STORAGE SWITZERLAND

BUILDING AFFORDABLE, SCALABLE STORAGE INFRASTRUCTURES

George Crump, Senior Analyst

The need to cost effectively store large amounts of data for long periods of time, in some cases forever, has never been greater. Organizations that offer online application, online storage and long term archives are all trying to build affordable, scalable storage infrastructures. But they are facing challenges because legacy storage solutions can't fulfill these conflicting requirements.

The problem is that legacy solutions don't typically offer the scale both in terms of capacity and performance that these modern era use cases require. Even many of the current scale out architectures can't scale to meet today's needs, or those of the future. In addition they suffer from inefficient resource and power utilization which, once again, hurts affordability. There is a need for a new storage paradigm that is not held back by the file system structures of the past.

The Challenges facing Legacy Systems

Scaling Challenges

Legacy systems typically fall into two categories. First there are scale up systems that are dependent on the horsepower of a finite number of storage controllers or CPUs. In this case the scalability of the system is 'pre-bought', meaning that for the most part the total scalability of the system is known upfront, it's just not used. It also means that in the early days of system use much of the performance capabilities go to waste because only the capacity component can actually "scale". The processing component has to be pre-paid. This of course also means a waste of budget dollars since the cost of performance and capacity will get cheaper over time and neither of the resources should be pre-bought.

The other type of storage is the legacy scale out architecture in which the controller function is distributed across a cluster of servers, called "nodes". As data is written to the cluster it's segmented and written across the nodes in the cluster. When more capacity and performance are needed more nodes are added to the cluster, data is segmented even further and more nodes participate in reads and writes. These systems seem like they should be able to deliver the scale needed for today's enterprise. The reality is that they may have limitations when implemented in online environments. They were not designed to handle the mass scale that these environments require and as their node count grows they become inefficient. A scale out storage cluster delivers the aggregate performance of the processors, I/O and capacity of its member nodes. When the node count reaches double and triple digits the total processing power and I/O capability is often overkill for the tasks at hand.

As is the case with the dual controller architectures processing power is wasted, as is IT budget, which makes the system less affordable upfront.

Retention Challenges

Another requirement of the modern storage infrastructure is the need to keep and retain data for many more years than was previously considered. This is more than just a capacity scaling issue, it's a flexibility and technology integration issue. For example most will agree that the hard drive used ten years from now will be significantly different than the hard drive used today. The manager of this online archive is not going to want to cut over to a new system overnight. In fact because of the sheer capacity, potentially hundreds of petabytes, any sort of complete change out, even from one disk system to another, is unthinkable. What's needed here is a scale out storage system that allows for a perpetual migration of nodes over an extended period. This will allow new technology to be implemented over time as the need justifies.

Data Availability Challenges

Beyond wasted resources the primary challenge facing legacy scale out systems is that they are still dependent on traditional data protection and file systems where the node count could easily reach triple digits within a few years. Scale up systems are also susceptible to these issues but may never encounter the file system limitation problem since often multiple systems have to be added (and managed) to deliver scale before those any file system problems are manifested.

Data protection in legacy systems is equally challenging. Those systems typically use RAID 5 or RAID 6 which consume up to 30% of the available disk capacity, wasting storage. More importantly, the time to rebuild a RAID group after a drive failure continues to increase with drive capacity. For projects like online applications, online storage and long term archives, the desire is to use as large a drive size as possible to help keep capacity costs in check. This means that the first drive failure could put data at risk (of a second drive failure) for a long time while the RAID system rebuilds. As a result the storage manager is faced with choosing cost savings or the possibility putting data recovery in jeopardy.

File System Challenges

Once these system level challenges are understood and workarounds applied (or in most cases lived with), the storage manager has to deal with the biggest challenge of all - the file system. While file systems have continued to grow in capacity, in most cases individual volume sizes are not adequate. This leads to multiple, in many cases hundreds, of volumes created to support a file system. Each volume that's added to a NAS increases management overhead, lowers capacity efficiency and makes it more difficult for users to find and access data.

