Scientific Computing with Amazon Web Services

Jamie Kinney
Principal Solution Architect, Amazon Web Services
jkinney@amazon.com
@jamiekinney

What is AWS?

2003

amazon.com

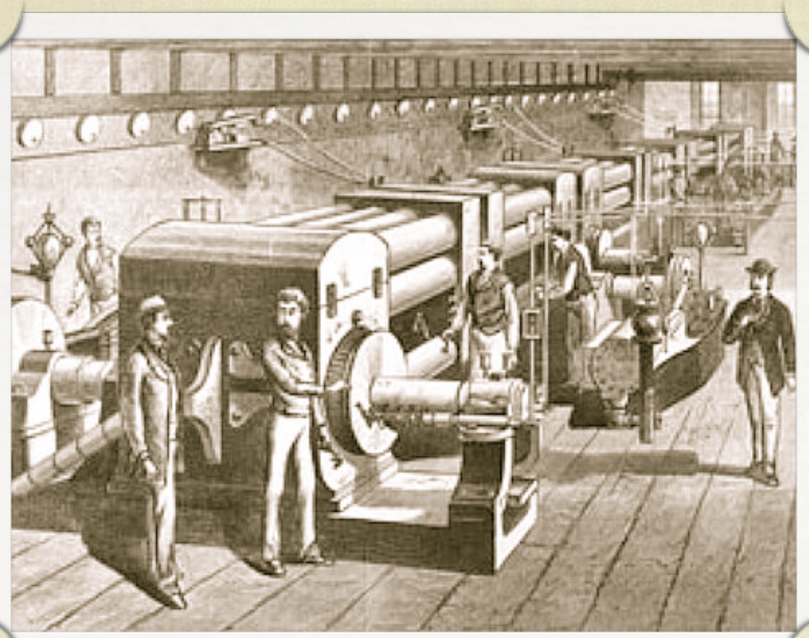
$5.2B retail business

7,800 employees

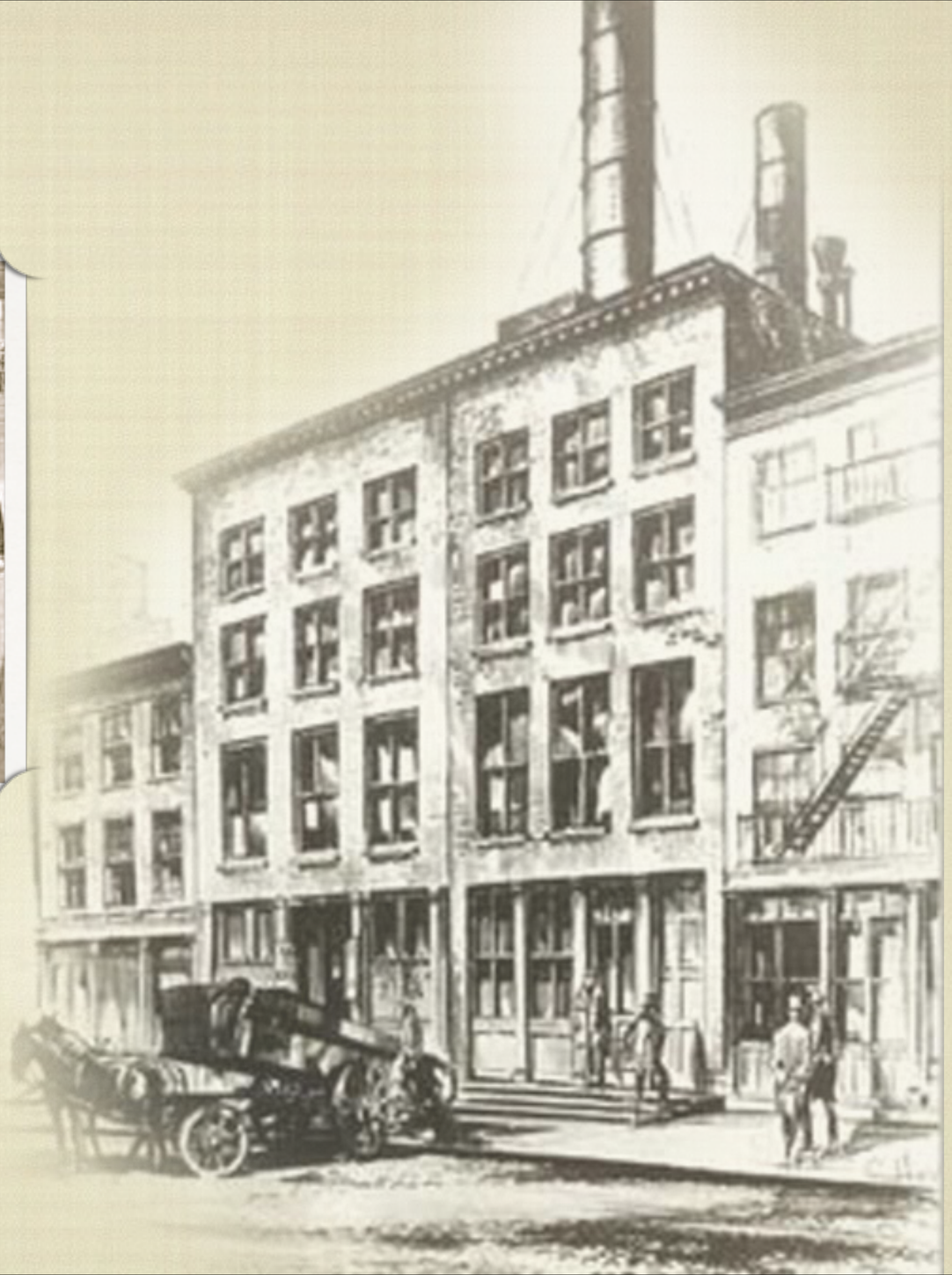
A whole lot of servers

2012

Every day, AWS adds enough server capacity to power this $5B enterprise



Pearl Street
Power Station



Your Applications

Web Interface

Management Console

Billing

Monitoring

CloudWatch

Deployment & Automation

BeanStalk

OpsWorks

Cloud Formation

Data Pipeline

Identity & Access

IAM

Federation

Hosted Desktops

Content Delivery

CloudFront

Applications

SES

SNS

SQS

CloudSearch

SWF

Parallel Computing

Kinesis

EMR

Libraries & SDK's

Compute

EC2

Storage

S3

EBS

Glacier

Storage Gateway

Networking

VPC

Direct Connect

ELB

Route53

Databases

RDS

Dynamo

ElastiCache

RedShift

Regions

Availability Zones

Edge Locations

AWS Global Infrastructure

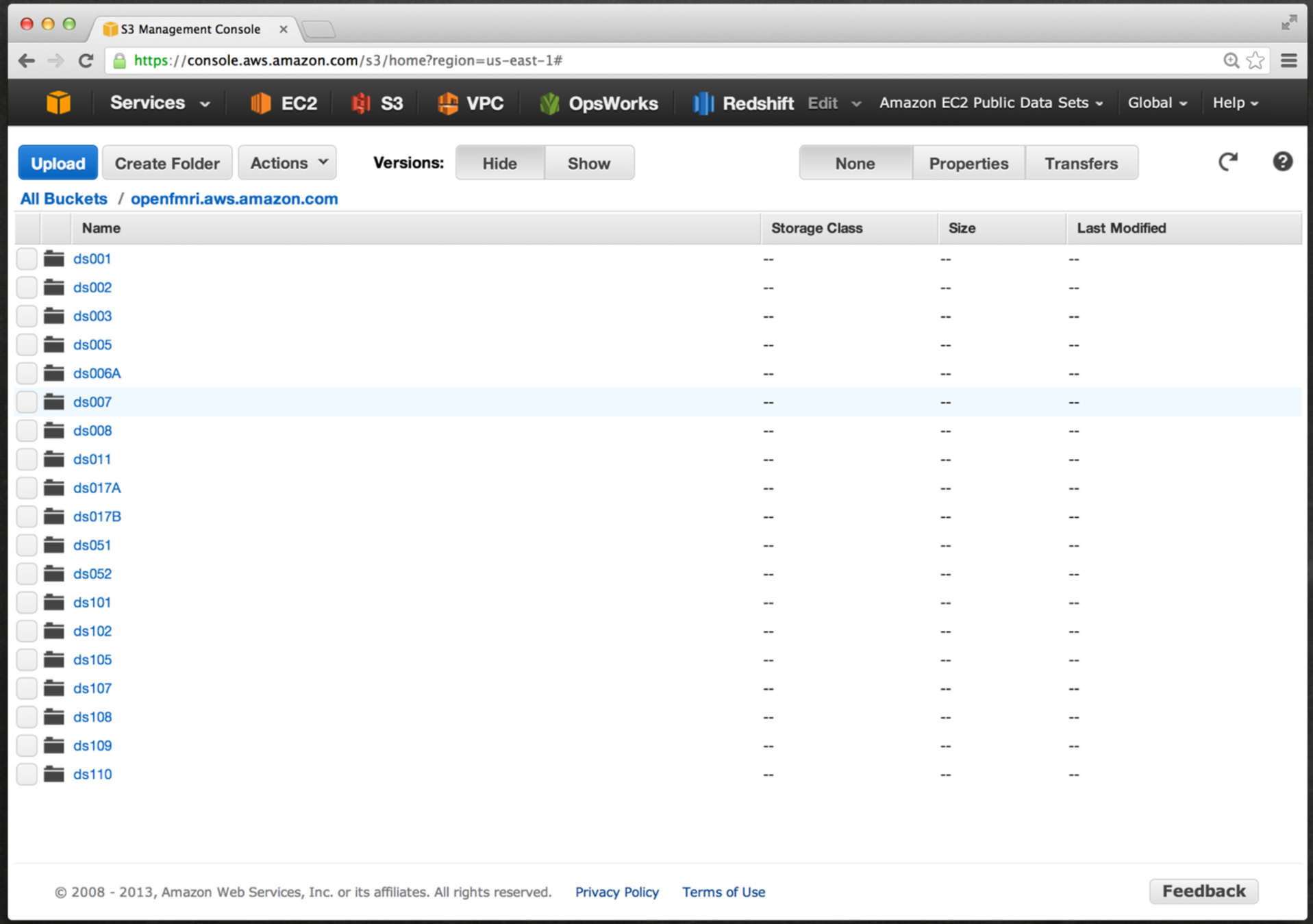
The AWS Big Data Stack

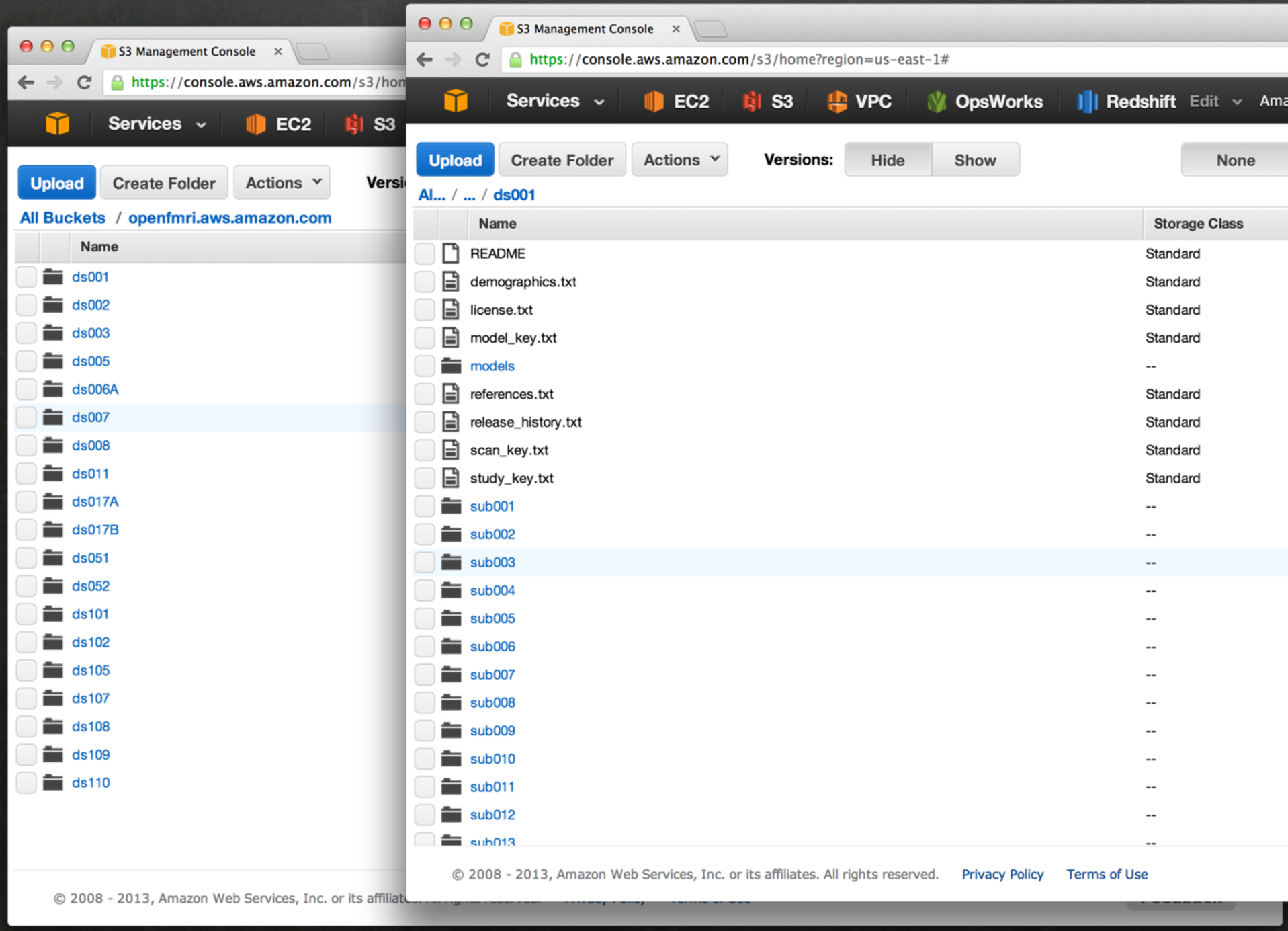
- Amazon S3 - Massively scalable, highly durable object store
- Amazon DynamoDB - SSD-backed NoSQL DB with geospatial APIs
- Amazon RedShift - Relational, columnar data warehouse
- Amazon Kinesis - Managed real-time streaming data processing
- Amazon Elastic MapReduce - Managed Hadoop
- Amazon RDS - Managed Relational DBs (Oracle, Postgres, MySQL, MS SQL)
- Amazon EC2 -
  - Spark/Shark
  - Cassandra
  - MongoDB
  - Many flavors of Hadoop
  - HTCondor
  - SLURM
  - StarCluster
  - ... and many more

Amazon S3

Amazon S3: Over 2 Trillion Total Objects

> 2,000,000 requests/second

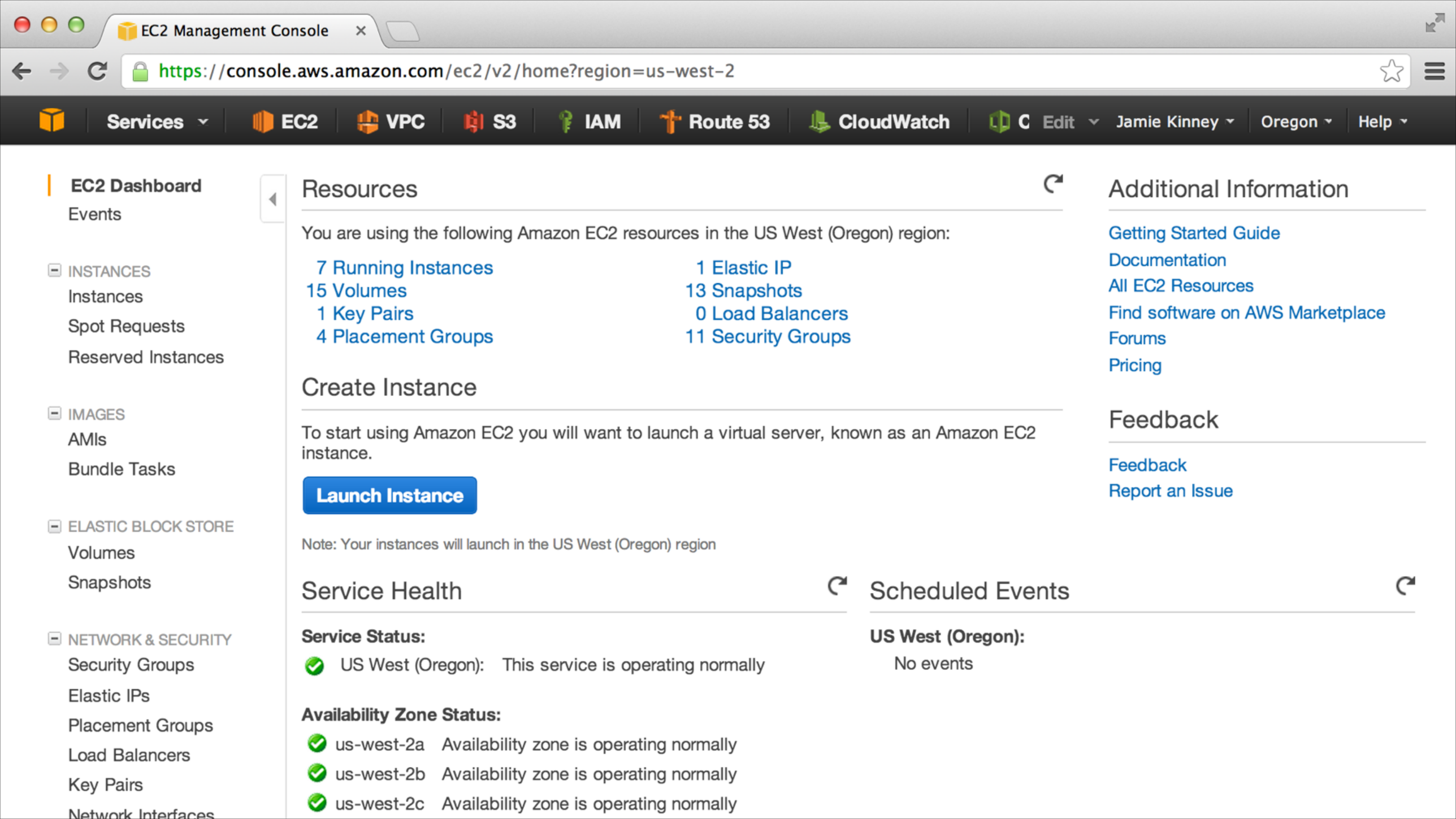

S3 Management Console
https://console.aws.amazon.com/s3/home?region=us-east-1#
Services
EC2
S3
VPC
OpsWorks
Redshift
Edit
Amazon EC2 Public Data Sets
Global
Help
Upload
Create Folder
Actions
Versions:
Hide
Show
None
Properties
Transfers
All Buckets / openfmri.aws.amazon.com
Name
Storage Class
Size
Last Modified
ds001 -- -- --
ds002 -- -- --
ds003 -- -- --
ds005 -- -- --
ds006A -- -- --
ds007 -- -- --
ds008 -- -- --
ds011 -- -- --
ds017A -- -- --
ds017B -- -- --
ds051 -- -- --
ds052 -- -- --
ds101 -- -- --
ds102 -- -- --
ds105 -- -- --
ds107 -- -- --
ds108 -- -- --
ds109 -- -- --
ds110 -- -- --
© 2008 - 2013, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Feedback


S3 Management Console
https://console.aws.amazon.com/s3/hom
Services
EC2
S3
Upload
Create Folder
Actions
All Buckets / openfmri.aws.amazon.com
Name
ds001
ds002
ds003
ds005
ds006A
ds007
ds008
ds011
ds017A
ds017B
ds051
ds052
ds101
ds102
ds105
ds107
ds108
ds109
ds110
© 2008 - 2013, Amazon Web Services, Inc. or its affiliates.
S3 Management Console
https://console.aws.amazon.com/s3/home?region=us-east-1#
Services
EC2
S3
VPC
OpsWorks
Redshift
Edit
Amazon EC2 Public Data Sets
Global
Help
Upload
Create Folder
Actions
Versions:
Hide
Show
None
Properties
Transfers
Al... / ... / ds001
Name
Storage Class
Size
Last Modified
README
Standard
1 KB
Wed May 01 15:56:56 GMT-700 2013
demographics.txt
Standard
289 bytes
Wed May 01 15:56:56 GMT-700 2013
license.txt
Standard
489 bytes
Wed May 01 15:56:56 GMT-700 2013
model_key.txt
Standard
15 bytes
Wed May 01 15:56:56 GMT-700 2013
models
--
--
--
references.txt
Standard
285 bytes
Wed May 01 15:56:56 GMT-700 2013
release_history.txt
Standard
28 bytes
Wed May 01 15:56:56 GMT-700 2013
scan_key.txt
Standard
7 bytes
Wed May 01 15:56:56 GMT-700 2013
study_key.txt
Standard
25 bytes
Wed May 01 15:56:56 GMT-700 2013
sub001
sub002
sub003
sub004
sub005
sub006
sub007
sub008
sub009
sub010
sub011
sub012
sub013
© 2008 - 2013, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Feedback

```
~> s3cmd ls s3://openfmri.aws.amazon.com/
                       DIR   s3://openfmri.aws.amazon.com/ds001/
                       DIR   s3://openfmri.aws.amazon.com/ds002/
                       DIR   s3://openfmri.aws.amazon.com/ds003/
                       DIR   s3://openfmri.aws.amazon.com/ds005/
                       DIR   s3://openfmri.aws.amazon.com/ds006A/
                       DIR   s3://openfmri.aws.amazon.com/ds007/
                       DIR   s3://openfmri.aws.amazon.com/ds008/
                       DIR   s3://openfmri.aws.amazon.com/ds011/
                       DIR   s3://openfmri.aws.amazon.com/ds017A/
                       DIR   s3://openfmri.aws.amazon.com/ds017B/
                       DIR   s3://openfmri.aws.amazon.com/ds051/
                       DIR   s3://openfmri.aws.amazon.com/ds052/
                       DIR   s3://openfmri.aws.amazon.com/ds101/
                       DIR   s3://openfmri.aws.amazon.com/ds102/
                       DIR   s3://openfmri.aws.amazon.com/ds105/
                       DIR   s3://openfmri.aws.amazon.com/ds107/
                       DIR   s3://openfmri.aws.amazon.com/ds108/
                       DIR   s3://openfmri.aws.amazon.com/ds109/
                       DIR   s3://openfmri.aws.amazon.com/ds110/
~> s3cmd ls s3://hcp.aws.amazon.com/q1/
                       DIR   s3://hcp.aws.amazon.com/q1/100307/
                       DIR   s3://hcp.aws.amazon.com/q1/103515/
                       DIR   s3://hcp.aws.amazon.com/q1/103818/
                       DIR   s3://hcp.aws.amazon.com/q1/111312/
                       DIR   s3://hcp.aws.amazon.com/q1/114924/
                       DIR   s3://hcp.aws.amazon.com/q1/117122/
                       DIR   s3://hcp.aws.amazon.com/q1/118932/
                       DIR   s3://hcp.aws.amazon.com/q1/119833/
                       DIR   s3://hcp.aws.amazon.com/q1/120212/
                       DIR   s3://hcp.aws.amazon.com/q1/125525/
                       DIR   s3://hcp.aws.amazon.com/q1/128632/
                       DIR   s3://hcp.aws.amazon.com/q1/130013/
                       DIR   s3://hcp.aws.amazon.com/q1/131621/
                       DIR   s3://hcp.aws.amazon.com/q1/137128/
                       DIR   s3://hcp.aws.amazon.com/q1/138231/
                       DIR   s3://hcp.aws.amazon.com/q1/142828/
                       DIR   s3://hcp.aws.amazon.com/q1/143325/
                       DIR   s3://hcp.aws.amazon.com/q1/144226/
                       DIR   s3://hcp.aws.amazon.com/q1/149337/
                       DIR   s3://hcp.aws.amazon.com/q1/150423/
                       DIR   s3://hcp.aws.amazon.com/q1/153429/
                       DIR   s3://hcp.aws.amazon.com/q1/156637/
                       DIR   s3://hcp.aws.amazon.com/q1/159239/
                       DIR   s3://hcp.aws.amazon.com/q1/161731/
```

Amazon EC2

ec2-run-instances


EC2 Management Console
https://console.aws.amazon.com/ec2/v2/home?region=us-west-2
Services
EC2
VPC
S3
IAM
Route 53
CloudWatch
Edit
Jamie Kinney
Oregon
Help
EC2 Dashboard
Events
INSTANCES
Instances
Spot Requests
Reserved Instances
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Load Balancers
Key Pairs
Network Interfaces
Resources
You are using the following Amazon EC2 resources in the US West (Oregon) region:
7 Running Instances
15 Volumes
1 Key Pairs
4 Placement Groups
1 Elastic IP
13 Snapshots
0 Load Balancers
11 Security Groups
Create Instance
To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.
Launch Instance
Note: Your instances will launch in the US West (Oregon) region
Service Health
Service Status:
US West (Oregon): This service is operating normally
Availability Zone Status:
us-west-2a Availability zone is operating normally
us-west-2b Availability zone is operating normally
us-west-2c Availability zone is operating normally
Scheduled Events
US West (Oregon):
No events
Additional Information
Getting Started Guide
Documentation
All EC2 Resources
Find software on AWS Marketplace
Forums
Pricing
Feedback
Feedback
Report an Issue

Programmable

```ruby
  # generate a key pair
  key_pair = ec2.key_pairs.create("ruby-sample-#{Time.now.to_i}")
  puts "Generated keypair #{key_pair.name}, fingerprint: #{key_pair.fingerprint}"

  # open SSH access
  group = ec2.security_groups.create("ruby-sample-#{Time.now.to_i}")
  group.authorize_ingress(:tcp, 22, "0.0.0.0/0")
  puts "Using security group: #{group.name}"

  # launch the instance
  instance = image.run_instance(:key_pair => key_pair,
                                :security_groups => group)
  sleep 1 until instance.status != :pending
  puts "Launched instance #{instance.id}, status: #{instance.status}"

  exit 1 unless instance.status == :running

  begin
    Net::SSH.start(instance.ip_address, "ec2-user",
                   :key_data => [key_pair.private_key]) do |ssh|
      puts "Running 'uname -a' on the instance yields:"
      puts ssh.exec!("uname -a")
    end
  rescue SystemCallError, Timeout::Error => e
    # port 22 might not be available immediately after the instance finishes launching
    sleep 1
    retry
  end

ensure
  # clean up
  [instance,
   group,
   key_pair].compact.each(&:delete)
end
```

```ruby
key_pair = ec2.key_pairs.create("ruby-sample-#{Time.now.to_i}")
puts "Generated keypair #{key_pair.name}, fingerprint: #{key_pair.fingerprint}"

# open SSH access
group = ec2.security_groups.create("ruby-sample-#{Time.now.to_i}")
group.authorize_ingress(:tcp, 22, "0.0.0.0/0")
puts "Using security group: #{group.name}"

# launch the instance
instance = image.run_instance(:key_pair => key_pair,
```

```ruby
# launch the instance
instance = image.run_instance(:key_pair => key_pair,
                              :security_groups => group)
sleep 1 until instance.status != :pending
puts "Launched instance #{instance.id}, status: #{instance.status}"

exit 1 unless instance.status == :running
```

