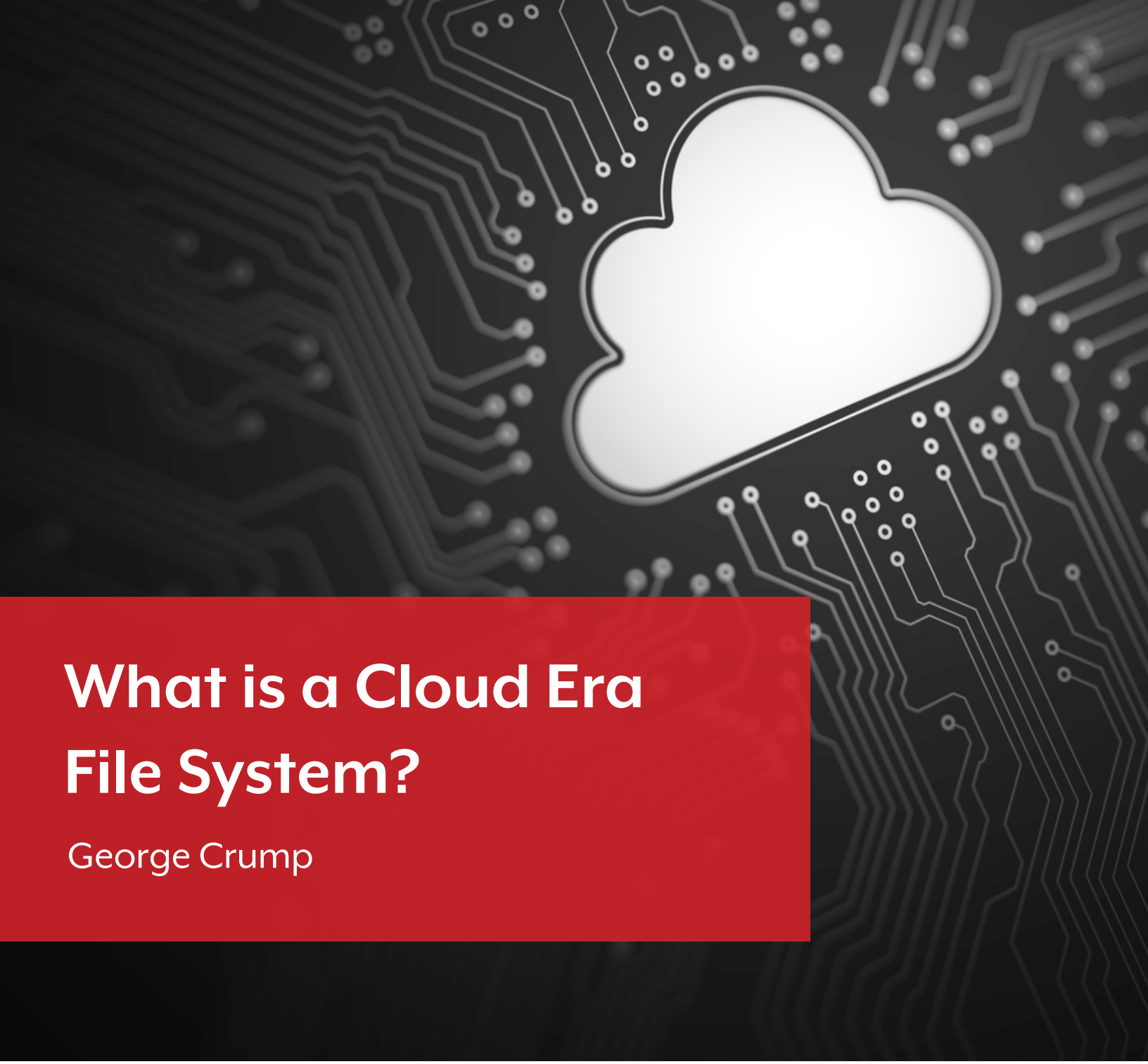

What is a Cloud Era File System?

George Crump

File systems originally started as a way to organize files. They've since evolved into the primary means of storing, sharing, and protecting all types of unstructured data. Now, with the rapid adoption of the cloud, the increase in both the number and size of files, and the need to support geographical distribution of data to global organizations and mobile users, the file system needs to evolve once again. Legacy file systems are simply not prepared for cloud storage, nor do they meet the scalability and mobility demands of the modern enterprise.

A QUICK HISTORY OF FILE SYSTEMS

File systems initially were used as a method to store files in a way that was more “human- friendly.” They were really a metaphor for their pre-technology counterparts - file cabinets, drawers and folders. Over time, though, file systems evolved to store much more than user data. Vendors enhanced them so their performance was similar to block-based storage, but still had the familiarity and ease of use that made them so popular. It was not long before file systems were used to store business-critical unstructured data.