As mentioned above there's also a file system issue as the storage environment scales. The problem has to do with the number of files that these systems have to manage. With traditional file systems each file has its own entry in a metadata table which is similar to a database and these file systems degrade in performance and stability as the metadata grows. The performance problem has become severe enough that some NAS vendors are relocating metadata to a solid state disk tier, which once again adds to the cost of the system. And metadata growth impacts reliability. Similar to a database, as metadata grows it's also more susceptible to corruption. For this reason most file systems have a hard limit on the number of files and or metadata entries that they will support. For the legacy data center these numbers seemed out of reach, but for data centers that need to support online applications, online storage or long term archives these limits can easily be reached and can cause significant challenges.

The Affordable, Scalable Storage Infrastructure

Data centers that support online applications, online storage and long term archives generally have the need to offer top end performance to thousands of users and store billions of files for decades. This new data center needs a new storage paradigm to address these challenges. This new generation of storage systems needs to perform well while being affordable and scalable, like those from Amplidata.

Affordability can be delivered in two ways. First by being less expensive and second by being more efficient. Scalability in this new paradigm means making sure that these affordable systems can grow to meet the demands of the online applications and projects. Lack of scale eventually drives up costs, so it's also an important factor in keeping the systems affordable. Both factors are interrelated and it is impossible to have a discussion on the new storage paradigm if both are not addressed in unison.

Affordable Scaling

The first step for these new systems is to use affordable, off the shelf hardware components and build them into a scale out storage infrastructure similar to the legacy storage systems of the past. The current scale out storage systems though, often used proprietary hardware and special node interconnects, which drove costs up and flexibility down. The other key to maintaining affordability is to make sure that this new storage paradigm does not waste node resources like memory and CPU, especially as the node count increases.

To accomplish this companies like Amplidata use the more efficient Intel ATOM processor instead of a full powered equivalent. Not only are these processors less expensive they also use substantially less power and since they operate at cooler temperatures, can be placed in denser enclosures which require less space. This one design decision saves hard costs, saves power and saves data center floor space, three critical factors driving affordability. Also, because their scale out architecture aggregates each node's CPUs, they still provide excellent performance.

Future Ready

Another key attribute for these systems is the ability to serve data for 10 or more years, and potentially much longer. This again means a break from the traditional scale out storage model where the nodes were tightly coupled and usually had to be identical in both disk size and processor type. The modern scale out architecture needs to support different types of nodes so that as years go by the latest hardware can be used, and allow for a self migration to new technology as it becomes available.

Better Data Protection

As the capacity of the individual drive increases as well as the capacity of the architecture, the scalable storage system has to use an alternative data protection technique than the classic RAID algorithms. For example, Amplidata's AmpliStor uses a modern variant of erasure coding data protection techniques, termed BitSpread. BitSpread encodes reliability into data at a check-block level, so that if a drive fails the system only needs to generate some new check-blocks, not the entire drive and not even the exact check-blocks that were lost. BitSpread can also protect against any number of disk failures and provides full protection against media errors, such as unrecoverable read errors (URE) also known as bit rot. This technique also extends efficiency, since the capacity lost to create this redundancy is less than with traditional RAID technologies. Better efficiency equals better affordability.

This level of data understanding also allows for better drive diagnoses. For example when a RAID system fails a drive, often much of the drive is still viable, but it has reported enough errors to fail the whole drive. A more intelligent system will only fail a portion (maybe a cylinder or just a sector) of the drive that is reporting the errors. This saves data copy time and increases capacity efficiency.

The key attribute is to move away from the overhead and constraints of a traditional file system and to more of an object based storage system. In an object based system data is accessed directly by object name, there is no metadata table overhead. Interestingly object based systems can still have metadata but it's stored with an object ID number in a flat file, not in a separate table. By accessing data directly and not going through a metadata table these systems can now scale in file/object count to match the ability to scale capacity and all under one volume. More importantly the object based environment has lower overhead and requires less processing power from the individual storage nodes, once again justifying the use of a lower power processor while maintaining performance.

Object based systems can also be manipulated directly by the application through an API set. This means that the application can store or read the data directly. It also means that the application can set metadata information like retention times, encryption levels, write status or number of copies. These are attributes that users should set but often don't. And the application is probably the next best qualified system to make a determination on what these settings should be.

The online application, online storage and long term archive projects that data centers are now embarking on have been roadblocked by legacy and even current storage systems. By using affordable, scalable storage infrastructures like those offered by Amplidata, organizations that have internal online applications, online storage and long term archives as well as organizations that provide these functions as a service, now have the ability to make sure storage remains manageable and cost effective.