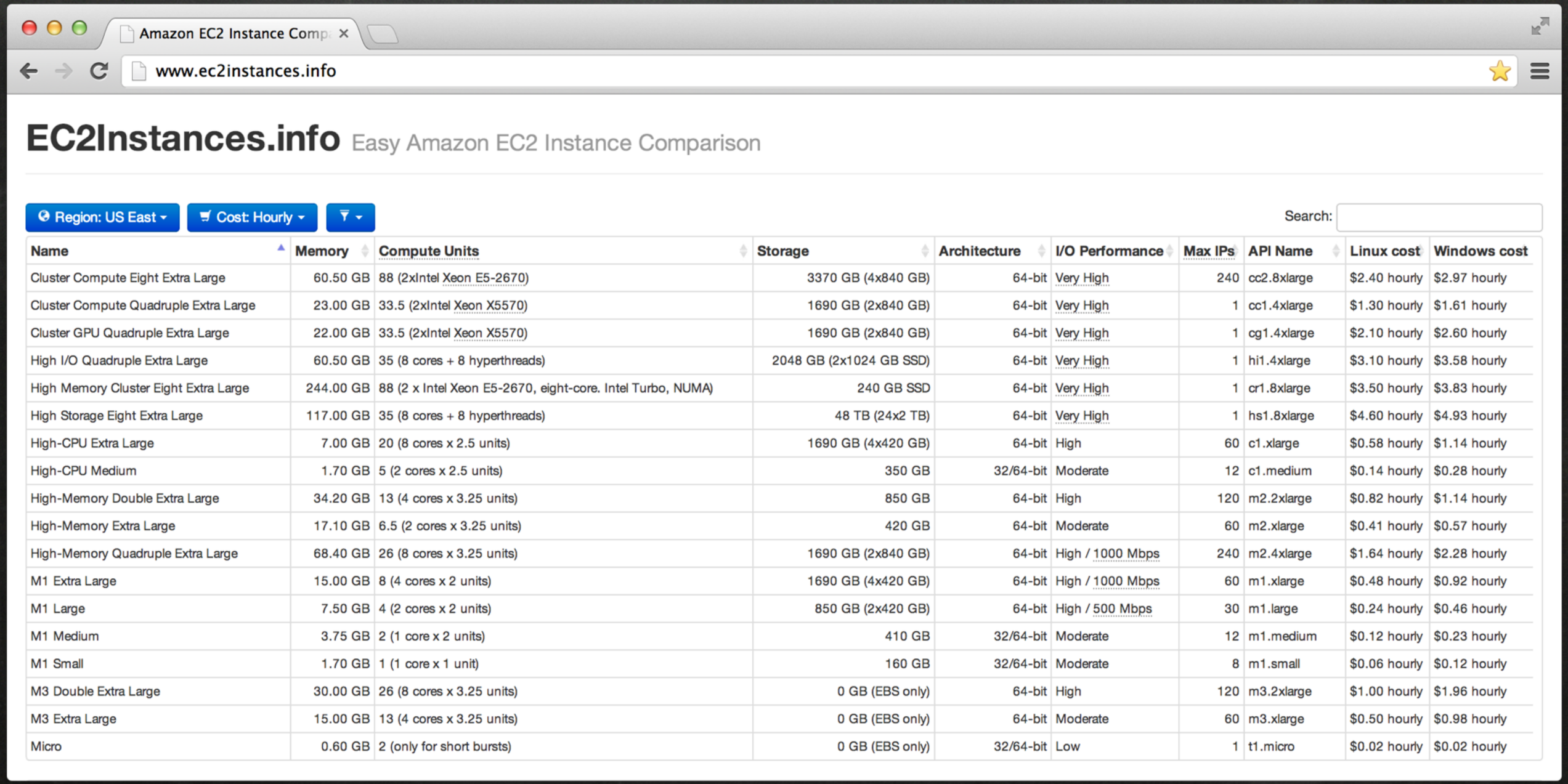
```ruby
    # port 22 might not be available immediately after the instance finishes launching
    sleep 1
    retry
  end

ensure
  # clean up
  [instance,
   group,
   key_pair].compact.each(&:delete)
end
```

Instance Types

Standard (m1)
High Memory (m2,m3)
High CPU (c1)

Cluster Compute

Intel Nehalem (cc1.4xlarge)
Nvidia GPUs (cg1.4xlarge and g2.2xlarge)
2TB of SSD 120,000 IOPS (hi1.4xlarge)
Intel Sandy Bridge E5-2670 (cc2.8xlarge)
Sandy Bridge, NUMA, 240GB RAM (cr1.4xlarge)
48 TB of ephemeral storage (hs1.8xlarge)
Intel Ivy Bridge E5-2680 (c3.2/4/8xl)

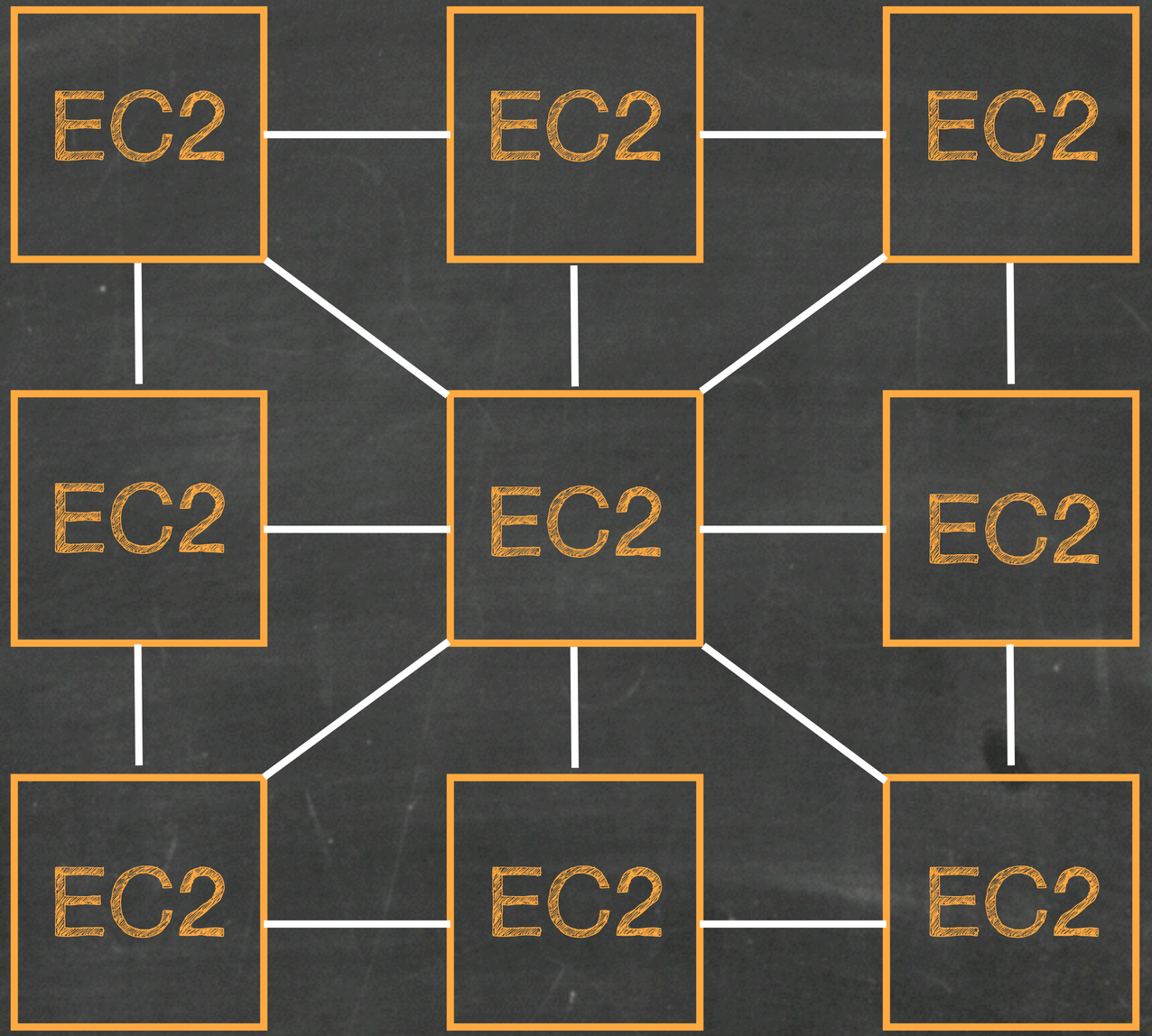


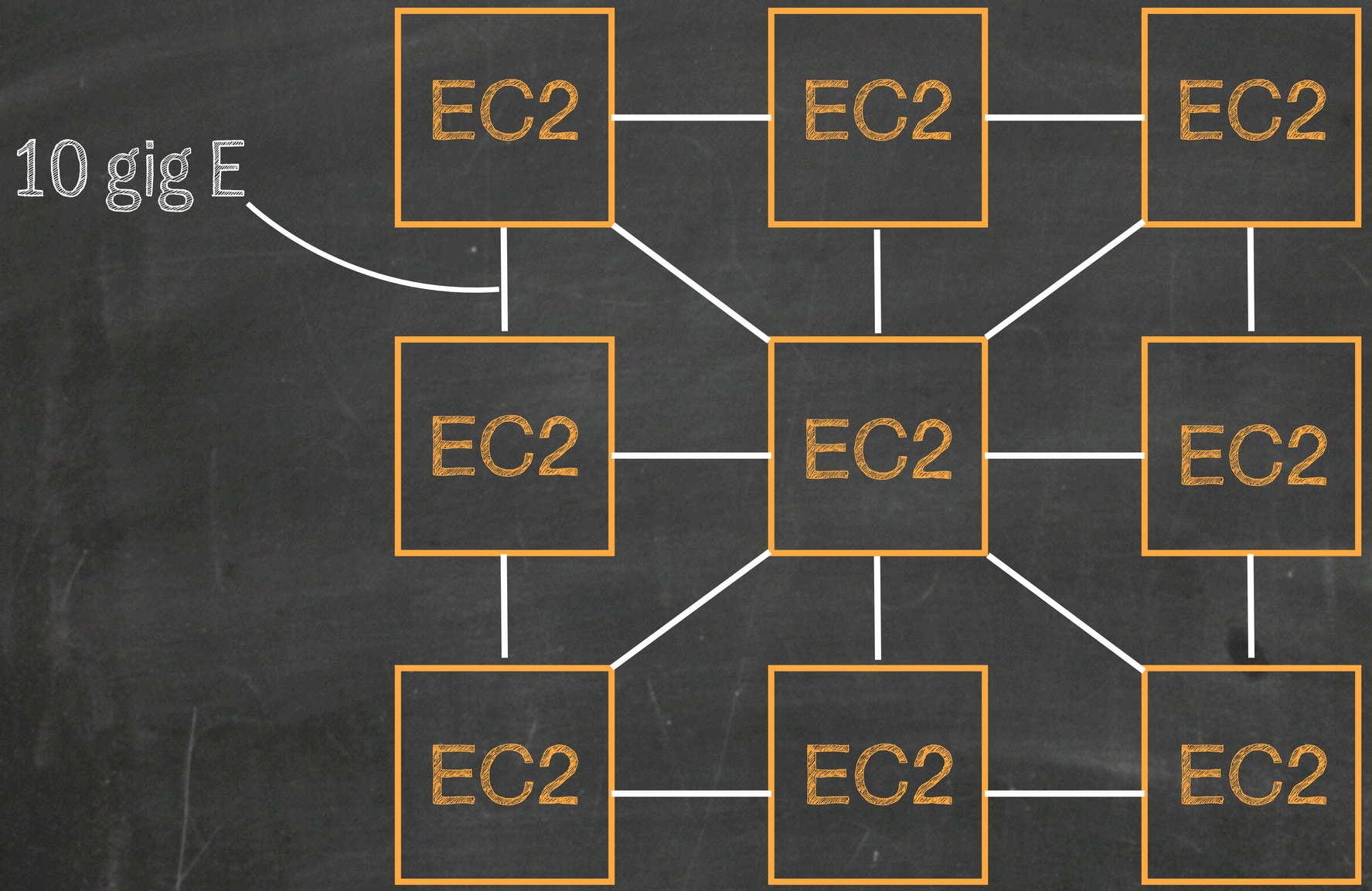
EC2Instances.info Easy Amazon EC2 Instance Comparison

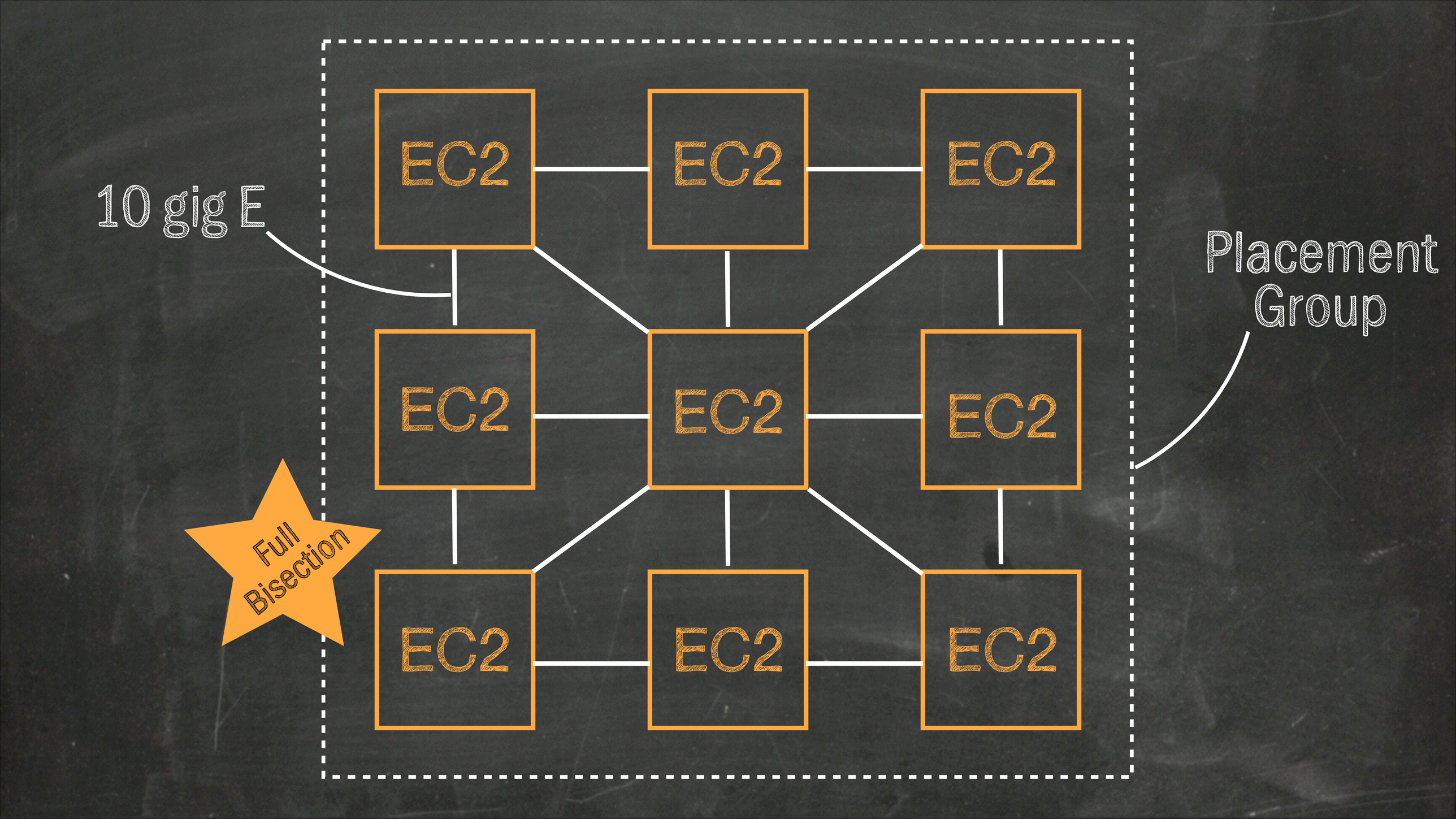
Region: US East | Cost: Hourly | Search:

Name	Memory	Compute Units	Storage	Architecture	I/O Performance	Max IPs	API Name	Linux cost	Windows cost
Cluster Compute Eight Extra Large	60.50 GB	88 (2xIntel Xeon E5-2670)	3370 GB (4x840 GB)	64-bit	Very High	240	cc2.8xlarge	$2.40 hourly	$2.97 hourly
Cluster Compute Quadruple Extra Large	23.00 GB	33.5 (2xIntel Xeon X5570)	1690 GB (2x840 GB)	64-bit	Very High	1	cc1.4xlarge	$1.30 hourly	$1.61 hourly
Cluster GPU Quadruple Extra Large	22.00 GB	33.5 (2xIntel Xeon X5570)	1690 GB (2x840 GB)	64-bit	Very High	1	cg1.4xlarge	$2.10 hourly	$2.60 hourly
High I/O Quadruple Extra Large	60.50 GB	35 (8 cores + 8 hyperthreads)	2048 GB (2x1024 GB SSD)	64-bit	Very High	1	hi1.4xlarge	$3.10 hourly	$3.58 hourly
High Memory Cluster Eight Extra Large	244.00 GB	88 (2 x Intel Xeon E5-2670, eight-core. Intel Turbo, NUMA)	240 GB SSD	64-bit	Very High	1	cr1.8xlarge	$3.50 hourly	$3.83 hourly
High Storage Eight Extra Large	117.00 GB	35 (8 cores + 8 hyperthreads)	48 TB (24x2 TB)	64-bit	Very High	1	hs1.8xlarge	$4.60 hourly	$4.93 hourly
High-CPU Extra Large	7.00 GB	20 (8 cores x 2.5 units)	1690 GB (4x420 GB)	64-bit	High	60	c1.xlarge	$0.58 hourly	$1.14 hourly
High-CPU Medium	1.70 GB	5 (2 cores x 2.5 units)	350 GB	32/64-bit	Moderate	12	c1.medium	$0.14 hourly	$0.28 hourly
High-Memory Double Extra Large	34.20 GB	13 (4 cores x 3.25 units)	850 GB	64-bit	High	120	m2.2xlarge	$0.82 hourly	$1.14 hourly
High-Memory Extra Large	17.10 GB	6.5 (2 cores x 3.25 units)	420 GB	64-bit	Moderate	60	m2.xlarge	$0.41 hourly	$0.57 hourly
High-Memory Quadruple Extra Large	68.40 GB	26 (8 cores x 3.25 units)	1690 GB (2x840 GB)	64-bit	High / 1000 Mbps	240	m2.4xlarge	$1.64 hourly	$2.28 hourly
M1 Extra Large	15.00 GB	8 (4 cores x 2 units)	1690 GB (4x420 GB)	64-bit	High / 1000 Mbps	60	m1.xlarge	$0.48 hourly	$0.92 hourly
M1 Large	7.50 GB	4 (2 cores x 2 units)	850 GB (2x420 GB)	64-bit	High / 500 Mbps	30	m1.large	$0.24 hourly	$0.46 hourly
M1 Medium	3.75 GB	2 (1 core x 2 units)	410 GB	32/64-bit	Moderate	12	m1.medium	$0.12 hourly	$0.23 hourly
M1 Small	1.70 GB	1 (1 core x 1 unit)	160 GB	32/64-bit	Moderate	8	m1.small	$0.06 hourly	$0.12 hourly
M3 Double Extra Large	30.00 GB	26 (8 cores x 3.25 units)	0 GB (EBS only)	64-bit	High	120	m3.2xlarge	$1.00 hourly	$1.96 hourly
M3 Extra Large	15.00 GB	13 (4 cores x 3.25 units)	0 GB (EBS only)	64-bit	Moderate	60	m3.xlarge	$0.50 hourly	$0.98 hourly
Micro	0.60 GB	2 (only for short bursts)	0 GB (EBS only)	32/64-bit	Low	1	t1.micro	$0.02 hourly	$0.02 hourly

Placement Groups


EC2
EC2
EC2
EC2
EC2
EC2
EC2
EC2
EC2


10 gig E
EC2
EC2
EC2
EC2
EC2
EC2
EC2
EC2
EC2


10 gig E
EC2
EC2
EC2
EC2
EC2
EC2
EC2
EC2
EC2
Full Bisection
Placement Group

On-Demand Pricing

Region: US East (N. Virginia)

	Linux/UNIX Usage	Windows Usage
Standard On-Demand Instances		
Small (Default)	$0.060 per Hour	$0.091 per Hour
Medium	$0.120 per Hour	$0.182 per Hour
Large	$0.240 per Hour	$0.364 per Hour
Extra Large	$0.480 per Hour	$0.728 per Hour
Second Generation Standard On-Demand Instances		
Extra Large	$0.500 per Hour	$0.780 per Hour

Reserved Instances

Linux Windows

Heavy Utilization Reserved Instances

Region: US East (N. Virginia)

Linux Windows

Medium Utilization Reserved Instances

Region: US East (N. Virginia)

Linux Windows

Light Utilization Reserved Instances

Region: US East (N. Virginia)

	1 yr Term		3 yr Term	
	Upfront	**Hourly**	**Upfront**	**Hourly**
Standard Reserved Instances				
Small (Default)	$61	$0.034 per Hour	$96	$0.027 per Hour
Medium	$122	$0.068 per Hour	$192	$0.054 per Hour
Large	$243	$0.136 per Hour	$384	$0.108 per Hour
Extra Large	$486	$0.271 per Hour	$768	$0.215 per Hour
Second Generation Standard Reserved Instances				

Spot Instances

Region: US East (N. Virginia)

	Linux/UNIX Usage	Windows Usage
Standard Spot Instances		
Small (Default)	$0.007 per Hour	$0.017 per Hour
Medium	$0.013 per Hour	$0.033 per Hour
Large	$0.026 per Hour	$0.066 per Hour
Extra Large	$0.052 per Hour	$0.132 per Hour
Second Generation Standard Spot Instances		
Extra Large	$0.0575 per Hour	$0.1375 per Hour

- Bid $X per hour
- If current price <= bid, instance starts
- If current price > bid, instance terminates
- Customers pay market rate, not bid

Scientific Computing Use Cases

Scientific Computing Use Cases

- Science-as-a-Service

Scientific Computing Use Cases

- Science-as-a-Service
- Large-scale HTC (100,000+ core clusters)

Scientific Computing Use Cases

- Science-as-a-Service
- Large-scale HTC (100,000+ core clusters)
- Large-scale MapReduce (Hadoop/Spark/Shark) using EC2 or EMR

Scientific Computing Use Cases

- Science-as-a-Service
- Large-scale HTC (100,000+ core clusters)
- Large-scale MapReduce (Hadoop/Spark/Shark) using EC2 or EMR
- Small to medium-scale clusters (hundreds of nodes) for traditional MPI workloads

Scientific Computing Use Cases

- Science-as-a-Service
- Large-scale HTC (100,000+ core clusters)
- Large-scale MapReduce (Hadoop/Spark/Shark) using EC2 or EMR
- Small to medium-scale clusters (hundreds of nodes) for traditional MPI workloads
- Many small MPI clusters working in parallel to explore parameter space

Scientific Computing Use Cases

- Science-as-a-Service
- Large-scale HTC (100,000+ core clusters)
- Large-scale MapReduce (Hadoop/Spark/Shark) using EC2 or EMR
- Small to medium-scale clusters (hundreds of nodes) for traditional MPI workloads
- Many small MPI clusters working in parallel to explore parameter space
- Small to medium scale GPGPU workloads

Scientific Computing Use Cases

- Science-as-a-Service
- Large-scale HTC (100,000+ core clusters)
- Large-scale MapReduce (Hadoop/Spark/Shark) using EC2 or EMR
- Small to medium-scale clusters (hundreds of nodes) for traditional MPI workloads
- Many small MPI clusters working in parallel to explore parameter space
- Small to medium scale GPGPU workloads
- Dev/test of MPI workloads prior to submitting to supercomputing centers

Scientific Computing Use Cases

- Science-as-a-Service
- Large-scale HTC (100,000+ core clusters)
- Large-scale MapReduce (Hadoop/Spark/Shark) using EC2 or EMR
- Small to medium-scale clusters (hundreds of nodes) for traditional MPI workloads
- Many small MPI clusters working in parallel to explore parameter space
- Small to medium scale GPGPU workloads
- Dev/test of MPI workloads prior to submitting to supercomputing centers
- Ephemeral clusters, custom tailored to the task at hand, created for various stages of a pipeline

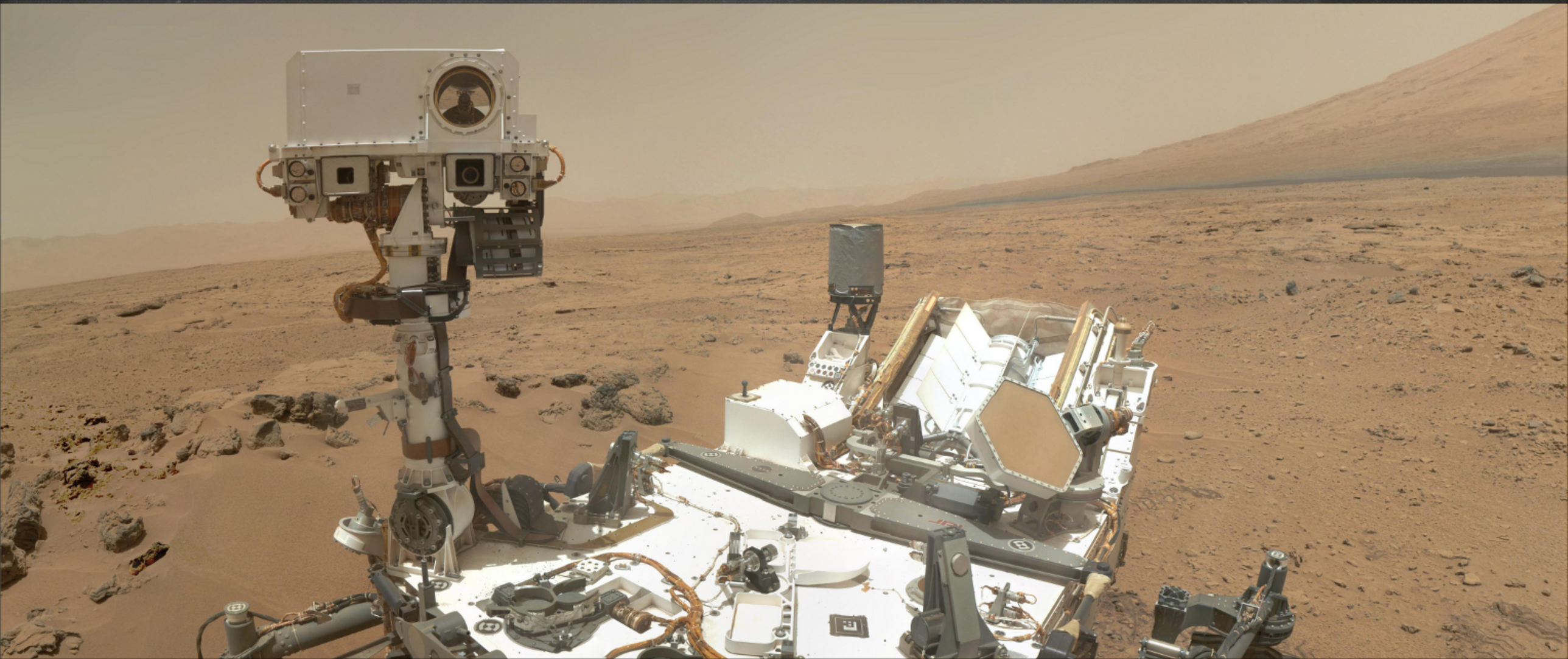
Scientific Computing Use Cases

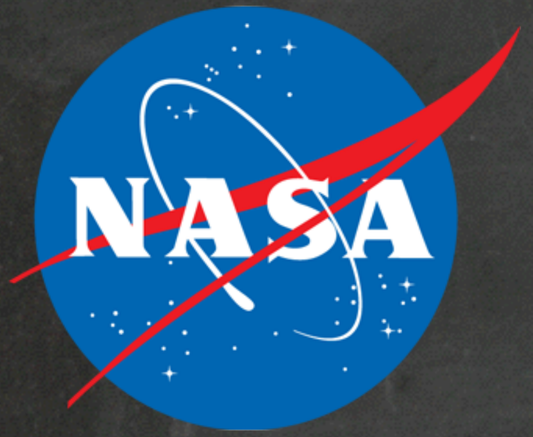
- Science-as-a-Service
- Large-scale HTC (100,000+ core clusters)
- Large-scale MapReduce (Hadoop/Spark/Shark) using EC2 or EMR
- Small to medium-scale clusters (hundreds of nodes) for traditional MPI workloads
- Many small MPI clusters working in parallel to explore parameter space
- Small to medium scale GPGPU workloads
- Dev/test of MPI workloads prior to submitting to supercomputing centers
- Ephemeral clusters, custom tailored to the task at hand, created for various stages of a pipeline
- Collaborative research environments

Scientific Computing Use Cases

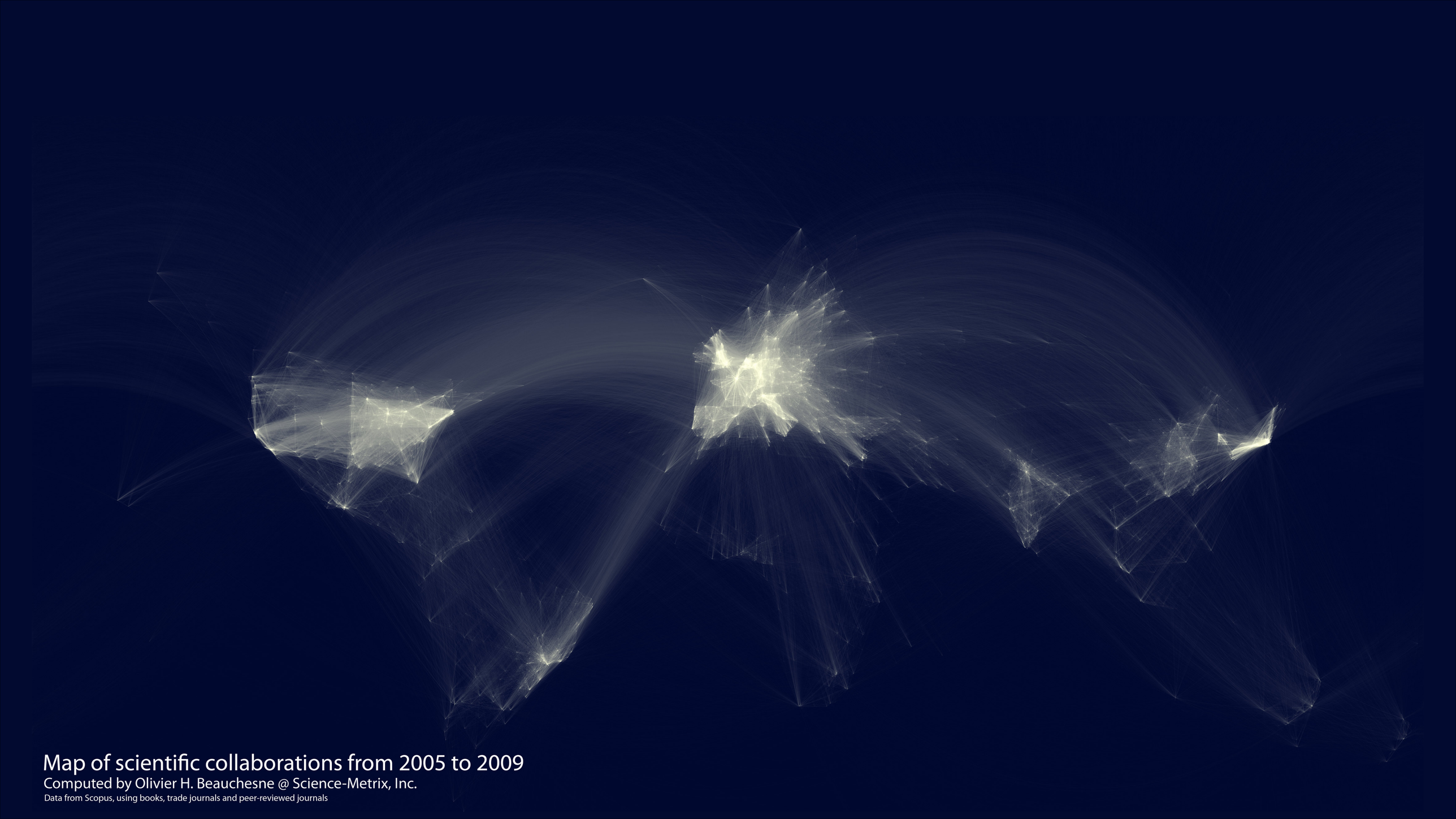
- Science-as-a-Service
- Large-scale HTC (100,000+ core clusters)
- Large-scale MapReduce (Hadoop/Spark/Shark) using EC2 or EMR
- Small to medium-scale clusters (hundreds of nodes) for traditional MPI workloads
- Many small MPI clusters working in parallel to explore parameter space
- Small to medium scale GPGPU workloads
- Dev/test of MPI workloads prior to submitting to supercomputing centers
- Ephemeral clusters, custom tailored to the task at hand, created for various stages of a pipeline
- Collaborative research environments
- On-demand academic training/lab environments

Who is Using AWS?

