

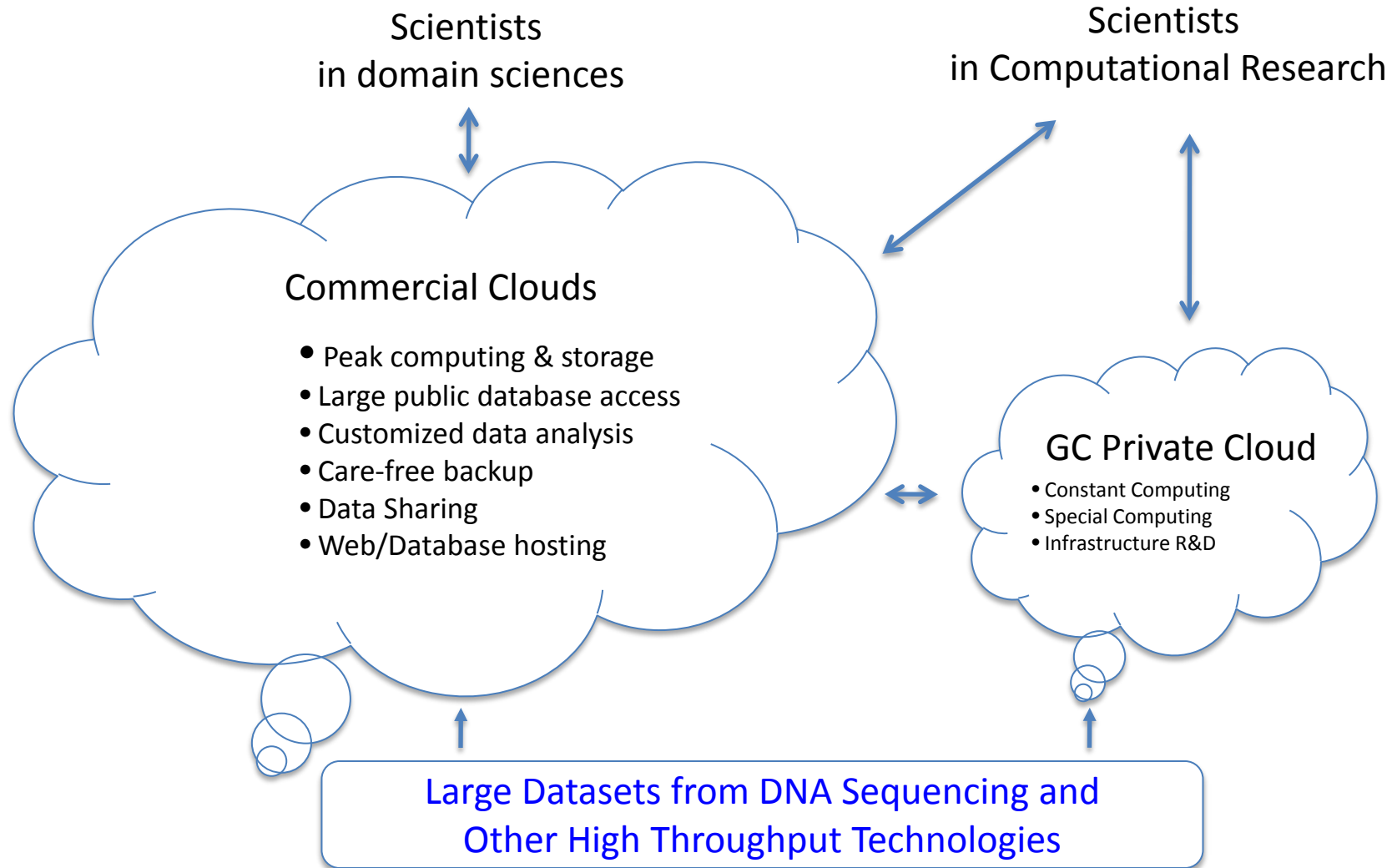


Bioinformatics In the Cloud

Dawei Lin and Brent Roger

Bioinfo-core Call, Thursday January 20, 2011

Cloud Computing at the Genome Center



Cloud Computing in Simple Terms

A Computer &
Operating System

+

Software tools
& Database

+

Storage

Computational Components

AMI
(Amazon Machine Image)

