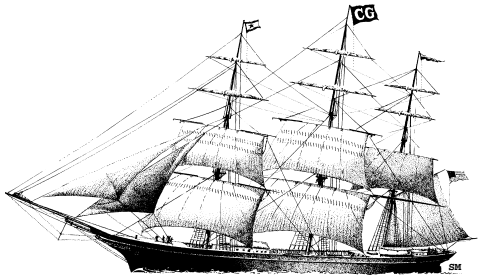
THE CLIPPER GROUP
Captain's Log™



Navigating Information Technology Horizons℠

Published Since 2001 Report #TCG2010020 April 22, 2010

In Search of Worry-Free Data Coherency Across the Enterprise

Analyst: Mike Kahn

Last week, I asked you to "Rethink Storage … Again".[1] Hopefully, you finished that bulletin with a sense that more might be possible. In this bulletin, I am going to be more specific and focus on a single thread in that more complex woven set of possibilities. First, let me reiterate some fundamental truths regarding IT, in general, and stored data, in particular.

There are four universal desires:

1. *Make it easier*
2. *Make it more transparent*
3. *Make it more efficient*
4. *Make it more effective*

In the most general case, the "it" is anything that falls under the IT umbrella, i.e., ***make IT easier, more transparent, more efficient, and more effective.*** You might call these the *Four IT Commandments*. In the storage arena, this becomes ***make storage easier, more transparent, more efficient, and more effective.*** Regardless of why you are involved in the ensuing discussion, the *Four Storage Commandments* always must be part of your equation.

Today, we embark on a *what-if* journey, specifically, *what would your world be like if we could change something to bring us closer to storage perfection (nirvana[2]), while achieving the Four Storage Commandments.* Like in the *Wizard of Oz*, on this journey there is a yellow brick road, a curtain, and maybe even a wizard. Let's begin!

Our yellow brick road might be described as the established path to where we want to go (with storage). We are warned neither to stray from it nor to think "off of the path". As easy, transparent, efficient, and effective as storage has become, we know that we need more of each of these four, if we are to survive on our continuing journey. *So, what keeps us from getting to where we want to go?*

Well, we tend to be burdened with a lot of distractions along the way. We must deal repeatedly with the familiar ones (like keeping everything safe and moving forward) and must pay close attention to everything new that confronts us (for example, the massive deployment of virtual machines and what that means for storage administrators). If only we could get to the Emerald City (the center of storage nirvana), our lives would be better (as would the state of data in the enterprise). The major problem is that many things are changing – from *where we were* to *where we are* to *where we think we are going* – not to mention that we are sharing the road (the networks and servers, etc.) with others (applications, users, data flows, etc.) who face the same dynamic stresses of change.

Understanding the Requirements

As an example of storage nirvana, in the world of file systems, nirvana is attained when the entire

[1] See the April 15, 2010, issue of *The Clipper Group Captain's Log* entitled *Rethinking Storage … Again*, available at http://www.clipper.com/research/TCG2010019.pdf.

[2] In this use, "a place or state characterized by freedom from or oblivion to pain, worry, and the external world". From *Dictionary.com Unabridged*. Random House, Inc.. <Dictionary.com http://dictionary.reference.com/browse/nirvana>.

world (of files in the enterprise) are cataloged in a *single namespace*. This singular, orderly logical representation solves many data integrity problems. Recognize that this is a virtual state of representation and not a physical statement of how or where files actually are stored; this is important. **We are much too accustomed to thinking about data being stored "here" and backed up or replicated to "there"**, where "there" might be several discrete locations. This is even true in a heavily virtualized world.

As we strive to satisfy the Four Storage Commandments, **users and their applications must abandon the concepts of *physical location*** (i.e., the need to know specifically which physical device(s) are holding the data) **in favor of *standards of delivery and quality*** (i.e., can the data be retrieved, stored, protected, etc., within acceptable parameters). Exactly how this is fabricated within the IT infrastructure is a detail left for its caretakers, most likely through policy-driven automation and real-time optimization. **For most situations, even the caretakers won't need to know where specific data is physically located.** This applies to blocks of data (think "databases"), as well as to files that are, of course, made up of blocks.

This is akin to the avionics in most modern airplanes and, now, increasingly in the systems of most passenger vehicles. Pilots of commercial aircraft don't fly the airplane anymore; they are flying a visually familiar, virtual representation of the airplane.[3] When the pilots manipulate the controls, they are really communicating with a computer system that actually engages the airplane's mechanisms (and usually can take actions at speeds faster than any human can execute).

For drivers, when you push on the brake pedal of your vehicle, these days almost always you are communicating with a "sensor" that measures the pressure on the pedal (and the speed of changing states, i.e., how quickly you changed from no pressure to a lot) and then communicates that to software that tells the braking mechanisms what to do to the mechanical brakes – at what pressure, in what manner (does it pump the brakes to prevent skidding), for how long, etc. Unless you are a car buff, you don't want to know how this is done but care only that it is done well and safely.

