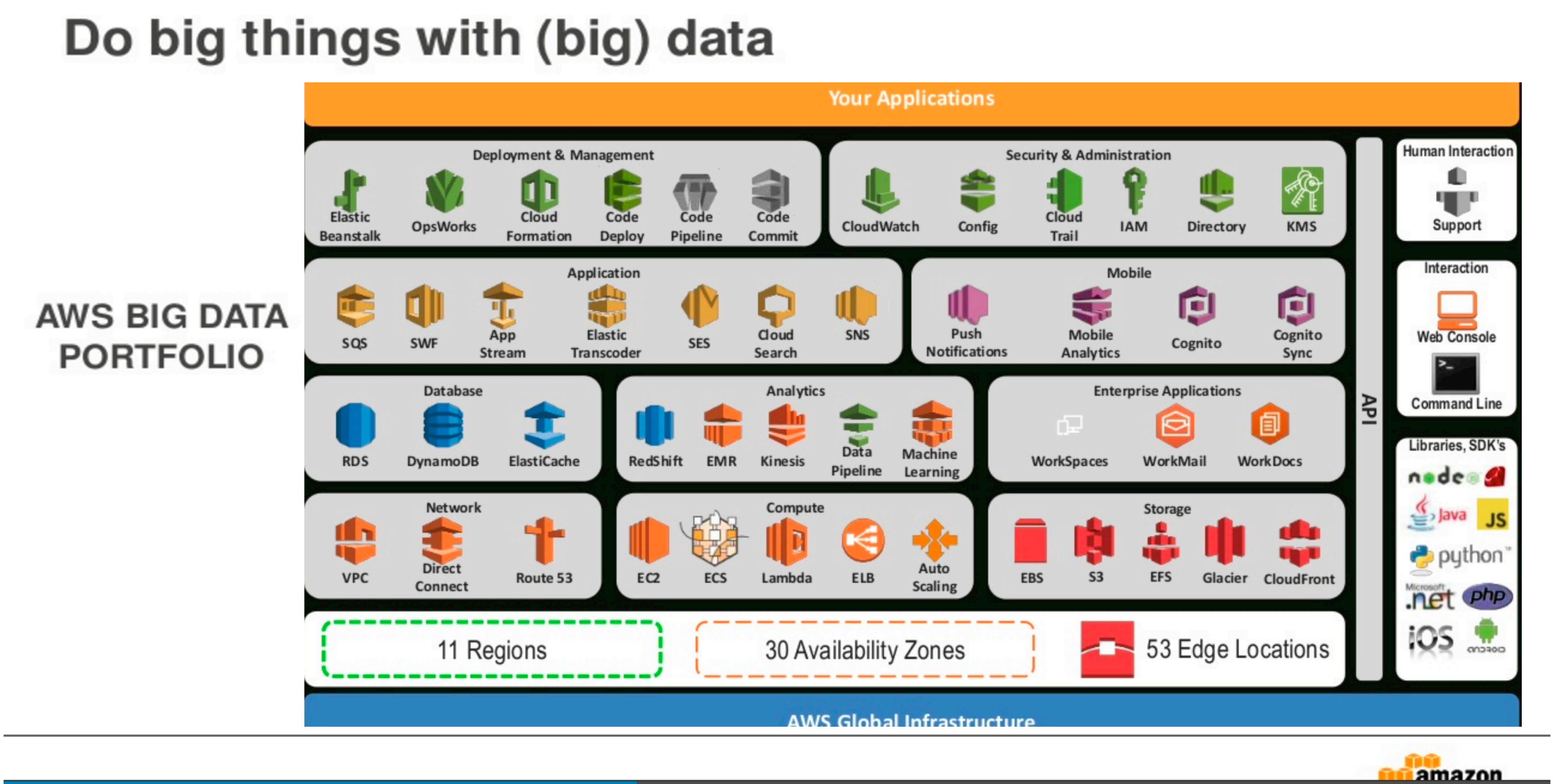
Big data Analytics in AWS Cloud

- Redshift
- EMR
- Kinesis
- Data Pipeline
- Machine learning
- …



Do big things with (big) data

AWS BIG DATA PORTFOLIO

IBM BigInsights

- BigInsights = analytical platform for persistent "big data"
  - Based on open sources & IBM technologies

- Distinguishing characteristics
  - Built-in Analytics

Big Data: Frequently Asked Questions for IBM InfoSphere BigInsights
http://www.youtube.com/watch?v=I4hsZa2jwAs

Google BigQuery

A fast, economical and fully managed data warehouse for large-scale data analytics

