



TechTalk Chat Transcript
**Cut Data Storage Use by As Much As 95%
Using Deduplication to Drive Down Storage Costs**
July 19, 2007

Contents

1. A-SIS Deduplication
 2. VTL Deduplication
 3. Deduplication and...
 4. Deduplication General
 5. NetApp General
-

1. A-SIS Deduplication

Question: What product does NetApp use for data deduplication?

Answer: A-SIS deduplication in the Data ONTAP® OS.

Question: Apologies for something probably obvious...what does A-SIS stand for?

Answer: Advanced Single Instance Storage.

Question: What is SIS deduplication? What is Data ONTAP? Can you give us some background so that we can understand this better ?

Answer: Data ONTAP is the operating system that runs on all NetApp storage systems. A-SIS deduplication is a new feature that will do a block level deduplication of redundant data stored within a volume. Here is a good Webcast that explains this technology with a customer example:
http://communications.netapp.com/p/Network_Appliance/20070605000000WOD.

Question: Are you able to integrate deduplication with a NetApp 3020, or do you need another type of NetApp system?

Answer: A-SIS deduplication is compatible with the R200 and all FAS3000 and FAS6000 systems. A NearStore® license is required on the FAS systems.

Question: So what software package from NetApp executes the dedupe process?

Answer: A-SIS deduplication is integrated into Data ONTAP and since deduplication is done at the WAFL® block level, it is completely application transparent.

Question: What is the Network Appliance approach to deduplication in their various product lines?

Answer: Deduplication is available on any of our FAS storage systems by licensing both a NearStore license and an A-SIS Deduplication license. We suggest using A-SIS deduplication on backup data, archive data, or light-duty file-serving data.

Question: OK, how do you define the data as being redundant? Exactly identical files? Content? Name?

Answer: Duplicate 4K blocks, as defined by digital fingerprints.

Question: What applications are best suited to deduplication?

Answer: In general, A-SIS deduplication is application transparent. Deduplication, however, depends on the ability to find duplicate blocks. Therefore, any volume with a large number of duplicate files (or blocks) is a good candidate.

Question: A-SIS looks like a post-process Snapshot™ copy. Hasn't this been around for a while?

Answer: Sort of – A-SIS technology is built around the NetApp “block referencing” technology first used in Snapshot. A-SIS takes Snapshot technology a step further by not only using one block to represent many blocks, but by also removing the redundant blocks from the volume.

Question: What is the storage entity boundary of your deduplication - volume, aggregate, or filer?

Answer: Volume level boundary (FlexVol®).

Question: The space reduction is most optimal when the dedupe is being performed before the actual writing takes place. Does A-SIS work this way? If not, will this be a future enhancement?

Answer: The space reduction is the same whether inline or post-processing. The trade-off is usually in the performance – inline deduplication requires much more processing power. A-SIS works as a post-process operation and therefore has a much lower impact on read/write performance.

Question: Can you explain why A-SIS is not expected to cannibalize NTAPs core product line?

Answer: Data growth is moving at such a fast rate that reducing space requirements through deduplication will not slow down our company growth. Also, if we did not offer deduplication, our customers and prospects might well look elsewhere for a solution. Deduplication will soon become a mandatory item, and companies that don't offer it will be at a competitive disadvantage.

Question: Does A-SIS do dedupe on the fly?

Answer: A-SIS deduplication is run as a post-process after data is stored on

disk. Other NetApp dedupe technologies provide in-line dedupe (SnapVault® for NetBackup™) or prevent duplicate data from being stored in the first place (nondupe). NetApp future VTL dedupe will support both inline and post-process dedupe for backup to disk applications.

Question: How is it priced? By filer? By filer type? By amount of storage on the filer?

Answer: A-SIS is free. System requirement is an R200 or a FAS 3000/6000 system with a NearStore license.

Question: How do you figure out duplicate 4K blocks ? How much of a performance overhead does this cause ?

Answer: We create fingerprints for each data block. We compare these for duplicate candidates. Then, before we delete anything, we do a byte-by-byte comparison.

Question: Is it necessary to have the NearStore option license enabled on a NearStore R200 to obtain the A-SIS deduplication license?

Answer: No, the NearStore option is inherent in the R200.

Question: How does enabling A-SIS affect the disk usage reporting from the client's side when using tools like df? What effect does enabling default tracking quotas and hard quotas have on the storage usage reporting?

Answer: df is a command that display "logical" results. After running A-SIS, you won't see any difference with df or with quota reporting. NetApp provides the df -s command to verify the cumulative space physical savings per volume. The results are listed in number of blocks and % saved. Using A-SIS to reduce physical volume space would potentially allow you reduce quotas without impacting user SLAs.

Question: What is the cost? Is it model dependent?

Answer: There is no additional charge for the A-SIS deduplication license. A-SIS deduplication is available for all FAS3000 and FAS6000 models and the R200. A NearStore license is required on FAS systems; there is a small charge for this license.

Question: How are fingerprints stored? Is there potential for data loss if the fingerprints get corrupted? How can they be backed up?

Answer: A-SIS deduplication creates fingerprint file metadata that requires 1% to 3% of the total volume size. These files are stored in FlexVol metadata. The fingerprints are used only during the actual deduplication process. The fingerprint files are very resilient; in fact you could actually remove all fingerprint files entirely before or after deduplication, and simply rescan the volume to recreate the fingerprints. Corruption is not an issue since we always do a byte-for-byte block scan before removing any duplicate data.

