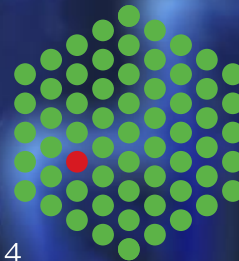


Large-Scale Genome Analysis on Helix Nebula – the Science Cloud Federated Big Data Processing On-Demand

Rupert Lueck | IT Services EMBL

Helix Nebula - The Science Cloud Public Event
CERN, Geneva
14 May 2014

EMBL
40 YEARS | 1974–2014



HELI
NEBULA
THE SCIENCE CLOUD

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

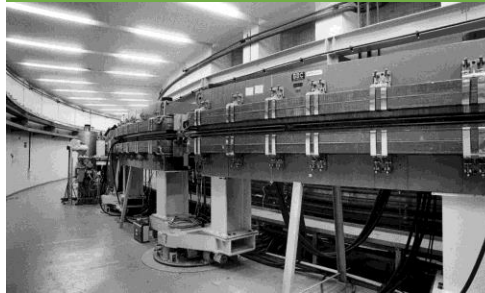
The Five Branches of EMBL

Heidelberg



Basic Molecular Biology
Research
Main Lab / Headquarters

Hamburg



Structural Biology
DESY

Hinxton



European Bioinformatics
Institute (EBI)
Sanger Centre

Grenoble



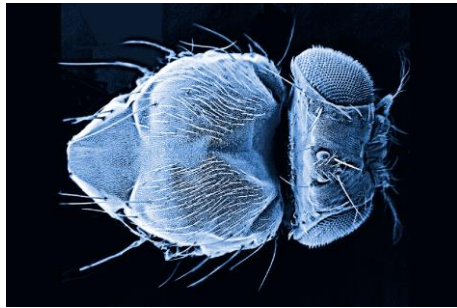
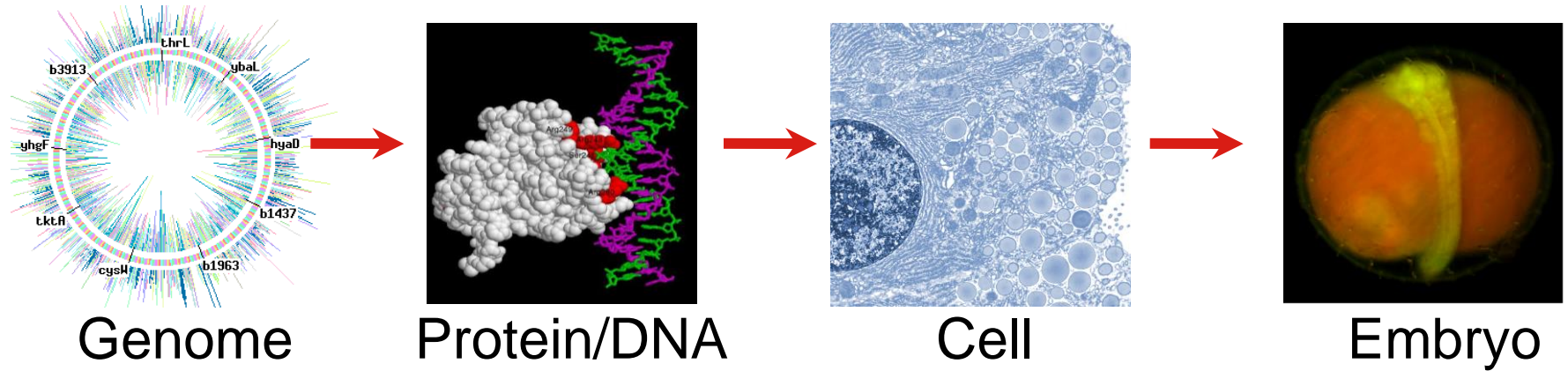
Structural Biology
ILL, ESRF, IBS, UVHCI

Monterotondo



Mousebiology
CNR, EMMA

Information Biology: From Molecules to Organisms



Development



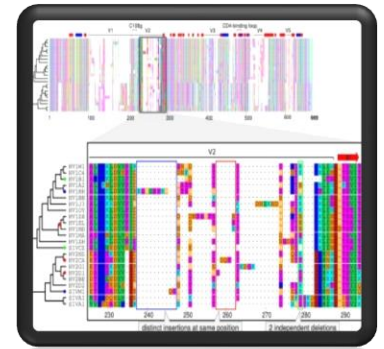
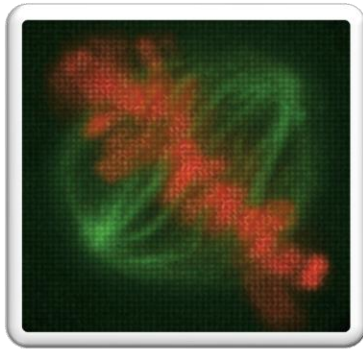
Organisms



Aging

Disease

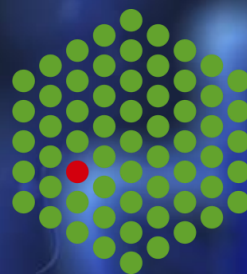
Why EMBL involves in Cloud Computing



- Information Biology research at EMBL is data-driven
 - Key technologies: Imaging, Computational Biology, Next Gen Sequencing, Modeling
 - Generate tens of TeraBytes of data every week
 - Require HPC IT infrastructures to move, store, analyse and share data
 - Project activities vary over time
- EMBL studies Cloud computing to provision on-demand and elastic IT resources to European scientists within and outside of EMBL

