Windows Azure Storage

A Highly available cloud storage service with strong consistency

Calder, et al.

Presenters: Pankaja A. and Howie Benefiel

Topics

- Introduction
- Design features
- WAS architecture
- Stream layer
- Partition layer
- Application throughput
- Workload Profiles
- Design choices
- Demo

Introduction

- WAS - Windows Azure Storage
- Scalable Cloud storage system, since 2008
- Store limitless amount of data
- Data accessible at all times
- Blobs, Tables and Queues
- Usage: social networking search, managing medical records

http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html

WAS design features

- Strong consistency, high availability,partition tolerance
- Global and scalable namespace/storage
- Disaster recovery
- Multi-tenancy and cost of storage.

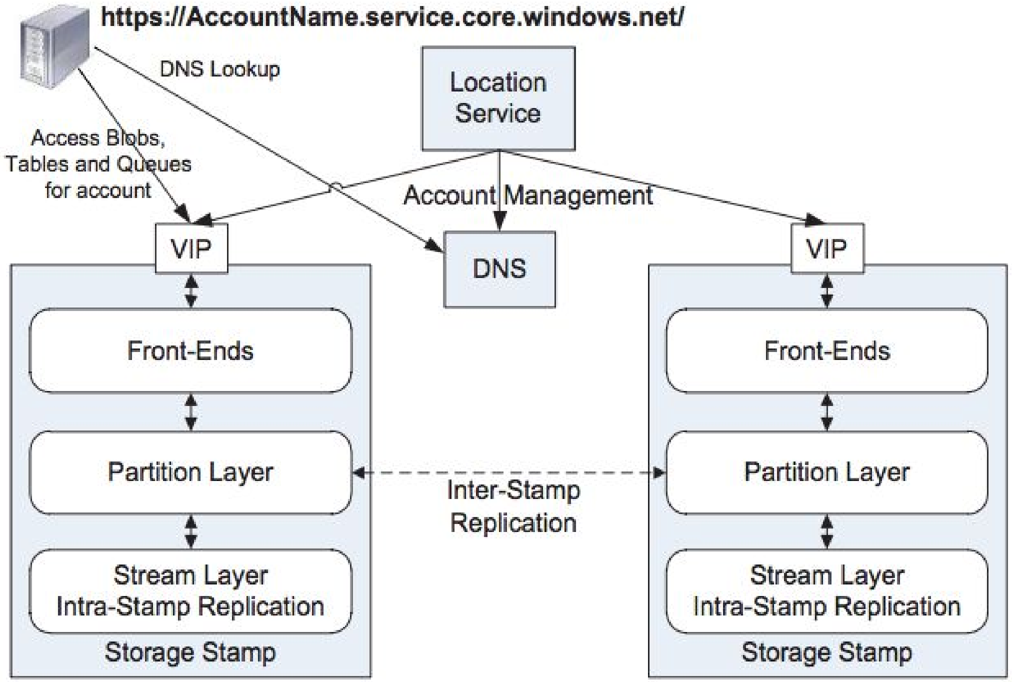
Global Partitioned Namespace

- Single global namespace

- Breaks the storage space into 3 parts -an account name, a partition name, and an object name

- http(s)://AccountName.1 .core.windows.net/PartitionName/ObjectName

High-level Architecture


https://AccountName.service.core.windows.net/
DNS Lookup
Location Service
Access Blobs, Tables and Queues for account
Account Management
VIP
DNS
VIP
Front-Ends
Front-Ends
Partition Layer
Inter-Stamp Replication
Partition Layer
Stream Layer Intra-Stamp Replication
Storage Stamp
Stream Layer Intra-Stamp Replication
Storage Stamp

http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html

WAS Architecture

- Storage stamps – is a cluster of N racks of storage nodes, holding upto 30PBs of data.

- Location services – manages all the storage stamps and account namespaces across all stamps.

WAS Architecture

- Three layers within a storage stamps:
- Stream layer – Stores data, keeps the data durable within the stamp.
- Partition layer - achieves scalability by partitioning all of the data objects within a stamp
- Front-end layer - consists of a set of stateless servers that take incoming requests

http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html

WAS Architecture

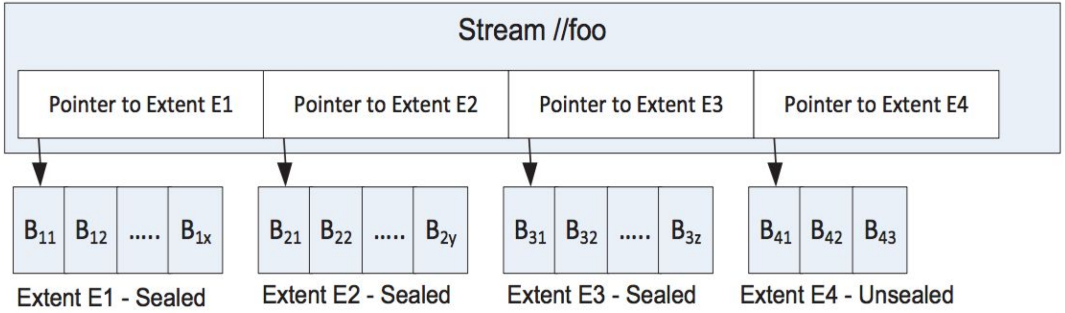
- Two replication engines:

- Intra stamp replication – provides *synchronous* replication. Performed by the stream layer.

