CISD - Data Domain Tech Day

Bruce Clarke, Director Product Management, bclarke@datadomain.com

Drew Piersol, Sr. Account Mgr, dpiersol@datadomain.com

April 21, 2009

datadomain

Agenda

Data Domain Overview

Why does Data Domain exist?

Deduplication Defined

What matters in deduplication storage?

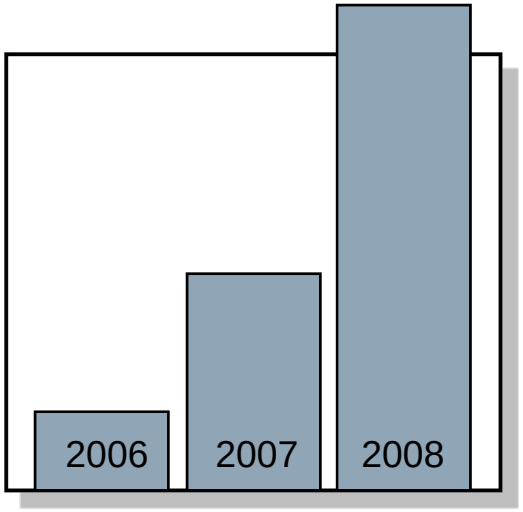
Beyond Backup: Active Archive storage

Disaster Recovery

Database Protection

datadomain

Data Domain Corporate Snapshot



Revenue increased 122% annually for FY2008

Offices in 24 countries

- Worldwide Headquarters: Santa Clara, CA
- EMEA Headquarters: Amsterdam

3 R&D locations

- Santa Clara, RTP, Princeton University

Global support & services

~ 3,000 enterprise customers

~ 500 channel partners

~ 800 employees

NASDAQ: DDUP

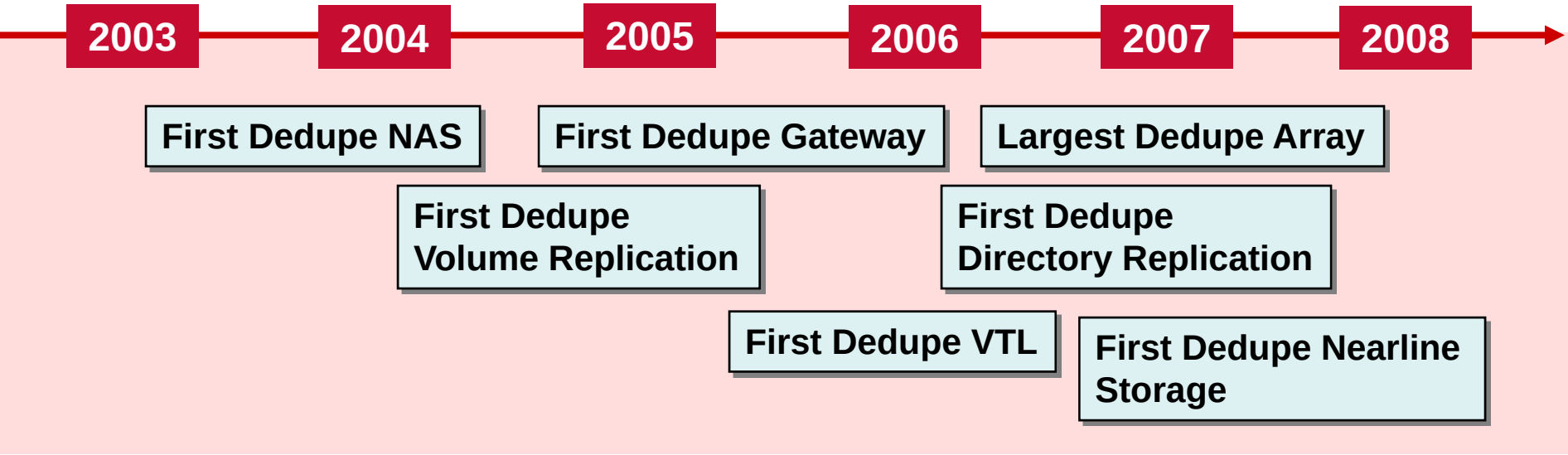
datadomain

Data Domain: Leadership and Innovation

Deduplication Storage Systems

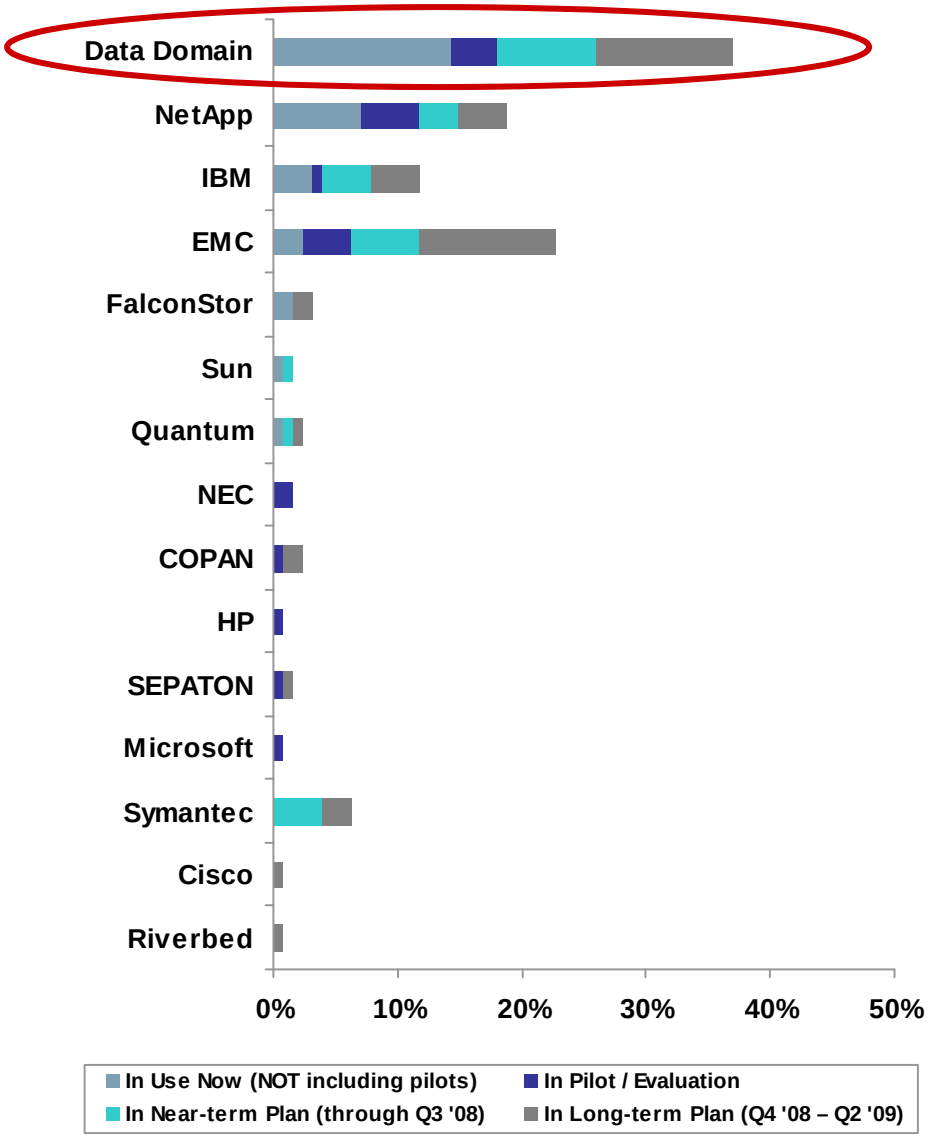
- ~ 7,700 systems installed
- ~ 3,000 customers
- ~ 950 petabytes under Data Domain protection worldwide