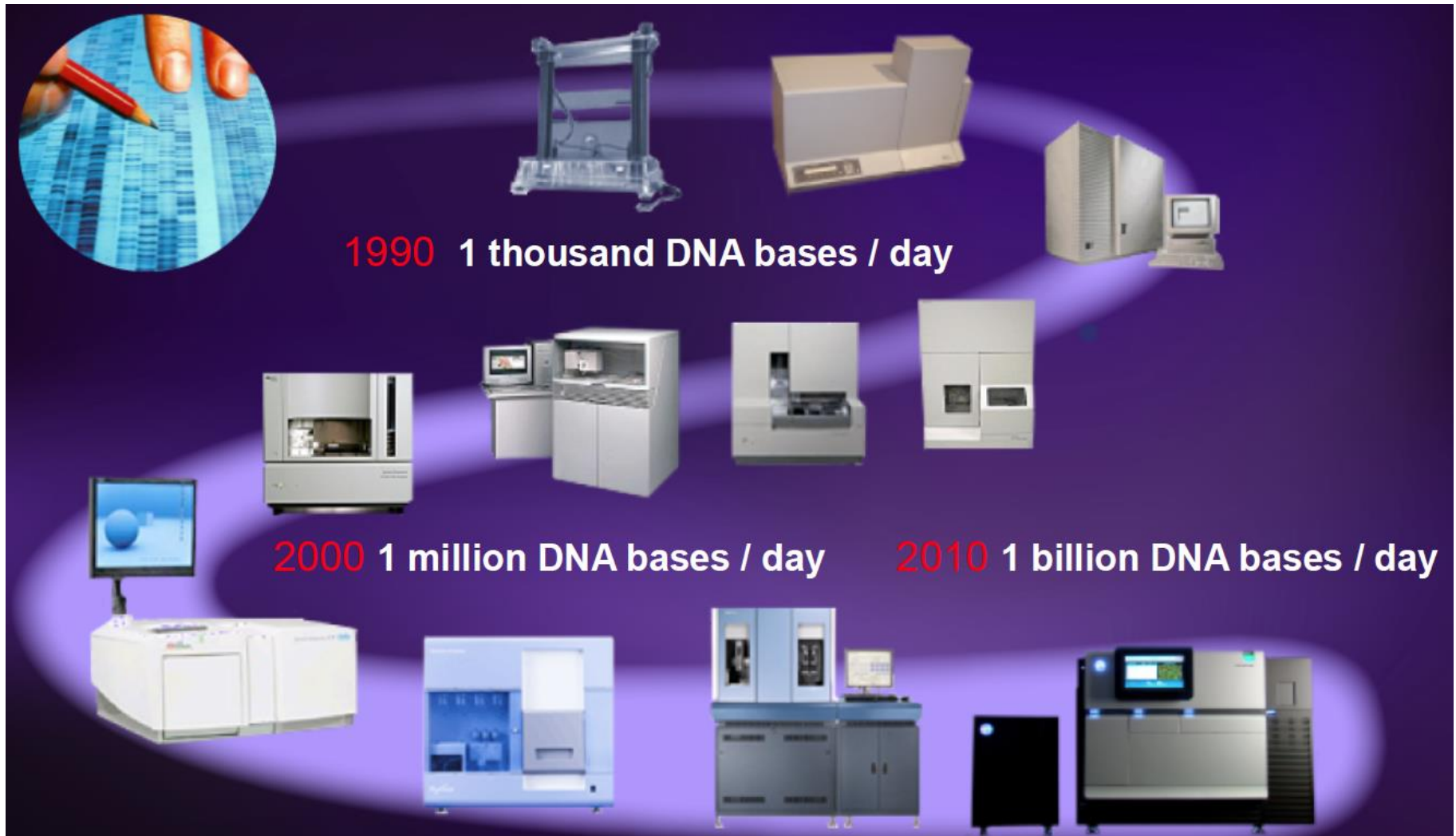
EMBL Flagship Project
*Genomic Assembly and
Annotation in the Cloud*

