Cloud Computing Resources, Grants, and Datasets

BD2K 2018 Workshop in Farmington

Today

- Research Computing Grants
  - XSEDE
  - AWS, Google, Azure
- Public Datasets
- Challenges

NSF XSEDE Program

- Originally a five-year, $121-million NSF-supported project in 2011 – XSEDE 2.0 funded in August 2016 for an additional five-years with $110-million
- 16 supercomputers and other high-end visualization and data analysis resources around the country are supported
- XSEDE Allocations are Zero-Cost (FREE) to the end user

XSEDE is the most advanced, powerful and robust collection of integrated advanced digital resources in the world. It is a single virtual system that scientists can use to interactively share computing resources, data, and expertise.

xsede.org

xsede.org/resources/overview

Credit: JAX IT Architecture Lunch - Shane Sanders

XSEDE Allocation Types

- **Trial** – rapid, but limited access to XSEDE resources
- **Campus Champions** – helps XSEDE Campus Champions get potential researchers familiar with XSEDE resources
- **Startup** – small, available for 1 year, after 1 year a Research allocation has to be requested
- **Education** – training or academic classes with specific start and end dates
- **Research** – full allocations, reviewed and awarded quarterly, merit based

XSEDE Allocation

Research Allocation Proposals:

- The research is summarized in context of the current state of the art; outlines the computational algorithms to be used; and relates those algorithms to subsections of the request.
- Provide sufficient information, without overwhelming details to the reviewers.
- The justification for the request is clear, and closely coupled to computational experiments and needs, so that if the committee needs to reduce the original request, it can be done rationally with minimum disruption to the investigator.
- Summarize results from relevant previous allocations, including manuscripts published, accepted, submitted, or in preparation, and relate these results to the current request.

XSEDE Science Gateways

Currently 33 XSEDE Science Gateways that provide web-portal interaction with XSEDE computational resources.

Some Gateways relevant to the researchers:

- Biodrugscore: A portal for customized scoring and ranking of molecules docked to the human proteome
- The CyVerse Collaborative Agave API
- IntegromeDE: Integrated database and search engine for systems biology
- ROBETTA: Automated Prediction of Protein Structure Interactions
- Neuroscience Gateway
- CIPRES: Portal for inference of large phylogenetic trees
- Computational Anatomy Gateway – image processing, visualization, and graphics

https://www.xsede.org/ecosystem/science-gateways/gateways-listing

XSEDE Campus Champions

The XSEDE Campus Champions program supports campus representatives as a local source of knowledge about high-performance and high-throughput discovery.

- Campus Champions serve as a:
  - Source of local, regional and national high-performance computing and cyberinfrastructure information
  - Source of information regarding XSEDE resources and services that will benefit research and education
  - Source of start-up accounts on your campus to quickly get researchers and educators using their allocations of time on XSEDE resources
  - Conduit for the campus high-performance computing needs, requirements and challenges, with direct access to XSEDE staff

https://www.xsede.org/web/campus-champions

Credit: IT Architecture Lunch - Shane Sanders

XSEDE Extended Collaborative Support Services

XSEDE ECSS provides XSEDE users with access to cyberinfrastructure experts with a variety of expertise, and these experts are available for collaborations lasting months to a year to help users utilize their XSEDE resources.

ECSS has expertise in a range of areas:

- Performance Analysis
- Petascale Optimization
- Coprocessor / GPU Acceleration
- I/O Optimization
- Data Analytics
- Visualization

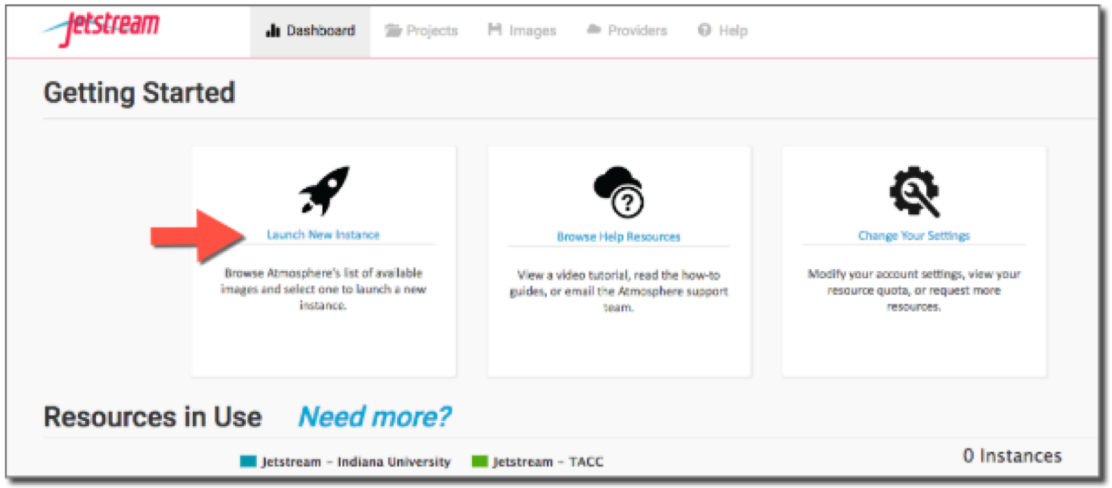
ECSS support can be requested through the standard XSEDE allocation request process.