Question: Is deduplication like a Snapshot a copy or replication? How are the different features used in different vertical markets? Are there regulations requiring that these be used?

Answer: A-SIS deduplication is a standalone feature; it can be used with or without Snapshot and replication. Basically, A-SIS uses "block sharing" technology and reduces space by allowing multiple files to reference the same data blocks.

Question: Does it take an outage to users to execute dedupe? Or will they just notice a performance issue?

Answer: The A-SIS process can be scheduled to execute during downtime, so the impact to users can be minimized. In addition, the overall impact of running A-SIS is minimal. Less than 10% write performance overhead is required during the process.

Question: So A-SIS is a post-write dedupe process that is actually working at the block level and not at the file level?

Answer: That is correct. In addition, each instance of A-SIS operates within a single FlexVol volume. You can choose which volumes to enable (or not enable) A-SIS on.

Question: Can A-SIS save me space on my LUNS if the data is being stored on a NetApp 3020 by a third-party VTL solution like Diligent, which does its own dedupe via their HyperFactor technology?

Answer: Can save space on LUNs. The end user might not see results, but the sys admin can reuse freed-up space for other applications. Unsure on the Diligent VTL solution.

Question: So do I understand you that it only works on NearStore and not on primary storage?

Answer: It works on a FAS storage system running a NearStore license. This system can be used for primary storage. We currently suggest using A-SIS deduplication for backup, archive, and light-duty primary storage.

Question: Darrin Chapman, A-SIS deduplication creates fingerprint file metadata which requires 1% to 3% of the total volume size. These files are stored in FlexVol metadata. How is this reported in terms of space utilization? Will it show as overhead in the df -s?

Answer: df does in fact utilize this data in the space-consumed calculations. It doesn't exactly point out those individual files, though. df -s reports the total net physical savings in the volume, including the space required by the metadata.

Question: If a customer wants to evaluate A-SIS and NearStore personality on a production box, what happens to the data after the evaluation is over and A-SIS and NearStore personality are turned off? Does an "undedupe" take place?

Answer: You can undedupe. But why would you want to? You'll love it...! ;^)

Question: Will dedupe span across multiple volumes? For example, if I have a duplicate block in a volume in a LUN and one on another volume on a NAS share, will the dedupe work on that block?

Answer: A-SIS is performed at the FlexVol level and will not span volumes or LUNs.

Question: What is the performance overhead of your A-SIS process?

Answer: Write overhead on an A-SIS-enabled volume is less than 10%, no negligible read overhead. The deduplication process is run as a post-processing (low-priority) operation.

Question: Storage companies tend to focus on what I like to refer to as "downstream deduplication." That is to say, dedupe that occurs during or as part of a backup or an archiving process. Yet much of the technology could be applied toward upstream deduplication at a level of granularity far better than existing solutions in the information management space. Is NetApp looking into this as an opportunity?

Answer: NetApp takes a different approach to deduplication with A-SIS. Consider it as general-purpose "volume" deduplication. We examine any data in any volume, and search for and remove duplicate data blocks. Because of this, A-SIS can be applied to all (upstream and downstream) tiers of storage: backup, archival, and even primary storage.

Question: When using A-SIS, I know that there are limitations on the volume size based on the hardware platform. However, what are the limitations on maxfiles when using A-SIS?

Answer: There are no limits on file sizes within volumes.

Question: So I imagine there is a lengthy process where you first define the volumes that you want the dedupe to run on, and then let it figure this all out?

Answer: Actually configuring A-SIS is trivial, two operations: sis on and sis start.

Question: How does NetApp deduplication technology differ from Symantec's or EMC's?

Answer: NetApp deduplication technology is similar in that we eliminate redundant blocks. It's different in that it is a part of our core storage software, not an add-on, and it works on more than just backup data. It is data agnostic on what it can deduplicate.

Question: Can A-SIS be suspended and pick up again where it left off, or does it need to run as one continuous process?

Answer: Yes, you can suspend the operation and then start again from the same point.

Question: I'd really like an in-depth talk about how exactly deduplication works. Is this merely for backup and recovery procedures, or is this also for live data storage?

Answer: A-SIS deduplication is application and content agnostic. It can be applied to backup, archival, and yes, live data. We simply examine the data in any volume and reduce the space requirements by removing duplicate blocks. For in-depth talk, I suggest you contact your NetApp sales rep or SE.

Question: So A-SIS is not free. We would need to purchase a license to turn on dedupe?

Answer: The A-SIS license itself is free. It requires a NearStore license also, but this is only a nominal charge. The exception is the R200 system, which can have the A-SIS license free without a separate NearStore license.

Question: Has anyone spoken about the dedupe reduction numbers? What I have seen from implementation is only a modest reduction due to the increase in Snapshot sizes

Answer: With A-SIS deduplication, customers have reported anywhere from 10% to 90% space savings. The amount of deduplication depends completely on the dataset. We have a Linux[®]-based space estimation tool that will "crawl" through any NFS volume and estimate the savings for you.

Question: What performance overhead has been experienced using deduplication on the FAS3020 with a NearStore option—for example, sharing iSCSI LUNs.

Answer: The performance impact of enabling A-SIS on a volume is write overhead, less than 10%; read overhead, no negligible overhead .

Question: Would A-SIS work on the volumes (with LUNS) that have fractional reserve lowered to 50%?

Answer: It can work on anything, but with LUNs and fractional reservations you need to be careful about not as much storage savings because of all the reserve space.

Question: Are there plans to have A-SIS find duplicate data across multiple volumes? Any timeline?

