

#### AWS RESEARCH - $1^{\text{ST}}$ OCTOBER 2024

# How to Use AWS Cloud in the Green Economy Era for Research

AWS Education & Research Real Piscitelli, PhD, MBA - <u>piscitr@amazon.com</u>

© 2023, Amazon Web Services, Inc. or its affiliates

# **AWS Global Infrastructure**

AWS REGIONS, LOCAL ZONES, EDGE LOCATIONS, AND GLOBAL BACKBONE

#### R E G I O N A L E X P A N S I O N

Available today: 32 RegionsComing soon: 5 Regions

### **Trends We have Observed**









Research is strategic for growth, recruiting and knowledge New and different types of users need large scale computation

Sustaining onpremise while keeping pace with technology Increasing grant

requirements around secure research



## Why using AWS Cloud for Research



# A data-driven approach to sustainability strategy, enabled by cloud computing



Sustainability

the cloud

Delivering a sustainable IT fleet – taking advantage of the cloud and AWS efficiency through migration Optimizing workloads on AWS with the Sustainability pillar of the Well-Architected Framework

Sustainability IN

the cloud

Sustainability THROUGH the cloud

Deploying cloudbased solutions and advisory support to accelerate customer sustainability objectives



# Sustainability OF the Cloud

**Opportunity to reduce impact** through migration



# **Europe: Carbon reduction opportunity**

AWS enables up to 99% lower carbon emissions when compute-heavy workloads are optimized on AWS vs. on-premises

**94%** lower carbon emissions when compute-heavy workloads are run asis on AWS vs. on-premises, due to:

**36%** from efficient hardware with improved utilization

**25%** from power and cooling efficiencies

33% from additional carbon-free energy procurement

aws



Find the report at https://sustainability.aboutamazon.com/carbon\_reduction\_aws.pdf



aws

### Reducing carbon emissions across our global operations data centers constructed using either or both lower-carbon concrete and steel

We use hydrotreated vegetable oil, a renewable fuel, to power our backup generators. HVO can reduce emissions by up to 90% over the fuel's lifecycle compared with fossil diesel.

Lower lifecycle impacts of silicon-based devices from materials to manufacturing

In 2023, electricity consumed in 22 AWS Regions was attributed to 100% renewable energy



aws

### Embracing a circular economy

**ECONOMY** Design. We eliminate excess materials, increase recycled and bio-based content, and plan for reuse from the start.

Operate. We extended the life of our servers and networking equipment from 5 years to 6.

Recover. In 2023, 14.6M hardware components were diverted from landfills by being recycled or sold into the secondary market for reuse.



# AWS water positive commitment

By 2030, AWS will return more water to communities than we use in our direct operations.

As of 2023, AWS is **41%** of the way toward water positive.



# Water stewardship

Water efficiency. 0.18 liters of water per kilowatt-hour water use efficiency for AWS data centers

Sustainable water sources. 24 data centers globally use recycled water for cooling

Water reuse. Reuse discharged water from our data centers in communities

Water replenishment. Amazon returns **3.9** billion liters of water to communities each year from replenishment projects completed or underway



# Sustainability IN the cloud

**Optimizing for sustainability Sustainable Pillar of the Well-Architected Framework Customer carbon footprint tool** 





### Sustainability Pillar of the Well-Architected Framework



aws

# Focus domains of the AWS Well-Architected Sustainability Pillar



aws



aws

# Graviton, Inferentia, and Trainium

Graviton-based Elastic Compute Cloud instances use up to 60% less energy than comparable Amazon EC2 instances

Models built on Trainium result in energy-consumption reductions of up to 25% versus comparable instances

Our Inferentia2 machine learning accelerator delivers up to 50% higher performance per watt

### New LLM development and refinement UAE Technology Innovation Institute



#### **CHALLENGE**

aws

TII researchers needed an easy way to develop, iterate, and distribute Falcon-40B LLM. Experimenting at such massive scale was slowed by limited computational resources and complex onprem infrastructure. Researchers sought more flexibility to efficiently enhance, evaluate, and provide access to Falcon under dynamic capacity requirements to empower their team's foundation model research.

#### **SOLUTION**

TII utilized SageMaker's managed machine learning infrastructure and tools to train, host, and deploy versions of Falcon with greater speed and experimentation. SageMaker provided on-demand access to training clusters along with modular deployment options for easy distribution to varying client workloads—key for customizing model access.

#### OUTCOME

✓ TII successfully implemented Falcon-180B by using SageMaker and custom innovation. This allows TII's contribution within UAE's 2031 National AI Strategy, fostering economic growth and social progress. Releasing UAE's Falcon 180B, World's Top-Ranked Open Source AI Model will further encourage AI academic research and scientific collaboration.

