Cloud and Virtual Data Storage Networking: Data Footprint Reduction

by Greg Schulz, IT Industry Advisor

Abstract: Data footprints – or information about how data are accessed, used, manipulated, and saved in a given system – are important to maintain as networks store increasingly large quantities of data. If not managed carefully, this metadata can become an additional storage burden. Many different data footprint reduction technologies are being implemented in order to improve efficiency: archive, compression, de-duplication, and thin provisioning. This article details the different technologies for data footprint reduction (DFR), as well as different techniques for data management. It also addresses data footprint impact, changing data access, changing lifecycles, and the economic benefits of DFR.

Introduction

The amount of active and inactive data that needs to be stored is continually growing. As this data growth rate climbs, your data footprint continues to expand and new business issues, challenges, and opportunities arise. Optimization, doing more with what you have or with less, and virtualization are both vital methods for managing your data footprint. These trends enable more data to be retained in a cost-effective manner without compromising quality of service or increasing associated infrastructure resource management (IRM) complexities. By improving the way vast amounts of data are stored, the overall storage network maintains its quality and efficiency, even as more and more data is acquired.

Getting Started

If it was not becoming increasingly necessary to process and store more data for longer periods of time in different locations, there would be little need for data storage, networking, IT clouds, or virtualization. Similarly, if there was not n ever-increasing dependence on information being accessible when and where needed—including data that was previously off-line or not even available in a digital format-there would be no need for business continuance (BC), disaster recovery (DR), or backup/restore as well as archiving.

However, as has been discussed in previous chapters, there is no such thing as a data recession and dependence on information continues to grow. Countering data growth and associated infrastructure IRM tasks as well as other data protection costs can be as simple as preventing data from being stored. Perhaps for a very few environments this is possible, along with implementing an aggressive data deletion policy, to counter data growth. However, for most environments, putting up barriers that inhibit business and economic growth are not the answer, although data management should be part of the solution and will be discussed later in this chapter.

Data footprint reduction is also about storing more data in a denser footprint. This includes storing more data managed per person, when the additional data being retained adds value to an organization. Also included is keeping more data readily accessible—not necessarily instantly accessible, but within minutes instead of hours or days—when access to more data adds value to the organization.

Another focus of DFR is to enable IT resources to be used more effectively, by deriving more value per gigabyte, terabyte, etc., of data stored. This also means alleviating or removing constraints and barriers to growth, or at least enabling those constraints to be pushed further before they become barriers. IT resources include people and their skill sets, processes, hardware (servers, storage, networks), software and

management tools, licenses, facilities, power and cooling, backup or data protection windows, services from providers including network bandwidth as well as available budgets.

Some aspects of addressing expanding data footprints and DFR include the following:

- Networks are faster and more accessible, but there is more data to move.
- More data can be stored longer in the same or smaller footprint.
- Some DFR can reduce costs while stretching budgets further.
- DFR can reduce the costs of supporting more information without negatively impacting service objectives.
- DFR allows existing resources to be used more extensively and effectively.
- DFR can be used to gain control of data rather than simply moving or masking issues that will pop up later.
- Unless optimized and DFR techniques are applied, data movement to clouds may not be possible in a timely or cost-effective manner.
- Desktop, server, and virtualization create data footprint opportunities.
- Consolidation and aggregation can cause aggravation without DFR.
- Backup/restore, BC, and migration to clouds are enhanced.
- More data can be moved in shorter time, enabling business resiliency.
- If you are going on a journey, what will you be taking with you, and how efficiently and effectively can you pack to move what you will need?

Organizations of all sizes are generating and depending on larger amounts of data that must be readily accessible. This increasing reliance on data results in an ever-expanding data footprint. That is, more data is being generated, copied, and stored for longer periods of time. Consequently, IT organizations have to be able to manage more infrastructure resources, such as servers, software tools, networks, and storage, to ensure that data is protected and secured for access when needed.

