SHARE
Technology • Connections • Results

Data Reduction Meets Reality
What to Expect From Data Reduction

Doug Barbian and Martin Murrey
Oracle Corporation

Thursday August 11, 2011
9961: Data Reduction Meets Reality

Introduction

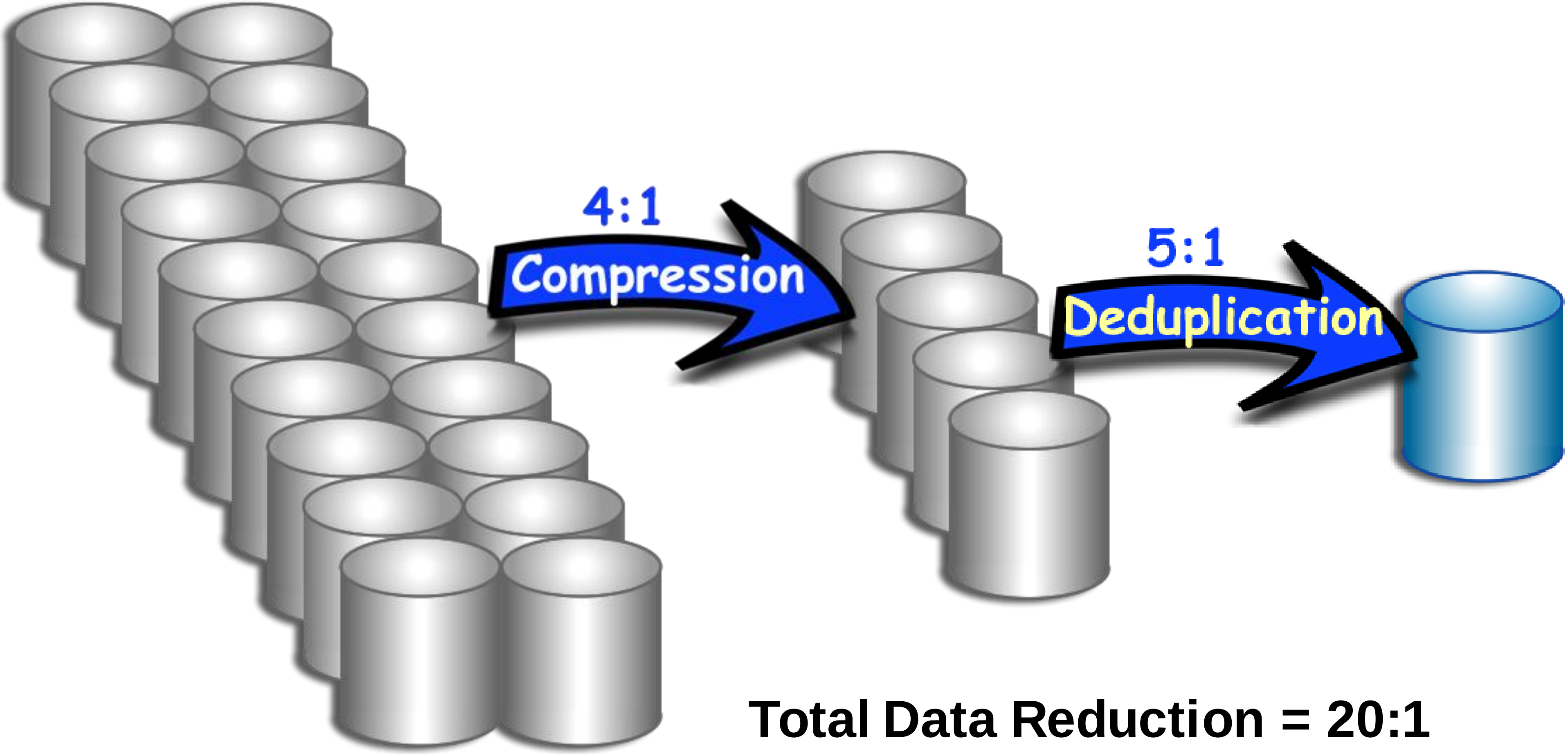
- **"Data deduplication may be the best thing that ever happened to backups!" - Rich Castagna, Storage Media Group**

- **Less is More!**
  - **In this presentation we will show you how data reduction is comprised of compression and deduplication technologies.**
  - **Effective use of these technologies can significantly reduce data storage and transmission costs.**

What is Data Reduction?