+

Configuration
& Installation

+

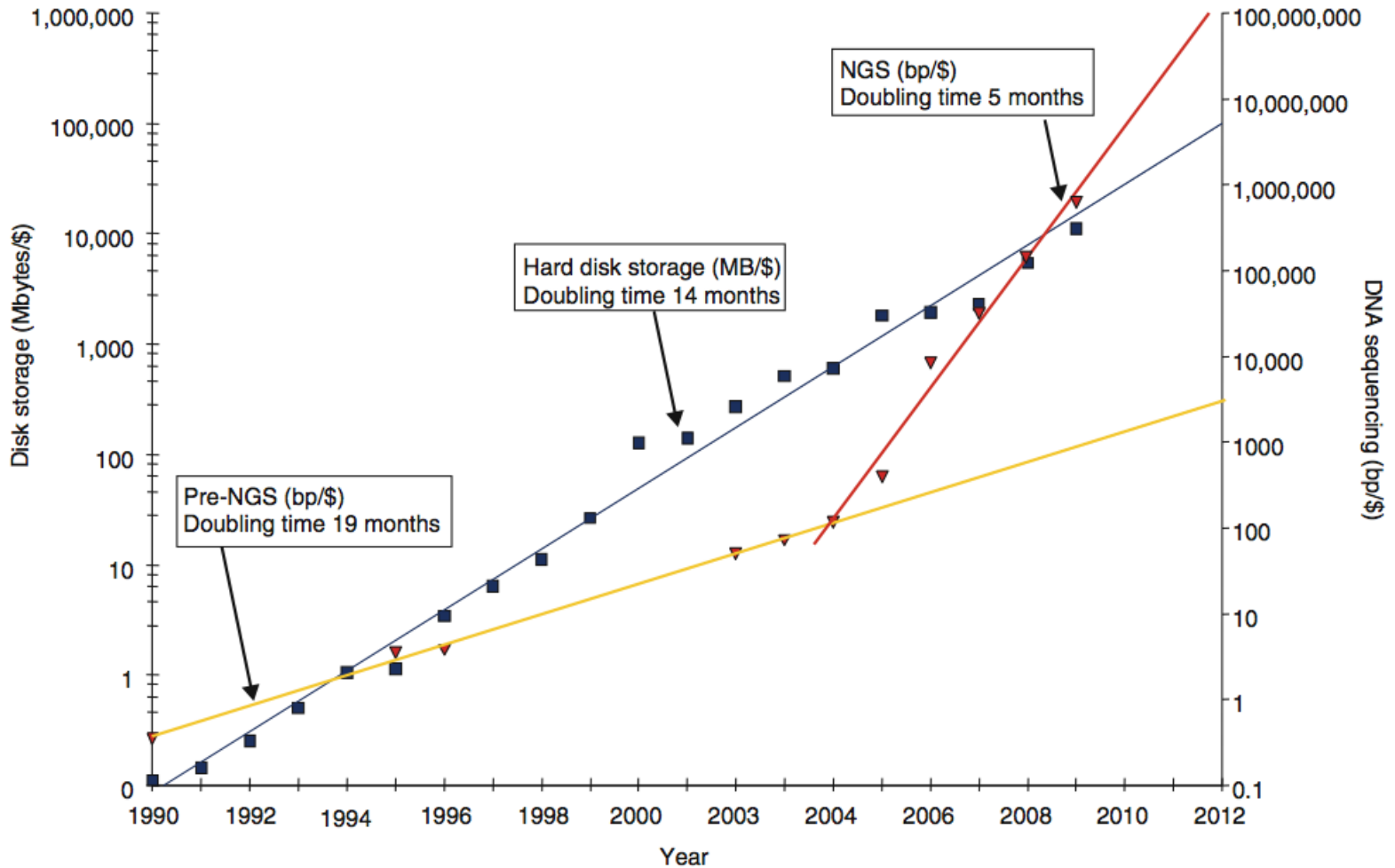
EBS/S3

Amazon Cloud Components

Key Features

- Scalability
- Community
- Affordability
- Reliability
- Security

Sequencing vs. Computing



Stein Genome Biology 2010, 11:207











Cloud Service Providers

Company	Services offered	Sectors/clients
Amazon	Application and data hosting Scalable public and private clouds	Big pharma (Eli Lilly) Large biotech (ABI) Small/startup (IXICO) Public sector (Sanger Centre/EBI)
Globus (Argonne, Illinois)	Grid services	Public sector (US Department of Energy (DoE))
Google	Software as a service Document hosting, e-mail	Large biotech (Genentech) Small/startup (De Novo)
IBM (New York)	Application and data hosting	Big pharma (Johnson & Johnson)
Microsoft	Application and data hosting	Big pharma (Johnson & Johnson) Large biotech (Genentech) Small/startup (Spirogen, London) Public sector (Sage)
Rackspace (London)	Application and data hosting, specializing in small companies	Small/startup (Spirogen)
Star Internet	Data hosting, networking	Small/startup (Science Warehouse)
Cycle Computing (Buffalo, New York)	Implementing Condor Grid workload management	Big pharma (Johnson & Johnson)
Geospiza	Sequence analysis and hosting	Large biotech (ABI)
rPath	Automating application development and maintenance	Public sector (US DoE)
Solcom	Cloud project management and software	Small/startup (Spirogen)
Univa UD (Chicago and Austin, Texas)	Managing public and private clouds	Small/startup (Pacific Biosciences)

Sansom Nature Biotech 2010, 28:13

<http://bioinformatics.ucdavis.edu>

Public Datasets in Cloud

-  **Human Liver Cohort (Sage Bionetworks)**
Human Liver Cohort characterizing gene expression in liver samples
Last Modified: Sep 8, 2010 1:56 PM