It is not only more data being generated and stored that causes an expanding data footprint. Other contributors to expanding data footprints include storage space capacity needed for enabling information availability and data protection along with supporting common IRM tasks. For example, additional storage capacity is consumed by different RAID levels to maintain data accessibility; high availability (HA) and BC to support site or systems failover; backup/restore, snapshots, and replication; database or file system maintenance; scratch or temporary areas for imports and exports; development, testing, and quality assurance; and decision support as well as other forms of analytics.

Debate is ongoing about the actual or average storage space capacity utilization for open systems, with numbers ranging from as low as 15–34% up to 65–85%. Not surprisingly, the lowest utilization numbers tend to come from vendors interested in promoting storage resource management (SRM) and systems resource analysis (SRA) tools, thin provisioning, or virtualization aggregation solutions.

What I have found in my research, as well as in talking and working with IT professionals in various sized organizations around the globe, is that low storage utilization can often be the result of several factors, including limiting storage capacity usage to ensure performance, to isolate particular applications, data, customers or users, to ease of management of a single discrete store system or for financial and budgeting purposes.

A point to keep in mind when consolidating storage is having insight as to where and how storage is being allocated and used (active or idle, updated or read) in order to know what policies can be set for when, where, and for how long to move data. Another important aspect of consolidation is leveraging newer, faster, and more energy-efficient storage technology as well as upgrading storage systems with faster processors, I/O busses, increased memory, faster HDDs, and more efficient power supplies and cooling fans.

Looking at storage utilization from the viewpoint of only space capacity consumption, particularly for active and on-line data, can result in performance bottlenecks and inability to service delivery. A balanced approach to data and storage utilization should include performance, availability, capacity, and energy in relation to the type of application usage and access requirements. When SRM and other storage management vendors talk to me about how much they can save and recoup from a storage budget, I ask them about their performance and activity monitoring and reporting capabilities. The frequent response is that it is not needed or requested by their customers or it will be addressed in a future release.

What Is Driving Expanding Data Footprints

There is no such thing as a data or information recession! Granted, more data can be stored in the same or smaller physical footprint than in the past, thus requiring less power and cooling per gigabyte, terabyte, petabyte, or exabyte. Data growth rates necessary to sustain business activity, enhance IT service delivery, and enable new applications are leading to continuously increasing demands to move, protect, preserve, store, and serve data for longer periods of time.

The popularity of rich media and Internet-based applications has resulted in the explosive growth of unstructured file data, requiring new and more scalable storage solutions. Unstructured data includes spreadsheets, PowerPoint, slide decks, Adobe PDF and Word documents, Web pages, and video and audio JPEG, MP3, and MP4 files.

The trend toward increasing data storage requirements does not appear to be slowing any time soon for organizations of all sizes.