Answer: Yes, A-SIS aggregate-level deduplication is an engineering goal. No timeline yet, but it is a high priority.

Question: Does the dedupe interact with the OS in any way? For example, does a file copy automatically dedupe, or is the duplicated data removed at the next dedupe pass?

Answer: Because A-SIS dedupe is post-processing, there's an OS hook per se.

Question: Can you specify which volumes to apply it to, or is the only option to run on the whole filer?

Answer: Volumes are specified.

Question: If production data on NetApp is deduped, what happens during backup to tape? Will it be reconstituted so that conceivably backup will be larger than the production amount?

Answer: When data that's been processed by A-SIS is dumped to tape, the data is expanded.

Question: If there are multiple volumes, can A-SIS be limited to only a subset of all the volumes on the filer?

Answer: A-SIS operates at the (flex) volume level. So, yes!

Question: what about dedupe on primary storage?

Answer: A-SIS is supported for light-duty primary storage today. Please work with your local account team for specific questions regarding sizing and applications that will best benefit from A-SIS dedupe.

Question: what version of Data ONTAP do we need for SIS? we're currently on version 7.2.

Answer: You need version 7.2.2.

Question: If A-SIS affects the blocks on the disk, should I expect my Snapshot utilization to go up (initially) after a dedupe operation has completed?

Answer: Hmm, yeah, it could. If the active file system is deduped and Snapshot copies existed prior, than the Snapshot copies will indeed consume additional blocks.

Question: A-SIS appears to be a single-instance store fixed-content reduction mechanism. Are you reducing the data stored by eliminating duplicate files, or are you actually applying an algorithm to streams of file data as it arrives?

Answer: A-SIS is a post-processing operation. Essentially, we scan for duplicate blocks and readdress the block pointers to reference only a single block, thus releasing the duplicate blocks back to the volume. This algorithm takes advantage of the existing WAFL file system structure and is a low-overhead operation since we are merely modifying block pointers.

Question: If I already have a strong disk-to-disk backup environment installed in my data centers, how would NetApp be best applied for my data dedupe requirements across all of my supported platforms?

Answer: I'd suggest you start with volumes that you feel have a large amount of duplicate files, for instance data archival volumes. NetApp has a space estimation tool that would help you identify the space savings you'd see on a given volume.

Question: Is dedupe in Data ONTAP transparent to the backup application?

Answer: Yes, it is transparent.

Question: Darrin, about the license add question, it doesn't seem to work on my R200 that is running 7.2.2 code. Where can I get the correct license code?

Answer: You can get an actual 7-character key from your account team that is the key for a_sis.

Question: Is there a tool or command than can determine how much system resources are being used by the A-SIS operation, and can the operation be throttled?

Answer: You can use sysstat to see the performance of the storage system, but there aren't any specific A-SIS commands.

Question: What volume size limits does A-SIS impose?

Answer:

R200, 4TB

FAS2020, 0.5TB

FAS2050, TB

FAS3020, 1TB

FAS3040, 3TB

FAS3050, 2TB

FAS3070, 6TB

FAS6030, 10TB

FAS6070, 16TB

Question: Does Data ONTAP include tools to demonstrate the efficiencies of A-SIS?

Answer: Yes, df -s will show utilization from A-SIS.

Question: Deduplication claims to save a great deal of space. How much time does the evaluation of a typical document take?

Answer: A-SIS deduplication will scan and dedupe at the rate of 30 to 50MB/sec. This is a background process that runs as a low-priority task.

Question: If we choose to use A-SIS dedupe on FAS, can you describe how that impacts performance? Is it a heavy up-front hit for the initial dedupe and then a lighter impact after that, or is it pretty much a consistent added load on the available performance?

Answer: There are two aspects of performance. First, just enabling A-SIS on a volume places little additional load on the system. Second, the actual deduplication process does consume CPU, I/O, and memory resources. For that reason, we suggest that the deduplication process be scheduled as resources permit.

Question: Is there an API or any way for an application to leverage the dedupe capabilities?

Answer: At the moment, there are no APIs for A-SIS.

Question: what about the older FAS server like the FAS270? Is this compatible?

Answer: A-SIS only supports FAS3000 and FAS6000 systems, and the R200. As newer FAS systems are released, they will also be supported.

Question: What other changes are in store for A-SIS with Data ONTAP 7.3?

Answer: Well, you know, future stuff we don't necessarily chat about.... but, trust me, there's good stuff coming.

Question: Do you have any large customers running A-SIS across their environment on all filers? Also, how many enterprise customers do you think are using this functionality?

Answer: Virtually all our current A-SIS users are enterprise-class customers. Typically, they implement A-SIS on their secondary storage volumes and then gradually increase its usage across more application volumes.

Question: Do you plan to introduce an in-line dedupe technology?

Answer: No plans for this with A-SIS deduplication.

Question: When will this product be available for prime-time file-serving duty?

Answer: Supports tier 2 now (and can run on anything, as long as the NearStore option is licensed).

Question: Sorry if I'm missing something obvious. I understand that deduplication requires the NearStore license, but I'm unclear...does it require additional NetApp hardware? Or can an isolated filer with the proper licensing utilize deduplication?

Answer: It's all license based, no extra hardware is required.

Question: Based on customer experience and your testing, which use cases or data sets work best with A-SIS?

Answer: I'd say volume archives. You'd only need to run A-SIS once and then resize the volume after dedupe is complete. Instant savings and minimal effort.

Question: Do you currently have any customers that are using the NearStore personality license with A-SIS turned on to serve their primary data? If not, is it possible, or will it be in the future?

