From Cloud to Edge:

Cloud Strategies for Building a Resilient Data Infrastructure

By distributing computing power closer to the data source (the "edge"), meteorological operations can achieve lower latency, faster responses, and localized processing. This is critical for weather forecasting, early warning systems, and disaster response, where timely and accurate data is essential. Cloud-to-edge solutions also enable better handling of massive datasets, ensuring that global meteorological services are resilient, scalable, and able to support **growing demands.**

Knowing the capabilities and tradeoffs of cloud and edge will help you build the most effective hybrid cloud strategy.

Topics:

- Moving Less Data, Doing More: Edge and regional compute, advanced compression, feature selection
- Compute Where the Data Lives: Containerization, Data and analytics and decision making
- Models that Move, Not Just Data: Distributed ML training, customizing AI

From Cloud to Edge:

Cloud Strategies for Building a Resilient Data Infrastructure

@SG-FIT Workshop
Brett Tackaberry

September 24, 2024

Google Cloud

Speaker introduction

Brett Tackaberry

Principal Architect

Google Cloud
Canada Public Sector

Brett Tackaberry is a Principal Architect at Google Cloud Canada, where he leads cloud and digital transformation initiatives for federal public sector clients. With over 20 years of experience in cloud-native infrastructure, data engineering, and security, Brett focuses on designing scalable, sustainable solutions for complex challenges. His passion for leveraging technology to address global issues extends to his work with community-driven initiatives. Brett is dedicated to using technology to drive environmental sustainability and improve outcomes in both public and private sectors.

Our mission

Support Increasing volume of data as needs of forecasting increase spatially and temporally

Importance of making atmospheric data accessible to a wider range of stakeholders, including researchers, policymakers, insurers, private industry and more.

01

From Cloud to Edge: Cloud Strategies for Building a Resilient Data Infrastructure

- Moving Less Data, Doing More
- Compute Where the Data Lives
- Models that Move, Not Just Data
- Breaking Down Barriers (Throughout)

Moving Less Data, Doing More

1

Moving Less Data, Doing More

The Problem: Traditional data movement bottlenecks. Costs, latency, and the impossibility of moving everything in real-time.

- Advanced Compression: Lossy & lossless, tailored to data types (imagery, sensor readings, etc.), not just generic algorithms.
- Intelligent Subsetting: AI-driven extraction of relevant features, not just crude downsampling. 'Smart' data for specific downstream tasks.

Data

- Edge Computing: Preprocessing closer to the source, sending only actionable insights to central systems.