- A combination of 2 technologies designed to reduce the amount of space needed to store data.
  - Data Compression
    - Well known technology, widely used, in use for decades. It replaces repeating data patterns with a smaller ‘symbol’.
  - Data Deduplication
    - Newer technology for data reduction. More widely used in open systems. It replaces redundant pieces of data with a reference to the original instance.
- These technologies can be combined and are multiplicative.
  - Data Compression * Data Deduplication = Overall Data Reduction

These Technologies are Multipliers



Total Data Reduction = 20:1

Data Compression

- Data Compression is the process of replacing repeating data strings in a data stream with smaller sized symbols. For example, using the Run Length Encoding (RLE) technique, the string “wwwwwwww” can be replaced with the symbol 8w.
- Data Compression can be done in the host or storage subsystem.
- Types of Data Compression include: LZ, ZLE, GZIP, BZIP, LZSTK. All accomplish the same basic task.
- Typical compression ratios range from 2:1 to 5:1.
- For some customers compression alone can provide enough space savings.

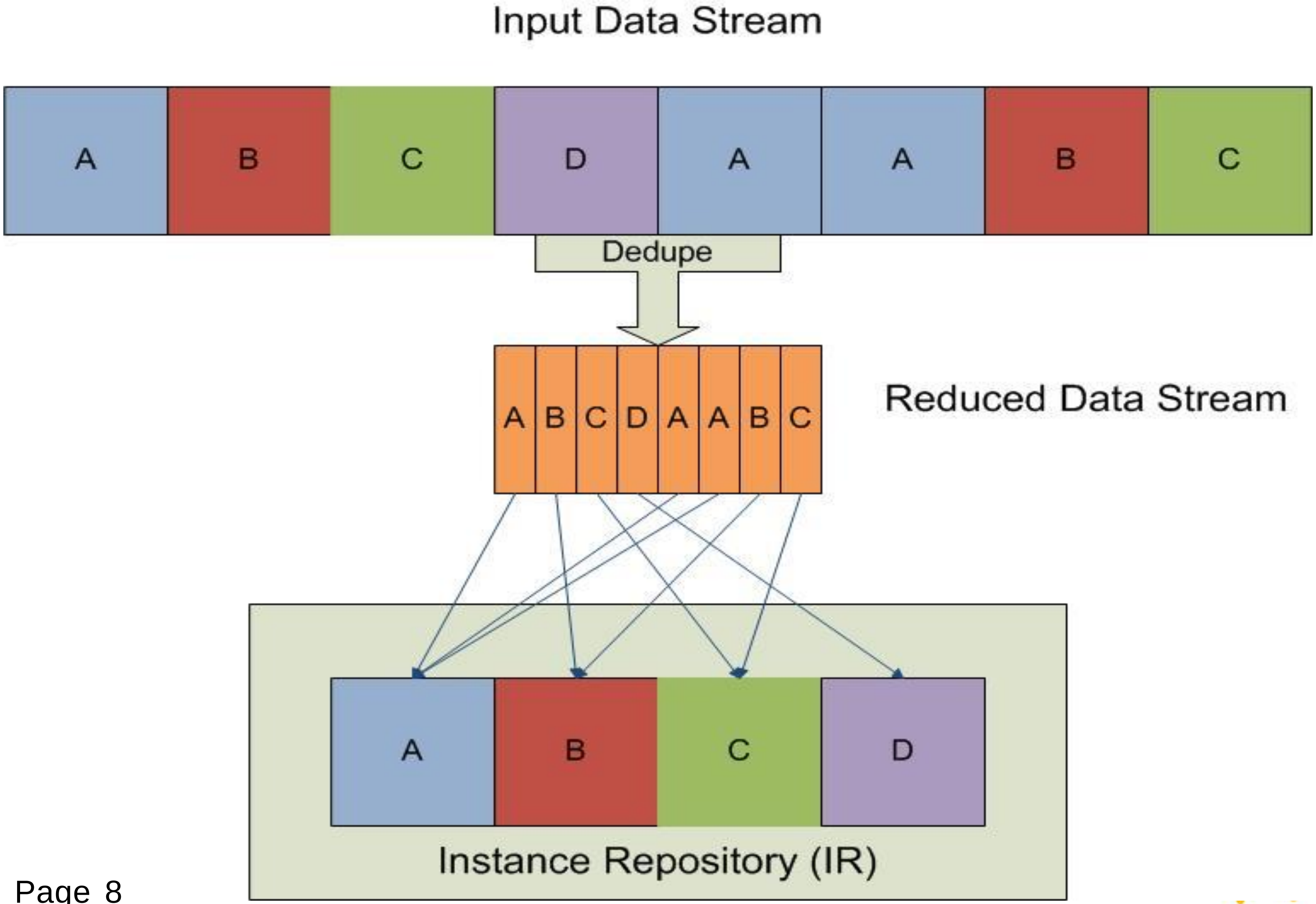
What is Data Deduplication?

