SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

Supercharging **OpenAI** Training with **Azure Blob Storage**

Presented by

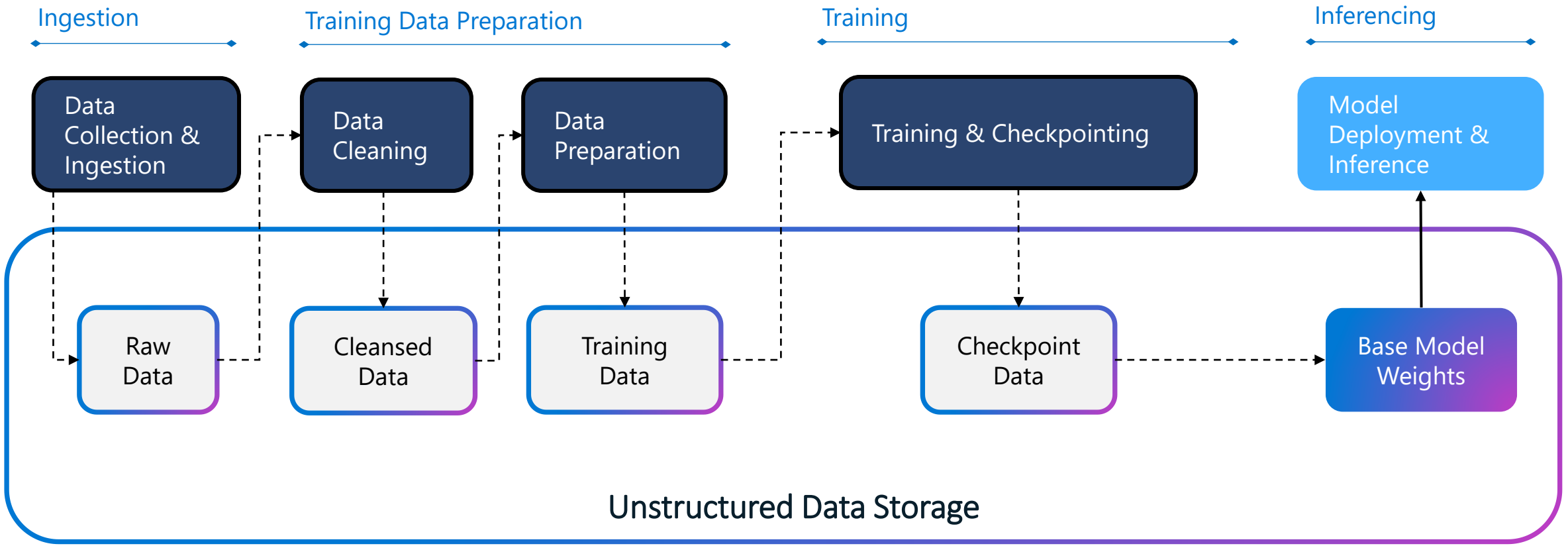
Jason Vallery – Group Product Manager (Azure Storage)

Jegan Devaraju – Group Engineering Manager (Azure Storage)

Agenda

- The AI Training Data Pipeline
- Scaling the AI Supercomputers
- Scaling the AI Data Platform
- Key Takeaways
- Q&A

