SCIENCE-BASED AI/ML INSTITUTE

EDX++: Migrating EDX to the Cloud, Unlocking Next Generation Data Infrastructure

8/28/2023

Vic Baker,
K. Rose, J. Obradovich, D. McFarland, TJ Jones, J. Mondello, E. Dean, J. Sarle

AVN

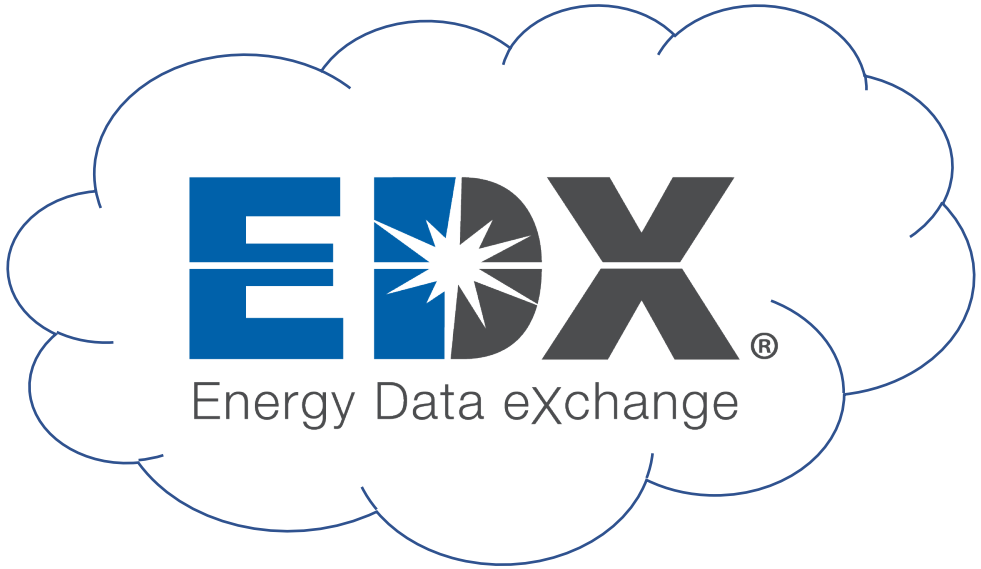
U.S. DEPARTMENT OF ENERGY | NATIONAL ENERGY TECHNOLOGY LABORATORY

Agenda

- Why go to the cloud?
- Why did we select Google Cloud Platform (GCP)?
- How did we migrate from On-Prem to Cloud?
  - Building a Cloud Native EDX
  - Architecting Cloud Infrastructure
- What's next?

Why Go to the Cloud?

- EDX growth is aligned with Cloud
  - Rapidly deploy new services
  - Data and Services Redundancy
    - Disaster Recovery and High Availability
    - Enhanced Uptime
  - Scalability for services
  - Storage
- Cloud Provider Managed Infrastructure
  - Guaranteed infrastructure uptime Service Level Agreements



Why GCP?

- Our team assisted with DOE HQ ATO evaluation of GCP via SmartSearch
- Conducted extensive deep dives into AWS and Azure for due diligence
  - *"Review & Recommendations of Cloud Platforms for FECM/NETL R&D – 2/9/2022"*
- GCP provides an ideal environment for cloud native, scalable innovation
- Google's Site Reliability Engineering (SRE) approach:
  - Use software as a tool to ***manage*** systems, ***solve*** problems, and ***automate*** operations tasks
- Google itself is Cloud Native -- they create cloud technologies the world uses

A Word on Infrastructure with Cloud

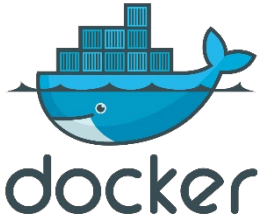
- Latest CPUs, GPUs available as soon as CSP can deploy
- Need a different CPU, more ram?
  - Modifications are immediately available to "right size" your compute
  - **Cut out months of bids, order delays, installation delays, setup delays**
- Turn compute machines on when you need them, off when you don't
- **This is development freedom** - to this presenter, one of the best features of cloud
  - Spin up what you want
  - When you want
  - Where you want
  - Automatically scale based on load

Some Definitions Before We Proceed

- **Zones:**
  - Geographically Dispersed Compute Facilities
  - Zones do not share a common point of failure
    - Independent power, storage, networking, compute, etc.
- **Region:**
  - Made up of three (3) or more zones
  - Regional disks are stored in two (2) zones
- **High Availability (HA):**
  - Two (2) or more zones
- **Disaster Recovery (DR):**
  - Two (2) or more regions