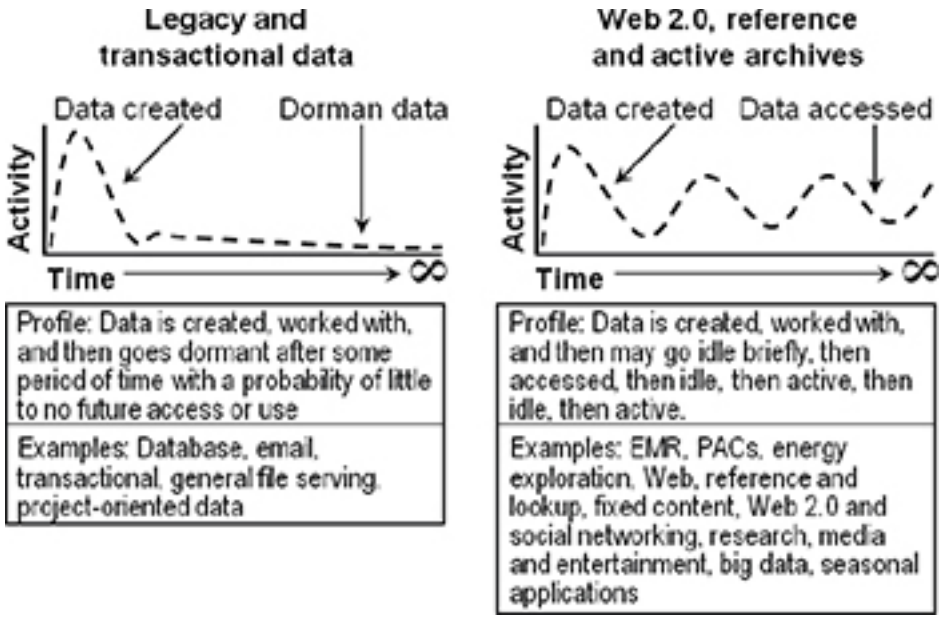

Figure 1 – Changing access and data lifecycle patterns.

Changing Data Access and Lifecycles

Many marketing strategies are built around the premise that, shortly after it is created, data is seldom, if ever, accessed again. The traditional transactional model lends itself to what has become known as information lifecycle management (ILM), by which data can and should be archived or moved to lowercost, lower-performing, and high-density storage or even deleted when possible. On the left side of Figure 1 is an example of the traditional transactional data lifecycle, with data being created and then going dormant. The amount of dormant data will vary by the type and size of an organization as well as the application mix.

However, unlike the transactional data lifecycle models, under which data can be removed after a period of time, Web 2.0, social media, unstructured digital video and audio, so-called big data, reference, PACS, and related data need to remain on-line and readily accessible. The right side of Figure 1 shows data that is created and then accessed on an intermittent basis with variable frequency. The frequency between periods of inactivity could be hours, days, weeks, or months, and, in some cases, there may be sustained periods of activity.

What Is Your Data Footprint Impact?

Your data footprint impact is the total data storage needed to support your various business application and information needs. Your data footprint may be larger than how much actual data storage you have, as in the example shown in Figure 2. This example is an organization that has 20 TB of storage space allocated and being used for databases, email, home directories, shared documents, engineering documents, financial, and other data in different formats (structured and unstructured) as well as varying access patterns.

The larger the data footprint, the more data storage capacity and performance bandwidth is needed. How the data is being managed, protected, and housed (powered, cooled, and situated in a rack or cabinet on a floor somewhere) also increases the demand for capacity and associated software licenses. For example, in Figure 2, storage capacity is needed for the actual data as well as for data protection using RAID, replication, snapshots, and alternate copies, including disk-to-disk (D2D) backups. In addition, there may also be overhead in terms of storage capacity for applying virtualization or abstraction to gain additional feature functionality from some storage systems, as well as reserve space for snapshots or other background tasks.

On the other hand, even though physical storage capacity is allocated to applications or file systems for use, the actual capacity may not be being fully used. For example a database may show as using 90% of its allocated storage capacity, yet internally there are sparse data (blanks or empty rows or white space) for growth or other purposes. The result is that storage capacity is being held in reserve for some applications, which might otherwise be available for other uses.

As an additional example, assume that you have 2 TB of Oracle database instances and associated data, 1 TB of Microsoft SQL data supporting Microsoft SharePoint, 2 TB of Microsoft Exchange Email data, and 4 TB of general-purpose shared NFS and CIFS Windows-based file sharing, resulting in 9 TB (2 + 1 + 2 + 4) of data. However, your actual data footprint might be much larger. The 9 TB simply represents the known data, or how storage is allocated to different applications and functions. If the databases are sparsely populated at 50%, for example, only 1 TB of Oracle data actually exists, though it is occupying 2 TB of storage capacity.

Assuming for now that in the above example the capacity sizes mentioned are fairly accurate in terms of the actual data size based on how much data is being backed up during a full backup, your data footprint would include the 9 TB of data as well as the on-line (primary), near-line (secondary), and off-line (tertiary) data storage configured to your specific data protection and availability service requirements.

For example, if you are using RAID 1 mirroring for data availability and accessibility, in addition to replacing your data asynchronously to a second site where the data is protected on a RAID 5–based volume with write cache, as well as a weekly full backup, your data footprint would then be at least (9 × 2 RAID 1) + (9+1 RAID 5) + (9 full backup) = 37 TB.