- Google BigQuery is a Restful web service that lets you do interactive analysis of massive datasets—up to billions of rows.
- Scalable and easy to use, BigQuery lets developers and businesses tap into powerful data analytics on demand
  - http://www.youtube.com/watch?v=P78T_ZDwQyk

content

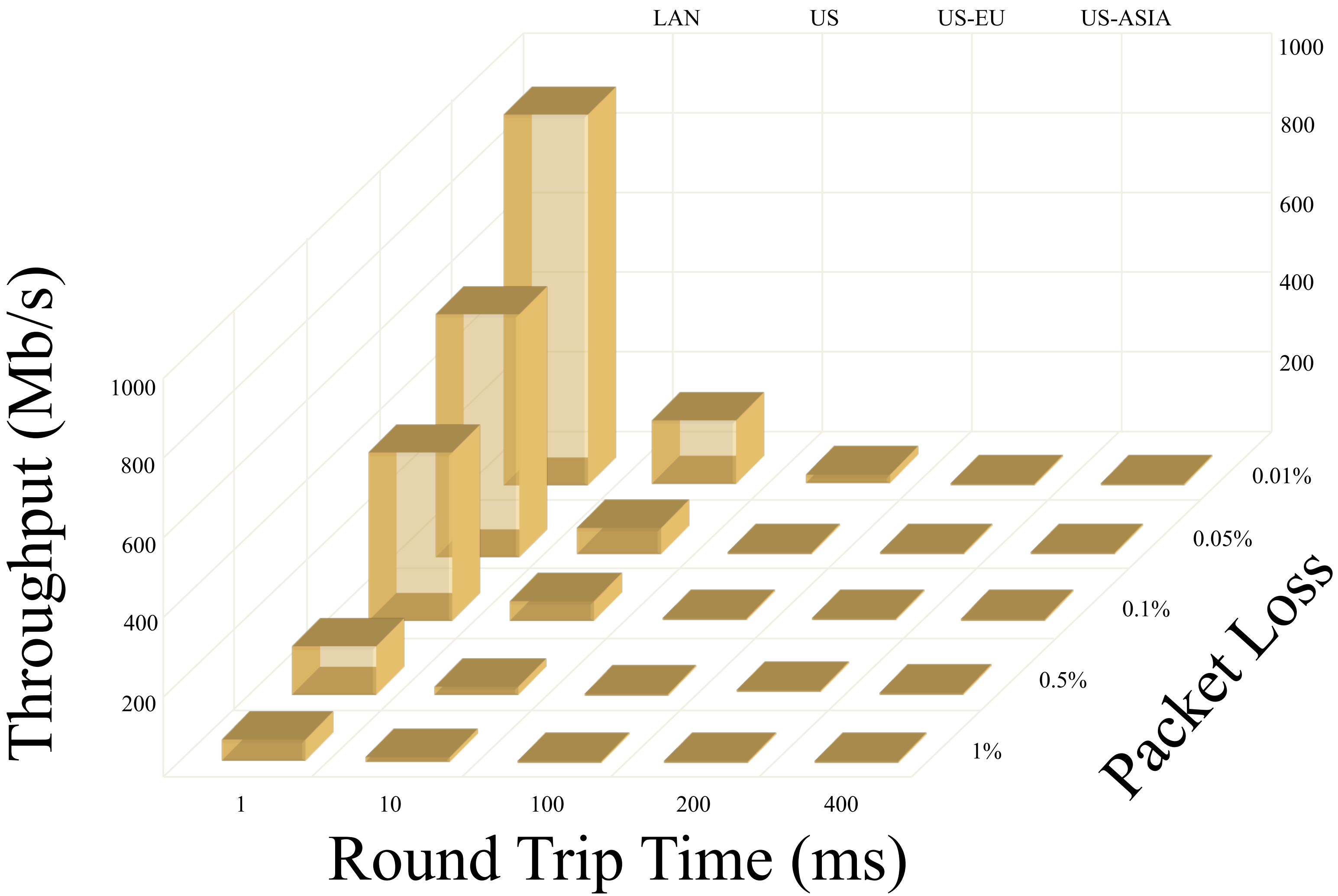
- General Introduction
- Definitions
- Data Analytics
- Solutions for Big Data Analytics
- The Network (Internet)
- When to consider BigData solution
- Scientific e-infrastructure – some challenges to overcome

The problem

- TCP Was never designed to move large datasets over wide area high Performance Networks.

- For loading a webpage,TCP is great.
- For sustained data transfer, it is far from ideal.
  - Most of the time even though the connection itself is good (let say 45Mbps), transfers are much slower.
  - There are two reason for a slow transfer over fast connections:
    - Latency
    - and packet loss bring TCP-based file transfer to a crawl.

TCP Throughput vs RTT and Packet Loss



Source: Yunhong Gu, 2007, experiments over wide area 1G.

