Deduplication Round Table Discussion

Excerpt… on effect of deduplication to reduce mainframe FICON attach backstore requirements

INNOVATION Data Processing
Thomas J Meehan

Session 9162

March 1 2011

Deduplication is a natural complement to z/OS mainframe and Multi-Platform z/OS Distributed Data protection…

FICON / ESCON

VTL/Source Recognition/ Deduplication/ Compression

GigE

Tape Data Store

Meta-data store

- Media generating headlines vs. analysts setting expectations
- z/OS backup already offers sophisticated data reduction…
  - Incremental Backup – Merge Backup – Last Tape Stacking
- Deduplication works… and dedupe appliances do more…

SHARE in Seattle

**Deduplication is a natural complement to z/OS mainframe…
and Distributed Enterprise Data Disaster Recovery.**

Mainframe backup deduplication generally implies employing a Virtual Tape Library (VTL).

There are definitely pros and cons relating to the choice of employing VTL technology as a backup repository, but I'm going to limit this discussion to deduplication and data reduction as it applies to those mainframe sites who have already made the choice to employ a VTL.

Articles are appearing in the media questioning the value of deduplication in conjunction with mainframe backup. The reason is that, while analysts are trying to set appropriate customer expectations, editors are all too often spinning words to create headlines.

z/OS backup solutions already offer sophisticated data reduction techniques that aim to only copy new data, i.e. copy only used space, only select changing files (incremental backup) and virtualize creation of a new full volume image, reading only the recently changed files from disk while copying unchanging data forward from prior incrementals (merge backup). Deduplication alone, consequently, may not garner the same dramatic reductions, when comparing mainframe backup reduction ratios, seen when applying deduplication to rote brute force full volume backup solutions.

z/OS customer experience and validation testing confirm deduplication works. Quoting one deduplication hardware vendor expert after testing with FDR/ABR; "The results were perfect. Data that we recognize (exists on previous backups) is not re-written."

Additionally and more important; deduplication appliances not only recognize data that already exists and don't record it again, they also compress the data they do record.

2

FDR, FDRABR, FDRSOS, UPSTREAM and RESERVOIR… all benefit.

FICON / ESCON — VTL/Source Recognition/ Deduplication/ Compression — GigE — Tape Data Store / Meta-data store

• INNOVATION is working with EMC/Bus-Tech/Data Domain, IBM & Luminex to insure you the best results..
• Customers report … "Adding a deduplication backstore to a mainframe VTL Solution is a no brainer".

SHARE in Seattle ······

Results will vary, but the addition of a deduplication backstore to a mainframe Virtual Tape Solution is a "no brainer".

INNOVATION in collaboration is sharing its data stream layouts, with Bus-Tech/EMC, Data Domain, IBM and Luminex, and is working to make all mainframe and open systems deduplication hardware solutions look their very best, especially when used in conjunction with INNOVATION software.

Customers regularly report data reduction averages of 20 to 1 and higher writing z/OS mainframe backup to a deduplication backstore.

Testing by INNOVATION, and by the major deduplication hardware vendors confirm; there is significant benefit to be gained using deduplication appliances as backstore for all INNOVATION data protection products; FDR, FDRABR, FDRSOS, FDR/UPSTREAM and RESERVOIR.

Efficiency Testing…
Shadow z/OS and Open Systems Backups

50 z/OS volumes w/130GB (3390-1, 3, 9 and 27)
- z/OS IPL packs, System & ISV executables, JCL, Source, Listings, sysdump, FTP server, application work, DB2 systems & data bases, BookManager and product softcopy manuals.
- ABR daily incr (7% changing) & weekly fulls.

50 heterogeneous servers w/2 TB of data
- Windows Servers, SOLARIS on X-86, SuSE, & RedHat Linux on x-86 and on System z.
- UPSTREAM daily incr (4% changing) & a weekly full merge backup

SHARE in Seattle ·······

Representative environments…

INNOVATION, to determine the benefit of deduplication for z/OS backup, began shadowing the FDR/ABR production backup of our z/OS system & the UPSTREAM backup of our distributed servers to a (FICON attached) Data Domain backstore while recording the resulting global compression (deduplication) and local compression (data compression) data reduction statistics.