Your data footprint could be even higher than the 37 TB in this example if we also assume that daily incremental or periodic snapshots are performed throughout the day in addition to extra storage to support application software, temporary work space, operating system files including page and swap, not to mention room for growth and whatever free space buffer is used for your environment.

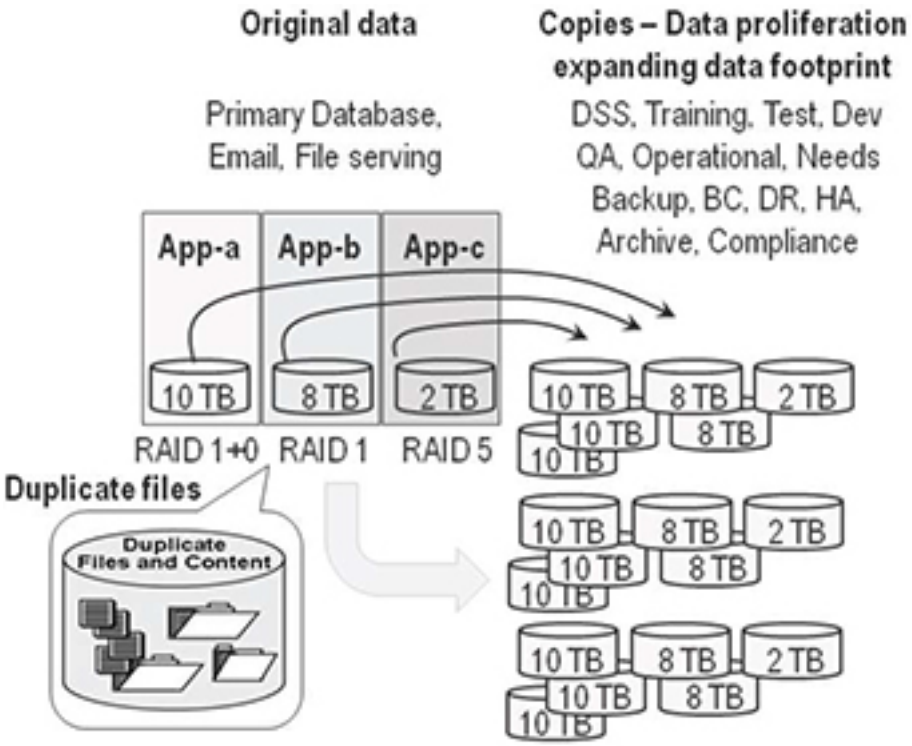

Figure 2 – Expanding data footprint impact.

In this example, 9 TB of actual or assumed data can rapidly expand into a larger data footprint, which only compounds as your applications grow to support new and changing business needs or requirements. Note that the above scenario is rather simplistic and does not factor in how many copies of duplicate data may be being made, or backup retention, size of snapshots, free space requirements, and other elements that contribute to the expansion of your data footprint.

