New situation: They are looking to now move not only new customers to the cloud but all of their existing customers. We need to make sure that each of the new types of Mo Vid customer can fit in your proposed architecture. If they do not, explain how you will change and or scale the architecture design to fit all the types of Mo Vid customers

Modified video play Solution:

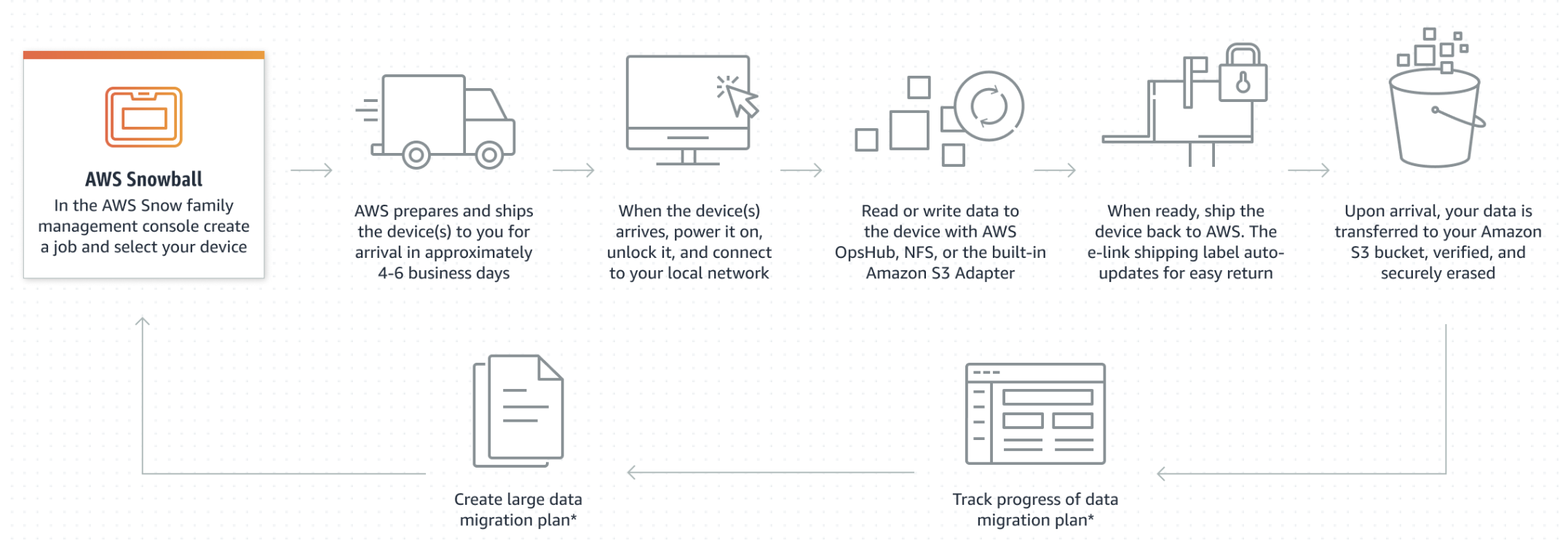
Step1

transfer data

Now we not only need to be able to store video data, but also help the original Large , Medium and Small 3 types of customers to migrate data to our new platform.

For large customer: They could use AWS Snowball. For datasets less than 10PB or distributed in multiple locations, you should use Snowball. In addition, you should evaluate the amount of available bandwidth in your network backbone. If you have a high speed backbone with hundreds of Gb/s of spare throughput, then you can use Snowmobile to migrate the large datasets all at once. If you have limited bandwidth on your backbone, you should consider using multiple Snowballs to migrate the data incrementally

In the AWS Snow Family console, select your preferred device, either Snowball Edge Compute Optimized or Snowball Edge Storage Optimized. Create a job with an Amazon S3 bucket, select Amazon Simple Notification Service (Amazon SNS) for tracking, and configure options like Amazon EC2 AMIs and a GPU. AWS prepares and ships the device to you, and you receive it in approximately 4-6 days. Once the device arrives, power it up and use AWS OpsHub to unlock it. Connect to your LAN. Use AWS OpsHub to manage the device, transfer data, or launch EC2 instances. When done, shut down and return the device to AWS. The shipping label automatically appears on the E Ink screen. When the device arrives at the AWS Region, any data stored in your on-board bucket(s) is moved to your S3 bucket and verified in about the same time it took you to load the device. All data is then securely erased from the device, and it is sanitized of any customer information.



