Don't be duped by dedupe - Modern Data Deduplication with Arcserve UDP

by Christophe Bertrand, VP of Product Marketing

Too much data, not enough time, not enough storage space, and not enough budget, sound familiar? Since the first mainframes, efforts have been made to optimize storage capacity requirements and data protection processes.

In the open systems world, the issues mentioned above are the same as years ago when the first data deduplication technology became mainstream. Backups are failing, taking too much space, and costing way too much.

Today, data volumes are growing exponentially and organizations of every size are struggling to manage what has become a very expensive problem. Cheaper storage helps, but is not an operationally efficient solution for many workloads. Data needs to be shrunk to more manageable levels. Too much data causes real problems for companies:

- Companies overprovision their backup infrastructure to anticipate rapid future growth. This is expensive.
- Legacy systems can't cope and backups take too long or are incomplete.
- Companies miss recovery point objectives and recovery time targets.
- Backups overload infrastructure and network bandwidth.
- Companies cannot embrace new technologies, such as cloud backup, because there is too much data to transfer over wide area networks.

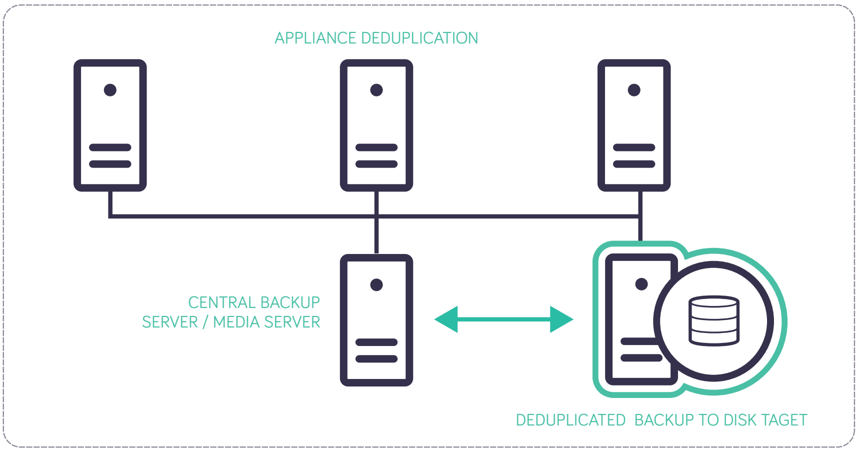
Today, new technology advances are needed to combat the unstoppable and exponential growth of virtual machines and data. In this paper, we will cover Arcserve's approach to data deduplication.

Target Deduplication... sooo yesterday

In recent years organizations realized they could not make their backup windows with traditional backup architectures. Besides backup performance, they struggled with the essential cost of storing large amounts of backup data. Storage costs exploded as back-up schemas imploded. Deduplication appliances became the preferred solution to address the issue. The process involved taking backup data, optimizing it through deduplication processes and storing it on disk. Let's "compress" your backup volumes and save you money. Target deduplication has worked very well and is still in use today in many environments.

Target deduplication is attractive because it does not require the user to drastically change their backup software configurations or policies, rather only to change the destination of the backup streams.

Target deduplication happens either on the fly, or as a post process (write it all on disk, then optimize after the fact to make backups go faster.)


APPLIANCE DEDUPLICATION
CENTRAL BACKUP SERVER / MEDIA SERVER
DEDUPLICATED BACKUP TO DISK TAGET

Target-based deduplication[1]

In one of its educational sessions from 2008, SNIA provides a view of target deduplication scenarios. In the case of target deduplication, one can see that the backup software acts as the data mover and sends all of the non-deduplicated backup data streams to target disk or VTL appliance.

Whether the processing happens in memory on the fly or post backup, why would you pay for another layer of technology? This is an approach and a technology of the past. A better way is available today.

Distributed deduplication… sooo expensive

Certain vendors introduce additional costs by essentially charging for each distributed deduplication agent on each system that needs protecting. While there might be some benefits or "boost" from a disaster recovery standpoint, the multiplication of expensive systems, software licenses and associated bandwidth requirements only add to the tab. In addition, while tools may exist, this type of architecture can be onerous from a management perspective.

Source-side deduplication… sooo Next Generation

Why deduplicate data after the fact if you could only backup the new and unique data you need at the source? As long as you do not impact the client, you can save yourself all of that bandwidth you use to send backup streams to the target.

Why not share all of the deduplicated intelligence across all of your clients? That is global source-side duplication, and it is where the industry needs to go. Global source-side duplication is what many end-users have recognized as the critical backup technology moving forward. The challenge of global source-side duplication is to ensure that the source (client) system is not bogged down by the deduplication software.

[1]Source: http://www.snia.org/sites/default/education/tutorials/2008/spring/data-management/Hamilton-D_Deduplication_Methods_Data_Efficiency.pdf

Arcserve Unified Data Protection

With Arcserve Unified Data Protection (UDP), backup administrators can optimize storage requirements and bandwidth as well as accelerate protection and recovery across multiple sites. In addition, the solution allows for in-place re-hydration of data, for fast granular restore, including from tape.