- It is the process of eliminating redundant pieces of information and replacing them with a reference to the original instance.
- Modes of Deduplication
  - Inline: Data is deduplicated as it is ingested. Requires more processing power to ingest data but less disk space, because. duplicate data is never written to the storage media.
  - Post: Data is deduplicated after it has landed on the storage system. Requires less processing power but more disk space.
- Levels of Deduplication
  - File: Entire files are deduped. It is faster and more efficient but sensitive to small changes in files.
  - Block: Requires more processing and generates more meta data, but can be more effective because increased granularity.
  - Byte: Strings of bytes are deduped. Very processor intensive.

How Does Data Deduplication Work?

- Take a data stream as input and:
  - Break data into pieces (files, or blocks, or bytes).
  - Create a unique digital signature for each piece.
  - Store the signature in a dictionary and the unique data in an Instance Repository (IR).
  - As new data arrives check to see if its signature has been seen before. If seen before, store a reference, if not store the new instance data in the IR.
  - Deduplication performs a transformation of the data resulting in data that is reduced and distributed.

Data Deduplication



Input Data Stream

A B C D A A B C

Dedupe

A B C D A A B C

Reduced Data Stream

A B C D

Instance Repository (IR)

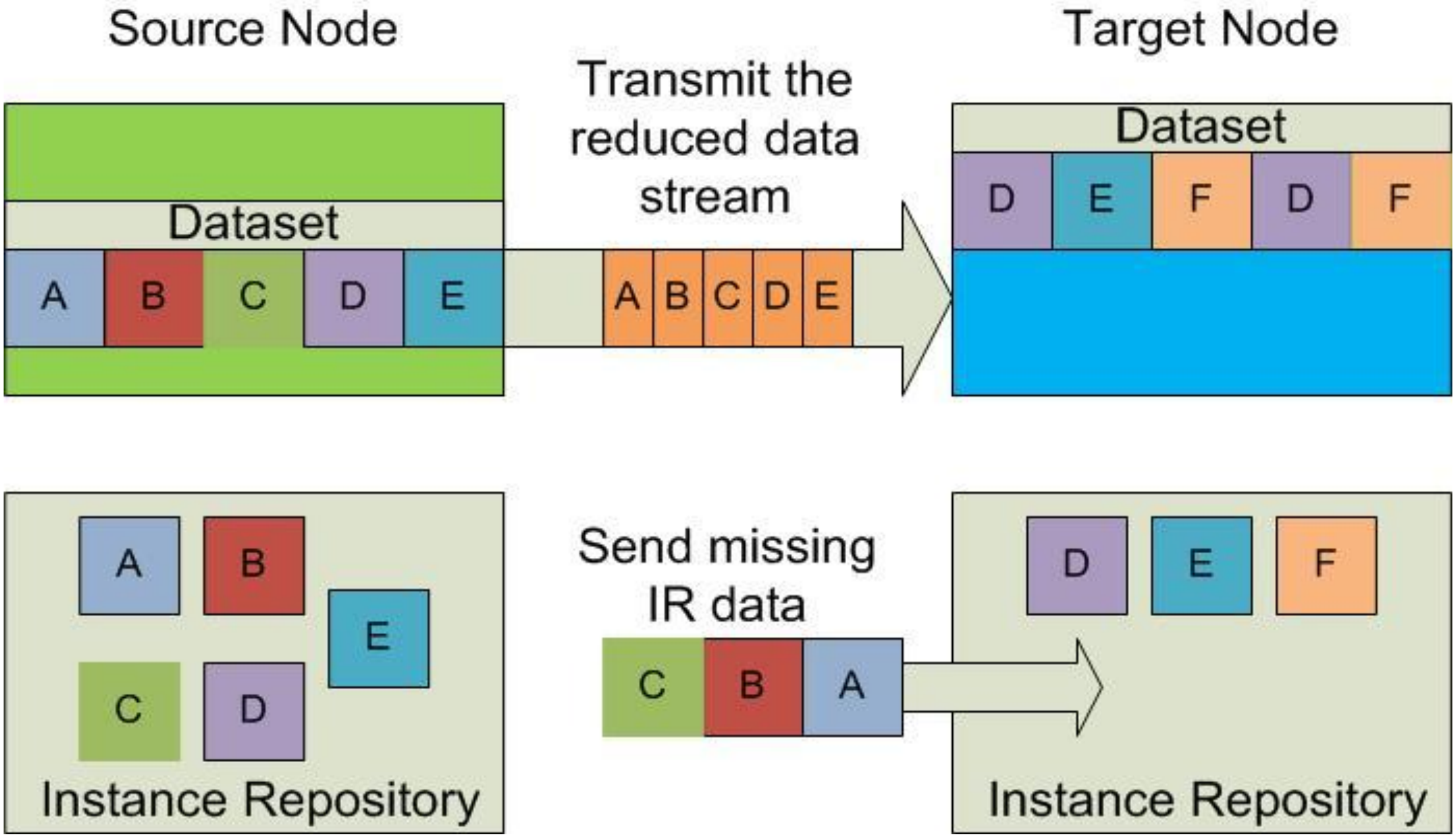
Creating Unique Digital Signatures

- Create a signature for each unique piece of data that will be stored on the system.
- To uniquely identify each 4K block on a 100TB system, you need approx 26.8 billion digital signatures!
- Using a Secure Hash Algorithm (SHA) is a common way to create digital signatures.
- SHAs include but are not limited to: SHA1, SHA256, SHA512. SHA256 generates a 32 byte signature. The probability of a SHA256 hash collision is less than $1*10^{-77}$.

