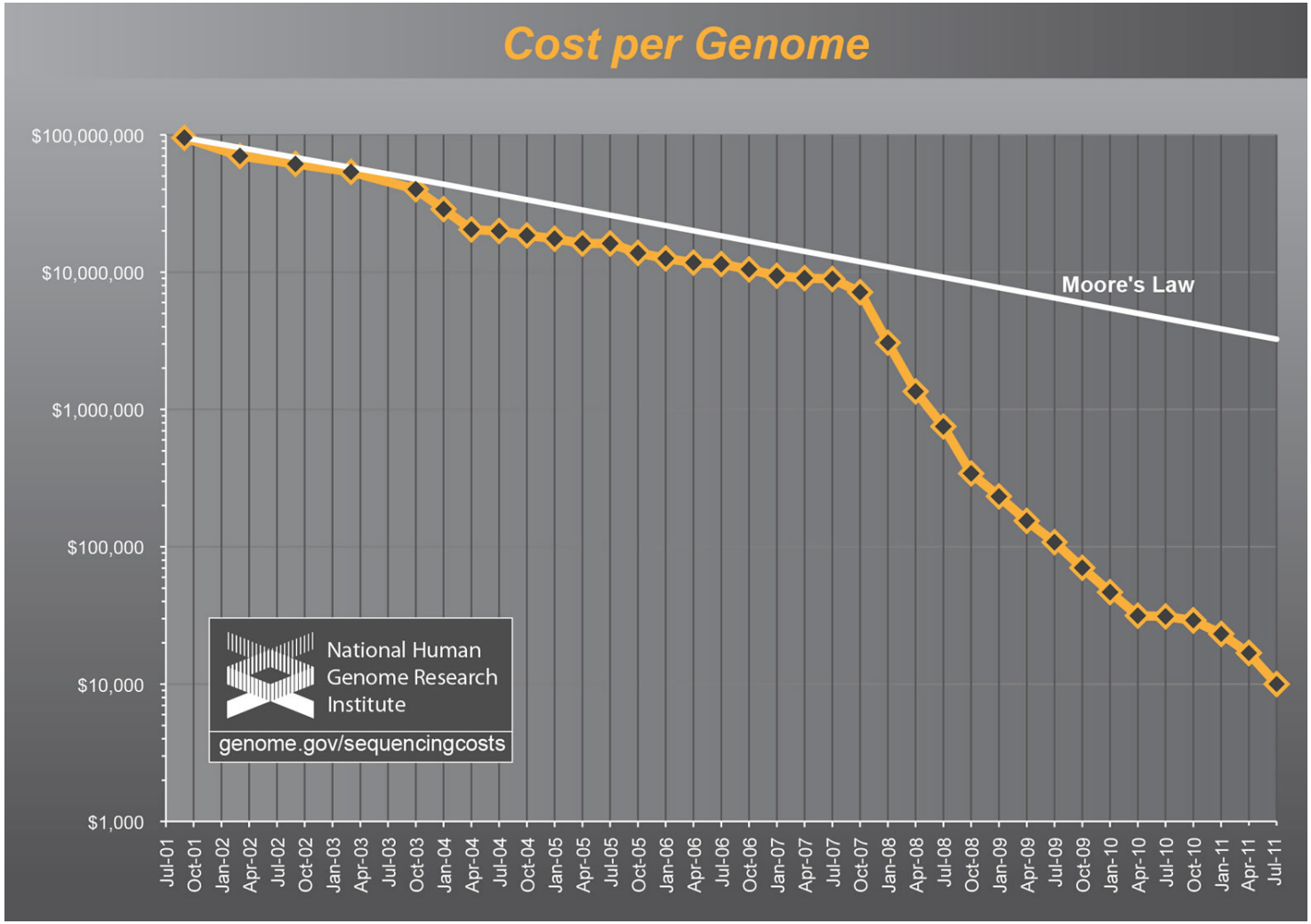
EMBL

Next Generation Sequencing (NGS) Revolution



EMBL

Cost of Sequencing Decreasing Rapidly

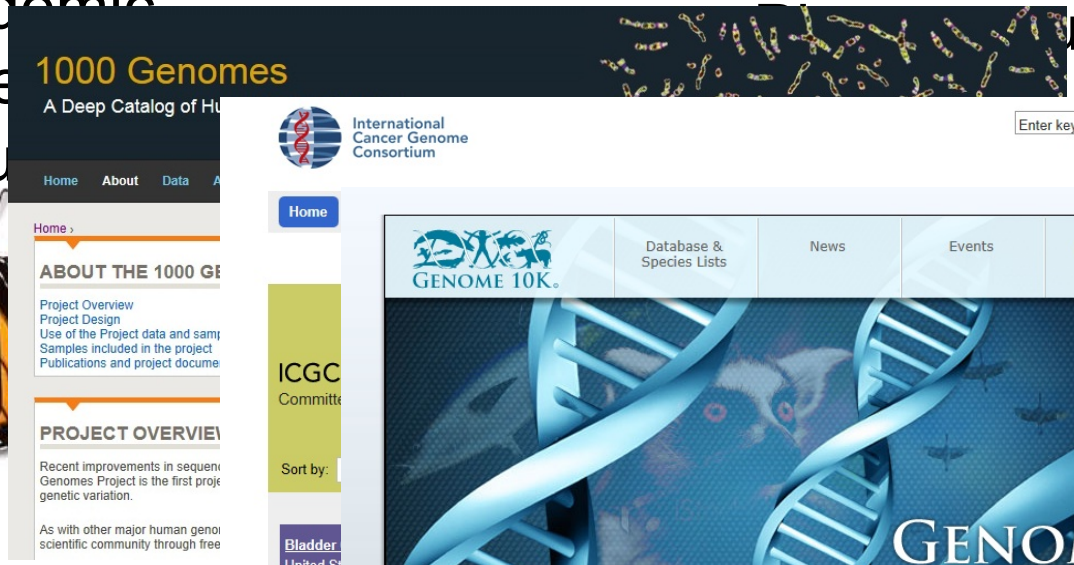


EMBL

Genomic Sequencing is Now an Affordable Solution

Academic
Rese
Grou

utical



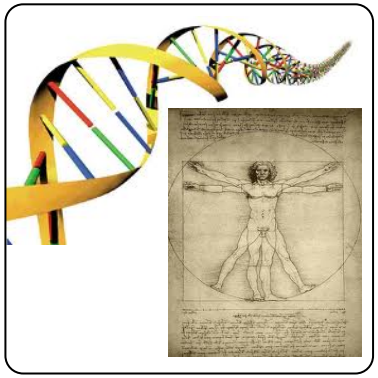
EMBL

Genomic sequencing is
now an affordable solution

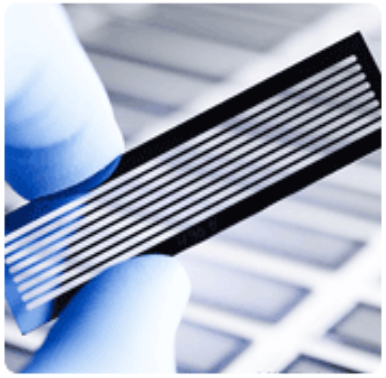
however ...