## Large Models built on AWS



Clibrain develops Lince Zero—the first large language model (LLM) optimized for Spanish using Amazon SageMaker Stability AI will build AI models on compute clusters with thousands of GPU or <u>AWS Trainium</u> chips, reducing training time and cost by 58%

stability.ai



BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance, utilizing AWS infrastructure services

# **Cloud Bursting**

**CLOUD BURSTING ON AWS** 



# **Optimize for sustainability: Quick wins**



Improve energy efficiency by switching to Graviton-based instances



Choose serverless when possible



Integrate AWS Instance Scheduler to shut down and terminate when not in use



Use AWS Compute Optimizer for right-sizing recommendations of workloads



# AWS customer carbon footprint tool



**Calculate** carbon emissions generated from your AWS workloads

**Understand** historical carbon footprint and review changes in emissions over time

**Forecast** reflects Amazon's achievement of 100% renewable energy in 2023

### Sustainability THROUGH the Cloud Leverage AWS technologies and data services to solve sustainability challenges



# **AWS Solutions Approach**

الله المعالم ا

Customer Use Case Identification & Prioritization





# **AWS Science Hub**



# Solutions and services for the entire research process



# Sustainability solutions are powered by data

Data is diverse, growing exponentially, and used by many applications. AWS storage and analytics services and data programs can help.

#### Open Data Sponsorship Program

#### **Amazon Sustainability Data Initiative (ASDI)**

AWS Data Exchange

Image from Landsat 8 satellite, courtesy of the U.S. Geological Survey



2023, Amazon Web Services, Inc. or its affiliates. All rights re

### **Open data on AWS**

AWS hosts a variety of public datasets to lower the cost and improve the speed of research.

#### https://registry.opendata.aws/



#### **Examples**

- 1000 Genomes Project
- The Cancer Genome Atlas
- International Cancer Genome Consortium
- Landsat 8
- Common Crawl
- SpaceNet
- OpenStreetMaps

#### ... Regularly updated

# FastMRI: Accelerating MRI Scans with AWS and AI

- Dataset: 7 million + brain and knee MRI scans
- Global Impact: Users from 79 countries, 11 publications
- Results: Up to 98% similarity to full scans in 15% of the time

Fully sampled

#### Undersampled





#### Amazon Open Data Sponsorship Program



https://avis.amazon.com/blogs/publicsector

in <u>ne sea principale evitare</u> a constituit de se a se se se su de la subaring research



# AWS investments supporting research and education



#### TRAININGS

AWS Training Academy Demo/Immersion days

#### SUPPORT

Technical Support AWS Research Scientists Grant Proposals

#### INCENTIVES

Proof of concepts AWS Cloud Credits for Research Amazon Open Data Sponsorship Program Data Egress Waiver

© 2024, Amazon Web Services, Inc. or its affiliates.

## **Voice of the researcher: requirements**



#### **Reduce time to run**

Access research environments in minutes



### Maintain consistent security, compliance, and governance





#### **Resilient data ingestion** Ingest datasets at scale



#### Integrated visualization

Post-analysis visualization and insights

#### **Spend controls**

Cost visibility, centralized budgeting, and chargeback management



#### End-to-end resource provisioning

Storage, compute (AWS ParallelCluster), and visualization from single pane of glass



#### **Universally accessible** Collaborate from any location

## **Research Gateway**

### Secure, performant, and scalable research workbench that . . .

... can be provisioned in minutes

. . . is backed by secure and resilient data ingestion framework

. . . is delivered with wide selection of compute resources, including HPC

. . . is delivered with role-based access to workbench and data

. . . has clear budget and consumption costs breakdown by project, user, and workload



## What are Research Gateway use cases?







**RONIN** is a cloud orchestration and collaboration platform, lowering the entry level to using the cloud for researchers and research IT.

- Deployed within one AWS account, serving many researchers
- Enables researchers selfservice access to AWS resources
- Enforces an institution's security policy





## **Research and Engineering Studio on AWS**

Open source, easy-to-use web-based portal for administrators to create and manage secure cloudbased research and engineering environments.