Credit: IT Architecture Lunch - Shane Sanders

XSEDE Jetstream

- Galaxy on Jetstream
- Apply for their own Startup and Research Allocations

- Once logged in *select* **Launch an instance**



Jetstream Cloud: jetstream-cloud.org/

Galaxy on Jetstream: galaxyproject.org/cloud/jetstream/

Allocations: portal.xsede.org/allocations-overview#types-startup

Thanks to Shane Sanders for providing all the XSEDE information.

AWS Cloud Credits for Research

1. Build cloud-hosted publicly available science-as-a-service applications, software, or tools to facilitate their future research and the research of their community.
2. Perform proof of concept or benchmark tests evaluating the efficacy of moving research workloads or open data sets to the cloud.
3. Train a broader community on the usage of cloud for research workloads via workshops or tutorials.

Proposal description should address the following topics:

1. Brief description of problem to be solved.
2. Proposed AWS solution (including specific AWS tools, timeline, key milestones).
3. Plan for sharing outcomes (tools, data, and/or resources) created during project.
4. Any potential future use of AWS beyond grant duration by individual research group or broader community.
5. Names of any AWS employees you have been in contact with (this is not a prerequisite for the application).
6. Any AWS Public Data Sets to be used in your research.
7. Keywords to facilitate proposal review.

https://aws.amazon.com/research-credits/

amazon web services

Slide from AWS Research Initiatives - Sanjay Padhi

Microsoft Azure for Research awards

- Provides researchers with Azure credits
- Good for initial proof of concept projects
- Azure for Research Award: Data Science

Learn more at: microsoft.com/en-us/research/academic-program/microsoft-azure-for-research/

Proposal submission: azure4research.azurewebsites.net

Data Science Award: microsoft.com/en-us/research/academic-program/data-science-award/

Google Cloud Research Credits Program

- GCP credits can be used for any computing services on Google Cloud Platform such as storage, compute, and data analysis.
- Awards are worth $5,000 (USD) in GCP credits

Blogpost: https://www.blog.google/topics/google-cloud/google-cloud-platform-announces-new-credits-program-researchers/

Application Form: https://lp.google-mkto.com/gcp-research-credits.html

Google Research Awards

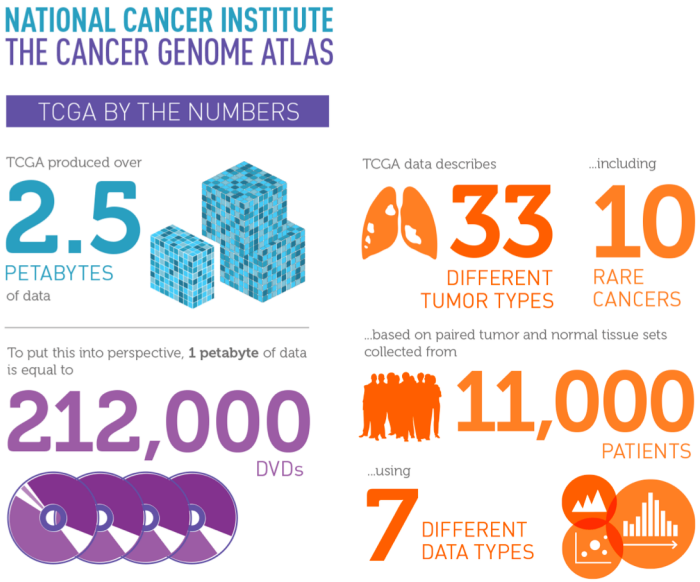
- Google Faculty Research Award
  - A Google employee must champion the proposal.
  - Seed funding to support research

