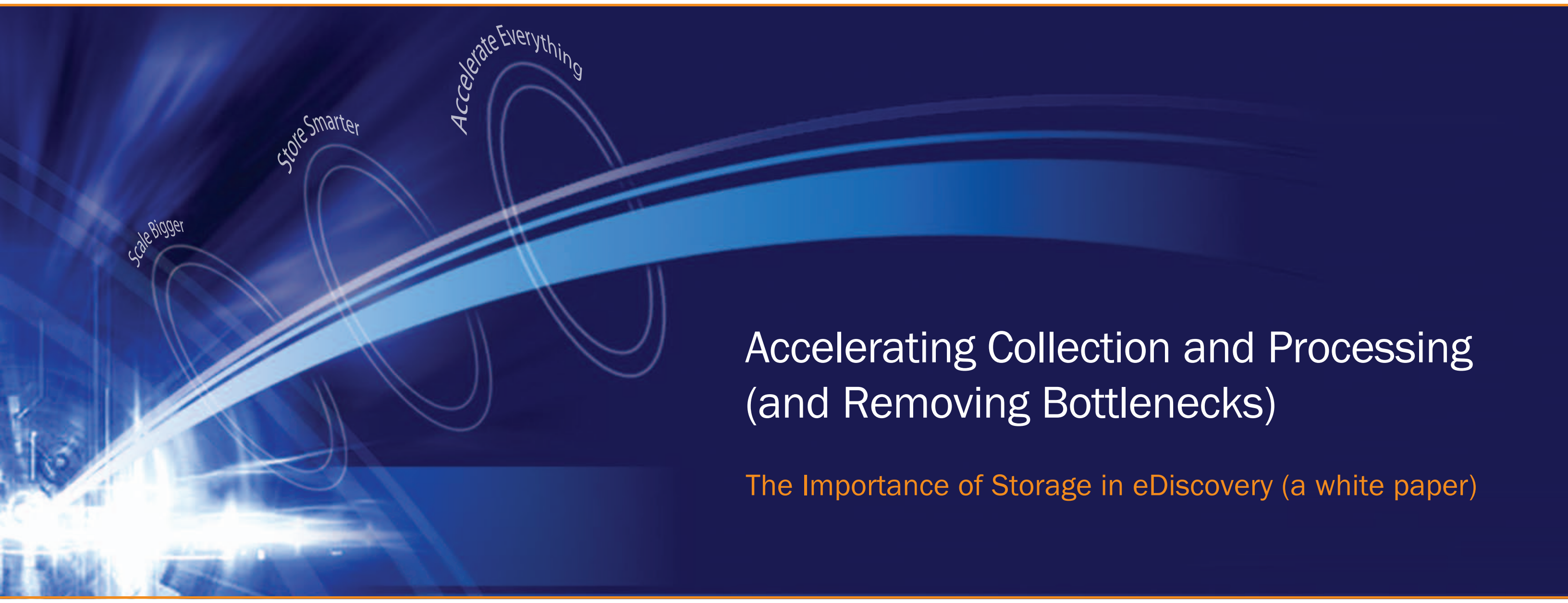
WHITE PAPER



Accelerating Collection and Processing (and Removing Bottlenecks)

The Importance of Storage in eDiscovery (a white paper)

Abstract

Corporations and government agencies are sitting on a time bomb and most don't know it. There is an increasing burden to collect and preserve information either in response to litigation or requests under regulations such as the Freedom of Information Act (FOIA). While legal hold and regulatory investigations were once straightforward, the explosion of digital information has created challenges never before seen. In many ways, there is a perfect storm forming that — if organizations don't prepare for — will leave behind devastation.