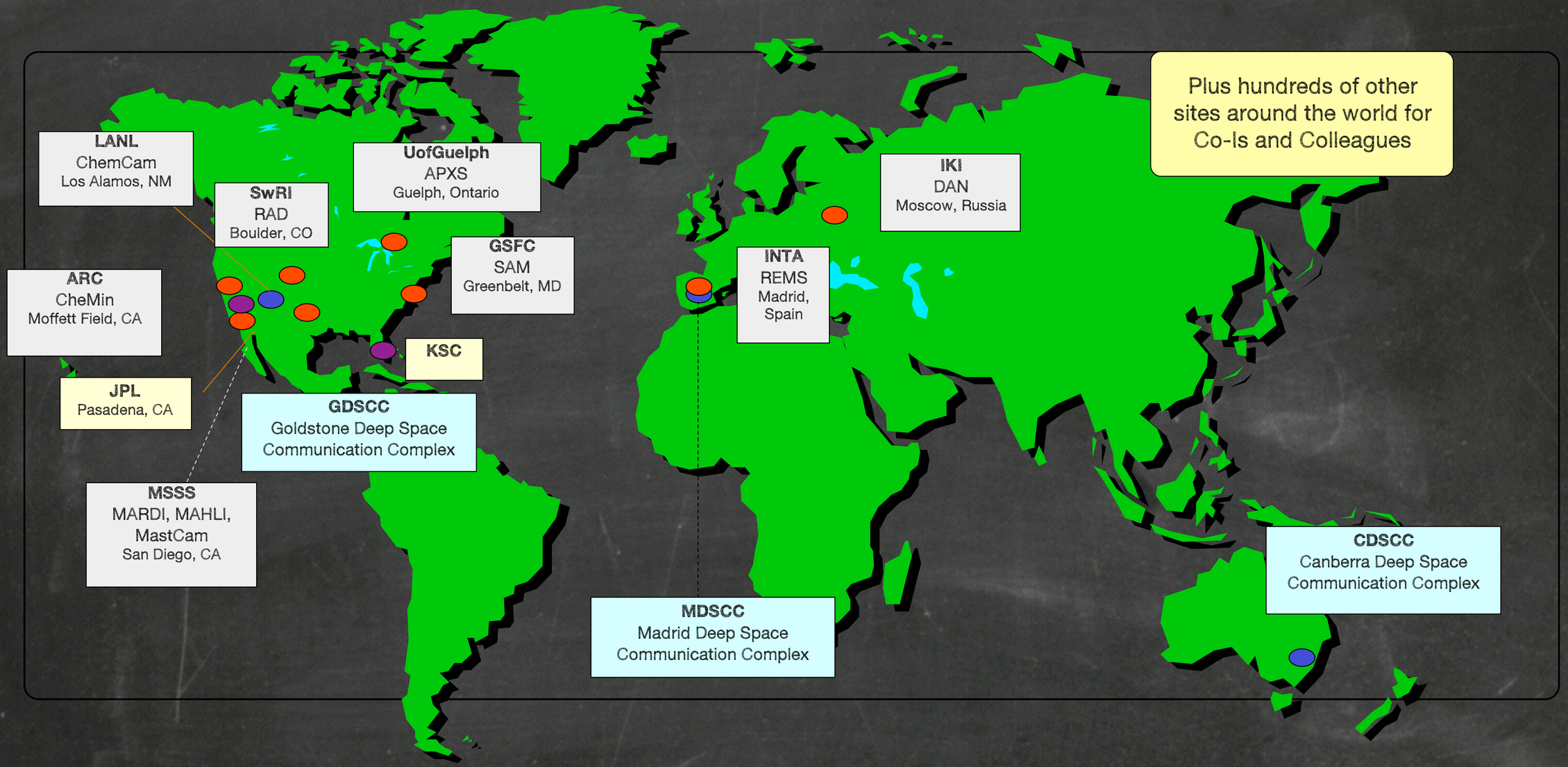




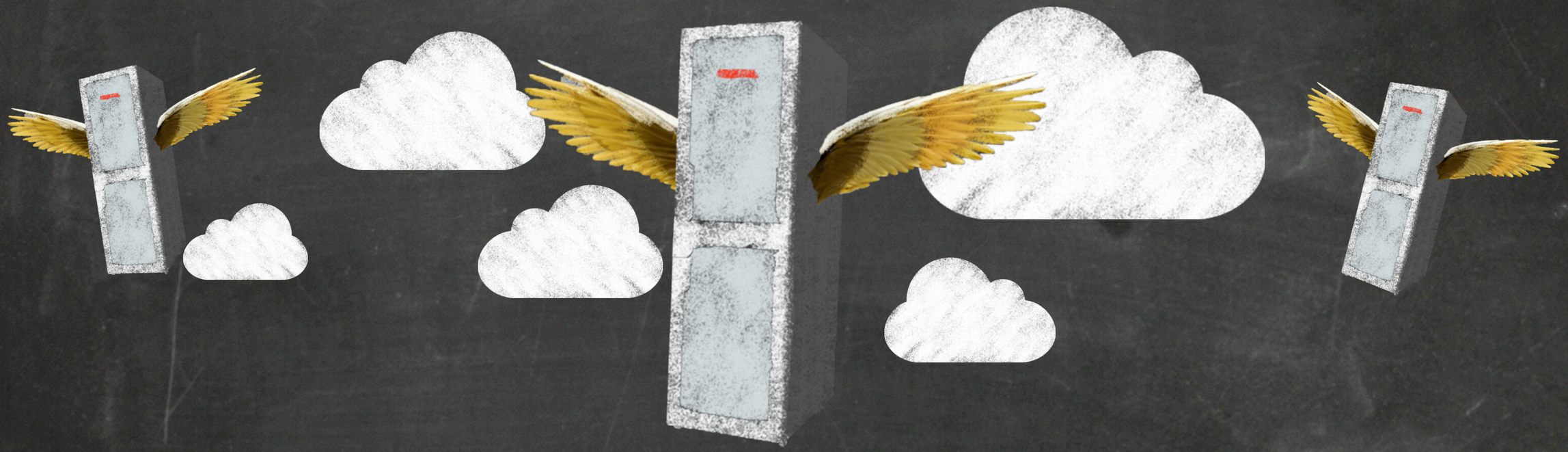
Jet Propulsion Laboratory
California Institute of Technology


Map of scientific collaborations from 2005 to 2009
Computed by Olivier H. Beauchesne @ Science-Metrix, Inc.
Data from Scopus, using books, trade journals and peer-reviewed journals

MSL Distributed Operations



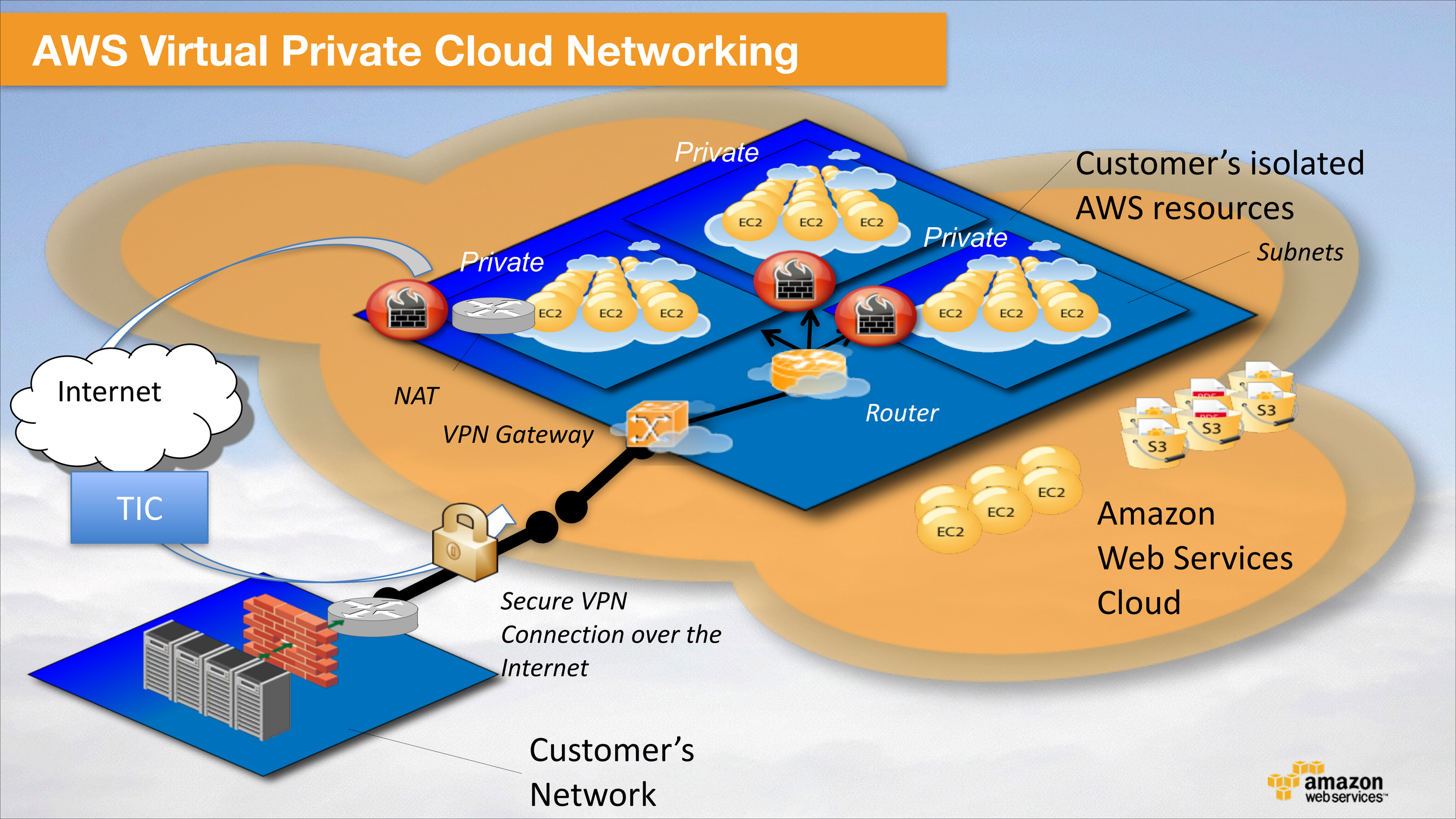
Integrated Architectures



AWS Direct Connect

Amazon VPC

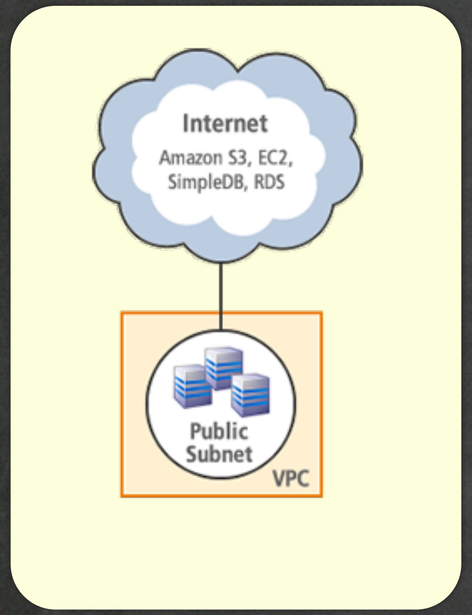
AWS Virtual Private Cloud Networking



VPC Deployment Models

VPC Deployment Models

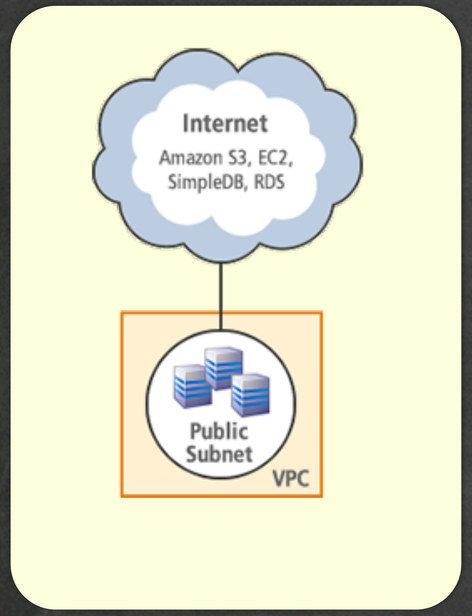
VPC With a Single Subnet



Simple web hosting, collaborative research environments, on-demand HPC/HTC clusters

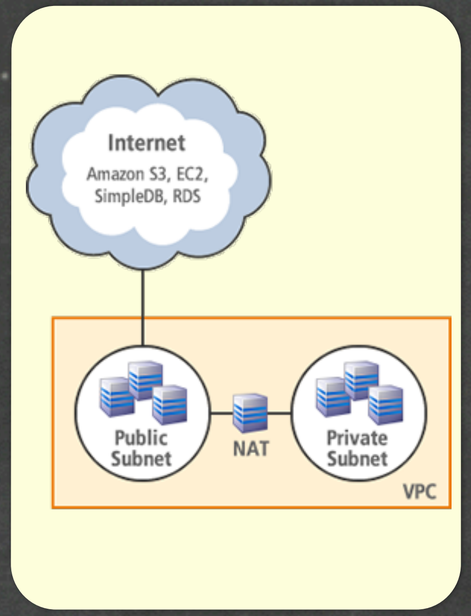
VPC Deployment Models

VPC With a Single Subnet



Simple web hosting, collaborative research environments, on-demand HPC/HTC clusters

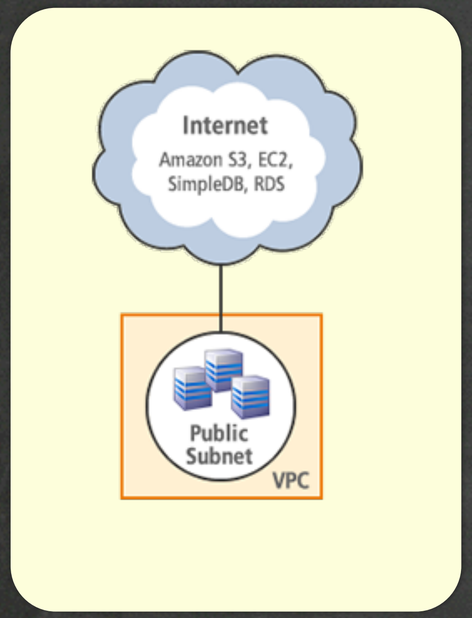
VPC With Private and Public Subnets



Multi-tier web hosting

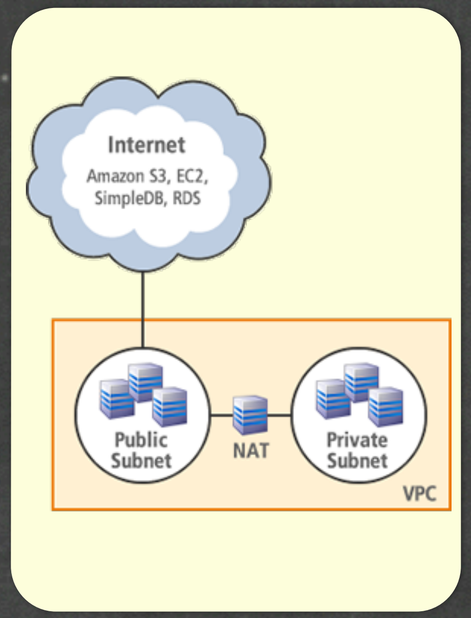
VPC Deployment Models

VPC With a Single Subnet



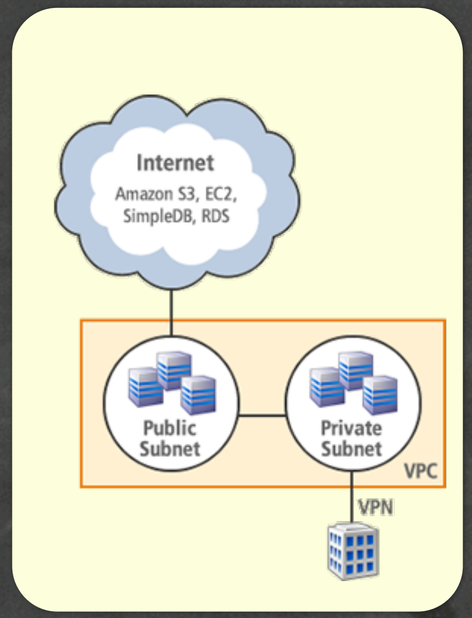
Simple web hosting, collaborative research environments, on-demand HPC/HTC clusters

VPC With Private and Public Subnets



Multi-tier web hosting

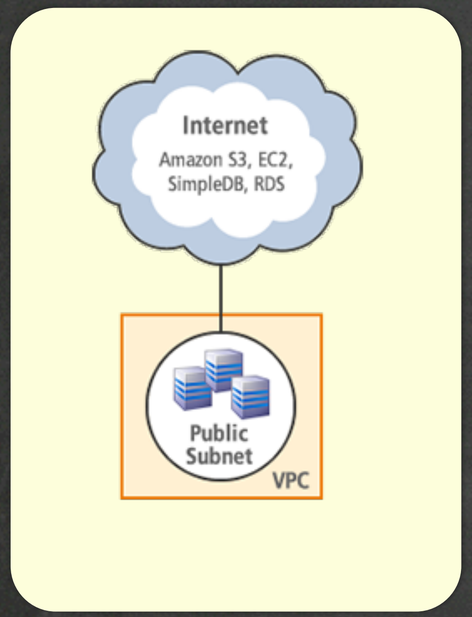
VPC With Private and Public Subnets & Hardware VPN access



Multi-tier web hosting with access to internal infrastructure

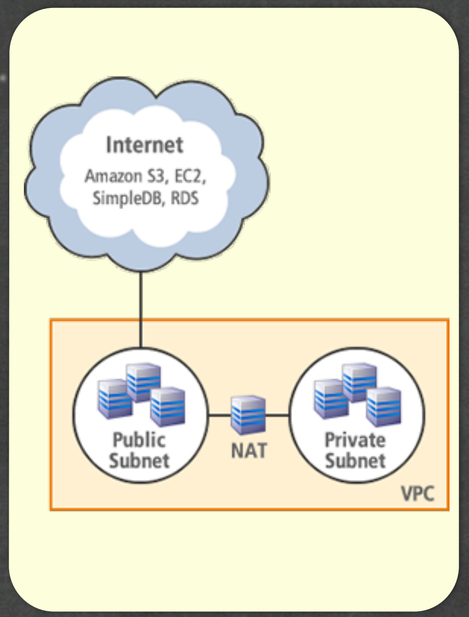
VPC Deployment Models

VPC With a Single Subnet



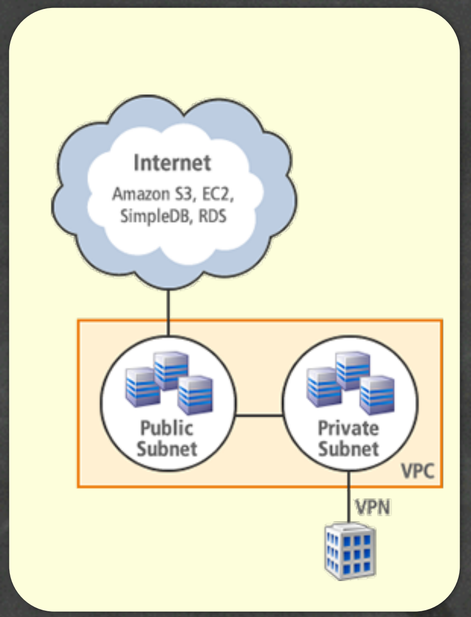
Simple web hosting, collaborative research environments, on-demand HPC/HTC clusters

VPC With Private and Public Subnets



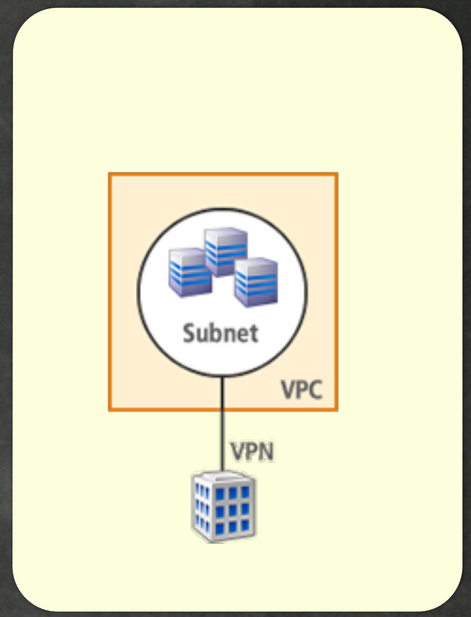
Multi-tier web hosting

VPC With Private and Public Subnets & Hardware VPN access

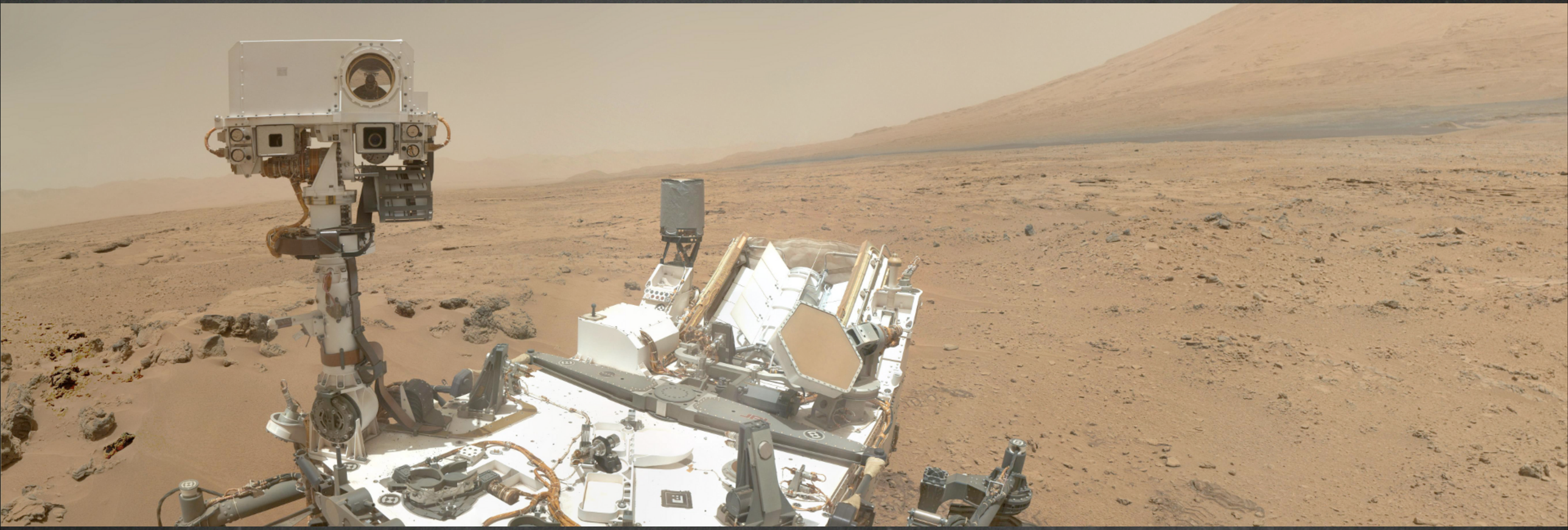


Multi-tier web hosting with access to internal infrastructure

VPC With a Private Subnet Only & Hardware VPN Access



Seamless private expansion of on-premise infrastructure. “Burst capacity” or dedicated cloud environments with connectivity to on-premise resources