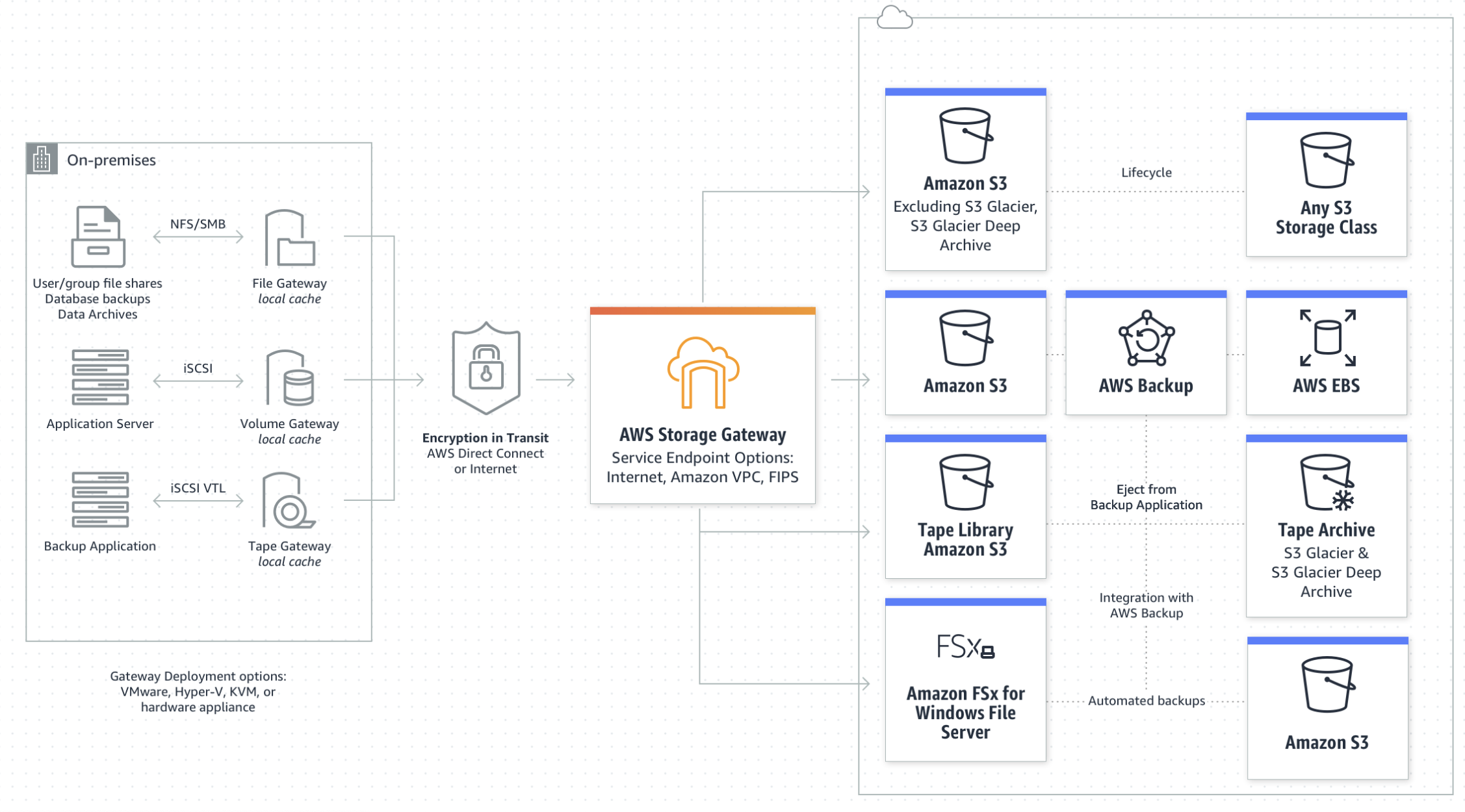
(Reference: <https://aws.amazon.com/cn/snowball/>)

As we are told: “Total video storage is greater than 60TB less than 150TB”, so We do not use snowmobile as: its dataset is not big enough, snowmobile is not suitable .When our data is over 10PB, we will consider it (reference: <https://aws.amazon.com/snowmobile/> )

For Medium and Small types customer: we use AWS Storage Gateway.

