EXAGRID® | Stress-free backup storage

Disk-Based Backup

The compute impact of data deduplication on disk-based backup

Data backup continues to become increasingly critical, paralleling the rise in frequency of data corruption events, natural disasters, systems failures, and other data-loss catastrophes. When data needs to be restored, organizations may need to roll back weeks or even months to find a readable copy based on when the corruption or deletion occurred. In addition, industry and government regulations (e.g., Sarbanes-Oxley, HIPAA, GLBA, etc.) are becoming more stringent. All of these imperatives combined are driving the need to keep many weeks, months, and years of backup retention.

Using straight disk for backup storage becomes cost prohibitive very quickly. For example, keeping 12 weeklies as well as monthlies for 3 years equals 45 backup copies. Due to backup retention requirements, data deduplication is necessary. Not only will deduplication greatly reduce backup storage, it will also reduce WAN replication to the disaster recovery (DR) site because only the changes from backup to backup are stored and replicated.

Data deduplication compares the amount of storage required with data deduplication enabled to the amount of storage required without data deduplication, and the result is the "deduplication ratio." If 20 copies of a 50TB backup are kept without data deduplication, 1PB of storage is required. The longer the retention period, the greater the storage required and, therefore, the greater the savings if data deduplication is used. For example, at 20 weeks of retention, a solution with data deduplication uses only 50TB to store 20 copies of a 50TB backup. The deduplication ratio is calculated as 1PB (without data deduplication) divided by 50TB (with data deduplication), which equals 20 and results in a deduplication ratio of 20:1.

When implementing data deduplication, there are some inherent challenges to take into account. The first is that although data deduplication reduces storage and WAN bandwidth, not all data deduplication is created equal. Each vendor has their own algorithmic approach and with the exact same data mix, backup size, and retention period, different solutions will achieve ratios of 2:1, 4:1, 6:1, 8:1, 10:1, 12:1, 20:1, and higher. Depending on the deduplication algorithm used, the amount of storage and bandwidth can vary greatly. A strong deduplication solution can achieve a deduplication ratio of 10:1 to as high as 50:1, with an average of 20:1 depending on the data mix and retention period. Typically, the backup application's deduplication results in a lower ratio and uses more disk and bandwidth than dedicated target-side appliances, since appliances have dedicated high-speed compute and can use more aggressive algorithms to achieve higher deduplication ratios.

Another challenge is that most vendors simply added data deduplication as a feature to a scale-up storage platform or to a backup application. This approach of adding deduplication as a feature versus creating an architecture specifically for data deduplication not only slows down backups, restores, and VM boots, but as data grows, the backup window expands; data deduplication is highly compute-intensive and when this processing is performed in the data stream of the backups, backup speed suffers.

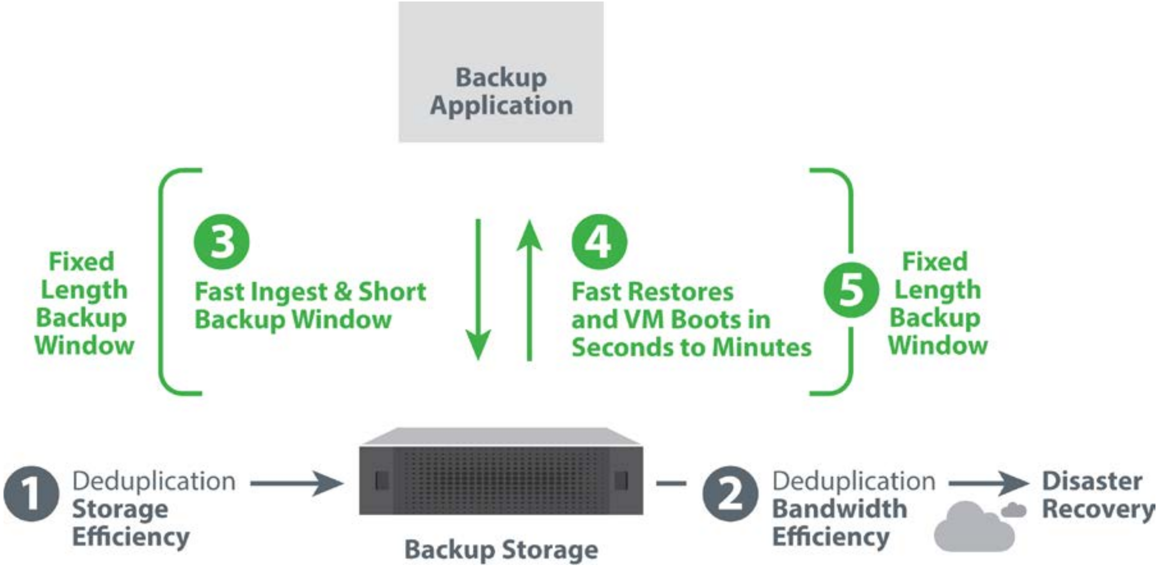
The impact of implementing an inferior solution may include:

- lower deduplication ratios, resulting in 1) additional storage and cost, and 2) increased bandwidth and cost due to replicating more data
- slow backups, resulting in long backup windows
- slow restores, offsite tape copies, and VM boots, impacting users and productivity
- a backup window that expands as data grows, requiring the regular upgrade to a bigger and faster front-end controller or media servers (known as "forklift upgrades," which are costly and disruptive)

When data is deduplicated during the backup window, the backups are slow resulting in a longer backup window. In addition, only deduplicated data is stored, so in order to perform restores and VM boots, data needs to go through a time-consuming reassembly or "rehydration" process each time. ExaGrid looked at these challenges differently than other vendors. They are not merely storage and replication challenges because data deduplication introduces a compute challenge that other approaches do not address.

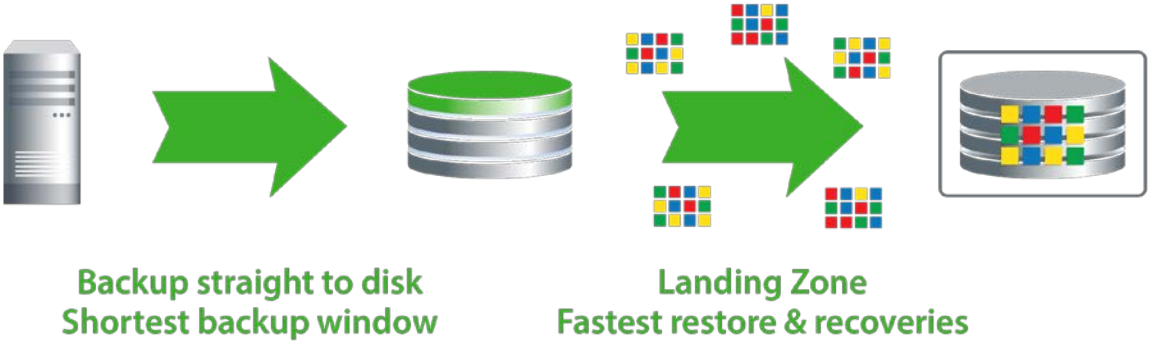
ExaGrid's approach meets all 5 challenges by delivering the:

1. best deduplication ratio for least amount of required storage
2. best deduplication ratio for least amount of bandwidth used
3. fastest backups for the shortest backup window
4. fastest restores, offsite tape copies, and VM boots to improve user uptime
5. backup window that remains fixed in length even as data grows