Remote Replication W/R to Deduplication

- **Suppose for instance you have one data center (source) and a remote data center (target) each with copies of all backup data. Both sites have dedupe capabilities. To replicate data from the source to the target you can:**
  1. **Reconstitute all the data at the source, send it to the target and have the target dedupe the data stream. (There must be an easier way.)**
  2. **Send the reduced version of the data stream from source to target. Send only the dedupe data from the source that does not already reside on the target.**
- **Reduced replication can significantly reduce transfer time and processor resources required to replicate the data from one site to another.**

Reduced Replication



Benefits of Deduplication

- It can dramatically reduce physical storage requirements because only unique data is written. This in turn reduces power, cooling and footprint requirements for the data center.
- For data that has a required retention period only one copy of the data needs to be saved not many copies.
- Coupled with compression, deduplication acts as a multiplier that yields even greater data reduction.
- Dedupe solutions work very well with backup data streams and allow you to use your existing backup applications and procedures.

Choose Your Data Wisely

- **"Enterprises need to make decisions based on a deduplication ratio that they can realistically achieve in their environment." – Jerome Wendt DCIG**
- **Select datasets that are appropriate for deduplication. Choosing the wrong data to dedupe can cause data expansion due to the generation of metadata.**
- **Types of data that will result in higher dedupe ratios include: Datasets with a low change rate; System volumes; libraries; and Full Volume Backups.**
- **Use a dedupe analysis tool before committing the data to deduplication. An automated tool that can analyze your datasets is preferable.**

