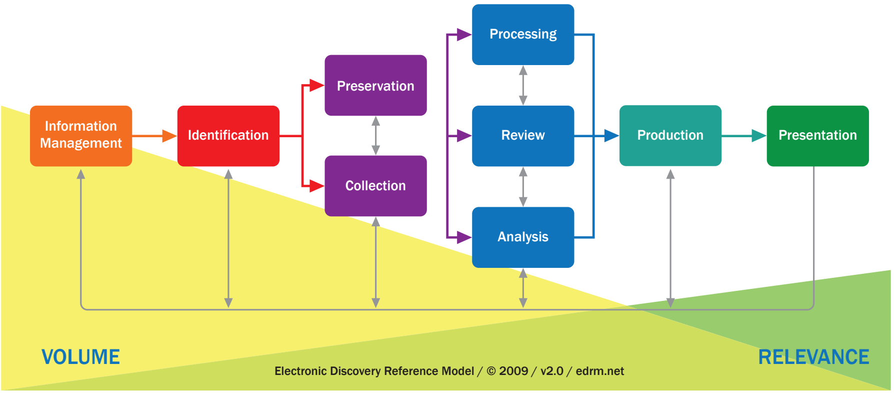
BLUEARC SOLUTION HIGHLIGHTS:

- Massive computing parallelism delivers the performance to support multiple, varied applications
- Scales up to 16 PB usable capacity, a 256 TB file system, and thousands of concurrent users
- Clusters up to eight nodes with Cluster Namespace to enable horizontal scalability
- Concurrent support for iSCSI, NFS and CIFS eliminates storage silos
- Dynamic, policy-based data migration and caching simplify management of infrequently accessed data
- Advanced Virtualization Framework delivers Thin Provisioning and Virtual Servers
- Integration with many legacy third-party network attached storage devices protects prior investment
- Support for multiple, best-in-class storage options for varied workloads

BlueArc—The Foundation For Cost-Effective eDiscovery

Electronic Discovery Reference Model



The Complexities of eDiscovery Lead To A Myriad Of Market Challenges

The rules governing eDiscovery and eDisclosure continue to grow more stringent, not only in the United States, but also in the United Kingdom, Germany, Australia, Canada, and other countries around the globe. Organizations must be able to identify, collect, process, analyze, and produce information requested in litigation or regulatory compliance in very short timeframes or face negative consequences such as sanctions, fines, or poor outcomes in legal matters. The ability to meet strict requirements has traditionally come at a large cost due to massive amount of electronically stored information (ESI) that exists. ESI volumes will continue to grow – IDC estimates that the universe of digital information will grow to nearly 1.8 zettabytes (1,800 exabytes) by 2011. That means costs will continue to rise unless organizations get proactive about managing information for eDiscovery and implement more efficient solutions.

Research firm Gartner estimates the average cost to defend a corporate lawsuit exceeds $1.5 million per case. Much of this cost comes from third party processing fees of up to $300-$1200 per GB and legal review costs of up to $500 per hour on information that is often completely irrelevant to a given matter. Historically, in order to meet necessary eDiscovery timeframes, organizations often simply took snapshots of all potentially responsive information and then sent the full collection out for processing and review. It was clear that much of this cost was unnecessary, but there was seemingly no way to avoid it. More recently leading organizations are improving ESI on a core infrastructure function allowing for more efficient identification and collection.

Organizations also need to mitigate risk within the eDiscovery process. More and more cases – including Pension Committee of the University of Montreal Pension Plan v. Banc of America Securities, LLC – point to the importance of legal hold and preservation efforts. In May, 2010, Piper Jaffray & Company was fined $700K for failure to preserve emails – this was in addition to a similar fine of $1.65 million in 2002. These costs are very real. What many don't understand, however, is that good preservation efforts need to be supported by the right storage platform. If the storage systems can't scale to hold a full collection, the data might be corrupted – leaving the organization subject to sanctions for spoliation despite the preservation efforts.

BLUEARC®

As a result, organizations now strive for litigation readiness in order to mitigate risk and control costs. Whether preserving all potentially responsive information for eDiscovery or deciding what to produce in eDisclosure, the back-end infrastructure must support the ability to quickly identify, collect, preserve, process, analyze, and review information. Organizations can benefit from early case assessment (ECA) to minimize the amount of information for downstream processing and review. ECA has been shown to reduce responsive data sets by over 80%, filtering out information that is clearly not responsive to a given matter. That could save both time and money per lawsuit. In the rush to reduce or avoid eDiscovery costs, some organizations have deployed tools without consideration to the underlying platform that powers them.