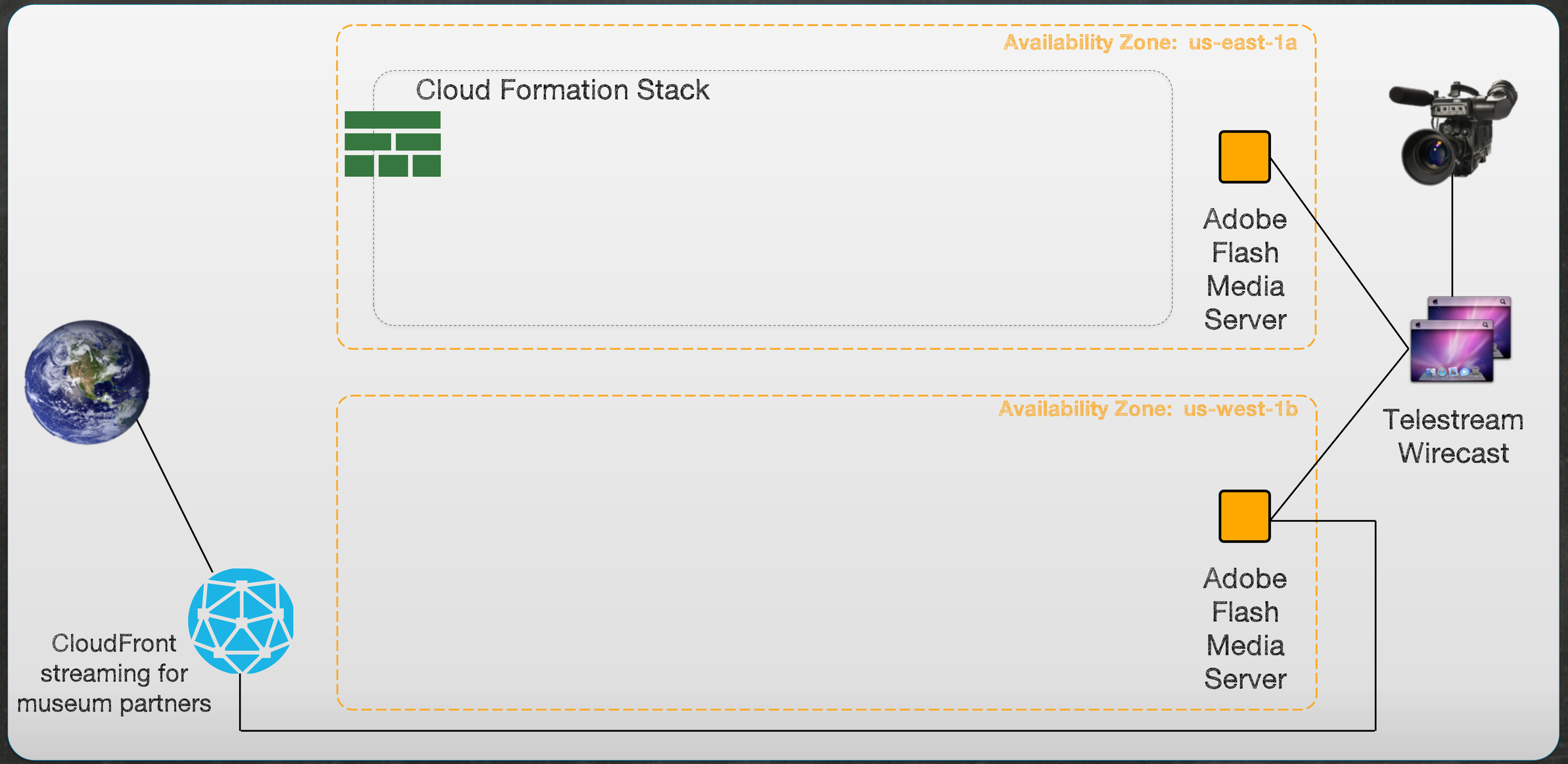


Mars Science Laboratory - Live Video Streaming

Architecture

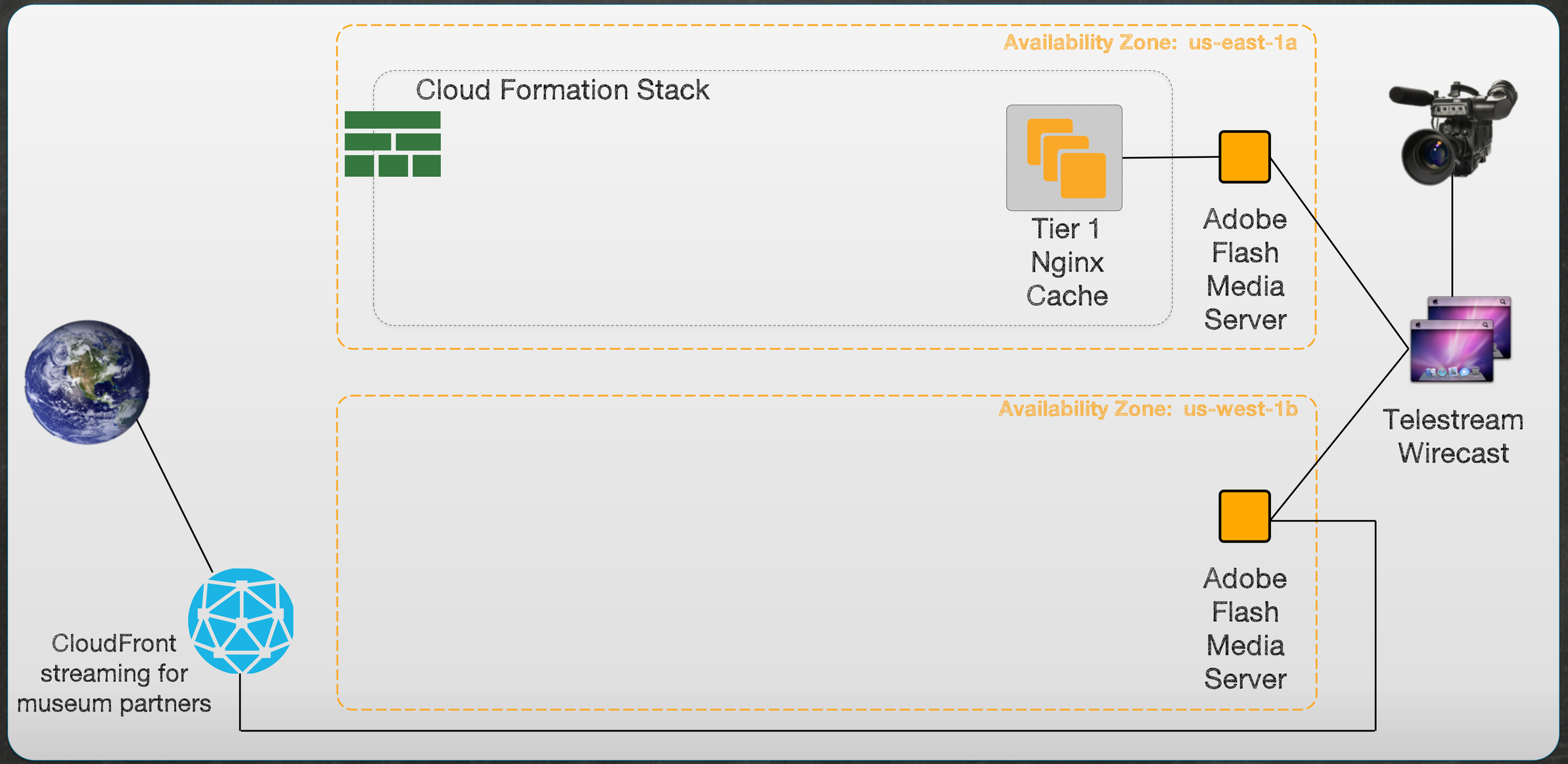
Mars Science Laboratory - Live Video Streaming

Architecture



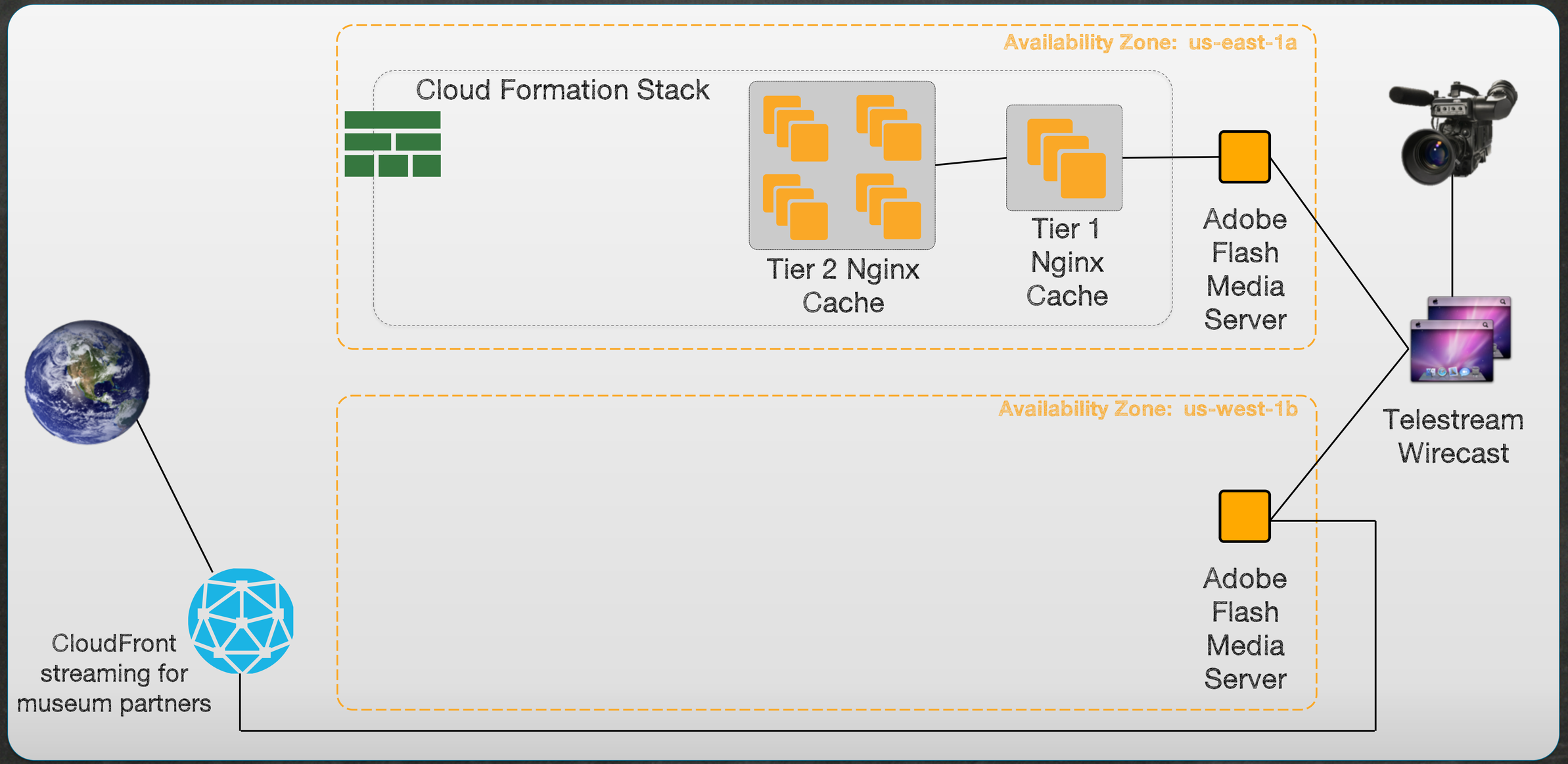
Mars Science Laboratory - Live Video Streaming

Architecture



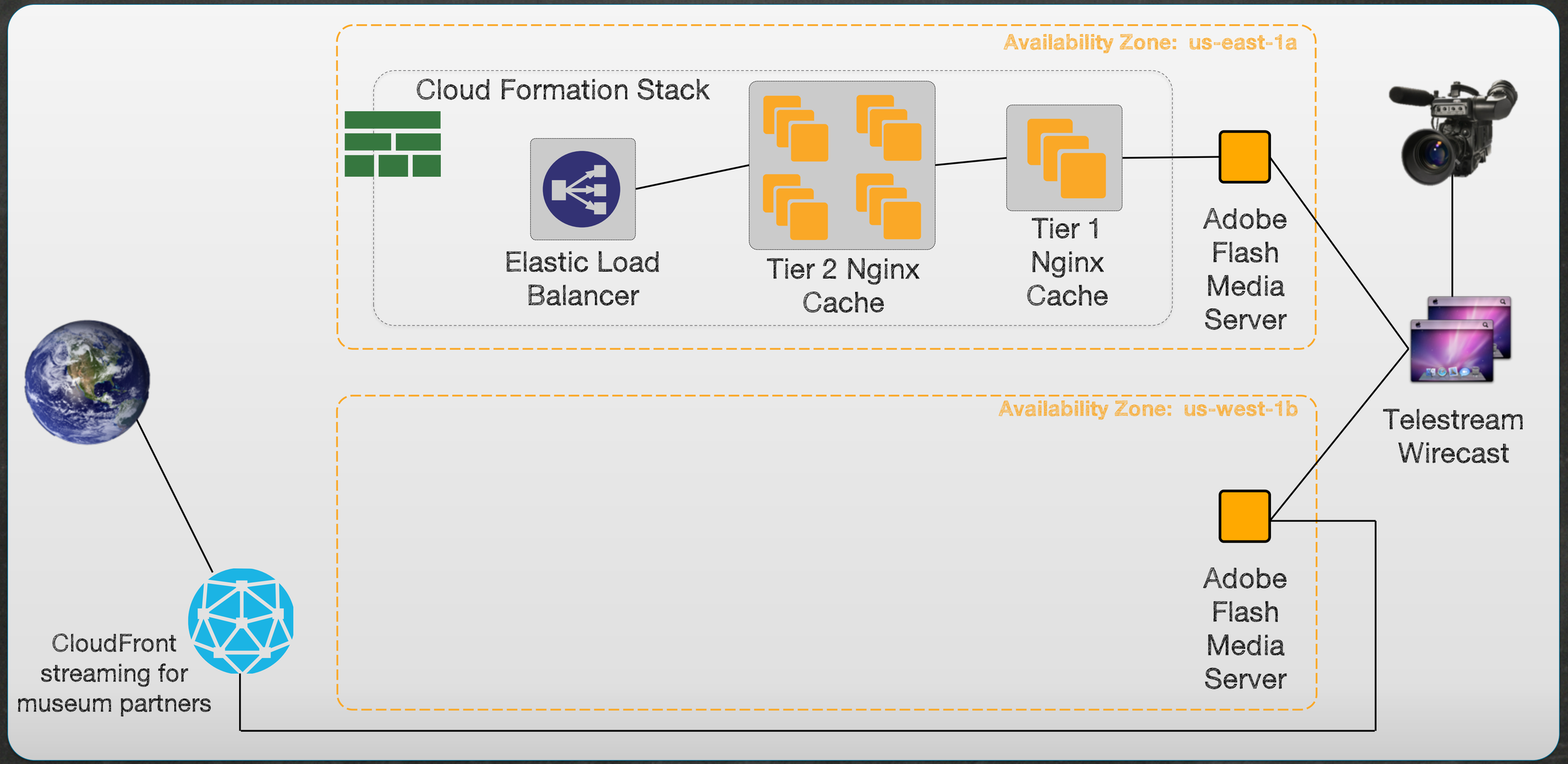
Mars Science Laboratory - Live Video Streaming

Architecture



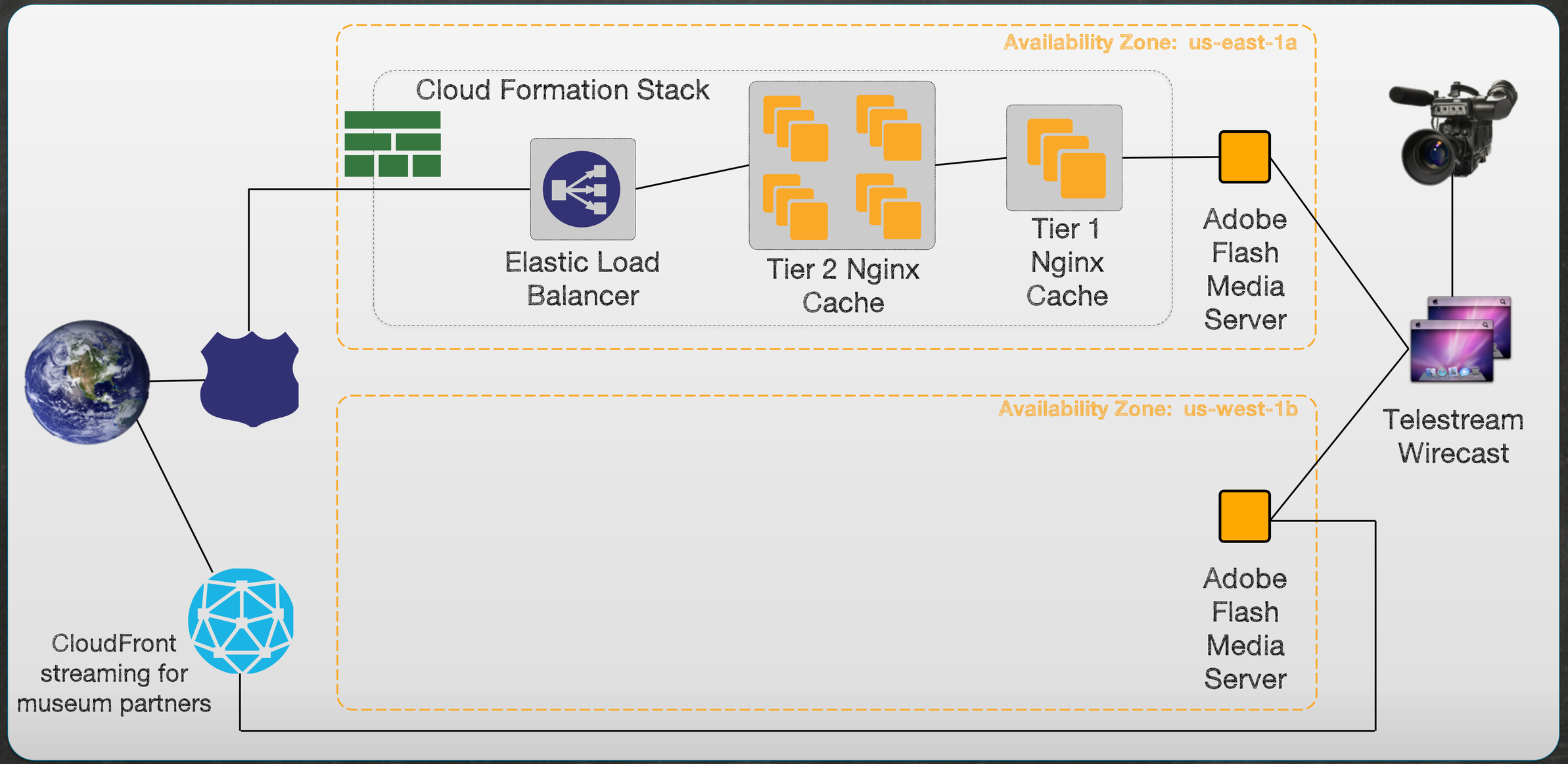
Mars Science Laboratory - Live Video Streaming

Architecture



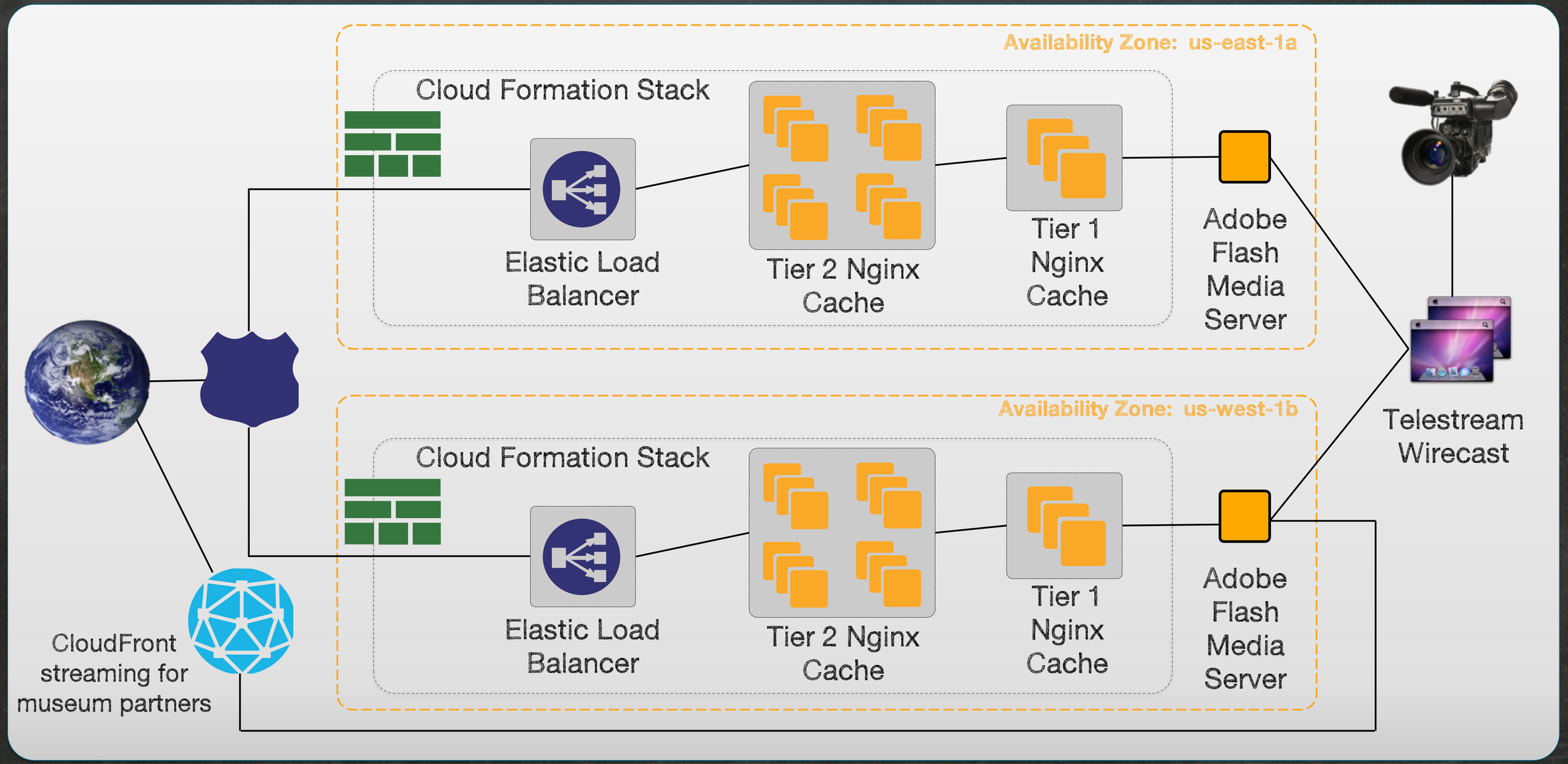
Mars Science Laboratory - Live Video Streaming

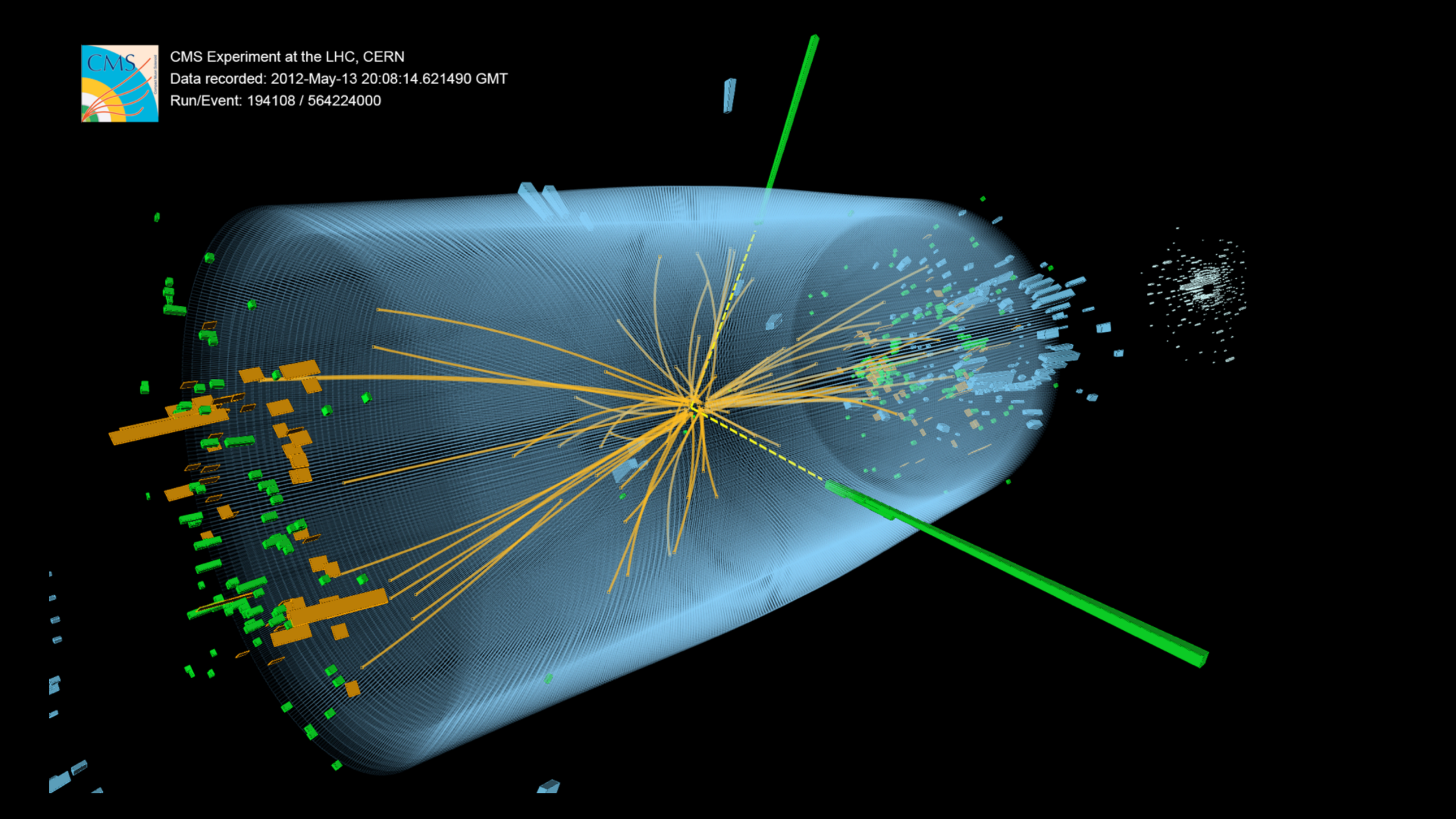
Architecture



Mars Science Laboratory - Live Video Streaming

Architecture




CMS Experiment at the LHC, CERN
Data recorded: 2012-May-13 20:08:14.621490 GMT
Run/Event: 194108 / 564224000

