SDC 20

BY Developers FOR Developers

Storage Developer Conference
September 22-23, 2020

Accelerate Artificial Intelligence IoT Use Cases with Storage Tiering and Shared Storage at the Edge

Presenter: Joey Parnell
SW Architect, NetApp ESG
Presenter: Dr. M.K. Jibbe
Technical Director, NetApp ESG

Agenda

- Current problem – architecture limitations
- Emerging AI data pipelines
- Shared storage for inference
  - Concepts and use cases
  - Logistics and considerations
- Key takeaways

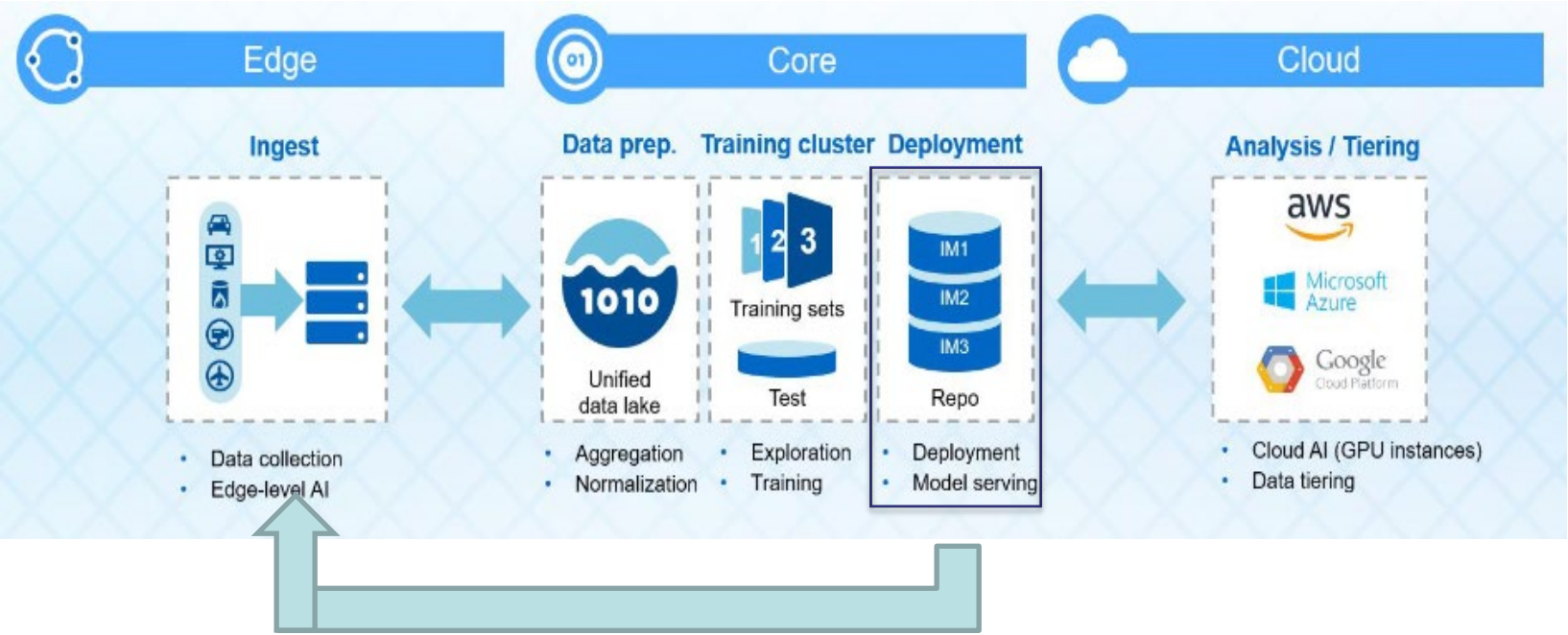
Current Problem

1. Transmitting data from Internet-of-Things (IoT) edge devices to core data centers to perform resource intensive AI/ML operations is costly in terms of network bandwidth and latency.
2. Placing hardware resources such as GPUs at the edge to perform those operations locally and reduce network congestion and latency can be prohibited by cost and power requirements.

This presentation describes a concept and framework for presenting shared storage to IoT devices so compute-intensive iterative refinement training can be performed either in the core data center or in cloud analytics platforms, and updated AI inference data transmitted back to IoT devices to provide customized training to improve reliability and reduce false positive rates.