The total costs of defending a lawsuit includes both the law firms time to review the data, and also the time and cost to processing the data. This total cost has often exceeded $1.0M per case. Therefore filtering out the information can have a significant affect on lowering costs.

One of the most complex types of litigation is intellectual property (IP) litigation, and according to The Sixth Annual Fulbright & Jaworski Litigation Trends Survey from 2009, IP litigation is expected to rise by 15%. When involved in IP litigation, you must collect data from all around the world across many diverse sources. But collection is not enough – the lawyers need the data very quickly in order to analyze it and create a case strategy. But, if your storage system doesn't have the performance and throughput to allow you to store a proper index for collection, it can take too long to get all the data; the result is a negative case outcome in addition to high eDiscovery costs.

No matter what steps your organization takes to address eDiscovery – legal hold, collection, process, ECA – it is critical that the right storage infrastructure be in place. eDiscovery applications deployed on inferior storage can actually increase costs and introduce more risk into the process. It is imperative that eDiscovery applications be deployed on the right storage foundation in order to achieve the potential benefits eDiscovery applications promise. To optimally address the challenges of eDiscovery, organizations must have:

- The ability to address and move large and ever increasing amounts of data, frequently in the range of terabytes and petabytes
- Extensive storage headroom both in file system size and transfer bandwidth 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 I/O
- 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 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

eDiscovery Applications Without The Right Underlying Storage Platform Will Deliver Limited Benefits

There are a variety of eDiscovery applications on the market – collection and processing tools, ECA tools, legal hold management tools – that can deliver benefits in the form of cost reduction/avoidance and risk mitigation. For a variety of reasons – disconnect between legal and IT, lack of eDiscovery experience, inability to get purchasing budgets – organizations traditionally attacked eDiscovery on a matter-by-matter basis. This often meant buying eDiscovery appliances with processing, review, and analytics software bundled with hardware because appliance costs can be lower than enterprise software licenses. One way that appliance costs could stay low was to bundle high-end software with low-end storage. While this made the entry level cost more palatable, it regularly caused more problems than it solved. When information volumes got large, these appliances were unable to either scale to the volume needed or support the processing throughput required to meet tight timeframes. So, essentially organizations spent money to fail at solving the eDiscovery problem.

Traditional File Systems Are Not Sufficient to Address eDiscovery Demands

eDiscovery requires the ability to process large volumes of information in very short timeframes. Information collected and preserved is stored on multiple file systems inside the corporation. The information tends to be comprised of millions of both large and small files. Most file systems support adequate levels of throughput when dealing with large files, but thousands or millions of small files degrade performance due to the high percentage of non-data transfer operations. Indexing and processing information means doing metadata look-ups for every file and that creates a lot of overhead that traditional storage systems cannot keep up with. This creates complexity for IT – managing tasks across multiple systems in addition to taking care of backup and replication for all this information.

By its nature, eDiscovery requires large amounts of storage. There can be multiple collections in any given matter and the average Global 2000 company has an average of one hundred and forty-three concurrent lawsuits with the average mid to large company managing over twenty ongoing law suits at any given point in time. With each custodian regularly generating 4 Gb-10 Gb of data each, and an average law suit consisting of 10-50 custodians at a cost of $50K-$100K per custodian for review and processing, that can quickly add up to a lot of data and costs can mount quickly. In addition, that information is often held for years at a time. It's necessary to have storage that can keep up and expand without constant administration, which would drive up both IT management complexity and cost.

eDiscovery involves putting collected sets of data onto file systems for downstream activities, e.g.ECA, processing, review, and production. 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, the operational limit 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. When failure occurs RAID rebuilds are executed to maintain 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 forindividual 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 and 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. In other words, more money; both more capital expense, and also more operational expense.

Too often, organizations make the short-sighted 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.

The Right Storage Foundation Is Key In eDiscovery

There is a lot of buzz in the eDiscovery market about the savings possible with deployments of applications like early case assessment (ECA), collection and processing tools, and legal hold management. But what most don't understand is that, no matter what eDiscovery applications an organization deploys, the benefits can be eliminated if the solution is not built on the right storage platform. It isn't just growth that impacts performance but the various activities that are run on the storage system that also contend for the same resources including RAID rebuilds, snapshot processes, backup jobs and remote replication. When you consider all of these variables, it is essential that organizations seeking efficient eDiscovery look at performance holistically and not just at the surface level. If they don't, the consequences can be dire resulting in eDiscovery applications slowing to a crawl or stopping altogether, or even worse, failure due to lost files.