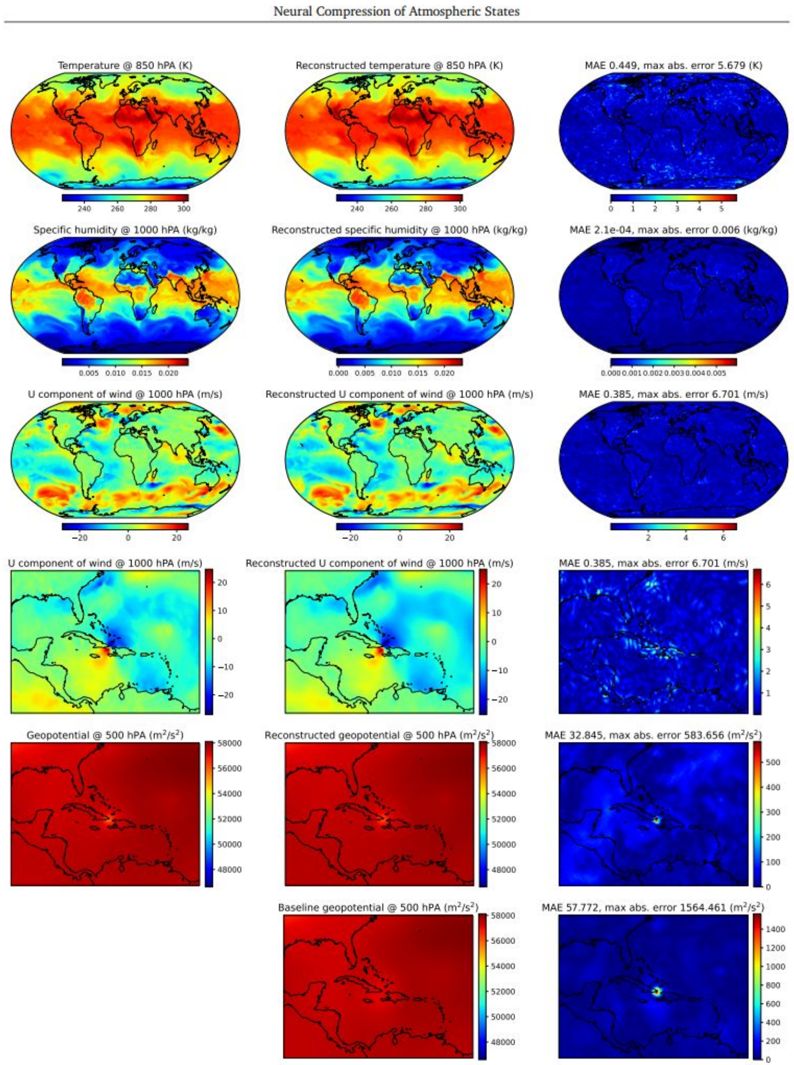
Compute Infrastructure

Google DeepMind

Neural Compression of Atmospheric States

https://arxiv.org/abs/2407.11666

July 2024



Neural Compression of Atmospheric States

Temperature @ 850 hPA (K)
Reconstructed temperature @ 850 hPA (K)
MAE 0.449, max abs. error 5.679 (K)
240 260 280 300
0 1 2 3 4 5
Specific humidity @ 1000 hPA (kg/kg)
Reconstructed specific humidity @ 1000 hPA (kg/kg)
MAE 2.1e-04, max abs. error 0.006 (kg/kg)
0.005 0.010 0.015 0.020
0.000 0.005 0.010 0.015 0.020
0.000 0.001 0.002 0.003 0.004 0.005
U component of wind @ 1000 hPA (m/s)
Reconstructed U component of wind @ 1000 hPA (m/s)
MAE 0.385, max abs. error 6.701 (m/s)
−20 0 20
2 4 6
U component of wind @ 1000 hPA (m/s)
Reconstructed U component of wind @ 1000 hPA (m/s)
MAE 0.385, max abs. error 6.701 (m/s)
20 10 0 −10 −20
6 5 4 3 2 1
Geopotential @ 500 hPA (m²/s²)
Reconstructed geopotential @ 500 hPA (m²/s²)
MAE 32.845, max abs. error 583.656 (m²/s²)
58000 56000 54000 52000 50000 48000
500 400 300 200 100 0
Baseline geopotential @ 500 hPA (m²/s²)
MAE 57.772, max abs. error 1564.461 (m²/s²)
1400 1200 1000 800 600 400 200 0

Google Cloud

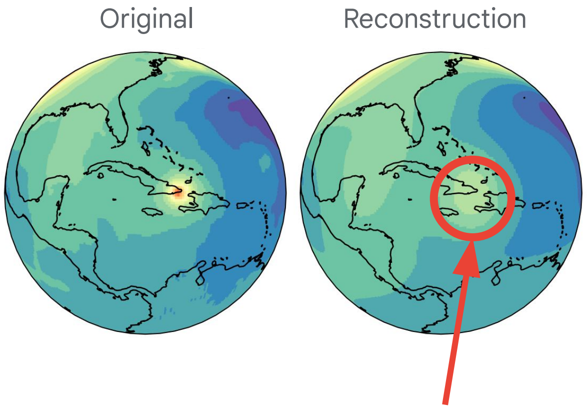
Neural Compression for Drastic Data Reduction