Emerging AI Data Pipelines

Edge – Core – Cloud AI Data Pipeline[1]



Workloads at Each Stage

- Ingest: raw data - sequential writes
- Data prep: labeling, normalize, transform
  - Workload varies
- Explore/Train: random, raw bandwidth, GPU-intensive
  - Training GPUs, powerful, saturate bandwidth
- Inference: testing and deployment - reads
  - Inference GPUs, lower wattage
- Where the work is done at each stage depends on the use case and industry

Decrease Time to Insight

- Transmitting large raw data sets to core data centers for inference is costly in terms of response latency and network congestion
- Moving the inference closer to the edge decreases the time to insight and action
- Quickly acting improves customer experience and drives business value

Emerging Pipelines

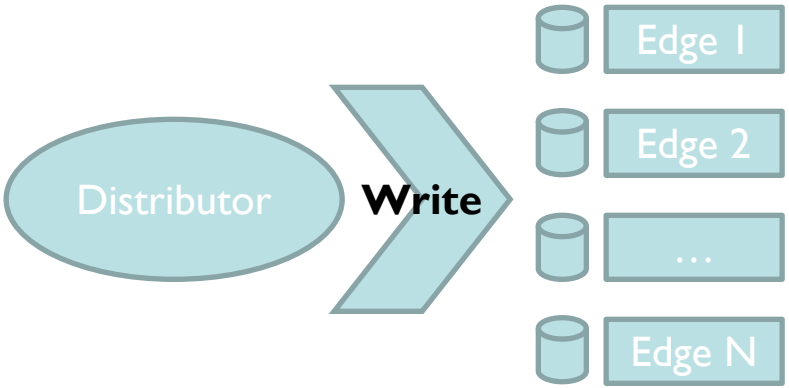
- Inference at the edge
  - Some industries are at different stages in this journey – some already doing; others just getting started
  - These are *not* fringe science projects
    - Clear ROI where ML is applied to mundane things - "boring AI"
    - Real inference use cases at the edge:
      - Does the patient need attention?
      - Is the person or crowd acting suspicious or threatening?
      - Does this area need to be cleaned?
      - Is the person's heart rate elevated?
      - Is there a manufacturing defect – should we stop the production line?
- Deep Transfer Learning (DTL) at the edge – emerging general ideas[2] and health care for COVID[3]

Shared Storage for Inference

Rethinking Architecture for Edge Inference

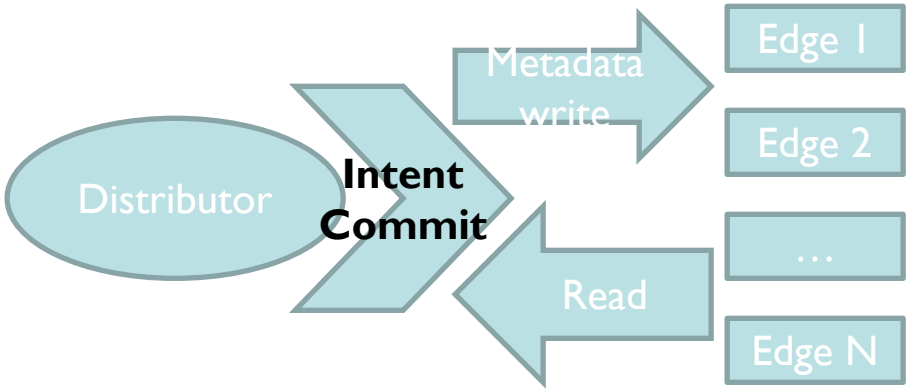
- Redesign architectures for (almost) "always connected" wireless devices with 5G and beyond
- IoT device proliferation[4] combined with increased deployment of AI at the edge[5] will exceed the FLOPS/watt increases over that same time[6]
- Some workloads are fine without GPUs, others absolutely require them
  - Urgency of the use case
    - Average consumer smartwatch monitoring vital signs
    - Stadium security camera system monitoring for threats
  - Locality of data
    - Network latency, expected performance of your inference use case
  - Storage space vs. AI inference use case storage size
    - Squeeze costs out of IoT devices - spend $$ solving the business problem

Commit Inference Model to Edge



- Many persistent copies of the same model
- Persist model on edge devices (flash parts)
  - Inefficient resource use
- Limited storage space for storing multiple model versions