The solutions

- Use parallel TCP streams
  - GridFTP
- Use specialized network protocols
  - UDT, FAST, etc.
- Use RAID to stripe data across disks to improve throughput when reading

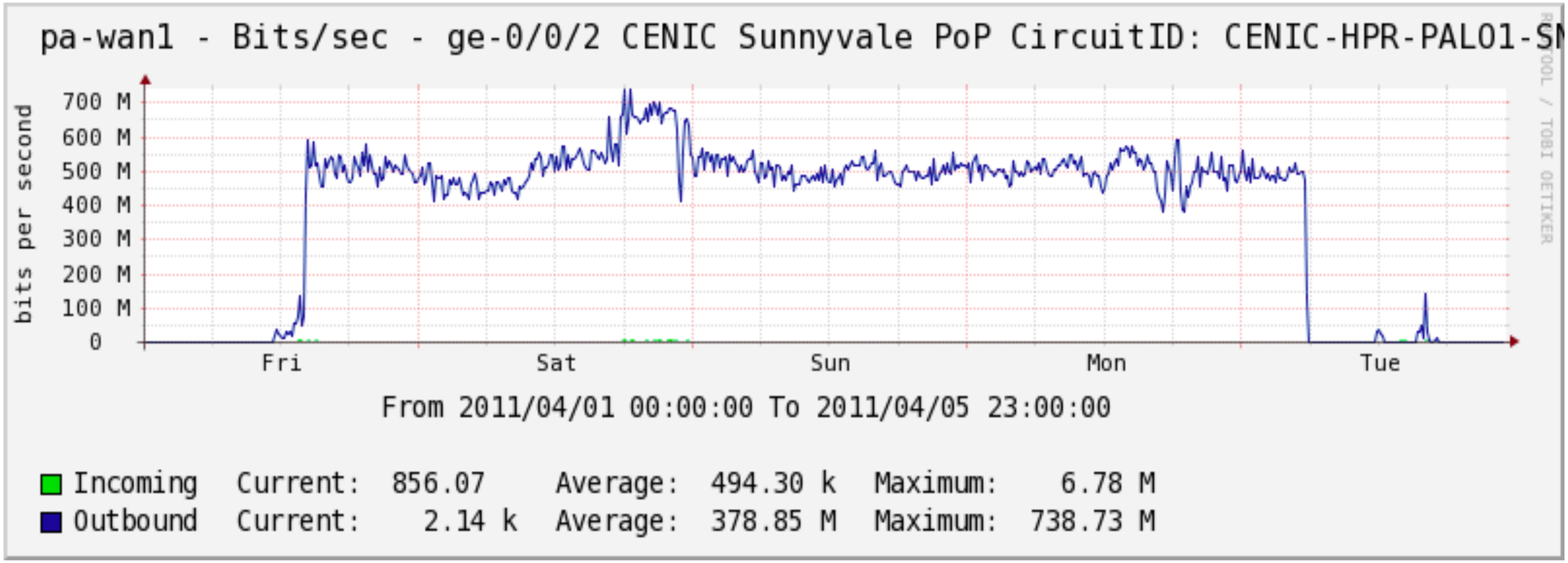
These techniques are well understood in HEP, astronomy, but not yet in biology

Moving 113GB of Bio-mirror Data

Site	**RTT**	**TCP**	**UDT**	**TCP/UDT**	**Km**
NCSA	10	139	139	1	200
Purdue	17	125	125	1	500
ORNL	25	361	120	3	1,200
TACC	37	616	120	55	2,000
SDSC	65	750	475	1.6	3,300
CSTNET	274	**3722**	304	12	12,000

- GridFTP TCP and UDT transfer times for 113 GB from gridip.bio---mirror.net/biomirror/ blast/ (Indiana USA).
  - All TCP and UDT times in minutes.
  - Source: http://gridip.bio-mirror.net/biomirror/

Case study: CGI 60 genomes



- Trace by Complete Genomics showing performance of moving 60 complete human genomes from Mountain View to Chicago using the open source Sector/UDT.
- Approximately 18 TB at about 0.5 Mbs on 1G link.