The INNOVATION z/OS data center doesn't compare in size to an enterprise installation nor does it have the ranging array of applications and customer data. However the fifty (3390-1, -3, -9 and -27) volumes that contain our z/OS system, application library and other associated volumes are representative of the core of any typical z/OS site large or small. We protect this representative z/OS core environment with daily ABR incremental backup (selecting about 7% as changing) and weekly full volume backups.

INNOVATION also maintains a diverse open systems development and support environment with over 50 heterogeneous enterprise servers utilizing over two terabytes of data. There are multiple levels of Microsoft Windows Server, SUN SOLARIS on X-86, SuSE and RedHat Linux on Intel and on System z. Again though not comparable in size with large enterprise sites it is diverse enough to be representative of the core of any enterprise server site. We protect this environment with daily UPSTREAM incremental backup (selecting about 4% as changing) and weekly full merge backups.
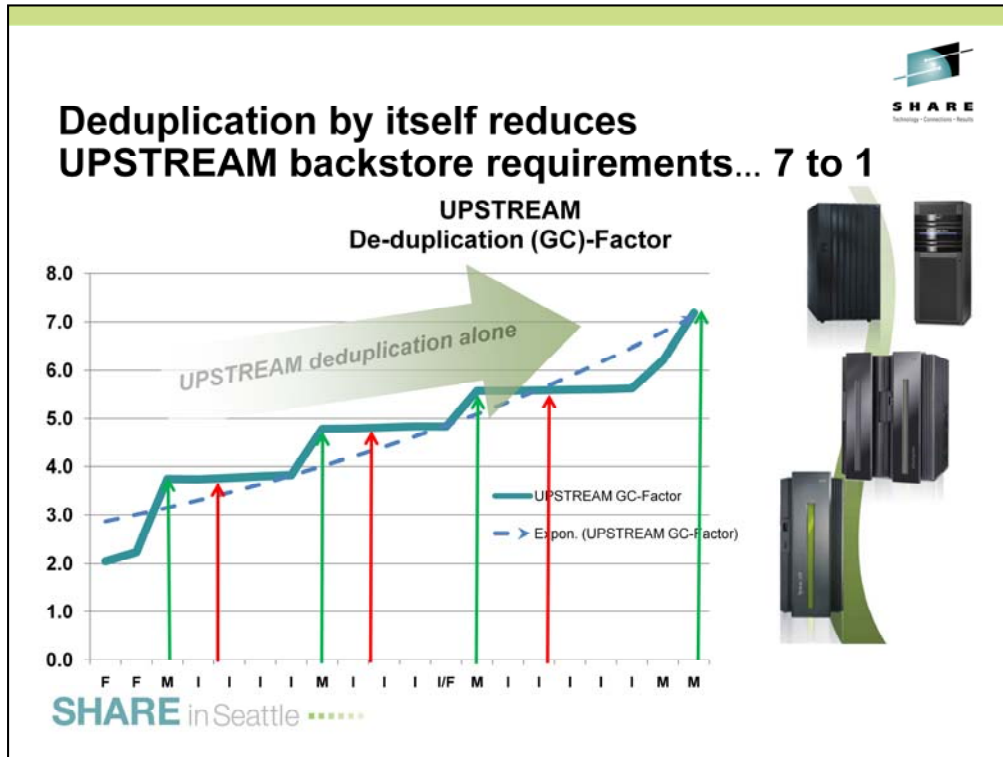
Deduplication technology by itself reduces the backstore…

The chart shows how the deduplication factor increases as full volume and merge backups containing mostly data that has previously been backed up is written to the backstore.

The root of the media articles hyping up speculations over the value of deduplication for mainframe backup is that , "Incremental Backup" solutions like ABR & UPSTREAM only backup new files and those they recognize as changing. Consequently the headlines ask; "what help can dedupe be, if the dedupe appliance is only given new data?"

The answer is simple. Typically after creation current applications and operating systems don't copy and rewrite master files. They only change a portion of the data in a disk file at any one time.

ABR & UPSTREAM incremental backups though they are only selecting new and changing files backup an entire file, no matter how much or how little of the data is actually new, with plenty of data in these files that didn't changes.

Consequently when a deduplication appliance can recognize chunks of data within a file, that it has previously recorded, deduplication can still reduce an incremental backup, and only need write out new "chunks".