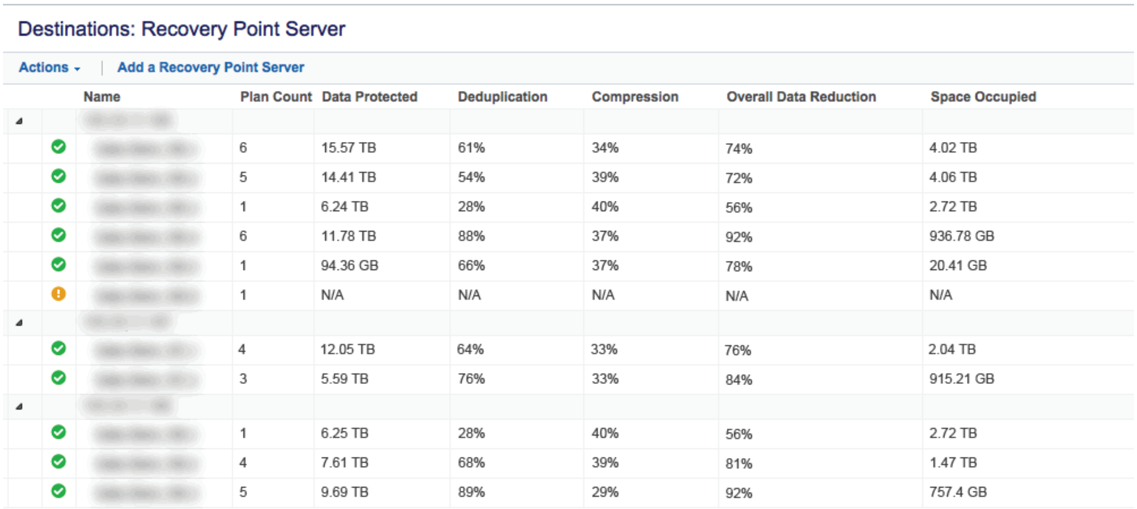
Arcserve UDP performs global source-side deduplication, on the node(s) being protected. It only transmits data from the source node that is unique to the UPD Recovery Point Server which serves as a central data store. After an initial full snapshot backup is taken, all other backups capture just the incremental changes to the data. Since this is a very efficient client-side process:

- The amount of data backed up is reduced
- The amount of data transferred over the network is reduced
- The load on the production server is reduced and the frequency of backups can be increased to provide great RPO and better protect the system, applications and data.

With Arcserve UDP global source-side deduplication, you never "backup" data twice – this includes OSs, hypervisors etc. Arcserve UDP global source-side deduplication dramatically reduces the amount of data transferred during backup cycles. This applies to UDP software and UDP appliances that run the exact same software.

To summarize, every computer, virtual machine or server that gets backed up communicates with the Arcserve UDP Recovery Point Server (RPS) that manages a global database index of files on all machines everywhere. The RPS server does the work of figuring out what needs to be backed up and pulls new data as required while eliminating duplicate copies. Combined with compression technology, RPS can compress data on the disk, by over 92%.

The RPS deduplication database index is stored on high performance SSD, to further improve performance, efficiency and reduce costs, compared to a pure memory-based approach.