How FedEx Has More Bandwidth Than the Internet—and When That'll Change

- If you're looking to transfer hundreds of gigabytes of data, it's still—weirdly—faster to ship hard drives via FedEx than it is to transfer the files over the internet.
- "
  Cisco estimates that total internet traffic currently averages **167 terabits per second**. FedEx has a fleet of 654 aircraft with a lift capacity of 26.5 million pounds daily. A solid-state laptop drive weighs about 78 grams and can hold up to a terabyte. That means FedEx is capable of transferring 150 exabytes of data per day, or **14 petabits per second**—**almost a hundred times the current throughput of the internet**.

http://gizmodo.com/5981713/how-fedex-has-more-bandwidth-than-the-internetand-when-thatll-change

content

- General Introduction
- Definitions
- Data Analytics
- Solutions for Big Data Analytics
- The Network (Internet)
- When to consider BigData solution
- Scientific e-infrastructure – some challenges to overcome

When to Consider a Big Data Solution User point of view

- You're limited by your current platform or environment because you can't process the amount of data that you want to process

- You want to involve new sources of data in the analytics, but you can't, because it doesn't fit into schema-defined rows and columns without sacrificing fidelity or the richness of the data

When to Consider a Big Data Solution

- You need to ingest data as quickly as possible and need to work with a schema-on-demand
  - You‘re forced into a schema-on-write approach (the schema must be created before data is loaded),
  - but you need to ingest data quickly, or perhaps in a discovery process, and want the cost benefits of a schema-on-read approach (data is simply copied to the file store, and no special transformation is needed) until you know that you’ve got something that’s ready for analysis?

When to Consider a Big Data Solution

- You want to analyse not just raw structured data, but also semi-structured and unstructured data from a wide variety of sources

- you're not satisfied with the effectiveness of your algorithms or models
  - when all, or most, of the data needs to be analysed
  - or when a sampling of the data isn't going to work

When to Consider a Big Data Solution

- you aren't completely sure where the investigation will take you, and you want elasticity of compute, storage, and the types of analytics that will be pursued—all of these became useful as we added more sources and new methods

If your answers to any of these questions are "yes," you need to consider a Big Data solution.

content

- General Introduction
- Definitions
- Data Analytics
- Solutions for Big Data Analytics
- The Network (Internet)
- When to consider BigData solution
- Scientific e-infrastructure – some challenges to overcome

Scientific e-infrastructure – some challenges to overcome

- Collection
  - How can we make sure that data are collected together with the **information** necessary **to re- use them**?

- Trust
  - How can we make informed judgements about whether certain data are **authentic** and can be **trusted**?

  - How can we judge which repositories we can trust? How can appropriate access and use of resources be granted or controlled

Riding the wave, How Europe can gain from the rising tide of scientific data

Scientific e-infrastructure – some challenges to overcome

- Usability
  - How can we move to a situation where non-specialists can overcome the barriers and be able to start sensible work on unfamiliar data

- Interoperability
  - How can we implement interoperability within disciplines and move to an overarching multi-disciplinary way of understanding and using data?
  - How can we find unfamiliar but relevant data resources beyond simple keyword searches, but involving a deeper probing into the data
  - How can automated tools find the information needed to tackle data

Riding the wave, How Europe can gain from the rising tide of scientific data

Scientific e-infrastructure – some challenges to overcome

- Diversity
  - How do we overcome the problems of diversity – heterogeneity of data, but also of backgrounds and data-sharing cultures in the scientific community?
  - How do we deal with the diversity of data repositories and access rules – within or between disciplines, and within or across national borders?
- Security
  - How can we guarantee data integrity?
  - How can we avoid data poisoning by individuals or groups intending to bias them in their interest?

Riding the wave, How Europe can gain from the rising tide of scientific data

Scientific e-infrastructure – a wish list

- Open deposit, allowing user-community centres to store data easily
- Bit-stream preservation, ensuring that data authenticity will be guaranteed for a specified number of years
- Format and content migration, executing CPU-intensive transformations on large data sets at the command of the communities

Riding the wave, How Europe can gain from the rising tide of scientific data

Scientific e-infrastructure – a wish list

- Persistent identification, allowing data centres to register a huge amount of markers to track the origins and characteristics of the information
- Metadata support to allow effective management, use and understanding
- Maintaining proper access rights as the basis of all trust
- A variety of access and curation services that will vary between scientific disciplines and over time

Riding the wave, How Europe can gain from the rising tide of scientific data

Scientific e-infrastructure – a wish list

- Execution services that allow a large group of researchers to operate on the stored date
- High reliability, so researchers can count on its availability
- Regular quality assessment to ensure adherence to all agreements
- Distributed and collaborative authentication, authorisation and accounting
- A high degree of interoperability at format and semantic level

Riding the wave, How Europe can gain from the rising tide of scientific data

References

- T. Hey, S. Tansley, and K. Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery, T. Hey, S. Tansley, and K. Tolle, Eds. Microsoft, 2009. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/
- Enabling knowledge creation in data-driven science https://sciencenode.org/feature/enabling-knowledge-creation-data-driven-science.php
- Science as an open enterprise: open data for open science http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf
- Realtime Analytics for Big Data: A Facebook Case Study http://www.youtube.com/watch?v=viPRny0nq3o