



# Deduplication

## Shrinking your Backup Footprint and your Backup Window

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Introduction



### **Dan Bailey, MBCP, FBCI**

Managing Director, The Global Continuity Advisors

[Dan.Bailey@GlobalContinuityAdvisors.com](mailto:Dan.Bailey@GlobalContinuityAdvisors.com)

- Involved in the Information Technology field since 1985
- Actively involved in the Business Continuity field since 1991
- CBCP (1999); MBCP (2002); FBCI (2006)
- Co-Founder and President (2001) of the Arkansas chapter of the Association of Contingency Planners
- 2002 President of the North Texas chapter of the Association of Contingency Planners
- 2003-2005 DRI International Certification Commissioner
- 2005-2007 DRI International Education Commission (Chair and Vice-Chair)

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## The Challenges and the Issues



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041

- **The Reality**

- Backups aligning with business requirements

- Business continuity

- Meeting SLAs

- Backup windows
- Restore timelines
- Continuity/Recovery expectations or requirements

- **The Problem with The Reality**

- The accelerating growth of data (50-100% year-over-year growth projections)
- The diversity (too many islands of media where the data is stored – tape, disk)
- The cost (do more with less)
- The shortcomings (low utilization rates)
- The GREEN initiatives (power/cooling/smaller footprints/etc...)
- The legacy (too much older technology – time and money)

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## What is Deduplication?



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041

Deduplication – by definition...

- Is a method of reducing storage needs by eliminating redundant data
- Can be called by other names:
  - Intelligent compression
  - Single-instance storage
- Redundant data is replaced with a pointer to the unique data copy
- Can generally operate at the file-, block-, and even the byte-level

Copyright 2009 – The Global Continuity Advisors – all rights reserved

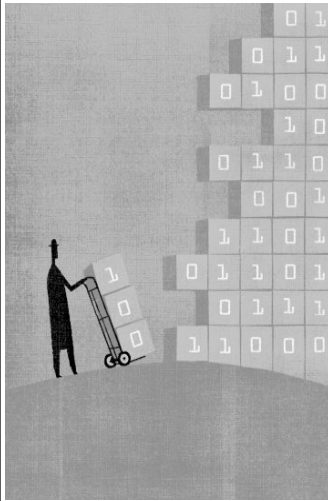
# What is Deduplication? – What's the Value?



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041



- **Storing backup data on disk**
  - 100TB of actual backup data reduced to only 30TB or less
- **Making your backup window more efficient**
  - High-speed data deduplication at between 100MBPS and 1GBPS...or faster
- **Improving the reliability of backup operations**
  - Eliminating the failures associated with tape – with both your onsite backups and your offsite backups
- **Driving the cost of disk-based backup down to the cost of tape...or even less**
  - Reducing the energy, cooling, and floor space required can be a significant part of cost reductions
- **Decreasing time to restore/recover critical data**
  - Restoring and/or recovering data at disk speeds

Copyright 2009 – The Global Continuity Advisors – all rights reserved

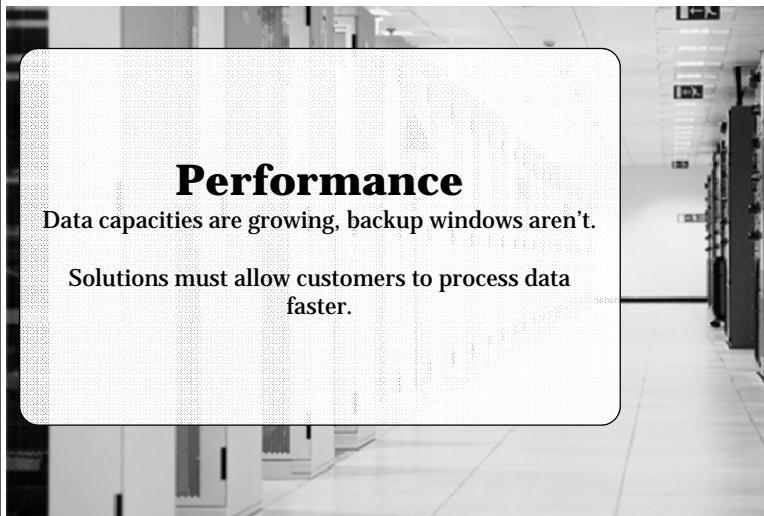
# Deduplication Requirements



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041



## Performance

Data capacities are growing, backup windows aren't.

Solutions must allow customers to process data faster.

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Deduplication Performance



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041

### Megabytes per second (MBPS) or gigabytes per second (GBPS)

- Always measure in AGGREGATE (not peak)
- Per appliance/instance (what is the sustainable rate?)
- How many nodes can have access to the same index and/or repository?
- How will the capacity of the index impact the performance of the unit, as the index is filled up?