U. Wisc.: CMS Particle Detector

Trial of Amazon Spot Instances for CMS MC

Dan Bradley <dan@hep.wisc.edu>

For AAA Team

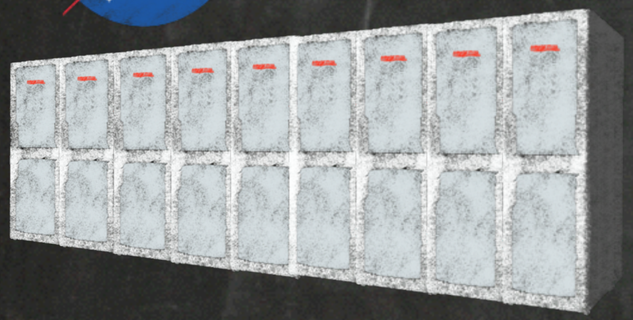
http://www.hep.wisc.edu/~dan/talks/EC2SpotForCMS.pdf

amazon web services

EC2

On-demand access to effectively limitless resources

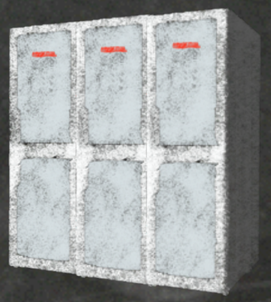
NASA Ames



Specialized supercomputing resources

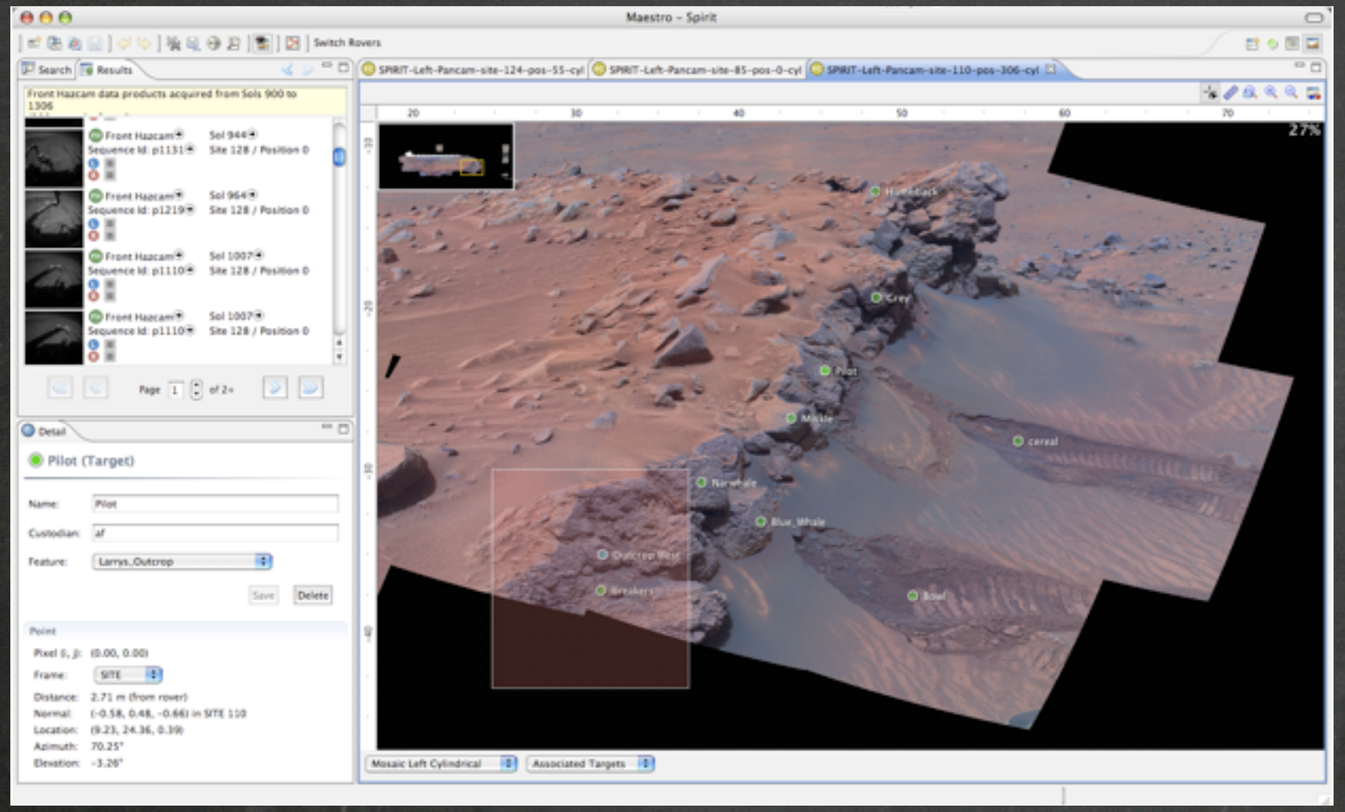
NASA Researcher

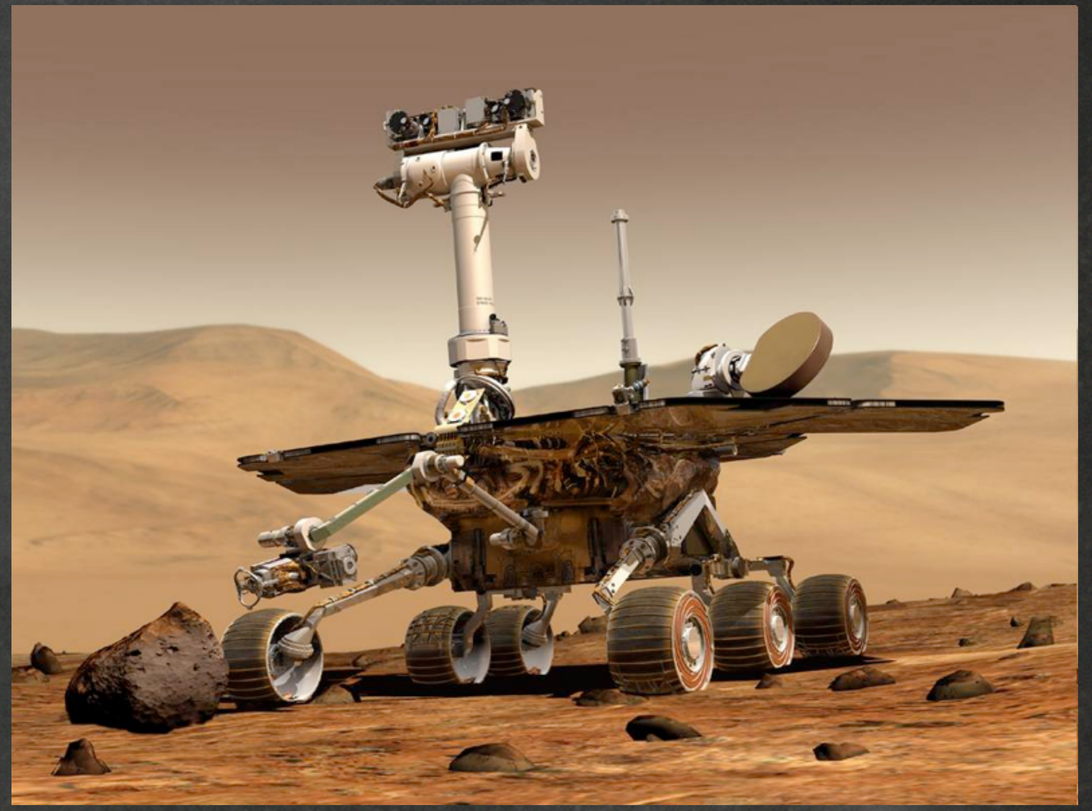
NASA JPL

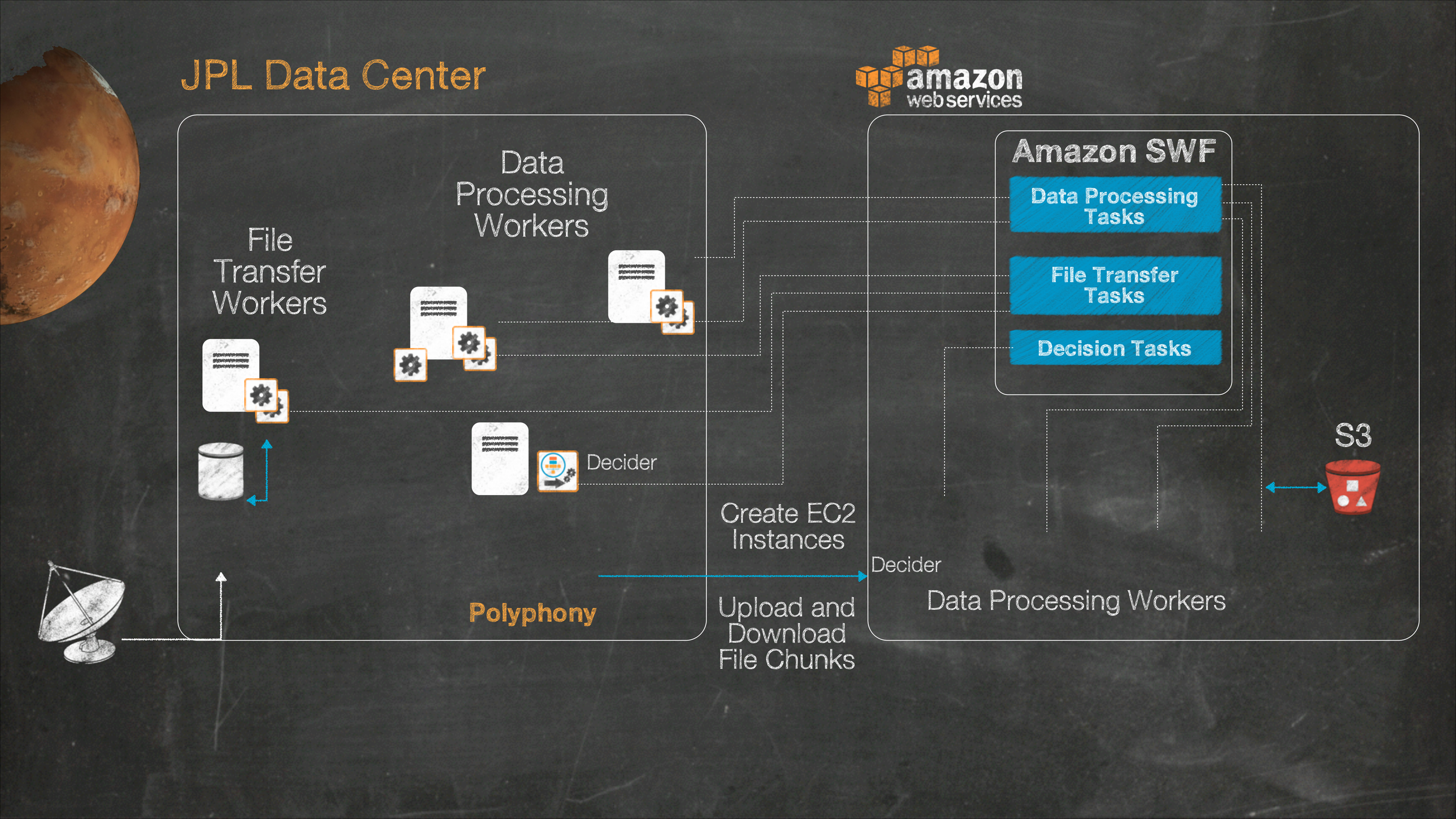


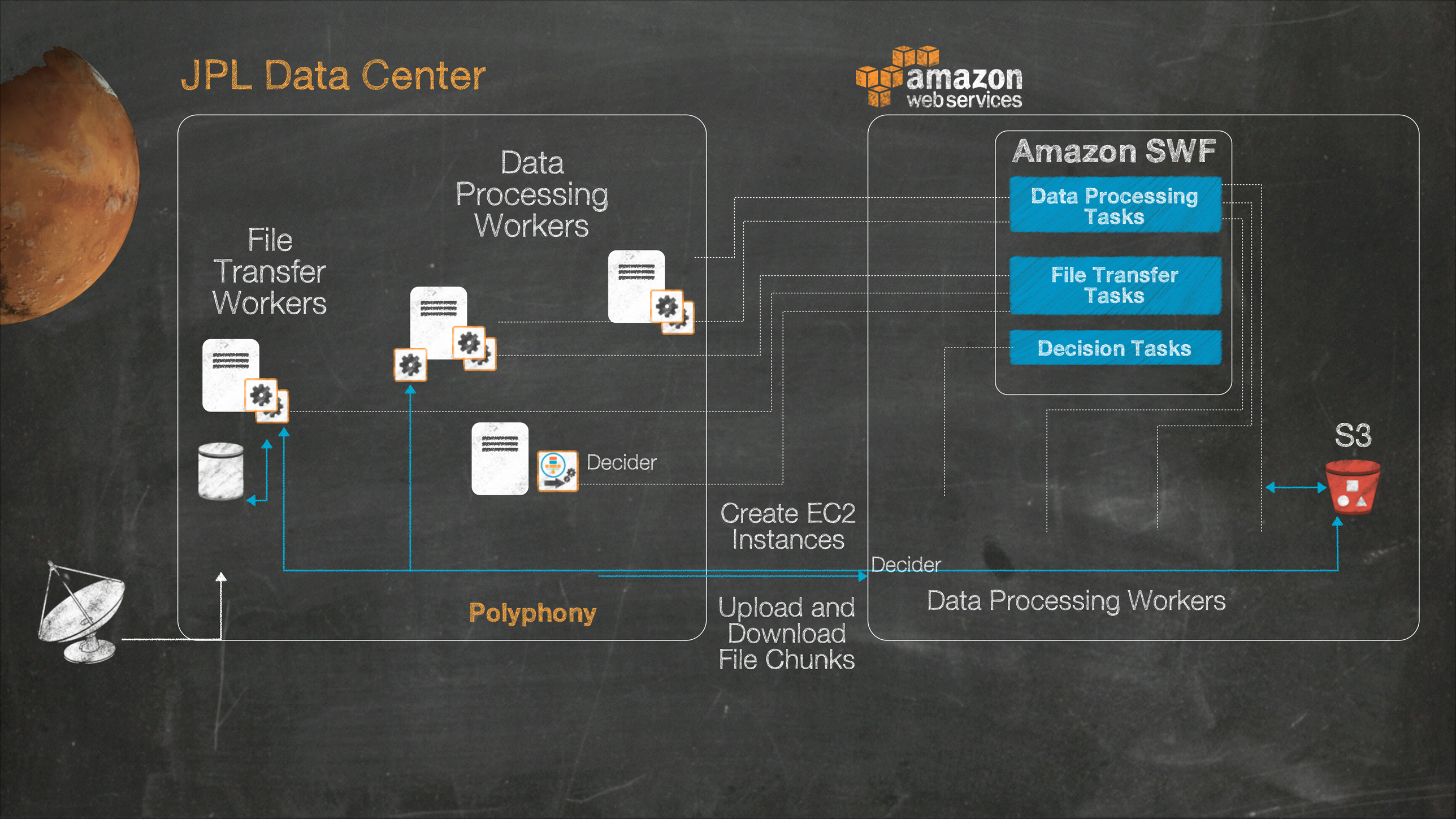
Small-scale shared compute

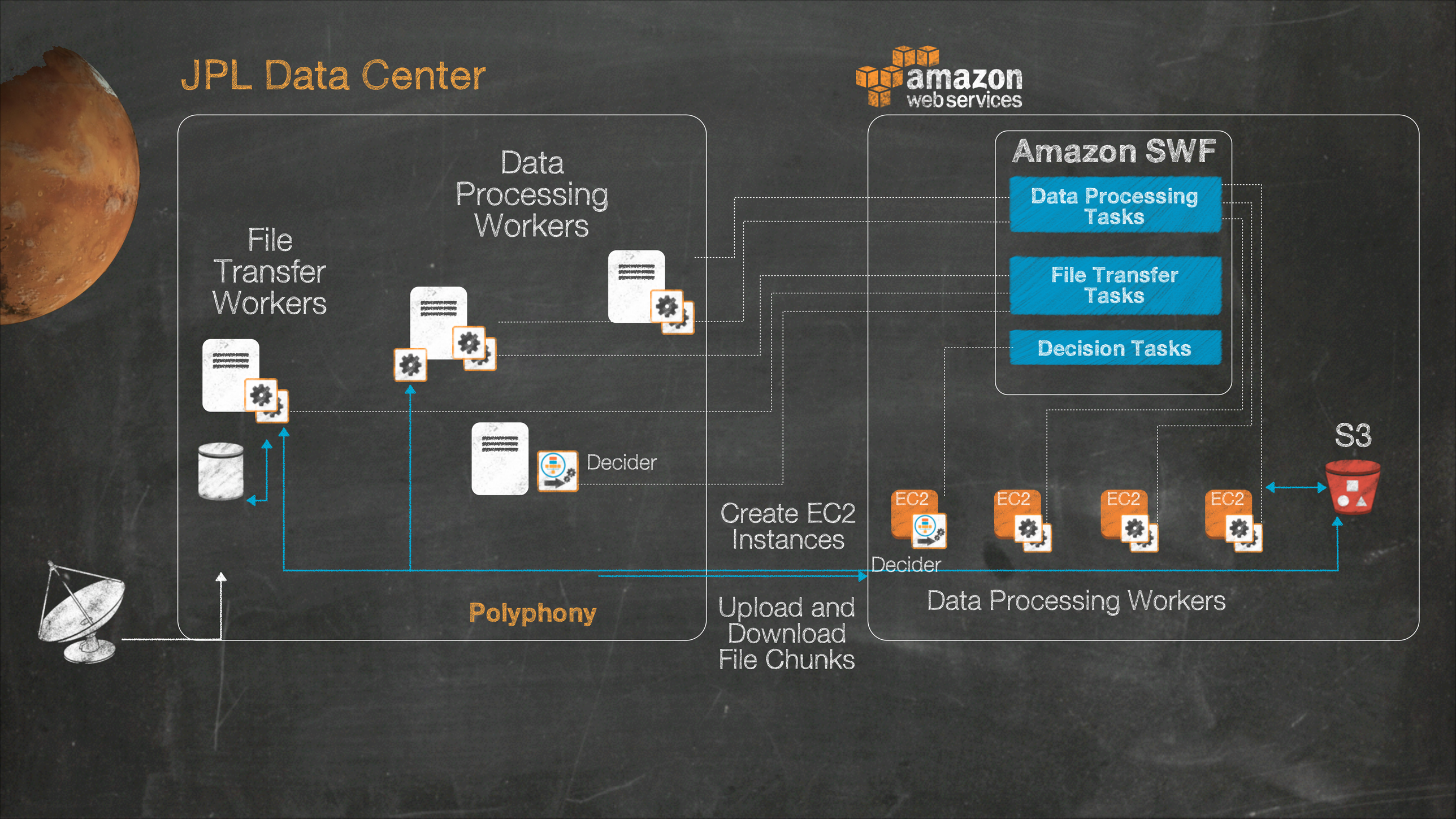
Mission-Critical Computing

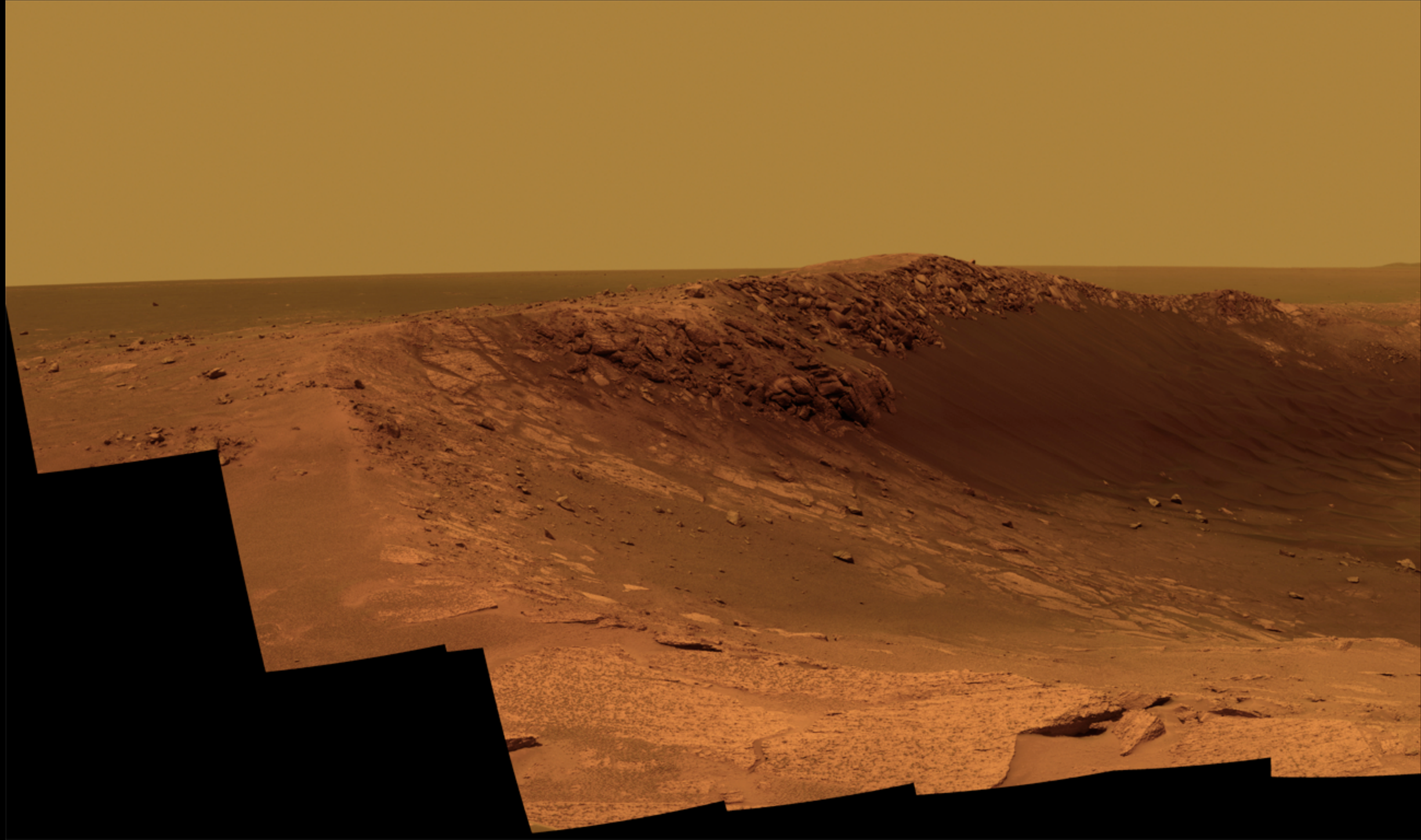





JPL Data Center
amazon web services
Amazon SWF
Data Processing Tasks
File Transfer Tasks
Decision Tasks
Data Processing Workers
File Transfer Workers
Decider
S3
Create EC2 Instances
Decider
Polyphony
Upload and Download File Chunks
Data Processing Workers


JPL Data Center
amazon web services
Amazon SWF
Data Processing Tasks
File Transfer Tasks
Decision Tasks
Data Processing Workers
File Transfer Workers
Decider
S3
Create EC2 Instances
Decider
Polyphony
Upload and Download File Chunks
Data Processing Workers

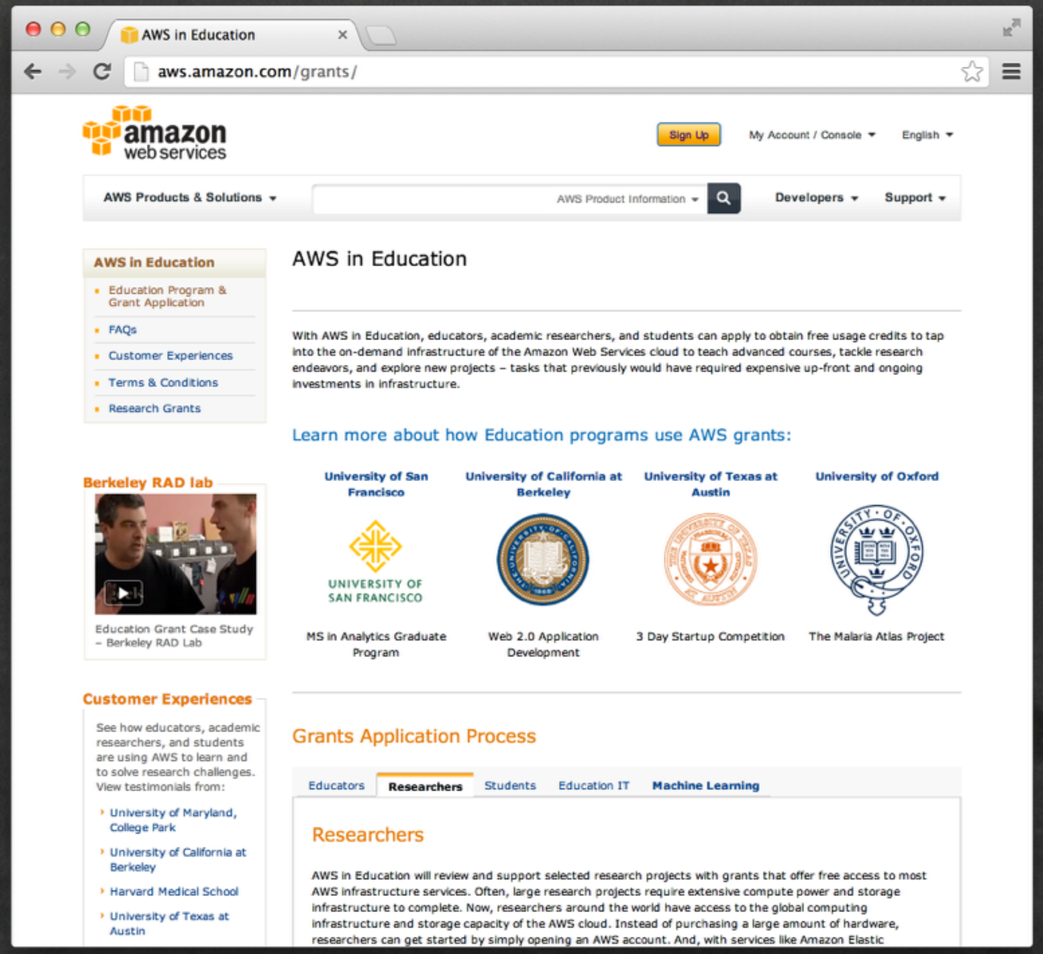

JPL Data Center
amazon web services
Amazon SWF
Data Processing Tasks
File Transfer Tasks
Decision Tasks
Data Processing Workers
File Transfer Workers
Decider
S3
EC2
EC2
EC2
EC2
Decider
Create EC2 Instances
Upload and Download File Chunks
Polyphony
Data Processing Workers



SWF
EC2
S3
SimpleDB
CloudWatch
IAMs
ELB

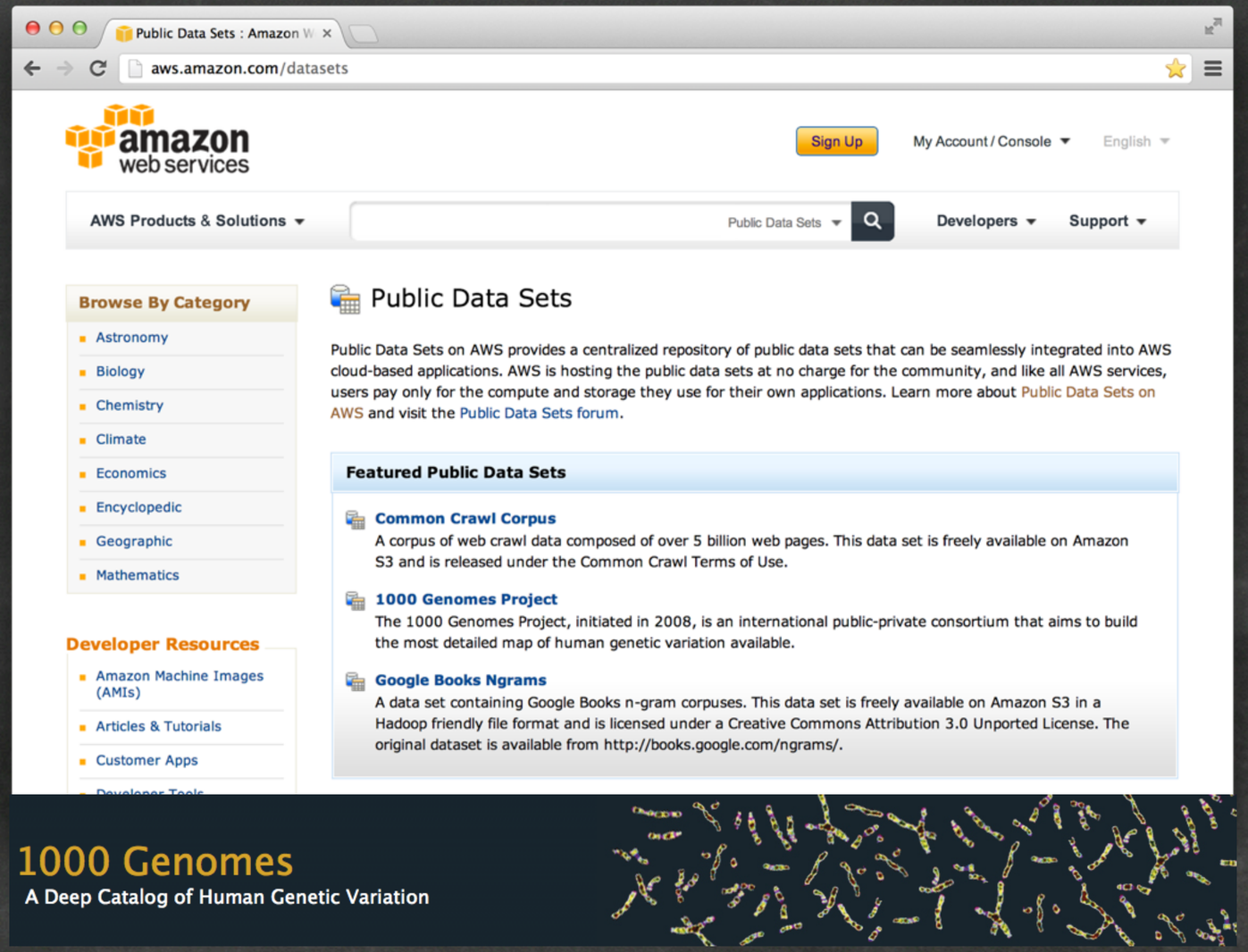
5 Giga-pixels in 5 minutes!