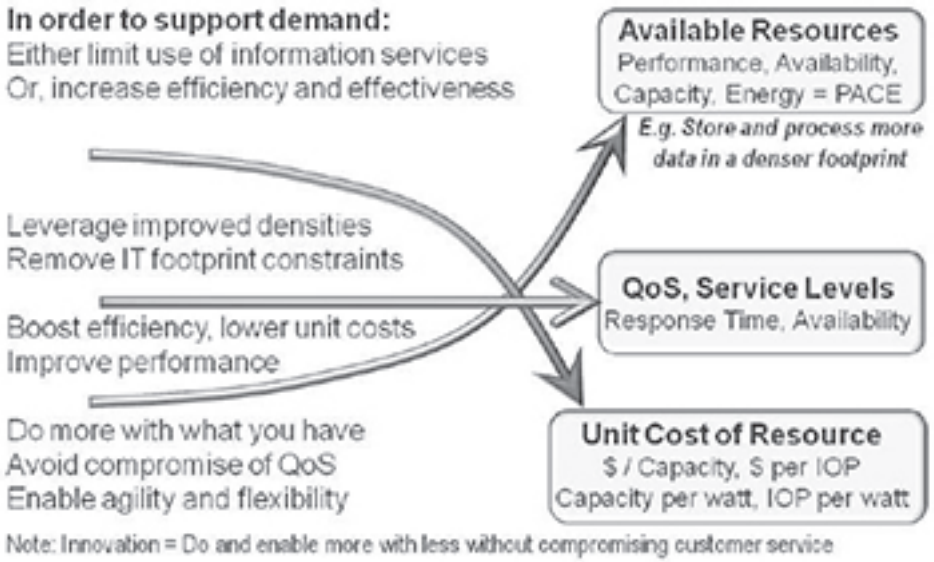

Figure 3 – IT resources, cost balancing, conflicts, and opportunities.

Business Benefits of Data Footprint Reduction

IT organizations of all sizes are faced with a constant demand to store more data, including multiple copies of the same or similar data, for longer periods of time. The result is not only an expanding data footprint but also increased IT expenses, both capital and operational, due to additional IRM activities to sustain given levels of application quality-of-service (QoS) delivery as shown in Figure 3.

Common IT costs associated with supporting an increased data footprint include:

- Data storage hardware and management software tools acquisition
- Associated networking or I/O connectivity hardware, software, and services
- Recurring maintenance and software renewal fees
- Facilities fees for floor space, power, and cooling, along with IT staffing
- Physical and logical security for data and IT resources
- Data protection for BC, or DR, including backup, replication, and archiving

As shown in Figure 3, all IT organizations are faced with having to do more with what they have—or even with less—while maximizing available resources. Additionally, IT organizations often have to overcome common footprint constraints (available power, cooling, floor space, server, storage and networking resources, management, budgets, and IT staffing) while supporting business growth.

Figure 3 also shows that to support demand, more resources are needed (real or virtual) in a denser footprint, while maintaining or enhancing QoS while lowering per-unit resource cost. The trick is improving on available

resources while maintaining QoS in a cost-effective manner. By comparison, traditionally, if costs are reduced, one of the other curves (amount of resources or QoS) is often negatively impacted, and vice versa.

The Expanding Scope and Focus of Data Footprint Reduction

Data footprint reduction is a collection of techniques, technologies, tools, and best practices that are used to address data growth management challenges. De-duplication ("dedupe") is currently the industry darling for DFR, particularly in the scope or context of backup or other repetitive data. However, DFR expands the scope of expanding data footprints and their impact to cover primary and secondary data along with offline data that ranges from high performance to inactive high capacity.

The expanding scope of DFR is moving beyond backup with dedupe to a broader focus that includes archiving, data protection modernization, compression, as well as other technologies. The scope expansion includes DFR for active as well as inactive, primary along with secondary, on-line and nearline or off-line, physical, virtual, and cloud using various techniques and technologies. Another aspect of the expanding focus of data footprint reduction is that a small percentage change on a large basis can have a big impact, along with the importance of rates in addition to ratios.

The main theme is that there is a bigger and broader opportunity for DFR across organizations to address different performance, availability, capacity, and economic or energy efficiency requirements using various techniques. In other words, avoid missing opportunities because you have become tunnelvisioned on just one or a few techniques. This also means you should avoid trying to use just one tool to address all issues or challenges and, instead, align the applicable techniques and tools to the task at hand.

While dedupe is a popular technology from a discussion standpoint and has good deployment traction, it is far from reaching mass customer adoption or even broad coverage in environments where it is being used. StorageIO research shows broadest adoption of dedupe centered around backup in smaller or small/medium business (SMB) environments (dedupe deployment wave one in Figure 4), with some deployment in remote office/branch office (ROBO) work groups as well as departmental environments.

There does continue to be early adoption in larger core IT environments, where dedupe complements already-existing data protection and preservation practices. Another current deployment scenario for dedupe has been for supporting core edge deployments in larger environments that provide support for backup and data protection of ROBO, work group, and departmental systems.