We see this on the chart , as the deduplication factor plateaus ,during the UPSTREAM incremental backups when new files and whole files the only contain some new data are written to the backstore.
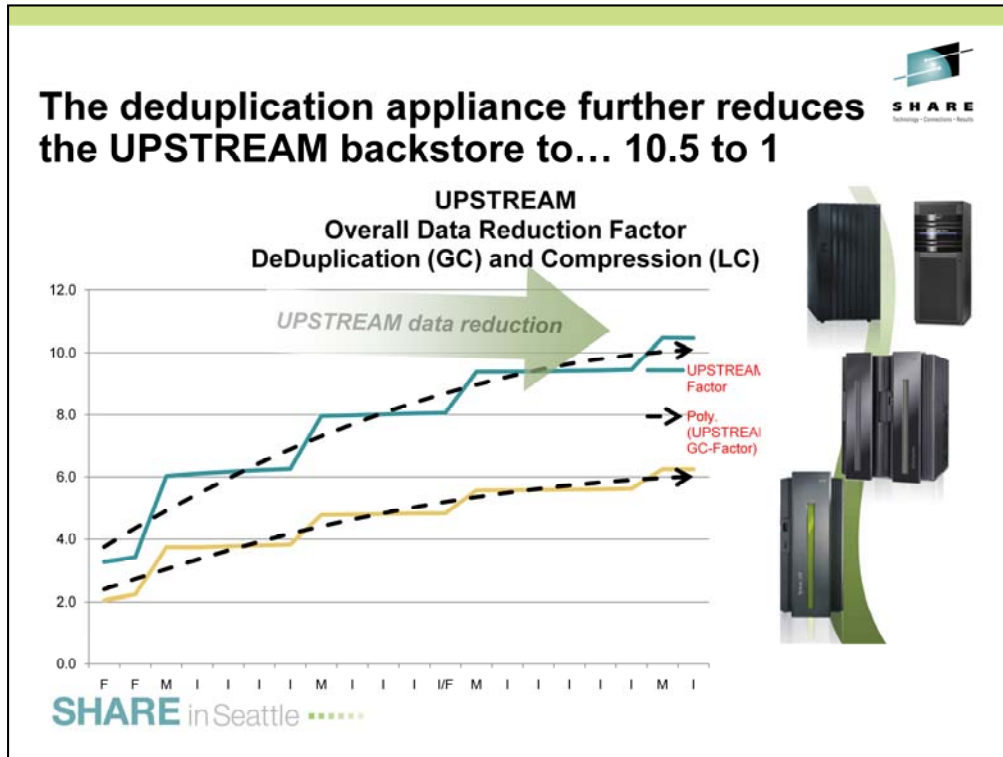
Deduplication alone reduces
FDR backstore requirements… 8 to 1

FDR
De-duplication (GC)-Factor
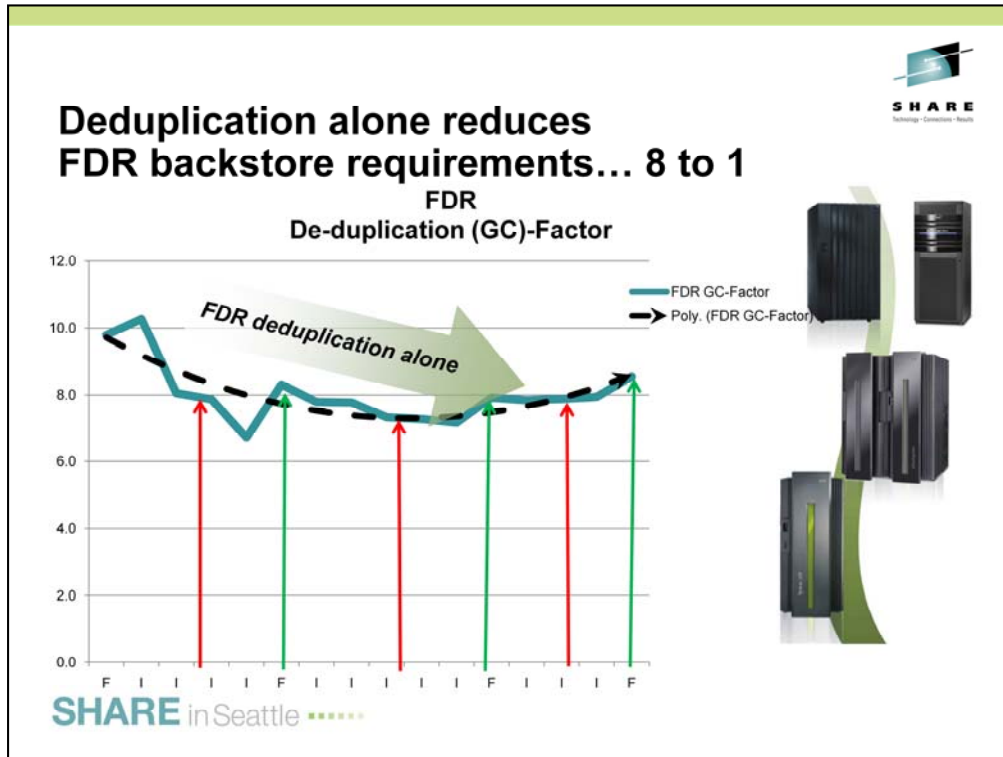
Deduplication technology by itself reduces the FDR backstore…