Shrink Weather Data, Not its Value: 1000x Compression

- Lossless compression offers minimal gains, and lossy methods often distort crucial features (hurricanes, etc.).
    - Rare and extreme events are important (unlike image compression)
- Neural network-based compression achieves 1000x+ reduction with low error AND preservation of extremes.
- Faster, More Accessible: ~1 second per global atmospheric state encoding/decoding time, enabling near-real-time use cases.
- Openness & Collaboration: Makes use of open standards (HEALPix). Working on evaluation with experts experts.

Baseline neural compression model

Huang & Hoefler (2023)



Original

Reconstruction

Hurricane erased

Intelligent Subsetting

Aka: Feature Selection, Dimensionality reduction

- Extracts critical climate indicators without needing to store entire simulations
- A technique where AI is used to selectively extract the most important or relevant features from a larger dataset.
- Reduce data size while retaining key information to ensure meaningful analysis.
- Targeted data retention:
  - Keeping high-impact regions
  - Retaining critical time periods

A Comparison

- **Uniform Reduction:**
  - Down-sampling the entire dataset uniformly.
  - May discard important information.
- **Principal Component Analysis (PCA):**
  - Reduces dimensionality by transforming data mathematically.
  - May lose interpretability.

Cloud Native file formats for climate, weather and geospatial data

- Cloud-native formats (like COG and Zarr) allow selective data access, minimizing latency by retrieving only the necessary information needing to store entire simulations
- Efficient compression (like Zarr and Parquet) reduce data size
- Designed for scalability
- Support both random access for real-time needs and sequential access for batch processing
- Formats with metadata and indexing capabilities (like COG and NetCDF) improve data organization and searchability

Cloud Option 1 - Global regions

- Regional and Zonal Compute Resources: By deploying compute instances in the same region or zone as your data, you can significantly reduce latency.
- Cloud Functions and Cloud Run: These serverless compute options allow you to execute code in response to events or triggers, minimizing the need to provision and manage dedicated infrastructure, further reducing operational overhead and latency.
- Reduced Costs: By minimizing data transfer and storage requirements, data-proximate compute can help reduce operational costs and improve efficiency
- Reduce data size through compression and inferencing, distribute globally reduced dataset

Cloud Option 2 - Edge

- Edge is any location where data is born, needing local processing. Including raw data or derived data
- Real-Time Response: Edge compute enables faster alerts, on-site decision support, crucial for extreme weather.
- Data Sovereignty: Process sensitive data locally, complying with regulations, while still benefiting from cloud capabilities.
- Empowerment of Members: Even smaller organizations can leverage advanced analytics at the edge, democratizing innovation.

Start with a strategy

- Understand your use cases
- Evaluate HW and SW infra
- Security & Compliance
- Scalability
- Skills & Expertise
- Cost optimization

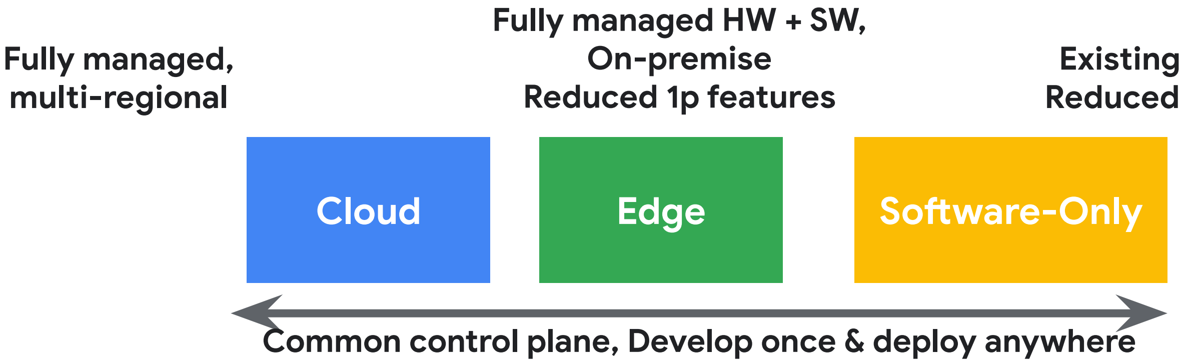
Edge Compute - how we do it.