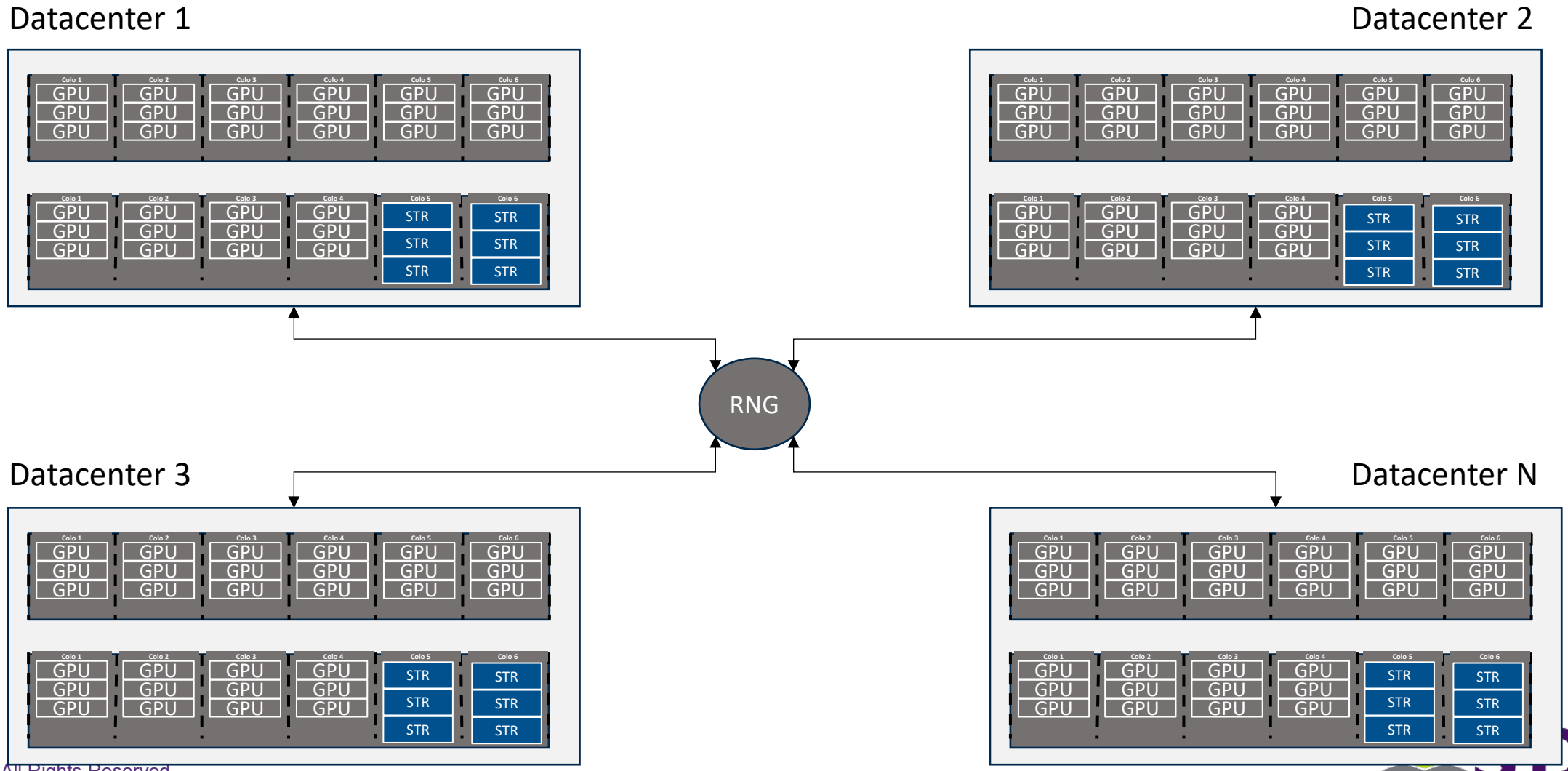
AI Pipeline – Storage centric view



AI Supercomputers

The ever-growing thundering herd

Physical Scale Out



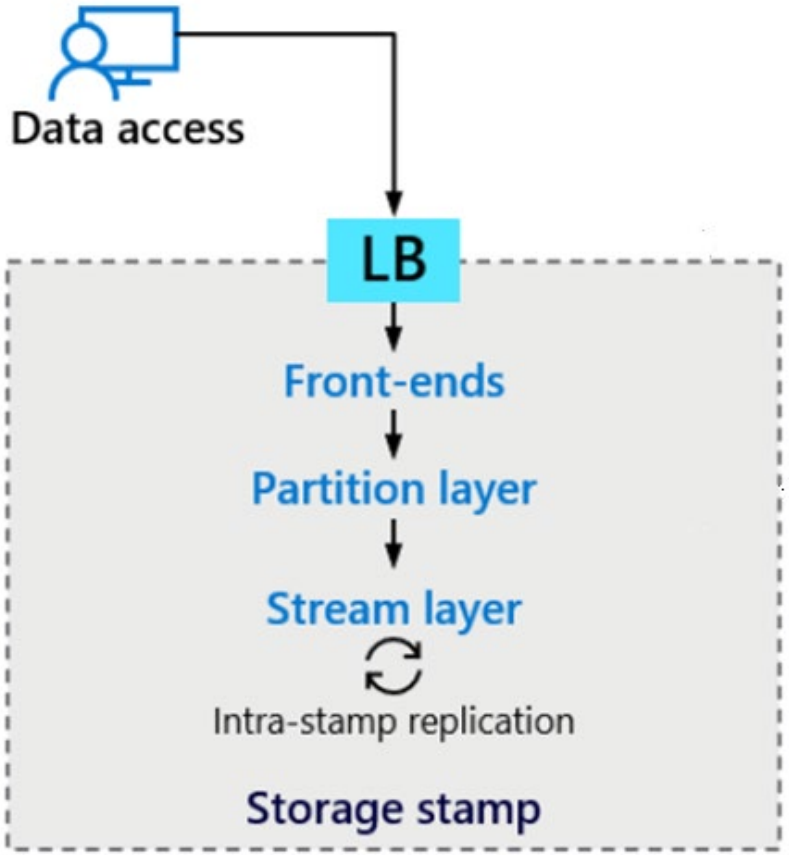
Checkpointing

- Checkpoint state of model during training
- Synchronous process
- Needed for:
  - Recovery from hardware failure
  - Job switch
  - Model evaluations
- LMT and LAT for retention policies

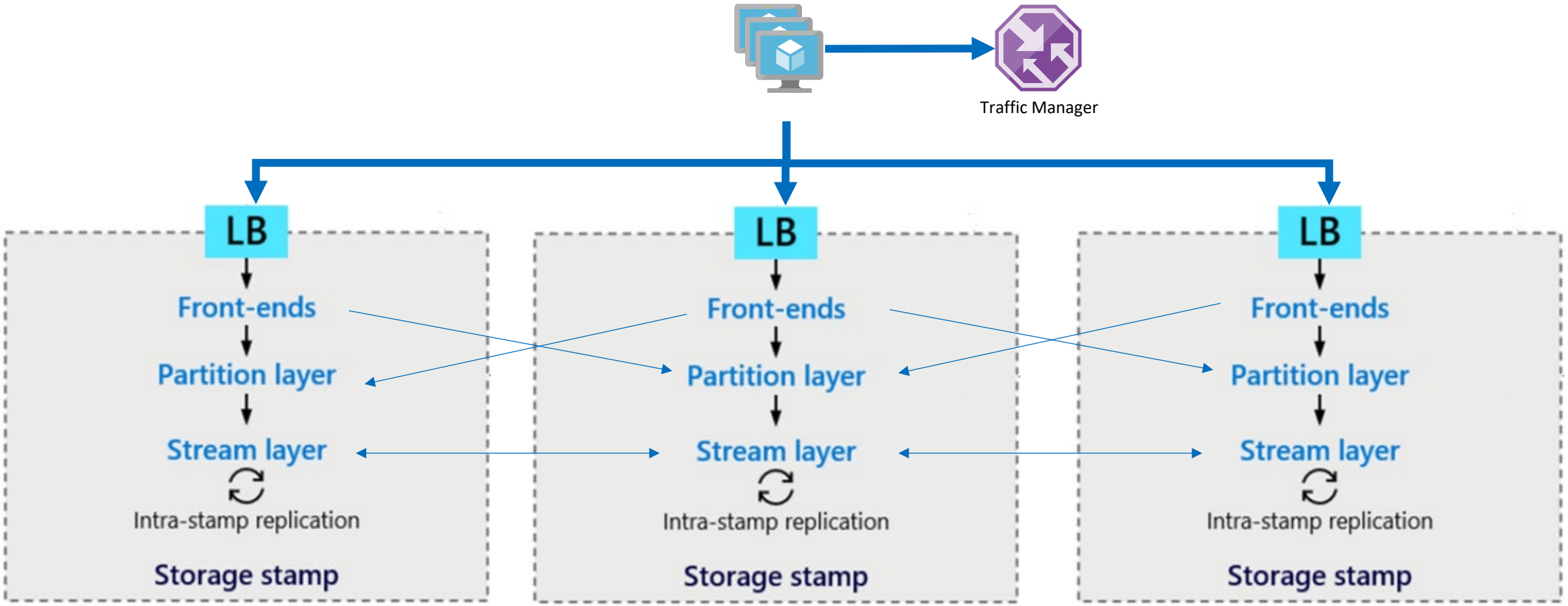
	Checkpoint Data
Capacity	• Function of model size, parameters, and GPU count • Python pickle • Capacity growth based on retention
Throughput	• Write local, drain to Blob (write back) • Write Tbps based on size and checkpoint frequency • “Cold Start” is high water mark for read from checkpoint
IOPS	• Inconsequential
Latency	• Inconsequential

Azure Storage Single Cluster Architecture

- Storage account limited by cluster boundary
- Scale limits tied to physical cluster limits
- Index and Data are together
- Fixed ratio of Tbps/TPS/PiB/Rack by hardware generation