EMBL

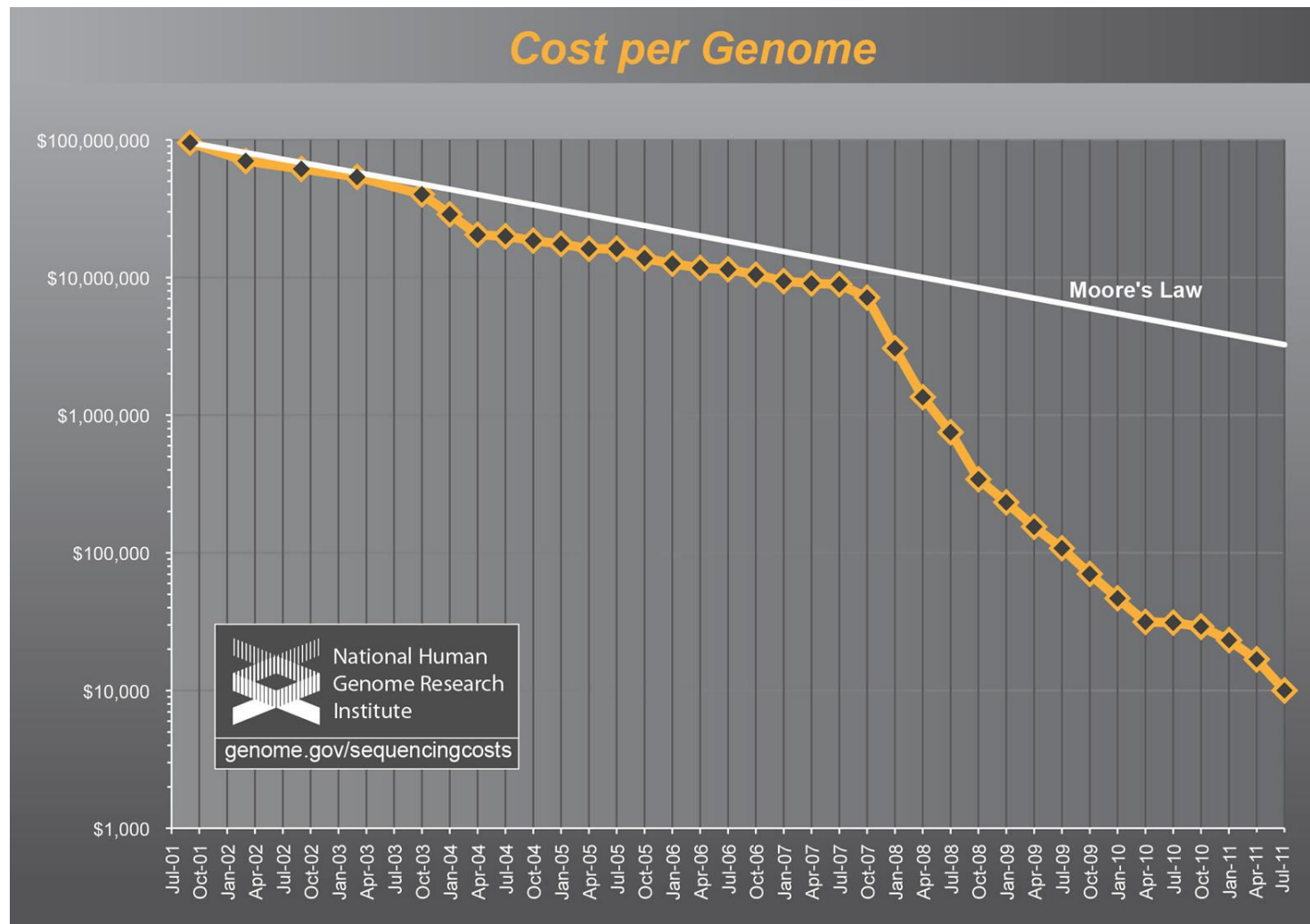


HOLIX
NBULA
THE SCIENCE CLOUD

Next Generation Sequencing (NGS) Revolution



Cost of Sequencing Decreasing Rapidly



Genomic Sequencing is Now an Affordable Solution

Academic
Research
Groups



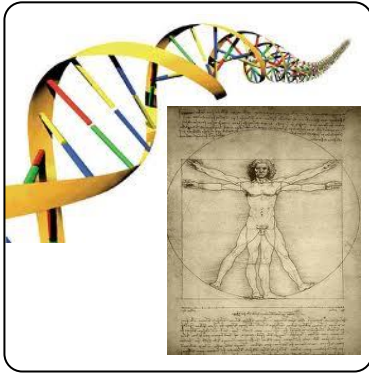
The image shows two overlapping web browser windows. The background window is the '1000 Genomes' website, featuring a dark header with the title '1000 Genomes' and the subtitle 'A Deep Catalog of Human Genetic Variation'. It includes a navigation bar with 'Home', 'About', and 'Data' links, and a sidebar with sections like 'ABOUT THE 1000 GENOMES PROJECT' and 'PROJECT OVERVIEW'. The foreground window is the 'GENOME 10K' website, which has a light blue header with the 'GENOME 10K' logo and navigation links for 'Database & Species Lists', 'News', 'Events', 'Publications', 'Participants', and 'For G10K Organizers (restricted)'. The main content area features a large blue DNA double helix graphic and the text 'GENOME 10K® Unveiling animal diversity'. Below this, it describes the 'Genome 10K Project' as a 'genomic zoo' of 10,000 vertebrate species. A 'Join us' button is visible in the bottom right corner of the GENOME 10K window.

Genomic sequencing is
now an affordable solution

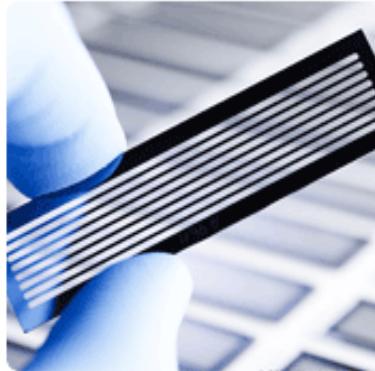
however ...

