Empowering the Cloud to Manage Non-Flash Data

November 2, 2015 by George Crump

An all-flash data center can respond instantly to user requests for information, and it requires less power and less physical floor space. The problem is moving to an all-flash storage infrastructure is expensive. However, it can be an affordable reality if the data center were only responsible for its active data. In reality, most of an organization's data is not active and should not be on flash.

Of course deleting the non-flash data set is not an option. This data has to be stored, managed and maintained, immediately eliminating those reductions in power and floor space requirements. Moving non-flash data to the cloud brings those reductions back into play, but it also introduces its challenges.

What is Non-Flash Data Center?

Data not accessed in the last 90 days is called non-flash data. 90% or more of the data that organizations store falls into this category. In fact, most data is never modified again a few weeks after its creation and is rarely accessed. With rare exceptions, flash is not the place to store this data. But just because this data has been inactive for more than 90 days does not mean that the data will never be accessed. In fact, some percentage of it will be accessed, at some point in the future, and when it is users will want to access it seamlessly and quickly.

The unpredictable need to access some portion of non-flash data is why the typical practice for most IT professionals is to continue to expand primary storage. This expansion is done by adding more shelves to storage systems or adding more storage systems into the data center. While it looks like the path of least resistance, the continued expansion of primary storage increases capital and operating costs significantly so that the "just in case" recovery demand is met.

A Whiteboard Full of Failures

In reality, inactive data has been a problem for IT professionals almost as long as there have been data centers. The rapid increase in the capacity requirements plus the desire to move active data to all-flash storage has only exacerbated the issue. In the past, technologies like hierarchal storage management (HSM), information lifecycle management (ILM), archive and file virtualization have tried to solve the inactive data problem.

The legacy solutions to the inactive data problem faced four key challenges. The first challenge was latency; the time it takes to return data to the requesting user or application. In most cases the response time or time to data was too slow, causing users to complain and applications to hang. The latency of secondary storage to respond to user requests lead to passive archiving, data had to be inactive for years before being removed from primary storage or, more typically, it lead to the abandonment of the project.

The second challenge was creating a seamless link between the original file location and the secondary file location. These systems used stub files or symbolic files to create the relationship between the two storage areas. Those stub files caused a variety of problems. Users didn't know what the files were, so they assumed they could delete them. Backup applications didn't know what to do with them, so they copied them into the backup storage, which meant the tracking of stub files in addition to the original file.

A third challenge was with the tertiary storage system that held the inactive data. It was still on-site and, as a result, had to be powered, cooled and managed, which lowered the potential savings that an archive system should deliver. It may have been less expensive per GB than primary storage, but it was not operationally less expensive.

Finally, these systems had to inspect, continuously, every file stored on primary storage. These solutions typically did this by continuously crawling through the file system, file by file, looking for files that met a certain criterion, typically last accessed date.

The Cloud Falls Short

Cloud storage seemed like an obvious solution. It does eliminate the need for the secondary storage systems to be on-site. The problem is most of the cloud archiving solutions do not address the other issues; they still had to inspect every file, repeatedly, to see if it qualified to be removed, and it had to create a seamless link to the files new location.

Latency is the big concern with a cloud-based archive solution. There is the obvious problem of the speed of the connection between the primary data center and the cloud, but bandwidth is increasing. A more pressing problem is the lack of optimization of that connection. Most vendors count on technologies like deduplication and compression as their optimization technique. In most cases non-flash data does not deduplicate well as it is storing the sole copy of that file.

Empowering the Cloud for Non-Flash Data

The key to empowering the cloud for non-flash data is to optimize it for this use case. Most data management solutions treat the cloud as a big disk drive; they fail to address issues with legacy solutions and they fail to address the new challenges that cloud presents.

A Network Controller

As stated above there are three problems that IT professionals will need to resolve if they decide they want to store their non-flash data in the cloud. Identifying the data to remove from primary storage is the first problem to overcome. Fundamentally, these solutions are implemented at the wrong point to identify, effectively, qualified data. They run at the backend of the infrastructure and have to “crawl” their way up the infrastructure to operate. An alternative solution is to implement the solution further upstream in the network as a data controller. Then this “network controller” can perform deep packet inspection of the files as they traverse the network, instead of waiting for files to be periodically crawled over for classification.

To be effective, the controller will need to build a table by walking the file system one time to build a baseline of the current data set. Then using deep inspection it can identify new files and files that are active. If the controller has not “seen” the file after an administrator-defined period, the controller can migrate that data to the cloud.

Deep packet inspection also resolves the challenges with trying to establish a seamless link to files. The appliance uses the initial file system crawl plus on-going deep packet inspection to maintain a rich meta-table of file information and attributes. One of those attributes can be file location. The result is that when a user requests a file they are routed to that file via the meta-data table. The rich meta-table eliminates the need for stub files and the management of those files.

Building a rich meta-table also improves the user experience. If the appliance keeps the meta-table locally and especially if the meta-table is flash-based, it has the information it needs to respond rapidly to a request for data with the current meta-table information and then stream the rest of the file's data from the cloud. Managing a local meta-table makes the user and application think that they are getting an instant response.

Finally, the controller should optimize the WAN/Internet transfer so that both sending and receiving can perform more efficiently. Again, the local meta-table helps with some of this but the appliance should also provide true WAN optimization for the use case. Additionally, deduplication won't typically have the same pay-off in this use case that it will in a backup. The files within the non-flash dataset are relatively unique. WAN efficiency is improved with compression and more typical WAN optimization techniques like IP packet optimization.

The Great Optimizer

Performing data analysis and data movement further upstream allows the controller to play a more powerful role in the enterprise than just moving data to the cloud. In can become a consolidation point of various NAS solutions in the data center as well as the potential secondary storage targets. The secondary storage could be multiple public cloud storage providers, a private cloud storage solution, and potentially an on-site disk archive.

Conclusion

Removing the 85%+ of data that is not actively being accessed not only makes the all-flash data center a reality it also reduces complexity throughout the data center. As a result, there is less data to backup, less data to replicate to a disaster recovery site and less data to expose to potential security breaches. The challenge has been how to identify this data, what to do with it after it has been identified and how to make sure it is easily retrievable after relocation. The cloud provides an excellent storage location, answering the "what to do with it" question. It needs a network controller type of solution to provide active identification and seamless connectivity to the data regardless of its location.

About George Crump

Eight years ago George Crump, founded Storage Switzerland with one simple goal. To educate IT professionals about all aspects of data center storage. He is the primary contributor to Storage Switzerland and is and a heavily sought after public speaker. With 25 years of experience designing storage solutions for data centers across the US, he has seen the birth of such technologies as RAID, NAS and SAN, Virtualization, Cloud and Enterprise Flash. Prior to founding Storage Switzerland he was CTO at one the nation's largest storage integrators where he was in charge of technology testing, integration and product selection.