- Inter stamp replication - provides *asynchronous* replication. Performed by the partition layer.

Stream Overview

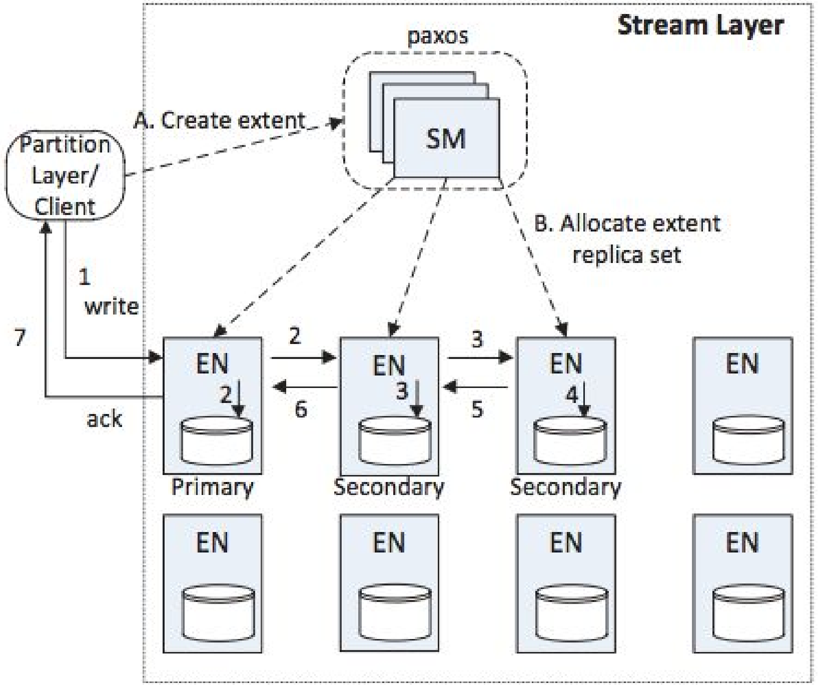
- A stream is an immutable ordered list of pointers which point to “extents”.
  - Looks like file to partition layer
- An extent is an ordered list of blocks which are the fundamental unit of replication.
- Blocks are the fundamental unit of data.
  - *N* bytes, e.g., 4 MB



-c-best-practices-guide/index.html

Stream Layer Architecture

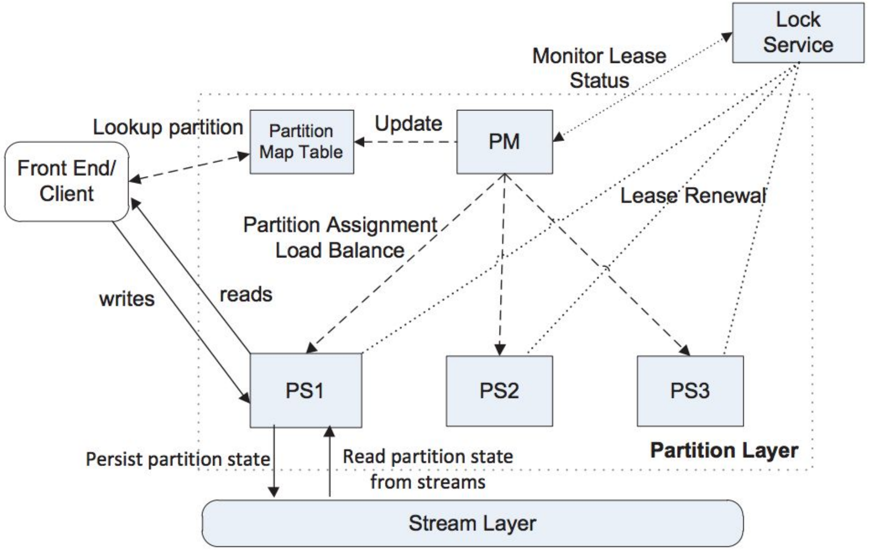
- Contract to client (partition layer) says:
    - if client gets ack on write request, all reads of that data will see same data
    - Once an entity is sealed, all subsequent reads of that entity will see same contents
- Stream manager (SM) sends three replicas of entity to Entity Nodes (ENs)