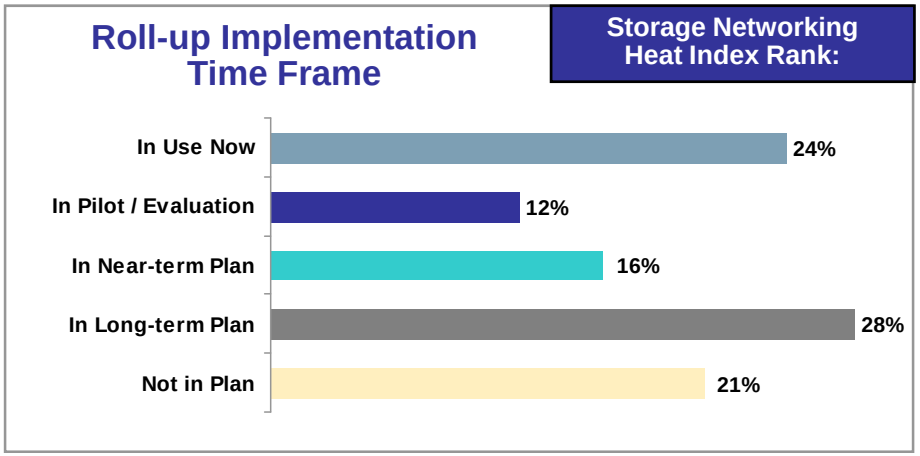
A History of Industry Firsts

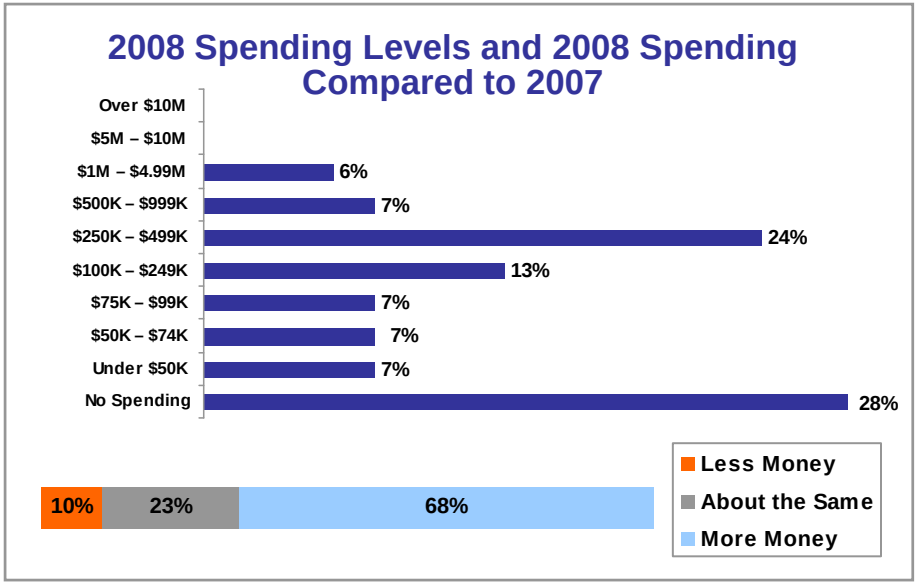


datadomain

Deduplication – F1000







Source: TheInfoPro Wave 11 Storage Study, September 2008

(9/16/08):F1000 Sample. **Left & Top Right Charts:** n=127. **Lower Right Chart:** n=54. **Lower Bottom Right Chart:** n=40

datadomain

Data Domain Basics

Easy integration with existing environment

Backup and Archive Applications

symantec

EMC²

commvault
solving forward

Tivoli. software

BakBone SOFTWARE

vizioncore
Enhancing Virtual Infrastructure

Data Domain DD690 Appliance

CIFS, NFS, NDMP, OpenStorage

Ethernet

VTL over FC

Replication

2U
2 - 6 ports
10 and 1 Gb Ethernet; 4Gb Fibre Channel
RAID-6
5.4 to 35.2TB usable capacity with shelves
1TB or 500GB 7.2k rpm SATA HDD in shelf
File system
NVRAM
N+1 Fans, Power Supplies

Gateway to:
HDS, Fujitsu, EMC, IBM, HP, Sun, 3PAR, Nexsan, Pillar, Xiotech

Confidential

datadomain

Data Domain: Why Do We Even Exist?

Data deduplication is not new, it predates us, goes back to academia in the late 80s/early 90s.

Data deduplication is not proprietary to any one company (contrary to some company's claims)

….but data deduplication was not attempted much as a commercially viable technology until Data Domain came along

Why?

Confidential

Data Deduplication

Data deduplication is by its very nature a highly look up intensive activity

Do the math: Transfer/write data at a (by today's standards) mid range speed of 200 MB/sec.

- Assume an 8K average segment size
- The system will perform 25,000 lookups each and every second
- If you have to go to disk for the lookups 25K/s, you have an inherently slow system for storing data.
- Storage vendors solve disk I/O limitations by over-provisioning, add more disk drives for many more concurrent accesses.
- Not a compelling option in data deduplication: we are trying to reduce the capacity foot print, and here we are expanding it to boost transfer rates. Net: one foot forward, one foot backward.
- Storage vendors decided to just stick with the status quo that storage vendors know and love: keep storing data using sheet metal, spindles and platters.

Confidential

Enter Data Domain

✓ What-If: We can break the disk bottleneck of data deduplication?

✓ What-If: We can shift the workload to the CPU/memory subsystem , away from the notoriously slow disk I/O subsystem?

✓ Then: Go to disk as sparingly as we possibly can, avoid disk access as much as possible, and when we do go to disk, make it count as much as possible.

✓ Bet on the Intel roadmap, avoid the Seagate roadmap. Historically, CPU and memory gets faster and cheaper by orders of magnitude compared to disk I/O which is barely moving the needle.

✓ The collection of patented inventions that represent this innovation has been termed SISL (Stream-Informed-Segment Layout)

SISL: An illustration

2004: 2 CPU platform with 15 drives (DD200)

✓ 30 MB/s

✓1.2 TB of usuable space

2008: same 2 CPU platform w/ 15 drives (DD690)

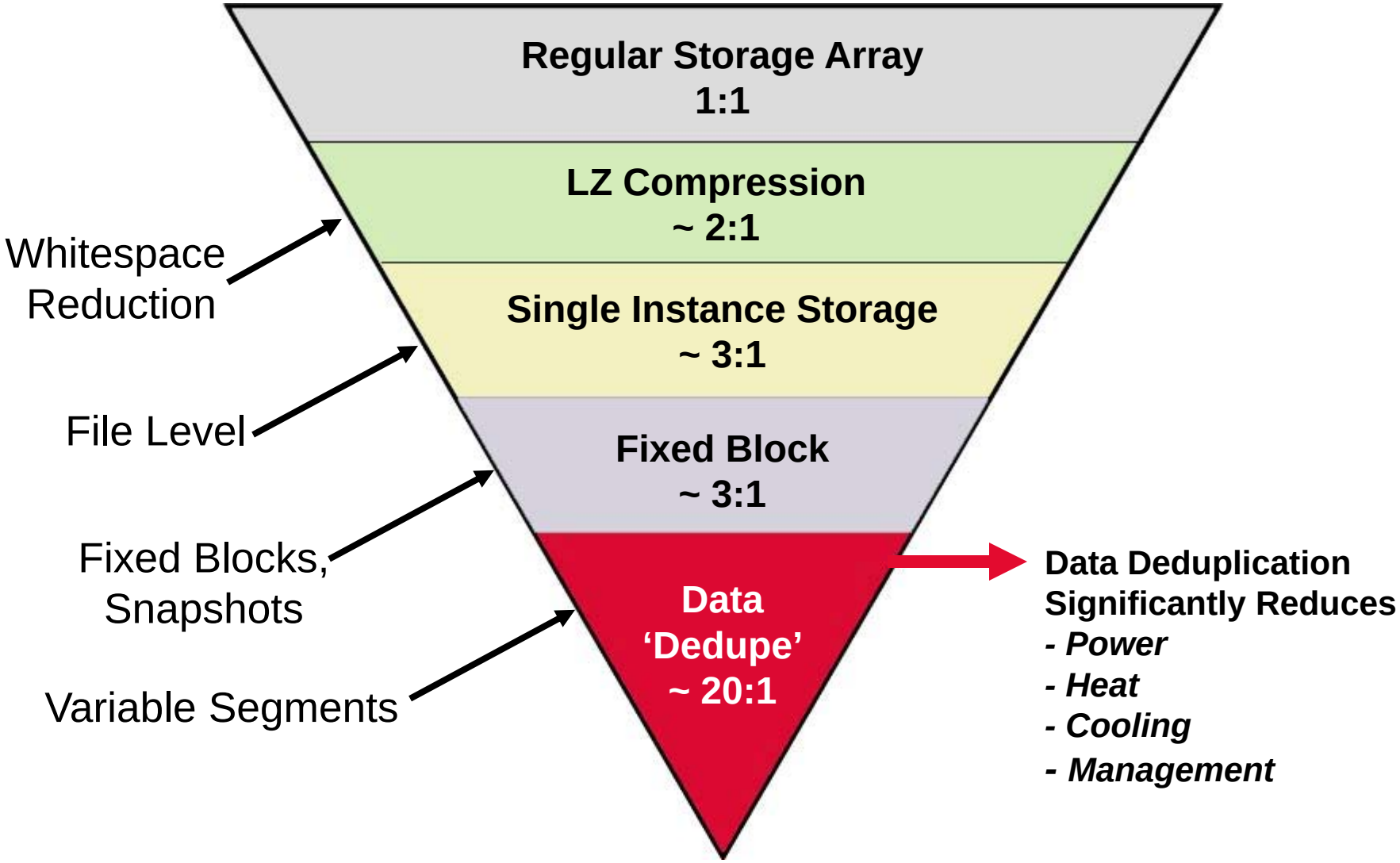
✓400 MB/s

✓35 TB of useable space

✓ Cost of the hardware platform has remained relatively the same in spite of these massive performance and capacity advances.