Read the Sequence to Study the Organism

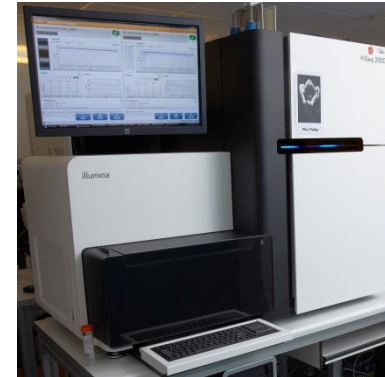
Extract DNA



Prepare



Sequence



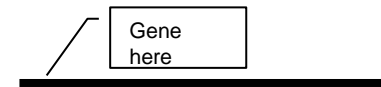
Lab

Assemble



Annotate

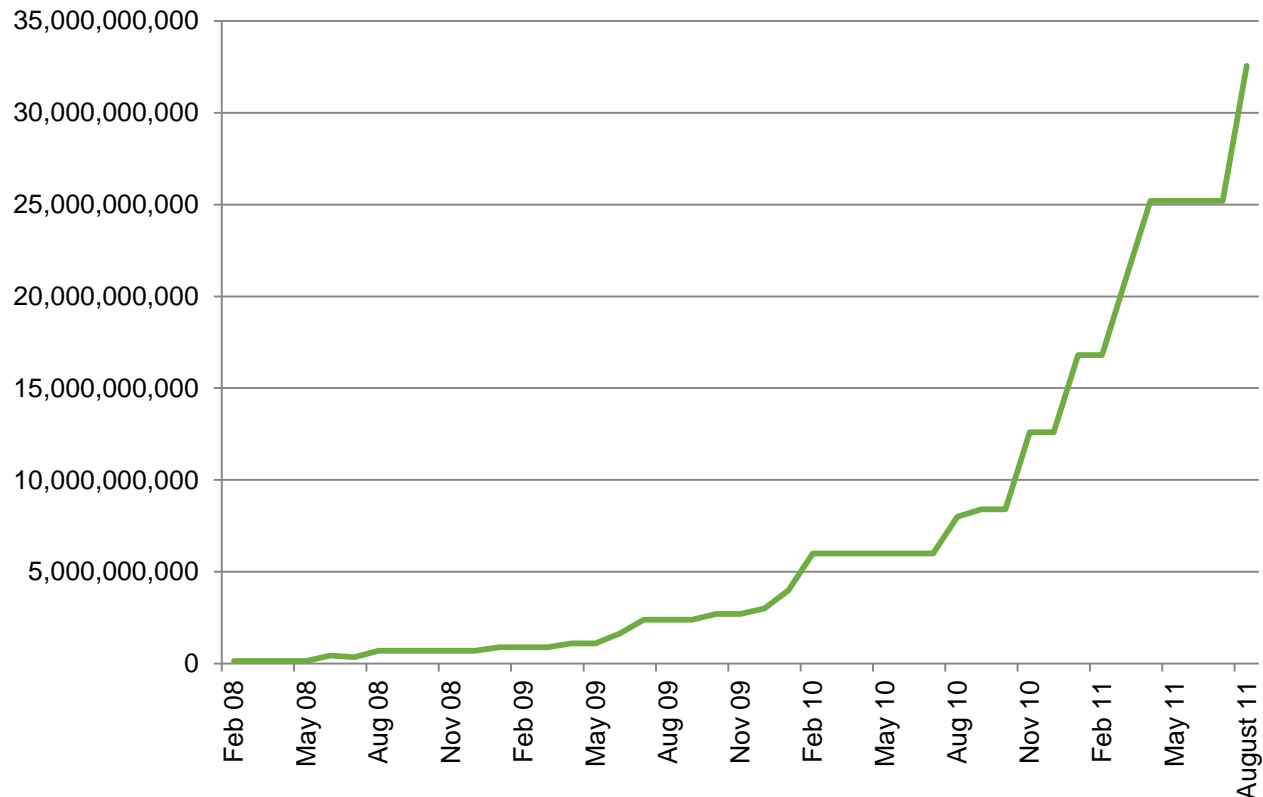
In Silico



Requires Computing Infrastructure & Expertise

Problem - Technology Explosion with NGS

**Bases Sequenced / Sample / Run @ EMBL
(Illumina)**



Sequence Production & IT Infrastructure at EMBL

5 x Illumina HiSeq2000 / 2500



1x NextSeq 500



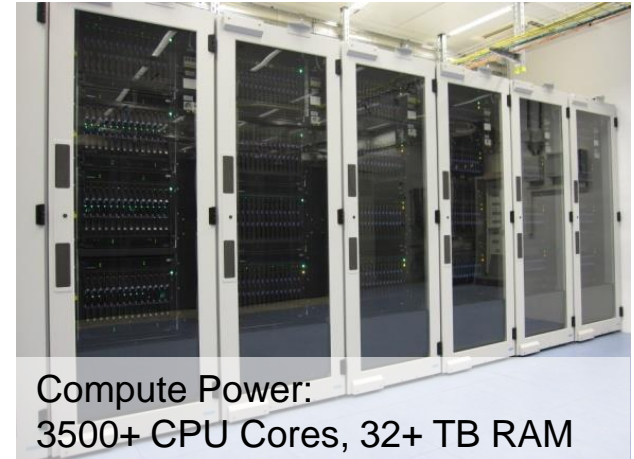
1 x MiSeq



1 x Ion Torrent



30+ TB data
each week



Compute Power:
3500+ CPU Cores, 32+ TB RAM

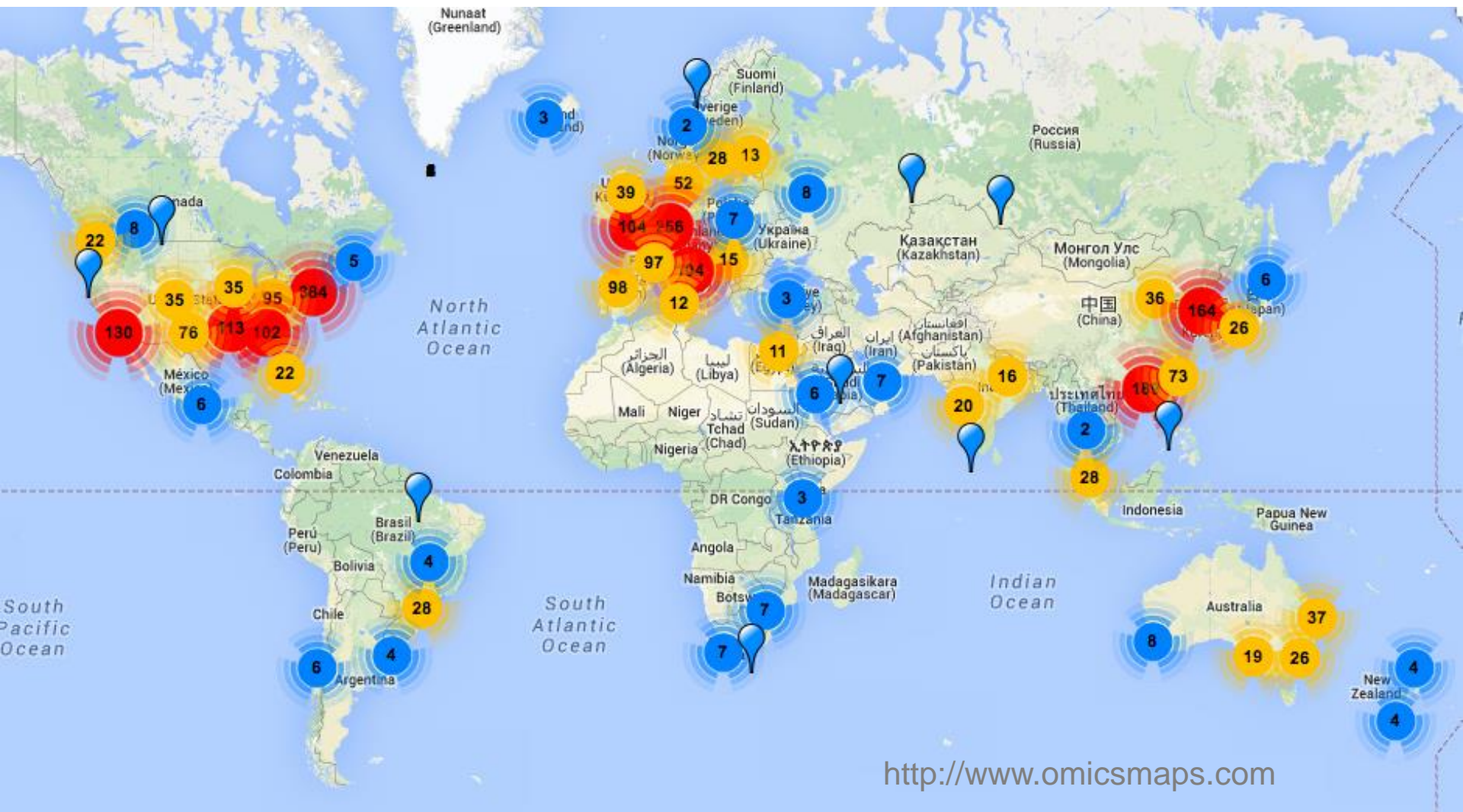


Storage:
1+ PB High Performance Disk

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+ Illumina 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