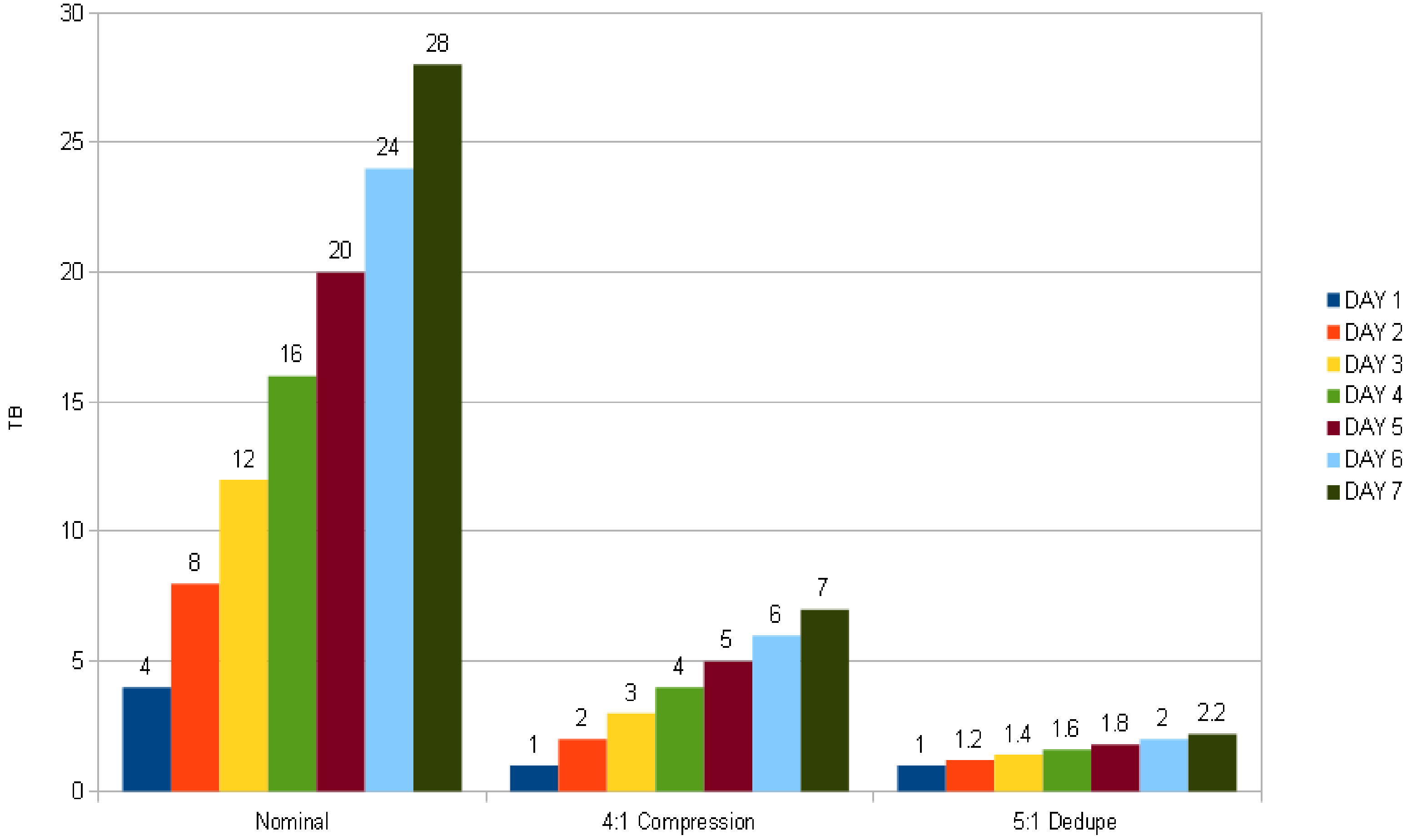
Typical Data Reduction Ratios

- Industry analysis shows:
  - Compression and dedupe ratios vary widely depending on:
    - the variability of the data; the type of backup;
    - the retention period; and the change rate.
  - Typical dedupe ratio claims in the open systems environment range from 4:1 to 50:1.
  - Typical dedupe ratio claims in the mainframe environment range from 2:1 to 20:1.
- Set reasonable expectations for your reduction ratio!

Example: Storage Requirements for a Week of Backups

- Assume a set of full backups one for each day for one week, given 4:1 compression ratio and a 20% data change rate (5:1 dedupe ratio). If you backup 4TB of data each day with no data reduction you would need 28TB of storage.
- If you realize a compression ratio of 4:1 the storage needed is reduced to 1TB per day or 7TB total.
- Adding in a 5:1 dedupe ratio there is 1TB/5 or 200GB of unique data per day. This results in a total storage requirement of 1TB + (6 * .2TB = 1.2TB) or 2.2TB. The overall reduction ratio is (28TB Nominal / 2.2TB Reduced) or 12.7:1. The storage space savings realized is 25.8TB!

Storage Requirements for the Week



Is Data Deduplication Free?

- Inline deduplication requires processing power to support the required ingest rates.
- Some solutions require additional hardware; some have deduplication embedded in the product.
- Data Deduplication generates Metadata in the 5% to 10% range.
- Data Deduplication requires a reconstruction process to render the data in its original format.
- So while data deduplication has great potential it does come with some costs.

Recommendations

- Set reasonable expectations for reduction ratios.
- Analyze your data first. Use a tool to identify good dedupe data sets.
- Select a solution that allows easy control over which data sets will be deduped and those that will not be deduped.
- Select a solution that allows easy monitoring of the overall compression and deduplication effectiveness.
- Consider solutions that incorporate both deduplication and compression.
- Select a data reduction solution that is easy to implement and manage.

Summary

- **Data reduction is comprised of compression and deduplication technologies.**
- **These technologies are multiplicative. When used together they will yield significant increases in storage efficiency.**
- **Proper analysis and selection of dedupe data sets will help ensure the highest reduction ratios will be achieved.**

- **Thanks for your time. We hope you have gained a better understanding of Data Reduction Technologies.**
- **Questions and Answers.**