datadomain

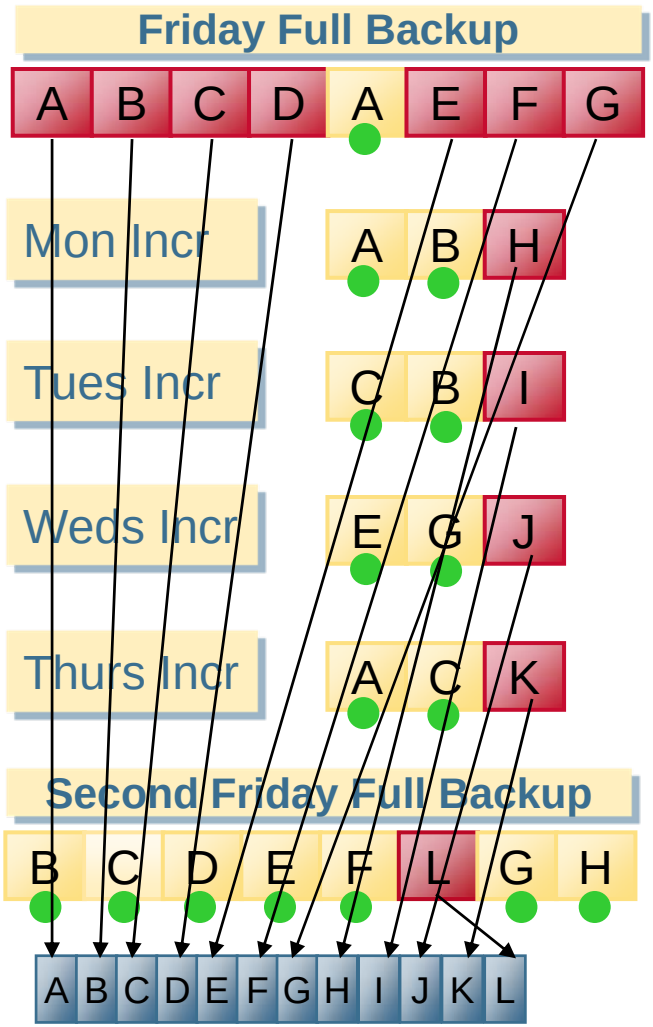
Hierarchy of Data Reduction Types for Backup



datadomain

Data Deduplication: Under the Hood

Store more backups in a smaller footprint



BACKUP DATA	LOGICAL	ESTIMATED REDUCTION	PHYSICAL
FRIDAY FULL	1 TB	2- 4x	250 GB
Monday Incr	100 GB	7-10x	10 GB
Tuesday Incr	100 GB	7-10x	10 GB
Wednesday Incr	100 GB	7-10x	10 GB
Thursday Incr	100 GB	7-10x	10 GB
2nd FRIDAY FULL	1 TB	50-60x	18 GB
TOTAL	2.4 TB	7.8x	308 GB

datadomain

Retain: Store More for Longer with Less

Over 1 year of retention in 3µ of Data Domain deduplication storage

Week 1

Week 2

Week 3

Month 1

Month 2

Month 3

Month 4

BACKUP DATA	CUMULATIVE LOGICAL	ESTIMATED REDUCTION	PHYSICAL
First Full	1 TB	4x	250 GB
April 7	2.4 TB	8x	308 GB
April 14	3.8 TB	10x	366 GB
April 21	5.2 TB	12x	424 GB
April 28	6.6 TB	14x	482 GB
May 31	12.2 TB	17x	714 GB
June 30	17.8 TB	19x	946 GB
July 31	23.4 TB	20x	1178 GB
TOTAL	23.4 TB	20x	1178 GB

datadomain

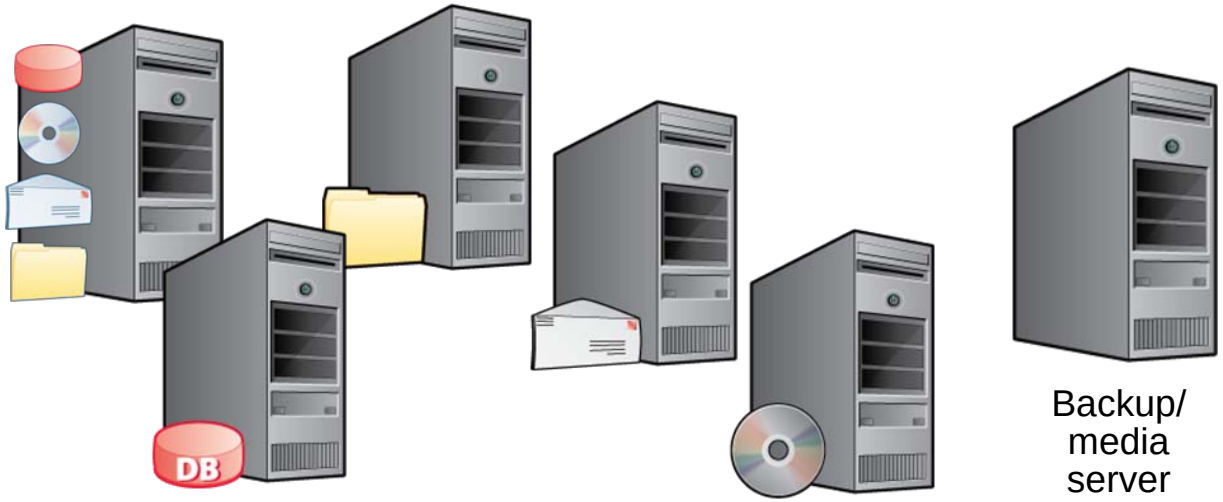
Dedupe All Data Formats, Not Just Some

Dedupe structured data and file data

- One approach for the whole data center and remote offices
- Avoid a patchwork of solutions

Dedupe independent of the backup software

- Support multiple backup packages
- Protect against backup data format changes
- Protect your right to change backup software

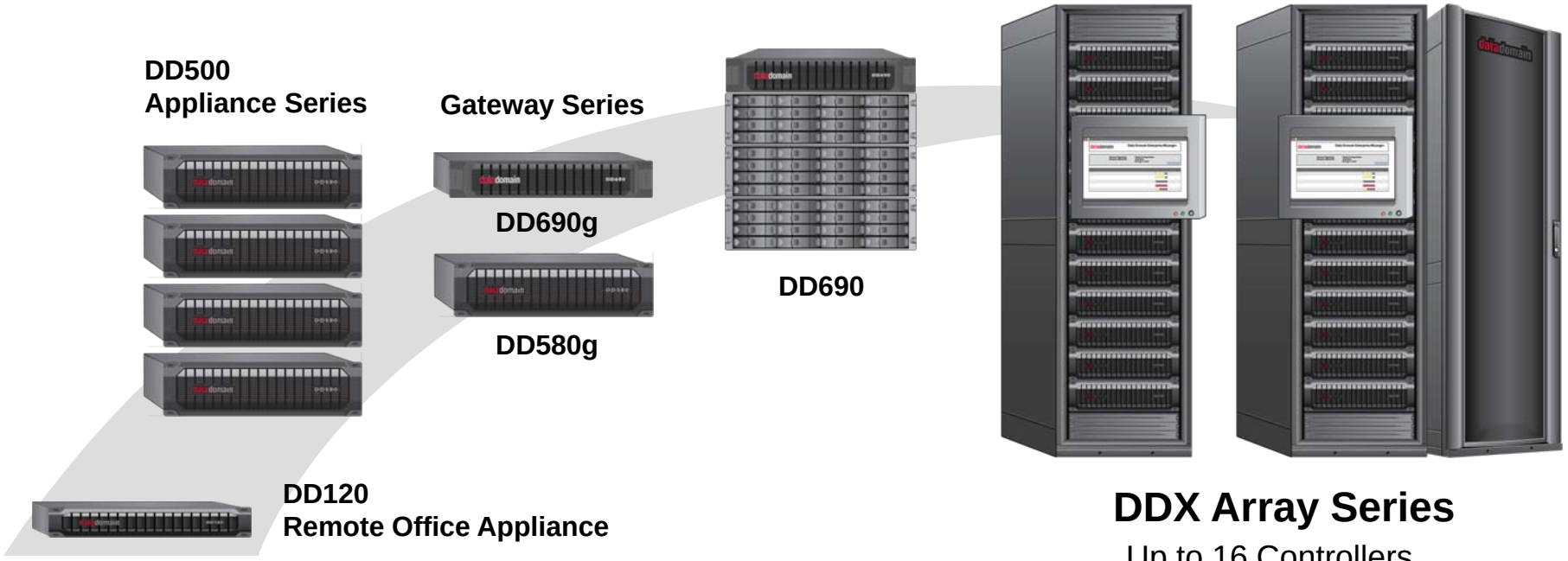


Backup/
media
server

Data Domain DDX Array

datadomain

Industry's Most Scalable Inline Dedupe Systems



OpenStorage, VTL, Replicator & Retention Lock software options

	DD120	DD510	DD530	DD565	DD580/g	DD690/g	DDX Array
Speed	300 GB/hr	435 GB/hr	540 GB/hr	1 TB/hr	1.2 TB/hr	2.7 TB/hr	43.2 TB/hr
Logical Capacity	7-18 TB	55-135 TB	110-285 TB	320-810 TB	430-1.08 PB	710-1.7 PB	11.3-28 PB
Raw Cap. (TB)	.750	Up to 3.75	Up to 7.5	Up to 23.5	Up to 31.5	Up to 48	Up to 768
Usable Cap. (TB)	.373	Up to 2.7	Up to 5.7	Up to 16.2	Up to 21.6	Up to 35.3	Up to 564

datadomain

What matters in deduplication storage?