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 Provides the Right Platform to Support eDiscovery Applications

BlueArc provides the network storage solutions of choice for eDiscovery Service Providers who 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. As a result, more ESI can be processed per day.

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.

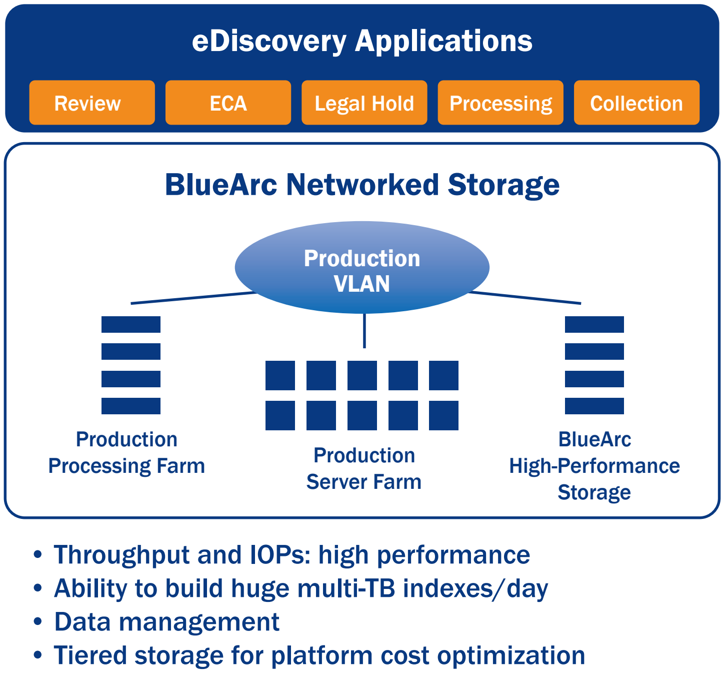

eDiscovery Applications
Review
ECA
Legal Hold
Processing
Collection
BlueArc Networked Storage
Production VLAN
Production Processing Farm
Production Server Farm
BlueArc High-Performance Storage
• Throughput and IOPs: high performance
• Ability to build huge multi-TB indexes/day
• Data management
• Tiered storage for platform cost optimization

Figure 1 – How BlueArc Powers eDiscovery

BlueArc networked storage is implemented under the hood of a typical file system, working within existing standards and traditional CIFS/NFS protocols. Within this traditional storage offering, BlueArc implements an object-based design utilizing an object store, with root and leaf onode hierarchies in a tree structure, with a high degree of parallelization and manipulation of object pointers to accomplish data management duties. By contrast, the data structure widely used in many UNIX or Linux-based filesystems – inodes – stores information about a file, directory, or other filesystem object. The inode is thus not the data itself, but rather the metadata that describes the data. inodes store such metadata as user and group ownership, access mode (i.e., file permissions), file size, timestamps, file pointers (usually links to this inode from other parts of the filesystem) and file type, for example. When a traditional filesystem is created there is a finite upper limit on the total number of inodes – this limit defines the maximum number of files, directories, or other objects the filesystem can hold. This limit leads to what is called the finite inode problem, and is why most traditional filesystems cannot scale easily to multiple petabytes or billions of files.

In object-based filesystems, objects are manipulated by the filesystem and correspond to blocks of raw data on the disks themselves. Information about these objects, or the object metadata, is called an onode, in much the same way inodes refer to file metadata in a traditional filesystem. In BlueArc's object store the underlying structure used to build up the filesystem is an "object", which is any organization of one or more of these raw blocks of data into a tree structure. The object is a container of storage that can be created, written to, read from, deleted, etc. Each element of the object is called an Onode, and while there are strong parallels to the normal use of the term onode in other object-based filesystems the concepts are not identical. In BlueArc's Object Store, objects are manipulated by logic residing in FPGAs located on the hardware modules. BlueArc achieves great acceleration of many filesystem operations though the use hardware acceleration using FPGA technology, and the design offers many benefits in the way of ehanced performance and scalability:

Scalability without impacting performance: BlueArc networked storage can support millions of files in a single directory, while keeping directory search times low and sustaining overall system performance. It can support many petabytes in a single unified namespace – presenting it all as a single filesystem accessible to many concurrent hosts, through a single mount point if desired.

Consolidation: Extreme scalability enables consolidation, particularly of older hardware and "storage islands." The ability to provide a unified, large-scale storage solution allows storage administrators to combine the functions of what were separately implemented file servers, reaping the cost-savings and ease-of-management benefits of a consolidated platform.