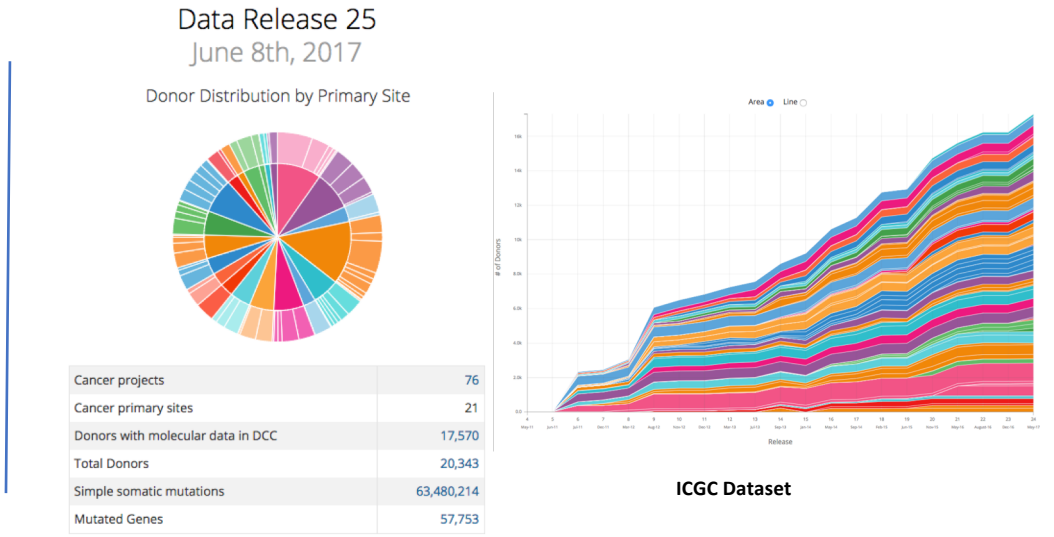
- Google Cloud Platform Education Grants for computer science
  - Teachers and faculty at universities which are regionally accredited

Google faculty research awards: research.google.com/research-outreach.html#/research-outreach/faculty-engagement/faculty-research-awards

TensorFlow Research Cloud: https://www.tensorflow.org/tfrc/

Big Genomics Datasets





ICGC Dataset

PCAWG: A Cloud-Based, Distributed Collaboration

- **International Cancer Genome Consortium (ICGC)**
- ~5,800 Whole Genomes
  - ~2,800 Cancer Donors
  - **~1,300 with RNASeq data**
  - ***Goal is to consistently analyze data***
- **8** sites storing and sharing data via GNOS
  - 300TB -> 900TB
- **14** Cloud (and HPC) environments
  - 3 Commercial, 7 OpenStack, 4 HPC
  - ~630 VMs, ~15K cores, ~60TB of RAM