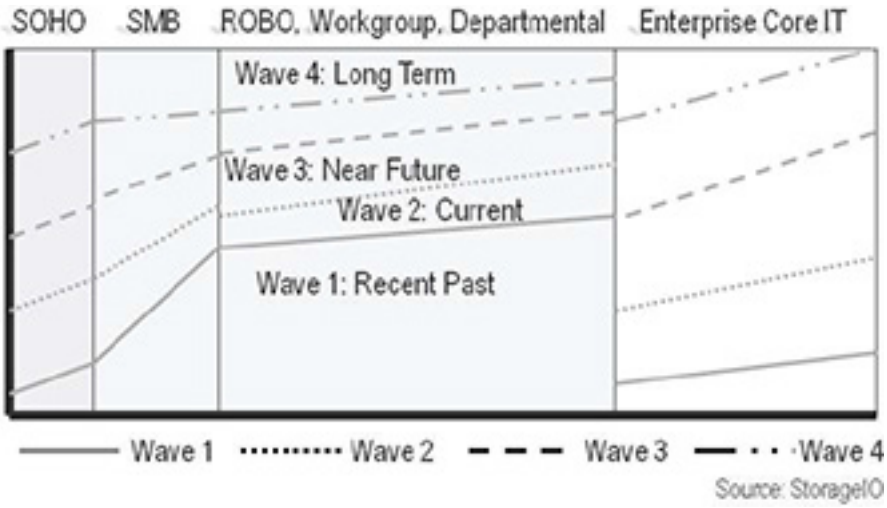

Figure 4 – Dedupe adoption and deployment waves over time.

Reducing Your Data Footprint

Storage has become less expensive per capacity, but as your data footprint expands, more storage capacity and storage management, including software tools and IT staff time, are required to care for and protect your business information. By managing your data footprint more effectively across different applications and tiers of storage, you can enhance application service delivery and responsiveness as well as facilitate more timely data protection to meet compliance and business objectives.

Reducing your data footprint can help reduce costs or defer upgrades to expand server, storage, and network capacity along with associated software license and maintenance fees. Maximizing what you already have using DFR techniques can extend the effectiveness and capabilities of your existing IT resources, including power, cooling, storage capacity, network bandwidth, replication, backup, archiving, and software license resources.

From a network perspective, by reducing your data footprint or its impact, you can also positively impact your SAN, LAN, MAN, and WAN bandwidth for data replication and remote backup or data access as well as move more data using existing available bandwidth.

Additional benefits of maximizing the usage of your existing IT resources include:

- Deferring hardware and software upgrades
- Maximizing usage of existing resources
- Enabling consolidation for energy-effective technologies
- Shortening time required for data protection management
- Reducing your power and cooling requirements

- Expediting data recovery and application restart for DR scenarios
- Reducing exposure during RAID rebuilds as a result of faster copy times
- Enabling more data to be moved to or from cloud and remote sites

Not All Data or Applications Are the Same

Data footprint reduction can be achieved in many ways with different implementations in different places to meet diverse application and IT requirements. For example, DFR can be done when and where data is created and stored, or it can be done after the fact, during routine IRM tasks including backup/restore, archiving, or during periods of application inactivity.

Not all applications and data are the same in terms of access or usage patterns as well as lifecycle patterns (as was seen in Figure 1). For example, some data is active-that is, being read or updated-while other data is inactive. In addition, some data and applications have time-sensitive performance requirements, while others have lower demands for performance.