Fetch Inference Model from Edge



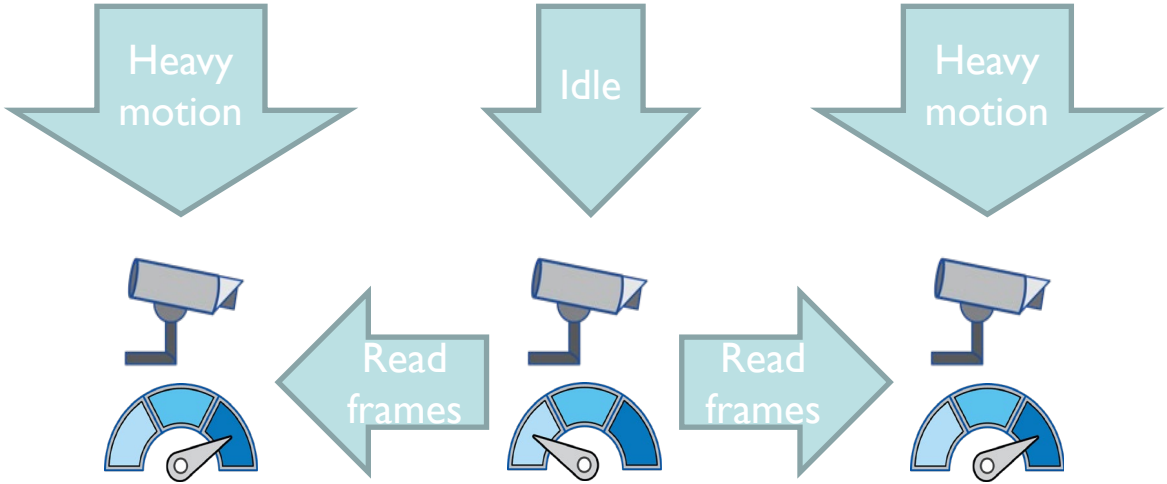
- Persist metadata only
  - Model persistence optional
- Version control with many accessible versions
  - Enable continuous deployment and roll-back
  - Flexible switch between models (e.g. day/night, weekday/weekend, COVID)
- Non-AI use cases also such as upgrades, patches, ...
  - Silicon to deployment from factory

Shared Storage Concepts and Use Cases

Centralized Shared Storage and Mesh Distributor Nodes - Concepts

- An extension of the model is to distribute work between edge nodes doing inference
- Create grouping of edge nodes to coordinate work instead of dedicated core node
- Fog computing concepts

IoT Mesh Work Distribution



"Read frames":

1) Write metadata on busy nodes indicating over-utilization
2) Idle nodes poll peer nodes for busy status
3) Read frames from idle nodes, as indicated by busy nodes, process inference
4) Busy nodes read result and merge

IoT Mesh Work Distribution

- Distributed edge devices with flash:
  - Applications divvy up work to idle or under-utilized devices, store results locally, send metadata to the originating device about where to read the results.
  - IoT devices coordinate and complete AI tasks that may exceed the computing capabilities or the latency requirements of the singular device
  - Minimizes the data traveling to and from data centers and clouds.

Disaster Recovery

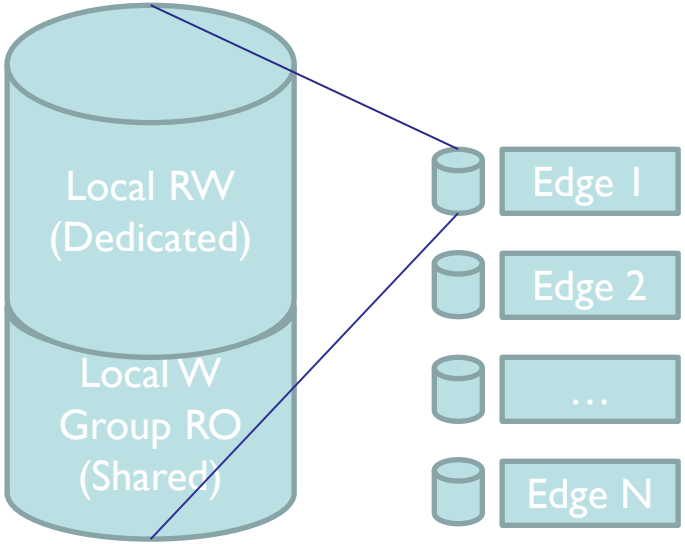
- Another use of shared storage is to present data from edge device as target for disaster recovery
- Mount edge device from cloud via gateway to perform analytics off-prem and return or archive results
- This can be a backup plan for when core or edge resources are unavailable or cannot keep up with demand

Flexible Mode Switching

- Where each part of the pipeline needs to be fulfilled depends on the use case
- Quality of Service (QoS) and Service Level Agreement (SLA) considerations
- Cost optimization
  - Compare costs between on-prem and cloud, and between cloud providers
  - For new inferencing of existing or new data switch providers and mount shared storage