This chart again shows the deduplication factor stepping up as subsequent ABR full volume backups, containing mostly data that has previously been backed up, are written to the backstore.

Interestingly this chart also shows the deduplication factor is decreasing slightly during the ABR incremental backups. This can be expected as the ABR incremental backups of new and changing files contain proportionally twice as much new data, i.e. 8% of a full backup ABR z/OS backup as compared to only 4% of a full UPSTREAM enterprise data backup.

Still the deduplication process is significantly reducing the size of the backup as it recognizes chucks of existing data within the ABR incremental backups.
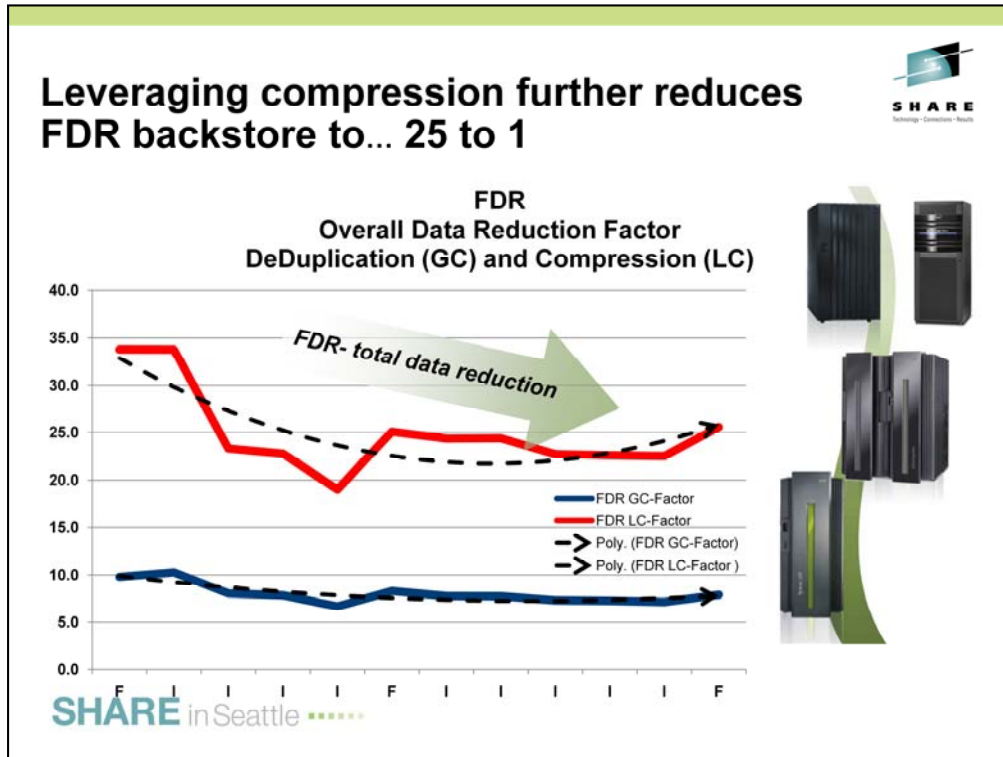
Sophisticated deduplication appliances leverage their position in the data path; i.e. they write the data to the backstore, to further reduce the storage requirement . The do this, after they account for all the chunks of duplicate data they recognize, by compressing the data that is left as they write it.