The service provides three different types of gateways – File Gateway, Tape Gateway, and Volume Gateway.File Gateway file data is stored in Amazon S3 as durable objects using Amazon S3 File Gateway or in fully managed file shares using Amazon FSx File Gateway.

Tape Gateway virtual tape library (VTL) configuration seamlessly integrates with your existing backup software for cost effective tape replacement in Amazon S3 and long term archival in S3 Glacier and S3 Glacier Deep Archive.Volume Gateway stores or caches block volumes locally, with point-in-time backups as EBS snapshots. These snapshots may be recovered in the cloud.



(Reference: https://aws.amazon.com/storagegateway)

Step2

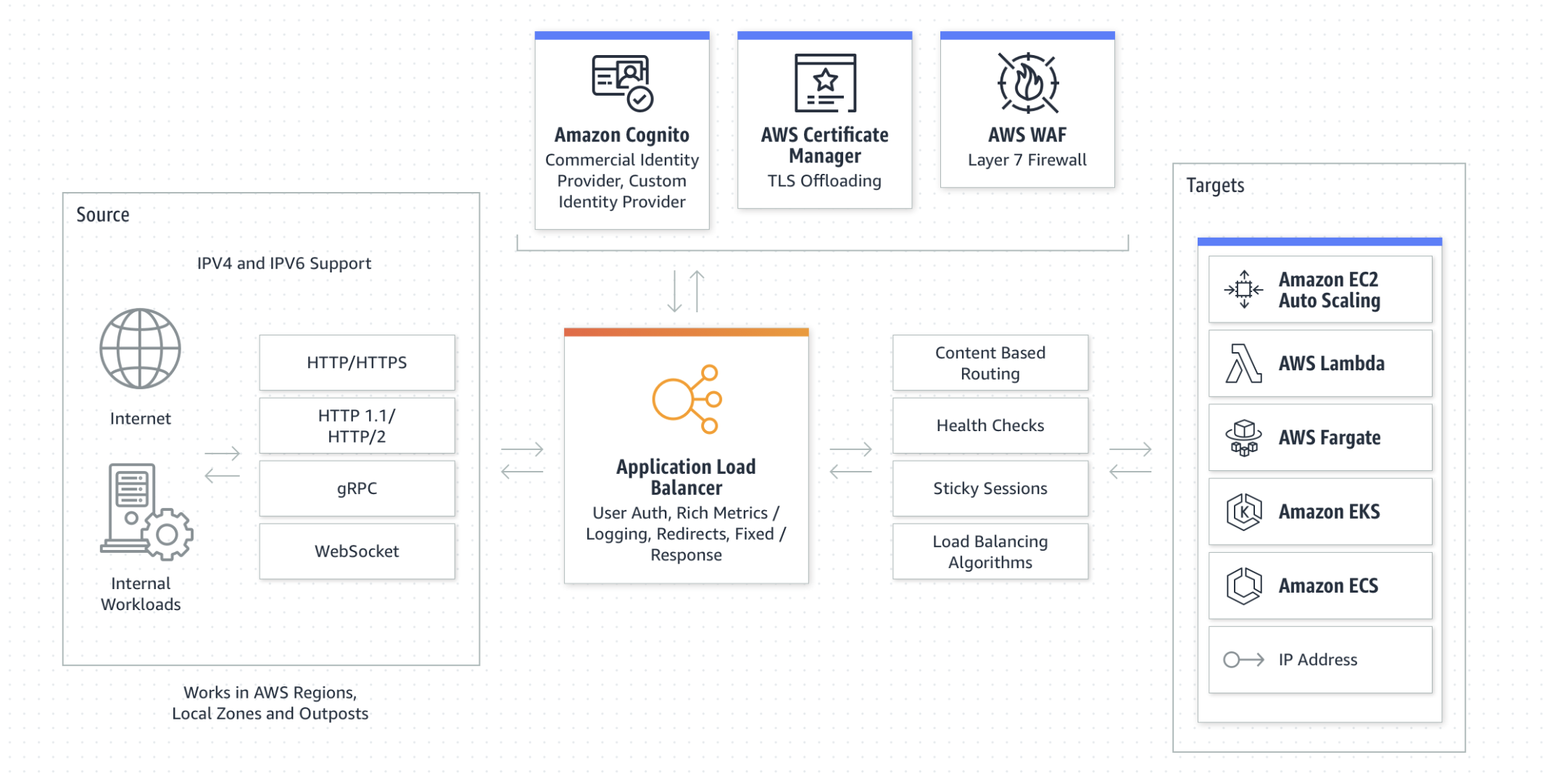
Increase robustness

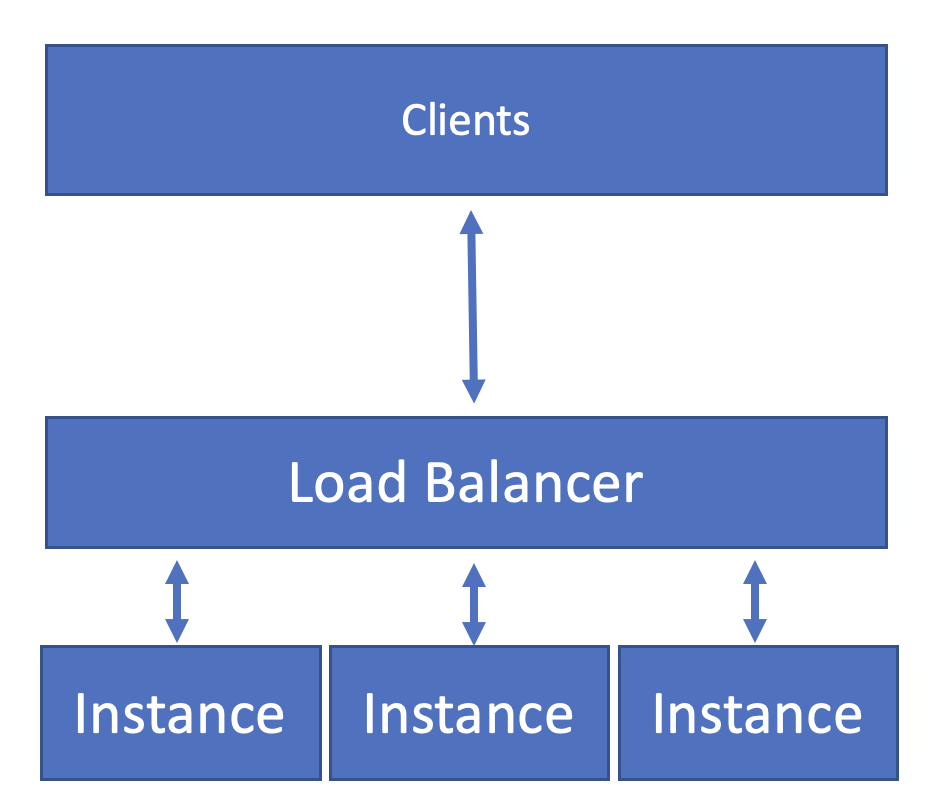
As all our customers are now on cloud, It greatly improves the requirements for service operation and maintenance robustness and automated maintenance.

We will consider to use ELB of our web systems part: user account login/register/comments/role managements/Authorization management for video playback and CRUD.

Elastic Load Balancing (ELB) automatically distributes incoming application traffic across multiple targets and virtual appliances in one or more Availability Zones (AZs).

It can realize automatic operation and maintenance management, automatically discover, identify, manage, and allocate disks, intelligently schedule applications and data according to affinity, and automatically monitor disk status and give timely warnings. Second, high-availability data support. Our video load balancing uses cross-node replicas to synchronize data to achieve high availability. When a problem occurs, it will automatically schedule the application to the high-availability data node to ensure the continuity of the application. Third, with a variety of data volume types, our video load balancing aggregates HDD, SSD, and NVMe disks to provide non-low-latency, high-throughput data services. Fourth, flexible and dynamic linear expansion can be achieved. Our video load balancing can dynamically expand capacity according to the size of the cluster to flexibly meet the data persistence requirements of applications.





(Reference: scaling.pptx and <https://aws.amazon.com/cn/elasticloadbalancing/> )

Step3

Support multiple situations:

Video playback and return visits, the application scenarios of various user operations mainly include the following three categories

1. Adapt to high-availability architecture middleware

Kafka, ElasticSearch, Redis, etc., such middleware applications have their own high-availability architecture, and at the same time have high requirements for IO access to data. The LVM-based single-copy local data volume provided by our overall solution can well meet their requirements.