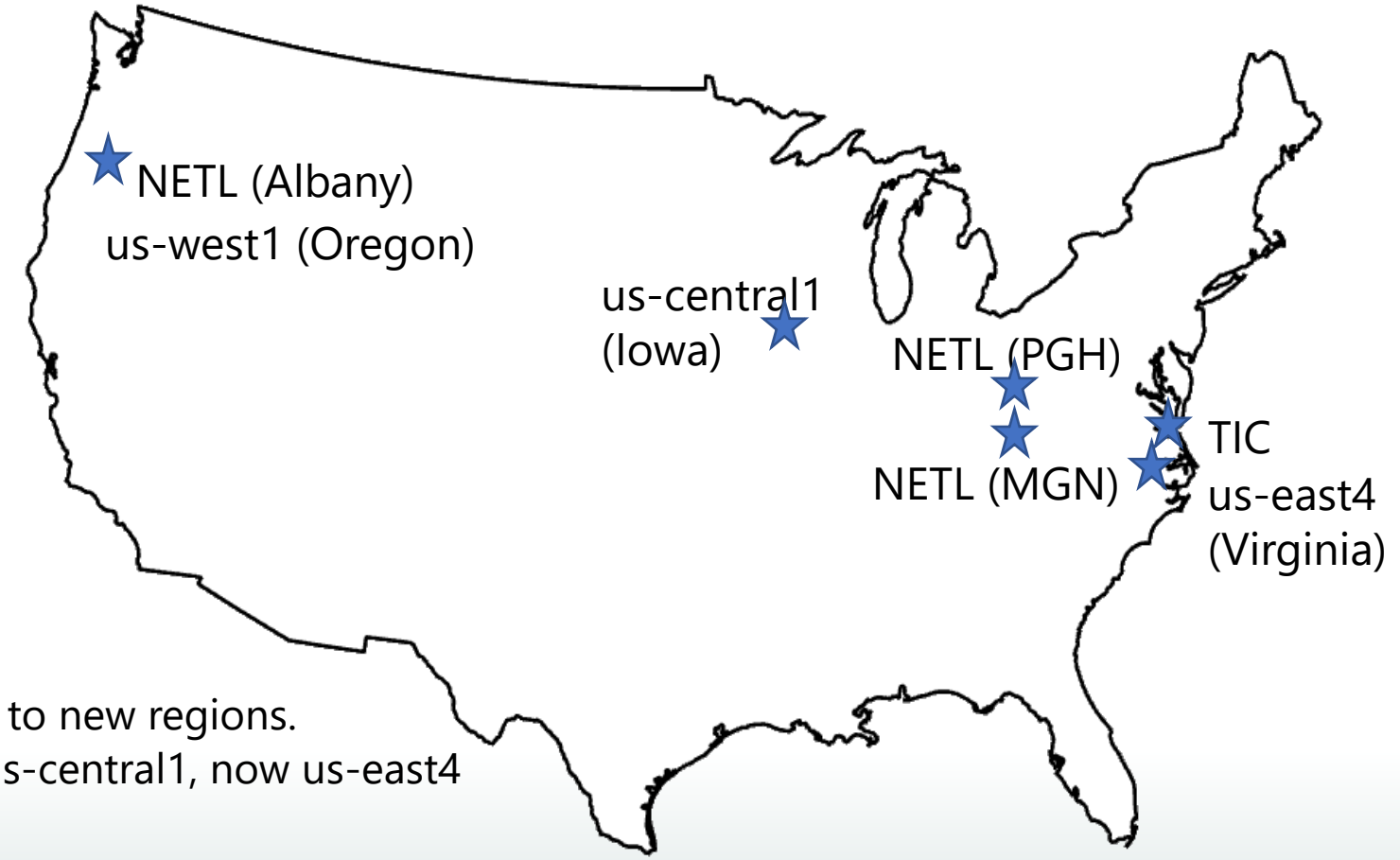
Navigating From On-Prem to Cloud

- **Follow "Cloud Native" pillars**
  - Microservices
  - Containerization and Orchestration
  - Dev/Ops agile delivery methodology
  - Continuous integration and continuous delivery (CI/CD)
  - Scalability
- **Container Orchestration:**
  - Migrate from Docker Swarm to Kubernetes
- **CI/CD:**
  - Automated Builds
  - Automated Deployments
  - Standardized structure of Git repositories
- **Infrastructure As Code:**
  - gcloud bash scripts
  - Terraform
- **Development:**
  - VS Code
  - Kubernetes Dev Starter