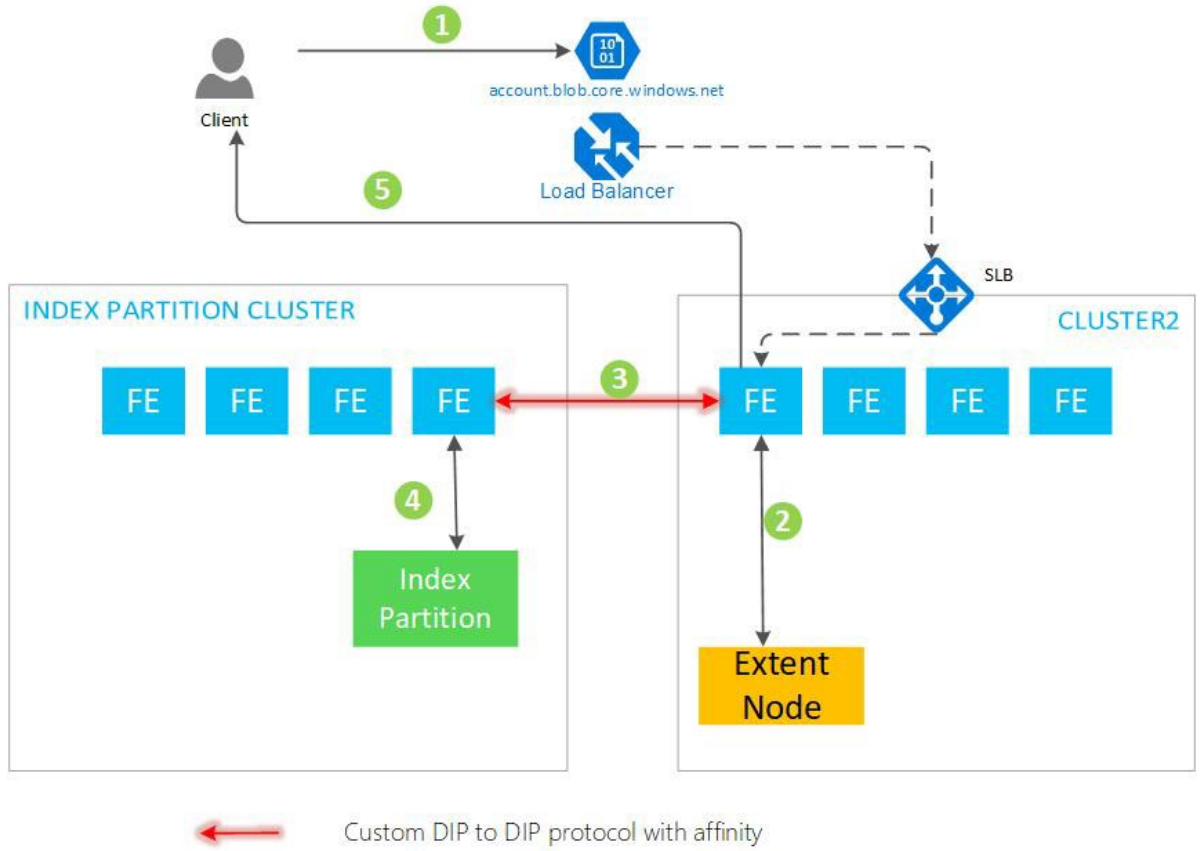


Scaled Accounts



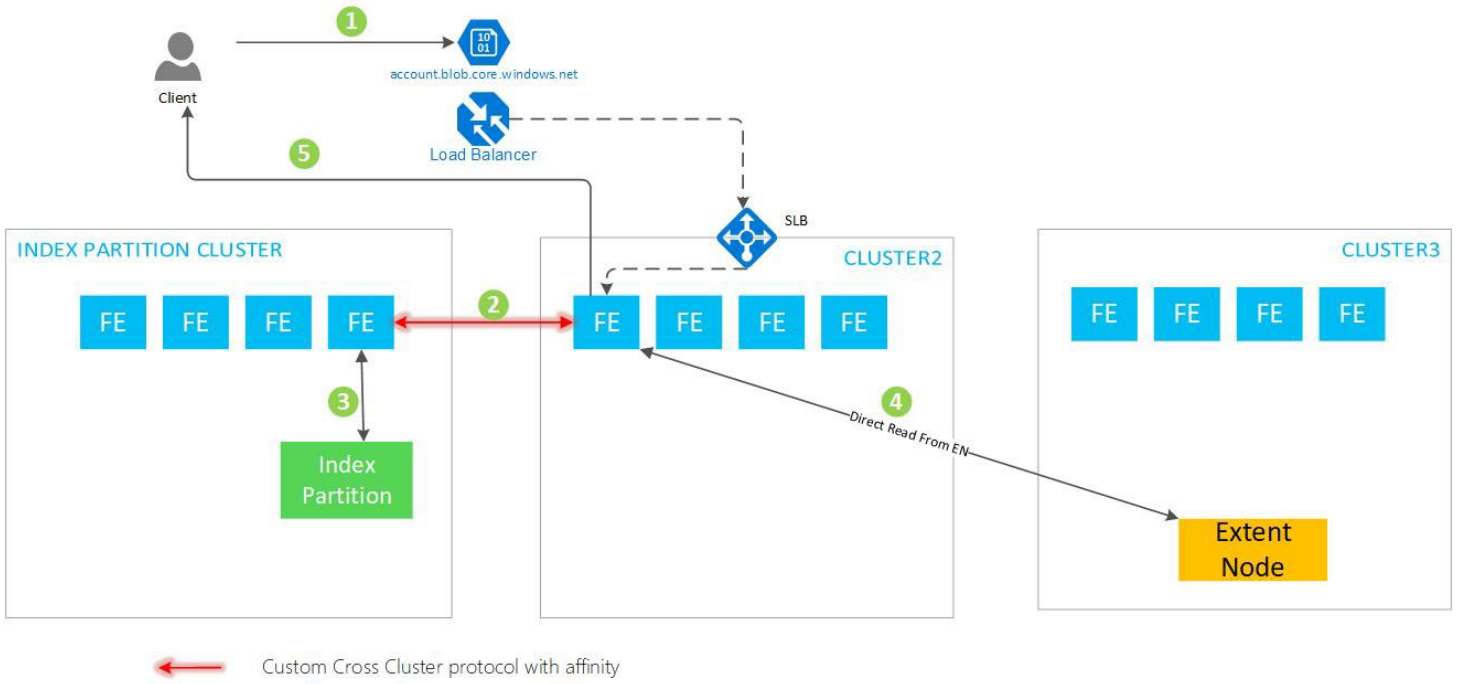
Scaled Accounts - Write

- Stateless FE persists data on the capacity node
- FE knows the authoritative cluster for serving the Index partition for the object
- FE routes the index update across cluster