Read the Sequence to Study the Organism

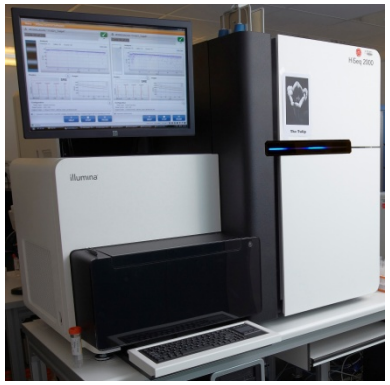
Extract DNA



Prepare



Sequence

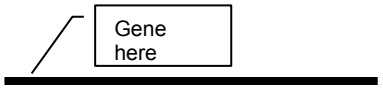


Lab

In Silico

Assemble

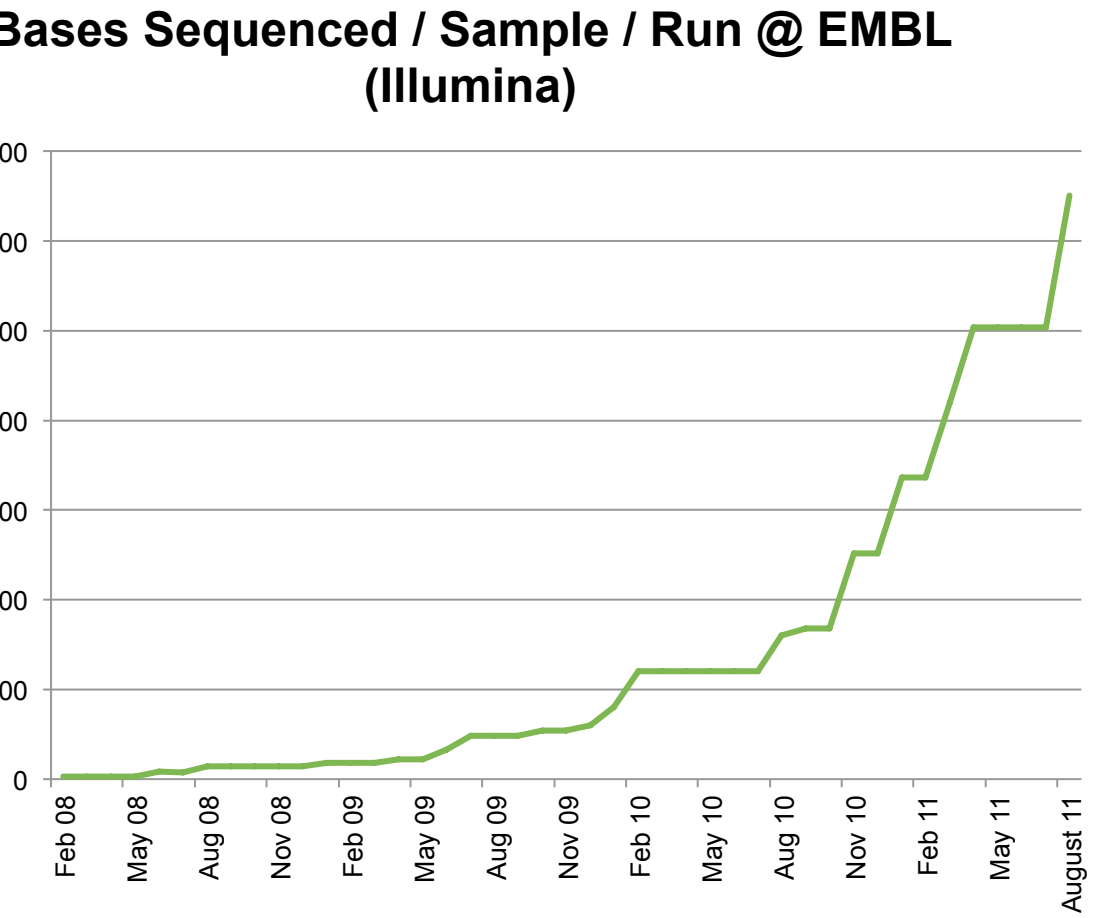
Annotate

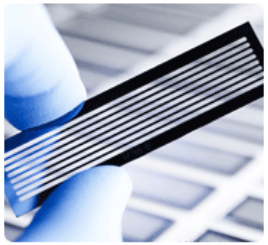


Requires Computing Infrastructure & Expertise

EMBL

Problem - Technology Explosion with NGS



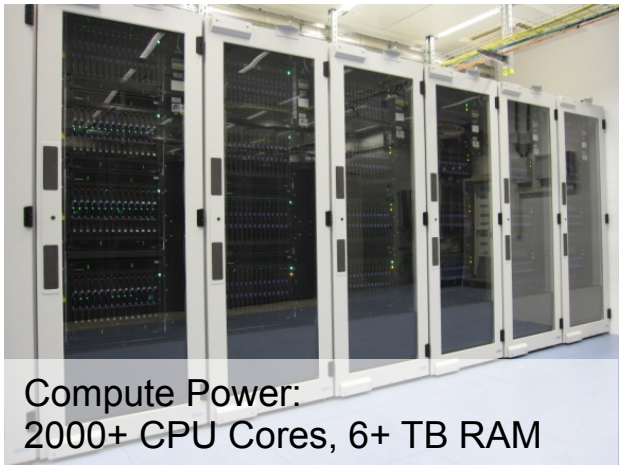


EMBL

Sequence Production & IT Infrastructure at EMBL

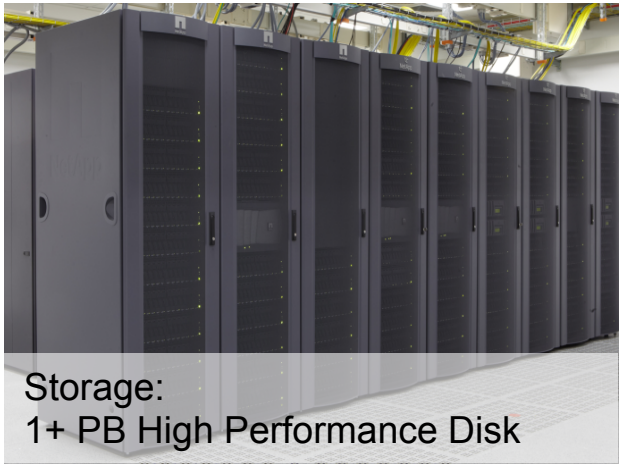
4 x Ilumina HiSeq2000

25+ TB data each week



1 x MySeq

1 x Ion Torrent

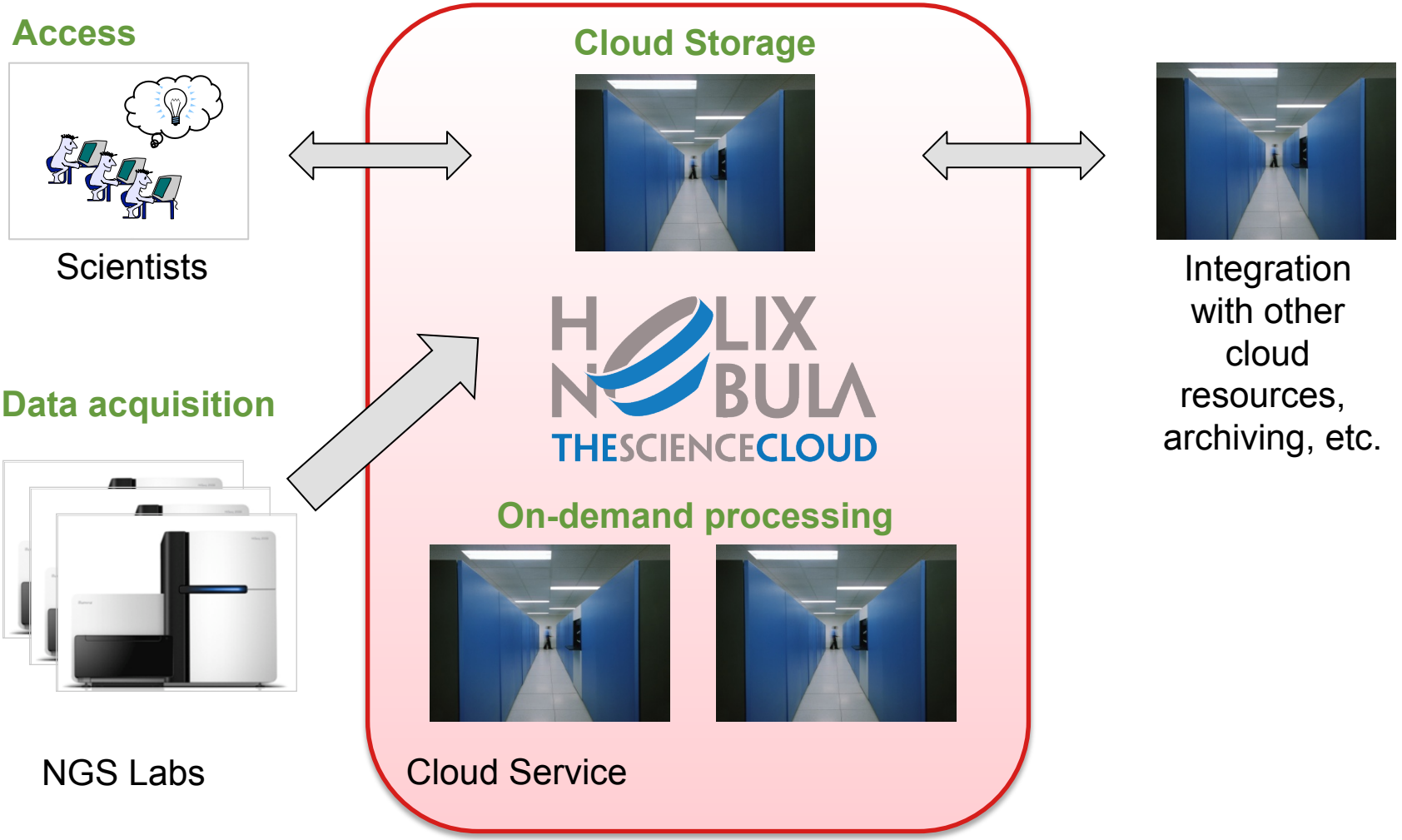


EMBL

NGS - The Big Picture

- ~ 8.7 million species in the world (estimate)
- ~ 7 billion people
- Sequencers exist in both large centres & small research groups
- 200+ Ilumina HiSeq sequencers in Europe alone
  - capacity to sequence 1600 human genomes / month
- Largest centre: Beijing Genomics Institute (BGI)
  - ~140 HiSeq
- ~1500 Hiseq devices worldwide today
  - 3-6 PB / day
  - 1.1 – 2.2 ExaBbytes / year

EMBL Flagship: Large-scale Genome Analysis



EMBL

Proof of Concept

- Multiple Cloud providers
  - ATOS / Sixsq
  - CloudSigma
  - T-Systems
- Each tested steps with increasing complexity
- Major software components to test

Assembly pipeline	SGA by Simpson, JT & Durbin, R http://genome.cshlp.org/content/22/3/549.long
Annotation pipeline	Ensemble
Shared File system	e.g. glusterFS
StarCluster	http://star.mit.edu/cluster/