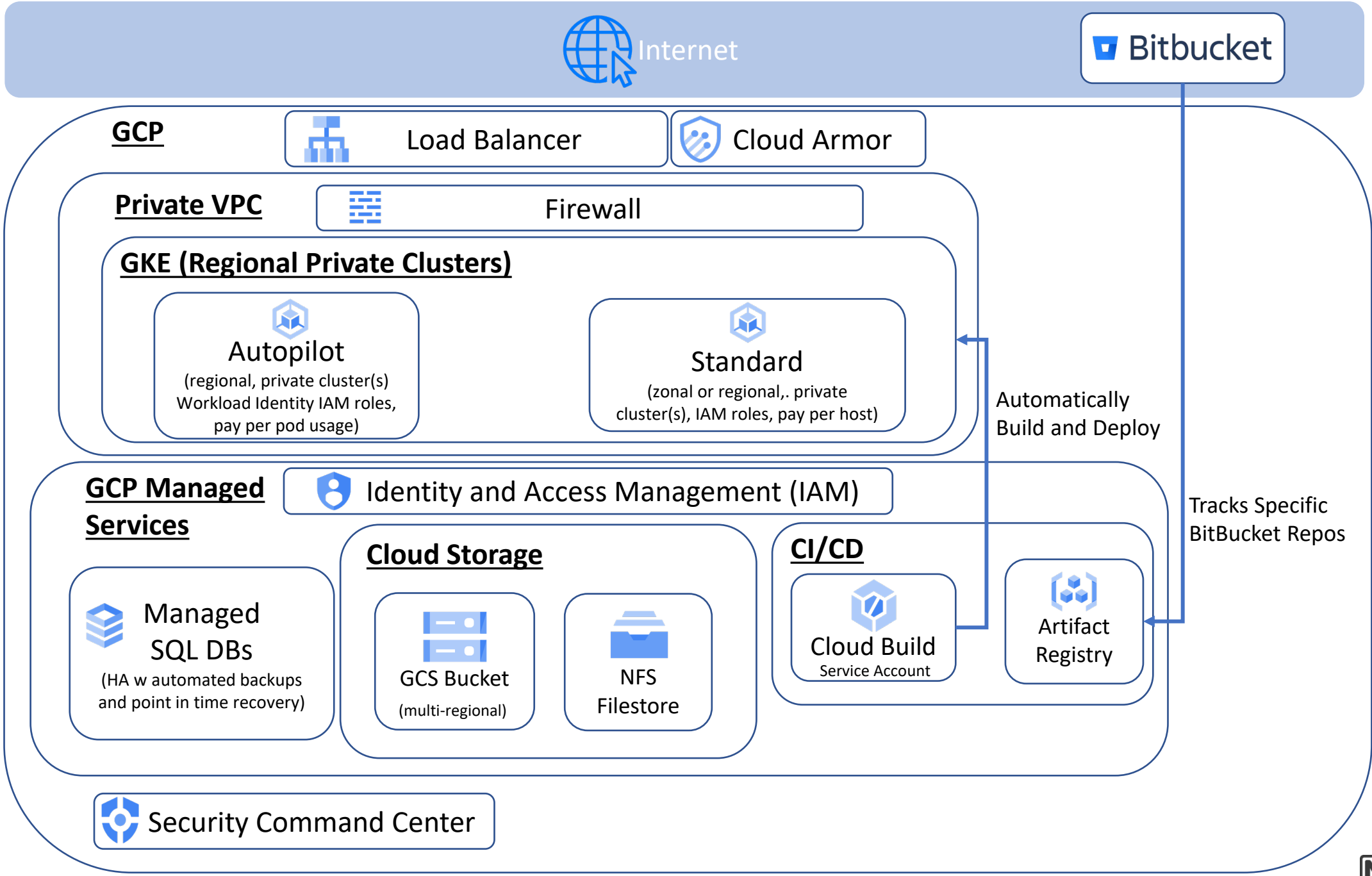


EDX Geographical Summary

- Kubernetes Services
  - us-east4
- Dual Region Bucket:
  - us-east4 and us-west1
- Virtual Private Cloud (network)
  - EDX-VPC:
    - us-east4 and us-west1 as subnets
- Managed Databases:
  - us-east4