datadomain

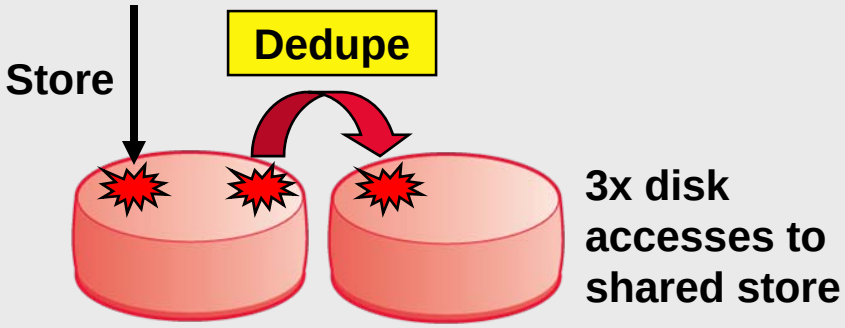
Decision criteria for Deduplication Storage?

- Methodology: Inline vs post-process
- Data Invulnerability: End-to-end verification
- Performance: CPU-centric vs spindle-bound
- Scalability: Multiple tiers of storage, all data types
- Integration: Dedupe in storage, not in backup software
- Replication: DR, consolidate backups, reduce/eliminate tape
- Lower TCO: Install, manage, resources, ‘Green IT’

datadomain

Methodology: Inline vs Post-Process Deduplication

Post Process: Dedupe After Storing

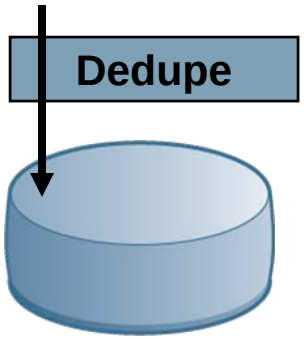


Process contention worse if more of them

- **Copy to tape:** Too slow to stream tape
- **Recovery:** SLA predictability
- **Replication**: Poor time-to-DR
- **Dedupe itself** if interleaved with backup or restore

More admin to fight these issues

Inline: Dedupe Before Storing



Other activities unimpeded

- **Predictable**
- **Simpler**

datadomain

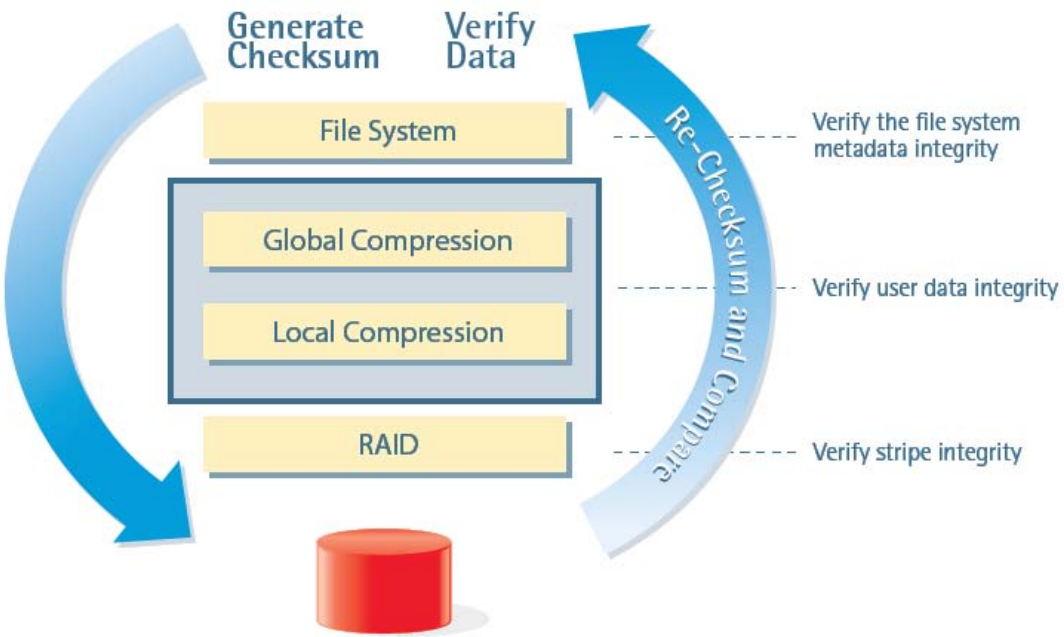
Recovery: Data Invulnerability Architecture

Trust but verify – hope is not a strategy

Data Verification
- **CheckSum**
- **Dedupe, write to disk**
- **Verify**

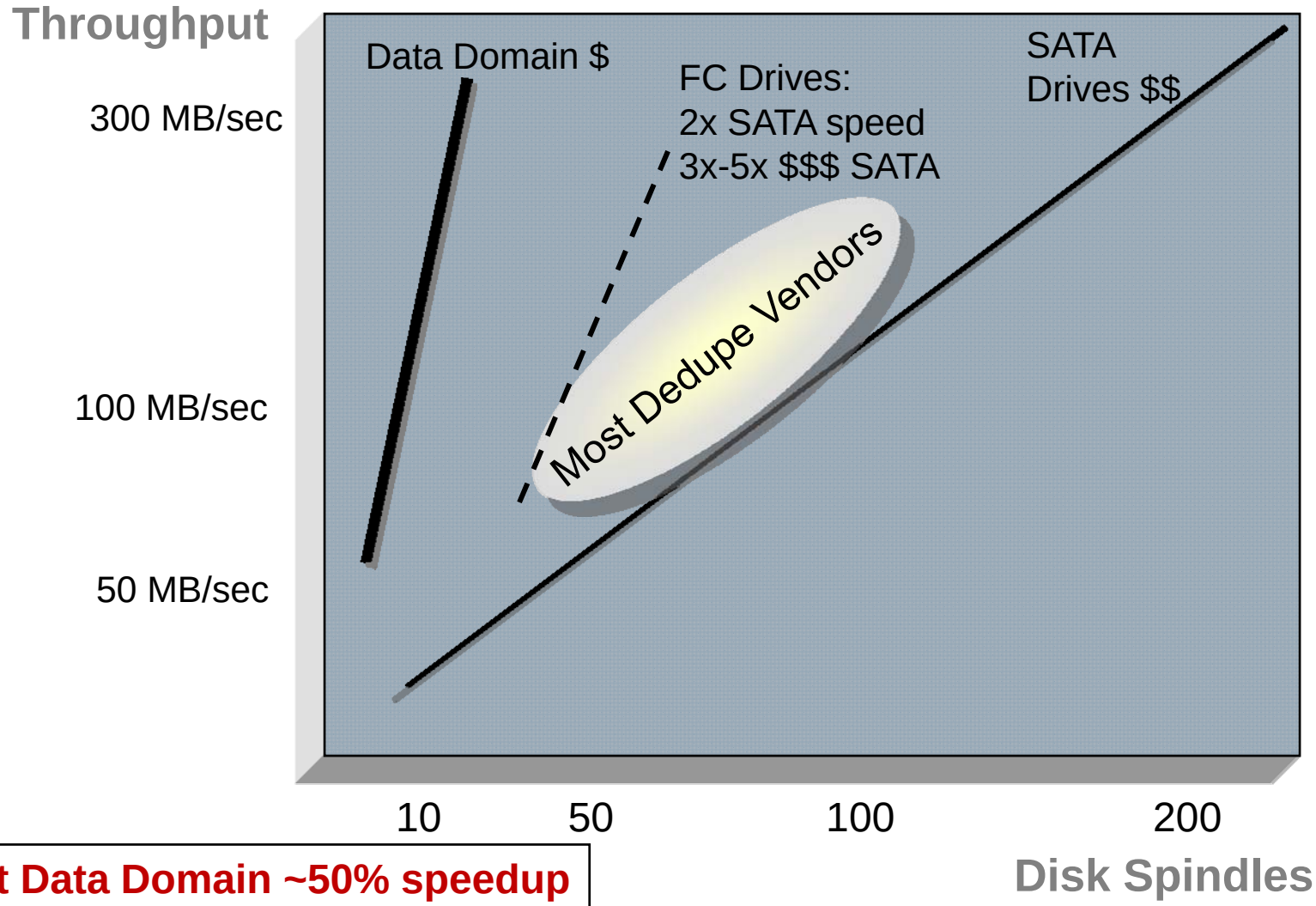
Self-healing file system
- **Cleaning**
- **Expired data**
- **Defrag**
- **Verify**