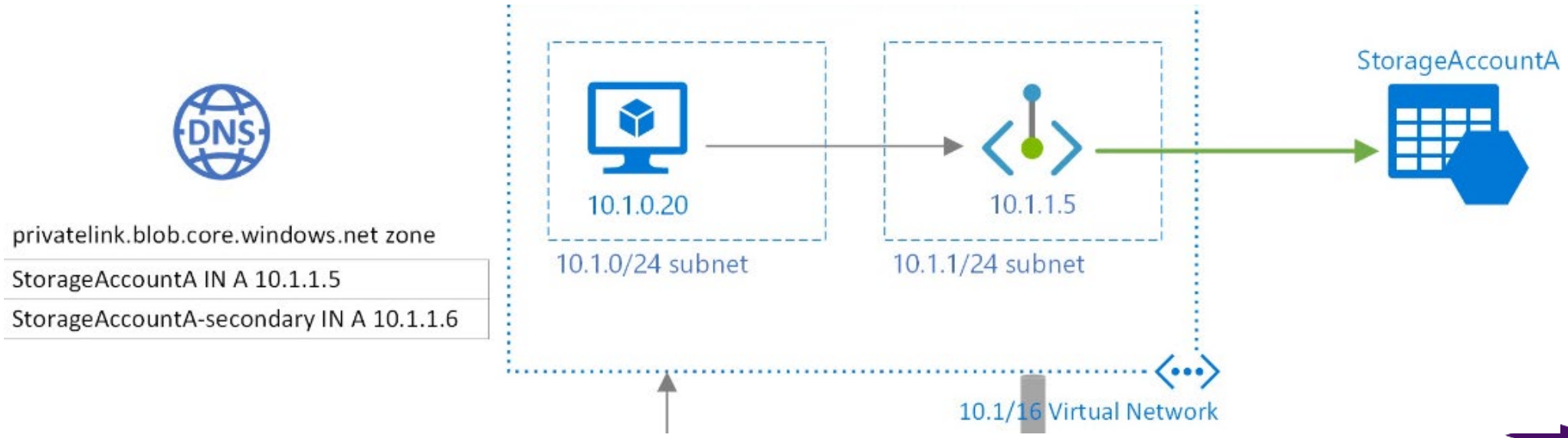
Scaled Accounts - Read

- Partition server identifies the capacity node
- Cross cluster communication on RDMA
- Support for all redundancy types



Private Endpoints

- Zero trust security posture
- Customers don’t want to expose Storage Account endpoints to internet
- Private Endpoint and Private Link technology injects Object Storage endpoints only into specific VNETs.
- Historically scale limited
- Scaled accounts orchestrates dynamically updating the private endpoints to provide seamless scale.
- Critical for AI training due to the private nature of the datasets and model weights.



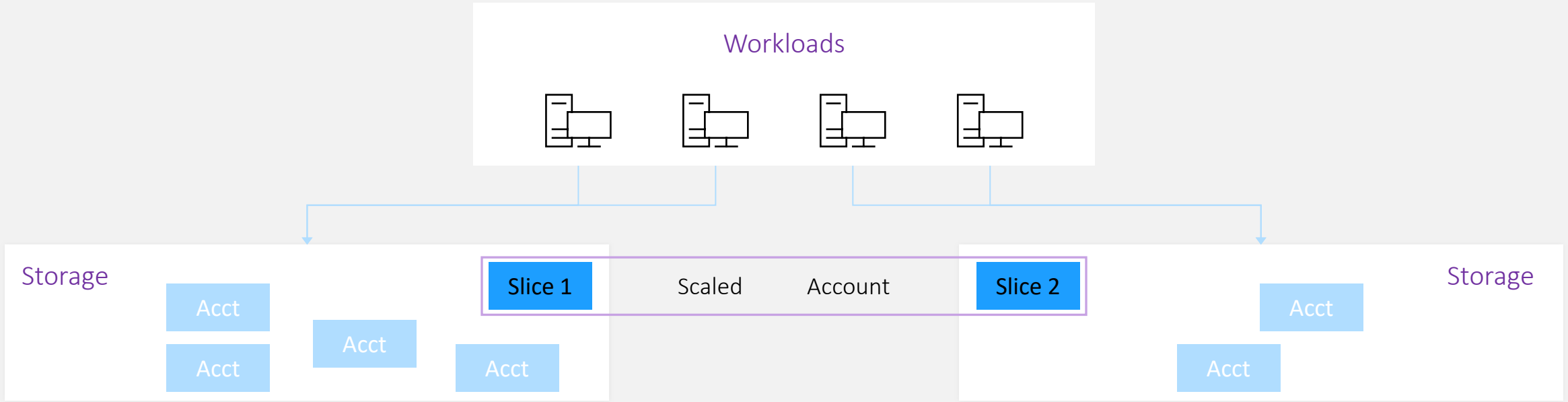
Scaled Accounts

Orders of magnitude increase in Capacity and Bandwidth to meet hyperscale customer asks

Horizontally Scale across any number of clusters

Eliminate the need for sharding large workloads across accounts

No change to pricing, not a new product/SKU, goal is to auto-scale



AI Data Platform

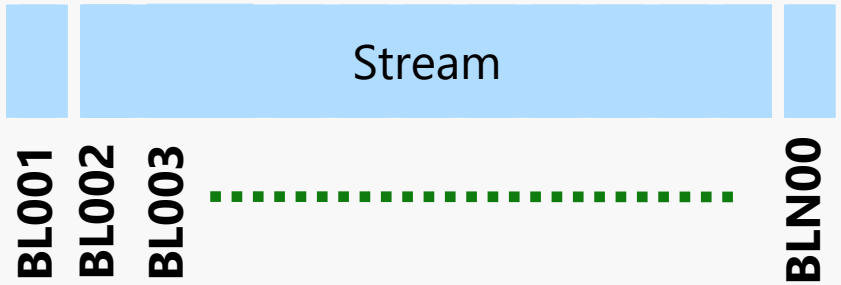
Data is the foundation of AI model development

