Thought Leadership Article
EDISCOVERY ARCHITECTURE AUTHORITY

The Case Against Hyper-Converged Infrastructure (HCI) for eDiscovery

APRIL 2020

eDiscovery. Perfected.

GEORGEJON

The Case Against Hyper-Converged Infrastructure (HCI) for eDiscovery

Among experienced professionals working in the eDiscovery industry, from software developers to infrastructure engineers, it is commonly understood that eDiscovery operates in a niche, though prolific, space in the technology landscape. However, within this tight-knit community of experts, there is a vast difference of opinion regarding how to best deploy and leverage software/hardware resources (commonly called a technology stack) to deliver performant, cost-effective eDiscovery platforms.

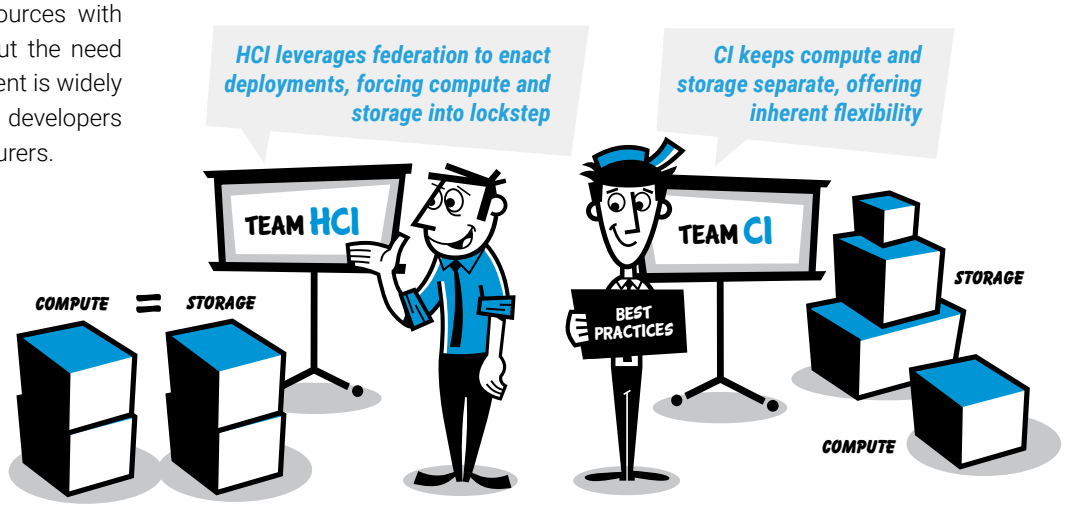
One of the most polarizing debates relates to the use of Hyper-Converged Infrastructure (HCI) versus Converged Infrastructure (CI, a.k.a. conventional/legacy hardware solutions) for the deployment of compute, network and storage resources.

The principal architectural difference between the two solutions is that HCI virtualizes compute/storage at or through a hypervisor, whereas CI physically separates these key functions. Simply put, HCI leverages federation to enact deployments, placing compute and storage in lockstep, whereas CI keeps them separate and flexible.

Proponents of HCI cite hardware/software integration management capabilities as the key advantage over conventional hardware. This "single pane of glass" operating model is an extremely marketable feature for technicians as it, in theory, addresses the age-old dilemma of balancing storage and compute resources with relative ease, full visibility, and without the need for specialized operators. This sentiment is widely endorsed and marketed by software developers and collaborating hardware manufacturers.

However, these developers/manufacturers built HCI solutions to address demand in the general IT market. As explained above, eDiscovery is far from generic, naturally unpredictable, workflow-intensive, and data-heavy. With this in mind, **it is not surprising to learn that HCI's value proposition for generic IT is specifically what makes it a sub-optimal solution for eDiscovery. This is what we will explore in detail in this article.**

To better understand why HCI is the wrong choice for eDiscovery systems, let's first look at the key components of the technology stack in isolation.



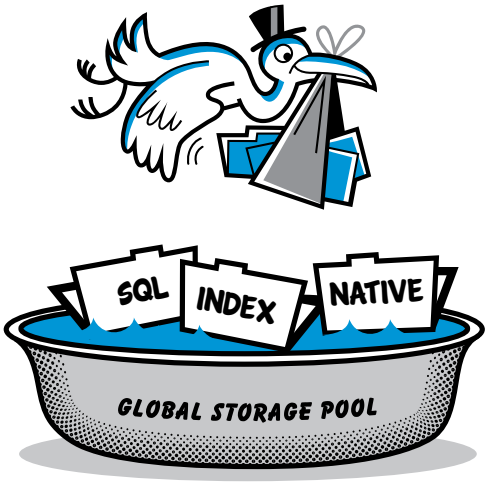
GEORGEJON

Storage

Storage in HCI environments leverages data locality (moving computations to where data resides versus moving data to computation) as a primary feature driving system performance, preventing data transfer over the network in order to maintain consistency and minimize network congestion.

eDiscovery data, however, is only sporadically used, and employing locality prevents system administrators from linking specific data repositories to performance tiers based on their workflow knowledge. This lack of flexibility in HCI environments results in volatile system performance and necessitates oversizing as a countermeasure, requiring an all-SSD configuration at a price point approximately ten times higher than traditional drives.