Can you achieve the benefits of edge computing AND take advantage of cloud native opportunities

- Managing infrastructure is hard
- Private cloud is hard to manage and difficult to stay modern and take advantage of new technologies

Google Distributed Cloud

Google's fully managed hardware and software product that delivers modern applications equipped with AI, security, and open source at the edge.



Fully managed, multi-regional

Fully managed HW + SW, On-premise Reduced 1p features

Existing hardware, Reduced 1p features

Common control plane, Develop once & deploy anywhere

Edge-ier devices

Compute Where the Data Lives

2

Compute Where the Data Lives

The Problem: Centralized processing is overwhelmed. Agility suffers, innovation is stifled.

- Serverless & Kubernetes: Scale compute on-demand, precisely where your data resides (on-prem, multi-cloud, at the edge).
- Data Analytics & ML at Scale: Tools like BigQuery, Vertex AI, not just for 'big tech', but accessible to any organization.

Compute Where the Data Lives

Kubernetes:

- **Container orchestration at scale:** Seamlessly manage complex, distributed applications.
- **Portability across environments:** Run on-premises, in the cloud, or at the edge - maintain consistency and avoid vendor lock-in.
- **High availability & resilience:** Ensure your applications are always up and running.

Serverless:

- **Focus on code, not servers:** Eliminate infrastructure management, scale automatically to handle any workload.
- **Pay for what you use:** Cost-efficient, ideal for sporadic or unpredictable workloads.
- **Rapid development & deployment:** Accelerate innovation, respond to changing needs with agility

A parallel effort - Our data centers work harder when the sun shines and wind blows

https://blog.google/inside-google/infrastructure/data-centers-work-harder-sun-shines-wind-blows/

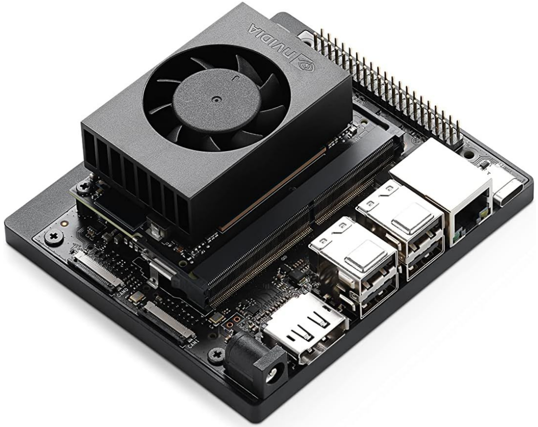
- Spatially and temporally shifting workloads to reduce carbon emissions and improve data center efficiency
- The platform uses forecasts to predict when low-carbon energy will be available.
- Carbon-intelligent computing platform was designed to shift the timing of compute tasks to when renewable energy is available.
- Strategies are about efficiency and sustainability, albeit with different primary resources in mind
- Take away: moving workloads around as the strategy

Analytics and AI/ML - Decision making at the edge

- Real-Time Data Processing
    - AI/ML models can run directly on edge devices, enabling real-time decision-making This is particularly valuable in time-sensitive meteorological applications like disaster detection or early warning systems.
- Reduced Latency, Faster Insights
    - Faster insights and quicker responses to critical weather events (e.g., storm predictions, flood warnings).
    - Immediate, actionable intelligence.
- Localized, Context-Aware Analytics
    - Localized data and environmental factors into AI/ML models, making analytics context-aware and more relevant.