Over time, as the capacity and number of files managed by file systems became greater, new capabilities spawned. File systems developed the ability to cluster many servers together, each contributing storage capacity and distributing the compute load so performance and feature sets could continue to improve, while meeting the never-ending appetite for capacity.

But with initiatives like Internet of Things (IoT) and Big Data Analytics becoming mainstream and organizations and users are becoming more distributed and mobile, file systems are starting to break down, or at least hit a wall. While scale-out systems can often meet the capacity demands of the data center, they typically cannot scale easily beyond the data center.

Scale-out systems also have the problem that performance and capacity scaling are inter-locked. Most organization don’t scale performance and capacity at the same pace and, in most cases, these resources either go underutilized, or worse, they are available only to the wrong group at the wrong time.

Finally, enterprises are moving to the cloud at an accelerated pace, and they want to leverage it as best they can. Most legacy file systems, if they support the cloud at all, can only use the cloud as a backup. They can’t use it as primary storage or even as an active archive for older data.

WHAT IS A CLOUD-ERA FILE SYSTEM?

A cloud-era file system is a file system that runs on top of cloud storage. But it does more than just store the files themselves in the cloud, a cloud-era file system also stores all its metadata in the cloud. Directory structures, access control lists (ACLs), really every part of the file system "lives" in the cloud, so it will benefit from the near limitless scalability, durability, and geo-redundancy of the cloud itself.

A cloud-era file system also leverages the cloud as the main hub for file transfer. By using the cloud as a highly economical backbone for moving files around the enterprise, more intelligent caching of files on-premises can be achieved, greatly minimizing the need for on-premises unstructured data storage arrays (e.g. NAS).

The combination of economical cloud storage and affordable cloud bandwidth allows the cloud-era file system to displace file backup software, replication technology, and WAN acceleration, significantly lowering data center costs and simplifying data management.

REQUIREMENTS OF A CLOUD-ERA FILE SYSTEM

There are several vendors claiming to have some form of a cloud-era file-system. Many of these vendors are actually just moving legacy file systems and instantiating them on cloud compute and storage. There are specific capabilities required for a file system to truly be "cloud era".

LOW LATENCY, HIGH PERFORMANCE ACCESS

The first requirement of a cloud-era file system is overcoming the key shortcoming of the cloud – latency. If data were only in the cloud and always being accessed remotely, users would revolt. The solution is for the file system to use the cloud as the authoritative source of all files, but have the ability to extend into edge locations and cache the most active data locally.

Some may be concerned about the impact of a cache miss. Without question, if a file has to be recalled from the cloud because it is not locally available, then access to the file will be slower than if it were on-premises. But the reality is, for most enterprises, the overwhelming majority of data has not been accessed in years. Even a cache sized to 15% of total capacity will almost never experience a miss. For most organizations there is a potential 85% reduction in the on-premises NAS. If the organization takes that a step further and accounts for hours many replicated arrays it has in branch offices and DR sites the potential savings is even larger.

If cache miss is really a concern, with all that cost savings, IT can size the local caches big enough to prevent cache misses, and still be saving a lot of its budget.

INDEPENDANT SCALING OF CAPACITY & PERFORMANCE

The cloud/edge concept also serves up the next requirement - the independent scaling of performance and capacity.

File IO performance is quite obviously needed when data is active. In the cloud/edge architecture, active data is now, by definition, at the edge. The physical or virtual edge appliance can be designed to deliver all-flash array class performance and enterprise-class reliability, but still be far less expensive than a full-blown NAS system, since it is a fraction of the size, as explained above.

With the cloud/edge hybrid architecture, performance can also be adjusted per location or department. A location needing to perform analytics on an active data set may require flash-class performance. Another location needing occasional user access to the same data set could use a less powerful (and less expensive) edge appliance, maybe one with hard disk drives. Both edge appliances are serving up the same volumes and capacity, but with different levels of performance.

The cloud/edge architecture also allows for independent capacity scaling. Capacity is now scaled without limit in the cloud; the cloud-era file system will elastically scale with the increase in the organization's cloud subscription. Scaling capacity in the cloud also means the organization only has to buy capacity as it is needed, sometimes referred to as "pay-as-you-grow." New capacity purchases can also be very granular - TBs at a time instead of entire storage systems or nodes within a cluster.

ACTIVE ARCHIVE

While today's unstructured data is growing at an incredible rate, it follows the same lifecycle as legacy unstructured data - it is very active for a few weeks after initial creation, then goes dormant for an extended period of time. Unlike legacy unstructured data, however, modern unstructured data tends to be reused, as enterprises seek to extract more value from their data through analytics. As a result users are less tolerant of delays in accessing this data, even though it is older.