Shared Storage Logistics and Considerations

Accessing Shared Data



- Need implementation of network file system to facilitate shared storage
  - NFS / CIFS / S3
- Adding support in edge device operating systems
  - Android, various RTOS
- Obvious trade-off of having minimal deployment in RTOS vs. rich features
  - Whether this makes sense depends on use case

Security

- Shared storage can be accessed by anyone!?
  - This is a non-starter for most business use cases - e.g. inference model or software is the valuable IP
  - Need a security model that can allow edge devices and the storage node (or other edge devices in mesh) to trust one another for communications – certificate chain of trust, etc.
  - Need data encryption model between those same devices
- Area of much deeper investigation – opportunities for increasing security

Other Considerations in AI/ML

- Ability to audit
- Compliance depending on industry
  - Personally identifiable information (PII)
  - On prem. vs. cloud
  - Retention requirements
- Reproduce-ability
- Federated learning capabilities
- Addressing fairness/bias concerns as well as social ethics

Key Takeaways

- Satisfy real-time IoT AI use cases with higher fidelity and without significantly increasing cost by providing shared storage in the IoT device to allow other devices to perform work and remotely update the local inference models.
- Add a layer of flash to IoT devices for data tiering provides the capability to perform AI tasks in a distributed fashion to utilize idle resources.
- Protect valuable IoT data at lower cost by selectively using cloud resources for archive and disaster recovery solutions.

Reference Citations

[1] Edge to Core to Cloud Architecture for AI:

https://www.netapp.com/us/media/wp-7271.pdf

[2] Cartel: A System for Collaborative Transfer Learning at the Edge:
https://acmsocc.github.io/2019/slides/socc19-slides-s1-daga.pdf

[3] A Survey on Deep Transfer Learning to Edge Computing for Mitigating the COVID-19 Pandemic:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7326453/

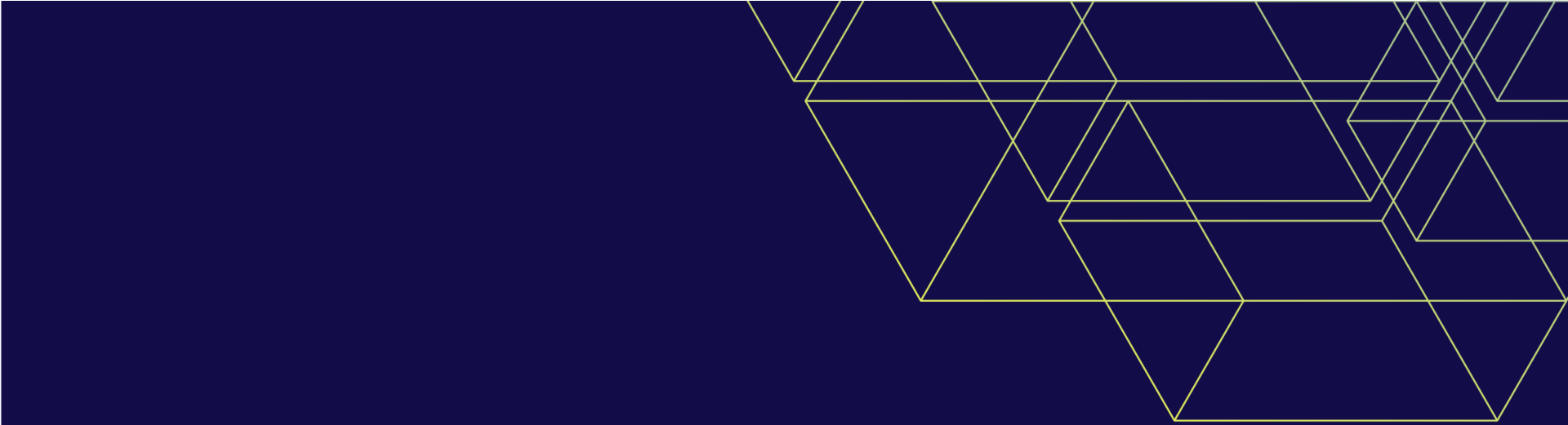
[4] The Growth in Connected IoT Devices Is Expected to Generate 79.4ZB of Data in 2025:

https://www.idc.com/getdoc.jsp?containerId=prUS45213219

[5] Bringing AI to the device: Edge AI chips come into their own:
https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2020/ai-chips.html

[6] Summarizing CPU and GPU Design Trends with Product Data

https://arxiv.org/pdf/1911.11313.pdf



Please take a moment
to rate this session.

Your feedback matters to us.