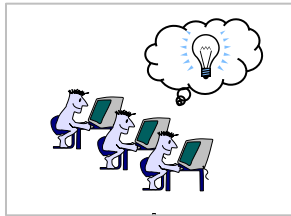
World Map of High-throughput Sequencers



<http://www.omicsmaps.com>

EMBL Helix Nebula Flagship Project

Access

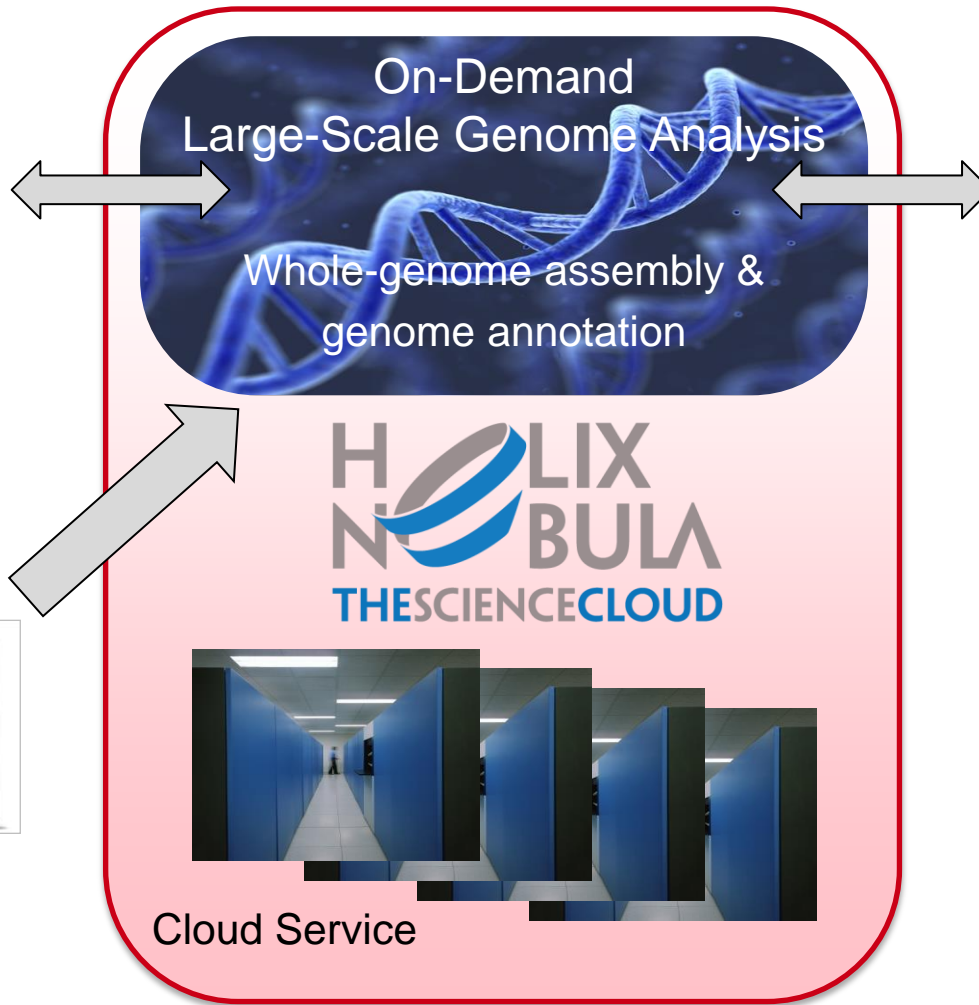


Scientists

Data acquisition



NGS Labs

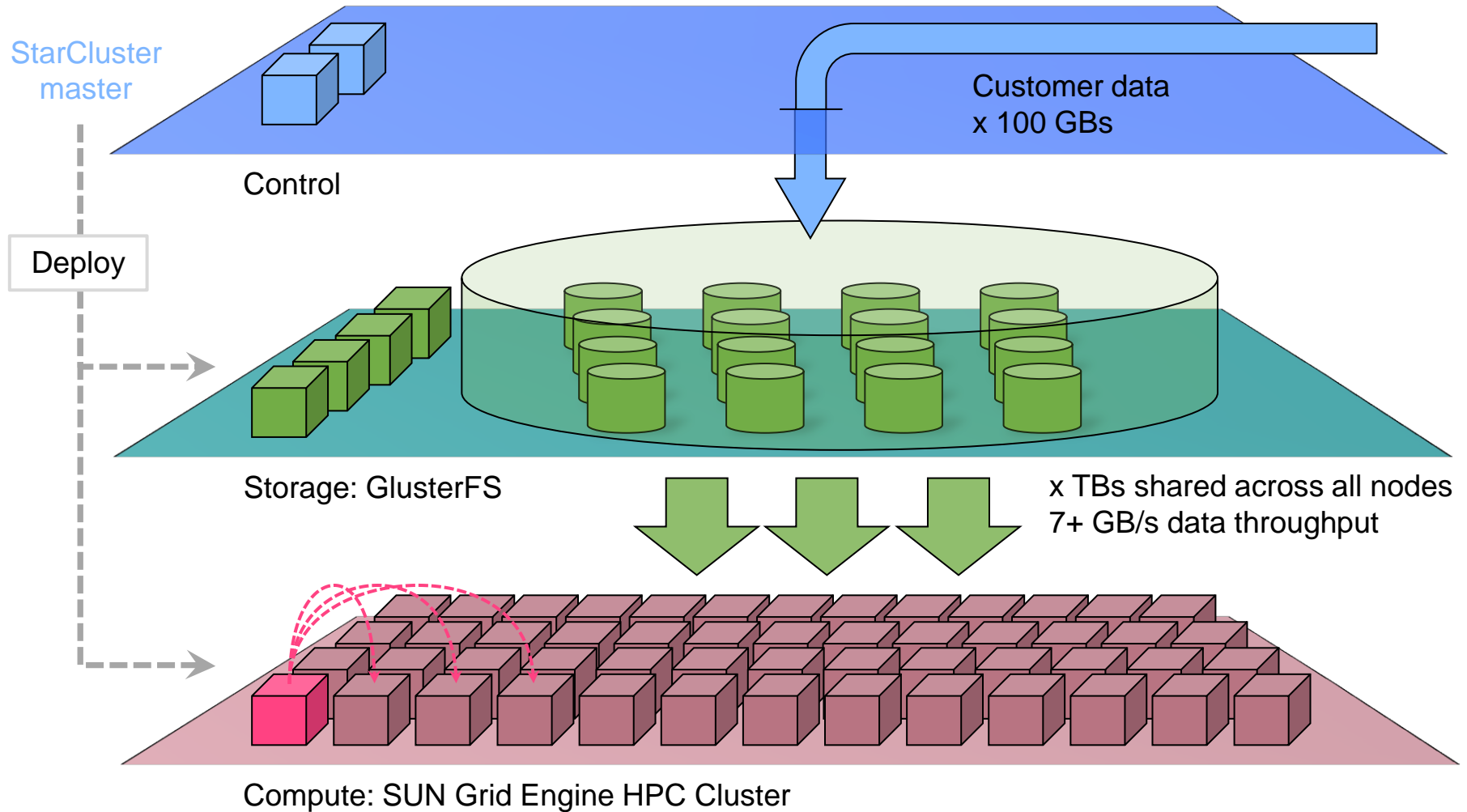


Integration
with other
cloud services