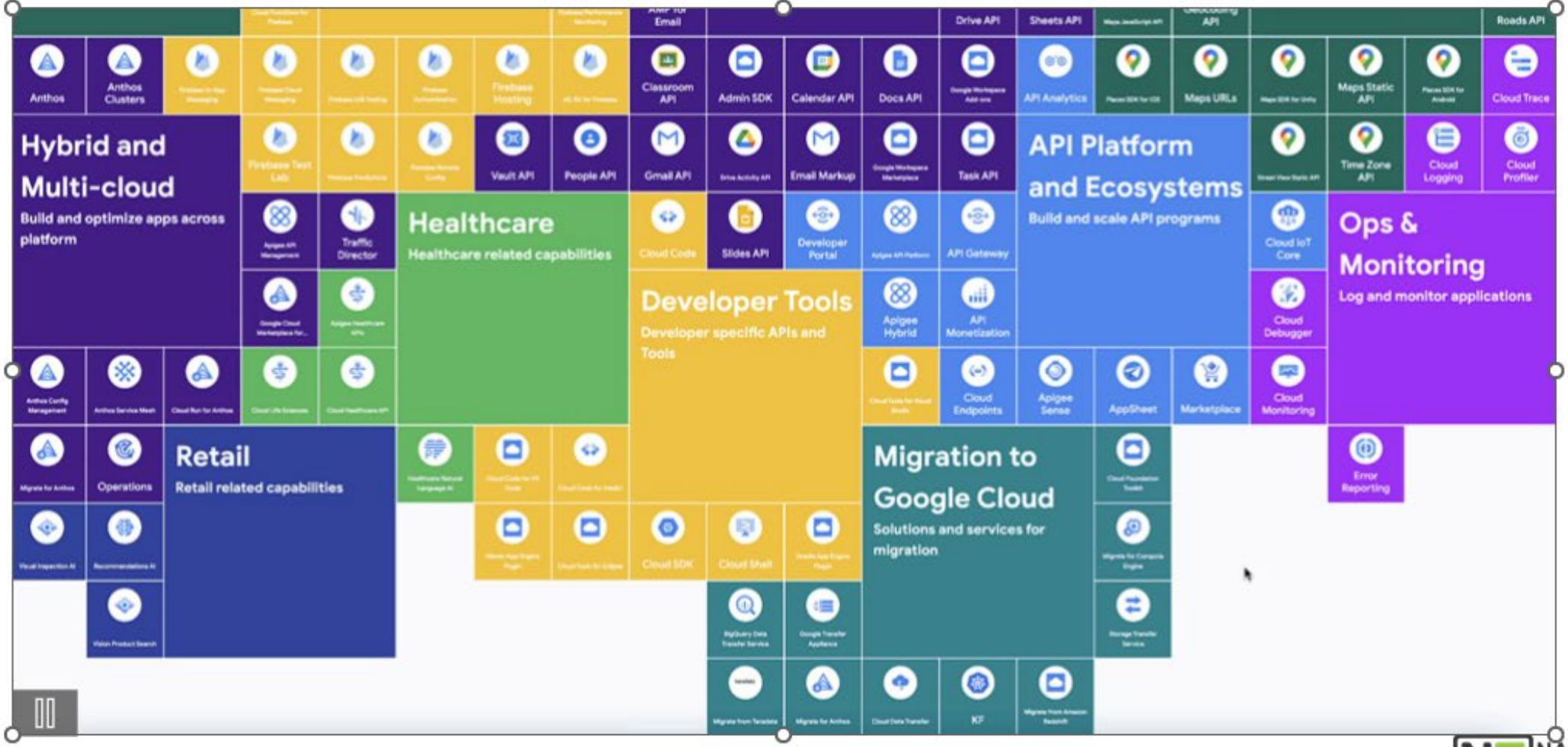


- Note:
  - GCP makes it easy to move deployments to new regions.
  - Initially EDX (Dev) was in us-east1, then us-central1, now us-east4

