File Direct – System Description

Volumez

August 18, 2024

Table of Contents

About Volumez 2

About File Direct 3

Benchmark Configuration 4

Benchmark Environment 4

Benchmark Configuration 5

Open Changes 5

Results 6

Discussion 7

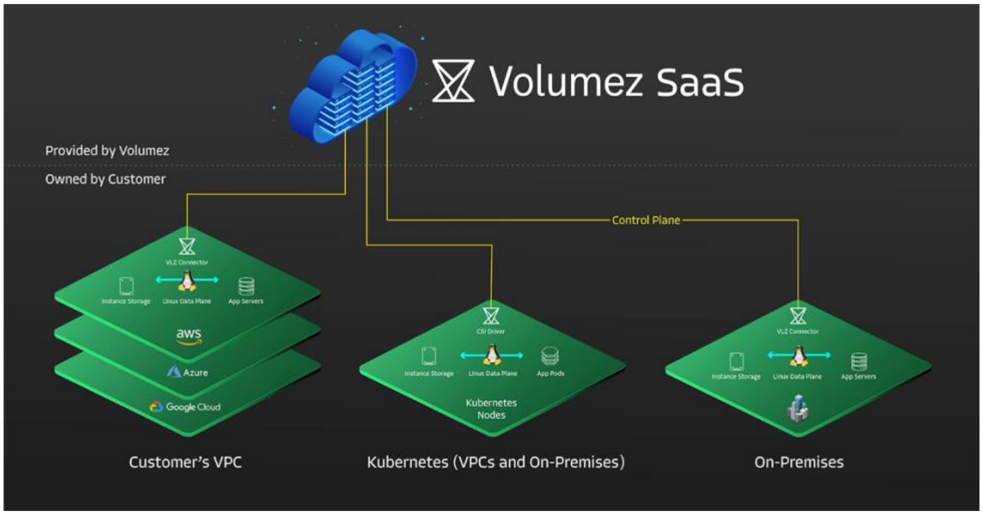
Custom Data loader 7

Data Format 7

Works Cited 8

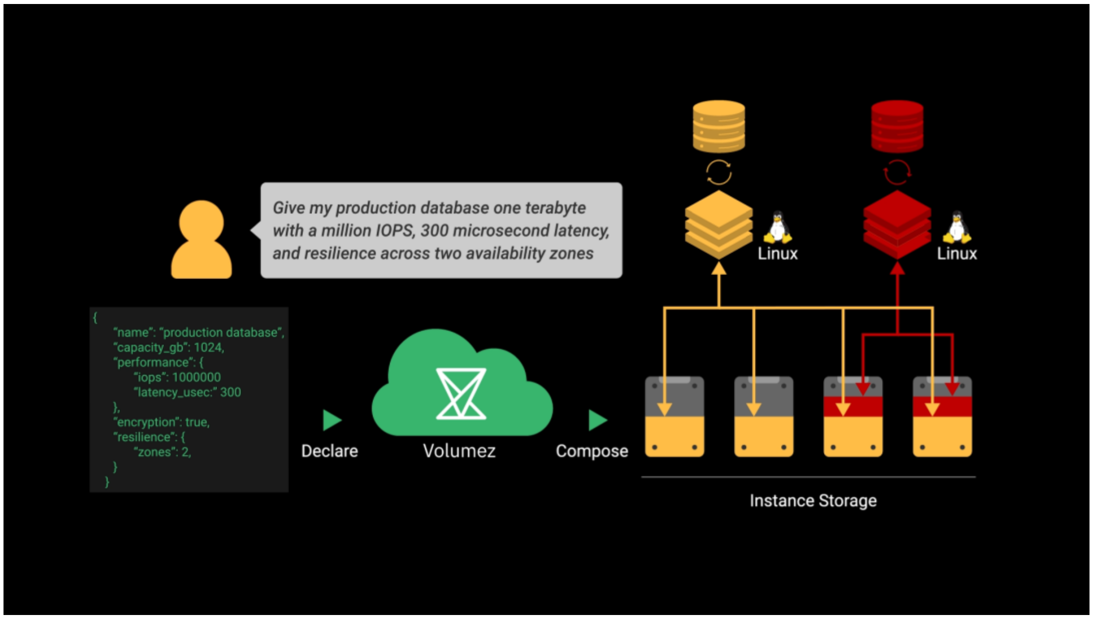
About Volumez

Volumez is a new architecture for block and file storage in the cloud. Volumez separates the storage control plane, hosted in the Volumez cloud, from the storage data plane, which runs in customer virtual private clouds and data centers.