Other
- **RAID-6**
- **NVRAM**
- **Snapshots**



datadomain

CPU-Centric Design Minimizes Disk and Costs

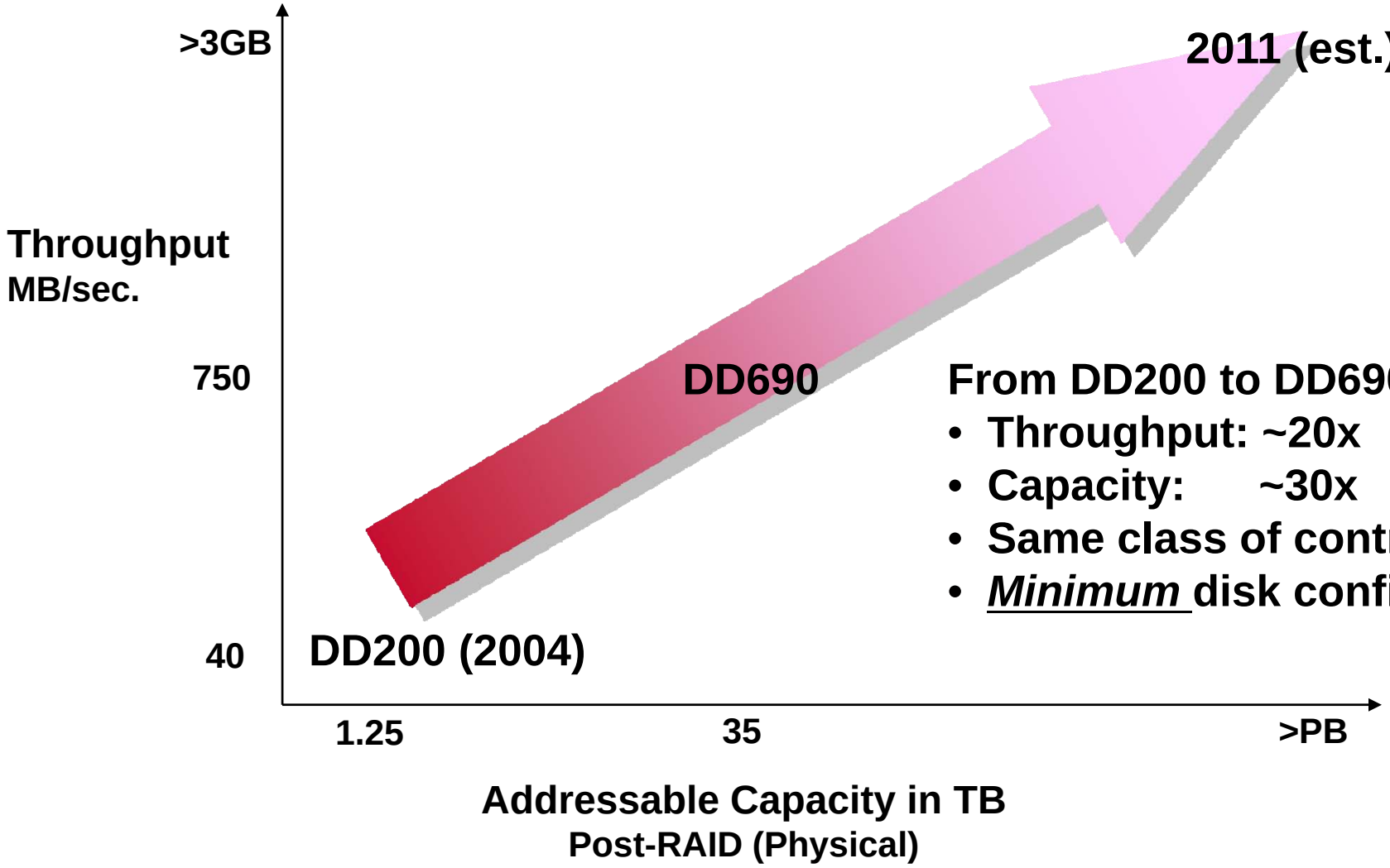


Expect Data Domain ~50% speedup each time # cores doubles

datadomain

Data Domain Scale: Per-Controller Scaling

Data Domain SISL™ Scaling Architecture



From DD200 to DD690:

- **Throughput: ~20x**
- **Capacity: ~30x**
- **Same class of controller**
- ***Minimum* disk config**