By Jeff Greenwald, senior solutions marketing manager
eDiscovery Markets
BlueArc Corporation
jgreenwald@bluearc.com

Organizations Feel the Bottom Line Impact of eDiscovery

Organizations Feel the Bottom Line Impact of eDiscovery

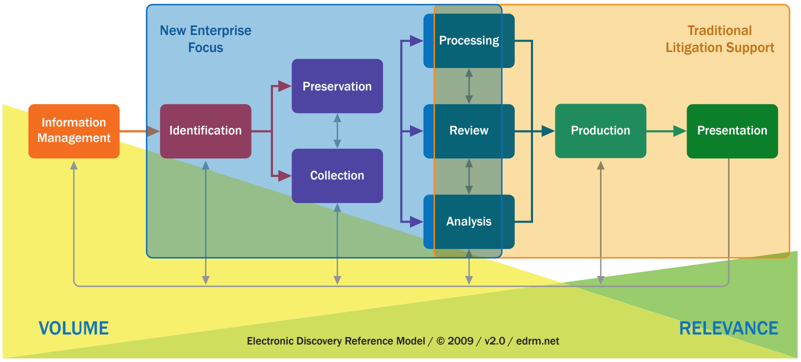
Corporations and government agencies are sitting on a time bomb and most don't know it. There is an increasing burden to collect and preserve information either in response to litigation or requests under regulations such as the Freedom of Information Act (FOIA). While legal hold and regulatory investigations were once straightforward, the explosion of digital information has created challenges never before seen. In many ways, there is a perfect storm forming that – if organizations don't prepare for – will leave behind devastation.

The rules governing electronic discovery (eDiscovery) are more stringent than ever. The Federal Rules of Civil Procedure (FRCP) place a broad and vague burden on companies to be able to quickly respond to discovery requests in civil litigation. The FRCPs call for companies to reasonable and good faith steps to preserve information and produce it to opposing parties. As a result, corporate lawyers will be looking for the widest possible retention policies and practices. They do not realize that overly broad retention efforts may have a negative impact on the daily operations within IT. The lawyers typically don't care if that does happen. They have learned over time that if they allow the IT people free reign, the data will be deleted and the case potentially compromised. Consequently, the response to every matter will be to "preserve everything."

But, preserving everything simply transfers an already difficult problem over to IT. The short timeframes for responding to discovery requests under the FRCPs has put IT teams in a nightmare cycle of reactive firefights – constantly scrambling to collect information from various sources. The incredible volume of digital information that now exists only exacerbates the problem. IDC estimates that the universe of digital information will grow to nearly 1.8 zettabytes (1,800 exabytes) this year. And as the universe of information grows exponentially, conducting eDiscovery only gets more challenging. For example, the federal courts are holding organizations to higher metadata management standards then ever before. In the case of National Day Laborer Organizing Network v. United States Immigration & Customs Enforcement Agency, 2011 U.S. Dist. LEXIS 11655 (S.D.N.Y. Feb. 7, 2011) the presiding judge held "that certain metadata is an integral or intrinsic part of an electronic record [and as] a result, such metadata is 'readily reproducible' in the FOIA [Freedom of Information Act] context."

The grim reality for organizations is that eDiscovery is expensive. While EDD processing (preparation of diverse data into a common, searchable set of information) costs have come down over the past five years, the fact is that the reduction in price has not kept up with the increase in the volume of information. With the "preserve everything" mentality, organizations have too much information to effectively minimize the amount of potentially responsive data without paying upwards of $500 per GB to a third-party processing vendor. As such, even a small case can cost thousands of dollars just for data processing alone. Add first-pass attorney review at $150 per hour and it's easy to see why companies want to reduce the cost of eDiscovery.

Electronic Discovery Reference Model



Electronic Discovery Reference Model / © 2009 / v2.0 / edrm.net

Figure 1. New Trends in eDiscovery

- Early review of potentially relevant records
- Some implementations inside the firewall — "ECA in Place"
- Potentially substantial savings in time and money for client and law firm

In order to truly decrease the cost of eDiscovery, organizations must be able to collect and process information in-house. It's simply too expensive to rely on a third-party EDD processing house to churn through huge amounts of information, a large percentage of which will most likely not even be relevant to the matter at hand. There are many components to the eDiscovery process, as seen in the Figure below – The Electronic Discovery Reference Model (EDRM). More and more organizations are proactively addressing the left side of the model, specifically the collection, preservation, processing, and analysis nodes.