#### Benefits:

- Minimize administrative overhead
- No cloud expertise required
- Flexible access to services

#### Use cases:

- Collaborate using shared research and engineering environments
- Define and manage projects
- Enable access to AWS without creating individual accounts



### The most comprehensiv e set of data services

#### DATA LAKES

DATABASES

Open Data Registry

LEARNING & **GENERATIVE AI** INTERACTIVE QUERY RELATIONAL **KEY-VALUE BIG DATA PROCESSING** DOCUMENT REAL-TIME ANALYTICS DATA WAREHOUSING GRAPH TIME-SERIES DATA INTEGRATION LEDGER **BUSINESS INTELLIGENCE** WIDE COLUMN OPERATIONAL ANALYTICS MEMORY

ANALYTICS

MACHINE

# A complete platform for building and deploying edge applications



## Agility and productivity -GROMACS: Max Planck Institute







https://aws.amazon.com/blogs/hpc/running-20k-simulations-in-3-days-with-aws-bat ch/ https://pubs.acs.org/doi/10.1021/acs.jcim.2c00044#





NICE DCV





Amazon S3

Amazon EC2 Auto Scaling AWS ParallelCluster

© 2023, Amazon Web Services, Inc. or its affiliates.

# **Clemson University - Natural Language**

Photomace.Socontrologs/aws/natural-language-processing-at-clemson-university-1-1million-vcpus-ec2-spot-instances/

> The researchers conducted nearly half a million topic modeling experiments to study how human language is processed by computers.

The 1.1 Million vCPU count usage is comparable to the core count on the largest supercomputers in the world.





"I am absolutely thrilled with the outcome of this experiment. The graduate students on the project [...] used resources from AWS and Omnibond and developed a new software infrastructure to perform research at a scale and time-to-completion not possible with only campus resources."

– Prof. Amy Apon, Co-Director of the Complex Systems, Analytics and Visualization Institute



Amazon EC2

#### NATIONAL INSTITUTE OF NUCLEAR PHYSICS AND AWS WORK TOGETHER TO ACCELERATE QUANTUM COMPUTING RESEARCH

#### "

We are pleased to partner with AWS in what is an important element of our global strategy in quantum computing

research.

Marco Pallavicini, executive board member of INFN.





## Sharing & collaborating UniTrento - ICLUS

SOLUTION

University of Trento's Department of Information Engineering and Computer Science used AWS to run a International Project (https://www.disi.unitn.it/iclus) where AI algorithms were used to analyze ultrasound images of lungs to determine possible Covid infections.



#### More than 29 joint pubblications

© 2024, Amazon Web Services, Inc. or its affiliates.

#### More than 60 institutes involved

#### Italy

- 118 Castelnuovo, Garfagnana (LU)
- APSS Trento, Trento (TN)
- Ausl Romagna Cesena, Cesena (NULL)
- Azienda Ospedaliera Universita di Padova, Padova (PD)
- Azienda Ospedaliera Universitaria Federico II, Napoli (NA)
- Azienda Ospedaliero Universitaria Policlinico Vittorio Emanuele, Catania (CT)
- Bresciamed, Brescia (BS)
- Cardiologia Ospedale Policlinico San Martino, Genova (GE)
- Emergency Department of Arzignano Hospital AULSS8 Berica, Vicenza (VI)
- Fondazione Policlinico San Matteo IRCCS, Pavia (PV)
- Fondazione Policlinico Universitario A. Gemelli IRCCS, Roma (RM)
- Mater Olbia Hospital, Olbia (SS)
- Ospedale Civile di Voghera, Voghera (PV)
- Ospedale dei Bambini Vittore Buzzi, Milano (MI)
- Ospedale di Sanremo Asl1 Imperiese, Sanremo (IM)
- Ospedale di Tione, Tione (TN)

#### **Other countries**

- Augusta University Department of Emergency Medicine, Augusta, United States of America
- Clinic of thoracic and vascular surgery, Gera, Germany
- Contra Costa Regional Medical Center, Martinez, California, USA
- DeepMed I/O, Mancester, United Kingdom
- Department of Obstetrics and Gynecology University Hospitals Leuven, Leuven, Belgium
- DSP Medea, Medea, Algeria
- Eindhoven University of Technology, Eindhoven, The Netherlands
- Hospital das Clínicas da Universidade de São Paulo, San Paolo, Brasile
- Hospital General de Catalunya, Barcellona, Spain
- IFC CNR LECCE, Lecce, ITALIA
- Indian Institute of Technology, Kharagpur, India
- Indian Institute of Technology, Jodhpur, India
- Indian Institute of Technology Patna, Patna, India
- Institute of Biomedical Engineering, University of Oxford, Oxford, United Kingdom
- Intelligent Ultrasound, Cardiff, United Kingdom
- Klinikum rechts der Isar der TU, München, Germany
- Leitat Technological Center, Barcellona, Spain
- Massachusetts General Hospital, Boston, USA
- Michigan State University, Lansing, USA
- North Carolina State University, Raleigh, United States of America



#### AWS RESEARCH $-17^{TH}$ SEPTEMBER 2024

# Services and case studies: cloud in action

# AWS Education & Research

© 2023, Amazon Web Services, Inc. or its affiliates.