Volumez is not a scale-up or scale-out storage architecture. It is not even a storage system in the conventional sense, although it replaces cloud storage systems in practice.

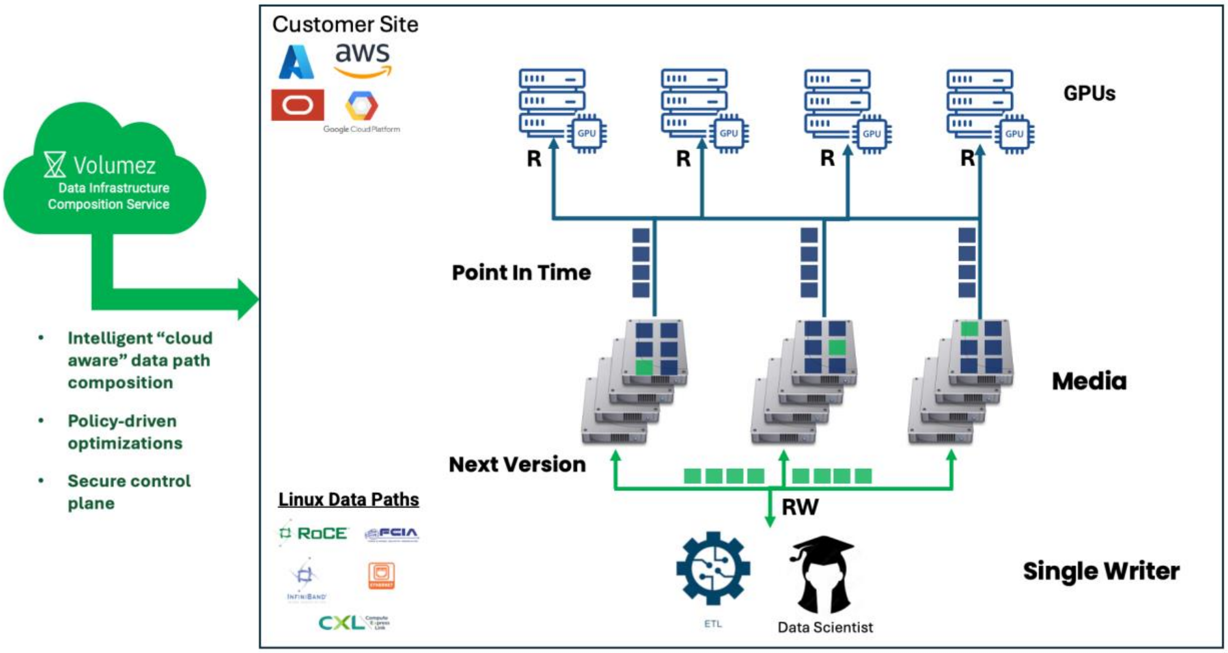
At its core, Volumez is a composable infrastructure software that makes it easy for developers to request storage resources, similar to the way they request CPU and Memory resources in Kubernetes.


Give my production database one terabyte with a million IOPS, 300 microsecond latency, and resilience across two availability zones
Declare
Volumez
Compose
Linux
Linux
Instance Storage

```
{
    “name”: “production database”,
    “capacity_gb”: 1024,
    “performance”: {
        “iops”: 1000000
        “latency_usec:” 300
    },
    “encryption”: true,
    “resilience”: {
        “zones”: 2,
    }
  }
```

About *File Direct*

File Direct is our AI composable data infrastructure solution designed to support large scale training jobs.