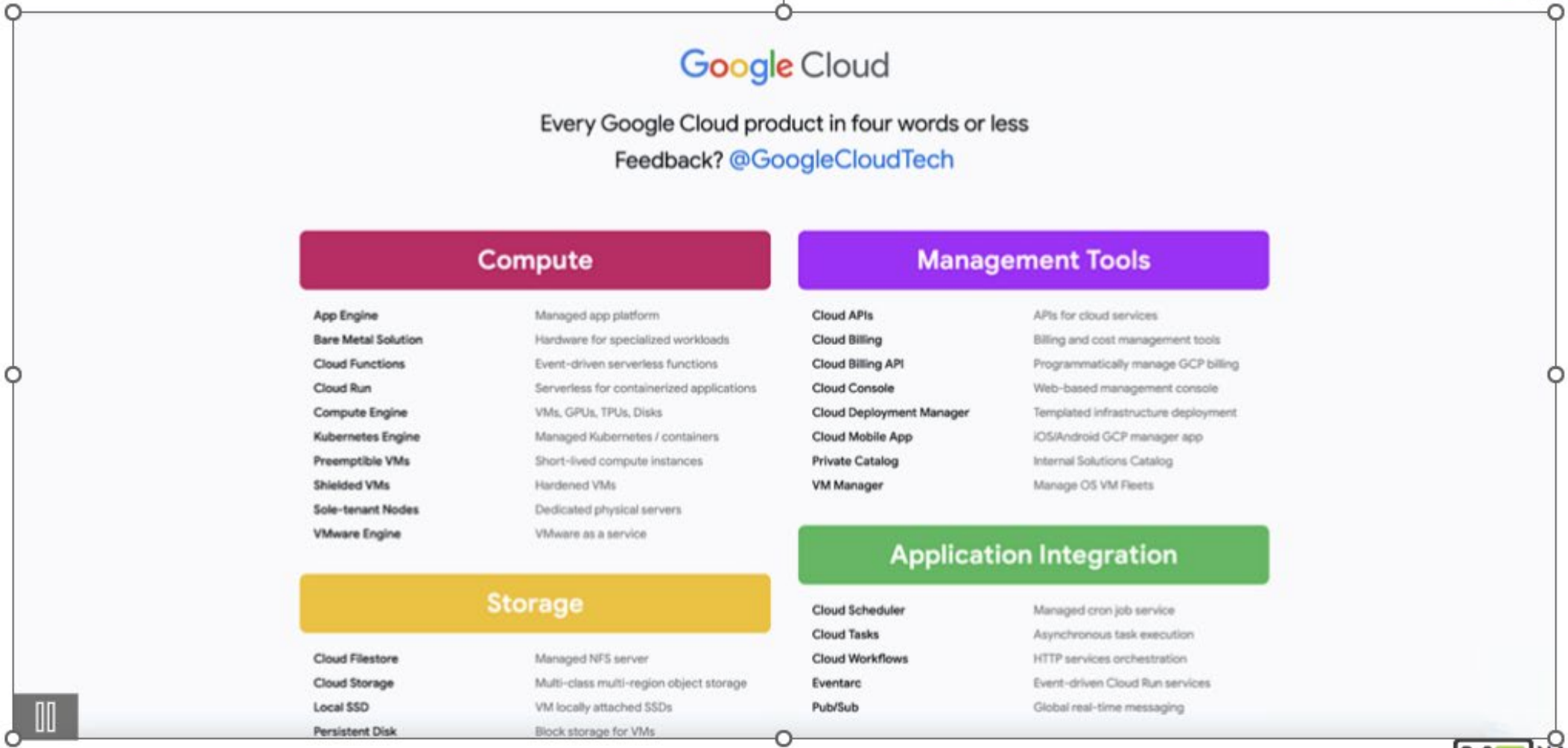

Internet
Bitbucket
GCP
Load Balancer
Cloud Armor
Private VPC
Firewall
GKE (Regional Private Clusters)
Autopilot
(regional, private cluster(s) Workload Identity IAM roles, pay per pod usage)
Standard
(zonal or regional,. private cluster(s), IAM roles, pay per host)
Automatically Build and Deploy
Tracks Specific BitBucket Repos
GCP Managed Services
Identity and Access Management (IAM)
Cloud Storage
CI/CD
Managed SQL DBs
(HA w automated backups and point in time recovery)
GCS Bucket
(multi-regional)
NFS Filestore
Cloud Build
Service Account
Artifact Registry
Security Command Center