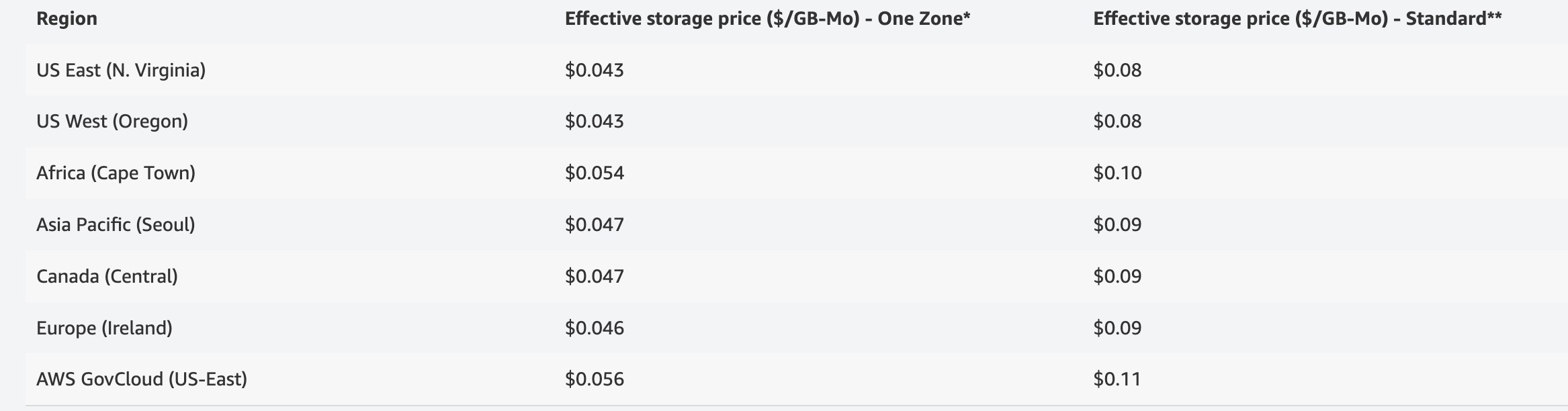
2. Provide highly available data volumes for applications

OLTP databases such as MySQL/MSSqlserver require the underlying storage to provide highly available data storage, which can quickly recover data when a problem occurs, and also require high-performance data access. The dual-copy high-availability data volume provided by our overall solution can well meet such needs.

3. Automated operation and maintenance of traditional storage software

We are considering use AWS EFS or MinIO, Ceph and other storage software need to use the disk on the K8s node. PVC/PV can be used to automatically use the single-copy local volume of our overall solution through the CSI driver to quickly respond to the deployment, expansion, and migration proposed by the business system. and other needs, realize automatic operation and maintenance based on K8s. Scale our video file system automatically as files are added, removed, and burst to higher throughput levels when necessary.

Amazon EFS file systems can automatically scale from gigabytes to petabytes of data without needing to provision storage. Tens, hundreds, or even thousands of compute instances can access an Amazon EFS file system at the same time, and Amazon EFS provides consistent performance to each compute instance. Amazon EFS is designed to be highly durable and highly available. With Amazon EFS, there is no minimum fee or setup costs, and you pay only for what you use.(reference: scaling.pptx and https://aws.amazon.com/efs/)

If we have more budgets , we could use ifs; if not , we can construct a similar file systems use open source systems like(Cph/K8s) on AWS. Following is price of Los:

Conclusion

In order to prevent server-sprawl, we consider to manager servers more automatically and accurately.

To automatic more actions: we use software-defined infrastructure, the top layer is a high-performance distributed file system, which is also the core of our overall solution. This user-mode file system is completely self-developed, fully compatible with the POSIX standard, and realizes zero data copy between user mode and kernel mode. Fully exploit the performance of high-speed media such as NVMe SSD/SCM/PM. he second layer is the data service layer. In addition to providing services between local storage of different nodes, between local storage and external storage, and between heterogeneous external storage, it can also provide services for single copy and multiple copies, strong consistency and weak consistency. Provide storage solutions separately. The third layer is the storage management layer, which is responsible for interacting with different storage devices, breaking device barriers and data islands through a unified data format, and allowing data and services to flow freely across heterogeneous devices. In addition, the system can also schedule storage devices through the CSI interface.

To manage more accurately: we use more Virtualize servers not real machines. Individual servers can also be consolidated and configured as standalone server workloads on powerful servers. Excessive reliance on virtualization brings its own virtual sprawl issues, so virtual servers should be provisioned carefully

Careful management of IT assets (reference clould\_capacity.pptx) and the use of capacity planning tools can help stop server sprawl before it begins. These tools can help track existing servers and determine when servers are running below capacity or when demand for servers has peaked and server expansion is required