Again, the cloud/edge architecture of the cloud era file system automatically solves this problem. Its ability to automatically recall inactive data and cache it on the local appliance when it becomes active is exactly what's needed for active archiving use cases. Once a data set becomes active, it automatically gets the accelerated access performance of being stored on the local edge appliance.

GLOBAL ACCESS & COLLABORATION

The next requirement is global access to data. Businesses today are much more distributed than they used to be. Traditionally, new offices were added as a result of acquisitions or mergers. Being able to quickly on-board these new offices by enabling them to share the same files is key to how quickly an acquired business will be accretive, and contribute to the bottom line.

Labor arbitrage is an even bigger reason why new offices are being opened in different locations. Many companies are now expanding in areas where talented potential employees are based, but where the cost of living – and therefore wages – are lower. These offices often need to work on the same files as the main office, and must be seamlessly integrated into design or content creation workflows.

There is also, of course, the road warrior who needs access to files. Files must be readily available to this mobile user.

Again, the cloud-era file system provides a strong foundation to meet these global access and collaboration requirements. The file system already distributes data from the cloud to the edge using cloud "pipes" for economical bandwidth. Doing this in many locations requires no additional effort.

GLOBAL ACCESS NEED FILE CONTROL

The new distributed office capability creates the need for two additional requirements. First, with the same file existing in multiple locations, a way of storing differnet versions created by different users is needed. The cloud-era file system needs a volume manager that can align the changes based on timestamps, store the gold copy of every version in the cloud, and enable users to retrieve any version at any time.

Second, the file system should offer a global file locking capability that enables the same file to be safely distributed to multiple locations without the concern of overwrite. A global file lock also allows for a distributed workflow that can eliminate the need for onerous check-in/check-out procedures. One office can start working on a file, then another can pick up that work. If the second office accesses the file before the first office is done, they will get a warning similar to the warning they would get if they are accessing the file from the same local NAS.

DATA PROTECTION (BACKUP AND DR)

The final requirement of the cloud-era file system is to provide more comprehensive data protection that is currently available with traditional file systems or even dedicated data protection solutions.

First is to enable IT to meet stricter recovery point and time objectives (RPO/RTOs). The cloud era file system should be able to take near-continuous snapshots of data as it changes. By sending just the deltas to the cloud and leveraging the favorable economics of cloud storage to store every change to every file, IT and end users can restore al most any version of any file in minutes.

These snapshots should also be immutable so if the environment is infected by Ransomware or malware, select files, entire volumes, or the entire file system can be rolled back to their state prior to the attack. With this capability, the cost and administrative overhead of file backup software and hardware targets can almost be eliminated.

The cloud-era file system, by its very nature, also offers built-in disaster recovery. Since the primary copy of all data is stored in the cloud, it is automatically off-site. Furthermore, every cloud provider automatically stores one or more copies of all data in other regions within their infrastructure so it is protected from any type of disaster in one region. Combine this with the ability to rapidly deploy edge appliances in any location that can connect to the cloud, and the organization benefits from near-instant DR anywhere, without the high cost of dedicated DR sites and duplicate file infrastructure.

CONCLUSION

Most enterprises are already in the cloud era in some way. They may be sending backups to the cloud or using a cloud-based application, but they are likely using the cloud. Leveraging the cloud for unstructured data is the logical next step.

But to take full advantage of cloud economics and the many emerging cloud object storage options, a new take on file systems is required. The cloud-era file system needs to look and act like a legacy file system, but provide users and IT with next-generation capabilities to meet the global file sharing, collaboration, and data protection requirements of the modern, cloud-era enterprise.

Storage Switzerland, LLC

NASUNI®

THE FIRM

Storage Switzerland is the leading storage analyst firm focused on the emerging storage cataegories of memory-based storage (Flash), Big Data, virtualization, and cloud computing. The firm is widely recognized for its blogs, white papers and videos on current appraoches such as all-flash arrays, deduplication, SSD's, software-defined storage, backup appliances and storage networking. The name "Storage Switzerland" indicates a pledge to provide neutral analysis of the storage marketplace, rather than focusing on a single vendor approach.

ABOUT OUR PARTNER

Nasuni ("NAS Unified") is the leading provider of cloud-scale enterprise file services. Powered by UniFS®, the first cloud-native file system, Nasuni leverages private and public cloud object storage to integrate and surpass the "silo" capabilities of traditional Network Attached Storage (NAS), distributed file systems, data protection software, file synchronization solutions, and disaster recovery infrastructure.

THE ANALYST

George Crump is the founder of Storage Switzerland, the leading storage analyst focused on the subjects of big data, solid state storage, virtualization, cloud computing and data protection. He is widely recognized for his articles, white papers, and videos on such current approaches as all-flash arrays, deduplication, SSDs, software-defined storage, backup appliances, and storage networking. He has over 25 years of experience designing storage solutions for data centers across the U.S.