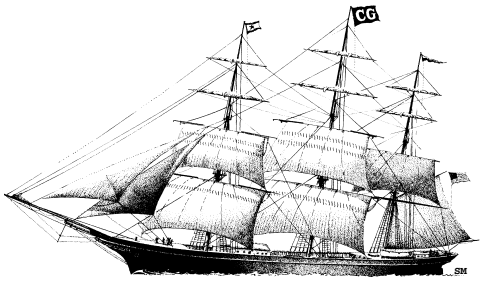
THE CLIPPER GROUP
Navigator™



Navigating Information Technology Horizons℠

Published Since 1993 Report #TCG2008044 September 12, 2008

NetApp Asks: *Why Not Deduplicate All Data?*

Analyst: Michael Fisch

Management Summary

The advertising campaign for the Florida Orange Juice Growers Association proclaims, "*Orange juice: It's not just for breakfast anymore*." While there is no real reason to limit orange juice consumption to breakfast time, people tend to think of it as a breakfast drink. It's habitual and cultural. Of course, the orange growers want to stimulate demand, so their ad campaign opens our minds to the possibility of drinking OJ at other times of day. Why not lunch or afternoon snack or even dinner? Seriously, why not?

Storage vendor NetApp is making a similar claim about data deduplication: *It's not just for backup anymore.* Data deduplication began several years ago as a special technology for eliminating redundancy in disk backup systems. Since backups are – by nature – highly redundant, deduplication is able to reduce the data by a factor of around 20:1, so enterprises can back up more data to disk and/or reduce the amount of storage purchased, while enjoying faster disk-based recoveries. This synergistic combination has propelled the rapid adoption of deduplication into the mainstream.

Now, NetApp is saying that orange juice isn't just for breakfast – and that deduplication can be applied to file shares and even primary application data. In support of this proposition, NetApp has made deduplication broadly available across its storage platforms by embedding it in its Data ONTAP operating system. The NetApp *FAS 2000*, *3000*, *3100*, *6000*, *V3000*, and *V6000* series now support this feature. There is no charge for the feature. Data deduplication provides greater storage efficiency, which translates into less equipment, less power, cooling, and floor space consumed, and less money spent on storage. NetApp's deduplication capability also is:

- **Content-agnostic** – supports any application
- **Protocol agnostic** – supports SAN and NAS connections to servers
- **Post-processing** – minimizes impact on performance by performing deduplication after data is written

Is there a catch? Depending on the type of data, the data reduction factor may not be as high as backup, though the efficiency benefits are still there. For instance, VMware virtual machines might experience a 70% data reduction and file shares a 35% reduction. Deduplication processing can also affect system performance, though NetApp's approach schedules it for off-peak hours.

Read on for more details about why NetApp is saying *deduplication isn't just for backup anymore*.

IN THIS ISSUE

Deduplication for Storage Efficiency

Why would enterprises want to deduplicate data? For the same reason many want cars with better gas mileage. For the same reason people drive straight to work, instead of taking long, circuitous routes. And for that matter, for the same reason your luggage has wheels: It is a more efficient use of resources.

Any discussion about deduplication has to begin with the ever-rising tide of data that enterprises contend with in the Information Age. Critical data keeps rolling in, and it must be stored, managed, and protected. Significant costs are associated with information storage. It consumes IT budgets, administrative staff time, power, cooling, and floor space. So, technologies that improve storage efficiency, like deduplication, increasingly are more valuable. They let enterprises store more for less.

One could describe deduplication as a bouncer who stands at the door and checks incoming data for redundancy. Only original data gets in, and the rest is asked to please step aside. The system detects and removes redundant data and inserts a pointer to existing data in its place. In this way, it helps minimize the number of bits and bytes stored.

The magnitude of reduction depends on the:

- Amount of redundancy in the data,
- Deduplication method used, and
- Scale of the data set to which it is applied

As a rule of thumb, backup data over time can be reduced by 90 to 95%, virtual machines by 70%, databases by 55%, file shares by 35%, e-mail PSTs by 30%, and document archives by 25%.