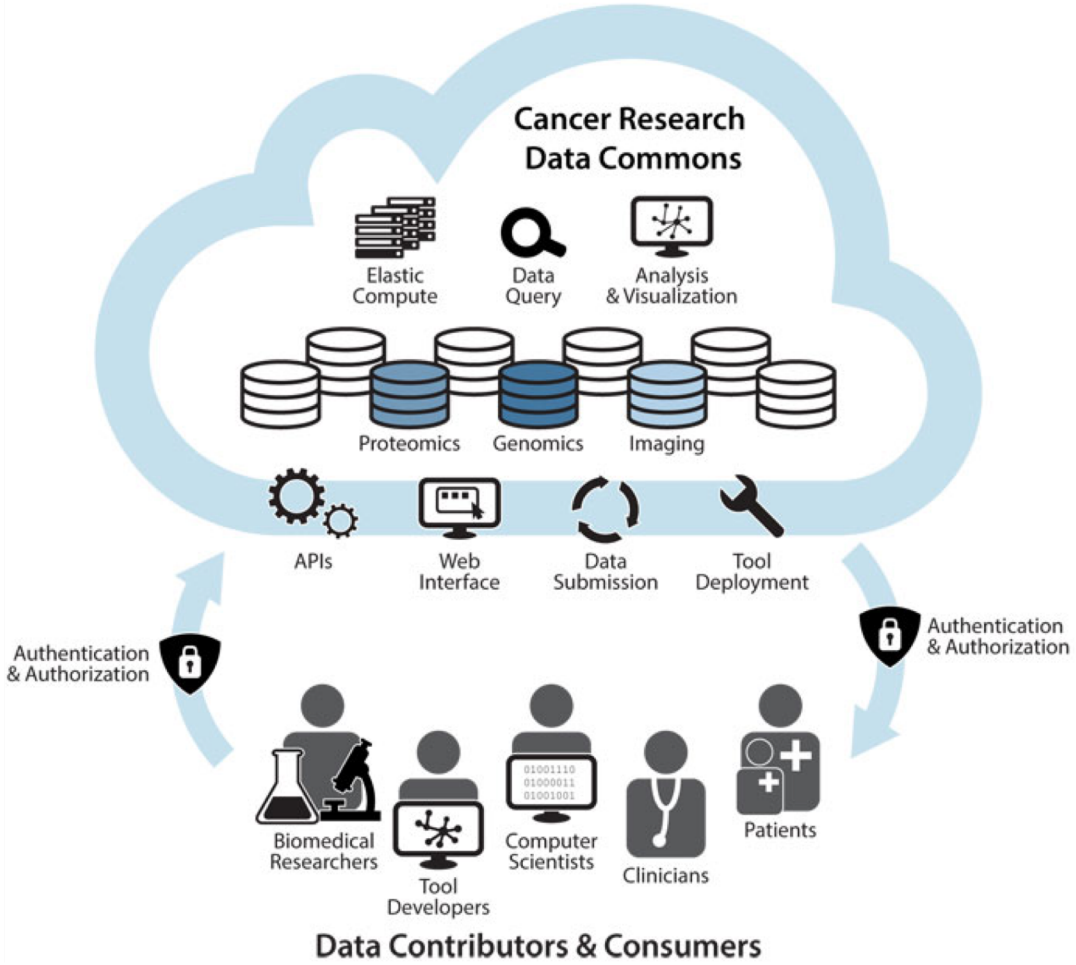
Slide from AWS re:Invent 2016: Large-Scale, Cloud-Based Analysis of Cancer Genomes: Lessons Learned from the PCAWG Project (LFS304). Credit: Brian O'Connor

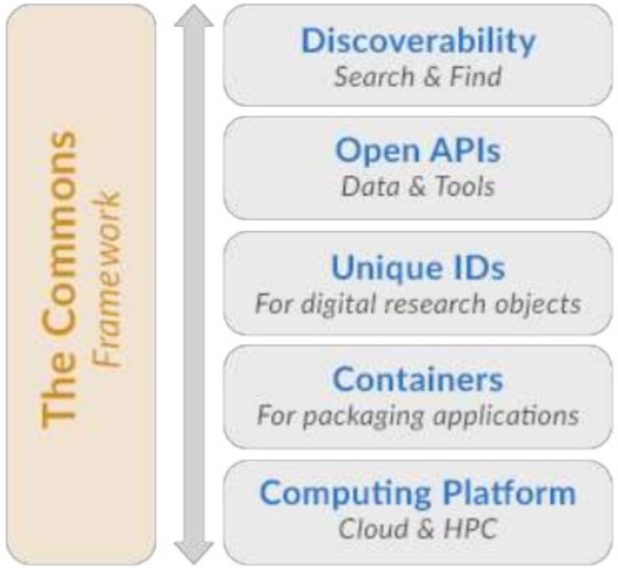
https://dcc.icgc.org/
http://docs.icgc.org/pcawg/
https://portal.gdc.cancer.gov/

Data Commons

Data commons co-locate data, storage and computing infrastructure with commonly used software services, tools & apps for analyzing and **sharing data** to create a resource for the research community.*

*Robert L. Grossman, Allison Heath, Mark Murphy, Maria Patterson and Walt Wells, A Case for Data Commons Towards Data Science as a Service, IEEE Computing in Science and Engineer, 2016. Source of image: The CDIS, GDC, & OCC data commons infrastructure at the University of Chicago Kenwood Data Center.





Presentation by Warren Kibbe: Precision Medicine in the Age of NCI MATCH and the Beau Biden Cancer Moonshot

Challenges

- High level of programming expertise
- Researchers are not IT experts
- A big learning curve
- Jargon used by each provider
- IT security, Legal and data security/privacy
- Controlled, Clinical, and HIPPA data
- Reproducibility of the analysis
- Management of resources, cloud costs and budgeting
- Data transfer rate remains a bottleneck
- Estimating costs
- Networking problems - 1GB and 10 GB
- Cloud vs On-Premise

Resources

- NCI Pilot handout

https://cbiit.nci.nih.gov/sites/nci-cbiit/files/Cloud_Pilot_Handout_508compliant.pdf