Answer: Yes, this is an emerging area for us. Today we have customers who are using A-SIS for light-duty primary storage, such as /home directories. Eventually, as we get more experience, I expect to see broader adoption into primary storage.

Question: So if this is block based, it's independent of file system (NFS, CIFS, whatever)?

Answer: Yes, it's independent of NFS/CIFS. Block determination is based on the WAFL file system in Data ONTAP.

Question: Is there any possibility to implement A-SIS deduplication on a FAS250 that is being used as a pure NAS box in a live production environment? CIFS and iSCSI licensed.

Answer: Sorry, A-SIS deduplication is not supported on older platforms such as the FAS250.

Question: It seems that numerous companies are trying to have a dedupe conversation; are there any differentiators between your solution and those of other tier 1 storage vendors?

Answer: No other tier 1 storage provider is shipping a general purpose deduplication feature on their systems. All other products focus only on backup data sets. The key differentiator for A-SIS is multitier deduplication: backup, archival, and light-use primary storage. Other tier 1 vendors have indicated they will "follow the NetApp lead."

2. VTL Deduplication

Question: Will deduplication be integrated into your VTL solution? And if so, how soon, hopefully?

Answer: Yes, it will. This will be available in early 2008.

Question: Is there a plan to introduce dedupe on non NearStore products? If so, what is the timeline?

Answer: Yes, NetApp will be introducing deduplication for the NearStore VTL platform early next year, and we are actively investigating dedupe for non NearStore storage platforms.

Question: How will the SIS dedupe in Data ONTAP integrate or be used with existing or new NetApp VTL customers? Can you elaborate on the architecture if there is a VTL solution with dedupe?

Answer: NetApp NearStore VTL will have its own version of deduplication early in 2008.

Question: I just bought a VTL1400 (2 heads) with 26 trays of 750GB drives. Will I be able to take advantage of dedupe?

Answer: Yes, NetApp plans to offer an upgrade to deduplication in a future software release. Customers who already have data stored on NearStore VTL disk will be able to run post-process dedupe to remove redundant data already stored, which will free up disk space. Please contact your NetApp sales rep for more information.

Question: I dedupe on my 6070 and use SnapMirror®/SnapVault to my VTL1400 then onto my tape library. Will the tapes be usable by a third party for restores, or will it need to be restored back to a NetApp device to reverse the dedupe?

Answer: In order to move the data to the VTL1400, you would need to back up the data via NDMP. Once the data is on the VTL, any tapes created will be in standard format.

Question: How will deduplication work with your VTL solution? Will it be like Data Domain and be in-band? And if so, what speed can my clients expect?

Answer: Deduplication for the NearStore VTL will be available in early 2008. Please contact your partner representative for further details.

Question: Should we wait on dedupe for VTL or go with Data Domain? They do in-band at up to 220 MB/sec. Will we see similar speed or greater with your VTL product and dedupe?

Answer: Since NearStore VTL will have both in-band and out-of-band dedupe engines, and it typically takes up to 30 weeks of full backups (with weekly full backups), you will be able to start using it today and dedupe data at rest when the feature becomes available. We expect that our in-band dedupe will outperform Data Domain because we don't have to capture everything. Anything that is missed will be deduped by the out-of-band engine.

Question: When data is stored as virtual tapes created by a backup application, what is the restore overhead on deduplicated data while it is reconstituted? Or what is the restore speed?

Answer: The restore performance, in general, is typically faster than the ability of most hosts to read the data.

Question: Do you have best practices and configuration information around using either NAS or VTL as a target for TSM?

Answer: Yes. TSM is different than other backup applications in many respects. NetApp provides both disk pool and VTL solutions and can provide best practices for both. Please contact your local NetApp sales office or partner for more information.

3. Deduplication and...

Question: Can SnapMirror and deduplication work hand-in-hand to mirror data to remote (COOP) sites?

Answer: Yes indeed!

Question: Does this work with Legato backup software?

Answer: A-SIS can deduplicate any data in a FlexVol volume of a NetApp storage array.

Question: Can data deduplication be accomplished in a global DFS environment with multiple DFS roots? If yes, how?

Answer: A-SIS doesn't care about the environment so much. How practical the actual space savings are may be an issue that is subject to the data type.

Question: Bill May, going off my previous question on SnapMirror and deduplication, what's about the greatest performance increase (as in the least amount possible being transferred over)? An issue we're currently experiencing with another product is that once there is a period greater than 6 hours, we have to replicate everything from the beginning (40TB+).

Answer: A-SIS and VSM work very well together. As long you dedupe first, than the amount of data across the wire/WAN/etc. is greatly reduced. I don't know of any issues with SnapMirror where we'd have to start from scratch.

Question: Does NetBackup need to be used in conjunction with SnapVault in order to take advantage of deduplication?

Answer: NetApp to NetApp SnapVault inherently deduplicates data and is not dependent on NetBackup. Our SnapVault for NetBackup solution is a unique integration with Veritas™ that deduplicates traditional NetBackup data streams in real time and stores them in a Snapshot/SnapVault type format. This solution requires NetBackup 6.0 or later and is used for non NetApp environments using NetBackup.

Question: How does running sis on a SnapMirror source impact the snap delta? In other words, how much change must be pushed over the network to the destination? Is it just pointers, or is it a large number of changed blocks?