Specific benefits of data deduplication include:

- Defer purchase of additional storage capacity
- Eliminate administrative overhead associated with additional capacity
- Reduce power, cooling and space requirements – i.e., greening the data center
- Protect more data and retain it longer on disk
- Lower bandwidth costs for replication and remote backups, and
- Extend disaster recovery to data not previously protected

NetApp Data Deduplication

Deduplication is part of the *Data ONTAP* operating system that serves as the foundation of all NetApp storage platforms. Therefore, deduplication is available in the *FAS 2000*, *3000*, *3100*, *6000*, *V3000*, and *V6000* series of storage platforms. They include primary storage for the entry level, midrange, high-end, and nearline storage, plus gateways for heterogeneous environments. There is no charge for the feature, and it takes about 10 minutes to deploy.

NetApp employs 4K block-level deduplication. This identifies redundant data in relatively small blocks, independently of the application or data structure. Thus, it can deduplicate any data, in contrast to other more limited approaches that identify redundancy at the file or object level.

NetApp uses a custom hash algorithm plus a bit-level check to identify redundancy. A bit-level check is more processing intensive, but it eliminates the possibility of a false positive in identifying redundant data.

NetApp deduplication is performed post-processing – also referred to as *out-of-band*. Customers schedule the processing for off-peak hours (assuming such times are available), so it does not significantly affect storage I/O performance during production hours. As with everything in system design, there are tradeoffs with this approach. It requires a small capacity buffer to store each day's non-reduced data. If an enterprise wants to replicate data for disaster recovery, it will either have to replicate non-reduced data, which requires more bandwidth, or schedule replication after deduplication, which may require adjustment to the data protection scheme. These factors should be weighed against the storage savings from deduplication, and in many cases, it will be worth it.

Deduplication is enabled volume by volume, so customers can choose the applications and file shares to which it applies. It works in both NAS and SAN environments, since

NetApp "unified" storage systems support NAS (CIFS/NFS), SAN (Fibre Channel/-iSCSI), as well as NDMP for NAS backup.

NetApp claims that 15,000+ of its systems and over 3,000 customers have licensed the deduplication capability. This implies a great deal of interest in deduplication from its customer base.

NetApp Efficiency Technologies

NetApp offers deduplication as part of a larger tool set of storage efficiency technologies that include:

- *Snapshot* – Differential point-in-time copies
- *SnapVault* – Differential backup to disk
- *RAID-DP* – Dual parity RAID that can recovery from two disk failures
- *FlexClone* – Writeable Snapshot copies
- Thin provisioning – Just-in-time capacity allocation to improve utilization, and
- Deduplication

Since they are part of Data ONTAP, owners of NetApp storage systems can deploy any or all of them. Their effects are additive: The more that are employed, the greater the storage efficiency.

Conclusion

Deduplication has shifted from being an interesting innovation to a core technology for information storage. Disk backup data generally should be deduplicated. The savings are just too great not to deduplicate. However, NetApp raises a valid point: Why stop at backup? Enterprises stand to gain even more efficiencies by applying it to at least some production data. Seriously, why not?

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

Michael Fisch **is Director of Storage and Networking for The Clipper Group.** He brings over 12 years of experience in the computer industry working in sales, market analysis and positioning, and engineering. Mr. Fisch worked at EMC Corporation as a marketing program manager focused on service providers and as a competitive market analyst. Before that, he worked in international channel development, manufacturing, and technical support at Extended Systems, Inc. Mr. Fisch earned an MBA from Babson College and a Bachelor's degree in electrical engineering from the University of Idaho.

- ***Reach Michael Fisch via e-mail at mike.fisch@clipper.com or at 781-235-0085 Ext. 211. (Please dial "211" 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.

Disclosure

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.

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.