In the past few years, organizations have deployed collection and processing applications – sometimes referred to as Early Case Assessment (ECA) tools – to lower the cost of eDiscovery and mitigate the risks of managing legal holds. But, software applications alone cannot solve the problem; the right storage environment to meet the specific demands of eDiscovery is a necessary foundational element.

Storage Is Part of the eDiscovery Foundation

Storage infrastructure is one of the pillars of eDiscovery. Specific applications such as identification, collection, and preservation software, ECA, and review management applications are important and offer feature and functions to make eDiscovery more efficient and defensible. Such applications, though, depend on the right storage for peak performance. And, while applications are typically mixed and matched as needed, storage infrastructure cannot be swapped out without considerable effort and expense. Such effort and expense can be a huge roadblock to eDiscovery success.

Thus, it is extremely important to choose the right storage foundation to support your eDiscovery activities and applications. Some applications are very demanding of storage infrastructure; if the storage foundation is inadequate, the whole eDiscovery process can be hobbled. It's easy to think of storage as simply the media where digital information sits. But, storage is so much more than that – it's the power behind an optimized eDiscovery infrastructure.

The performance and scalability of eDiscovery applications is highly dependent on the underlying storage platform. Organizations store huge amounts of information. When a discovery request comes in, IT must collect a subset of that data. The subset itself is often hundreds of gigabyes. Processing that much data requires real computing horsepower. To optimally address the challenges of eDiscovery, organizations must have:

- The ability to handle and move large and ever increasing amounts of data, frequently in the range of terabytes and petabytes
- Extensive storage headroom to handle unpredictable and growing flows of large amounts of new data
- The ability to provide high performance for mixed workloads that may vary widely and quickly, between shorts bursts of reads and writes, long sequential reads, and heavy CPU-generated input/output
- Network systems robust enough to handle the size and speed of data movements and not buckle under peak loads
- Very fast connection to computational capability to crunch through and analyze multiple large datasets running in parallel
- A flexible IT environment that is not overly tuned for a specific workload but can grow and adapt to changing workload patterns, including:
  - Ability to deliver high throughput and high input/output per second (IOPS)
  - Powerful and flexible data management capabilities to cost efficiently manage the information life cycle
  - Multiple file and networking protocols
  - Mix of several device types and technologies – rapid growth implies that new storage systems live alongside older ones in a heterogeneous environment
  - Multiple users and applications accessing the same data sets simultaneously

- Growing need to manage the aging of data, especially for matters that spread over many years
- Optimized use of storage subsystems for data with differing needs of performance, cost, high availability and data retention

If the storage system can't keep up with the demands imposed on it by a high-volume processing, both the speed and the quality of the final output will suffer. Data throughput performance will be stressed, for example, in enterprise-scale operations like eDiscovery that utilize tens or hundreds of clients hosted on modern multi-threaded, multi-core CPUs. Whereas most systems require a relatively steady state of storage and computing power, an eDiscovery infrastructure must support fast, unexpected bursts of computing needs to churn through massive collections of information. With eDiscovery, there is no way to forecast when these resources will be necessary. Instead, it is prudent to be sure the eDiscovery infrastructure is built for high-performance from the ground up.

eDiscovery involves putting collected sets of data onto file systems for downstream activities, e.g. processing and ECA. The file system size is a logical limit connected to the number of objects it can effectively support. This number is often arbitrary and in some cases may only be a best guess estimate. In other words, a storage system can have a stated 2 TB or 16 TB file system limit but depending on the number of objects and the stress put on that file system, it could be far less.

Storage systems do much more than deal with primary I/O operations. There are a number of background tasks that are fairly common within most, if not all, storage systems. For example, a common occurrence within storage systems is RAID rebuilds. Disk drives do three things – they read and write and they break. Disk drives are mechanical devices and as such can suffer physical failures. Once this occurs then RAID rebuilds are executed to ensure data integrity. However, RAID rebuilds consume storage system resources that often contend with primary I/O and can have an impact on performance and scalability. Naturally, the more disk drives you have the greater chance for disk failure creating a cycle between performance and scalability.

The number of disk drives plays an important role in both performance and scalability. Often customers will add more disk drives in order to improve performance by striping data across a large number of spindles. However, this leads to under utilization of capacity drives up costs. On the other hand, when capacity is the priority requirement many storage systems have physical and logical capacity limitations, which require customers to buy multiple storage systems, again, driving up costs.