datadomain

Dedupe in the Storage, Not Apps Software

Keep backup environment the same

- No new client agents; no new backup software
- Dedupe all backup applications

Keep load off the clients

- Perform dedupe in dedicated hardware

Software independent: use best-of-class

- Don’t have different dedupe silos

Going beyond backup

- Optimization applies to other applications, e.g. archive

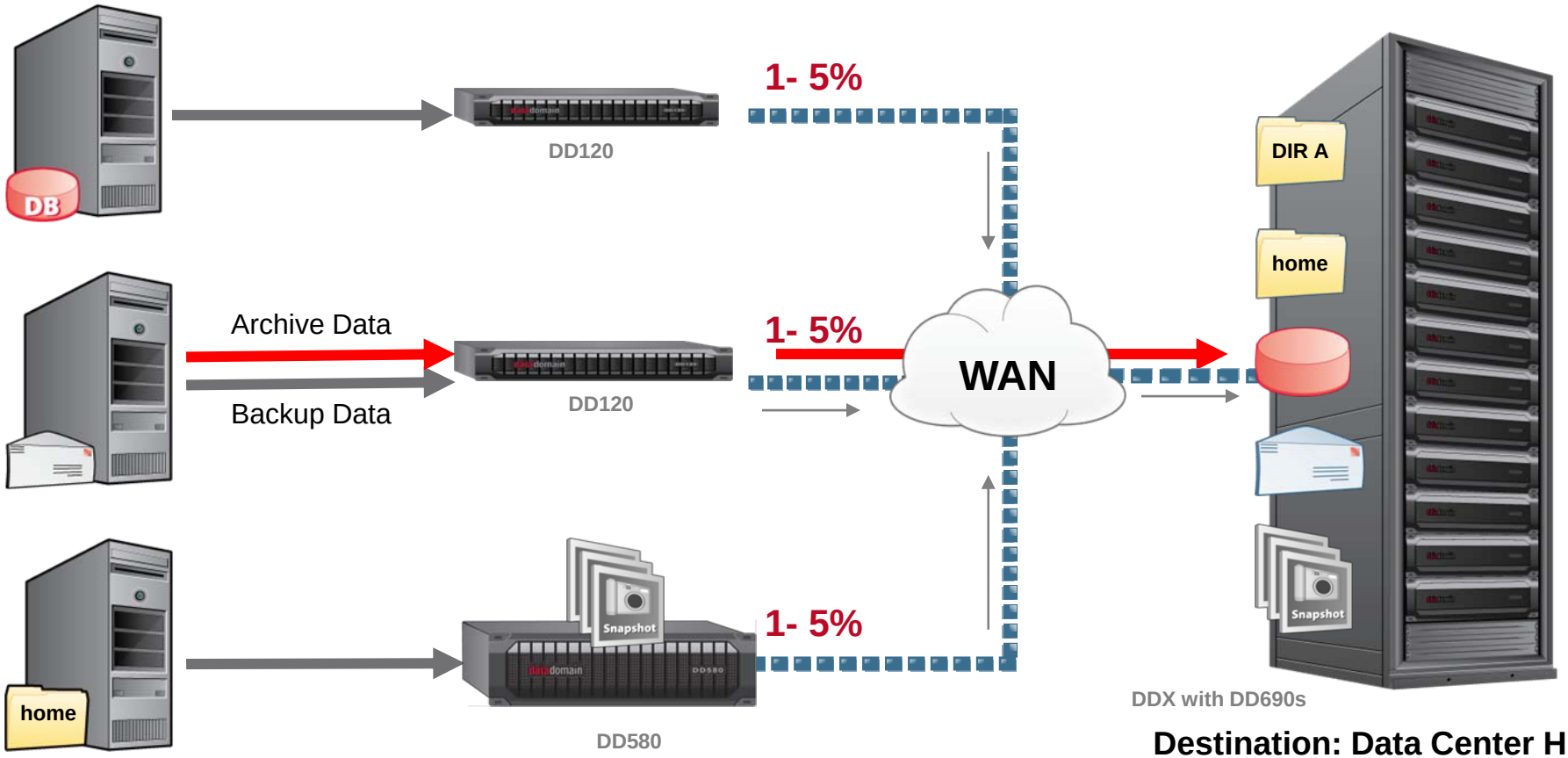
datadomain

Network-Efficient Replication for True DR

True DR; lowers WAN costs; improves SLAs

95- 99% Cross-site Bandwidth Reduction

Flexible Replication
- Many-to-one
- Bi-directional
- System-to-system



Source: Remote Sites

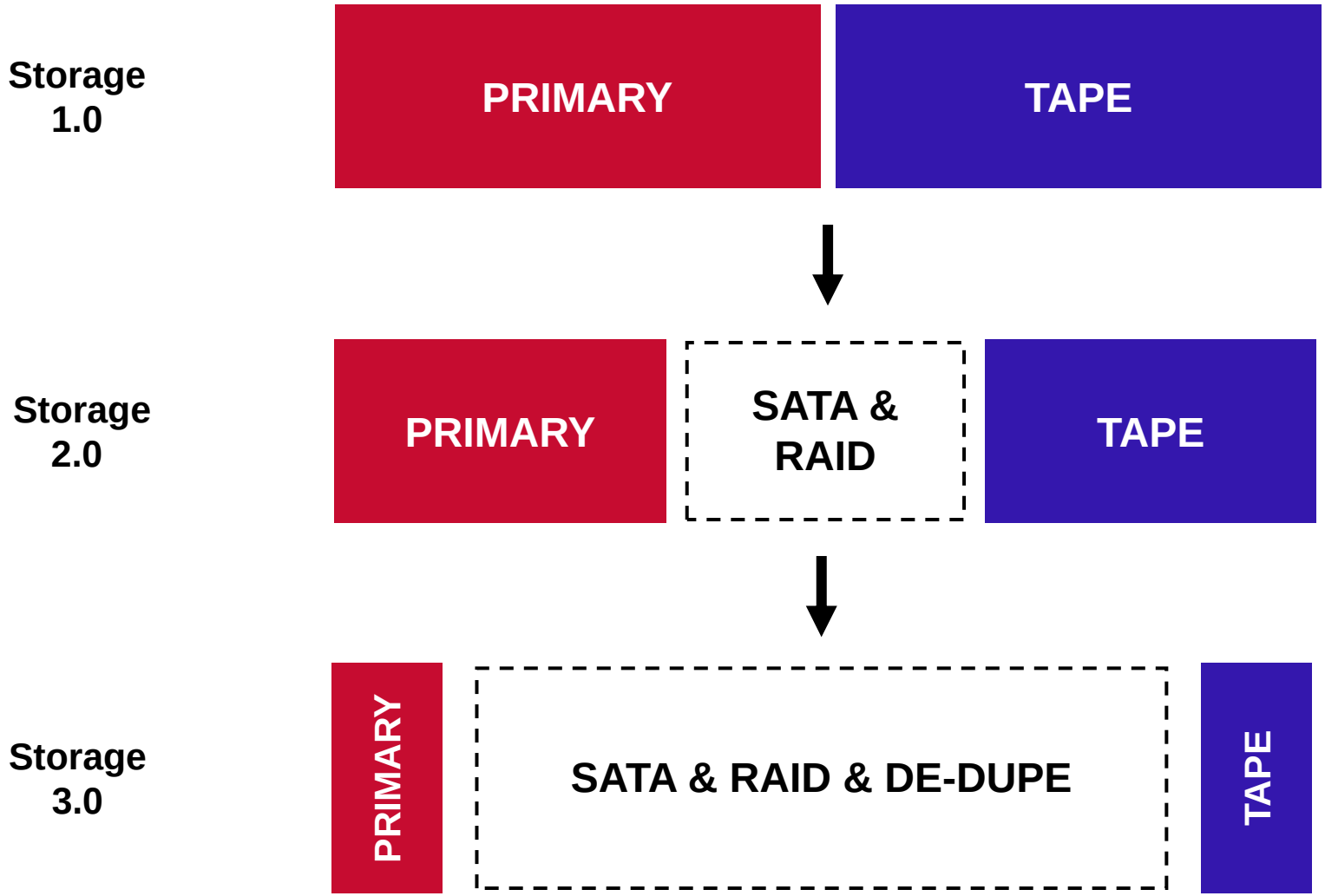
Destination: Data Center Hub
Supports hundreds of remote sites

datadomain

Beyond Backup: Active Archive

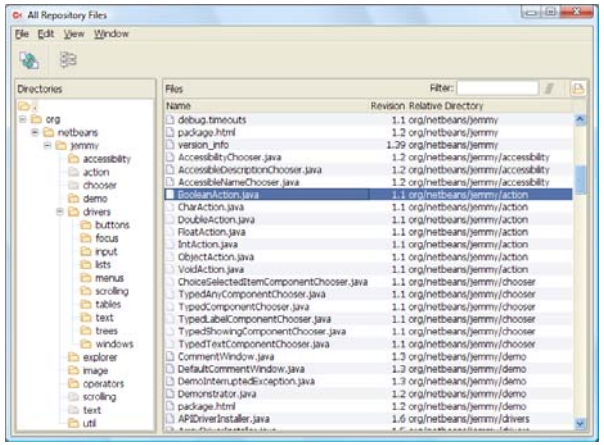
datadomain

Deduplication Defining a New Tier of Storage



datadomain

Online Reference Data – Direct Access

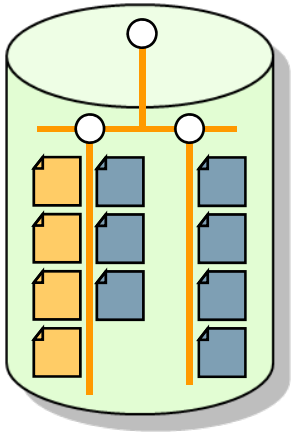


'Mothballed' Directories

Many Reasons to Keep or Move Data

- Engineering software builds
- Legal discovery
- Reference

Completed Projects



Software Build Repository

Resilient & Protected Deduplication Storage

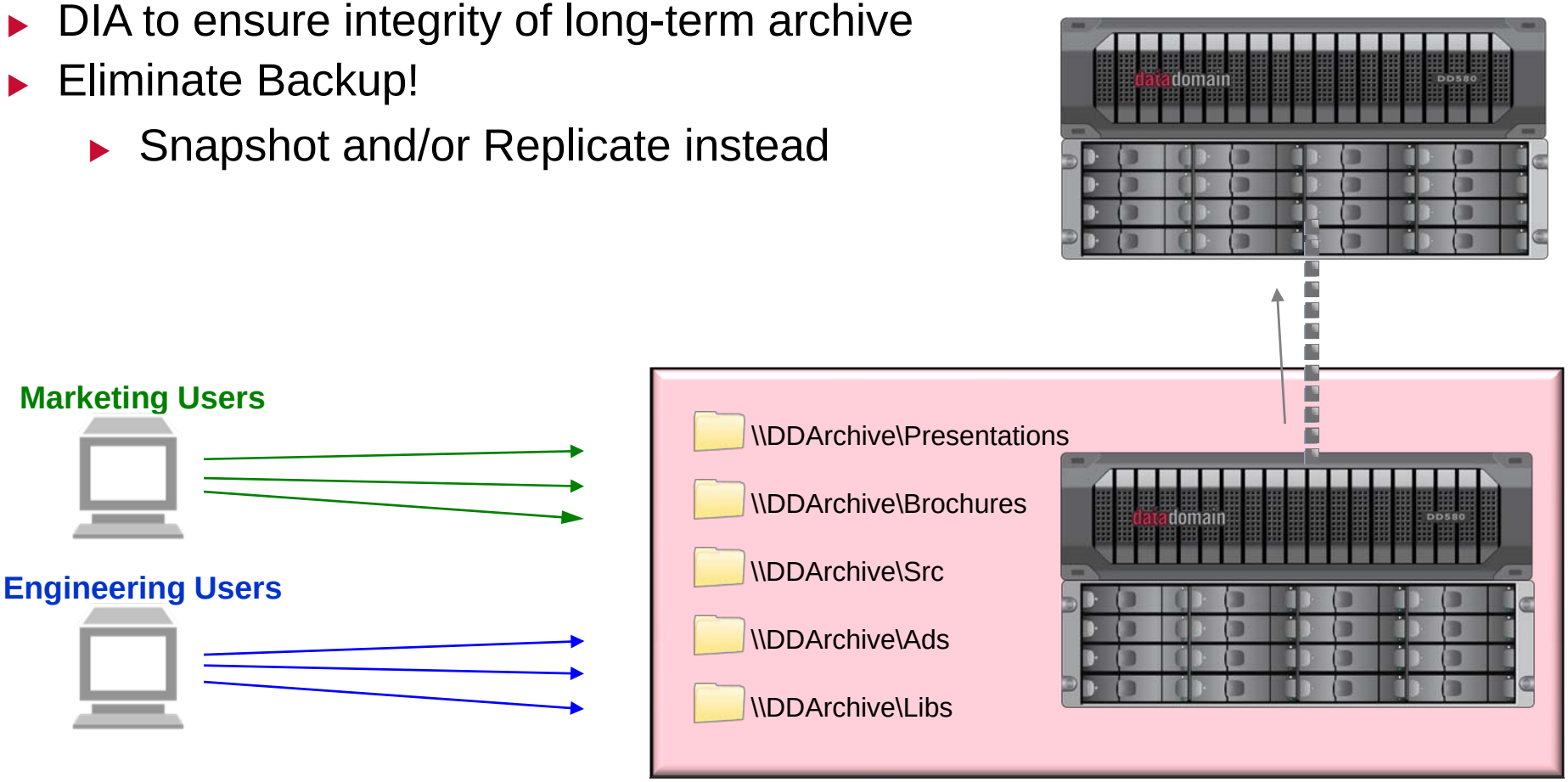
- Data is accessible but on lower cost storage
- Versions can be kept longer
- More granular iterations can be stored for frequently-updated project data