AWS Academic Grants



AWS.amazon.com/grants

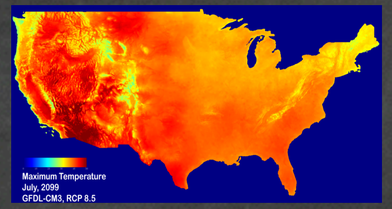
AWS Public Data Sets

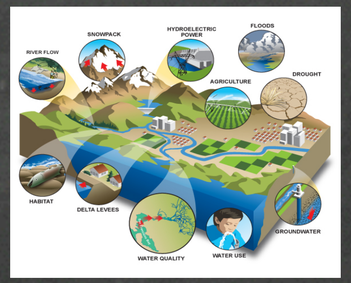


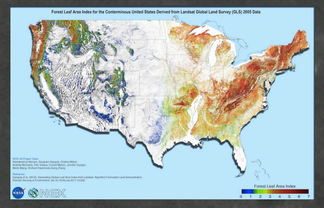
AWS.amazon.com/datasets

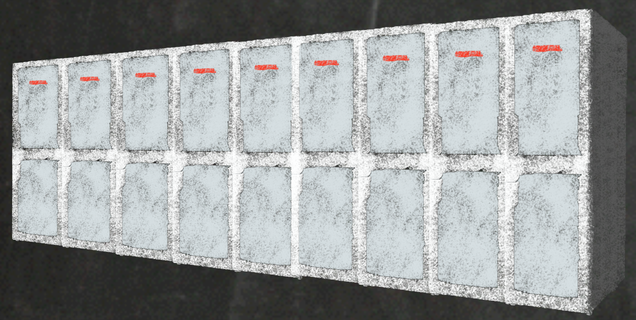
AWS and the NASA Earth eXchange (NEX)

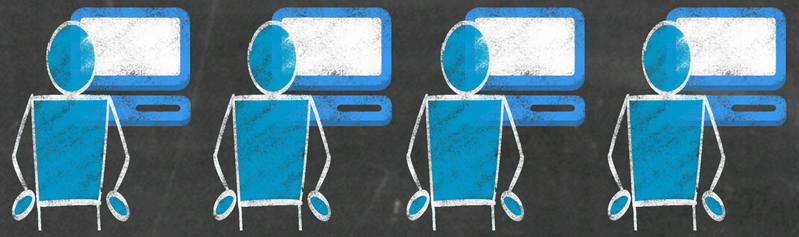
- National Climate Assesment datasets hosted on AWS
- Machine images, tutorials and hosted workshops provided by NASA
- Data and Software now available to those without @nasa.gov email addresses
- Enables crowd-sourced citizen science applications like those found on the Zooniverse





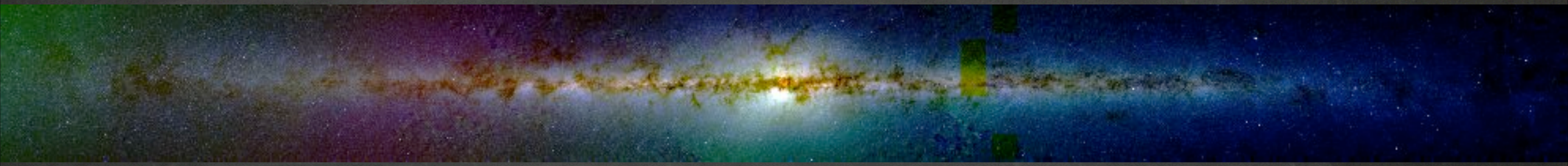




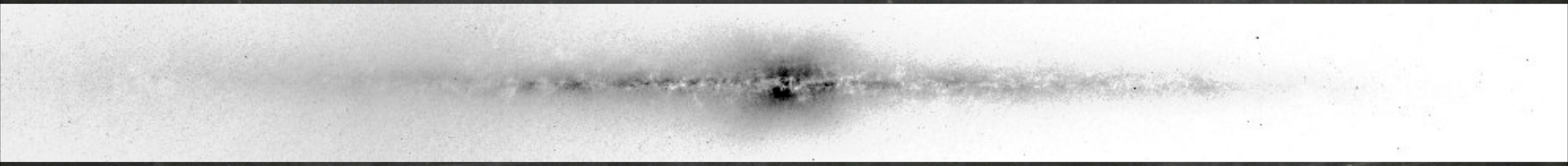


Climate Researchers

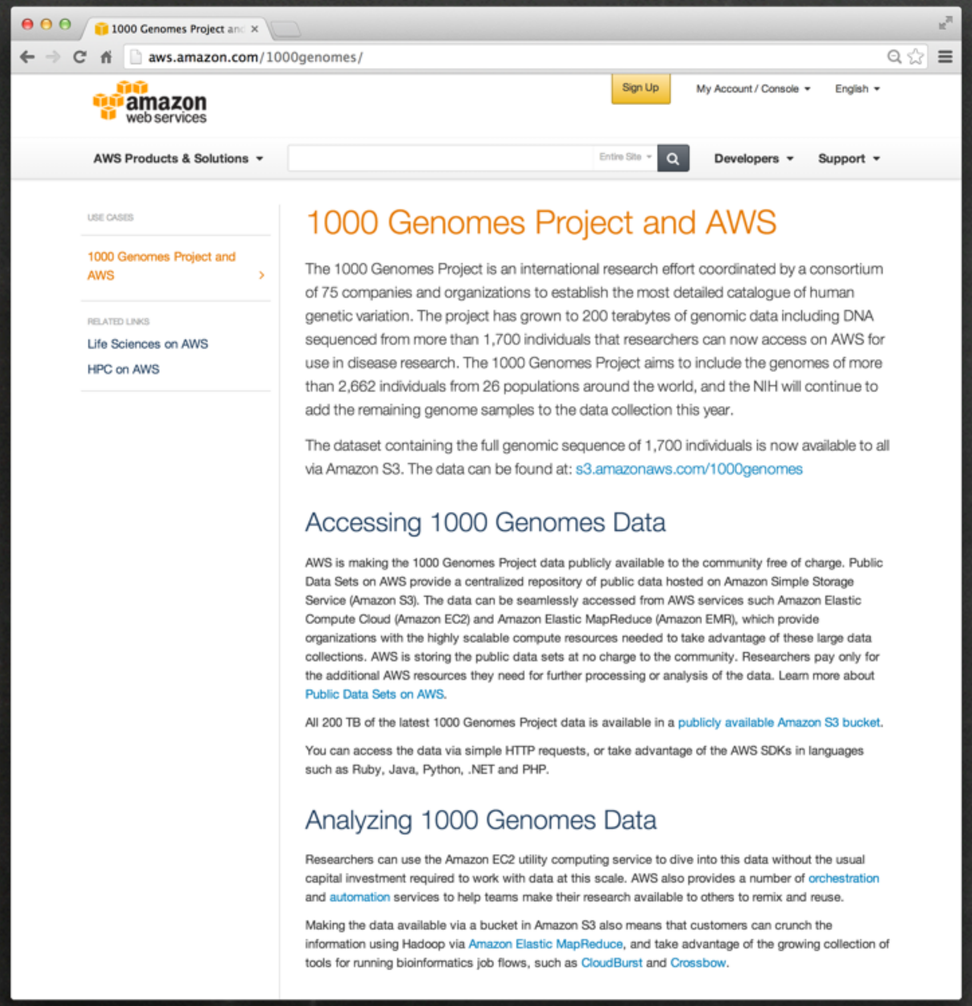
Multi-Spectrum Atlas of the Galactic Plane

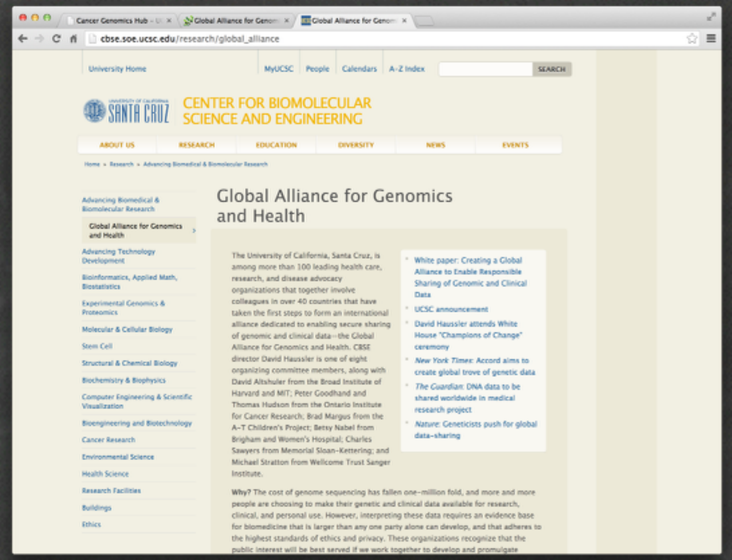


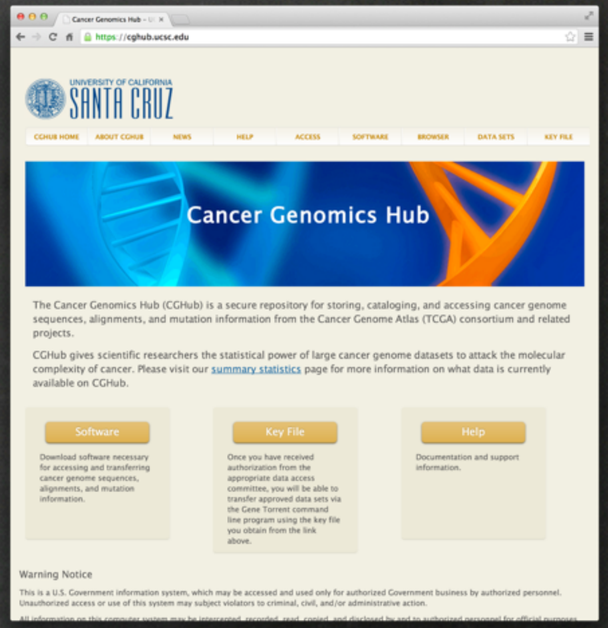
- Collaboration between AWS, Caltech/IPAC and USC/ISI
- All images are publicly accessible via direct download and VAO APIs
- 16 wavelength infrared atlas spanning 1μm to 70μm
- Datasets from GLIMPSE and MIPSGAL, 2MASS, MSX, WISE
- Spatial sampling of 1 arcsec with ± 180° longitude and ± 20° latitude
- Mosaics generated by Montage (http://montage.ipac.caltech.edu)

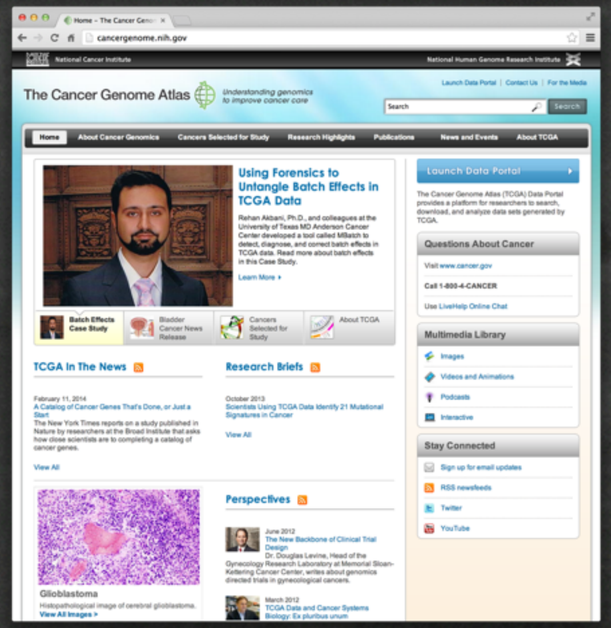


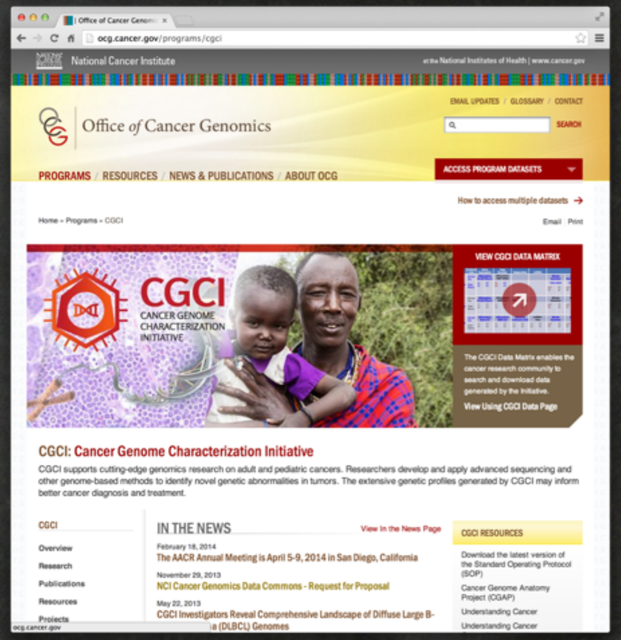
Cancer Research with AWS











National Database for Autism Research

http://ndar.nih.gov/cloud_overview.html

All autism research funded by NIMH must be publicly accessible

NDAR provides a web interface to query the aggregate data set

National Database for Autism Research - Data - Data from Labs - Mozilla Firefox

File Edit View History Bookmarks Tools Help

National Database for Autism Research -...

www.ndar.nih.gov/data_from_labs.html

Google

NDAR National Database for Autism Research

Serving the autism research community

Home | Query | Harmonization Tools | ntribute | est Access | Policy | Tutorials | About | FAQ | login

Query Data | Data from Labs | ata from Papers | Query Instructions

Expand

Collection Title: *1/2 Development of a Screening Interview for Research Studies of ASD*

Investigators: Christoper Monk, Somer Bishop and Cathy Lord (Owner: Lord, Catherine)

Collection Description: This collection contains phenotypic data on subjects diagnosed with ASD and subjects with non-ASD diagnoses.

Download Data

Genomics | Neuroimaging | Phenotype

Expand

Collection Title: *3/3-Multisite RCT of Early Intervention for Spoken Communication in Autism*

Investigators: Dr. Rebecca Landa (Owner: Landa, Rebecca)

Collection Description: School-Age Children with Autism with Limited Expressive Language Skills: An Intervention Study

Download Data

Genomics | Neuroimaging | Phenotype

Expand

Collection Title: *AGRE - U24 Collection*

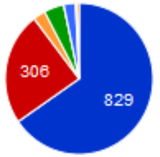
Investigators: Clara Lajonchere (Owner: Lajonchere, Clara)

Collection Description: AGRE is a DNA repository and family registry, housing a database of genotypic and phenotypic information that is available to autism researchers worldwide.

Download Data

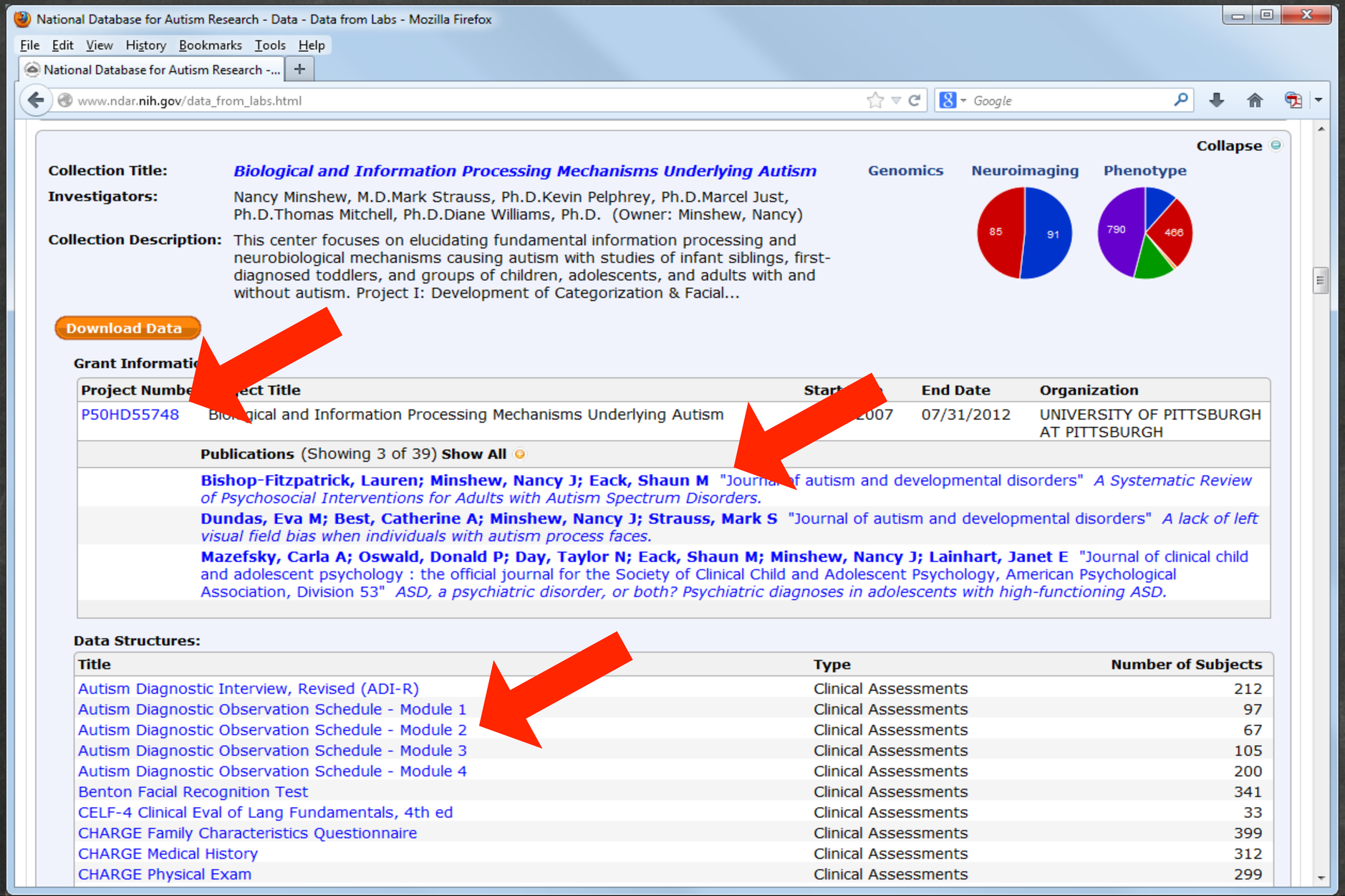
Genomics | Neuroimaging | Phenotype

No Data Shared



Find: Next Previous Highlight all Match case

Data Associated With a Particular Laboratory



Data From Papers

NDAR National Database for Autism Research
Serving the autism research community

Home | Query | Harmonization Tools | Contribute | Request Access | Policy | Tutorials | About | FAQ

Query Data | Data from Labs | Data from Papers | Query Instructions

Study Title: *Autism Genome Project* NDAR
Investigators: Scherer, S.; Devlin, B;
Study Abstract: The Autism Genome Project (AGP) Consortium represents more than 50 centers in North America and Europe. In an ongoing effort, the international AGP Consortium is collecting ASD families for ongoing genetic studies. The first phase of this initiative involved examining genetic linkage and...
Results: *Genotype Calls*, *Genotype Calls, part 2*
Documents:
Study Cohorts: Test - Non-verbal (483 subjects); Control - Parents (2933 subjects); Test - Verbal subjects (1000 subjects)
Study Measures: Primary Measures (2); Secondary Measures (4)
Data Analysis: Genotyping
Expand

Study Title: *Gastrointestinal Dysfunction in Autism: Parental Report, Clinical Evaluation, and Associated Factors* NDAR
Investigators: Levitt, Pat; Gorrindo, Phillip Williams, Kent C. Lee, Evon B. Walker, Lynn S. McGrew, Susan G. Levitt, Pat
Study Abstract: The objectives of this study were to characterize gastrointestinal dysfunction (GID) in autism spectrum disorder (ASD), to examine parental reports of GID relative to evaluations by pediatric gastroenterologists, and to explore factors associated with GID in ASD. One hundred twenty-one children...
Results: *Results published in Autism Research*
Documents:
Study Cohorts: Control - GID-only (28 subjects); Test - ASD-GID (40 subjects); Control - ASD-only (44 subjects)
Study Measures: Primary Measures (2); Secondary Measures (5)
Data Analysis: Statistical
Expand

Study Title: *Somatic copy-number mosaicism in human skin revealed by induced pluripotent stem cell* NDAR
Investigators: Vaccarino, Flora; Abyzov, Alexej; Mariani, Jessica; Palejev, Dean; Zhang, Ying; Haney, Michael Seamus; Tomasini, Livia; Ferrandino, Anthony; Rosenberg Belmaker, Lior; Szekely, Anna; Wilson, Michael; Kocabas, Arif; Calixto, Nathaniel E.; Grigorenko, Elena L.; Huttner, Anita; Chawarska, Katarzyna; Weissman, Sherman; Urban, Alexander Eckehart; Gerstein, Mark; Vaccarino, Flora
Study Abstract: Reprogramming human somatic cells into induced pluripotent stem cells (iPSC) has been suspected of causing de novo copy number variations (CNVs). To explore this issue, we performed a whole-genome and transcriptome analysis of 20 human iPSC lines derived from primary skin fibroblasts of 7 individuals using next-generation sequencing. We find that, on average, an iPSC line manifests two CNVs not apparent in the fibroblasts from which the iPSC was derived. Using qPCR, PCR, and digital droplet PCR (ddPCR) to amplify across the CNVs' breakpoints, we show that at least 50% of those CNVs are present as low frequency somatic genomic variants in parental fibroblasts and are manifested in iPSC colonies due to their clonal origin. Hence, reprogramming does not necessarily lead to de novo CNVs in iPSC, since most of line-manifested CNVs reflect somatic mosaicism in the human skin. Moreover, our findings demonstrate that clonal expansion, and iPSC lines in particular, can be used as a discovery tool to reliably detect low frequency CNVs in the tissue of origin. Overall, we estimate that approximately 30% of the fibroblast cells have somatic CNVs, suggesting widespread somatic mosaicism in the human body. Our study paves the way to understanding the fundamental question of the extent to which cells of the human body normally acquire structural alterations in their DNA post-zygotically.
Results: *publication in Nature*
Documents: *R01MH89176*, *R33MH087879*, *SFARI GRANT 137055 R09345*
Study Cohorts: Control - Control (7 subjects); Age: 0 to 700 months; Gender: Both; Test - Test (2 subjects); Age: 90 to 200 months; Gender: Both
Study Measures: Primary Measures (2); Genomics: Genomics Sample - v02, Genomics Sample - v03; Secondary Measures (2); Genomics: Genomics Subject - v01, Genomics Subject - v02
Data Analysis: Genotyping; Data Transformation, QA/QC: Chromosomal Aberration Screening, Family-Based QC, Gender Misidentification, Quartile Summary Statistics, Wave Detection and Correction; Genetic Model: Additive model, Allele test, Recessive model; Genotype Statistics by Marker and Sample: Hardy-Weinberg Equilibrium (HWE) p-value, Minor allele frequency; Software: BWA, CNVnator, FBAT, GATK, SAMtools, SpectralGEM; Test Statistics: Correlation/Trend test, F-test, Fisher's exact test, Pearson's chi-square test; Type: CNV, Exome Sequencing, GWAS, SNP, Whole Genome sequencing
Microarray; Algorithm: Bayesian Method; Normalization Method: Global; Significance; p-value: <0.01; Software: GenomeStudio, TM4 - MeV MultiExperiment Viewer; Type: Gene Expression
Collapse

Study Title: *Using Whole-Exome Sequencing to Identify Inherited Causes of*
Study Cohorts: Control - Parents and siblings (411 subjects)
Expand

Data from Labs | **Data from Papers** | Autism Publications

Gastrointestinal Dysfunction in Autism: Parental Report, Clinical Evaluation, and Associated Factors NDAR

Investigators: Levitt, Pat; Gorrindo, Phillip Williams, Kent C. Lee, Evon B. Walker, Lynn S. McGrew, Susan G. Levitt, Pat

Study Abstract: The objectives of this study were to characterize gastrointestinal dysfunction (GID) in autism spectrum disorder (ASD), to examine parental reports of GID relative to evaluations by pediatric gastroenterologists, and to explore factors associated with GID in ASD. One hundred twenty-one children were recruited into three groups: co-occurring ASD and GID, ASD without GID, and GID without ASD. A pediatric gastroenterologist evaluated both GID groups. Parents in all three groups completed questionnaires about their child`s behavior and GI symptoms, and a dietary journal. Functional constipation was the most common type of GID in children with ASD (85.0%). Parental report of any GID was highly concordant with a clinical diagnosis of any GID (92.1%). Presence of GID in children with ASD was not associated with distinct dietary habits or medication status. Odds of constipation were associated with younger age, increased social impairment, and lack of expressive language (adjusted odds ratio in nonverbal children: 11.98, 95% confidence interval 2.54-56.57). This study validates parental concerns for GID in children with ASD, as parents were sensitive to the existence, although not necessarily the nature, of GID. The strong association between constipation and language impairment highlights the need for vigilance by health-care providers to detect and treat GID in children with ASD. Medications and diet, commonly thought to contribute to GID in ASD, were not associated with GID status. These findings are consistent with a hypothesis that GID in ASD represents pleiotropic expression of genetic risk factors.

Results: *Results published in Autism Research*

Study Cohorts: Collapse

Test - ASD-GID (40 subjects)
- Age: 60 to 215 months
- Gender: Both
- Autism-like Developmental Disorders: Affected, Mildly Affected, Severely Affected

Control - GID-only (28 subjects)
- Age: 60 to 215 months
- Gender: Both
- Non-ASD Control: Typical

Control - ASD-only (44 subjects)
- Age: 60 to 215 months
- Gender: Both
- Autism-like Developmental Disorders: Affected, Mildly Affected, Severely Affected

Study Measures: Primary Measures (2)
- Clinical Assessments: Questionnaire on Pediatric GI Symptoms: Rome III Parent - v01, Social Responsiveness Scale (SRS) - v02

Secondary Measures (5)
- Clinical Assessments: Autism Diagnostic Observation Schedule - Module 1 - v02, Autism Diagnostic Observation Schedule - Module 2 - v02, Autism Diagnostic Observation Schedule - Module 3 - v02, Autism Diagnostic Observation Schedule - Module 4 - v02, Diet Diary - v02

Data Analysis: Statistical
- Method: ANOVA, Chi-square test, Regression analysis
- Significance
  - p-value: <0.05
- Software: SPSS

Gastrointestinal Dysfunction in Autism: Parental Report, Clinical Evaluation, and Associated Factors NDAR

Investigators: Levitt, Pat; Gorrindo, Phillip Williams, Kent C. Lee, Evon B. Walker, Lynn S. McGrew, Susan G. Levitt, Pat

Study Abstract: The objectives of this study were to characterize gastrointestinal dysfunction (GID) in autism spectrum disorder (ASD), to examine parental reports of GID relative to evaluations by pediatric gastroenterologists, and to explore factors associated with GID in ASD. One hundred twenty-one children were recruited into three groups: co-occurring ASD and GID, ASD without GID, and GID without ASD. A pediatric gastroenterologist evaluated both GID groups. Parents in all three groups completed questionnaires about their child`s behavior and GI symptoms, and a dietary journal. Functional constipation was the most common type of GID in children with ASD (85.0%). Parental report of any GID was highly concordant with a clinical diagnosis of any GID (92.1%). Presence of GID in children with ASD was not associated with distinct dietary habits or medication status. Odds of constipation were associated with younger age, increased social impairment, and lack of expressive language (adjusted odds ratio in nonverbal children: 11.98, 95% confidence interval 2.54-56.57). This study validates parental concerns for GID in children with ASD, as parents were sensitive to the existence, although not necessarily the nature, of GID. The strong association between constipation and language impairment highlights the need for vigilance by health-care providers to detect and treat GID in children with ASD. Medications and diet, commonly thought to contribute to GID in ASD, were not associated with GID status. These findings are consistent with a hypothesis that GID in ASD represents pleiotropic expression of genetic risk factors.