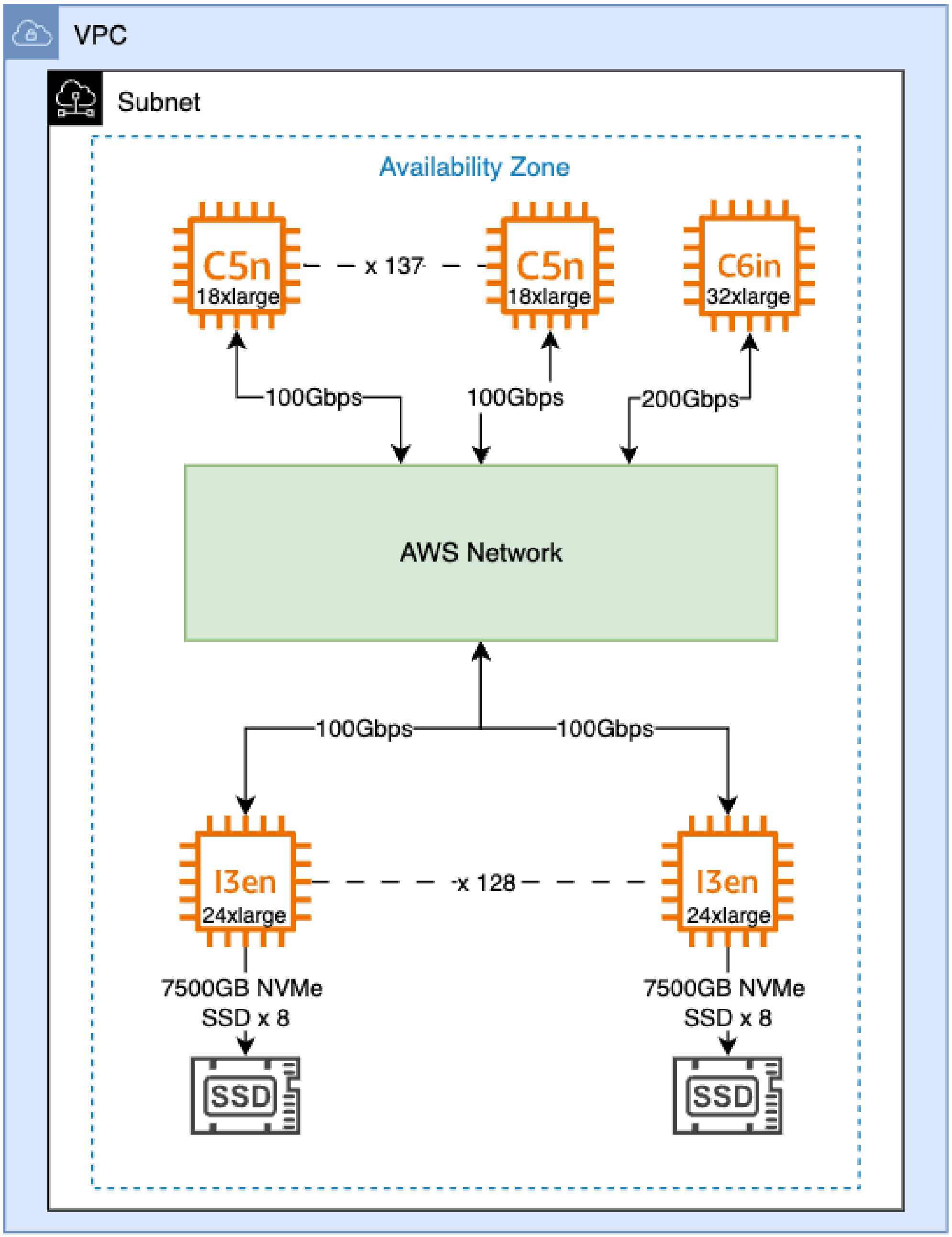
File Direct allows hundreds of GPUs to read simultaneously from the same volume and the same dataset.

It is also possible for a single writer to prepare the next generation of the dataset or prepare for the next experiment while the training job is running with no effect.

The entire data pipeline is Linux native and resides within the customer's environment. Volumez SaaS is only managing the control plane.

Benchmark Configuration

Benchmark Environment



Benchmark Configuration

- 137 host nodes (c5n.18xlarge), running 3 h100 accelerators for unet3d workload.
- 128 media nodes (i3en.24xlarge) orchestrated by the Volumez control plane.
- 1 c6in.32xlrage for data generation only.
- 2 Volumes were created:
    - Data volume
        - Size = 350 TiB
        - Policy = “Performance Optimized”
    - Checkpoint volume
        - Size 100GiB
        - Policy = “Performance Optimized”