# AWS grant acceleration support package





# **Build research**



### **Lightsail for Research**

https://aws.amazon.com/lightsail/research/















# Acquiring data

# Data analysis



## **Multimodal analytics**

PURPOSE-BUILT SERVICES FOR HEALTHCARE AND LIFE SCIENCES



#### The AWS ML Stack Broadest and most complete set of machine learning capabilities



🥱 GLUON Deep K Keras mxnet **TensorFlow** GPUs & Learning Elastic Inferentia **FPGA** Trainium AMIs & CPUs Inference **O** PyTorch DeepGraphLibrary Containers lea



# Amazon

# Build, train/ and deploy ML models at scale

Automatic model fine-tuning & distributed training

Flexible model deployment options

Tools for ML operations

Built-in features for responsible AI

### University of Oxford Introduces a Sector-Leading Image Recognition ML Prototype to Augment Digitization in Numismatics



First stage Improve base image quality





Second stage Visually search for items in collection

I thought this project would be complex and time consuming, but using AWS made it easy.

Anjanesh Babu

11

Systems architect and network manager, Gardens and Museums IT, University of Oxford's Gardens, Libraries & Museums







Amazon SageMaker Amazon EC2

Amazon S3

## Large Models built on AWS



Clibrain develops Lince Zero—the first large language model (LLM) optimized for Spanish using Amazon SageMaker Stability AI will build AI models on compute clusters with thousands of GPU or <u>AWS Trainium</u> chips, reducing training time and cost by 58%

stability.ai



BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance, utilizing AWS infrastructure services

### New LLM development and refinement UAE Technology Innovation Institute



#### **CHALLENGE**

TII researchers needed an easy way to develop, iterate, and distribute Falcon-40B LLM. Experimenting at such massive scale was slowed by limited computational resources and complex onprem infrastructure. Researchers sought more flexibility to efficiently enhance, evaluate, and provide access to Falcon under dynamic capacity requirements to empower their team's foundation model research.

#### **SOLUTION**

TII utilized SageMaker's managed machine learning infrastructure and tools to train, host, and deploy versions of Falcon with greater speed and experimentation. SageMaker provided on-demand access to training clusters along with modular deployment options for easy distribution to varying client workloads—key for customizing model access.

#### OUTCOME

✓ TII successfully implemented Falcon-180B by using SageMaker and custom innovation. This allows TII's contribution within UAE's 2031 National AI Strategy, fostering economic growth and social progress. Releasing UAE's Falcon 180B, World's Top-Ranked Open Source AI Model will further encourage AI academic research and scientific collaboration.

# Innovating at the silicon level



## Graviton, Inferentia, and Train



Graviton-based Elastic Compute Cloud instances use up to 60% less energy than comparable Amazon EC2 instances

Models built on Trainium result in energy-consumption reductions of up to 25% versus comparable instances

Our Inferentia2 machine learning accelerator delivers up to 50% higher performance per watt

# Amazon Q Developer

Al-powered code suggestions in the IDE and the command line



# **Amazon Bedrock**

# simplifies







Choice

Customization

Integration



### **How AWS supports Gen AI research**



**Citizen services** 

# Publishing and sharing research data



### Reach out to your contacts at Amazon Web Services



### **Roberta** Piscitel

Tech BDM Research | EMEA piscitr@amazon.com | +44 793 3576 787



#### Andreas Wagner

Sr. Account Manager | Research & Nonprofits <u>wagneand@amazon.com</u> | +49 151 0913 5033



# Thank you!

AWS Education & Research

Reparts Piscitelli - piscitr@amazon.com

© 2023, Amazon Web Services, Inc. or its affiliates.

# Responsible Al

# **Key dimensions**



# AWS announces new tools for responsible AI innovation across the generative AI lifecycle

#### Implementation of guardrails for Amazon Bedrock

Implements predefined safeguards and filters based on customer policies to avoid harmful content Model Evaluation for Bedrock & SageMaker

Compares foundation models on custom metrics like accuracy, safety, and subject criteria

#### **Fight misinformation**

Invisible watermarks on AI-generated images to combat disinformation in amazon titan

#### **Building Trust & Transparency**

IP Indemnification - Protects customers from infringement lawsuits related to outputs of generative AI services

#### AWS's Commitment to Responsible AI

Collaborating with policymakers and the AI community on responsible practices Investing in purpose-built tools like SageMaker Clarify and ML Governance to operationalize responsible AI

https://aws.amazon.com/blogs/machine-learning/announcing-new-tools-and-capabilities-to-enable-responsible-aiinnovation/



© 2023, Amazon Web Services, Inc. or its affiliates