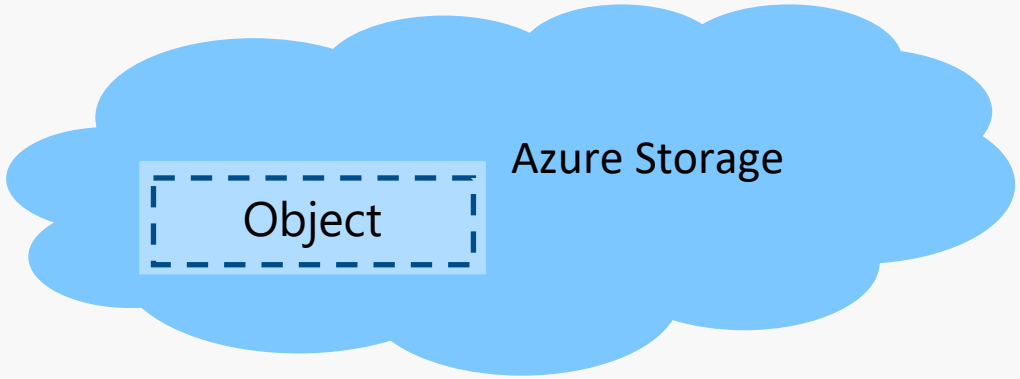
Data Collection and Preparation

- Ingesting streams of data
- Staging blocks can replace tools like Kafka
- Analytics tools like Spark, Pandas, Dask, Databricks used for processing, classification, and transformation for training

	RAW Data
Capacity	• Text, image, video, parquet • “Small” unstructured files (KiB to MiB) • Growth based on dataset acquisition
Throughput	• Primarily write heavy during ingestion • Primarily read heavy during pre-training
IOPS	• High during ingestion • Small files can push limits
Latency	• Low and consistent latency useful for iterative and small file processing

Staging Streams as Blocks





Up to 4GiB per block, variable size

Up to 50,000 blocks

Benefits:

Efficient Continuation

Parallel, out of order uploads

Easy bin packing

```
REST Calls
PutBlock(blobName, BL001, BL001Data);
PutBlock(blobName, BL002, BL002Data);
……
PutBlock(blobName, BLN00, BLN00Data);
PutBlockList(blobName,
             BL001….BL00N);
```


Partition Load Balancing

- Thundering herd of GPUs and CPU hit a new prefix (read or write)
- Huge surge in IOPS on a new partition/range
- Keys are based on object name logographically sorted
- Blob Storage Index is stored on the distributed filesystem
- Blob storage can load balance/split partitions in 1 to 5 seconds
- Thousands of load balancing/splits every day
- Scalable Dynamism

/dataset1_092024/file1.json …. file1000000000.json

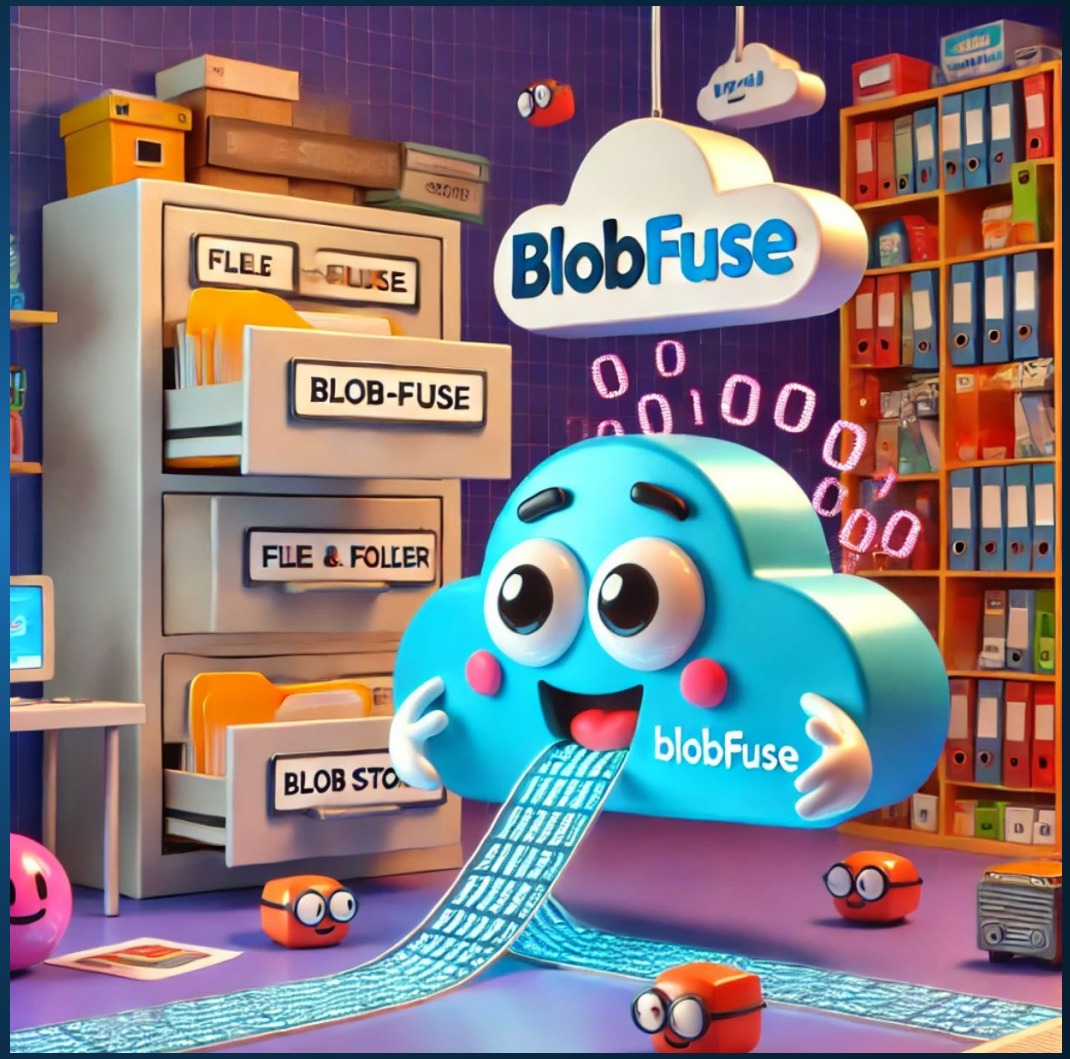
/dataset1_092024/file1.json …. file499999999.json
/dataset1_092024/file499999999.json …. file1000000000.json