- Google Public Datasets

https://cloud.google.com/public-datasets/

- AWS HPC

https://aws.amazon.com/hpc/

- Google Genomics

https://cloud.google.com/genomics/

- XSEDE Jetstream cloud platform

http://research-it.berkeley.edu/blog/17/02/17/jetstream-cloud-support-multi-institutional-data-science-workshops-and-research

- Google Codelabs
- XSEDE Flyer:
    - https://www.xsede.org/documents/10157/169907/WhatIsXSEDE_flyer2012.pdf%20

Resources

- AWS pricing calculator

https://calculator.s3.amazonaws.com/index.html

- Microsoft Azure Pricing Calculator

https://azure.microsoft.com/en-us/pricing/calculator/

- Google Pricing Calculator

https://cloud.google.com/products/calculator/

- Data Egress Waivers
  - https://azure.microsoft.com/en-us/blog/azure-egress-fee-waiver-for-the-academic-community/
  - https://aws.amazon.com/blogs/publicsector/aws-offers-data-egress-discount-to-researchers/
- Total cost of ownership calculators
  - https://aws.amazon.com/tco-calculator/
  - https://azure.microsoft.com/en-us/pricing/tco/calculator/

Resources

- Research & Technical Computing on AWS

https://aws.amazon.com/government-education/research-and-technical-computing/

- AWS Research Cloud Program

https://aws.amazon.com/government-education/research-and-technical-computing/research-cloud-program/

- AWS Educate

https://aws.amazon.com/education/awseducate/

- Google Cloud Platform Education Grants

https://cloud.google.com/edu/

- Global Alliance for Genomics and Health

http://genomicsandhealth.org/

- Galaxy Capacity Planning

https://galaxyproject.org/cloudman/capacity-planning/

- Teaching Cloud Computing
  - https://ieeexplore.ieee.org/document/7562336/

Thank you

AWS Public Datasets

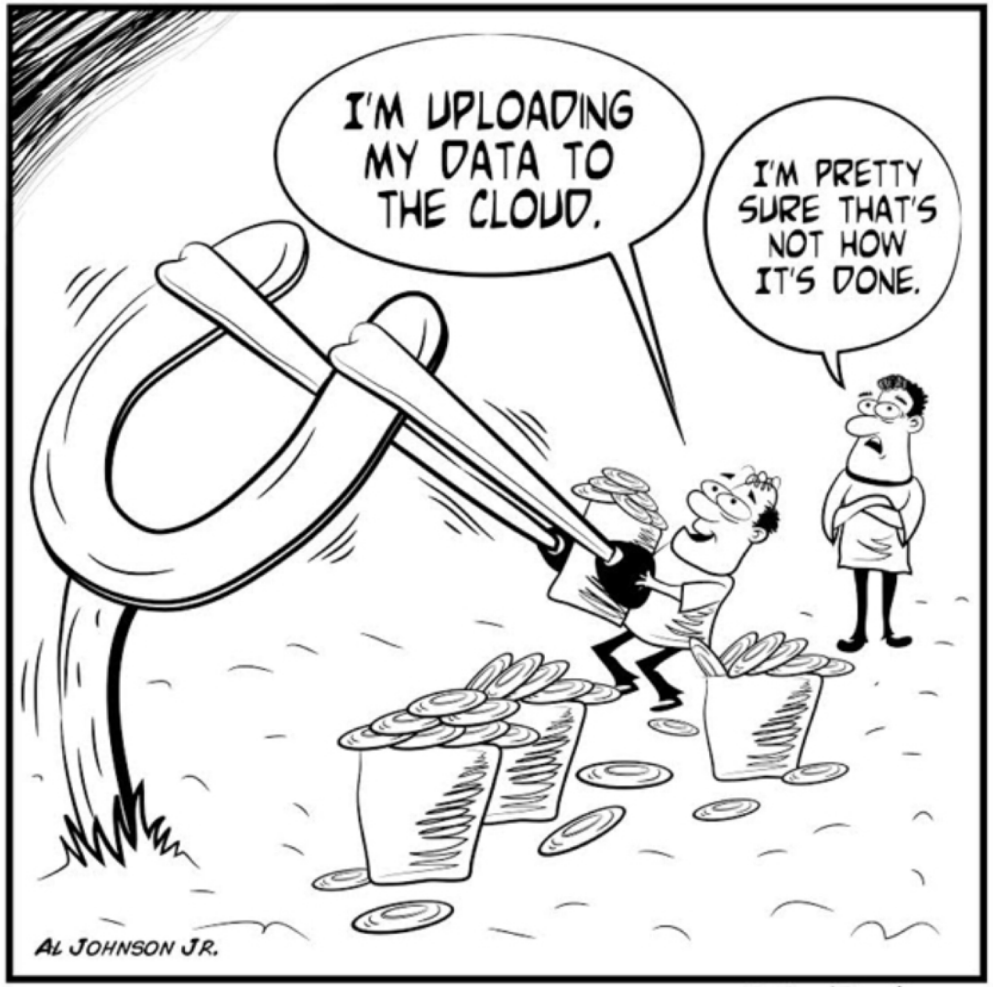
- 1000 Genomes Project: A detailed map of human genetic variation.
- TCGA on AWS: Raw and processed genomic, transcriptomic, and epigenomic data from The Cancer Genome Atlas (TCGA) available to qualified researchers via the Cancer Genomics Cloud.
- ICGC on AWS: Whole genome sequence data available to qualified researchers via The International Cancer Genome Consortium (ICGC).
- 3000 Rice Genome on AWS: Genome sequence of 3,024 rice varieties.
- Genome in a Bottle (GIAB): Several reference genomes to enable translation of whole human genome sequencing to clinical practice.