This chart shows the result with the INNOVATION UPSTREAM backups is that the overall data reduction almost doubles.

Interestingly thought the data is coming from fifty different physical servers running many different operating systems and multiple applications the data is still compressible.

Leveraging compression further reduces FDR backstore to... 25 to 1

FDR
Overall Data Reduction Factor
DeDuplication (GC) and Compression (LC)

The result with our ABR z/OS backups is even more impressive as compression better than triples the already high deduplication reduction factor for a significant overall data reduction in the backstore.

This shows that there must be a significant amount of compressible data on the volumes that make up a core z/OS system.

This significantly higher reduction ratio however is the result of leverage, or multiplier effect, that comes with combining two complementary data reductions techniques; deduplication and compression.

Simply compressing the raw data might have resulted in a 3 to 1 reduction; similar to the hardware reduction we typically see on tape today and possibly will see on disk in the near future. However by first reducing the data, with deduplication, the benefit of the compression is leveraged resulting in a substantially higher overall savings.
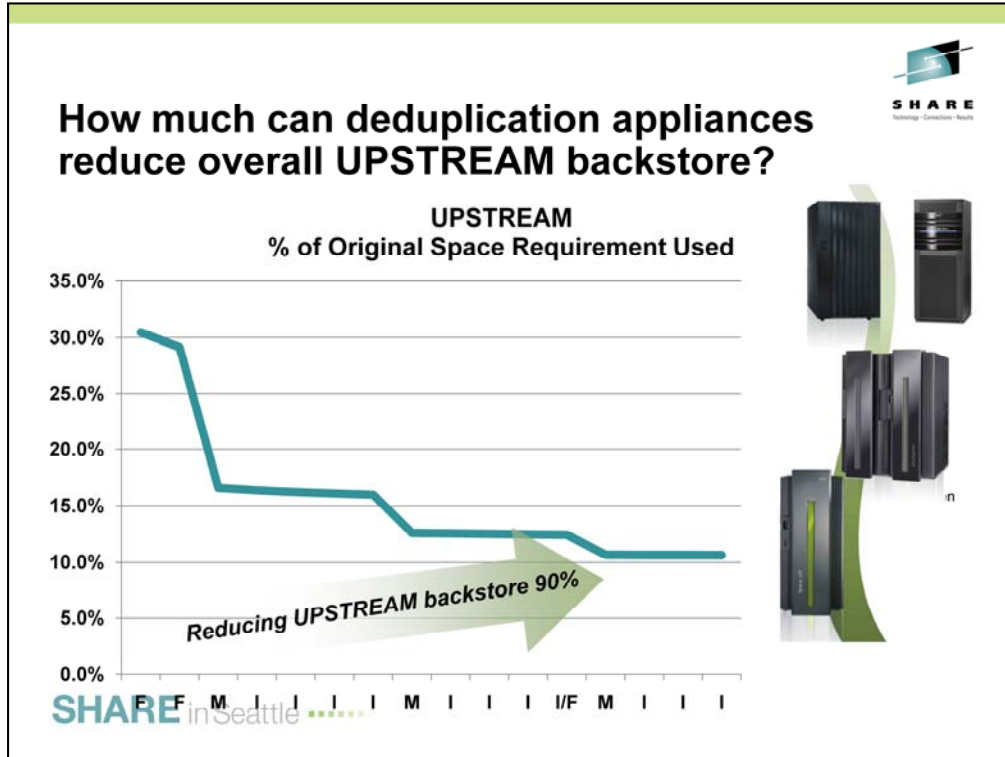
Results will vary, but the addition of a deduplication backstore to a mainframe Virtual Tape Solution is a "no brainer".

Just as its important to recognize the multiplier effect that comes of using compression to leverage the benefits of deduplication, it important to understand the relative value of a particular data reduction ratio number.

Its obvious that a 10 to 1 reduction is a 90% savings. Less obvious is that anything greater than 10 to 1 can result in additional saving of only a few percentage points more and that as the deduplication ratio continues to increase, the incremental increase in saving diminishes.