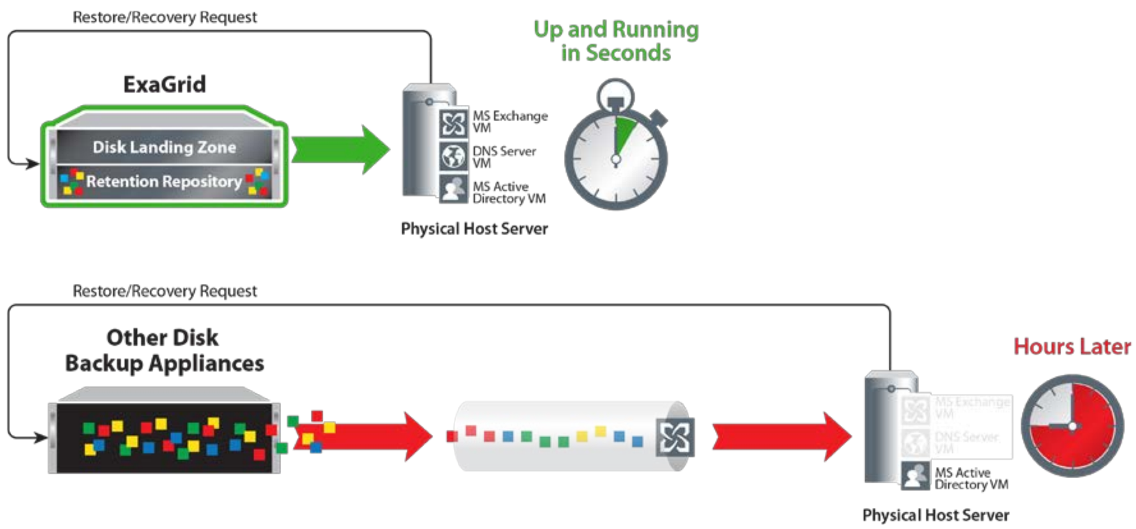

Backup Application
Fixed Length Backup Window
3
Fast Ingest & Short Backup Window
4
Fast Restores and VM Boots in Seconds to Minutes
5
Fixed Length Backup Window
1
Deduplication Storage Efficiency
Backup Storage
2
Deduplication Bandwidth Efficiency
Disaster Recovery

ExaGrid deploys zone-level deduplication that compares zone stamps and then identifies only the bytes that have changed. ExaGrid then stores and replicates only the changed bytes. Depending on data mix and retention periods, ExaGrid achieves or exceeds the industry's best average deduplication ratio of 20:1 with ratios of 10:1 to as much as 100s:1. ExaGrid requires the least amount of storage for deduplicated data and the least bandwidth to replicate data offsite for disaster recovery.

In order to avoid the compute-intensive deduplication “tax” during the backup window, ExaGrid writes straight to disk for the fastest possible backups, resulting in ingest rates that are the highest in the industry. In contrast, appliances that perform deduplication inline are slow because they deduplicate data through a single controller; therefore, they need to deploy software on backup media servers and database servers to offload some of the deduplication. Even with doing some of the processing on the production servers, these solutions still cannot come close to ExaGrid’s performance. In larger installations, ExaGrid is at least three times the speed of its closest competitor.



Writing direct to disk allows ExaGrid to keep the most recent backups in both their deduplicated and original non-deduplicated form. The most recent non-deduplicated backups are kept in a “landing zone,” quickly accessible for the fastest restores, offsite tape copies, and VM boots. ExaGrid’s approach avoids the time-consuming data rehydration process and is five to ten times faster than solutions that only store deduplicated data. When booting a VM, backup software deduplication can take a few hours up to a full day, inline scale-up appliances can take hours, but ExaGrid takes just seconds to minutes because the most recent backups are readily available in their non-deduplicated form in the landing zone. Behind that, ExaGrid keeps weeks, months, and years of deduplicated data for long-term retention storage.



It stands to reason that as the volume of backup data grows, so too does the amount of data to be deduplicated, and it follows that if additional compute is not added, the backup window will continue to grow indefinitely. However, ExaGrid is the only solution that uses a scale-out storage architecture, so instead of adding just storage behind a fixed resource controller, ExaGrid adds full appliances with all required resources – processor, memory, network ports, and storage – to a scale-out GRID. Therefore, when data doubles, triples, quadruples, ExaGrid doubles, triples, quadruples all necessary resources, not just storage capacity.

ExaGrid is the only solution that meets all 5 backup storage challenges with the:

1. highest storage efficiency
2. lowest bandwidth usage
3. fastest backups
4. fastest restores, offsite tape copies, and VM boots
5. fixed-length backup window despite data growth

United States: 2000 West Park Drive | Westborough, MA 01581 | (800) 868-6985
United Kingdom: 200 Brook Drive | Green Park, Reading, Berkshire RG2 6UB | +44 (0) 1189 497 051
Singapore: 1 Raffles Place, #20-61 | One Raffles Place Tower 2 | 048616 | +65 6285 0302

ExaGrid reserves the right to change specifications or other product information without notice. ExaGrid and the ExaGrid logo are trademarks of ExaGrid Systems, Inc. All other trademarks are the property of their respective holders.