Another detrimental byproduct of using HCI is that workload predictability (matter size, type and urgency) is not native to the eDiscovery industry, and thus HCI's requirement for a single, global pool of storage is neither economical nor performant for multifaceted workloads (SQL, natives, indexes, source files).



Finally, it is well known that scaling eDiscovery environments is not a parallel endeavor for compute and storage. In an HCI solution, eDiscovery data will quickly outgrow the storage capacity per host, requiring the purchase of additional licenses and compute resources (hardware) every time additional storage is needed. This inability to scale in a non-linear fashion will inevitably put decision makers into a difficult position, requiring:

- Unnecessary capital investments to concurrently scale both compute and storage hardware resources;
- Implementation of innovative data management and archiving methodologies (easier said than done); and/or,
- Outsourcing eDiscovery operations to a secondary provider/platform when only storage scaling is required.

These unsavory scenarios are compounded when you factor in per-gigabyte (GB) pricing models (currently ranging from $3-$20) that drive up your cost base with the unused GBs, diluting your margins. At the end of the day, when you are unnecessarily spending money on hardware resources, operational improvements and outsourcing, your storage costs will be significantly higher with an HCI solution.

STORAGE FOR EDISCOVERY VIA HCI REQUIRES:

Unnecessary capital investments to concurrently scale both compute and storage hardware resources

Implementation of innovative data management and archiving methodologies (difficult)

Outsourcing eDiscovery operations to a secondary provider/platform when only storage scaling is required

Absorbing expensive per-GB costs for storage you do not need or use

Compute

Using HCI for eDiscovery forces administrators to either purchase more compute than necessary or leverage resources such as VMware DRS to shuffle virtual machines across multiple, less utilized hosts. Both scenarios are problematic, unnecessarily driving up platform costs and degrading performance.

HCI requires you to proactively purchase compute to ensure eDiscovery performance

HCI can utilize up to 50% of available compute resources to support integrated storage and overhead, with overhead consumption varying based upon machine activity (usually 10%-20%). In eDiscovery, however, the virtual machines that are sharing CPU and RAM resources are "always on" and continually taxing the system. With an HCI setup, this means that integrated storage, compute and overhead are all performing parallel to the eDiscovery application (Relativity, Nuix, Ipro, etc.) compute. During peak load, this will result in poor performance from a core count ratio unless additional compute is purchased beforehand. To illustrate the significant amount of additional compute that would be needed, research conducted by an industry-leading cloud computing company specializing in HCI software, cloud services, and software-defined storage states that the Storage Controller should provide either the same number of cores as a single socket in the host, or roughly 8-12 CPUs.

On top of purchasing additional compute to offset system overhead and storage compute requirements, service providers and platform owners must also license these unused core counts, including base software, up to the OS. In an industry where Microsoft SPLA is a major budget item for service providers, it makes little economic sense to pay for licenses and additional compute when you don't have to.

Using VMware DRS (and similar programs) to shuffle virtual machines in real-time to less utilized hosts consumes resources, perpetuates system failure, and drives up costs.

For eDiscovery providers and users, platform performance and stability are paramount to success. For this reason, we always disable automated intra- or cross-host balancing of VMs, as this usually hampers the OS and results in job/agent task failure and making it difficult for support teams to troubleshoot issues caused by automation.

In mission-critical eDiscovery environments, where prescriptive balancing and intimate knowledge of the platform ecosystem are prerequisites, just one misguided move based on CPU metrics and enacted by automation can bring the application to a standstill. Imagine a system-wide failure on a Friday evening with a production deadline looming, and you can see the latent danger.

As a best practice, George Jon leaves a specified amount of compute resources free in all environments to avoid compute contention, which is very costly for an HCI solution.

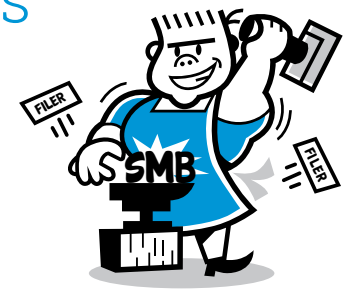
COMPUTE FOR EDISCOVERY VIA HCI:

Unnecessarily consumes costly resources

Requires that purchased overhead, which cannot be leveraged for eDiscovery performance, be fully licensed, a lose-lose cost proposition

Forces scaling in parallel between compute and storage regardless of need

Network Attached Storage (NAS) Filers



NAS-based filers, which are required for HCI solutions (but optional for CI) are used to write, protect, and organize data into large blocks. These filers are created through complex reverse-engineering of Microsoft's Server Message Block (SMB) protocol.

This imperfect process presents obvious and significant challenges when enhancements are made to the SMB protocol, as the source code is not released, thus leaving NAS software engineers in an endless cycle of reverse-engineering to accommodate enhancements/improvements. Even when the technicians manage to successfully recreate the SMB update, the overall lag in innovation is ever-present and can potentially lead to major issues.

When functioning, the large data blocks that NAS filers create artificially inflate the overall size of data growth in the environment. We have seen this data inflation range from 30% to 100%, which is an extremely high cost proposition considering that data growth and storage for eDiscovery tends to only scale up.