-  **C57BL/6J by C3H/HeJ Mouse Cross (Sage Bionetworks)**
C57BL/6J by C3H/HeJ mouse cross from the Jake Lusis lab at UCLA
Last Modified: Sep 8, 2010 1:53 PM
-  **Illumina - Jay Flatley (CEO of Illumina) Human Genome Data Set**
Jay Flatley (CEO of Illumina) human genome data set.
Last Modified: Jan 20, 2010 1:54 PM
-  **YRI Trio Dataset**
Complete genome sequence data for three Yoruba individuals from Ibadan, Nigeria
Last Modified: Oct 19, 2009 9:57 AM
-  **Ensembl - FASTA Database Files**
Ensembl sequence databases of transcript and translation models
Last Modified: Oct 1, 2009 3:34 PM
-  **Influenza Virus (including updated Swine Flu sequences)**
NCBI Influenza Resource Center Data.
Last Modified: Jun 4, 2009 1:18 PM
-  **AnthroKids - Anthropometric Data of Children**
Anthropometric data on children from two studies in 1975 and 1977
Last Modified: Jun 4, 2009 1:19 PM
-  **Ensembl Annotated Human Genome Data - for MySQL**
The Ensembl project produces genome databases for human as well as almost 50 other species, and makes this information freely available.
Last Modified: Aug 10, 2010 10:48 AM
-  **GenBank**
An annotated collection of all publicly available DNA sequences including more than 85.7B bases and 82.8M sequence records.
Last Modified: Dec 8, 2009 6:49 PM
-  **Unigene**
UniGene: An Organized View of the Transcriptome.
Last Modified: Jun 4, 2009 1:19 PM

<http://bit.ly/amazonpublicdata>

Open Source Bioinformatics Applications in Cloud

- Galaxy
- Crossbow
- Cloudburst
- Myrna
- Clovr
- Bioperl Max
- VIPDAC
- Cloud BioLinux
- Bionimbus
- DNAnexus
- CloudNGS (GC Bioinformatics Core)
- ..

<http://bit.ly/d66Drm>

Commercial Efforts

- BioTeam
- DNAnexus
- Complete Genomes
- Pacific Biosciences
- Geospiza
- GenomeQuest