29

datadomain

Direct File Serving

- Archived project data accessed via CIFS shares
- Save up to 80% of space/power/cooling required
- DIA to ensure integrity of long-term archive
- Eliminate Backup!
  - Snapshot and/or Replicate instead



datadomain

Data Domain Nearline Storage Advantages

Deduplication for all data, no matter what the source

- Deduplication across all data stored on the system
- Backup Applications, Archive Applications, & general NAS
- Deduplication is automatic – no management required!

Industry-leading Snapshot Feature

- Efficient space utilization
- Only small incremental changes are stored as data is updated

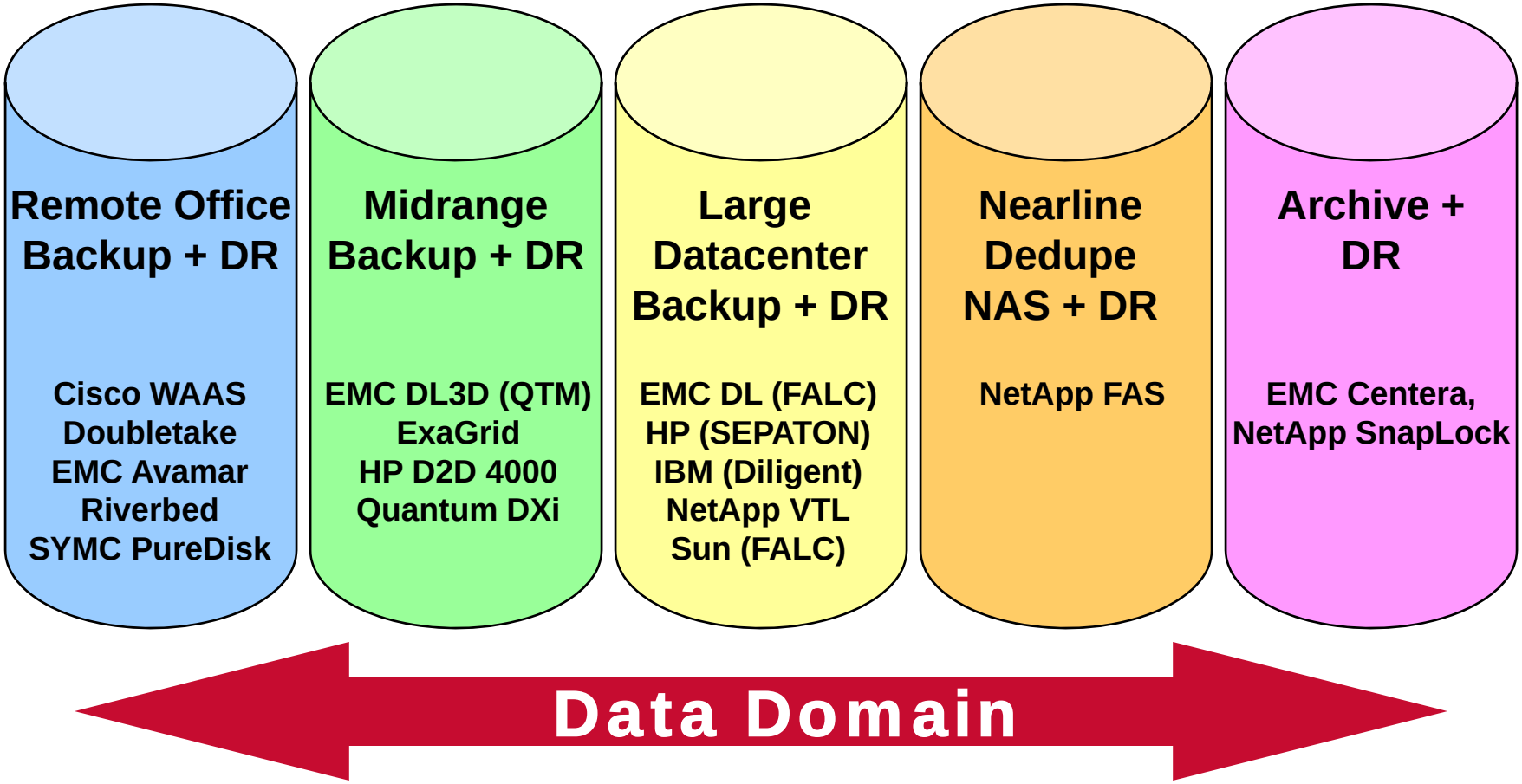
Efficient Replication

- Only unique data is replicated
- Short time-to-recovery, even over slow links

datadomain

Dedupe Flexibility

Existing Application Silos

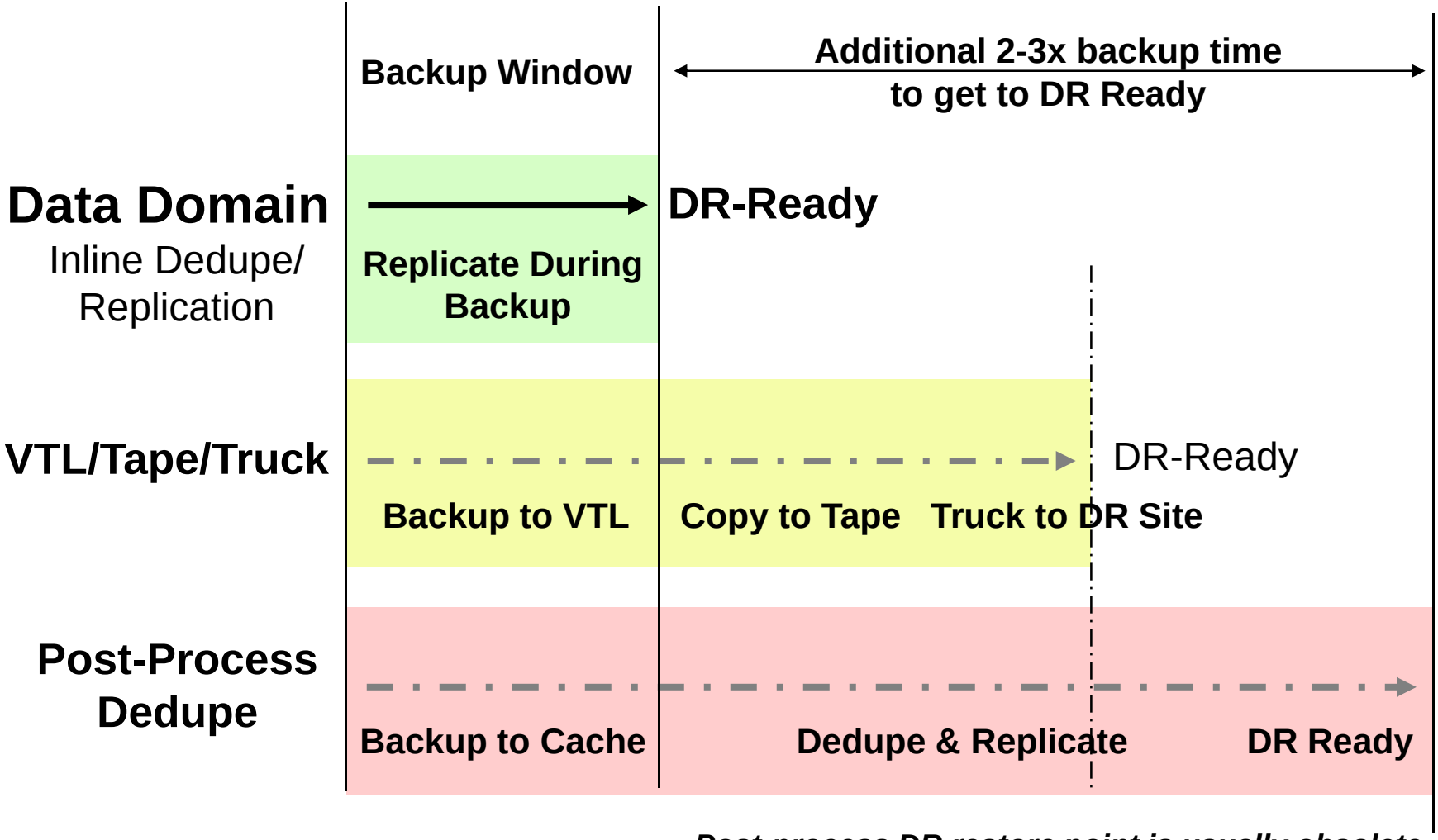


datadomain

Disaster Recovery

datadomain

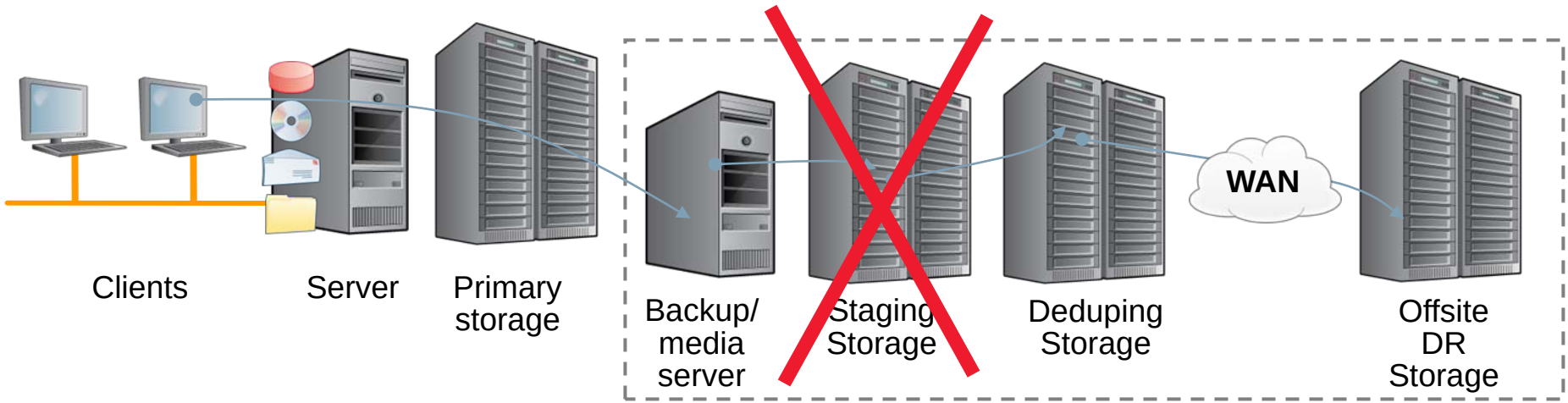
Inline Deduplication for Optimized Time-to-DR



Post-process DR restore point is usually obsolete

datadomain

Dedupe Inline - Why Wait to Replicate?



Shortcomings of post-process dedupe

- Staging disk is as big as primary disk
- Requires 3 local copies of your data – the idea is to eliminate!
- Delays offsite replication for DR; can't replicate until you dedupe