Meaningful virtualization: virtualization is about making more efficient use of a single server. The more powerful the individual server is, the better suited it is to virtualize a larger number of less capable, under-utilized devices. BlueArc's implementation of virtual servers allows groups to retain "ownership" of their virtual entity within a single physical server. And thin provisioning makes it possible for multiple virtual servers to share a single pool of storage devices.

Widest applicability to changing workloads, data sets, and access patterns: Fine-grained parallelism, off-loading of specific filesystem operations to FPGAs, and data pipelining all contribute to BlueArc networked storage' optimized handling of both throughput and metadata processing. Both attributes have been principal design criteria from the beginning with BlueArc networked storage.

Flexible performance scalability due to separation of function between servers and storage: BlueArc networked storage delivers great performance with relatively small storage systems. The filesystem also allows for performance to increase granularly as more disks are added. Typically this benefit will be felt immediately, even before "restriping" the data across both old and new spindles, as writes will automatically be spread immediately. As a result BlueArc customers may start small and scale performance by adding storage when needed. Additional file servers are not necessarily required for additional performance. As performance requirements grow even further, customers may also take advantage of clustering technology within BlueArc networked storage to add more servers while maintaining a single namespace, providing easy management of very large pools of data. But again, BlueArc networked storage offers true separation of function between storage and servers. Each may be scaled independently to meet a customer's needs; there is no requirement to purchase one to get the other as with many competing NAS products.

Best-in-class namespace scalability: Scaling beyond a single NAS server is essential for high performance storage solutions. Many parallel filesystem implementations rely on clustering multiple servers together for greater aggregate performance. The difference with BlueArc networked storage is the scale: individual servers are much more powerful than traditional CPU-based architectures, meaning fewer servers are needed in a given cluster to achieve some specified level of performance. BlueArc networked storage also makes it possible to create a single, unified namespace across the entire cluster of BlueArc file servers – making it appear as a single filesystem to all network hosts. This functionality is known as Cluster Namespace,™ or CNS. CNS satisfies the most common scalability requirements, allowing network hosts to access data on any BlueArc server in the cluster, regardless of physical location. BlueArc networked storage takes advantage of BlueArc's unique architecture to move data seamlessly between multiple cluster nodes with minimal impact to performance.

Advanced multi-tier storage mechanisms: Since data has an assigned value (by age, data type, owner, etc.) the ability to transparently relocate data to an applicable storage "tier" is a key feature of BlueArc networked storage. Transparency requires that applications and users do not have to be pointed to new locations following data migration. BlueArc networked storage provides policy-driven data migration mechanisms that allow data to be migrated transparently between many storage tiers. Individual storage tiers may also include 3rd-party, or foreign, filesystems accessible from the BlueArc servers via NFSv3 and HTTP. This ability to extend BlueArc networked storage to external devices allows integration with many 3rd-party appliances for deduplication, compression, or archival for example. Such data migration mechanisms also allow for repurposing of existing storage devices as external storage tiers, lowering total costs and offering easier platform transitions.

Robust data protection: BlueArc networked storage provides various mechanisms for ensuring data protection. The storage used by the filesystem is protected by traditional hardware mechanisms, such as the use of redundant arrays of inexpensive disks (RAID) to provide fault-tolerance. BlueArc networked storage also adds layers of functionality for further assurance of data preservation: enterprise features such as snapshots, replication, and high-availability cluster options are all part of the BlueArc networked storage data resiliency framework.

Advanced storage virtualization framework: A key advantage of NAS architectures over Storage Area Network (SAN) designs is the ability to more readily virtualize storage, simplifying data management and making storage provisioning much easier. BlueArc networked storage provides an advanced virtualization framework that includes a global namespace (CNS), file server virtualization, storage pools, thin provisioning, and robust quota support.

Storage Solutions From BlueArc For eDiscovery

Central to the BlueArc solution is a hardware accelerated file system that uses parallel processing and aggregates performance of multiple drives to eliminate the typical constraints of traditional software-based solutions. The file system and directory structure supports millions of files per directory, and thousands of concurrent users can be handled by the BlueArc file system.

Hardware acceleration allows each of these systems to handle up to 200,000 IOPS and move large amounts of data in file sizes up to 256 TB, and offer storage capacity of (currently) up to 16 PB. As data and user population grows, or as workstation and application server performance accelerates, systems can scale up to eight nodes in a single cluster. A single, global, Cluster Name Space capability keeps all this storage manageable and in control. NGS users will be happy to know that crunching at performance and analyzing large data sets which are frequently in the range of terabytes and petabytes, becomes possible with the BlueArc systems. Yet, these systems leave the door open for future expansion.

