caringo®

Protecting Data in the Cloud Age with Caringo Swarm Object Storage

by Don Baker, Caringo Swarm Lead Developer

With the onslaught of data that organizations are experiencing in the cloud age, protecting all that data has never been more important. Storing and providing access to your stored data may be the most visible thing that Swarm does, but Swarm also protects your data. Our founder, Jonathan Ring, uses a metaphor for Swarm as a ship carrying and protecting your data as the river of hardware changes over time, either through upgrade or eventual failure.

KEEPING ALL BASES COVERED

It's worth explaining the various data protection mechanisms in Swarm and how a Swarm cluster can serve as the unit data protection even if parts of the overall system fail. Swarm has the bases covered when it comes to protecting data. Even before there's a failure, Swarm is actively protecting your data. Our basic strategy is to make multiple replicas of your data so that we are never putting "all our eggs into one basket." A central function of the Swarm health processor is to maintain the right number and placement of replicas of all objects in the cluster despite changing conditions. The number of replicas Swarm stores for an object is up to the cluster administrator. More replicas gives higher protection at the expense of more space used.

Many customers choose three replicas, which protects all data in a cluster against 2 simultaneous disk losses at a cost of 3 the logical data stored. All replicas are completely equivalent and so there is no risk of some special replica being more vulnerable than any other. Those three replicas are placed in locations within the cluster that are unlikely to fail at the same time and we make sure that all replicas are faithful copies by checking computed hashes during transfer.

These mechanisms extend to erasure-coded objects which may allow for high protection and a smaller data footprint at the expense of many parts that we call segments that must be assembled to reconstruct the object. Segments, too, must have independent failure modes and they are subject to similar health processor checks to make sure all the segments comprising an object are present and properly placed for optimal protection.

As an aside, a great feature of Swarm is that we can protect different objects with different encoding schemes and different levels of protection all in the same Swarm cluster. This is a patented feature of Swarm that no other storage provider has. We even allow these protections to change over time, so you can protect your data more when your data is more valuable and less when it's not as valuable.

PROTECTING DATA IN TRANSIT

Swarm can protect your data in transit over the network using integrity seals (hashes) that prevent tampering or transmission errors. These seals are stored with each object and are re-verified when the object is retrieved from Swarm. A write or update request can even request that the full complement of replicas is made before the requests finishes.

MANAGING DISK FAILURES

Ultimately, Swarm data is stored on mechanical disks which provide high data density, inexpensive persistence, and fast transfer speeds. Although it's a rare thing, disks can drop data due to bad sectors. Those bad sectors may hold one or more of your objects, all of which are unreadable. The Swarm health processor will periodically read every object on disk to verify its data integrity. If this check fails, the replica is declared invalid and a new replica is made in its place using duplicate data residing elsewhere in the cluster. Here, we rely on the fact that we check the data much faster than the failure rate, so these sorts of disk read/write errors don't result in lost data.

More commonly though, disks just go bad. They can do so slowly over time or quickly and catastrophically. For slow failures, Swarm watches for an accumulation of disk errors that is usually an indication of an impending failure. When a disk is determined to be near failure, we "retire" it. Retiring involves the cluster performing an active recovery of the questionable disk without increasing the load on it. As sufficient replicas are made elsewhere in the cluster, the retiring disk can erase its replicas, ultimately resulting in an empty disk in the cluster than can be replaced at the administrator's leisure. All of this is done automatically and without any potential for data loss.

RECOVERING DATA

Even when a disk failure is catastrophic, Swarm has a few tricks up its sleeve. The entire cluster participates in an active recovery that quickly restores the cluster to full replication in a relatively short amount of time. By making the recovery fast, we shorten the window of time during which another disk failure might impact the cluster. If a customer chooses 3 replicas for an object, the two remaining replicas effectively guarantee that a third replica is rapidly made. Yes, these active recoveries have a small impact on cluster performance, but they are often just minutes long, which is a small price to pay for closing the window of vulnerability. Usually, the only thing that slows active recovery is a customer who does not leave enough empty space in a cluster for active recovery to replicate the content of failed disks.

Larger clusters raise some interesting questions. It is true that larger clusters can expect to see more frequent disk failures, that is, a failure of some disk in the cluster. This is simply because larger clusters will have more disks and even highly reliable disks, taken as a large collection, will have a relatively frequent failure of some disk in the collection.

But Swarm's active recovery uses the entire cluster in the recovery which means recovery time in a larger cluster is shorter due to the parallelism of cluster resources devoted to it. While the math is a bit complex, larger Swarm clusters offer comparable data protection to smaller ones, despite the higher likelihood of an individual disk failure in a larger cluster.

SURVIVING CATASTROPHIC FAILURES

What about more catastrophic failures? If a chassis is lost, it's most likely due to a power supply or motherboard failure. Swarm disks can just be moved to another chassis without much effort and no data will be lost. Active recovery may be counterproductive in this case as the recovery may fully replicate the contents of the chassis' disks faster than the administrator can provision another chassis and move the disks. In this case, suspending recovery may be the prudent thing. But even if all the disks are gone, active recovery will do the right thing and recover all disks that were on the chassis. Swarm's health processor will prevent the co-location of replicas or segments of the same object on the same chassis, so even though multiple disks are involved, all the data lost is just a single replica or segment of a possibly larger number of objects. So with more recovery time, Swarm easily handles a chassis loss.

Swarm allows a cluster administrator to define logical subclusters. This feature allows cluster administrators to tell Swarm of even larger units of failure, involving say, power supplies or network structure. When this feature is used, Swarm will spread replicas and EC segments across the subclusters so that data remains accessible and recoverable, even if there's a subcluster outage or loss.

You might think that's the end of the story, but Swarm can even protect your data from the loss of an entire cluster! We allow the entire contents of one cluster to be replicated to another cluster for disaster recovery or other purposes. It takes just minutes to set up a replication feed which can, depending only on network bandwidth between the clusters, maintain a near perfect backup of the original cluster in a distant physical location.

Data protection is a core function of Swarm that leverages cluster resources to protect your data all the way from bit errors to natural disasters. You can do so with full knowledge of the protection and resource trade-offs and make the best decision for your needs. You can even make different decisions for different types of data. If you are serious about protecting your data, check out Swarm 8 with even more features to protect you from common user errors.