Capacity Scaling

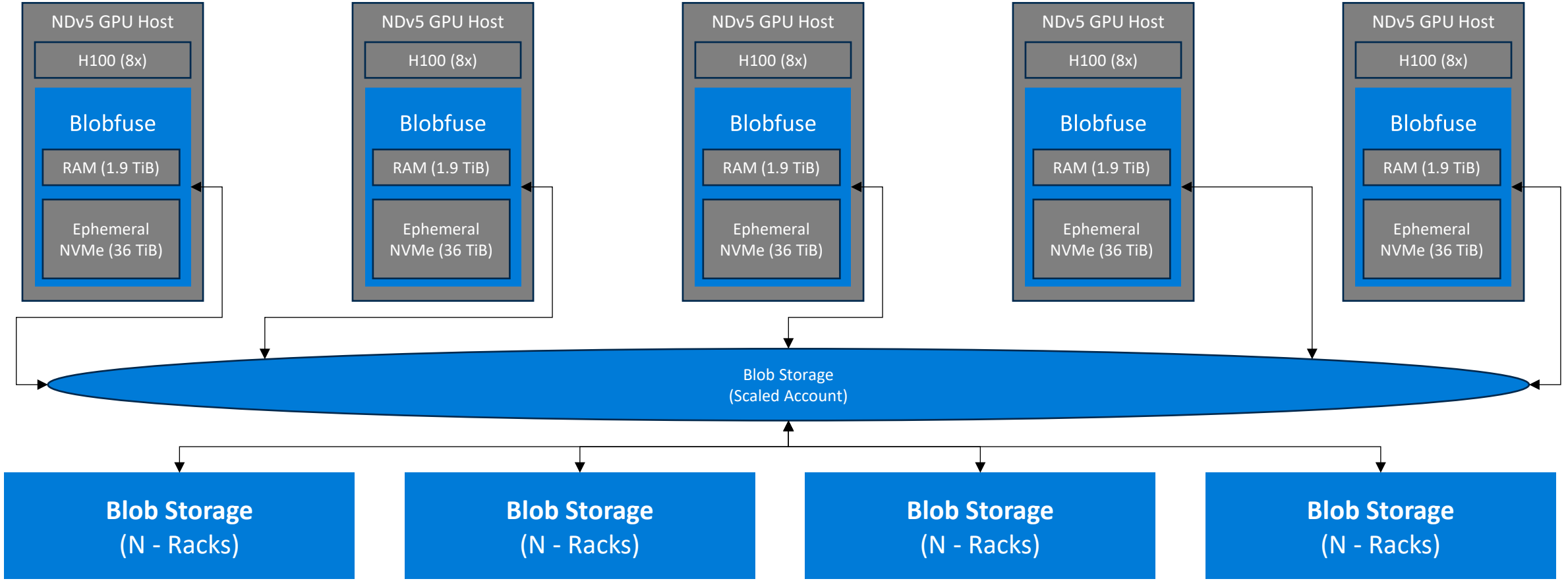
- Seamless RDMA cross connectivity between clusters
- Ability to migrate at Extent Level
- Movement orchestration is scaled out. The target ENs pull the data directly from source
- Scenarios
  - Redirect writes to a cluster that has free capacity
  - Balancing capacity and load across clusters
  - Increasing read bandwidth for existing data