Storage	Tier Tier 0 and Tier 1 Primary On-Line	Tier 2 Secondary On-Line	Tertiary Near-Line Or Off-Line
Characteristics focuses	Performance focused Some capacity needed	Some performance Emphasis on capacity	Less performance Much more capacity
	Active changing data Databases, active file systems, logs, video editing, or other timesensitive applications	Less active or changing data, home directories, general file shares, reference data, online backup and BC	Static data with infrequent access, on-line or active archives, off-line backups, archive or master copies for DR purposes
Metric	Cost per activity Activity per watt Time is money	Activity per capacity Protected per watt Mix of time and space	Cost per GB GB per watt Save money
DFR approach	Archive inactive data Space-saving snapshots Various RAID levels Thing provisioning and I/O consolidations, realtime compressions, dedupe if possible and practical	Archive inactive data Space-saving snapshots, RAID optimized, modernized data protection, thin provisioning, space and I/O consolidations, compression and dedupe	Data management, target for archived data, tiered storage including disk, tape, and cloud. Dedupe and compress. Storage capacity consolidation

Table 1 – Different Applications Have Various Data Footprint Reduction Needs

The importance of active vs. inactive data in the context of DFR is to identify the applicable technique to use in order to gain some data reduction benefit without incurring a performance penalty. Some data lends itself to compression, while other data is well suited for dedupe. Likewise, archiving can be applied to

structured databases, email, and Microsoft SharePoint and file systems, among others. Table 1 shows various applications and data characteristics as they pertain to DFR.

DFR Techniques

As previously mentioned, there are many different DFR approaches and technologies to address various storage capacity optimization needs. Likewise, there are different metrics to gauge the efficiency and effectiveness of the various approaches, some of which are time (performance) centric whereas others are space (capacity) focused.

In general, common DFR technologies and techniques include:

- Archiving (structured database, semi-structured email, unstructured file, NAS, multimedia, and so-called big data)
- Compression including real-time, streaming, and post processing
- Consolidation of storage and data
- Data management including cleanup and deletion of unnecessary data
- Data de-duplication, also known as single instancing or normalization
- Masking or moving issues elsewhere
- Network optimization
- Spacing-saving snapshots
- Thin provisioning and dynamic allocation

Conclusion

Organizations of all shapes and sizes are encountering some amount of growing data footprint impact that needs to be addressed either now or in the near future. Given that different applications and types of data with associated storage mediums or tiers have various performance, availability, capacity, energy, and economic characteristics, multiple data footprint impact reduction tools or techniques are needed. What this means is that the focus of data footprint reduction is expanding beyond that of just de-duplication for backup or other early deployment scenarios. For some applications, reduction ratios are an important focus, so the need is for tools or modes of operations that achieve the desired results. For other applications, the focus is on performance with some data reduction benefit, so tools are optimized for performance first and reduction second.

Greg Schulz

Greg Schulz is an independent IT industry advisor, author, blogger (http://storageioblog.com), and consultant. He has a B.A. in computer science and a M.Sc. in software engineering from the University of St. Thomas. Greg has over 30 years of experience across a variety of server, storage, networking, hardware, software, and services architectures, platforms, and paradigms. After spending time as a customer and a vendor, Greg became a Senior Analyst at an IT analysis firm covering virtualization, SAN, NAS, and associated storage management tools, techniques, best practices, and technologies in addition to providing advisory and education services. In 2006, Greg leveraged the experiences of having been on the customer, vendor, and analyst sides of the "IT table" to form the independent IT advisory consultancy firm Server and StorageIO (StorageIO). He has been a member of various storage-related organizations, including the Computer Measurement Group (CMG), the Storage Networking Industry Association (SNIA), and the RAID Advisory Board (RAB), as well as vendor and technology-focused user groups.

Greg has received numerous awards and accolades, including being named a VMware vExpert and an EcoTech Warrior by the Minneapolis-St. Paul Business Journal, based on his work with virtualization, including his book, The Green and Virtual Data Center (CRC Press, 2009). In addition to his thousands of reports, blogs, twitter tweets, columns, articles, tips, pod casts, videos, and webcasts, Greg is also author of the SNIA-endorsed study guide, Resilient Storage Networks-Designing Flexible Scalable Data Infrastructures (Elsevier, 2004).

Contributions

- Cloud and Virtual Data Storage Networking: Data Footprint Reduction
- Cloud and Virtual Data Storage Networking: Virtualized Desktops and Servers
- Cloud and Virtual Data Storage Networking: Being Secure Without Being Scared