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: 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

- Services
- Basic Research
- Technology Transfer
- Advanced Training
- European Integration
- Instrument and Technology Development
Systems Biology: From Molecules to Organisms

Genome → Protein/DNA → Cell → Embryo

Development → Organisms → Complexity → Aging → Disease
Next Generation Sequencing (NGS) Revolution

1990 1 thousand DNA bases / day

2000 1 million DNA bases / day

2010 1 billion DNA bases / day
Cost of Sequencing Decreasing Rapidly

Cost per Genome

Moore's Law

National Human Genome Research Institute
genome.gov/sequencingcosts
Genomic Sequencing is Now an Affordable Solution

Academic Research Groups

1000 Genomes
A Deep Catalog of Human Variability

Genome 10K Project
To understand how complex animal life evolved through changes in DNA and use this knowledge to become better stewards of the planet.

The Genome 10K project aims to assemble a genomic zoo—a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus.
Genomic sequencing is now an affordable solution

however ...
Read the Sequence to Study the Organism

Extract DNA

Prepare

Sequence

Assemble

Annotate

Requires Computing Infrastructure & Expertise
Problem - Technology Explosion with NGS

Bases Sequenced / Sample / Run @ EMBL (Illumina)

February 2008 to August 2011
Sequence Production & IT Infrastructure at EMBL

4 x Ilumina HiSeq2000

Compute Power: 2000+ CPU Cores, 6+ TB RAM

25+ TB data each week

Storage: 1+ PB High Performance Disk

1 x MySeq

1 x Ion Torrent
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

- Access
  - Scientists
- Data acquisition
  - NGS Labs
- Cloud Storage
- On-demand processing
- Integration with other cloud resources, archiving, etc.

Cloud Service

- NGS Labs
- Integration with other cloud resources, archiving, etc.
Proof of Concept

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

<table>
<thead>
<tr>
<th>Assembly pipeline</th>
<th>SGA by Simpson, JT &amp; Durbin, R <a href="http://genome.cshlp.org/content/22/3/549.long">http://genome.cshlp.org/content/22/3/549.long</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotation pipeline</td>
<td>Ensemble</td>
</tr>
<tr>
<td>Shared File system</td>
<td>e.g. glusterFS</td>
</tr>
<tr>
<td>StarCluster</td>
<td><a href="http://star.mit.edu/cluster/">http://star.mit.edu/cluster/</a></td>
</tr>
</tbody>
</table>
StarCluster & Sun Grid Engine
Dynamic cluster provisioning

• **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

Storage: GlusterFS

Compute: SUN Grid Engine HPC Cluster

Customer data x 100 GBs

x TBs shared across all nodes
7+ GB/s data throughput
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
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

![Genome Analysis Interface](image)

- **Running Jobs**
  - Total Jobs Submitted: 0
  - Jobs Running: 0
  - Total Jobs Completed: 0

- **Job Completion by Stage (%):**
  - All Stages: 0%
  - DNA Align: 0%
  - Splice Sites: 0%
  - Genes: 0%
  - Proteins: 0%
  - Filtering: 0%

- **Current Status Summary:**
  - Recently Completed: 0
  - Resubmitted: 0
  - AWOL Jobs: 0
  - Failed: 0
  - Avg. Job per Sec: 0

- **Time Now:** Mon Jan 14 16:45:09 2013
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