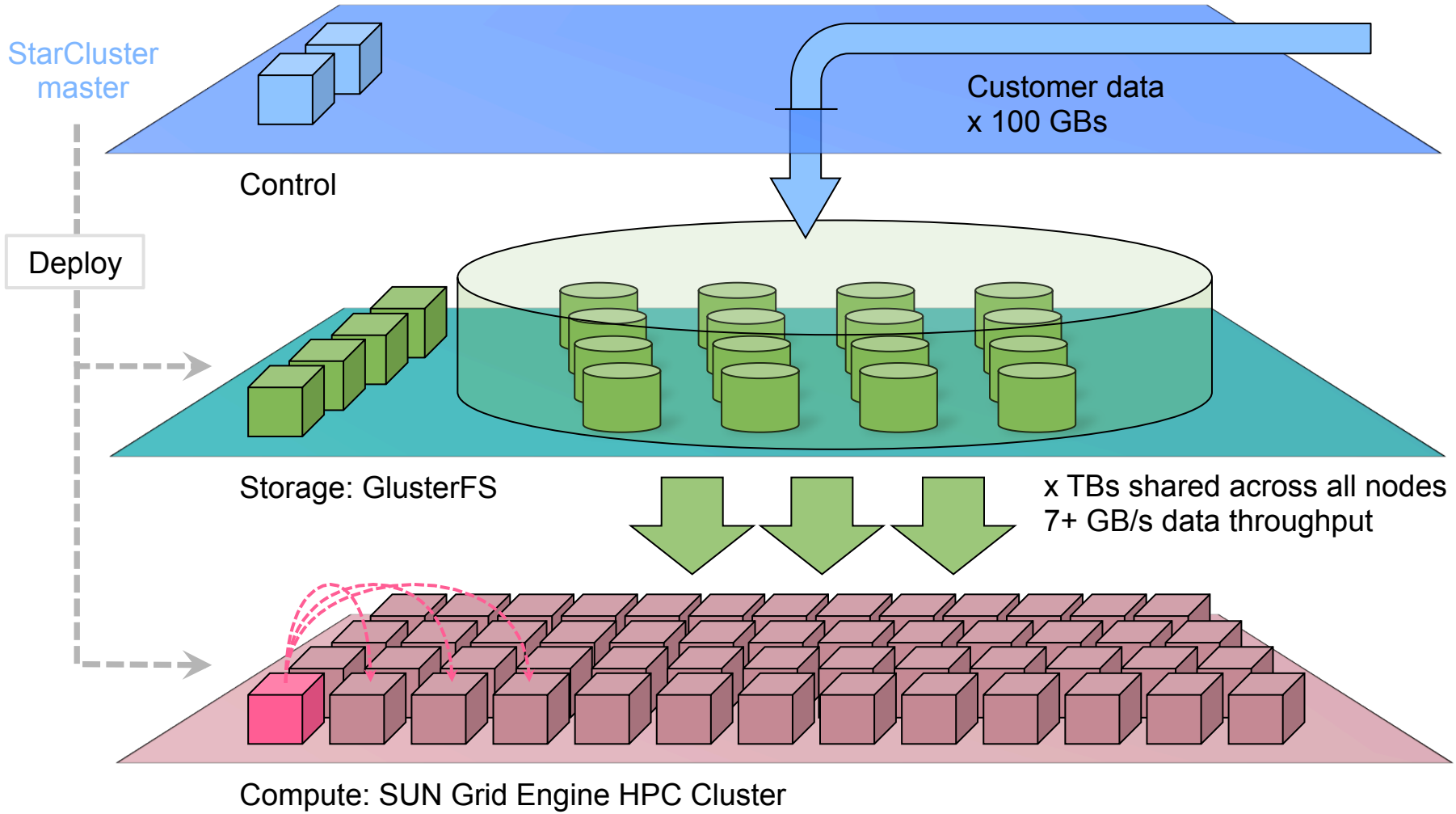
StarCluster & Sun Grid Engine
Dynamic cluster provisioning

HELIX NEBULA THESCIENCECLOUD

- StarCluster – Dealing with the Fluctuating Workload
  - Manages provisioning of images and setting up of cluster
  - Requires sets of EC2 APIs to work
  - It monitors the number of jobs in the queue and launches more instances
  - Terminates them when no longer required