Answer: With volume-based SnapMirror, A-SIS deduplication is performed only on the source volume. Since the source volume is replicated block for block to the VSM destination volume, this results in deduplication space savings at the source and the destination, as well as reduced bandwidth during the replication data transfer. A NearStore license and an A-SIS deduplication license are required on both the source and destination VSM volumes. With qtree SnapMirror, A-SIS deduplication can be enabled at the source volume, the destination volume, or both the source and destination volumes. If the qtree SnapMirror source volume is deduplicated, you will not see any bandwidth savings, and the volume will be "unduplicated" as it is stored at the destination. To deduplicate the destination, A-SIS deduplication would need to be run on that volume after the qtree SnapMirror data transfer.

Question: How does your disk A-SIS technology interact with the A-SIS implementation already integrated in NetBackup?

Answer: SnapVault for NetBackup deduplication is referred to as SIS (single instance storage). It is application specific to NetBackup, and only deduplicates NetBackup tar files that are stored on NearStore systems. Because A-SIS deduplication works at the WAFL block level and is not tied to any particular application, it provides saving across a wide range of environments.

Question: I am buying your new FAS2050HA with SAS disks. I was wondering when the expansion trays might be available and how I could use deduplication with this box for data storage?

Answer: The FAS2050HA supports A-SIS deduplication when this system is delivered. A NearStore license is required. SATA expansion shelves are available for order now.

Question: Do you have a timeline as to when Open Systems SnapVault backups will be deemed compatible with deduplication?

Answer: OSSV backups are inherently deduplicated at the source. Data that does not change is not backed up again, making it very network and storage efficient. Deduplication levels will increase in subsequent releases. Data ONTAP 7.3 will be the next version with these improvements.

Question: Do you have best practices on how large a volume should be? I understand that dedupe is across a volume only.

Answer: Volume size is very dependent on what your data requirements are.

Question: Will dedupe work on iSCSI LUNs?

Answer: Yes.

Question: With deduplication, the benefits are really seen in replication—so you replicated the compressed data with A-SIS?

Answer: Yes. Deduplicated data at the source will reduce the network requirements when using volume SnapMirror for replication.

Question: What does the Snapshot churn look like when it is running (after it has run)?

Answer: Snapshot copies prior to process will reduce the effectiveness of A-SIS deduplication on the volume. Best practice is to take Snapshot copies after running A-SIS deduplication and delete any unnecessary Snapshot copies.

Question: With Windows[®], dedupe might be fairly easy, but with nonstandard installations of all flavors of UNIX[®] potentially out there, how will dedupe handle these types of scenarios?

Answer: Nope, dedupe with UNIX is an easy too! ;^)

Question: Please identify the deduplication products that are available for the NetApp 3050 filer.

Answer: A-SIS deduplication is available for the FAS3050. A NearStore license is required.

Question: I notice in using A-SIS on a SnapMirror destination that the daily Snapshot copies are still quite high, based on the amount of data transferred. Is there a way to prevent that, since SnapMirror transfers all changes before the deduplication process runs?

Answer: If you're only running A-SIS on the destination (so it must be qtree SnapMirror), the Snapshot copies can "lock" data, but eventually the data comes back. This could be an hour-long white-board conversation (so if you are someplace nice, have your SE ask and I can come do that).

Question: Is A-SIS effective on iSCSI Exchange stores? Would it provide any benefit to try to dedupe a filer-based Exchange environment?

Answer: Some Exchange data dedupes well, and some doesn't. Dedupe can be very data-set dependent.

Question: Let me clarify my question about VSM and A-SIS. If you have a large VSM set up already (~4TB) with a relatively low change rate (~10GB/day), and you turn sis on the source, will the next SnapMirror update after sis completes be very large? In other words, if you get 25% dedupe, are you going to push 1TB worth of SnapMirror update? Or will the sis process take many days or weeks to complete and generate lots of smallish daily mirror updates?

Answer: With VSM, if you dedupe on the source, you don't need to send *any* of the dedupe blocks. It rocks. Really.

Question: I had heard that you need to have A-SIS turned on on both the source and destination in a qtree SnapMirror relationship, but it appears that this is not true. I could have A-SIS on our R200 without it being on the primary?

Answer: Yes, you can have it turned on just on the qtree SnapMirror secondary.

Question: What kind dedupe ratios have you seen with SnapVault secondary, since it already dedupes?

Answer: Depends on the change rate and retention time. We did a recent case study of a customer with a 30-day retention and they saw a 10:1 compression. The longer the retention time, the higher the compression levels.

Question: Will the dedupe for virtual tape work with all major backup applications? Or just specific ones? (NetBackup/Backup Exec™, ARCserve, CommVault, NetWorker, etc.)

Answer: Yes, both A-SIS and VTL dedupe will work with all major backup applications. NearStore with A-SIS acts as a disk target, while the NearStore VTL acts as a tape target. In addition, NetApp provides integration options with backup partners such as Symantec®, CommVault, and SyncSort.

Question: Currently we keep 90 days worth of Snapshot copies on R200 and others on tape. How does dedupe work?

Answer: Snapshot copies and deduplication are separate processes that are run independently.

Question: Can dedupe and SnapMirror be used in combination to reduce the bandwidth required for replication?

Answer: Yes, it requires volume SnapMirror.

Question: If I run A-SIS against a volume that I want to replicate via SnapMirror, would I schedule the A-SIS, then schedule the SnapMirror to the remote site?