http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html

Partition Layer Components

- **Object Table (OT)** - Contains data about objects
- **RangePartition (RP)** - Non-overlapping partition of OT
- **Partition Manager (PM)** - delegate RPs to Partition Servers
- **Partition Server (PS)** - Serves requests for assigned RPs
- **Lock Service** - Used to elect leader and ensure one PS per RP



http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html

Partition Layer Operation

- RP composed of in-memory and persistent (stream layer) data structures
- Data Flow
  - FE makes write request -> PL writes to commit log & stores write in cache -> return 200
- If a RP becomes too large or too small, the RP is split or merged by the PM.
- Partition layer handles asynchronous inter-stamp replication

http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html

Application Throughput

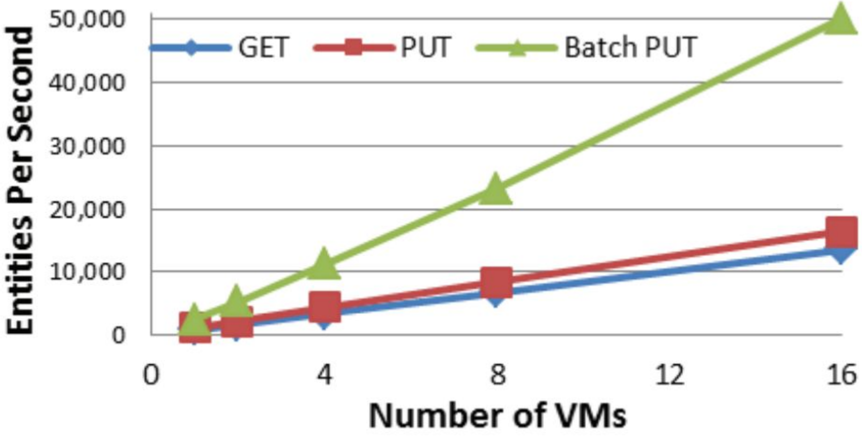


Figure 6 Table Entity Throughput for 1-16 VMs

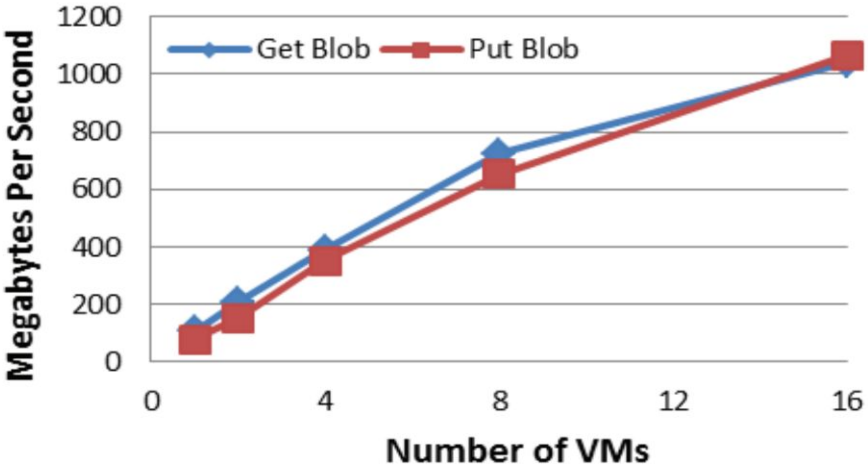
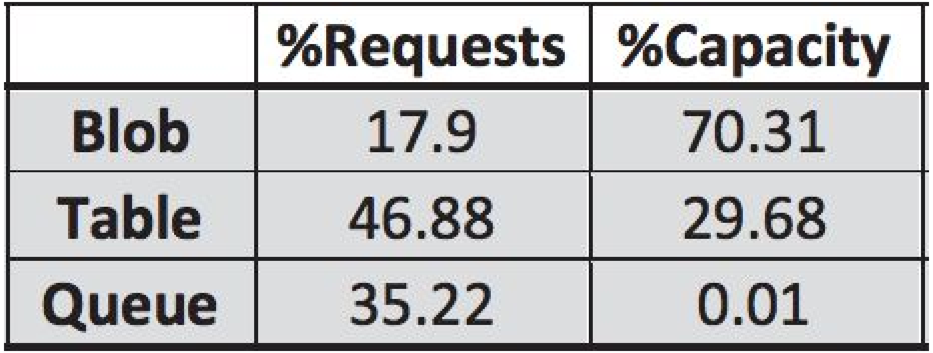


Figure 7: Blob Throughput for 1-16 VMs

http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html

Workload Profiles

- Powers the Zune music storage service!
- Across four workload profiles:



	%Requests	%Capacity
Blob	17.9	70.31
Table	46.88	29.68
Queue	35.22	0.01

http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html

Design choices

- Scaling compute separately from storage
- Range vs. Hash-based partitioning
- Throttling
- Automatic Load Balancing
- Append-only stream layer is extremely important
- Violates CAP theorem?

Demo

http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html

Questions?

http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html