Clustering and Cluster Name Space (CNS) functionality pulls together the power of 2 to 8 nodes in a single cluster. There is a unified directory structure that provides a single logical view of the data regardless of where it resides in physical storage. When the application demands data throughput that requires additional processing power, the CNS option allows multiple nodes to act on the problem. File systems can be assigned and reassigned to virtual servers and physical nodes as usage and performance requirements change, without impacting user access to files and data global accessibility to data resources.

A comprehensive virtualization framework comprised of Enterprise Virtual Servers, Virtualized File Systems with the above-mentioned global CNS and Virtual storage pools with parallel RAID striping provide a storage infrastructure that is fast and easy to expand. The physical server can be divided into virtual servers, which in turn can be moved around with changing performance or high availability needs. A file system can be attached to a specific virtual or physical server. This virtualization capability is built into the server platform and runs without any performance degradation and provides the ability to dynamically allocate and respond to spikes in demand for high data throughput or data processing. For example, eDiscovery projects requiring a high level of storage resources can now be moved to a dedicated server in the storage pool to balance the workload.

Dynamic Read Caching functionality dramatically increases (up to 400%) read-intensive aggregate workload performance by caching read data across a cluster. eDiscovery customers with read-intensive workload profiles and the ability to stage data in an optimized workflow process can leverage read caching as a way to scale performance when and how they need it. Active files are immediately copied from lower performance storage tiers to a high performance Fibre Channel or SSD cache storage tier for use across physical or virtual servers. This aggregates bandwidth and improves response time to prevent quality of service issues during spikes in demand.

Tuned for unpredictable workloads, BlueArc storage systems are built from the ground up to handle the variability in performance, size and longevity of eDiscovery data in situations where many applications need access to the data at different points in time. At the early processing stage, a very large data set comprising of several hundred TB may be required. This data will often be accessed in high performance I/O bursts. On the other hand, data in later stages of the process may have lesser peak storage and performance requirements but with a mixture of reads and writes. This presents a mixed workload to the system, with different scalability requirements at various stages of the data lifecycle, making it very difficult to tune the system to any particular workload.

BlueArc storage systems adapt to changing workloads and scalability requirements to deliver optimal performance. This is made possible through the BlueArc storage management software that controls the hardware accelerated file system and the virtualization of servers and storage. BlueArc provides highly concurrent access and performance; virtualization partitions servers from storage thereby isolating the differing workloads preventing them from affecting each other, and automatically provisioning storage only as needed.

BlueArc Network Storage vs Traditional Networked Storage	
Hardware-accelerated, Object-based file system	**Software-based architecture**
Consistency of file system performance, regardless of block size or workloads	**Tuned for either small-block, random I/O or large-block, sequential workloads**
Support up to 256 TB of data per file system	**Capped at 16 TB of data per file system**
Supports millions of files in a single directory, while keeping directory search times to a minimum and sustaining overall system performance	**Scaling up impacts overall file system performance**
Automated data migration to tiered storage	**Manual and expensive data migration**
Policy-based, Secured Storage	**Unsecure or reliance on 3rd-party solutions**

BlueArc: The Right Solution for eDiscovery Environments Requiring High Performance and Scalability

Deploying eDiscovery applications without BlueArc networked storage is like building your dream home on a foundation of sand. eDiscovery applications like ECA promise fast ROI via lower processing and legal review costs. However, if you build eDiscovery applications on networked storage that is not high performance, is difficult to scale, and does not provide the necessary throughput, you will find that timely data identification and collection for litigation or early case assessment is difficult. Deploying your eDiscovery applications on BlueArc's next-generation storage enables that you can achieve ROI and respond quickly to eDiscovery demands. BlueArc systems provide excellent throughput and IOPS for mixed workloads that have varied and unpredictable characteristics. BlueArc bridges the gap between Legal and ITorganizations by ensuring that eDiscovery applications work at the speed Legal needs while remaining manageable and scalable for IT.

> **"BlueArc is blindingly fast. Ultimately it's the IOPS of BlueArc's silicon-based technology that outperforms NetApp and EMC's processor-based technologies. We wanted something that was engineered for performance.**
>
> Warren Roy
> President and CEO of Global Relay

BlueArc Corporation
Corporate Headquarters
50 Rio Robles
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