The 20 to 1 reduction we customers are reporting is already in the rarified atmosphere of a 95% savings in the storage requirement. Yet a five or 6 point drop in reduction, for say a 15 to 1 value, is only about a 1% drop in the overall saving and still very much in that same rarified atmosphere.

The overall reduction in UPSTREAM backstore requirement from our testing, you might only need about 10% of the storage you now use to store the same UPSTREAM backup data in its raw form.
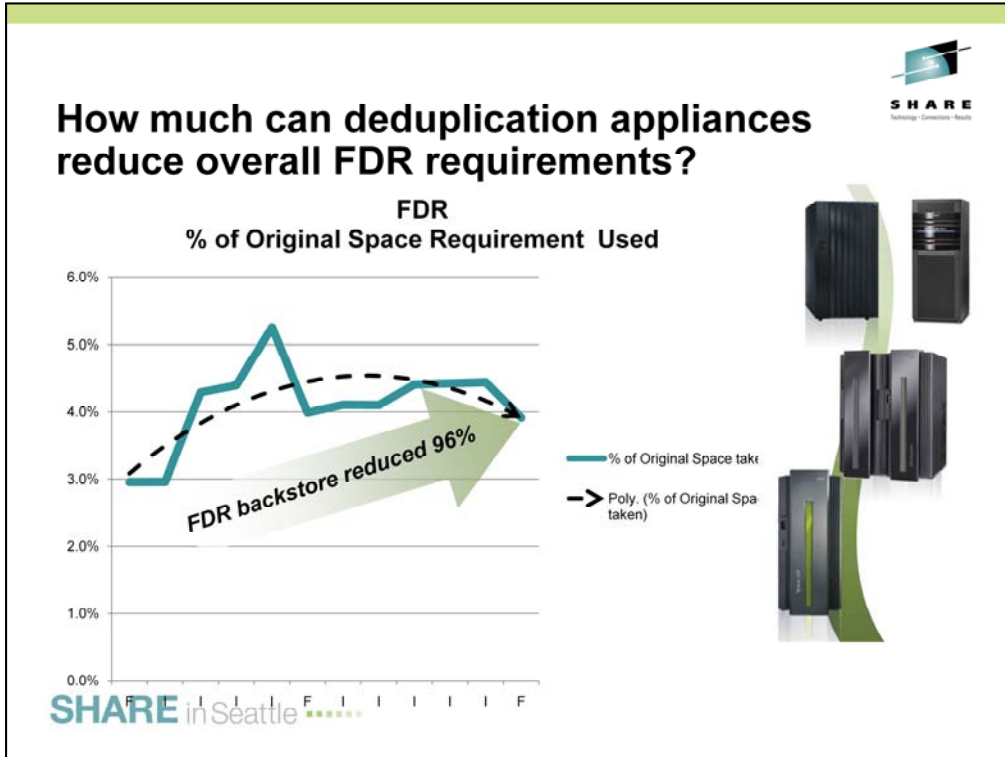
Our testing shows FDR/ABR customers may need as little as 4% or 5% of the amount of storage they now use to store the same amount of ABR backup data in its raw form.

What does 20 to 1 mean in practical terms?

Normal Backup Storage requirement

Resulting 20 to 1 Backup Storage Reduction

Results will vary but the addition of a deduplication backstore to a mainframe Virtual Tape Solution is for certain a "no brainer"…

SHARE in Seattle

Customers are reporting a 20 to 1 reduction…
You can see in this example that a 20 to 1 reduction is significant…

Your actual savings would depend on what kind of backup procedures you use today, incremental and full versus only full, and how long you are retaining your backup copies. How many full volume back copies do you keep? How many daily incremental copies?

The more full volume backups you take and the longer amount of time you keep any kind of backup the greater the savings deduplication can provide.

There are also other savings that relate to the use of virtual tape you can explore. Just a few ideas to consider… might be

Are you constrained by the number of physical tape drives you have? Do you have to run many of your backup jobs serially? Consider the advantage of running all you backup jobs/tasks in parallel to virtual tape drives.

Are you using merge backup? Do you have to copy data from one tape to another? Consider the advantage of stacking incremental backups on a single virtual tape and not having to physically copy the data forward, while having no concern about serial tape searches during a recovery.

Results will vary but the addition of a deduplication backstore to a mainframe Virtual Tape Solution is certainly a "no brainer".