The same is true for all of the many physical copies of data that we have created, either explicitly (say, by sending a copy of a database to each branch office) or implicitly (e.g., by backup copies automatically made locally or remotely or extra copies of data strategically placed to improve performance). **Having to worry about all of these originals and copies has made storage management into a tedious and time-consuming set of chores that have come to dominate and even overwhelm storage administrators. This is anything but worry-free.** It fails the test of the Four Storage Commandments. It is stressful, time-consuming, and never-ending, plus it is getting worse under the new demands fostered by server virtualization.

We need a storage architecture and underlying infrastructure that is easy to use; that encourages transparency; that is efficient in storing, retrieving, and protecting data; and, hopefully, does this with greater impact (more effectiveness to the users and applications). What we desire is *informational consistency*, regardless of our location along the road. **In technical jargon, this is *data coherency*.** From the point of view of the users (and involved applications), if data is both coherent and accessible, a lot of things get better, even if we no longer know exactly where it is stored.[4]

Having data remain coherent is not too great of a challenge, as long as everyone is near where the data is stored and their behaviors are tidy. Unfortunately, much of the time neither is true. Add distance to the equation and you have more challenges. In solving these challenges, often we usually are faced with even more challenges, like managing replicated copies. **If only we could ignore the challenges brought on by distances and by the way we store, retrieve, and protect data, then we could treat data as if it were local. This now may be possible.** To consider and enjoy that possibility, please read on.

The Search for Data Coherency

What we need is a generalized solution to dynamic data storage (writing) and retrieval (reading), one that also enhances data availability, all while making its location

[3] This is 100% true for pilots flying drones remotely.

[4] This is one of those chasms that one most first recognize and then cross over to accept it as a foundational concept. We now accept the transparency of server virtualization and are willing to live with the accompanying vagaries regarding an application's physical existence. We must do the same for stored data.

irrelevant (i.e., transparent to those involved). To achieve this, we need a virtual storage architecture that:

- Appears locationless (i.e., is location transparent and may be somewhat fluid),
- Satisfies our quality-of-service requirements (it needs to be stored and retrieved reliably and fast enough),
- Is easy to administer (flexible), and
- Is transparent to applications (agile), and
- All while being cost-efficient (bringing down the total cost of storage, on a per unit of storage basis).

Thus, we must consider how dynamic our data is, as we seek to have real-time data coherency across our geographically dispersed domains of existence. Some data is inherently hard to keep coherent across geographic distances, for example online transactional data with many simultaneous writers (updating or adding data to the same data sets) processing many thousands of orders per second. We need to exclude these large-scale, real-time transactional data coherency challenges from the discussion that follows. (Maybe someday we'll be able to include them, but for now, they are a significant distraction.) Fortunately, those are a small fraction of the data volumes and use cases that enterprises leverage every day to run their operations.

Achieving Data Coherency

The Need to Inventory

First, everything needs to be inventoried (i.e., who (person, program, etc.) has what, where, and when was it last changed). It is even better if you know who uses it how often and when. While you can imagine what this might mean for the millions (billions) of files in your enterprise, that is the neater part of challenge, because there usually is context and coherency to files (starting with the name and its owner but also including the date and time it was last changed). Ultimately, all data is stored in bigger chunks call *blocks*, whose context and coherency may only be understandable to the programs that generated it; good examples are databases and file systems. **Their coherency depends upon all of their parts being consistent, usually aligned to the same point in time.**

Maintaining coherency over a distance just makes this task harder and illustrates why this is not a task suited for human handling. With potentially hundreds of thousands of additions or changes to non-transactional data per day in the largest enterprises, **keeping the data coherent (aligned and accurate) and well positioned needs to be done automatically (and driven by policies). Only then do you get the agility and flexibility that enterprises require.**

The Need to Optimize

Second, you need to realize that making this happen within the Four Storage Commandments is a mathematical puzzle. **Determining what should be kept where and in what state of currency is a complex, multi-dimensional optimization problem.** Not only does the optimizer require knowledge of everything stored (providing the coherency of data), it also needs to know how to anticipate where, how, and by whom it is going to be used, based on policy rules and patterns of usage. It also needs to be able to optimize the use of available resources dynamically, i.e., when they are available.

New Needs from Server Virtualization

Third, this all needs to be considered under a new, broadened definition of what data really is. Historically, we think of data as fields, files, documents, databases, and binary objects (like x-rays, photos, and videos). **However, in an era of widespread deployment of applications on virtualized servers, the server images (i.e., what is in each of the partitions) and the virtual machines (that contain the in-play partitions) also are another kind of data that needs to be managed well**. Being able to move "mobile virtual partitions" around easily (e.g., to rebalance the servers, take one or more out of service, or to relocate the processing to another geography) addresses many of the new challenges of managing a collection of applications on virtualized servers. **Being able to move or restart a running application, including all of its data and network connectivity, is part of today's storage challenge.** Just solving this challenge might actually be a sufficient reason to move forward with alacrity.