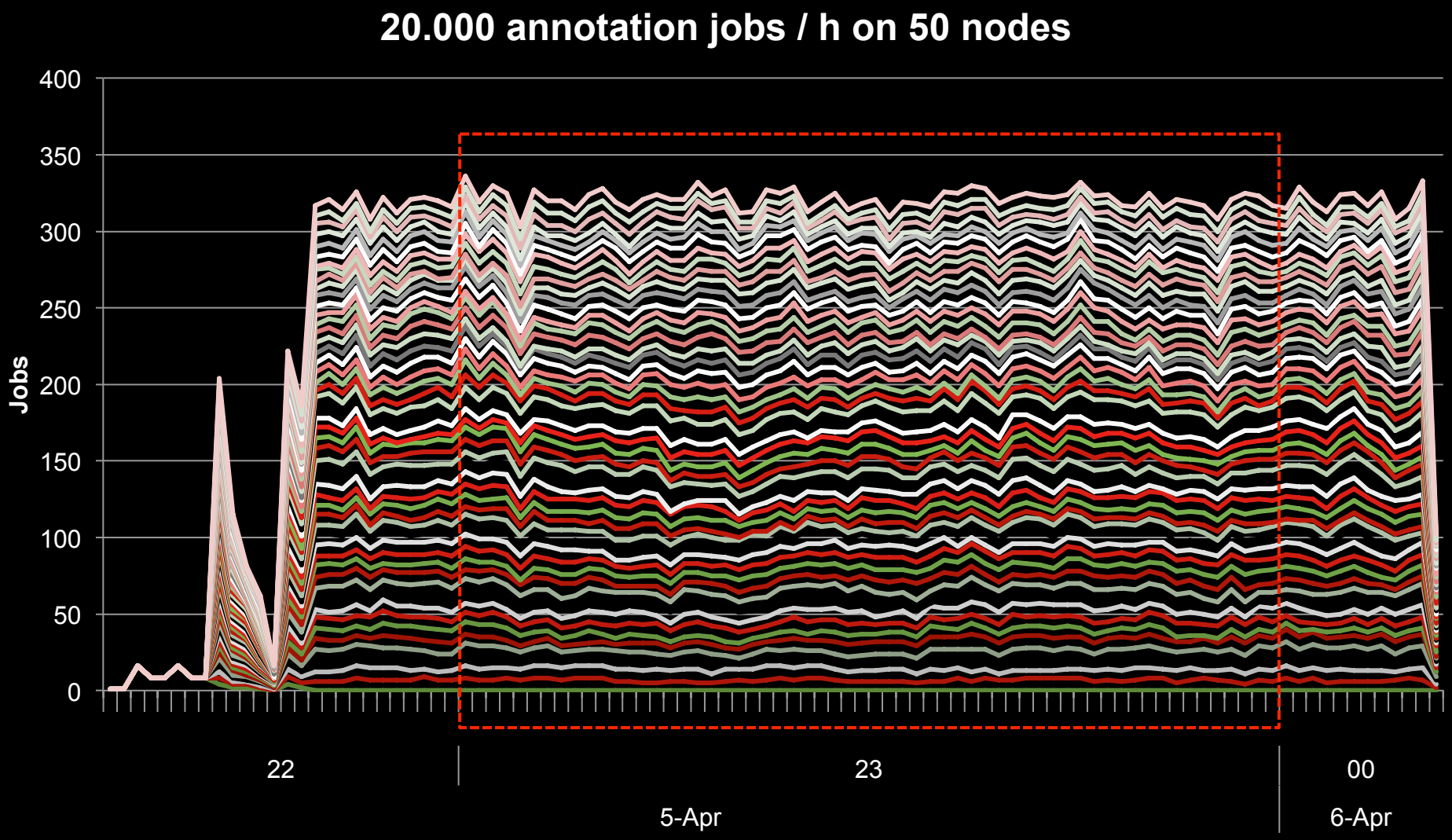
- Sun Grid Engine
  - Single image running in two modes – master/worker
  - Post-launch configuration