/ Archiving

Key software components

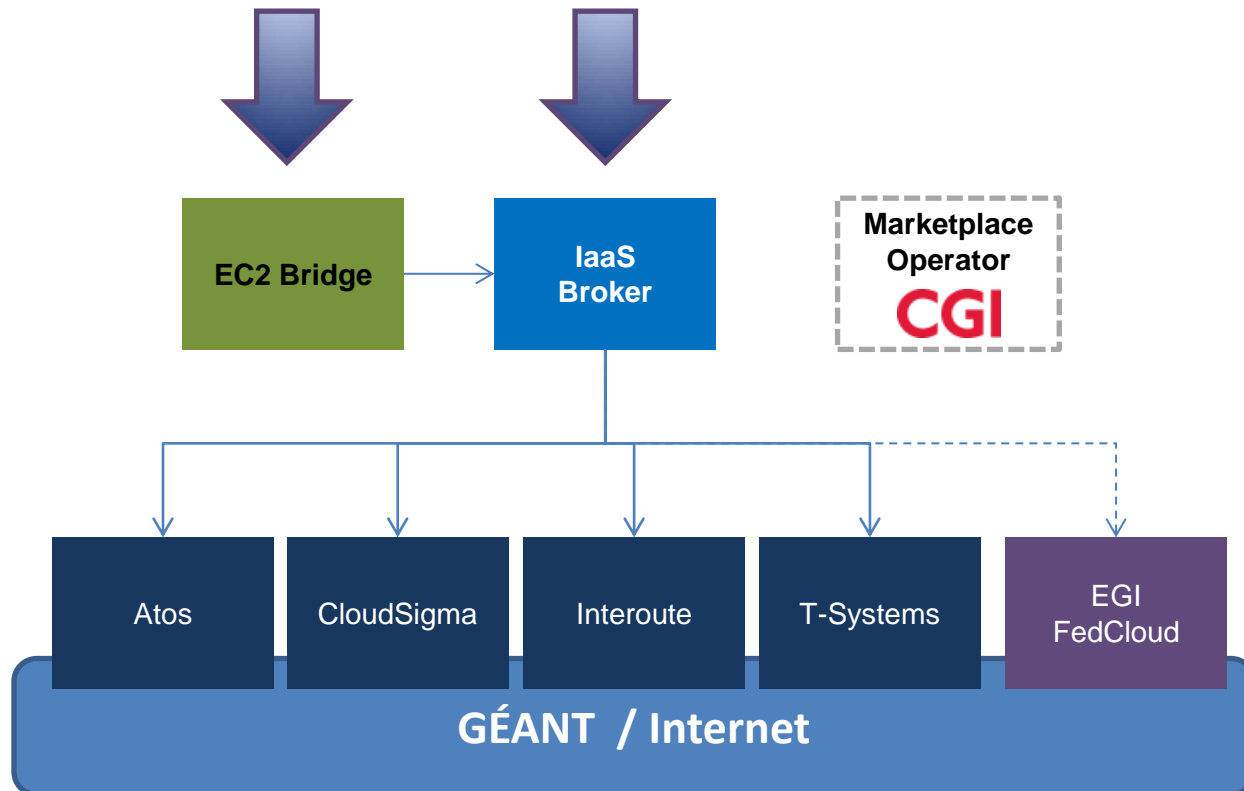
Assembly pipeline	SGA by Simpson, JT & Durbin, R http://genome.cshlp.org/content/22/3/549.long
Annotation pipeline	Ensembl
Shared file system	glusterFS, Fraunhofer FS
On-demand dynamic HPC cluster provisioning	StarCluster http://star.mit.edu/cluster/