The number of controllers also plays a role in both scalability and performance. There are a number of scale-out clustered architectures that utilize more storage controller nodes to scale. This requires more CPUs, memory, physical space, power and cooling.

Too often, organizations make the shortsighted assumption that eDiscovery applications like ECA can be deployed on existing storage infrastructure. However, file access and directory lookups often degrade as more files are added to a directory. While this may not be a problem when there are several thousands of files in a directory, with the millions of files involved with eDiscovery, this can become a crippling impediment. It is better – and more cost-effective – to look at storage platforms with management software specifically designed for high-volume scalability, high throughput, and massive processing capacity for millions of small and large files.

Collections of electronically stored information (ESI) often include a number of large (~300GB or greater) source files, such as logical evidence files (LEF) accumulated from disk imaging. Converting these files to smaller target files requires a storage platform with capacity for high IOPS as well as the capacity to perform many multiple concurrent read/write operations. Processing the millions of small files that are extracted from containers like LEFs creates a storage IOPS challenge since the ratio of the file protocol handling to the size of the data is so significant.

When application vendors speak about high processing performance numbers, a major part of the equation depends upon the storage used in the testing. It's important to realize that you might not get the same performance if you just use existing storage that you have in your data center. Scalability is also highly dependent on storage. Organizations that conduct collection, preservation, analysis, and review on a case-by-case basis and tend to have very small matters might be able to get away with using an application built for limited scale. However, for the enterprises, law firms, and service providers conducting large-scale eDiscovery operations, scalability is a must. With eDiscovery, storage requirements for a given matter are not known until the identification and collection phases are complete; at that point, there is no time for a typical information technology infrastructure purchase cycle. The right infrastructure – with the ability to scale and process data at the right speeds – is a critical upfront investment if there is going to be a chance to optimize both the effectiveness and efficiency of eDiscovery.

Chain-of-Custody Management

One of the key components of ESI processing involves managing the chain of custody of the data. If files or metadata are changed in any way, the collection itself could be challenged or called into question. In addition to providing much of the necessary horsepower for processing digital information, storage software is critical in maintaining defensible collection and preservation. While eDiscovery applications run mechanisms such as hashing and provide reports to prove that ESI has not been altered during eDiscovery operations, it is the storage software that must support combining collection of data sets from different platforms and support the migration of data from those platforms to a specific matter set. For example, some storage devices have file allocation tables that are FAT 16 whereas others are FAT 32. The right storage infrastructure must support collecting data when there are potential differences in file system fields/properties from the various sources of data in a way that allows you to show the defensibility of the collection, preservation, and management of the data. eDiscovery applications are ultimately responsible for managing chain-of-custody, but it is important to be aware of the differences in the storage platforms of various ESI sources and use the collection software to track operating system (OS) fields before they are altered. What is important is that the actual storage be homogeneous, so that there are no secondary/tertiary changes after the initial collection. Storage administrators need to be aware of potential issues and provide expertise to support the legal team in proving that data is not altered.

Not Just Any Storage

To process information at the speed and volume that eDiscovery demands requires massive scalability, high throughput, and processing capacity. This is where storage systems become critically important. While any storage system can be configured to provide speed and throughput, enabling processing capacity requires the next-generation NAS. It means being able to support high input/output per second (IOPS) on file systems.

BlueArc is the network storage of choice for organizations that value processing time as money because it allows thousands of parallel operations in the processing of data and metadata, maintaining performance even with the heavy demands of eDiscovery. BlueArc's hardware-accelerated architecture, based on an object-based file system, can process hundreds of millions of files across deep directory structures for a quantum improvement in eDiscovery application performance. Traditional software based file systems need to create extraneous directories or folders to effectively handle this many files, while BlueArc network storage can easily scale and maintain perforce to 16 million files in a single directory.

The BlueArc NAS storage system has a unique architectural advantage that off-loads file system operations utilizing internal high-performance silicon. This distinct design element enables BlueArc NAS to scale in performance and capacity simply, cost-effectively, and with minimal physical footprint. Both performance and capacity are inextricably linked and BlueArc is designed to efficiently scale without degrading as the system grows. This is critically important for ensuring that eDiscovery activities are never halted due to the stress put upon the storage system.