Analytics and AI/ML - Decision making at the edge (#2)

- Distributed AI Model Deployment
    - Using tools like TensorFlow Lite and AutoML Edge from Google Cloud, complex AI/ML models can be efficiently deployed and run on edge devices.
    - This ensures that advanced analytics, such as predictive weather models, are available even in remote or low-connectivity areas.



Models That Move, Not Just Data

3

Models That Move, Not Just Data

The Problem: 'One size fits all' models are inefficient. Local conditions, rapidly evolving situations demand adaptation.

- Distributed ML Training: Build models where the data is, then federate learnings for global insights without raw data sharing.
- Customizable AI: Tools to fine-tune models for specific regions/phenomena, empowering local experts.

"One-Size-Fits-All" Models: Inefficiencies and Challenges

- Inefficiency of Large Centralized Models
  - "One-size-fits-all" models can be overly large and cumbersome.
  - May include unnecessary complexity that is not relevant for all data sources or use cases, leading to inefficient processing.
- Resource-Intensive Model Training
  - Moving massive amounts of data to the cloud for centralized training consumes bandwidth and increases latency.
  - Real-world meteorological conditions are diverse, so a single model may not adapt well to local variations.
- Unwieldy Model Deployment
  - Large models are harder to deploy and manage on edge devices with limited compute and storage capacity.

Train at the Edge, Share via the Cloud

- Local Training at the Edge
  - Edge devices collect and process real-time data specific to their environment, enabling localized model training.
  - Train the model directly on the edge device with real-time weather data for greater relevance and adaptability.
- Sharing Models Through the Cloud
  - Once trained, the edge can share the optimized model with other devices through the cloud, reducing the need to transmit raw data.
  - This approach optimizes bandwidth use and facilitates collaborative learning across regions.

Foundational Models in the Cloud, Fine-Tuned or Customized at the Edge

- General Foundational Models
  - Start by training a general-purpose, foundational model in the cloud that is capable of handling a broad range of meteorological data.
- Fine-Tuning at the Edge
  - Deploy this foundational model to edge devices where it can be fine-tuned using local data to increase accuracy and relevance.
  - This hybrid approach allows edge devices to benefit from cloud-powered generalization while incorporating location-specific knowledge for enhanced precision.

Building Expertise and Customization at the Edge

- Leveraging Local Information
  - Edge devices can build expertise based on hyper-local environmental data.
  - Enable researchers at the edge to leverage their deep understanding of local weather patterns to further refine and customize models.
- Customization of Models
  - Edge training allows models to specialize in the unique patterns or anomalies present in a particular location (e.g., local climate, terrain, or infrastructure).
  - Customized models can evolve continuously based on new data without needing constant intervention from centralized cloud servers.

Adaptive Learning Models

- Continuous Learning at the Edge
  - Edge devices can adapt models based on changing conditions, such as evolving climate patterns or unexpected environmental anomalies.
- Feedback Loops Between Cloud and Edge
- Cloud models can be updated based on edge-learned refinements, ensuring a continuous feedback loop for model improvement.

Wrap Up

1

Edge compute strategy building for AI, real-time analytics, flexibility, scalability.

Neural compression for data reduction.

Data lifecycle management and data governance will also be important.

2

Build with containers and cloud-native technology to be able to scale.

Combined edge strategy with global cloud strategy

Build models that can respond quickly, run on smaller, low-power devices.

3

Start with shifting workloads through global regions to prepare for shifting models around the world.

Deploy model training capability and analytics at the edge

4

Foster collaboration: Share data, share models. Build for interoperability.

Secure & Compliant Data Flows: Ensure robust security and regulatory compliance safeguarding sensitive meteorological data throughout its lifecycle.

Thank you

Brett Tackaberry
tackaberry@google.com
Google Cloud, Public Sector Canada

https://www.linkedin.com/in/tackaberry/

Google Cloud