Answer: With volume SnapMirror, A-SIS deduplication is performed only on the source volume, so you would schedule it to run prior to your SnapMirror transfer. With qtree SnapMirror, A-SIS deduplication can be enabled at the source volume, the destination volume, or both the source and destination volumes. If the qtree SnapMirror source volume is deduplicated, you will not see any bandwidth savings, and the volume will be “undeduplicated” because it is stored at the destination. To deduplicate the destination, A-SIS deduplication would need to be run on that volume after the qtree SnapMirror data transfer.

Question: Does A-SIS work in combination with V-Filer solutions?

Answer: We do not currently support V-Filers with A-SIS.

Question: Using the 3020c's NearStore personality, do you take a greater hit on performance when you're also using this for SAN/NAS/iSCSI activity as well?

Answer: Obviously, the more stuff you run on a filer head, the more impact there could be.

Question: Are your management tools (like Operations Manager) A-SIS aware?

Answer: At the moment, Operations Manager isn't A-SIS aware.

Question: If you are running SnapVault from target to source, do you need a NearStore license on both target and source, or just on one or the other?

Answer: You will need a SnapVault primary license (or OSSV for non-NetApp storage) on the source and a SnapVault secondary license on the target.

Question: I'm currently using FLM for ILM. If A-SIS gets implemented, what happens to data that has already been migrated to second-level storage? Is it remigrated and then deduplicated?

Answer: Hmm, well I would run it on the secondary storage location, versus the primary location where the stub is.

Question: I have existing SnapVault or SnapMirror secondaries on NearStore systems that are not deduped. How do I go about using A-SIS? Do I need to reinitialize my snap relationships?

Answer: SnapVault not supported today. SnapMirror is. Volume SnapMirror requires A-SIS to run at the source system. You could run dedupe at source and then update SnapMirror without reinitializing. With qtree-level SnapMirror you can run A-SIS at either filer, no reinitializing required. Stay tuned for SnapVault support.

Question: Can we integrate A-SIS into an existing environment ? We have a few NearStore systems with which we SnapVault (including OSSV) and SnapMirror.

Answer: You can enable A-SIS on your NearStore systems at no charge if the

systems are R200s. A-SIS won't further deduplicate the already deduplicated SnapVault data. You can impact the SnapMirror data, depending on the configuration and data profile.

Question: Is there significant latency when using deduplication and SnapMirror together for a backup?

Answer: A-SIS deduplication has negligible effect on SnapMirror latency.

Question: So this is part of the OS now? How does it compare with the EMC Avamar offering? I know that this is not part of the EMC OS, but from a feature functionality perspective?

Answer: The two products are completely different. Our deduplication is embedded within the storage system, whereas Avamar requires that a backup agent be installed on all source data system.

Question: I still have questions about this dedupe technology. I have 2 FAS3020c 's in a cluster and a R200 NearStore. I run SnapVault to snap the data from OSSV clients and iSCSI LUNs to the NearStore system. How would this dedupe technology improve my environment?

Answer: Today, SnapVault/OSSV is not supported. However, like SnapMirror, you would simply run A-SIS on the destination system that contains your SnapVault backups (no need to do new level 0). You would potentially achieve space savings. Depends on the number of backups, dataset types, etc.

Question: It probably goes without saying that A-SIS and SnapLock[®] don't (or won't) work or play nice with each other?

Answer: Jason, Jason, Jason, it's good I know ya buddy—asking a question like that!!!! ;^) SnapLock isn't supported with A-SIS until 7.3. Happy now?!?!? ;^

Question: How is A-SIS affected by using SnapRestore[®]?

Answer: It shouldn't be affected.

Question: Is A-SIS supported on V-Series?

Answer: No, not today. Data ONTAP 7.3 will support this.

Question: What is the performance impact of a SnapVault for NetBackup integrated solution that deduplicates backup data on the fly?

Answer: There's a Tech Report that goes into that (it's only internal right now, so your SE will have to share it with you). It depends on which platform, how many streams, network connections, etc.

Question: Does this affect local and remote replication?

Answer: A-SIS deduplication will lower your bandwidth requirements for data replicated with volume SnapMirror.

Question: If we need to copy data that has been deduplicated to another filer, is it reconstituted, copied, then rededuped? (Assuming dedupe on both back ends.)

Answer: If you are copying data via SnapMirror. With volume SnapMirror, A-SIS deduplication is performed only on the source volume. Since the source volume is replicated block for block to the volume SnapMirror destination volume, this results in deduplication space savings at the source and the destination, as well as reduced bandwidth during the replication data transfer. With qtree SnapMirror, A-SIS deduplication can be enabled at the source volume, the destination volume, or both the source and destination volumes. If the qtree SnapMirror source volume is deduplicated, you will not see any bandwidth savings, and the volume will be "unduplicated" as it is stored at the destination. To deduplicate the destination, A-SIS deduplication would need to be run on that volume after the qtree SnapMirror data transfer.

Question: Can we use A-SIS on a 7600?

Answer: Yes, as long as you have 7.2.2 and purchase the NearStore license from IBM.

Question: If we dedupe and then back up the FAS, is the data stored deduped on tape?

Answer: No, it will be expanded again on tape. No pointers on tape.

Question: TSM works in base an incremental back-up. Only the first back-up is full, TSM doesn't have duplicate files, is that true?

Answer: TSM does have an incremental forever backup mode. NetApp technology is different in that the deduplication is transparent. The restore process is completely transparent to the fact that the data has been deduplicated. TSM requires the data to be reconstructed from the full and relevant incrementals during a restore.

Question: Does this work on a gateway appliance using Data ONTAP?

Answer: Support for our gateway appliance (V-Series) is slated for early in 2008.