Other architectures risk metadata corruption (and therefore spoliation, and the potential sanctions that result) because when the file protocol processing cannot keep up with the file flow, it leads to potential dropped files and file collisions on the bus – that can look like bad metadata. And, as the National Day Laborer case points out, metadata is a critical component of electronic records.

BlueArc Mercury 55 with two shelves of disks, although capacity can grow to 8 PB
(Prices with 10 TB can start as low as $42,000)

Evaluating Storage For E-discovery

The storage used in eDiscovery operations can be the difference between moderate efficiency gains and drastic cost reduction and risk mitigation. Many organizations have existing storage systems in place and assume those systems can handle the load that eDiscovery operations will demand. While in some cases that may be true, it is a better practice to have a dedicated storage platform to support eDiscovery. Not every organization will need the same eDiscovery storage platform, but there are some best practices that all should consider:

- **Understand the amount of eDiscovery operations that will take place in your data center.** Storage requirements depend upon how much of eDiscovery will be managed in your data center. For corporations, service providers, and law firms that will be processing large amounts of data, storage is a critical element for success. When it comes to storage, there are many options. Some will opt for storage area networks (SAN) which can require more resources to manage, while others will want to use network attached storage (NAS), which can be administered and managed with considerably less effort. Don't just think about near-term eDiscovery; rather forecast the long-range activity and capacity needs in order to determine performance requirements.
- **Ensure centralized, historical tracking of ownership and access rights for shares.** Data will be sitting on multiple file shares – trying to manage the ownership and access rights for each share is hard to do in a distributed manner. It's important to have that centralized management capability and centralized reporting of any changes to the shares.
- **Make sure any information on legal hold can remain on legal hold with full metadata and context intact on the new storage.** One of the worst things that can happen in eDiscovery is spoliation, where data is altered somewhere in the process of collection, preservation, review, or production. It is important to recognize the differences between storage platforms of collected data sets and ensure that your eDiscovery storage platform be able to accommodate all the diverse collections.
- **Plan for end-of-life, both of data and storage.** How does data get defensibly disposed of? Does legal have a sign off? How does data get migrated from EOL storage? How do you actually destroy drives so that the data is not recoverable or sold to a recycling company with confidential information?

About BlueArc

BlueArc develops and sells clustered NAS systems for storing and managing digital content and unstructured (file-based) data. The company's products use a scalable file system that allows for multiple storage appliances (nodes) to be managed as a single pool. As customer capacity requirements grow, new storage nodes can be added without disrupting operations and legacy storage systems can be incorporated into the BlueArc namespace for better utilization of assets. The EDRM application areas that include identification, processing and collection specifically benefit from the BlueArc architecture, and many industry benchmarks demonstrate the value add of accelerating eDiscovery in the corporation, at the law firm, and at the service provider.

BlueArc's products are important for high-end file repositories in eDiscovery. BlueArc has been increasingly positioning its branded products toward mainstream enterprise virtualization environments.

BlueArc is a 12 year old company with its engineering in San Jose, California and Bracknell UK, and BlueArc markets its products internationally, and sells through a variety of OEM and Reseller partnerships as well as directly from BlueArc.

We enable companies to expand the ways they explore, discover, research, create, process and innovate in data-intensive environments. Our products replace complex and performance-limited products with high performance, scalable and easy to use systems capable of handling the most data intensive applications and environments. Further, we believe that our energy efficient design and our products' ability to consolidate legacy storage infrastructures, dramatically increases storage utilization rates and reduces our customers' total cost of ownership.

BLUEARC®

BlueArc Corporation
Corporate Headquarters
50 Rio Robles Drive
San Jose, CA 95134
t 408 576 6600
f 408 576 6601
www.bluearc.com

BlueArc UK Ltd.
European Headquarters
Queensgate House
Cookham Road
Bracknell RG12 1RB, United Kingdom
t +44 (0) 1344 408 200
f +44 (0) 1344 408 202