GCP Services Summary

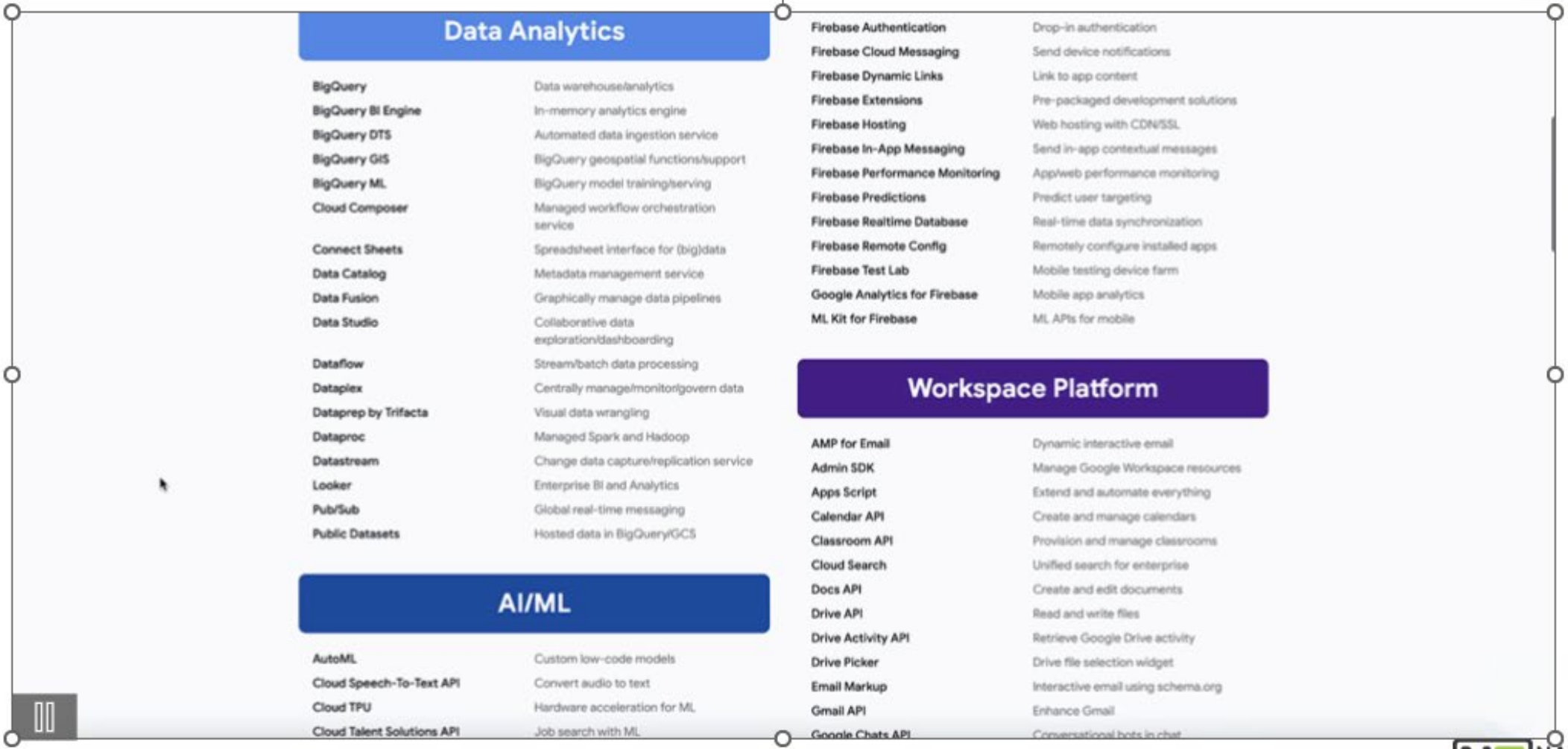
GCP Services Summary



GCP Services Summary



GCP Services Summary



Data Analytics

BigQuery	Data warehouse/analytics
BigQuery BI Engine	In-memory analytics engine
BigQuery DTS	Automated data ingestion service
BigQuery GIS	BigQuery geospatial functions/support
BigQuery ML	BigQuery model training/serving
Cloud Composer	Managed workflow orchestration service
Connect Sheets	Spreadsheet interface for (big)data
Data Catalog	Metadata management service
Data Fusion	Graphically manage data pipelines
Data Studio	Collaborative data exploration/dashboarding
Dataflow	Stream/batch data processing
Dataplex	Centrally manage/monitor/govern data
Dataprep by Trifacta	Visual data wrangling
Dataproc	Managed Spark and Hadoop
Datastream	Change data capture/replication service
Looker	Enterprise BI and Analytics
Pub/Sub	Global real-time messaging
Public Datasets	Hosted data in BigQuery/GCS

AI/ML

AutoML	Custom low-code models
Cloud Speech-To-Text API	Convert audio to text
Cloud TPU	Hardware acceleration for ML
Cloud Talent Solutions API	Job search with ML

Firebase Authentication	Drop-in authentication
Firebase Cloud Messaging	Send device notifications
Firebase Dynamic Links	Link to app content
Firebase Extensions	Pre-packaged development solutions
Firebase Hosting	Web hosting with CDN/SSL
Firebase In-App Messaging	Send in-app contextual messages
Firebase Performance Monitoring	App/web performance monitoring
Firebase Predictions	Predict user targeting
Firebase Realtime Database	Real-time data synchronization
Firebase Remote Config	Remotely configure installed apps
Firebase Test Lab	Mobile testing device farm
Google Analytics for Firebase	Mobile app analytics
ML Kit for Firebase	ML APIs for mobile

Workspace Platform

AMP for Email	Dynamic interactive email
Admin SDK	Manage Google Workspace resources
Apps Script	Extend and automate everything
Calendar API	Create and manage calendars
Classroom API	Provision and manage classrooms
Cloud Search	Unified search for enterprise
Docs API	Create and edit documents
Drive API	Read and write files
Drive Activity API	Retrieve Google Drive activity
Drive Picker	Drive file selection widget
Email Markup	Interactive email using schema.org
Gmail API	Enhance Gmail
Google Chats API	Conversational bots in chat