Destinations: Recovery Point Server

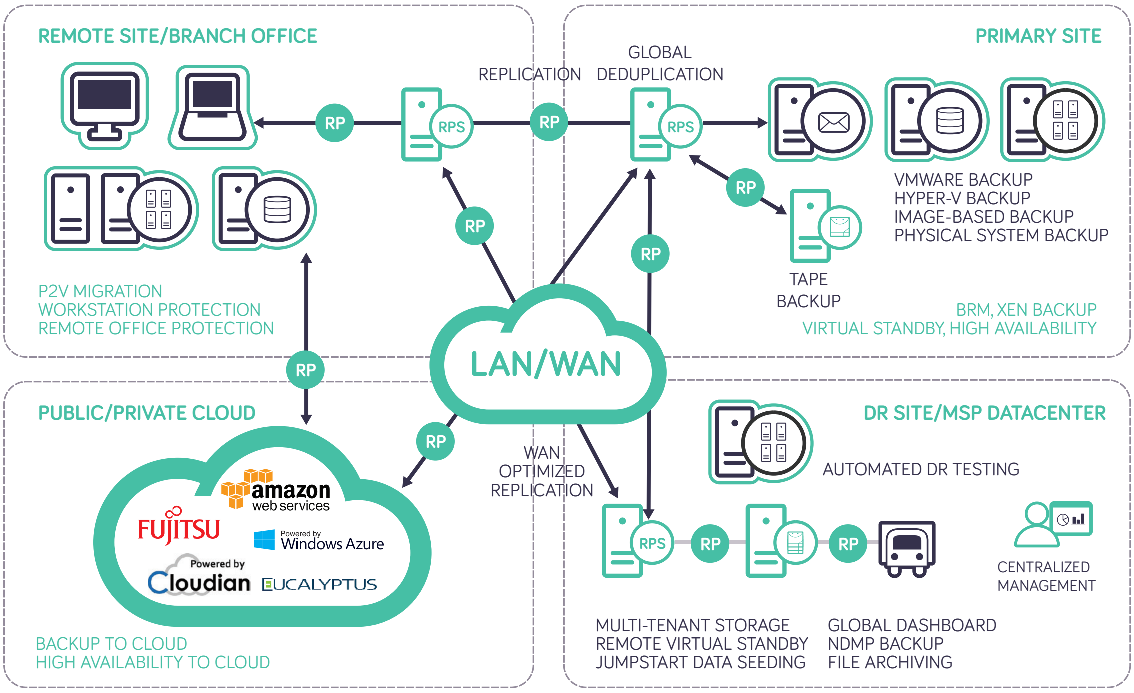
		Name	Plan Count	Data Protected	Deduplication	Compression	Overall Data Reduction	Space Occupied
◢		[redacted]						
	✓	[redacted]	6	15.57 TB	61%	34%	74%	4.02 TB
	✓	[redacted]	5	14.41 TB	54%	39%	72%	4.06 TB
	✓	[redacted]	1	6.24 TB	28%	40%	56%	2.72 TB
	✓	[redacted]	6	11.78 TB	88%	37%	92%	936.78 GB
	✓	[redacted]	1	94.36 GB	66%	37%	78%	20.41 GB
	!	[redacted]	1	N/A	N/A	N/A	N/A	N/A
◢		[redacted]						
	✓	[redacted]	4	12.05 TB	64%	33%	76%	2.04 TB
	✓	[redacted]	3	5.59 TB	76%	33%	84%	915.21 GB
◢		[redacted]						
	✓	[redacted]	1	6.25 TB	28%	40%	56%	2.72 TB
	✓	[redacted]	4	7.61 TB	68%	39%	81%	1.47 TB
	✓	[redacted]	5	9.69 TB	89%	29%	92%	757.4 GB

In this real-life customer screen shot, you can observe the efficiency of the deduplication combined with the additional compression capabilities we apply to the data store, resulting in very high overall data reduction levels.

True Global Deduplication… sooo Awesome

"Global" deduplication across all the clients in the infrastructure is central to limiting the unnecessary storage and transfer of duplicate backup data. Data is deduplicated across nodes, jobs and across sites. The global deduplication in Arcserve UDP goes beyond the limitations of some vendors whose deduplication only applies to the WAN replication cache, and not what is actually stored on disk, thus reducing the overall potential benefits, in terms of bandwidth and storage savings.

For data protection across local and remote sites (private Cloud), the Arcserve UDP global deduplication database index is replicated and distributed so that all source and target data is deduplicated across all RPS servers. Since backup data is globally deduplicated before it is transferred to the target RPS, only changes are sent over the network, which improves performance and reduces bandwidth usage. This entire process is secured with data store-level encryption and per-session passwords.



Arcserve UDP's global deduplication architecture

The “net net”

Arcserve UDP’s deduplication is not just better technology, there are business benefits too:

- **Complete backups faster.** With less data to transmit and store, backups are faster. This is important for situations where the total volume of data threatens to take so long to backup that one backup isn’t finished before the next one is due to start.
- **Improve client performance.** The Arcserve agent has a minimal impact on client performance because the majority of processing is done in the Recovery Point Server. For virtualized environments, agentless VMware and Hyper-V backup reduces the risk of bottlenecks and performance problems at the hardware level. In other words, it ensures that backups don’t stall virtualized servers.
- **Simplify backup infrastructure.** By routing backups through the Recovery Point Server, it’s easier to direct backup data to different places, for example to the cloud via Azure or Amazon Web Services or to local tape or an offsite private cloud.
- **Improve resilience and availability.** Because the deduplication data is stored centrally in the Recovery Point Server, it is easy to protect the backup infrastructure. For example, all the information in the data store can be replicated to the cloud, another server in the same data center or offsite.
- **Reduce bandwidth required for backups.** The Recovery Point Server only pulls new or changed data from a client with high levels of granularity – even down to 4kb chunks. This makes backups extremely efficient.
- **Meet recovery point and recovery time SLAs.** Thanks to Arcserve’s global source-side deduplication it’s easier to meet RPO and RTO targets. With less data to transfer, backups can be more frequent and restores are faster.

See for yourself and try Arcserve UDP for free for 30 days – just go to **arcserve.com/free-backup-software-trials**

Check our UDP deduplication calculator at: **arcserve.com/calculator**

arcserve®
assured recovery™

For more information on Arcserve UDP, **please visit arcserve.com**