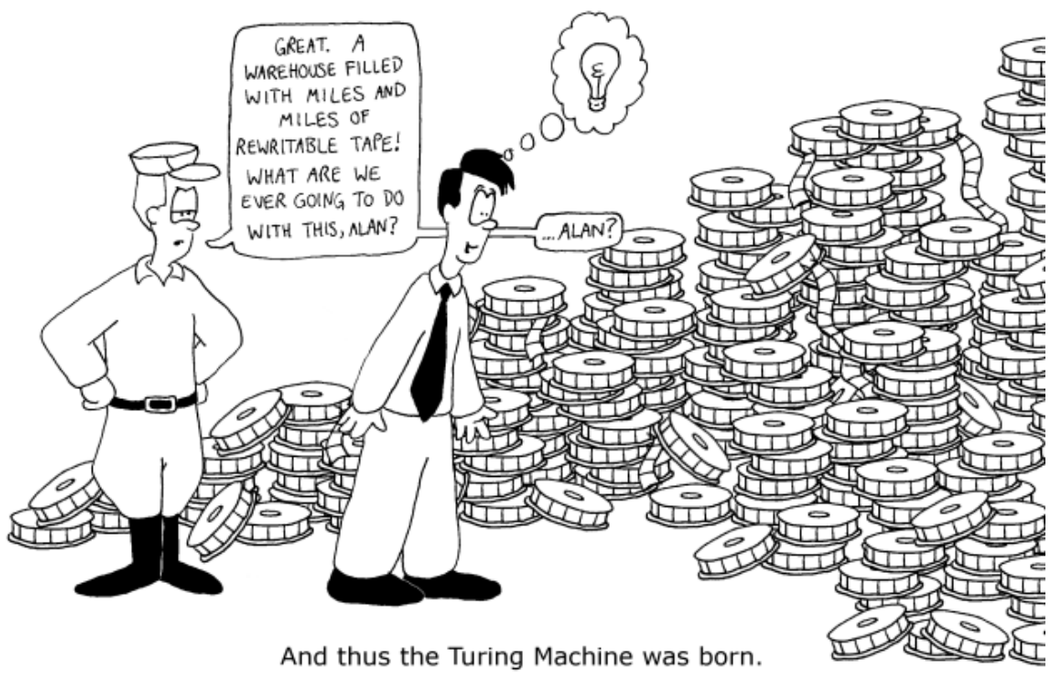
Challenges in Building a Commercial Deduplication Storage System

Kai Li
Paul and Marcia Wythes Professor, Princeton University
&
Chief Scientist, Data Domain at EMC