EMBL Dynamic Cloud Architecture



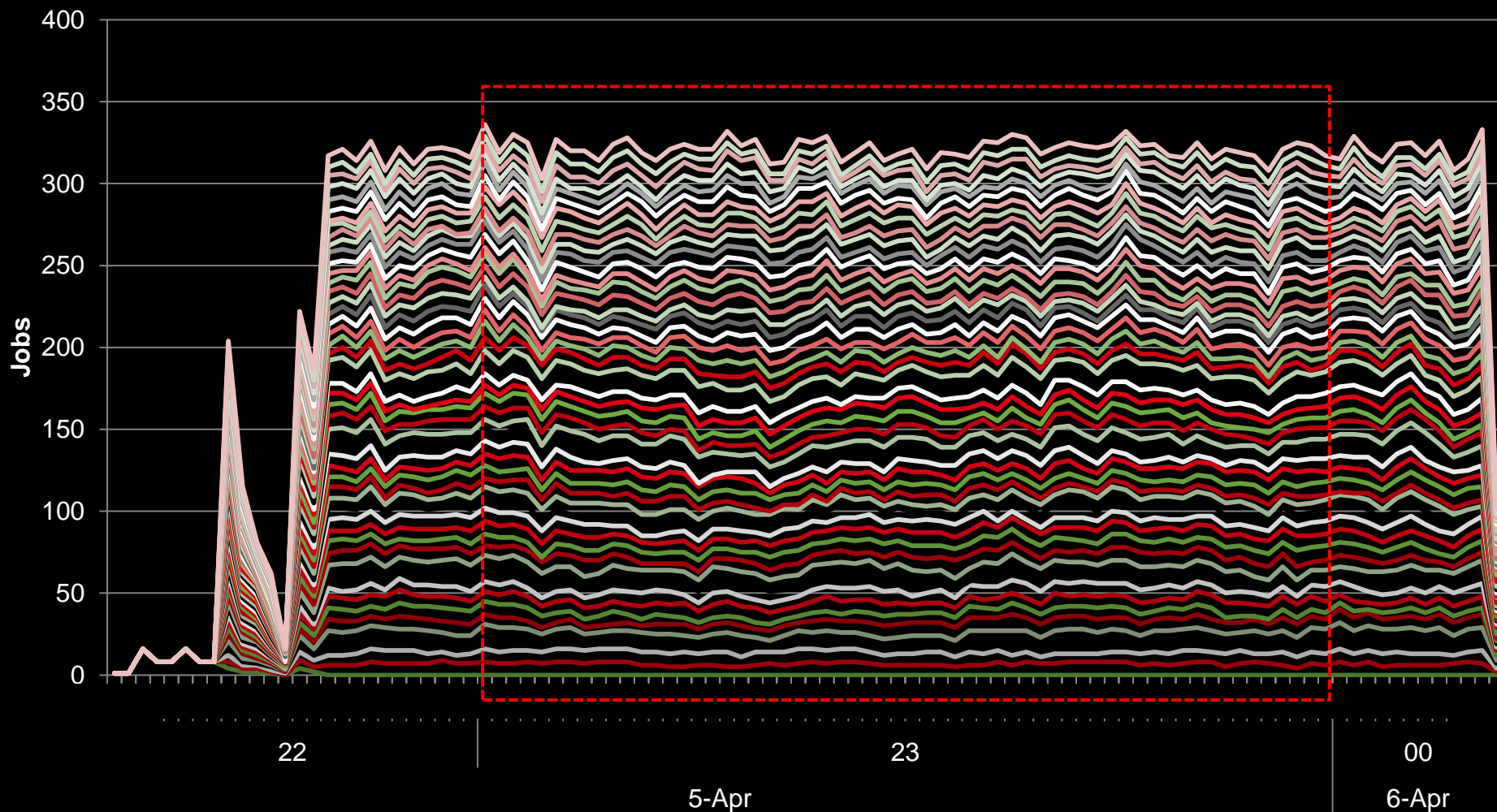


HNX v1.0



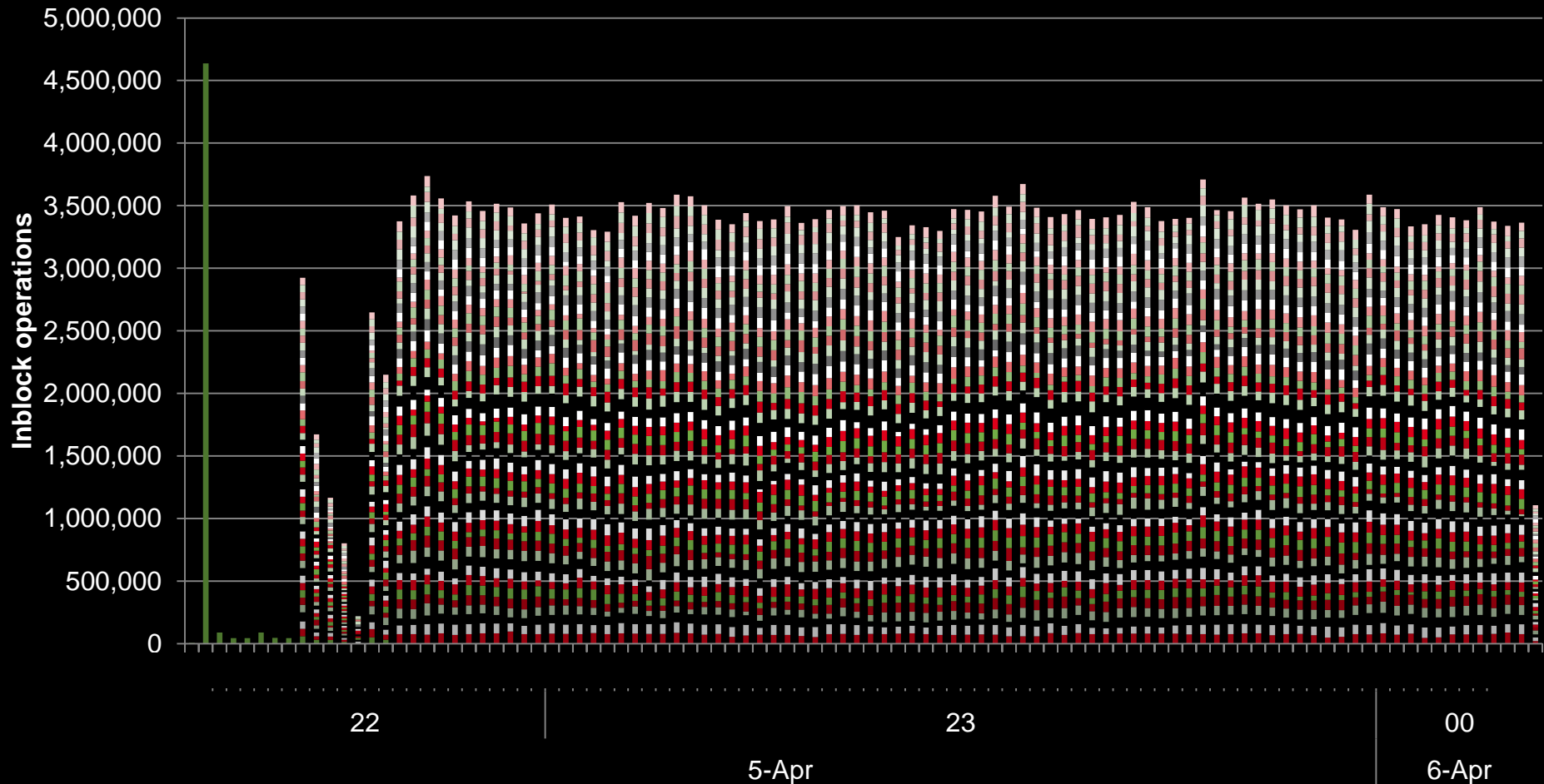
Sun Grid Engine cluster throughput

20.000 annotation jobs / h on 50 nodes



GlusterFS throughput

**60.000 inbound block I/Os / sec
from annotation jobs on 50 nodes**



Pilot test results

Successful tests of all vendors deployed so far

- StarCluster API integration
- auto-provision 50-node cluster setups
- Successful end-to-end tests of bioinformatics pipelines
- Using 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

HNX Deployment so far

- Dynamically launched cloud based HPC clusters using a set of EC2 calls initiated by StarCluster
- EMBL NGS web portal-based cluster launch is fully automated
- Initial performance test matching pre-HNX findings

EMBL NGS Cloud Portal

localhost:9000/#/ localhost:9000/#/ EMBL.dashboard

Messages 8 albert@embl.de Settings Logout

Navigation

- Initialise Cluster
- Annotation Pipeline
- Assembly Pipeline
- Pipeline Status
- File Manager