Open Changes

- Custom Data loader. -- see description below

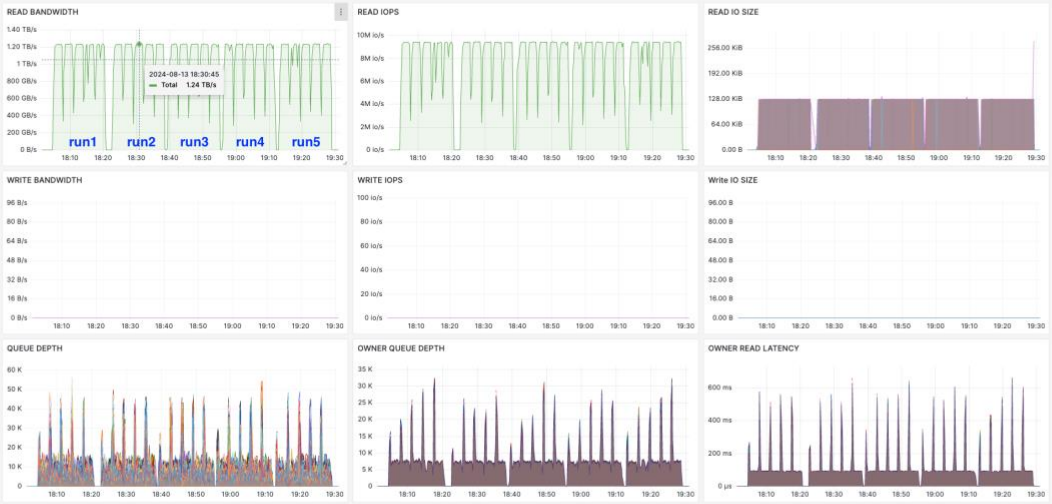
- NPY data format. -- see description below

Results

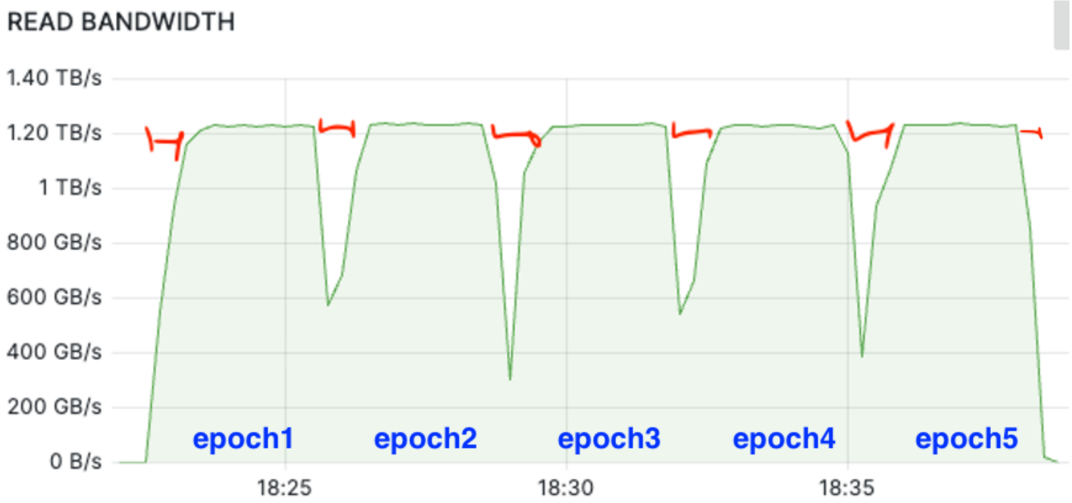
Metric	Average (5 runs)	Std
Throughput	1.075TB/s	5.88GB/s
Samples/s	7718	42
IOps	9.44M	-
AU Utilization	92.2%	0.53%

File Direct was able to achieve an outstanding aggregated throughput of 1.075TB/s, while maintaining an average 92.2% AU utilization as measured by the benchmark.

Volumez monitors are measuring a consistent peak performance rate of 1.25TB/s, as seen in the following image:



The source of the difference is in the benchmark, measuring the throughput over the entire time, including time between epochs that is not yet utilizing the storage.



Discussion

Custom Data loader

Our custom data loader issues O_DIRECT IOs to the storage system. This allows the IOs to be issued to storage directly without passing through the OS page-cache which creates redundant cpu bottleneck and memory pressure.

Data Format

The data was generated in .npy format instead of .npz format. The difference between the format types is that .npz represents compressed tensors, while .npy does not compress the data.

The decompression of the .npz files is compute intensive, preventing the benchmark from scaling and achieving the required AU utilization although the storage can keep up.

The necessity to separate samples online pre-processing (decompression in this scenario) from the compute nodes (e.g GPU nodes) in high scale is discussed in [1].

Works Cited

[1] M. Zhao, N. Agarwal, A. Basant, B. Gedik, S. Pan, M. Ozdal, R. Komuravelli, J. Pan, T. Bao and H. a. N. Lu, "Understanding data storage and ingestion for large-scale deep recommendation model training," *In Proceedings of the 49th Annual International Symposium on Computer Architecture (ISCA),* pp. 1042-1057., 2022.