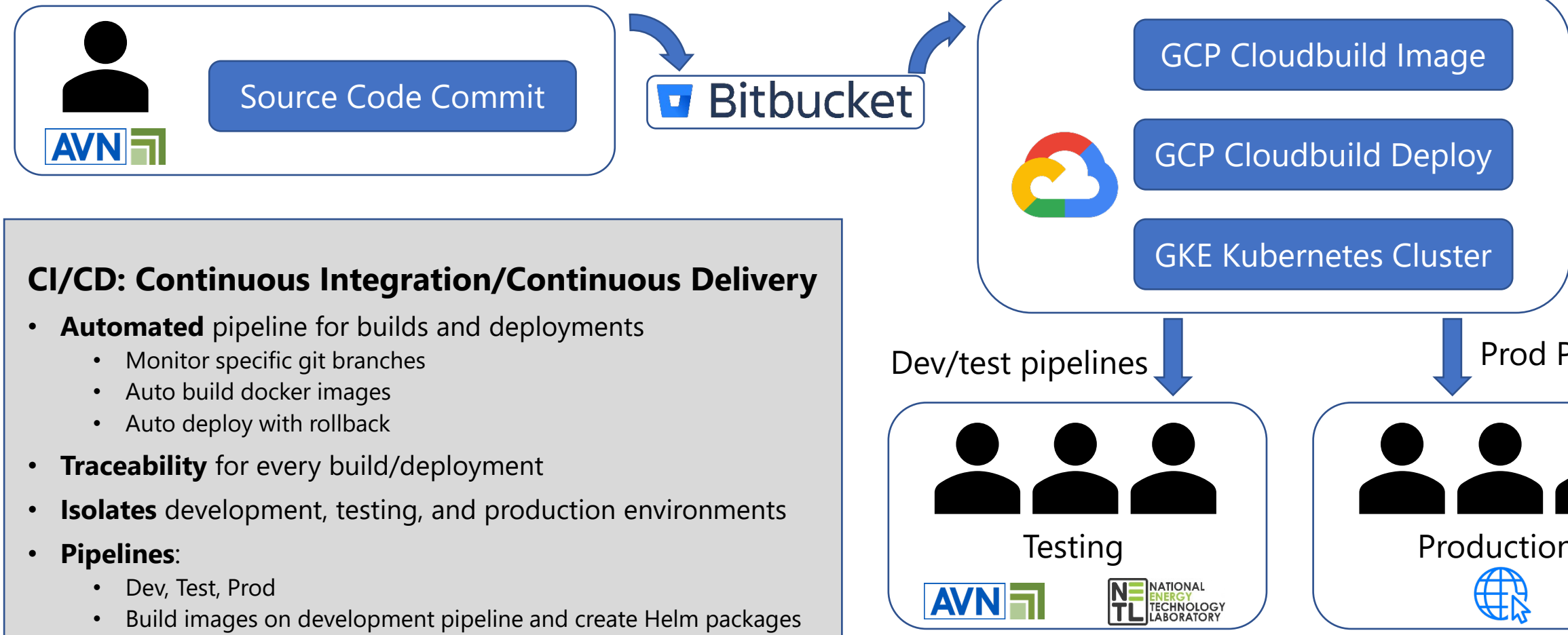
GCP Benefits for EDX Dev/Ops

- **Security**
  - Artifact Registry Vulnerability Scanner
  - Security Command Center and Cloud Armor
  - Data Security – encryption in transit and at rest
  - Secrets via Secret Manager
- **CI/CD**: Automated Builds and Deployments tied to commits!
- **Unlimited storage** via Buckets
- **Stability**: reliable infrastructure and networking
- **GCP Network**: Fast, reliable, zonal / regional / global
- **Deep Bench**: broad tech community
- **Infrastructure**: Managed databases, CPU / RAM / GPU options, etc.

Security

- **Data Security**:
  - File data stored on buckets across 2 regions
  - Managed Databases with High Availability and Point in Time recovery
  - GKE Backups of running services
  - GKE Autopilot Persistent volumes that are regional (multi-zones)
- **Secrets**:
  - Google Secret manager
    - We inject secrets into deployments via CI/CD automation scripts
- **Process**:
  - Dev / Test / Production Environments
    - Build and deploy via automation
    - Deployment traceability at each stage via git commits or pull requests
    - Role attrition: Reduce backend user access transitioning from dev > test > prod

(Simplified) Build & Deployment Pipelines



CI/CD: Continuous Integration/Continuous Delivery

- **Automated** pipeline for builds and deployments
  - Monitor specific git branches
  - Auto build docker images
  - Auto deploy with rollback
- **Traceability** for every build/deployment
- **Isolates** development, testing, and production environments
- **Pipelines**:
  - Dev, Test, Prod
  - Build images on development pipeline and create Helm packages
  - Promote Helm packages on test and prod pipelines.

https://edx.netl.doe.gov/sami/

EDX Core Changes

- **File storage and retrieval via GCS bucket(s):**
  - GCS buckets are 'limitless' in size, secure, cost effective, allow lifecycle policies, fast, and can span multiple regions (EDX is dual region)
  - We've built scalable Kubernetes services for GCS read and write operations
- **Scalability:**
  - Split EDX into a collection of scalable microservices that can each scale as needed
- **Automated builds and deployment:**
  - GCP Cloud Build
  - Helm Charts for service configuration
- **Reliability:**
  - Utilizing managed databases (High Availability with “Point in Time” recovery)
  - Using GKE Autopilot (regional) – services spawn across multiple geographic zones
  - File storage using GCS Buckets spanning dual regions – fast, replicated, fault tolerant