For the vast majority of general IT applications, NAS filers function without issue. But when they are employed for eDiscovery systems, they can promote inconsistent platform performance and compromise stability. When you have thousands of files running simultaneously, with large workload surges, the filers can fail and break the workflow process, forcing you to start over. These vulnerabilities manifest themselves as processing, production and imaging failures related to locked files and/or a systemic inability to locate files in a timely manner, commonly referred to as "time to first byte".

NAS FILERS FOR EDISCOVERY VIA HCI:

Lag in updates and improvements, which can lead to major environmental issues

Inflate overall data growth, driving up costs

Can compromise deadlines through inconsistent performance and platform instability

CASE STUDY

STUDY VARIABLES
George Jon conducted a real-world test, comparing an enterprise client system using HCI to our standard enterprise offering, running Relativity on both.

BENCHMARK TESTING RESULTS
The dataset was a group of mixed PSTs totaling nearly 300GB. The GJ Kit was discovered and published the files in 17 hours, while the HCI client environment completed the task in 47 hours. Both the discovery and publishing phases on the GJ Kit outpaced the client's HCI environment.

KEY FINDINGS
The 30-hour difference in performance is attributed to the Conventional Hardware "Kit" not competing for CPU resources and proper SQL sizing.

30 HOUR
DIFFERENCE BETWEEN HCI & CI SYSTEMS

Conclusion

It is our professional opinion, based upon fifteen years of real-world, industry-specific experience, that HCI is the wrong choice for eDiscovery systems

- Compute and storage for eDiscovery NEVER scale in a linear fashion. Yet HCI solutions must scale compute and storage in lockstep because of its integrated stack, which leads to significantly higher platform costs, both upfront and throughout the life of the system.
- Variable eDiscovery workloads, defined by function and data type, thrive when administrators can specify storage performance tiers to optimize resources. HCI only provides a single, global pool of storage, which is inflexible and does not allow for customization.
- eDiscovery environments DO NOT require the performance delivered by costly all-SSD (or "flash") solutions, but HCI solutions rely heavily on the all-SSD configuration to deliver acceptable performance.
- NAS-based filters, employed by all leading HCI solutions, ALWAYS lead to sporadic and inconsistent eDiscovery platform performance.
- eDiscovery operations yield unpredictable and inevitable data growth, large caseloads, heavy text volume, limited control of data queries, and 24/7/365 access to and maintenance of matters without expiration dates. These complex, mission-critical requirements are not aligned to HCI's inflexible "single pane of glass" model, from either a financial or operational standpoint.
- Committing to an HCI solution forces a "vendor lock-in" scenario, whereby users are unable to mix-and-match compute and storage resources for future scaling, architectural updates, or technology improvements. Once implemented, altering an HCI solution requires you to rebuild your systems from scratch, a costly proposition.
- The proprietary storage protocols inherent to HCI solutions typically result in complicated troubleshooting, requiring technology-specific, high-priced experts to ably maintain the system.

Do your business (and yourself) a favor by avoiding HCI. Adopt a more sustainable, scalable, cost-effective, and proven solution using conventional hardware for your on-premise or cloud-hybrid eDiscovery environment.

About this Document

George Jon (GJ) is an eDiscovery infrastructure, product and process specialist, delivering performant, scalable, fault tolerant environments for users worldwide. GJ works with global corporations, leading law firms, government agencies, and independent resellers/hosting companies to quickly and strategically implement large-scale eDiscovery platforms, troubleshoot and perfect existing systems, and provide unprecedented 24/7 core services to ensure optimal performance and uptime.

The Research

George Jon's (GJ) conclusions are informed by fifteen-plus years of conducting enterprise-class eDiscovery platform assessments, application implementations and infrastructure benchmark testing for a global client base. GJ has compiled extensive quantitative and qualitative insights from the research and implementation of these real-world environments, from single users to multinational corporations, and is a leading authority on eDiscovery infrastructure.

Contact Information

PHONE
(312) 850.4320

EMAIL
sales@georgejon.com

WEB
georgejon.com

The Authors

Jordan McQuown

VICE PRESIDENT OF TECHNOLOGY, GEORGE JON

Jordan McQuown is an authority in information technology, cyber security, electronic discovery, and digital forensics. He has written Thought Leadership articles for the American Bar Association's *Cybersecurity Handbook* and *Information Security Magazine*, and he is a regular speaker as a subject matter expert on the eDiscovery security, application and legal conference circuits.

Reynolds Broker

CHIEF OF STAFF, GEORGE JON

As a strategist, consultant and implementer for George Jon, Reynolds positions the company for sustainable growth and long-term success through business and technology roadmap development, cross-functional process and structural improvements, strategic initiatives management, and holistic data-driven reporting. His eclectic work and scholastic backgrounds are rooted in and delineated by a successful track record in the Technology (eDiscovery), Corporate Finance, and Government Affairs sectors. Reynolds holds an International MBA in Corporate Finance & Spanish (University of South Carolina) and a bachelor's degree in International Affairs (University of Georgia).