The Need for Automation

Fourth, if this sounds too tedious to imagine and potentially more of a problem than a solution, consider this. Most of us drive a vehicle with an automatic transmission. We don't think about changing gears; it happens *automatically*. On the other hand, there is an art to

driving a vehicle with a manual transmission. It involves many senses. You can hear (the whirring drive train), feel (the vehicle's vibrations), and/or see (on the tachometer) when you need to shift. Once you have made the determination to shift gears, you have to execute a multi-pedaled operation smoothly, at the right times, while otherwise driving the vehicle in a sea of other vehicles, whose behavior generally might be predictable but, at times, is not. Not having to worry about the many factors and operations required with a manual transmission makes sense for most of us, especially now that we have additional chores to manage, like our cell phones, music, GPS, and cup of coffee.

Just like the automatic transmission, maintaining data coherency becomes a sophisticated optimization game, with complex constraints (including resources and capabilities available at various locations, bandwidth and expected traffic, etc.) and priorities (e.g., this user or application or data is more important than are others). Just keeping the indices aligned across the locations can be a big challenge. Optimizing the data placements and updates – without planned pre-staging – is an even greater challenge. **Storage administrators need to move from focusing on physically managing storage to focusing on data policies. They need an affordable, worry-free automatic transmission that delivers the data coherency to many points along the way.**

Conclusion

Overwhelming, isn't this? That's why this is a challenge best left to storage management experts, especially those who can offer a packaged solution to make all of this happen without having to look under the hood. Such a solution would be multi-nodal, i.e., data would exist in multiple nodes – for protection, availability, and convenience. Some nodes would be local, others might be in the same metro area, and some would be at a distance, say in a remote data center.

This point is where the curtain and wizard come into play. Once you make it to the Emerald City, you need to believe in the experts that promise to take you to where you want to go, without sweating all of the hidden details. This requires a lot of trust in the Wizard. You want to believe, you need to believe, and if the Wizard delivers, you will end up much better off than before; but, please, don't look behind the curtain! That not only is unnecessary but also a distraction from what is important.

Thus, if you can believe, i.e., comfortably assume (without worry) that maintaining data coherency over a distance is not only desirable but also possible (at least most of the time), then you can begin to see how the storage game will change. If you can appreciate the benefits that will result from storage optimization via real-time analytics and policy-driven automation, it may be time to explore the possibilities.

Transparent, policy-driven placement of data is the *next big thing in storage*. (The last was storage tiering.) Together, these are the needed building blocks for success well into the future. *You had better find your way to the yellow brick road and begin your journey!*

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Mike Kahn **is Managing Director and a cofounder of The Clipper Group.** Mr. Kahn is a veteran of the computer industry, having spent more than four decades working on information technology, spending the last 17 years at Clipper. For the vendor community, Mr. Kahn specializes on strategic marketing issues, especially for new and costly technologies and services, competitive analysis, and sales support. For the end-user community, he focuses on mission-critical information management decisions. Prior positions held by Mr. Kahn include: at International Data Corporation - Director of the Competitive Resource Center, Director of Consulting for the Software Research Group, and Director of the Systems Integration Program; at Power Factor Corporation, a Boston-based electronics startup – President; at Honeywell Bull - Director of International Marketing and Support; at Honeywell Information Systems - Director of Marketing and Director of Strategy, Technology and Research; at Arthur D. Little, Inc. - a consultant specializing in database management systems and information resource management; and, at Intel Corporation, Mr. Kahn served in a variety of field and home office marketing management positions. Earlier, he founded and managed PRISM Associates of Ann Arbor, Michigan, a systems consulting firm specializing in data management products and applications. Mr. Kahn also managed a relational DBMS development group at The University of Michigan, where he earned B.S.E. and M.S.E. degrees in industrial engineering.

- ***Reach Mike Kahn via e-mail at Mike.Kahn@clipper.com or via phone at (781) 235-0085 Ext. 121. (Please dial "121" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group *Captain's Log***, **The Clipper Group Voyager**, ***Clipper Notes***, and "*clipper.com*" are trademarks of The Clipper Group, Inc., and the clipper ship drawings, "*Navigating Information Technology Horizons*", and "*teraproductivity*" are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosures

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

After publication of a bulletin on *clipper.com*, The Clipper Group offers all vendors and users the opportunity to license its publications for a fee, since linking to Clipper's web pages, posting of Clipper documents on other's websites, and printing of hard-copy reprints is not allowed without payment of related fee(s). Less than half of our publications are licensed in this way. In addition, analysts regularly receive briefings from many vendors. Occasionally, Clipper analysts' travel and/or lodging expenses and/or conference fees have been subsidized by a vendor, in order to participate in briefings. The Clipper Group does not charge any professional fees to participate in these information-gathering events. In addition, some vendors sometime provide binders, USB drives containing presentations, and other conference-related paraphernalia to Clipper's analysts.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.