Question: How does A-SIS deduplication compare to file-level compression? Can one expect comparable results, or is it too dependent on the datasets to say?

Answer: File-level compression like that found on tape environments typically achieves around a 2:1 compression. Deduplication can achieve much higher compression ratios, especially with backup data.

Question: Hmmm, so if SnapVault isn't currently supported, would that mean OSSV is also not currently supported?

Answer: Correct, OSSV isn't currently supported.

Question: Even though you have a 3020c, could you still purchase a NearStore appliance, keep that in a remote location, and use that as your "off-site" backup?

Answer: Yes. That is a very common configuration when using SnapVault for backups because it is very network efficient, making replicating to an off-site location very practical.

Question: With the recent announcement of CommVault Simpana, do you guys have a timeline on when you'll white paper the performance with the new software and A-SIS?

Answer: We are working with CommVault to provide integration support for A-SIS and Simpana. Expect full support and white paper early in 2008.

Question: Since I've found that NetBackup (any version) is not appropriate for 75% of my multi-PB environment, what plans, if any, does NetApp (dedupe, or any other component) have to integrate with my TSM enterprise environment?

Answer: A-SIS deduplication is application agnostic and will work well with TSM. We have been testing and are working with IBM to publish these results.

Question: Some answers seem to indicate that running A-SIS on tier 1 production data such as Exchange, SQL, etc. is possible. How do you know when this is viable?

Answer: It is possible, as long as you have the NearStore option licensed. Also, you'll want to consider how often Snapshot copies are required, etc. If highest performance is required, or if frequent Snapshots are maintained for a long time, it may not be a good fit.

Question: Sorry if this has already been asked - my question is, when can we expect integration of A-SIS with FilerView®?

Answer: Soon—I hope/wish. Integration with some sort of UI is a key next/future step. We always do the techie stuff first.

Question: One of my clients uses Legato DiskXtender and another couple use Symantec Enterprise Vault™ for archiving. How does dedupe work with these tools?

Answer: It works fine, since it just dedupes the blocks on the storage medium. How well the data actually dedupes depends on the specific apps.

Question: We use SyncSort BackupExpress to create OSSV backups of our data. SyncSort has informed us that they have not tested deduplication with their product and do not support it at this time. Any ongoing activity with SyncSort or other backup vendors (in addition to Veritas)?

Answer: SyncSort and OSSV data are already deduplicated, in that data that doesn't change is not backed up again. Each backup is effectively a full, but only the changed data is transferred and stored. A-SIS does not increase the deduplication level with this data, but upcoming versions of Data ONTAP will further deduplicate the data.

4. Deduplication General

Question: How are you defining deduplication?

Answer: Block-based space reduction by removing duplicate blocks.

Question: Can you explain deduplication and how it works?

Answer: Deduplication is the process of finding duplicate data objects and removing them, resulting in substantial disk space savings.

Question: Can you please explain about this technology?

Answer: A-SIS deduplication is integrated into FAS systems that have NearStore licenses. Deduplication removes duplicate blocks and reduces space requirements.

Question: Do you have tools or logs that report on how well the deduplication is working?

Answer: We have a tool our SEs can bring out that predicts how well data sitting *anywhere* (that is, not on NetApp storage) will dedupe with A-SIS. On the NetApp storage array, when you're actually running A-SIS, `df -s` shows what you need.

Question: How does it do what it does?

Answer: A-SIS deduplication really is an extension of our Snapshot functionality. Here is a Webcast that goes into more detail:
http://communications.netapp.com/p/Network_Appliance/20070605000000WOD.

Question: Did anyone provide a primer for this discussion that we are unaware of?

Answer: This chat event is a follow-up to a previous NetApp Webcast held in June. The following is a link to replay that Webcast:
http://communications.netapp.com/p/Network_Appliance/20070605000000WOD.

Question: Bill, going off the tool that will show how well the deduplication is working, I'd definitely like to see this in play. I am currently working on scheduling a demo at NetApp @ RTP. What should I ask the SE in the demo for?

Answer: Ask 'em to see me (I'm in RTP)!!! ;^) Ask 'em about the A-SIS storage savings tool.

Question: When does the deduplication take place?

Answer: Post-processing, at either scheduled times, manually, or automatically when a certain amount of extra data exists.

Question: We do not have NetApp technology in house today. Can we still use and take advantage of the data deduplication technology?

Answer: To take advantage of NetApp deduplication, you would need to purchase a NetApp FAS system.

Question: Which storage tier are you targeting for deduplication—for example, archive?

Answer: We are targeting secondary storage applications such as backup data, archive data, and light-duty file serving. We see this expanding over time.

Question: If deduplication takes place post-processing, how much data can be cached prior to deduplication? And is this memory- or disk-based?

Answer: The data all has to fit in the flexible volume. And then you can dedupe.

Question: What are the average disk space savings?

Answer: It depends on the dataset. Customers are reporting anywhere from 10% to 90%.

Question: When dedupe is running, is its priority changeable (via an option, etc.)?

Answer: No.

Question: Is the dedupe run automatically or on demand?

Answer: A-SIS deduplication is a scheduled post-processing activity. Our backup product, SnapVault, is a backup solution that both transfers and stores deduplicated backup data and is done on the fly.

Question: How much does deduplication help if you're using compressed or encrypted online disk backup?

Answer: When data is already encrypted, deduplication will not provide further space reduction.

Question: How do you calculate the fingerprints?

Answer: It's a hash of the offset and the data contained in the data block.

Question: Are there limitations on the maximum number of files in a volume (not max size of files) for it to be deduplicated?