EMBL Dynamic Cloud Architecture



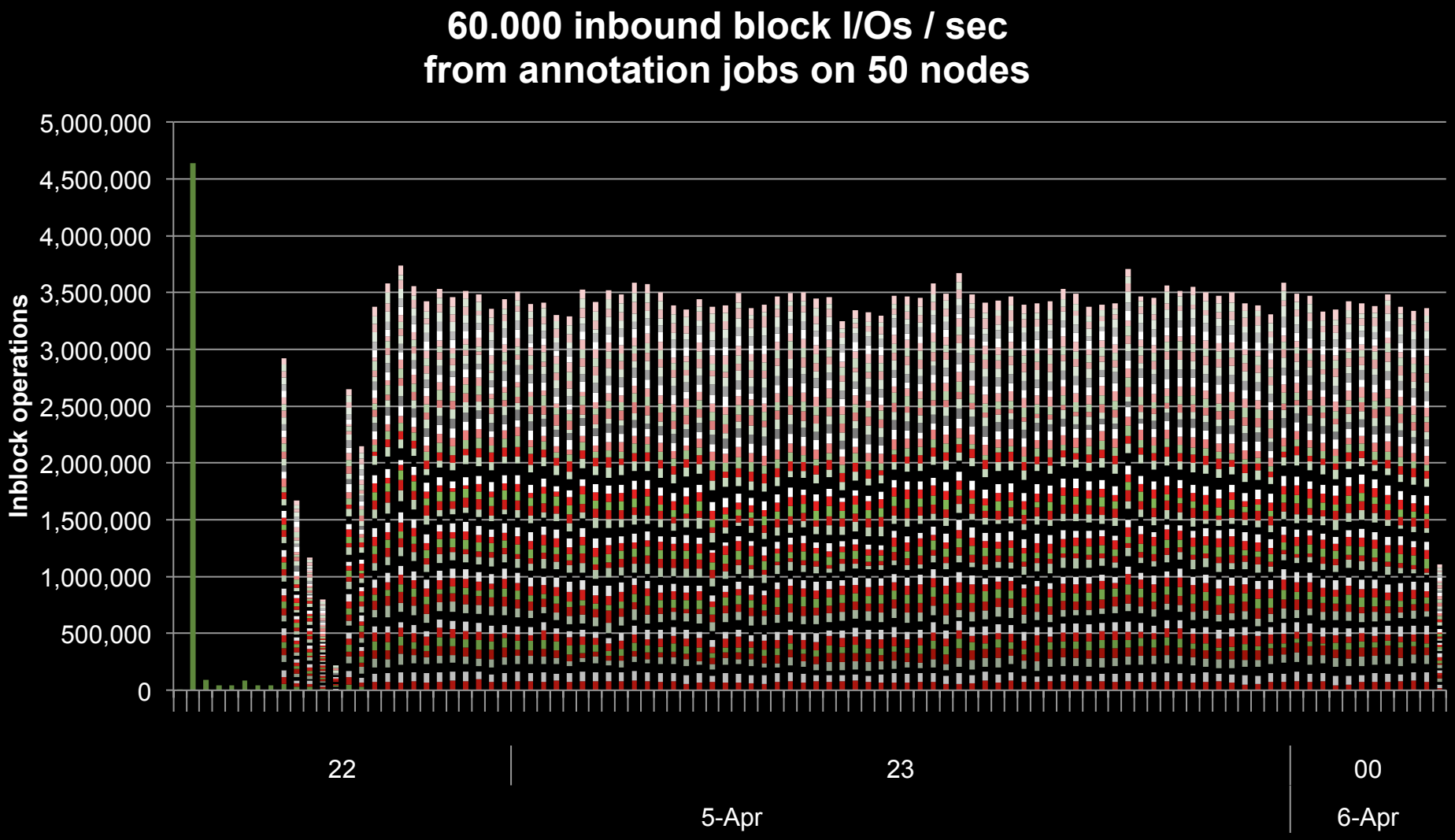
Sun Grid Engine cluster troughput

HELIX NEBULA THESCIENCECLOUD



GlusterFS throughput

HOLIX NEBULA THESCIENCECLOUD

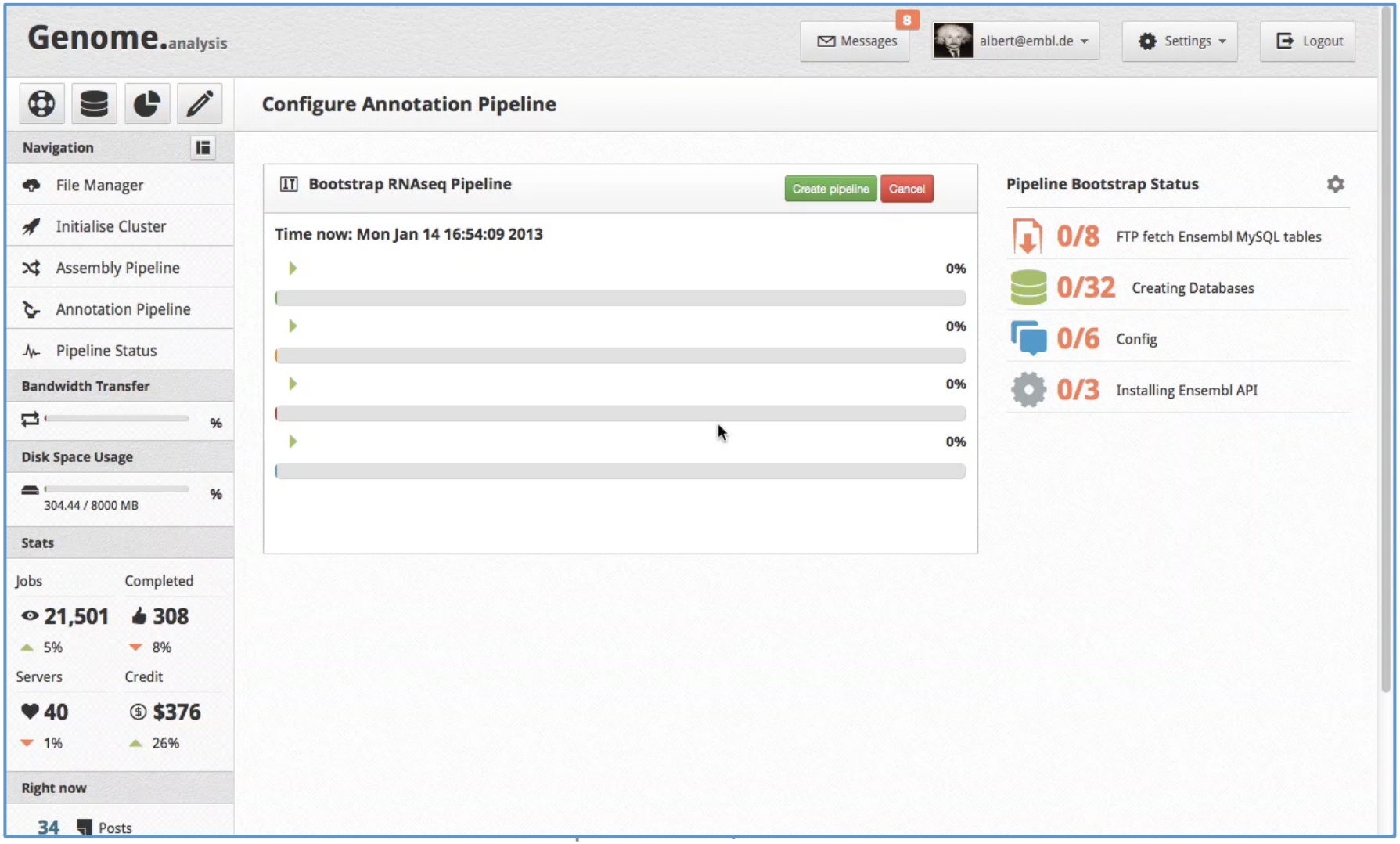


PoC results

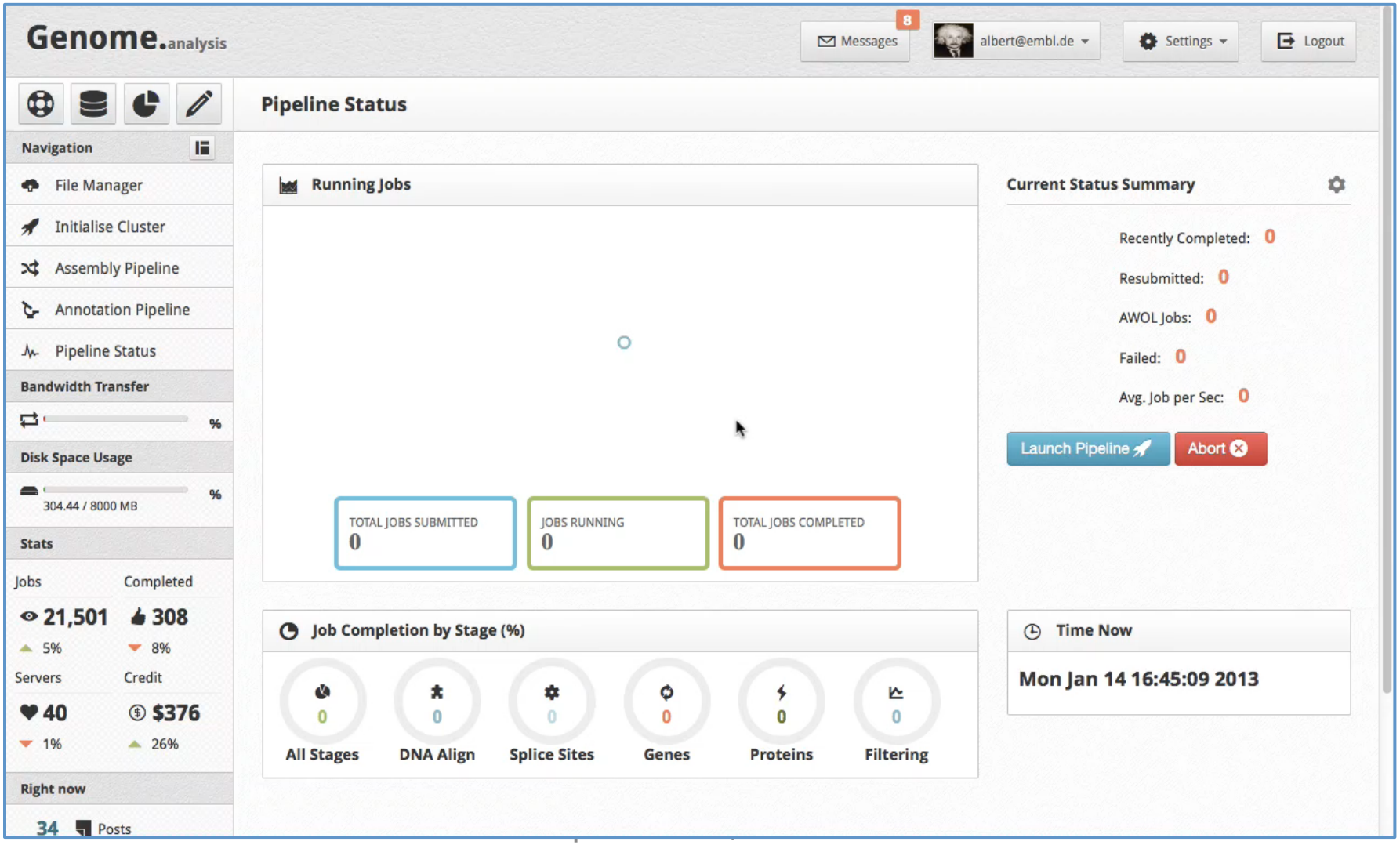
Successful tests of all vendors deployed so far

- StarCluster API integration
- auto-provision 50-node cluster setups
- real world large genome sequencing data
- 100,000s of jobs
- mix of quick parallel jobs and long running serial jobs
- glusterFS stability under high I/O levels
- Initial hurdles (e.g. image deployment, StarCluster integration, network setup) solved

User Interface Genome Analysis



User Interface Genome Analysis



Next steps

- Integration of EMBL flagship with first implementations of the Helix Nebula Blue Box
- Extensive testing of this federated cloud
- Feed back results to Helix Nebula requirements management
- Continue work on the UI
- Prepare for putting EMBL genome analysis pipeline into production in 2013

Helix Nebula PoC Acknowledgements

EMBL

Michael Wahlers
Jonathon Blake
Tobias Rausch
Jürgen Zimmermann
Vladimir Benes
Christian Boulin
Rupert Lueck

EMBL- EBI

Stephen Keenan
Paul Flicek

Atos

CloudSigma

T · · Systems · · ·

sixsq