aws.amazon.com/public-datasets/

Google Public Datasets

- Reference Genomes: Reference Genomes such as GRCh37, GRCh37lite, GRCh38, hg19, hs37d5, and b37.
- Illumina Platinum Genomes: This dataset comprises the 17 member CEPH pedigree 1463.
- Personal Genome Project Data: This dataset comprises roughly 180 Complete Genomics genomes.
- ICGC-TCGA DREAM Mutation Calling Challenge synthetic genomes: This dataset comprises the three public synthetic tumor/normal pairs created for the ICGC-TCGA DREAM Mutation Calling challenge.
- Simons Genome Diversity Project: This dataset comprises 25 genomes from 13 diverse populations serving as the pilot project dataset
- TCGA Cancer Genomics Data in the Cloud: Open-access TCGA data including somatic mutation calls, clinical data, mRNA and miRNA expression, DNA methylation and protein expression from 33 different tumor types.

cloud.google.com/public-datasets/

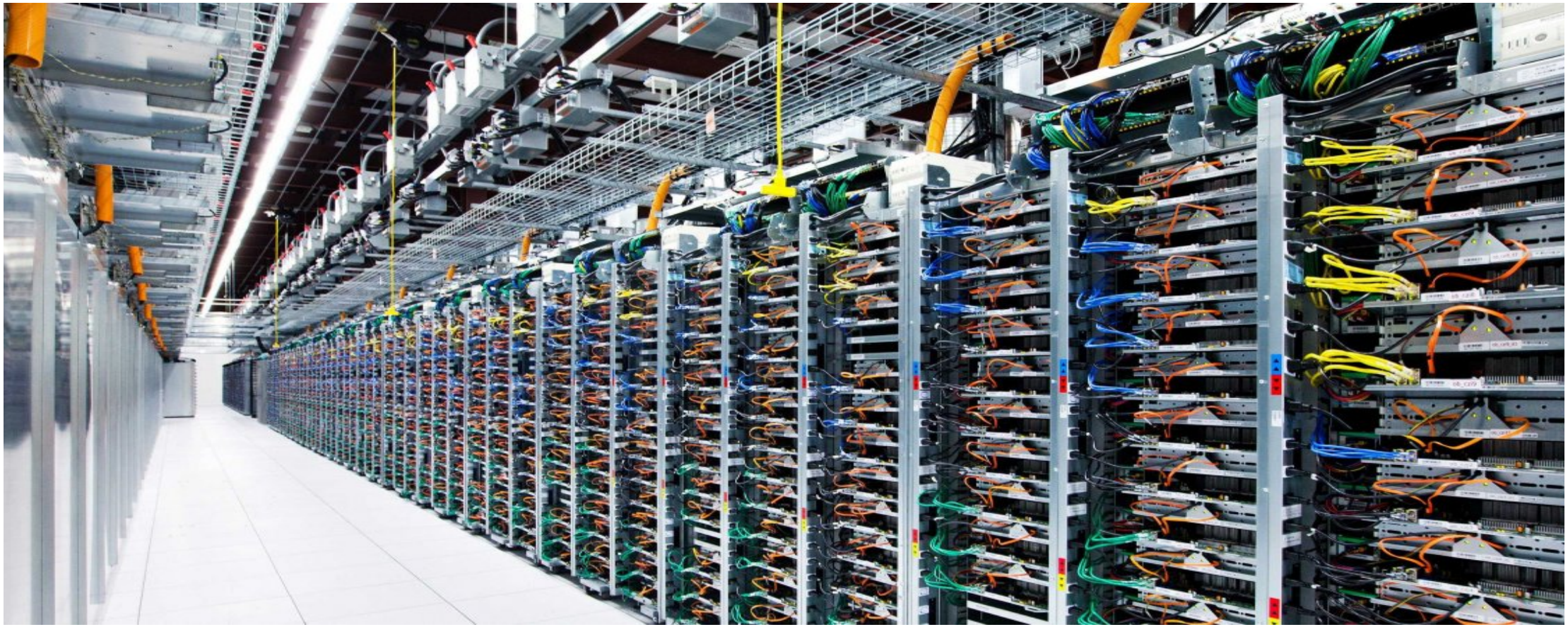


Into the future..

- Machine Learning, Deep learning, and AI
- Highly distributed, scalable systems
- Edge computing
- Highly interconnected networks
- Software defined architectures
- New programming languages

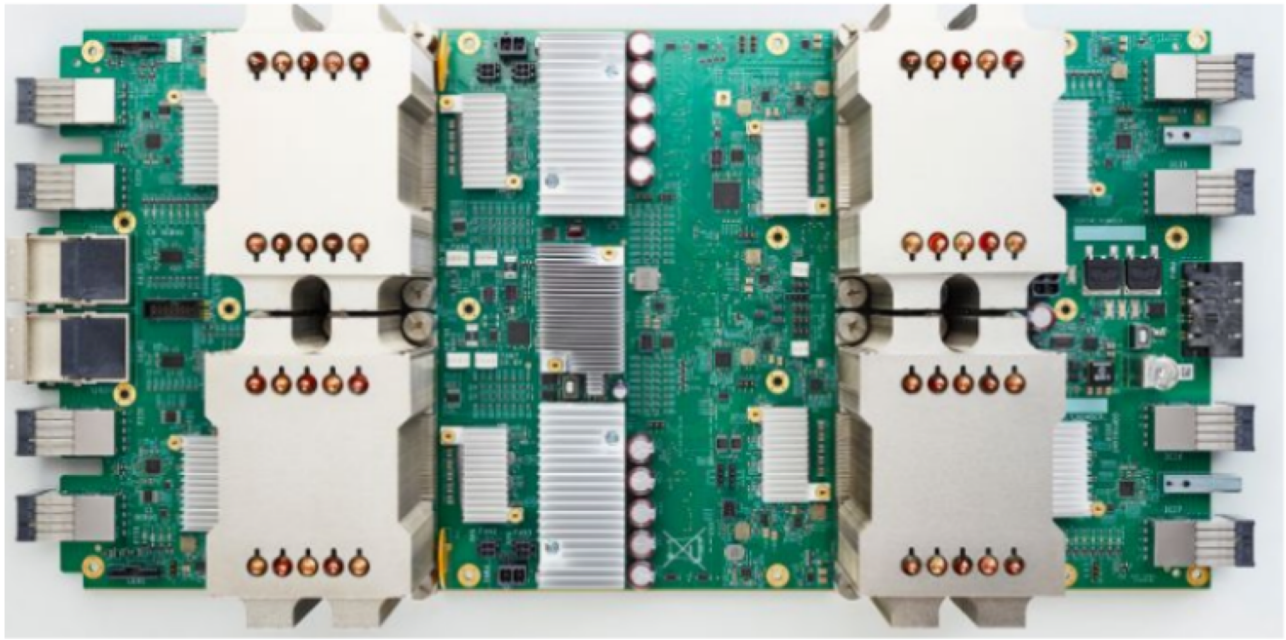
➔ Build a rich ecosystem for collaboration among researchers