Answer: No max number of files within a volume.

Question: How does this deduplication work with databases and messaging systems? Could I use it to decrease their sizes?

Answer: Yes, you can. Actual results depend upon the type of data.

Question: How big a problem is data duplication?

Answer: Deduplication is a solution to storage costs. Deduplication simply addresses the elimination of redundant data, which results in less storage being required.

Question: Is there a tool to predict savings before implementing?

Answer: Yes. Ask your SE.

Question: Could you please outline the pros and cons of file-level dedupe versus block-level dedupe?

Answer: File-level dedupe will only eliminate redundancy in two completely identical files. Block-level dedupe allows the elimination of blocks within a file.

Question: After dedupe, is a file system's space automatically reclaimed?

Answer: Yes, that assuming blocks aren't locked in Snapshot copies.

Question: Is there any way to do a "dry run" to view the possible storage savings before implementing deduplication?

Answer: Your SE can use the tool I mentioned earlier against the data.

Question: What happens if FlexVol metadata gets corrupted? Is the deduped data still valid?

Answer: Deduped data is still okay. You could rerun the dedupe process to recreate the metadata.

Question: How can you say that dedupe has little impact, since it's a post-process event and has little impact? This is primary data - we don't have little impact times.

Answer: That's why we don't say it's for all primary apps. It will work there, but specific customers' requirements are key to consider.

Question: after a volume is deduped, then doing an NDMP backup to tape using CommVault or Veritas, is the amount of data written to tape the "nondeduped" size? or will the tape backup benefit from the deduping?

Answer: Data written to tape is "nondeduped".

Question: We like to replicate the deduped data to a remote location. I assume that the metadata needs to be sent at the same time. Right? Is there a best practice or white paper regarding the management of deduped data at off-site locations?

Answer: Yes, the metadata gets replicated too. There is a Tech Report coming out soon (end of the month) that will help.

Question: Does deduplication work for NAS, FC, and iSCSI file systems?

Answer: Yes, deduplication works with all protocols.

Question: I got this error when deduplicating a volume:

```
$ rsh filer1 sis on /vol/volume1 Volume or maxfiles exceeded max allowed for SIS: /vol/volume1
```

What does that mean?

Answer: It means "Volume or maxfiles exceeded max allowed"! ;^) Seriously though, it means that the maximum volume size A-SIS supports was/is exceeded.

Question: How did you arrive at the 95% reduction number? Can you supply a real-world example?

Answer: Backup data deduplicates the most. An example of 95% reduction would be a policy of daily incrementals and weekly fulls retained with a 1% daily change rate retained for 6 months. Or daily fulls retained for 30 days.

Question: EMC touts that you can get 300 to 1 better bandwidth utilization when implementing a solution that takes into account dedupe, such as Avamar. What is a more realistic number? And does dedupe only play a role in file systems and e-mail?

Answer: Those "big" numbers refer to repetitive backups of highly redundant data. (We can achieve those too, but they are really corner cases.) What we find customers asking more often is "how much space can we save per volume?" This is a more realistic way to describe space saving. If you reduce a 1TB volume by 50%, the value of deduplication is evident. Typical A-SIS volume space reduction is 10% to 90%.

Question: Do the freed up (duplicate) blocks go into the Snapshot copy, or are they immediately available?

Answer: Freed blocks are immediately available.

Question: As a post-process (dedupe), do I presume that it requires a quiescent file system?

Answer: No, the file system is not required to be quiescent.

Question: Per the answer about compression rates—doesn't 10% to 90% seem like a broad range? What will my clients see on average? And what does it depend on?

Answer: It depends on the data set. Some data dedupes substantially better than others.

5. NetApp General

Question: What is NearStore? And is it simply a license key for the filer, or is there something like another server needed for it (like Operations Manager)?

Answer: NearStore is a license key for a FAS system and does not require additional hardware.

Question: Is there a white paper that describes deduplication, how it is implemented, and the costs?

Answer: There is a Tech Report that will be posted in a week or so on the Tech Library. Look for TR-3505. There is also lots of information on netapp.com, under NearStore on FAS.

Question: What OS do you use on the 3020, Linux?

Answer: The NetApp OS for FAS systems is Data ONTAP.

Question: Where can one go to see actual demonstrations of this technology?

Answer: Please work with your local NetApp office or reseller for a demonstration.

Question: When NearStore is added to a FAS device, what functionality does that bring, other than it is required for A-SIS? Does that mean you can dedupe tier 1/production data?

Answer: The “NearStore option” also increases the number of SnapMirror/SnapVault replication streams per platform. Typically, a NearStore system is a target system, and therefore more streams are required for replication operations.

Question: Can you please point us to some online presentations or white papers?

Answer: <http://www.netapp.com/products/storage-systems/near-line-storage/asis-dedup.html>

© 2007 Network Appliance, Inc. All rights reserved. Specifications subject to change without notice. NetApp, the Network Appliance logo, Data ONTAP, FlexVol, NearStore, SnapLock, SnapMirror, SnapVault, and WAFL are registered trademarks and Network Appliance and Snapshot are trademarks of Network Appliance, Inc. in the U.S. and other countries. Linux is a registered trademark of Linus Torvalds. Windows is a registered trademark of Microsoft Corporation. Symantec is a registered trademark and Backup Exec, Enterprise Vault, NetBackup, and Veritas are trademarks of Symantec Corporation. UNIX is a registered trademark of The Open Group. All other brands or products are trademarks or registered trademarks of their respective holders and should be treated as such.