Results: *Results published in Autism Research*

Direct access to the NDAR Collection for users with access to the NDAR portal

List of project investigators

Study abstract

Link to study results/publication

Direct access to the NDAR Collection for
AR portal

Gastrointest… Evaluation,

Investigators: Levitt, Pat; Lynn S. McG…

Study Abstract: The objecti… dysfunction … reports of GI… to explore fa… children were… without GID… evaluated b… questionnaire… dietary journ… in children … concordant … in children … medication … age, increas… (adjusted od… 2.54-56.57). … ASD, as par… the nature, … language im… providers to … diet, commonly thought to contribute to GID in ASD, were not associated with GID status. These findings are consistent with a hypothesis that GID in ASD represents pleiotropic expression of genetic risk factors.

Results: *Results published in Autism Research*

AUTISM RESEARCH

RESEARCH ARTICLE

Gastrointestinal Dysfunction in Autism: Parental Report, Clinical Evaluation, and Associated Factors

Phillip Gorrindo[1,*], Kent C. Williams[2], Evon B. Lee[2,3], Lynn S. Walker[2], Susan G. McGrew[2], Pat Levitt[4,*]

Article first published online: 9 JAN 2012

DOI: 10.1002/aur.237

© 2011 International Society for Autism Research, Wiley Periodicals, Inc.

Issue

Autism Research

Volume 5, Issue 2, pages 101–108, April 2012

Link to study results/publication

Subjects defined to a cohort

Outcome measures

Analysis methods defined

Study Cohorts: **Test - ASD-GID (40 subjects)** **Collapse**

Age: 60 to 215 months
Gender: Both
Autism-like Developmental Disorders: Affected, Mildly Affected, Severely Affected

Control - GID-only (28 subjects)

Age: 60 to 215 months
Gender: Both
Non-ASD Control: Typical

Control - ASD-only (44 subjects)

Age: 60 to 215 months
Gender: Both
Autism-like Developmental Disorders: Affected, Mildly Affected, Severely Affected

Study Measures: Primary Measures (2)

Clinical Assessments: Questionnaire on Pediatric GI Symptoms: Rome III Parent - v01, Social Responsiveness Scale (SRS) - v02

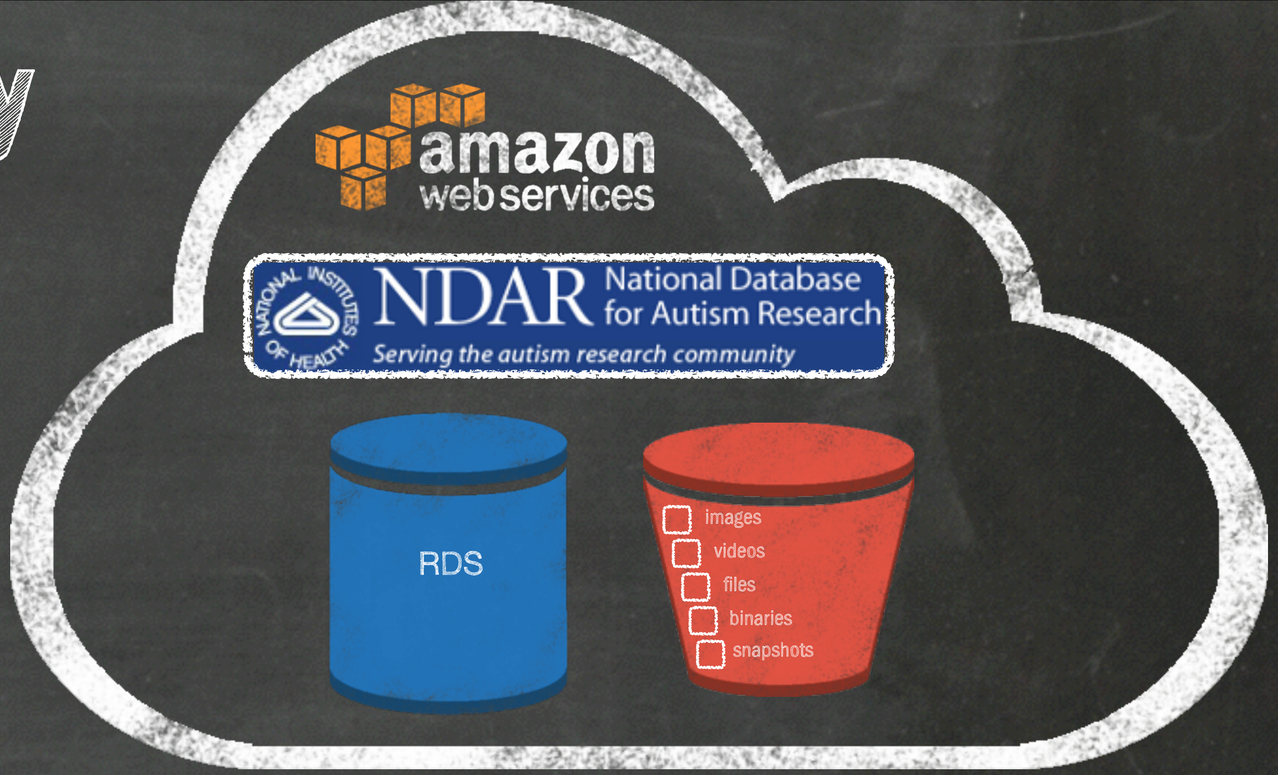
Secondary Measures (5)

Clinical Assessments: Autism Diagnostic Observation Schedule - Module 1 - v02, Autism Diagnostic Observation Schedule - Module 2 - v02, Autism Diagnostic Observation Schedule - Module 3 - v02, Autism Diagnostic Observation Schedule - Module 4 - v02, Diet Diary - v02

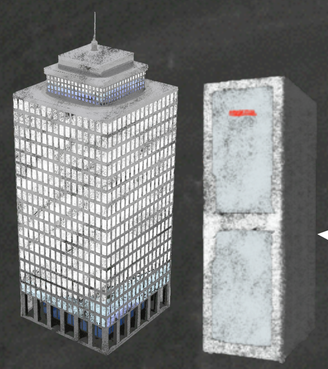
Data Analysis: Statistical

Method: ANOVA, Chi-square test, Regression analysis
Significance
p-value: <0.05
Software: SPSS

Download and Copy



Download and Copy



amazon web services

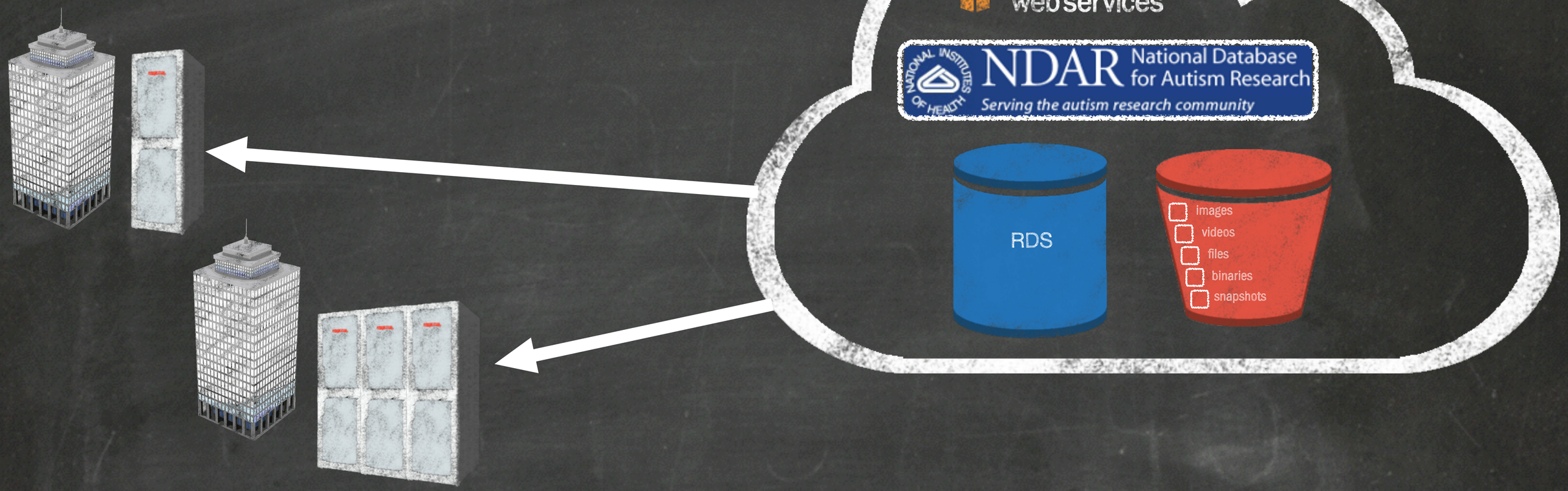
NDAR National Database for Autism Research