Blobfuse

File folder semantics on an exabyte scale



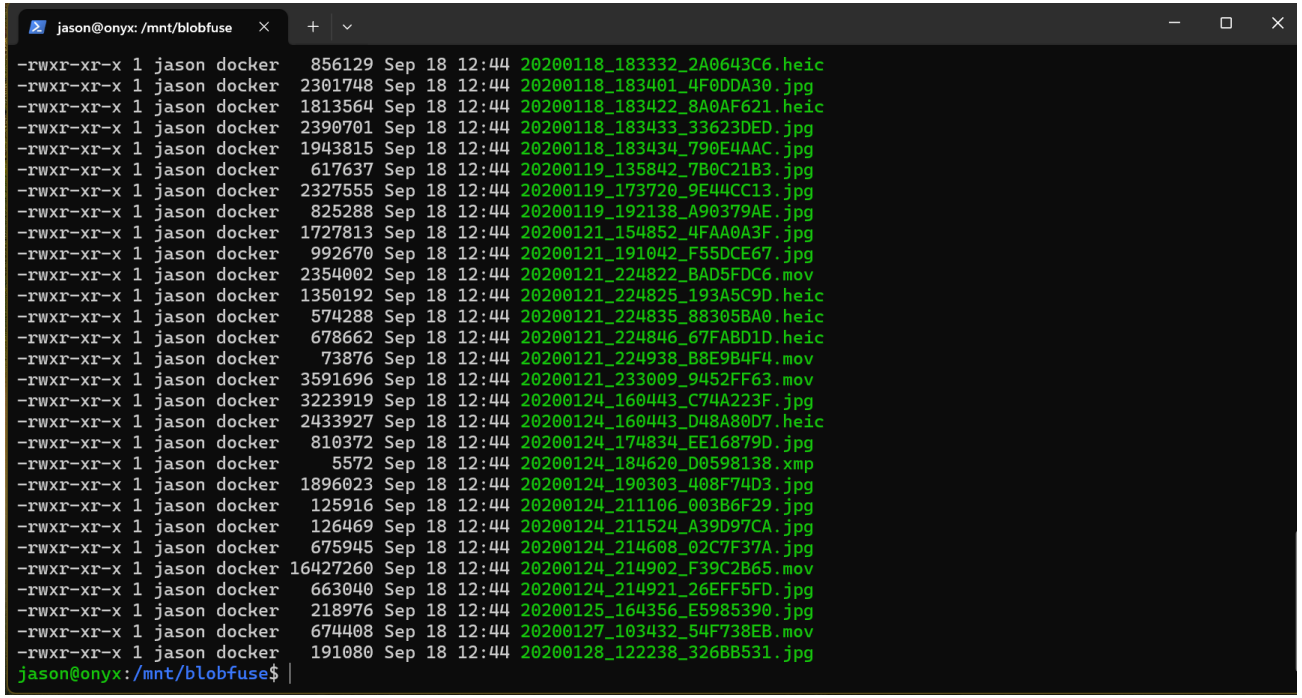
Physical to Logical



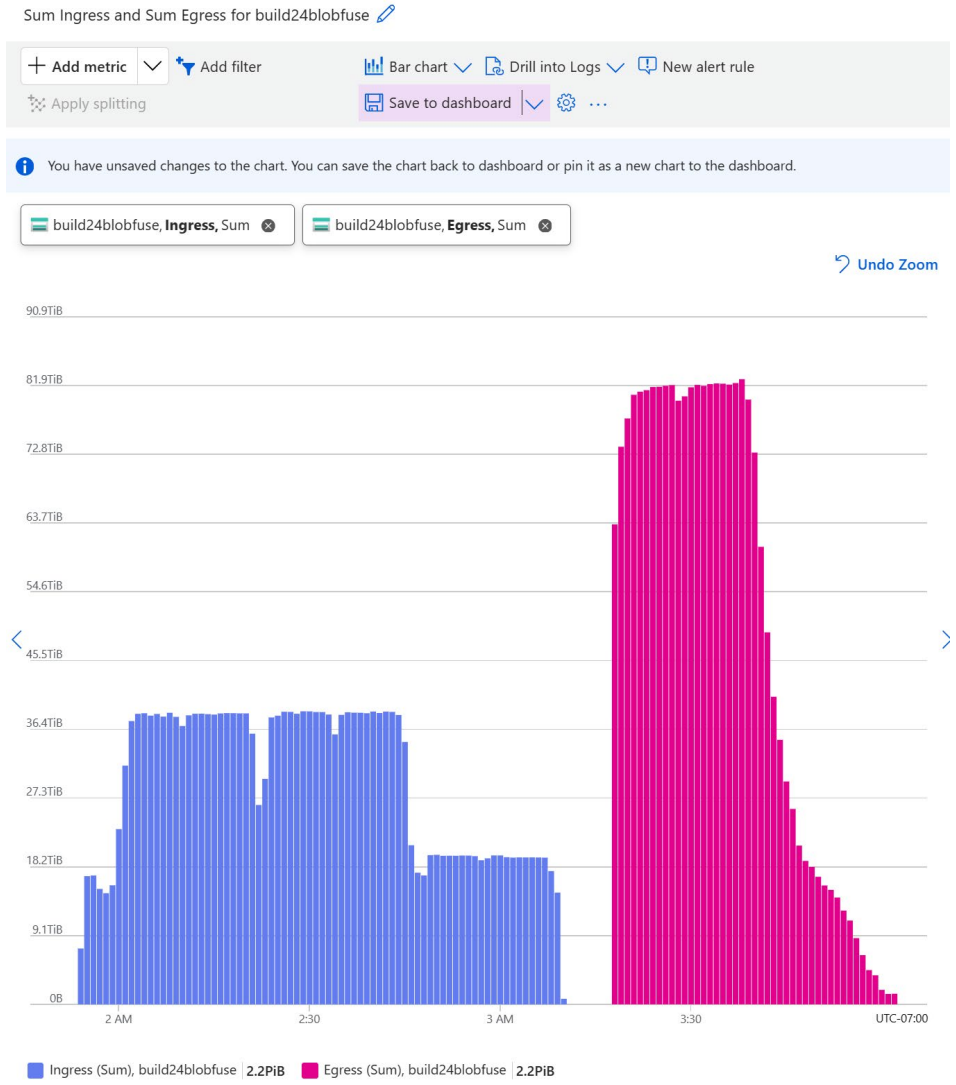
Blobfuse

- Blobfuse provides “filesystem like” semantics and mount point access
- Translates local filesystem calls to REST API
- Hierarchical namespace + Blobfuse enables folder and file level delete/rename

- Caching:
  - Block cache stages data on GPU host memory/NVMe for lazy write back
  - Prefetching of training data to GPU local NVMe *(future)*
  - Cross node distributed caching *(future)*



IOR Benchmark – 16800 vCores



Peak Throughput during the Performance Run on a single Blob container

7-0 Sa or

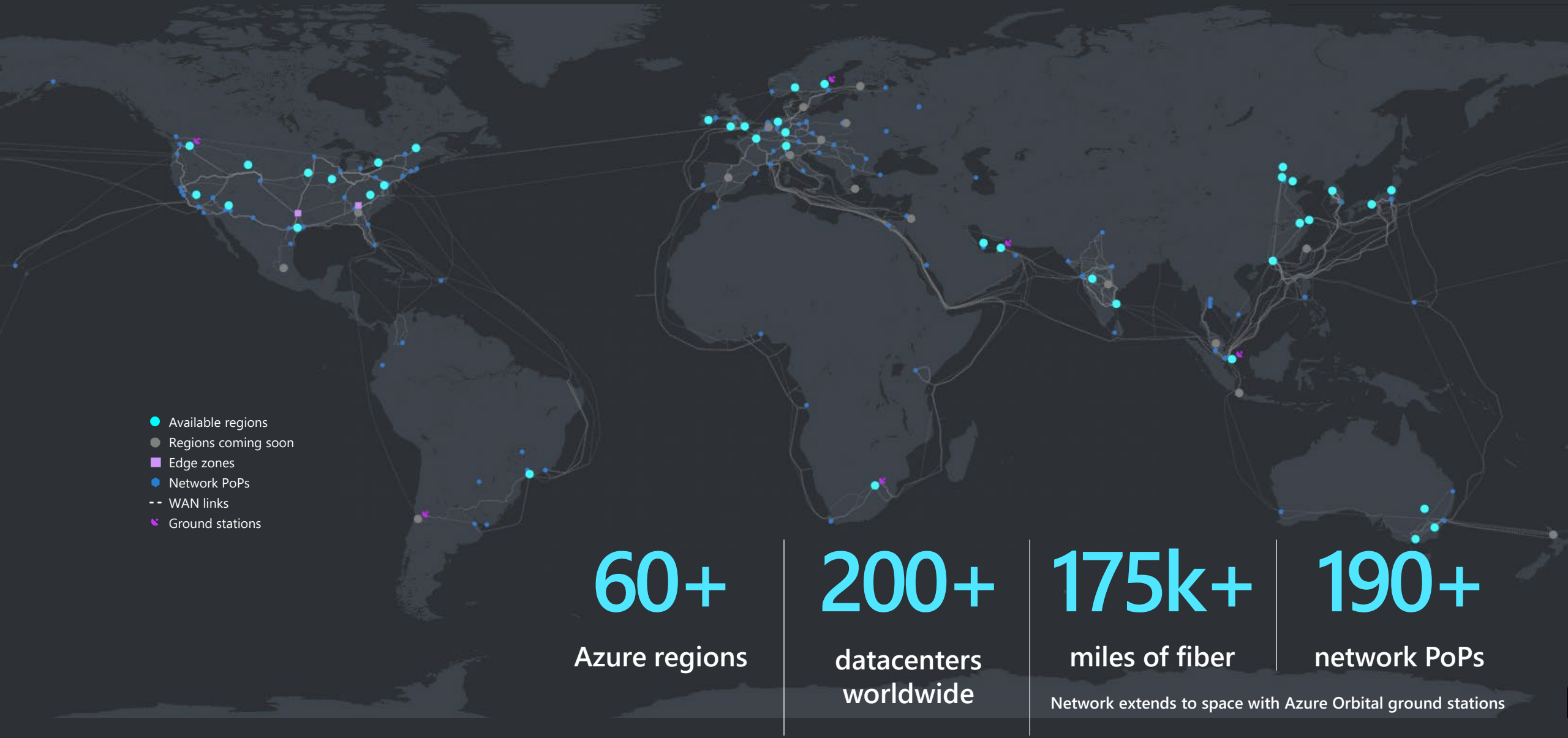
Ingress

02-4 Sa or

Egress

Global GPUs

The herd goes global



60+
Azure regions

200+
datacenters worldwide

175k+
miles of fiber

190+
network PoPs

Network extends to space with Azure Orbital ground stations

Networking

Traffic moving between our datacenters, and between any one datacenter and the edge, all stays on Microsoft's network