Disclaimer…

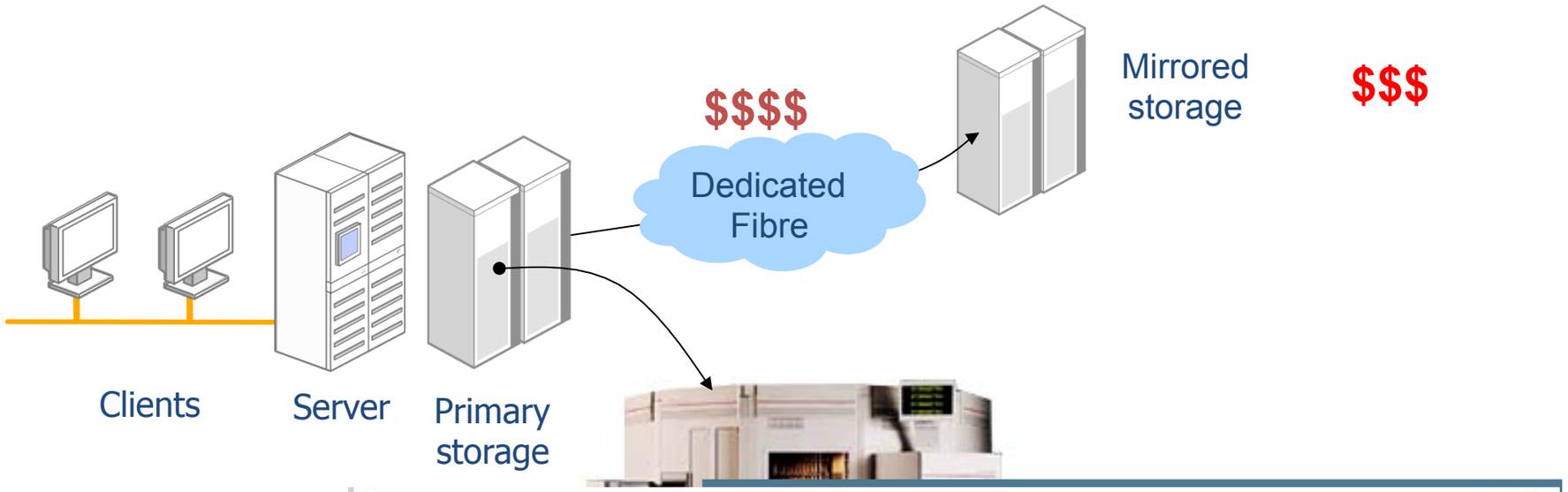
The following is my opinion, not EMC

Why Did We Do It…



Tape Sucks

A Traditional Data Center



US bank loses details of 4.5 million customers

Social security numbers and birthdates are among the data lost by the Bank of New York Mellon Corp

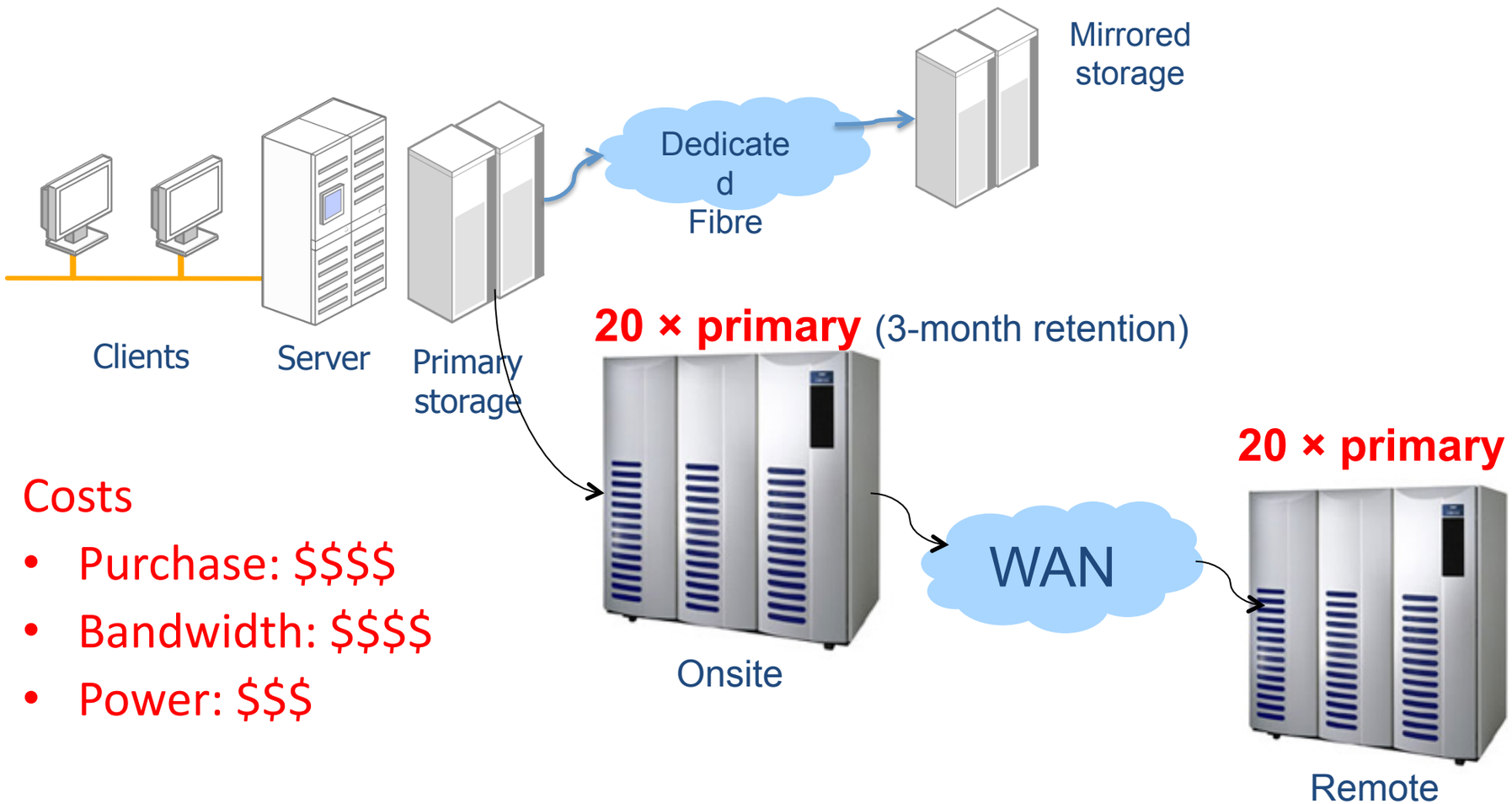
Written by Neon Kelly
Computing, 02 Jun 2008

The details of over 4.5 million customers have gone missing at the Bank of New York Mellon

The Bank of New York Mellon Corporation has admitted to misplacing the details of 4.5 million customers, following the loss of a data tape earlier this year.