- If dedupe is slow, data doesn't get offsite fast enough

Best to dedupe inline and get data offsite fast

datadomain

Database Protection

datadomain

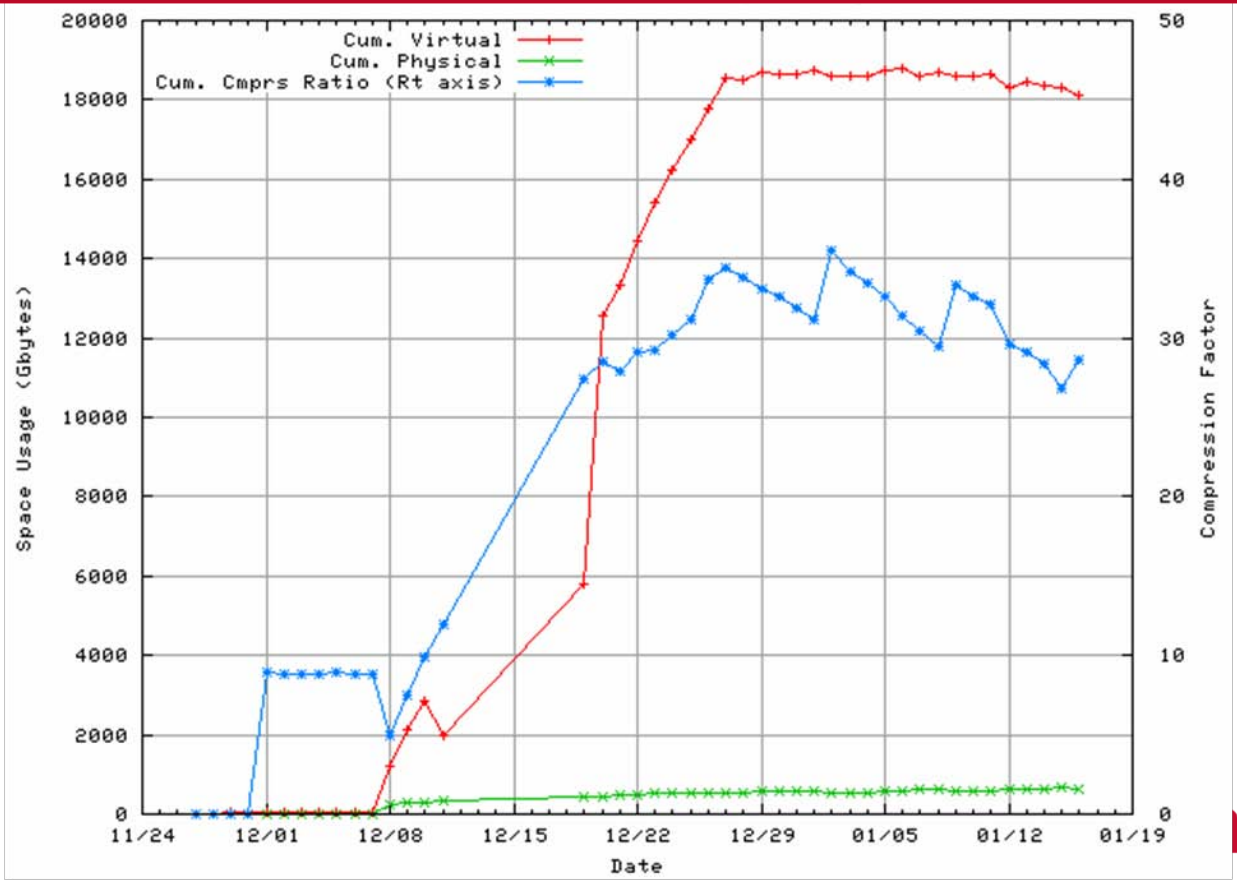
Attraction to Database Administrators

- Daily fulls instead of incrementals
  - Faster recoveries and cloning
  - Space management / compression performed by DDR, not Oracle/MSSQL server thus saving on valuable database server CPU cycles
- No requirement for other Backup software plugins
  - Simplifies environment, saves $$$
  - Storage location on DDR can be used as source for Enterprise backup if necessary
  - DBA is in control without other products involvement
- Non disruptive introduction into the backup process
  - Presents as just another Disk-type storage location
- Simplified replication using Data Domain replication
  - Easy for DBA to implement, validate, and test remote backup data
  - Provides an alternate means for shipping logs to the remote site

datadomain

Data Domain Oracle Customers

datadomain

DDR usage graph in Customer Oracle Environment

Autosupport Date	Compression Ratio	Virtual Space (GB)	Physical Space (GB)	Max Write MB/sec
12/8/2007	4.9	1,193.70	244	98.2
12/9/2007	7.5	2,115.80	282	107.9
12/10/2007	9.9	2,859.00	289	107.5
12/11/2007	12	4,009.40	334	117.1
12/19/2007	27.4	11,568.80	422	156.7
12/20/2007	28.2	12,563.10	446	106
12/20/2007	28.9	12,536.10	434	105.1