190+

Network PoPs in proximity for low-latency network performance

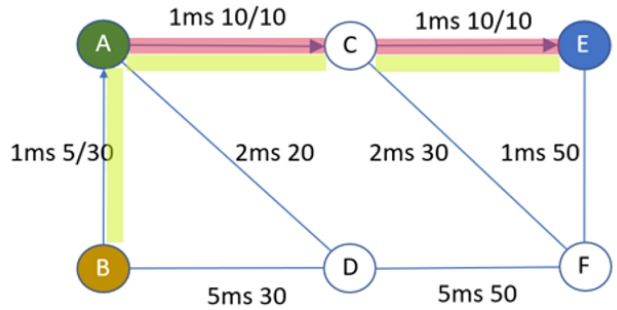
Brokered WAN Traffic

- Splitting traffic among different paths in different proportions
- Pushing the traffic peaks into the valleys
- Sending it on longer, under-utilized paths
- Realize maximum throughput using existing links
- Multi-Tbps cross-region replication

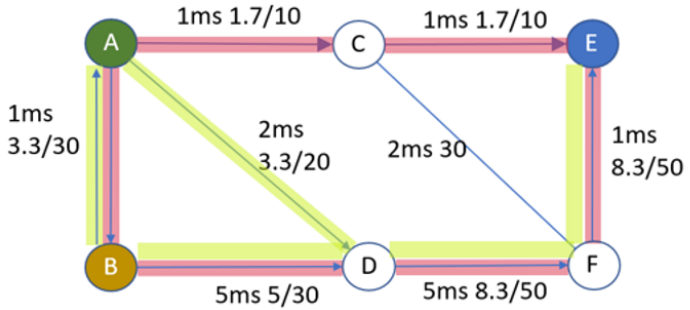
T0
(Optimizes for latency)

Utilization Comparison

T1
(Optimizes for throughput/utilization)



Peak Utilization – 100%
Headroom on hottest link – 0Gbps



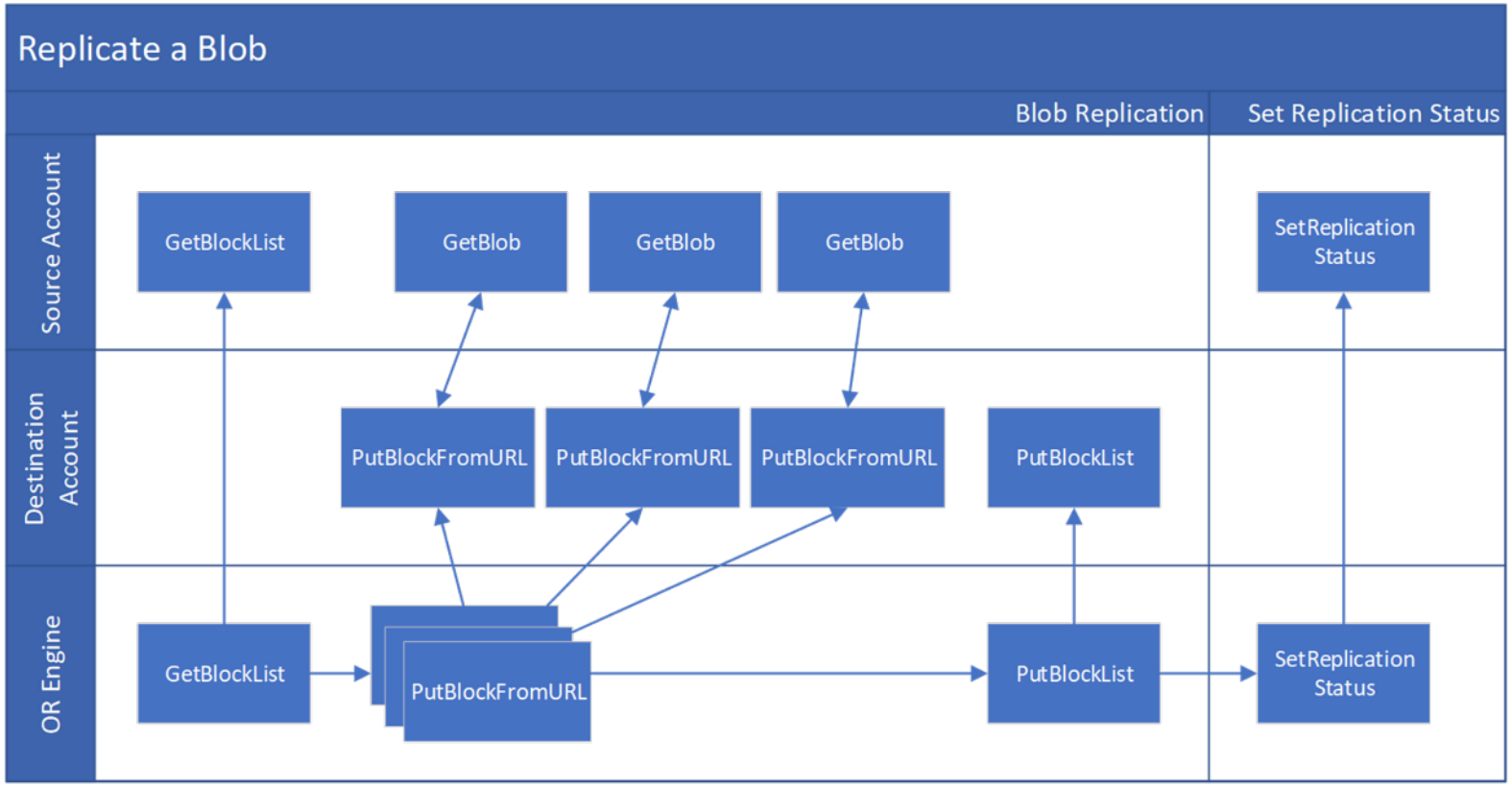
Peak Utilization – 17%
Headroom on hottest link – 8.3Gbps

PutBlobFromURL && PutBlockFromURL

Flow

1. List of objects to replicate
2. Create scoped and signed URL to source object
3. Parallelize call to PutBlobFromURL with signed URL as payload
4. Destination cluster calls GetBlob on source cluster, returning bytes
5. Bytes are committed, and client gets HTTP 200

AzCopy & Object Replication



Key Capabilities for Success

- Scaled accounts enables secure horizontal capacity scale out
- Blob API enables efficient data ingestion pipelines
- HNS enables file and folder rename and atomic delete
- Dynamic partition splitting minimizes hot spots
- Data movement APIs and global WAN

Please take a moment to rate this session.

Your feedback is important to us.