Copyright 2009 – The Global Continuity Advisors – all rights reserved

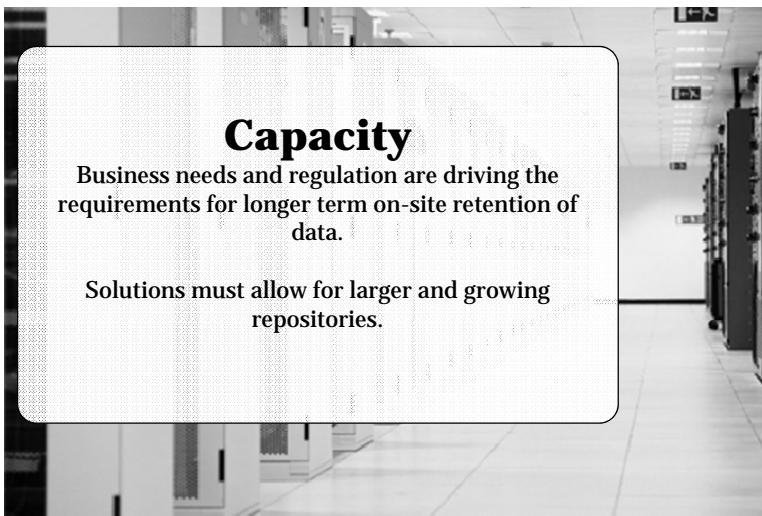
## Deduplication Requirements



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041



Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Deduplication Capacity



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041

### Terabytes or Petabytes?

- Per appliance/instance
- How will the capacity impact the performance as it fills up?
- How is the deduplication ratio being measured – always measure in AGGREGATE for the total size of the solution

Copyright 2009 – The Global Continuity Advisors – all rights reserved

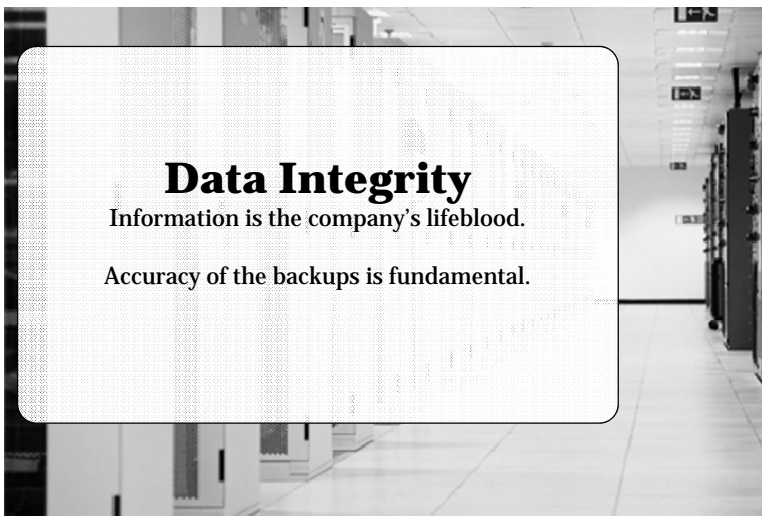
## Deduplication Requirements



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041



### **Data Integrity**

Information is the company's lifeblood.

Accuracy of the backups is fundamental.

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Deduplication Integrity



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041

### RESEARCH!!!

Make sure you are aware of deduplication methodologies and their associated risks/benefits

- File-level (also commonly called Single-Instance Storage)
  - compares a file to be backed up or archived with those already stored by checking its attributes against an index
  - If the file is unique, it is stored and the index is updated; if not, only a pointer to the existing file is stored
  - The result is that only one instance of the file is saved and subsequent copies are replaced with a "stub" that points to the original file
- Block-level
  - Typically uses MD (2, 5) and/or SHA (1, 2, 256, 382, 512) hashing algorithms
  - Hash collisions are a *potential* problem with deduplication
  - Storing unique IDs in an index can slow the inspection process as it grows larger and requires disk I/O
- Byte-level
  - A byte-by-byte comparison of new data streams versus previously stored ones can deliver a higher level of accuracy
  - Before space reclamation is performed, an integrity check can be performed to ensure that the deduplicated data matches the original data objects

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Deduplication Requirements



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041

### **Minimally Disruptive**

The solution must install with minimal downtime and must fit with existing practices, policies and SLAs.

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Deduplication – Ease of use



Dan Bailey, MBCP, FBCI

www.GlobalContinuityAdvisors.com

Office: 972.914.5041

### Installation

- Integrate w/ existing environment?
  - Plug & Play or Rip & Replace

### Daily Operation

- Integrate w/ existing practices, policies, and SLAs?
  - Minimal changes or entirely new structure
- Define the point of completion for the backup window
  - Inline versus Post-Processing

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Deduplication Requirements



Dan Bailey, MBCP, FBCI

www.GlobalContinuityAdvisors.com

Office: 972.914.5041

### Knowledge of what already exists!

- Having knowledge of what exists is key
  - How this knowledge is kept is critical
  - The scale of the knowledge map is pivotal
- 
- If no knowledge exists, the entire data mart must be scanned with every backup instance
  - A catalog or index is created to host the knowledge
- (Word of Caution** - how the catalog/index grows proportionally to the data inhibits some deduplication approaches)