Infrastructure As Code

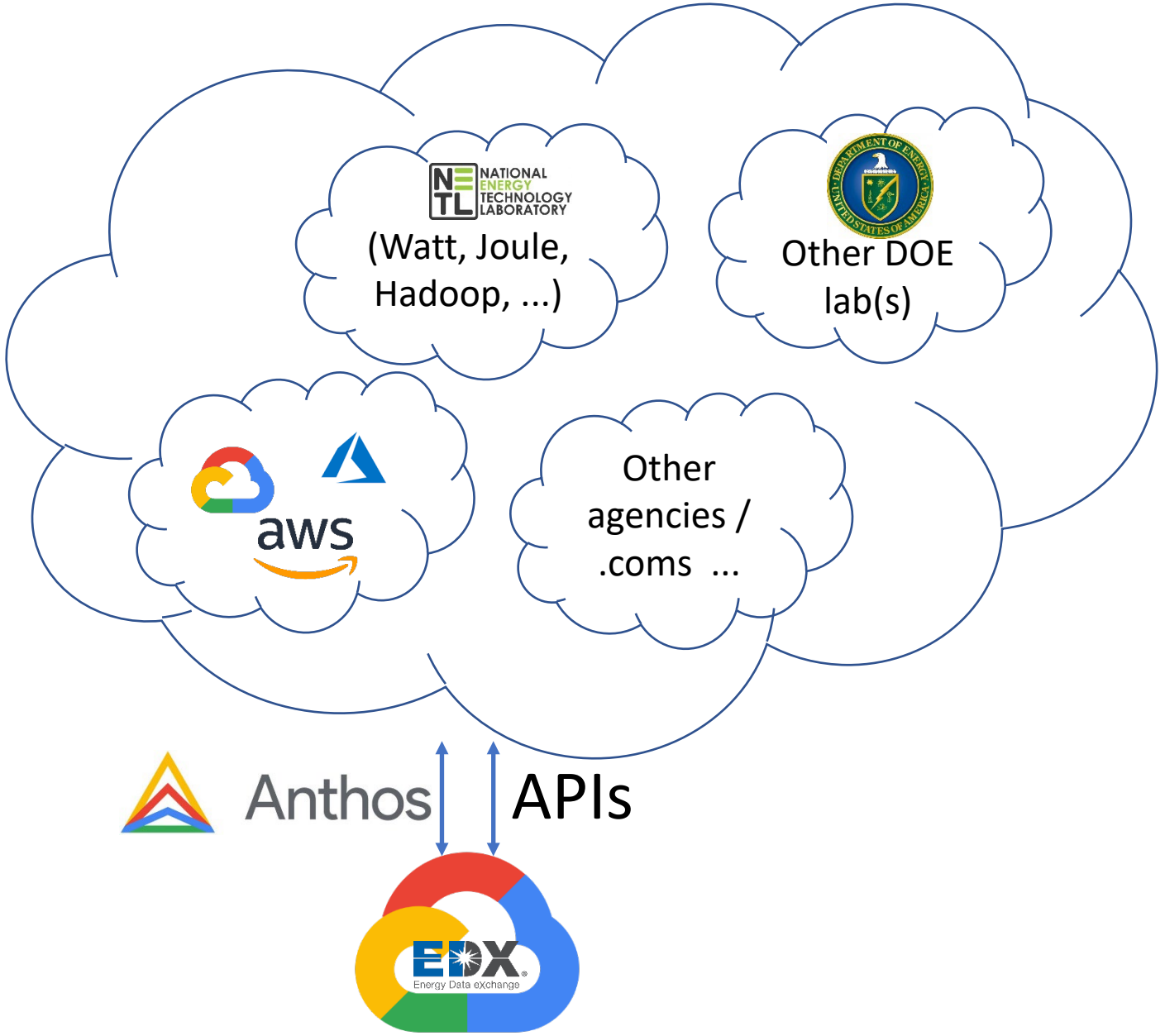
- Repeatable automation for configuring networks & subnets, Kubernetes clusters, databases, etc.
- Configure dev/test/prod environments with consistent architecture
- Useful for blue/green deployments
- Terraform vs gcloud:
  - Terraform configurations vs gcloud (bash)
  - Both have benefits / drawbacks
  - We primarily use gcloud bash scripts
  - We have limited terraform usage

Helm: Kubernetes Package Manager

- "Helm is the best way to find, share, and use software built for Kubernetes"
  - http://helm.sh
- Helm is a configuration suite for Kubernetes deployments (secrets, mounts, etc.)
- EDX container orchestration was migrated from Docker Swarm to Kubernetes
- EDX services have been migrated to Helm
- Cloudbuild Automation deploys EDX updates via Helm

Multi-Environment: Cloud + On Prem

- **EDX on GCP** ideally positioned for hybrid computing
- **GCP Anthos**
  - manage Kubernetes clusters across platforms
  - create hybrid clusters, deploy and manage workloads
  - *"I want to deploy my Kubernetes ML application to AWS, or GCP, or Watt … or a hybrid GCP + Azure... etc."*
- **APIs** allow us to create, manage, access services for additional workloads



Connecting to Compute Services

- **EDX on GCP** allows compute service connectivity flexibility.
- **Utilizing APIs** allow us to create, manage, access services for additional workloads
- **AI/ML integration** allows us to create customized language models, analytics capabilities
- **Scalable, On Demand Compute Infrastructure** enables cloud-based notebooks and processing workflows

Build, Run, Share Custom ML Workflows

SME: Run & monitor results of ML Pipelines

- Pull data from EDX
- Publish Transformed Data

- Publish and Share models stored on Mlflow

Developer: Create & publish ML pipelines

- Store models trained in Kubeflow Pipelines
- Serve models stored in Mlflow

- Store and share models
- Compare Model Versions
- Compare experiment metrics

What's Next

- **EDX Production Deployment on GCP Summer 2023**
  - Development nearing completion
  - Configuring Test and Production environments on GCP
  - NETL ATO for GCP pending review and signature

- **Ongoing Efforts:**
  - Additional Fast Bandwidth integrations (e.g. ESnet)
  - EDX and SmartSearch integration
  - Architecting integration of additional compute tools
  - Kubernetes Dev Starter (KDS)

Contact Us

VISIT US AT: **www.NETL.DOE.gov**

@NETL_DOE

@NETL_DOE

@NationalEnergyTechnologyLaboratory

CONTACT:

SAMI – SAMI (doe.gov)

EDX Support, edxsupport@netl.doe.gov

Vic Baker, Vic.baker@netl.doe.gov, vic.baker@AVNcorp.com

Disclaimer: *This work was funded by the United States Department of Energy, National Energy Technology Laboratory, in part, through a site support contract. Neither the United States Government nor any agency thereof, nor any of their employees, nor the support contractor, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.*