Bandwidth Transfer

Disk Space Usage

304.44 / 8000 MB

Cluster Status

Uptime N/A

Launch Time N/A





Total Nodes N/A


Genomic Assembly
Assembly from Gene Core


EMBL HELIX NEBULA THE SCIENCE CLOUD


EMBL Helix Nebula NGS platform


Globus-based Data Transfer into the Cloud





Navigation 

 Initialise Cluster



 Annotation Pipeline

 Assembly Pipeline


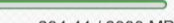
 File Manager

 Account


Bandwidth Transfer


  %


Disk Space Usage

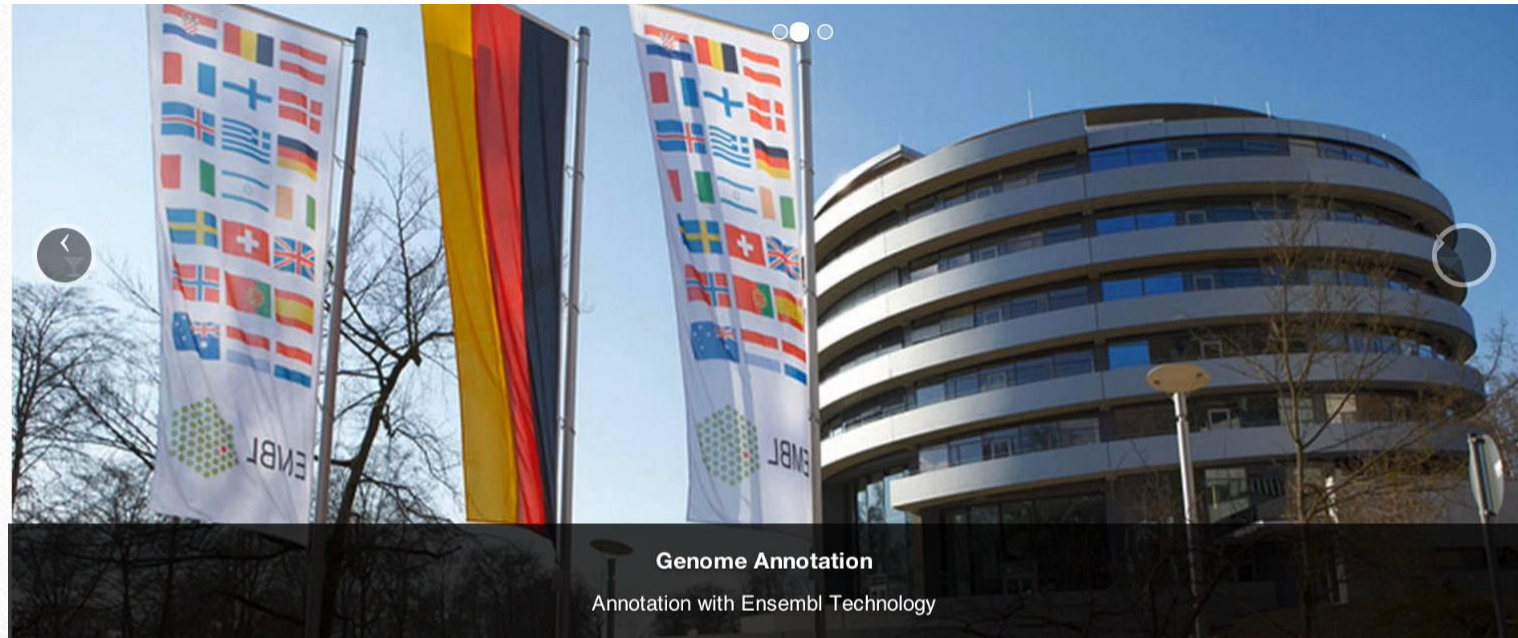
  %
304.44 / 8000 MB

Cluster Status

Uptime
 0 days, 01h:16m:52s

Launch Time
 Today at 10:44 AM


Total Nodes
 16








EMBL Helix Nebula NGS platform


StarCluster-driven Cluster Launch


EMBL dashboard


8 Messages  albert@embl.de Settings Logout





Navigation 

 Initialise Cluster

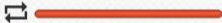
 Annotation Pipeline

 Assembly Pipeline


 File Manager

 Account


Bandwidth Transfer


 %


Disk Space Usage


 304.44 / 8000 MB %

Cluster Status

Uptime
 0 days, 00h:00m:00s

Launch Time
 Not yet launched

Total Nodes
 0



Genome Annotation
Annotation with Ensembl Technology

EMBL Helix Nebula NGS platform

Annotation Pipeline Run

The screenshot displays the EMBL dashboard interface for running an annotation pipeline. The browser window shows the URL `localhost:9000/#/annotation`. The dashboard includes a navigation sidebar on the left with options like 'Initialise Cluster', 'Annotation Pipeline', 'Assembly Pipeline', 'Pipeline Status', 'File Manager', 'Bandwidth Transfer', 'Disk Space Usage', and 'Cluster Status'. The main content area features a progress bar at the top with steps: Summary, FastQ config, Bootstrap, BAM indexes, SRA, Merging, 1st Models, Introns, and Refine Genes. The 'Summary' step is currently active. Below the progress bar, the 'Choose a Species To Get Started' section shows a dropdown menu with 'caenorhabditis_elegans' selected, and buttons for 'Get Started', 'Cancel', and 'Reset'. A message indicates '0 of 8 Steps Completed'. To the right, the 'Pipeline Summary' section lists the steps with checkboxes: 0 Summary Page (checked), 1 Fastq metadata, 2 Bootstrap Databases, 3 Make BAM Indexes, 4 make BAM Alignments, 5 merge BAM files, and 6 First pass Gene Models.

Chrome File Edit View History Bookmarks Window Help

localhost:9000/#/annotat: x

localhost:9000/#/annotation

EMBL dashboard

Messages 6 albert@embl.de Settings Logout

Navigation

- Initialise Cluster
- Annotation Pipeline
- Assembly Pipeline
- Pipeline Status
- File Manager
- Bandwidth Transfer
- Disk Space Usage
- Cluster Status

Uptime
N/A

Launch Time
N/A

Total Nodes
N/A

Summary

Choose a Species To Get Started

Choose Species caenorhabditis_elegans

Get Started Cancel Reset

0 of 8 Steps Completed

Pipeline Summary

- 0 Summary Page
- 1 Fastq metadata
- 2 Bootstrap Databases
- 3 Make BAM Indexes
- 4 make BAM Alignments
- 5 merge BAM files
- 6 First pass Gene Models

Next Steps

- Further evaluation of EMBL software stack on HNX production platform
- Testing with more suppliers
- End-to-end tests of bioinformatics pipelines
- Scale up and stability testing
- Further develop EMBL genome analysis pipeline towards production

Acknowledgements

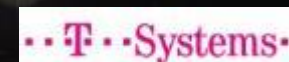
EMBL

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

EMBL- EBI

Stephen Keenan
Paul Flicek

EMBL



In Memoriam

Christian Boulin

EMBL Director of Core Facilities and Services

† 27 April 2014



EMBL