Serving the autism research community

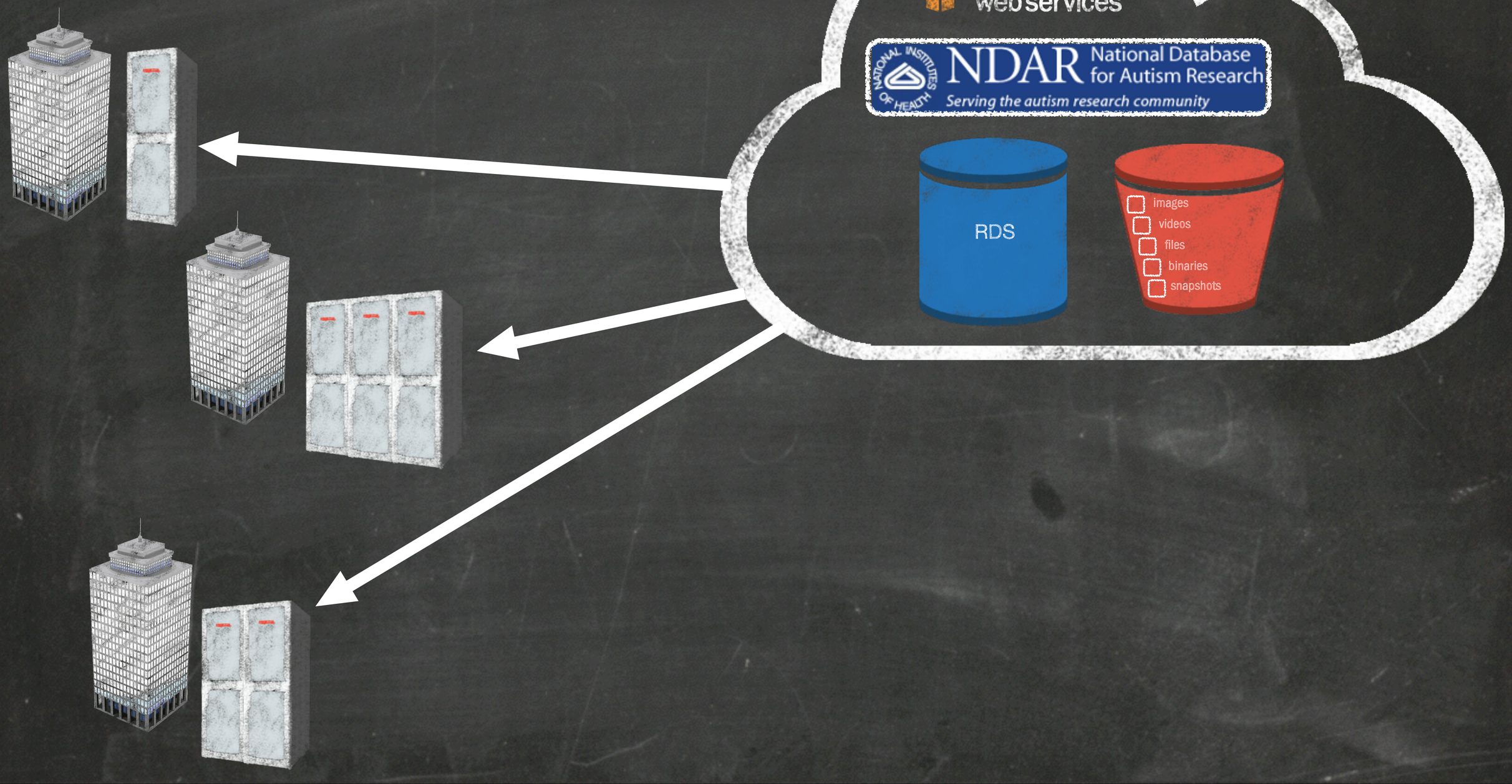
RDS

- images
- videos
- files
- binaries
- snapshots

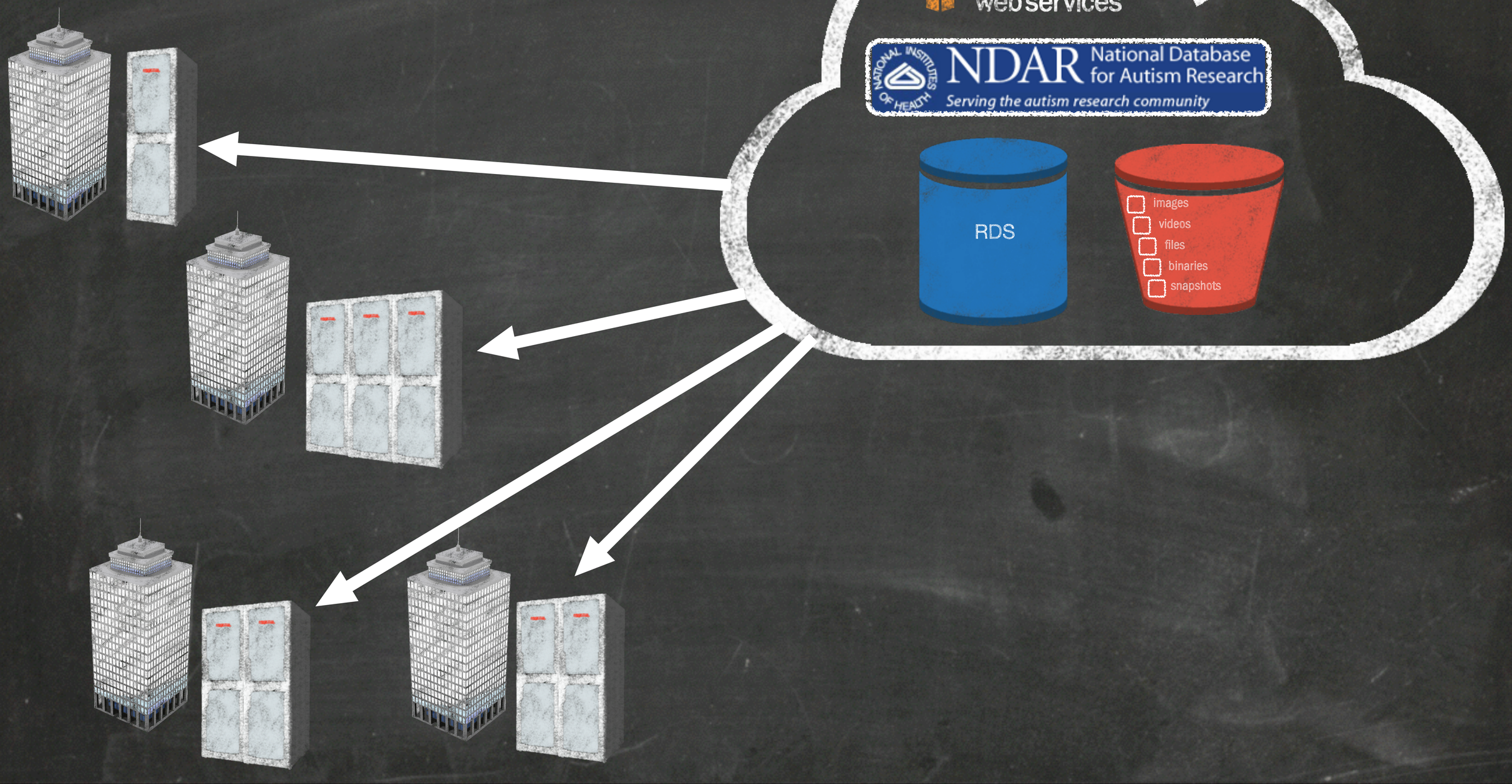
Download and Copy



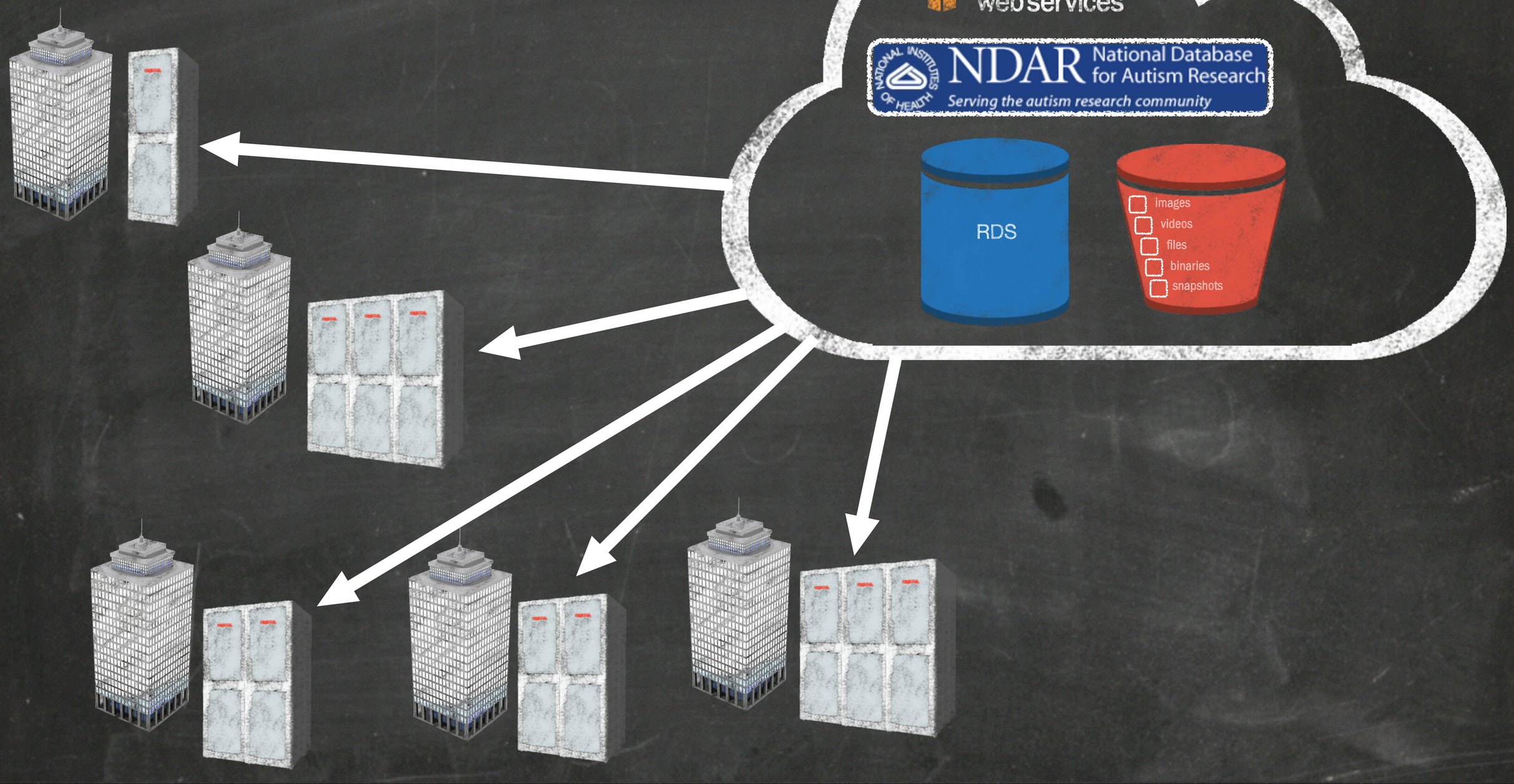
Download and Copy



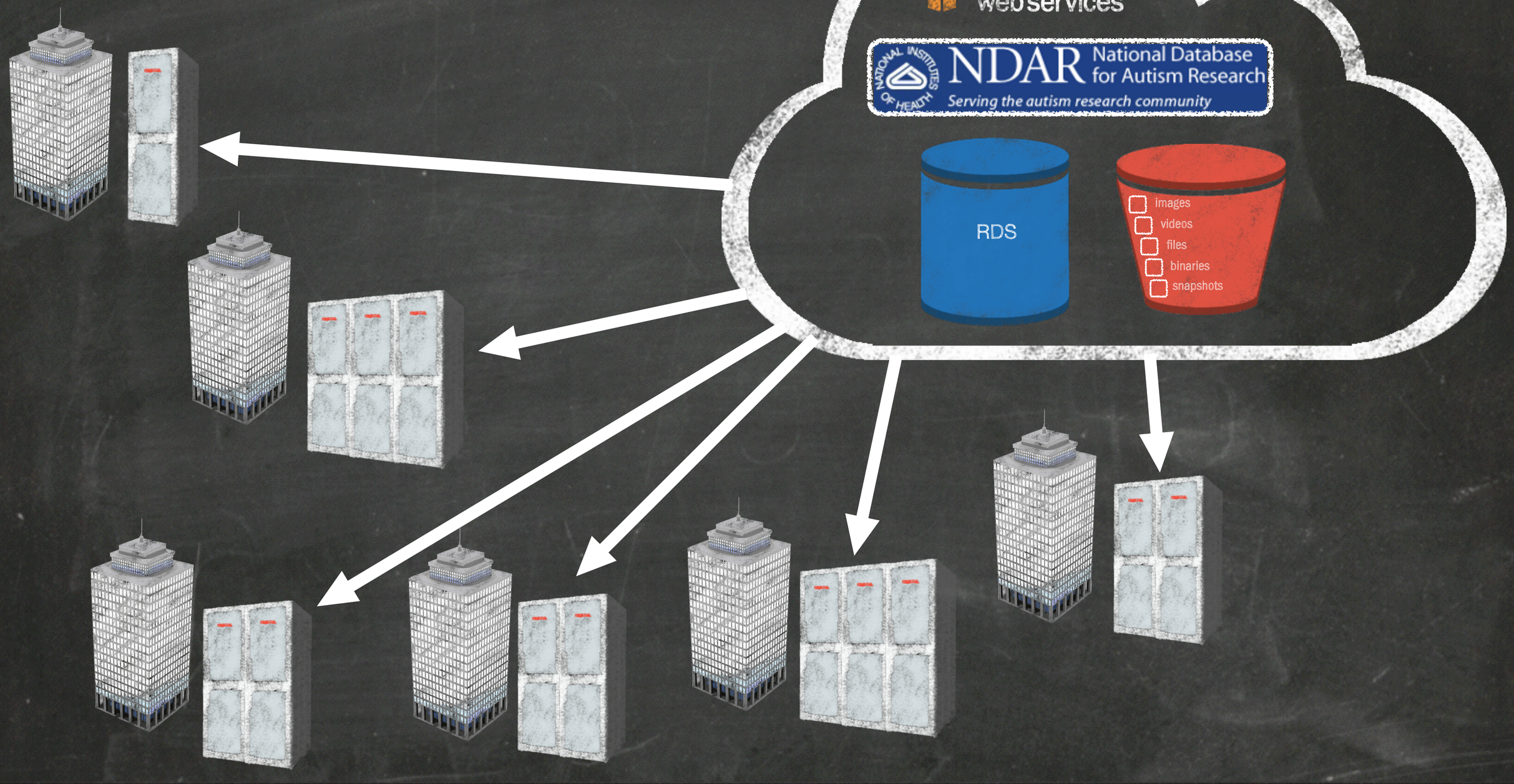
Download and Copy



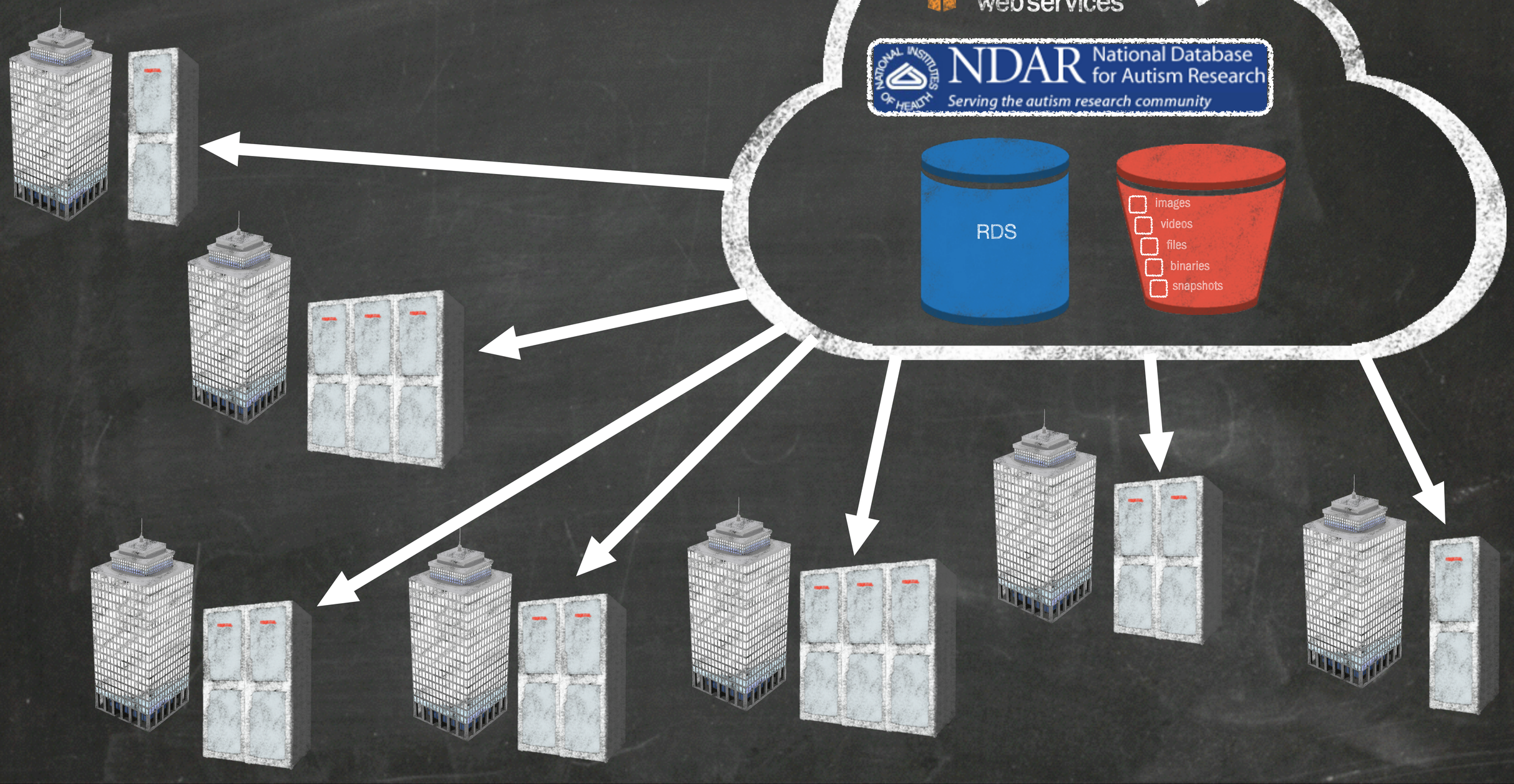
Download and Copy



Download and Copy

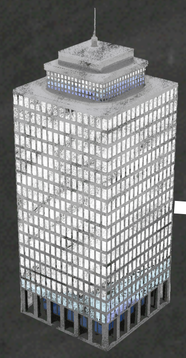


Download and Copy



Access in the Cloud

Access in the Cloud



amazon web services

NDAR National Database for Autism Research

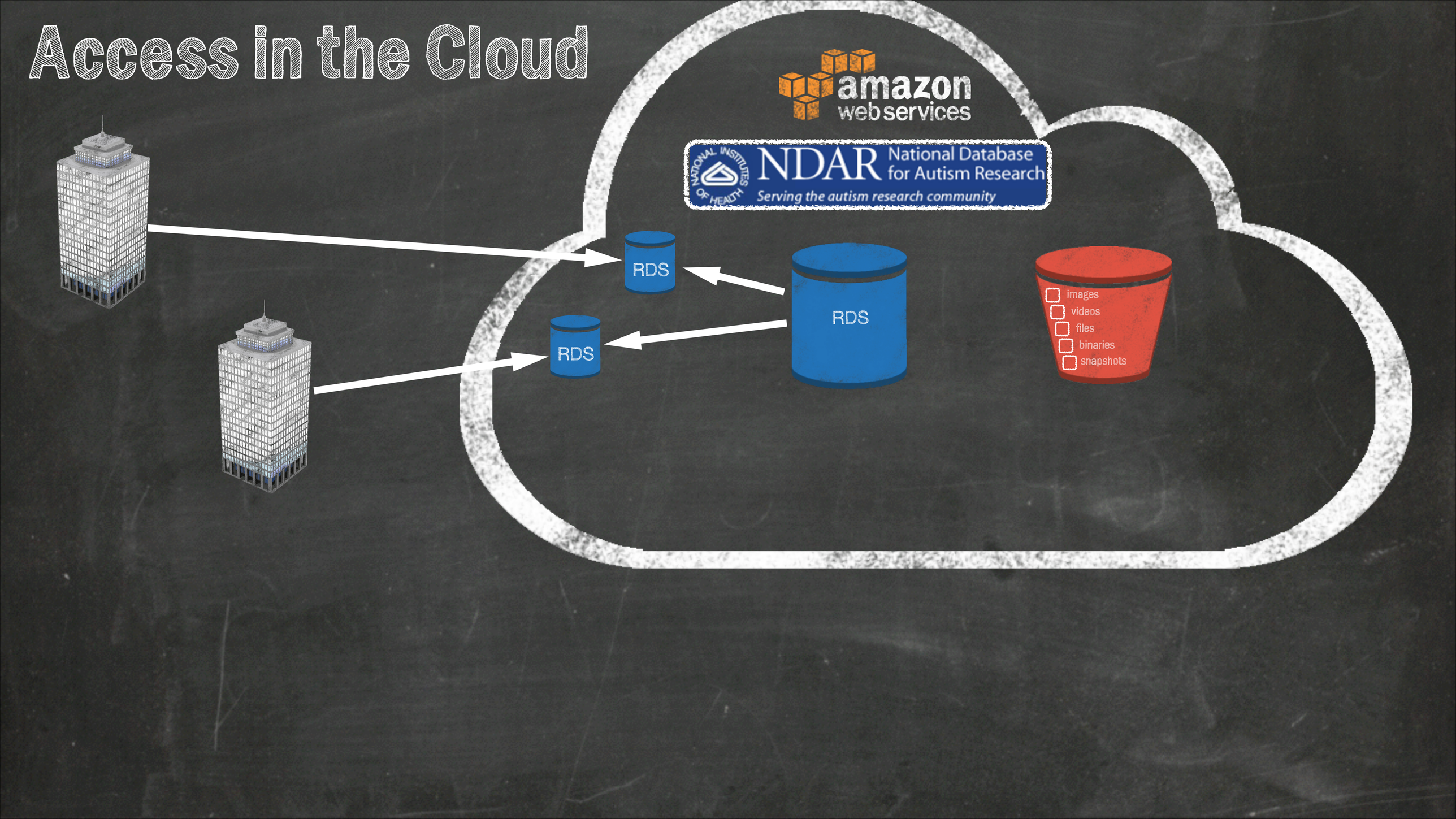
Serving the autism research community

RDS

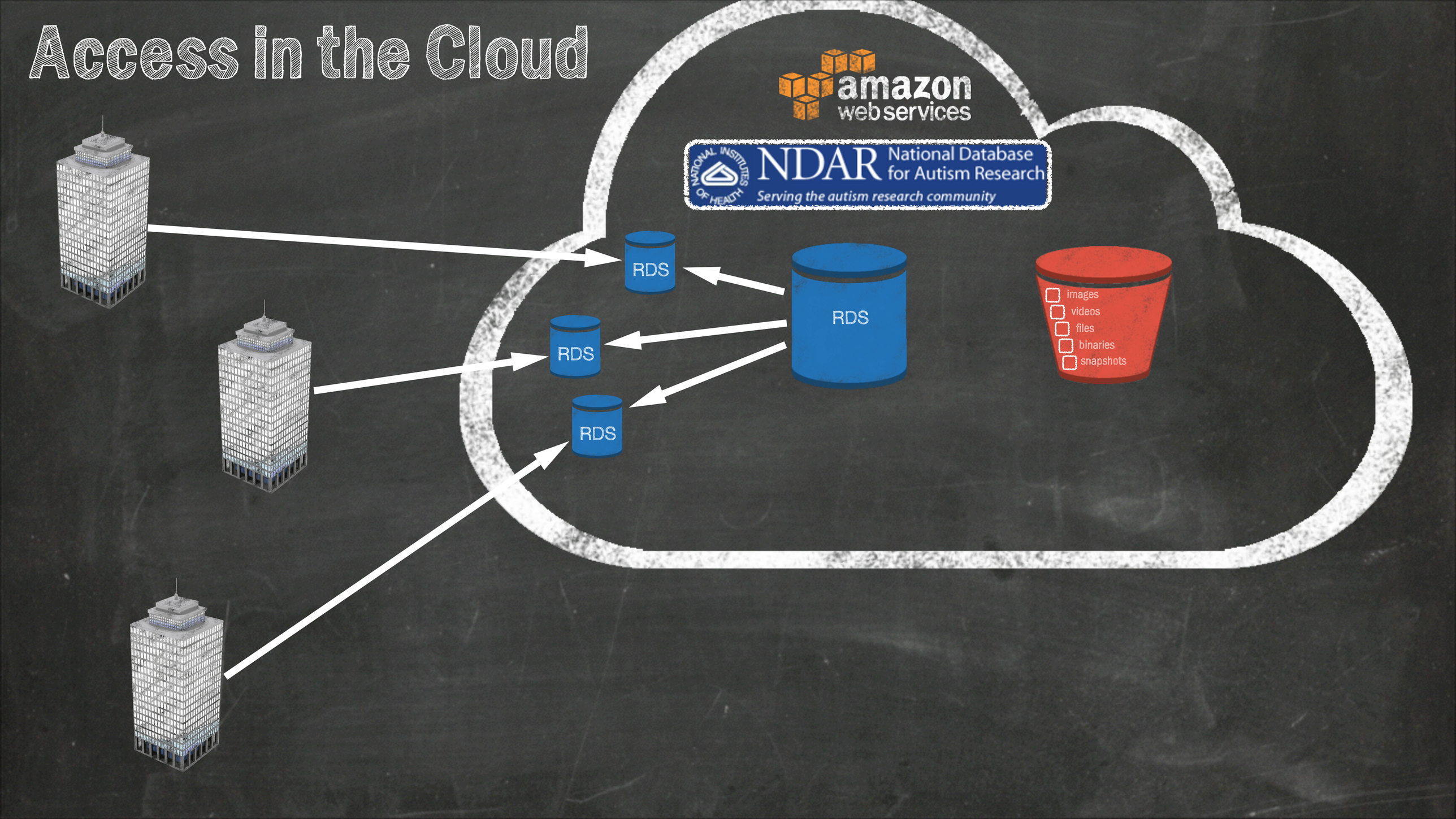
RDS

images
videos
files
binaries
snapshots

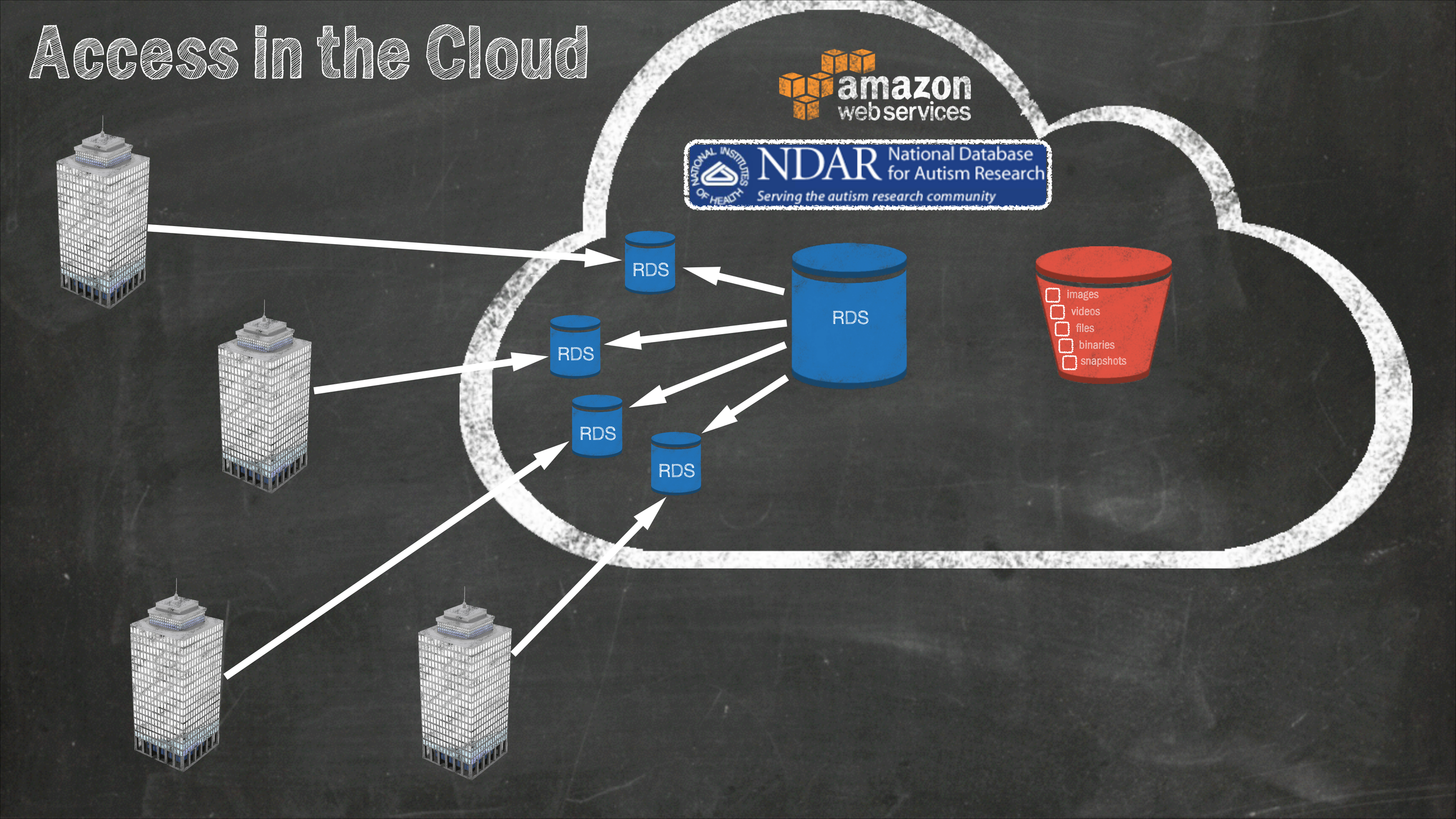
Access in the Cloud



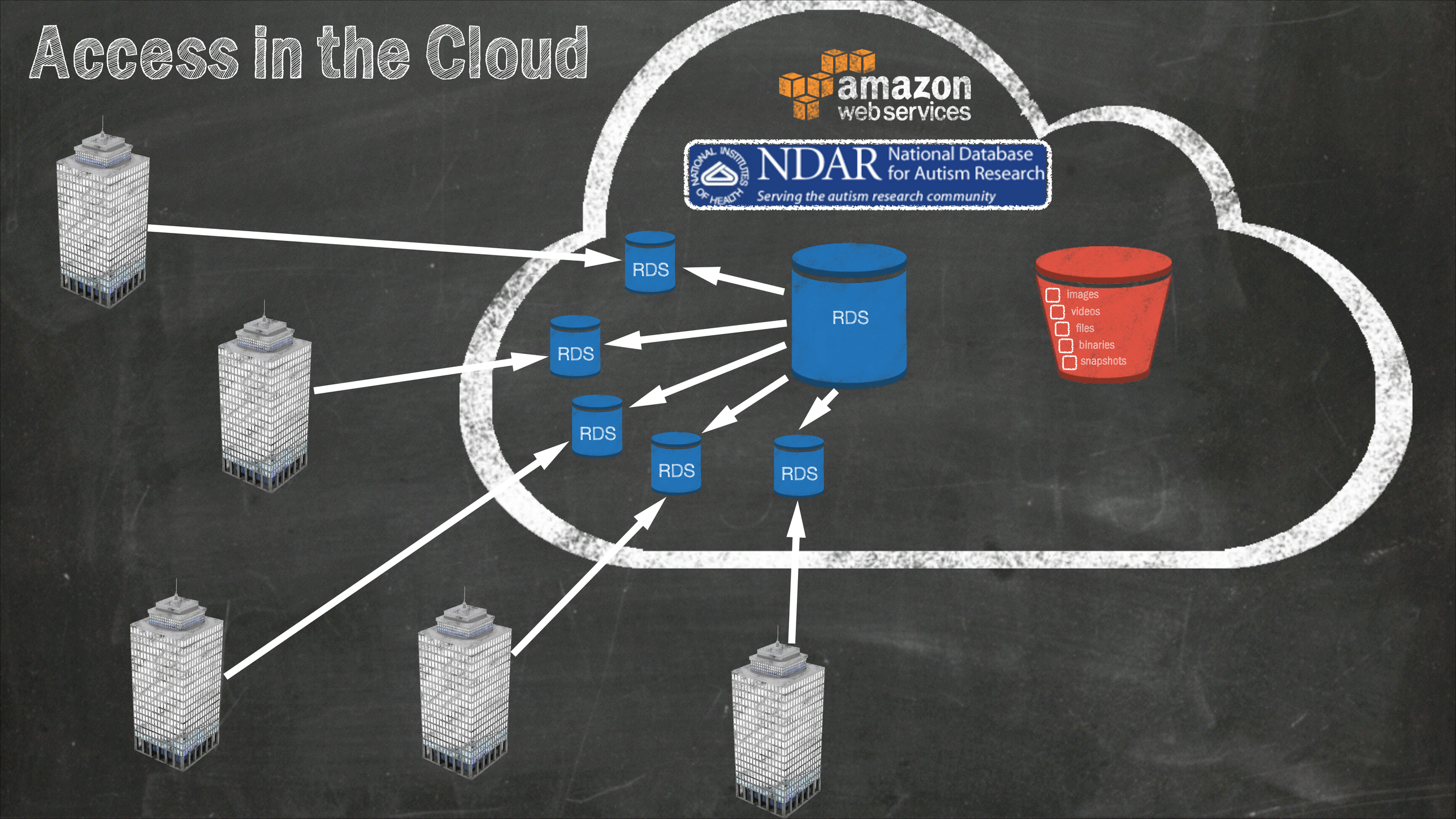
Access in the Cloud



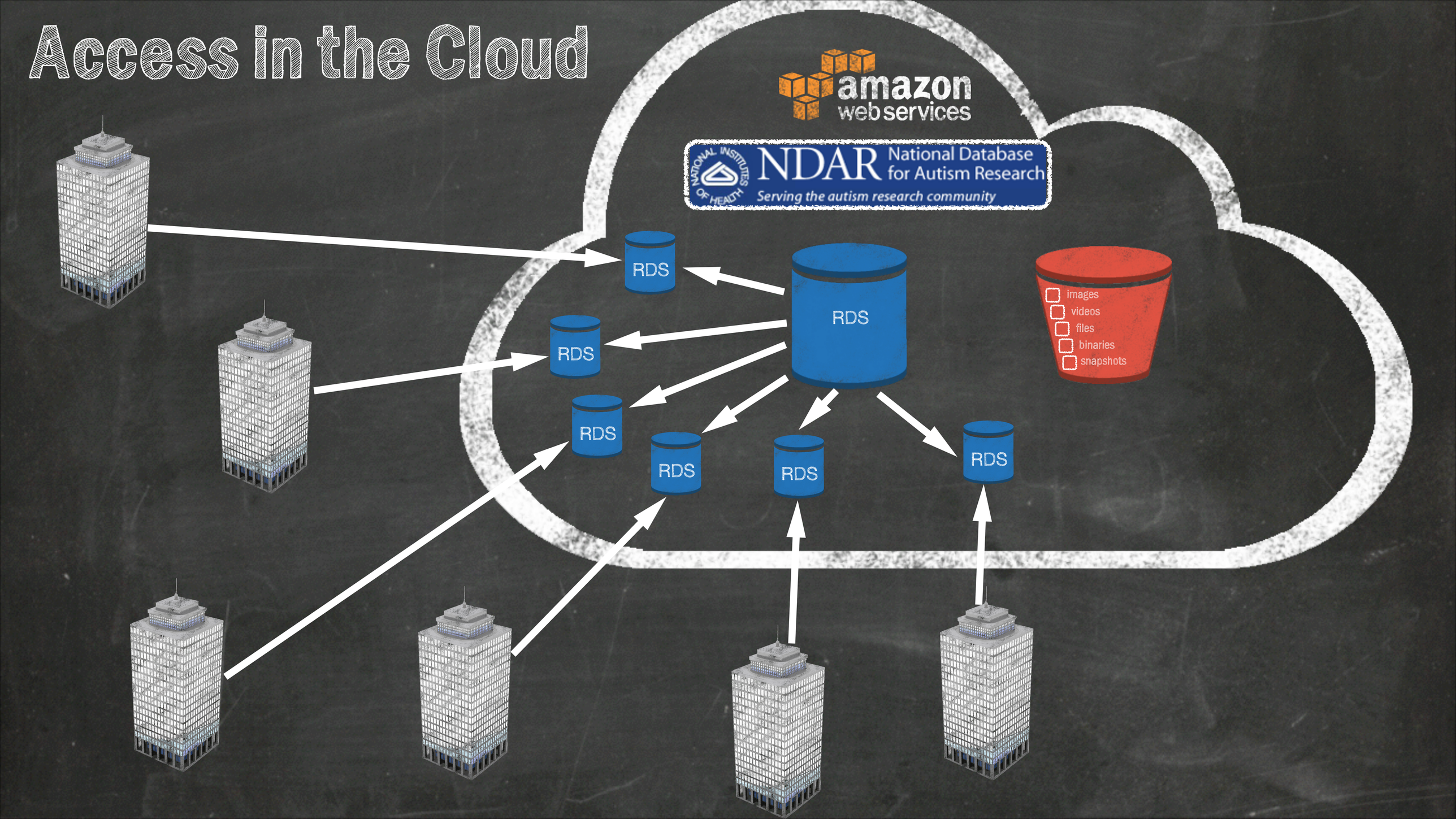
Access in the Cloud



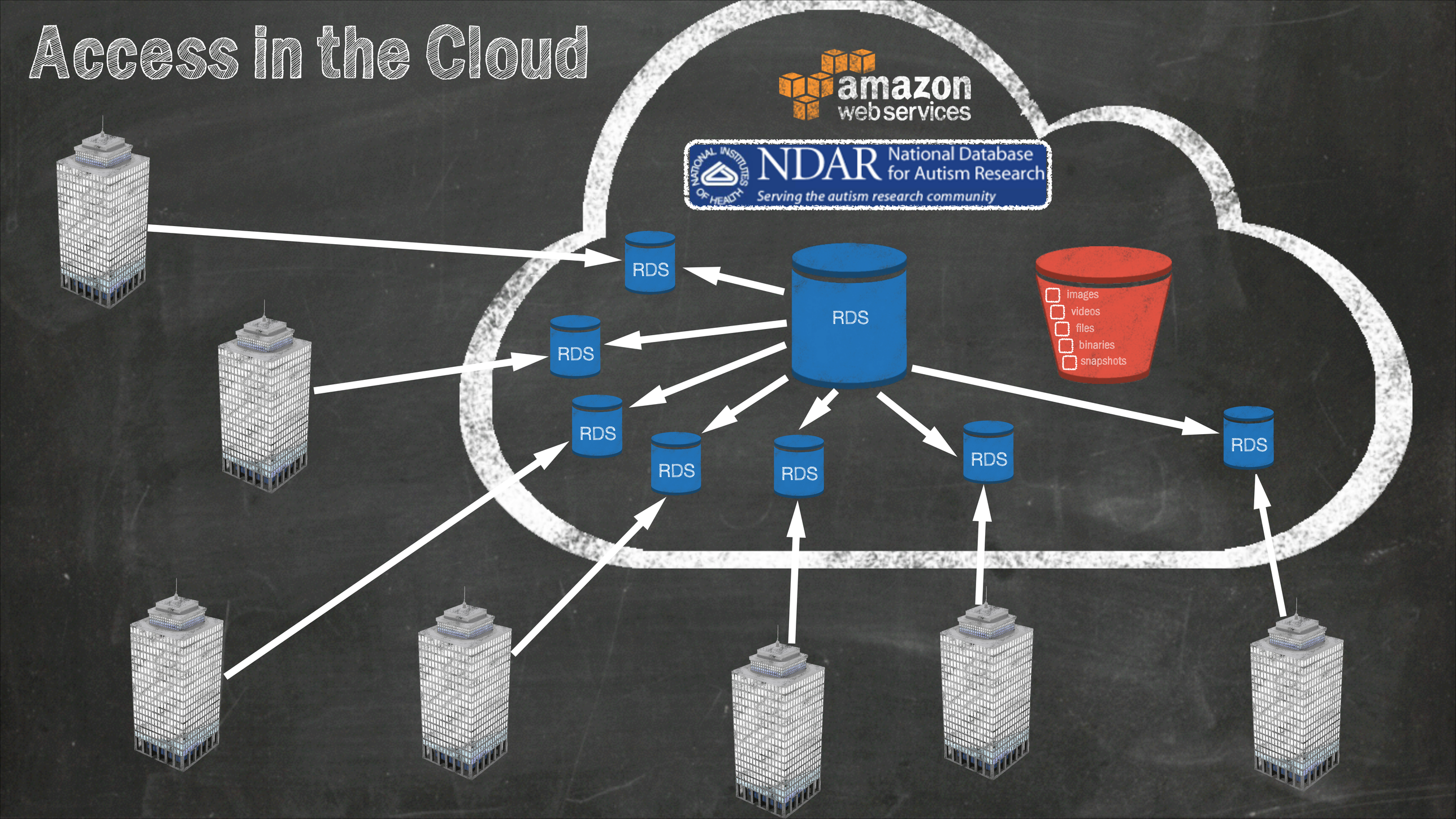
Access in the Cloud



Access in the Cloud



Access in the Cloud



Computation in the Cloud

Where is this Going?

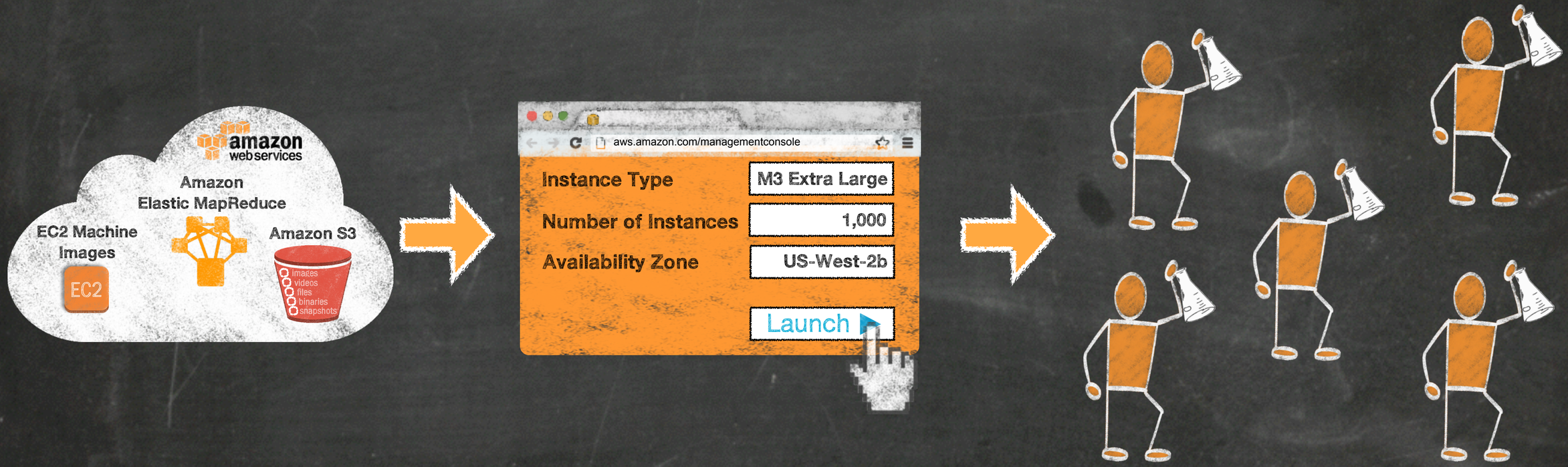
1. Researcher conducts experiment

Where is this Going?

1. Researcher conducts experiment
2. Experimental data and results uploaded to the cloud along with reproducible machine images

Where is this Going?

1. Researcher conducts experiment
2. Experimental data and results uploaded to the cloud along with reproducible machine images
3. Reviewers leverage cloud resources to reproduce and validate results.

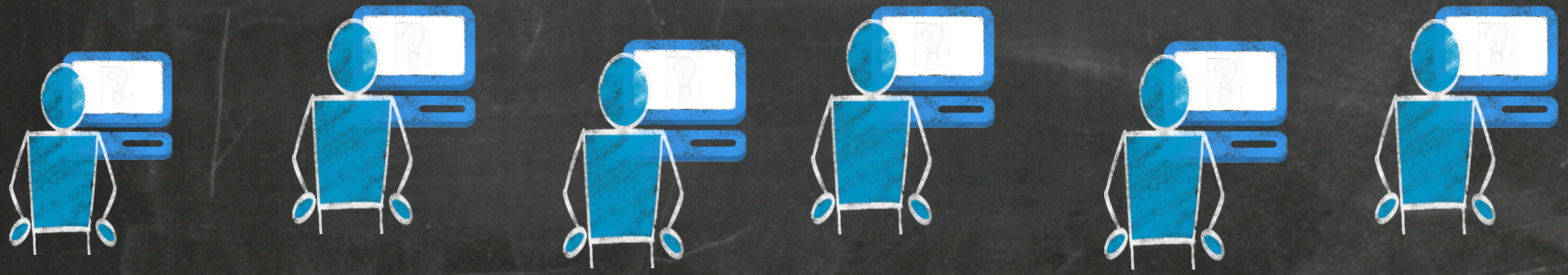


Where is this Going?

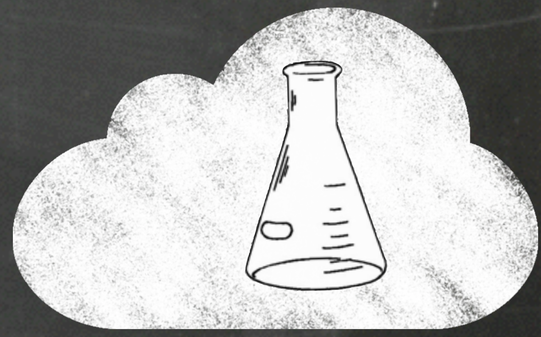
1. Researcher conducts experiment
2. Experimental data and results uploaded to the cloud along with reproducible machine images
3. Reviewers leverage cloud resources to reproduce and validate results.
4. Results published in a peer-reviewed journal, including references (e.g. DOIs) to cloud data and AMIs

Where is this Going?

1. Researcher conducts experiment
2. Experimental data and results uploaded to the cloud along with reproducible machine images
3. Reviewers leverage cloud resources to reproduce and validate results.
4. Results published in a peer-reviewed journal, including references (e.g. DOIs) to cloud data and AMIs
5. Other researchers use these resources as a jumping off point for further research, also publishing their results in the cloud.



Where is this Going?



1. Researcher conducts experiment
2. Experimental data and results uploaded to the cloud along with reproducible machine images
3. Reviewers leverage cloud resources to reproduce and validate results.
4. Results published in a peer-reviewed journal, including references (e.g. DOIs) to cloud data and AMIs
5. Other researchers use these resources as a jumping off point for further research, also publishing their results in the cloud.
6. Automated workflows re-run the original researcher's experiments on the new data using the original machine images. Interesting results trigger notifications and further review.

Thank You