Google Data Centers

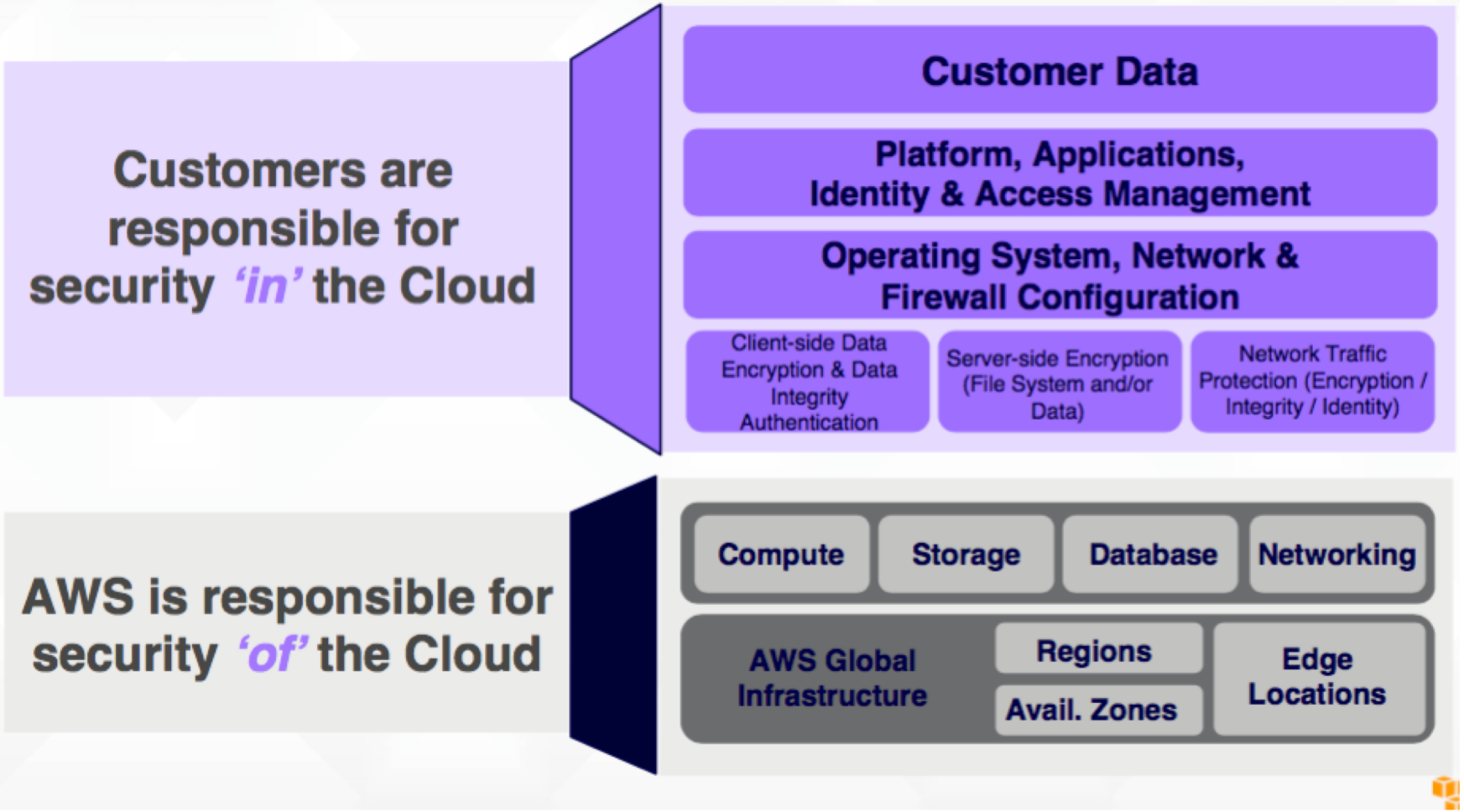


Google TPU

- 180 teraflops
- Designed for ML training and prediction



With AWS, Security Is a Shared Responsibility



Current JAX Campus Champion Allocations:

System	Allocation (SUs)
Gordon	50,000
OSG	200,000
Stampede	50,000
Maverick	3,000
SuperMIC	50,000
Comet	50,000
Jetstream	50,000
Bridges	2,500
Xstream	3,000

Thanks to Shane Sanders for providing all the XSEDE information.

XSEDE Allocations Overview

All allocations are for a period of 1 year.

- **Computational Resources**: XSEDE SPs offer a variety of high-performance computing (HPC) and high-throughput computing systems for allocation. Computing platforms include clusters, scalable-parallel systems, and shared-memory systems with various CPU, memory, communication, and storage configurations.
- **Visualization Resources**: SPs provide a variety of visualization resources and software services to the XSEDE user community. These systems provide a powerful way to interact with and analyze data at any scale.
- **Storage Resources**: Several XSEDE SPs host storage platforms providing services such as data management, data collections hosting, and large-scale persistent storage. Data cannot be controlled access or HIPPA or PHI data.

Example allocation requests: portal.xsede.org/allocations/research