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Analysts and Data Center Customers Agree



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041

### Top Requirements in selecting a deduplication solution\*

- High Cumulative Performance/Throughput\*\*
- Enterprise-Class Data Integrity
- Excellent Capacity/Scalability
- Proven Vendor Reputation/Experience

\*Results from an online deduplication survey conducted by *the 451 group* of in from March to May 2007

\*\* Cumulative performance takes both the ingest and any post-processing into consideration

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Deduplication Design Considerations



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041

### Redundant Data Elimination

The grain of redundancy, 8KB, 16K, 32K, 64K or 1MB+

### Capacity

- Backup to disk has to cope with potentially 100s of TB
- All designs can grow in capacity...but many do so at the cost of performance

### Data Integrity

- Performance battles with Capacity
- Performance is challenged/ curtailed by disk I/O

### Truckless Backup

A major by-product of reducing data

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Two Basic Approaches



Dan Bailey, MBCP, FBCI

www.GlobalContinuityAdvisors.com

Office: 972.914.5041

### #1 Inline

- As data is received by the target device it is:
  - deduplicated in real time
  - not temporarily stored on disk
- Data written to the disk storage is deduplicated

### #2 Post Processing

- As data is received by the target device it is:
  - temporarily stored on disk storage
- Data is subsequently read back in to be processed by a deduplication engine

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Real World Impact of Deduplication



Dan Bailey, MBCP, FBCI

www.GlobalContinuityAdvisors.com

Office: 972.914.5041

<i>Results may not be typical...</i>	<b>Nominal Data Protected (TB)</b>	<b>Physical Capacity - Repository (TB)</b>
<b>Telecom</b>	<b>190</b>	<b>13</b>
<b>Retailer</b>	<b>880</b>	<b>40</b>
<b>Financial</b>	<b>850</b>	<b>50</b>
<b>Manufacturing</b>	<b>450</b>	<b>37</b>

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Who are the market players?



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041

- In alphabetical order...

- Avamar (EMC)
- CommVault
- Data Domain (EMC)
- Diligent (IBM)
- Exagrid
- FalconStor
- NetApp
- Quantum (Dell is partnered; EMC was)
- SEPATON
- Symantec PureDisk & OST

(this list is not necessarily comprehensive...)

Copyright 2009 – The Global Continuity Advisors – all rights reserved



Dan Bailey, MBCP, FBCI

[www.GlobalContinuityAdvisors.com](http://www.GlobalContinuityAdvisors.com)

Office: 972.914.5041

## Q & A

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Value Add - Questions to Consider



Dan Bailey, MBCP, FBCI

www.GlobalContinuityAdvisors.com

Office: 972.914.5041

- How fast is the deduplication process in an operational environment?
- If deduplication is done in parallel to ingest, what is the impact on ingest speed?
- Does capacity scale without impacting performance?
- How does the solution scale in performance?
- Does the system need 'quiet' times for space management?
- Will deduplication impact operational/production activities?
- How long has their system been in production?
- (If appropriate to your environment...) How many customers do they have who backup more than 10 TB per night?

When looking into deduplication based solutions make sure you ask the critical questions

Copyright 2009 – The Global Continuity Advisors – all rights reserved

## Value Add - Data Protection Logistics



Dan Bailey, MBCP, FBCI

www.GlobalContinuityAdvisors.com

Office: 972.914.5041

Technology	RPO Range	RTO Range	Minimum # Disk Copies	Distance	Regional Disaster Support
Tier 6 – Tape BU <sup>4</sup>	24-168 hours	2-168 hours	N/A	Any	Yes
Tier 5 – Virtual Tape <sup>4</sup>	12- 48 hours	1 – 24 hours	Disk Pool	Any if replicated	Yes
Tier 4 – Disk PiT's <sup>4</sup>	Minutes-36 hours	15 mins-12 hours	3 <sup>1</sup>	Any	Yes
Tier 3b – Synch	0-2 minutes	1-8 hours	2 <sup>1</sup>	Limited	No
Tier 3a – Synch W/Failover	0-2 minutes	5-60 minutes	2 <sup>1</sup>	Limited	No
Tier 2b – Async	0-5 minutes <sup>2</sup>	30 minutes-8 hours	2 <sup>1</sup>	Any	Yes
Tier 2a – Async W/Failover	0-5 minutes <sup>2</sup>	30 – 90 mins	2 <sup>1</sup>	Any	Yes
Tier 1 – 3DC	0-2 minutes	1-8 hours	3-7 <sup>1,3</sup>	Any	Yes

Note 1 Best Practice is one additional copy for doing DR testing without impacting the ongoing replication session

Note 2 Network Problems will extend the RPO

Note 3 Depends on vendor and method deployed

Note 4 Depends on how much data being recovered. Local recovery will probably be less and DR significantly more

Copyright 2009 – The Global Continuity Advisors – all rights reserved