The backup tape went missing on 27 February while being transported to an off-site archive by a third-party vendor. The lost data includes the names, birthdates and social security numbers of customers of the Bank of NY Mellon and the People's United Bank in Bridgeport, Connecticut.

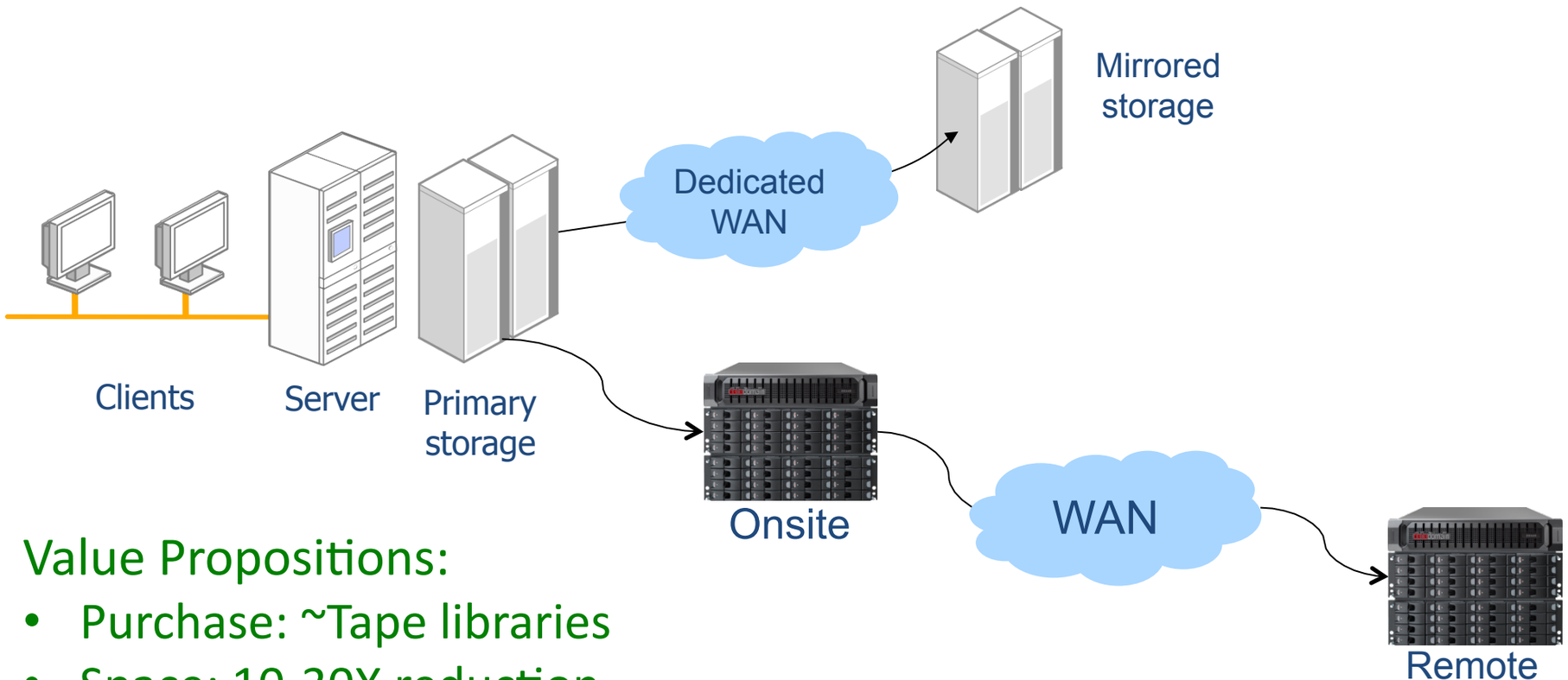
A Data Center using Disk Storage?



Costs

- Purchase: $$$$
- Bandwidth: $$$$
- Power: $$$

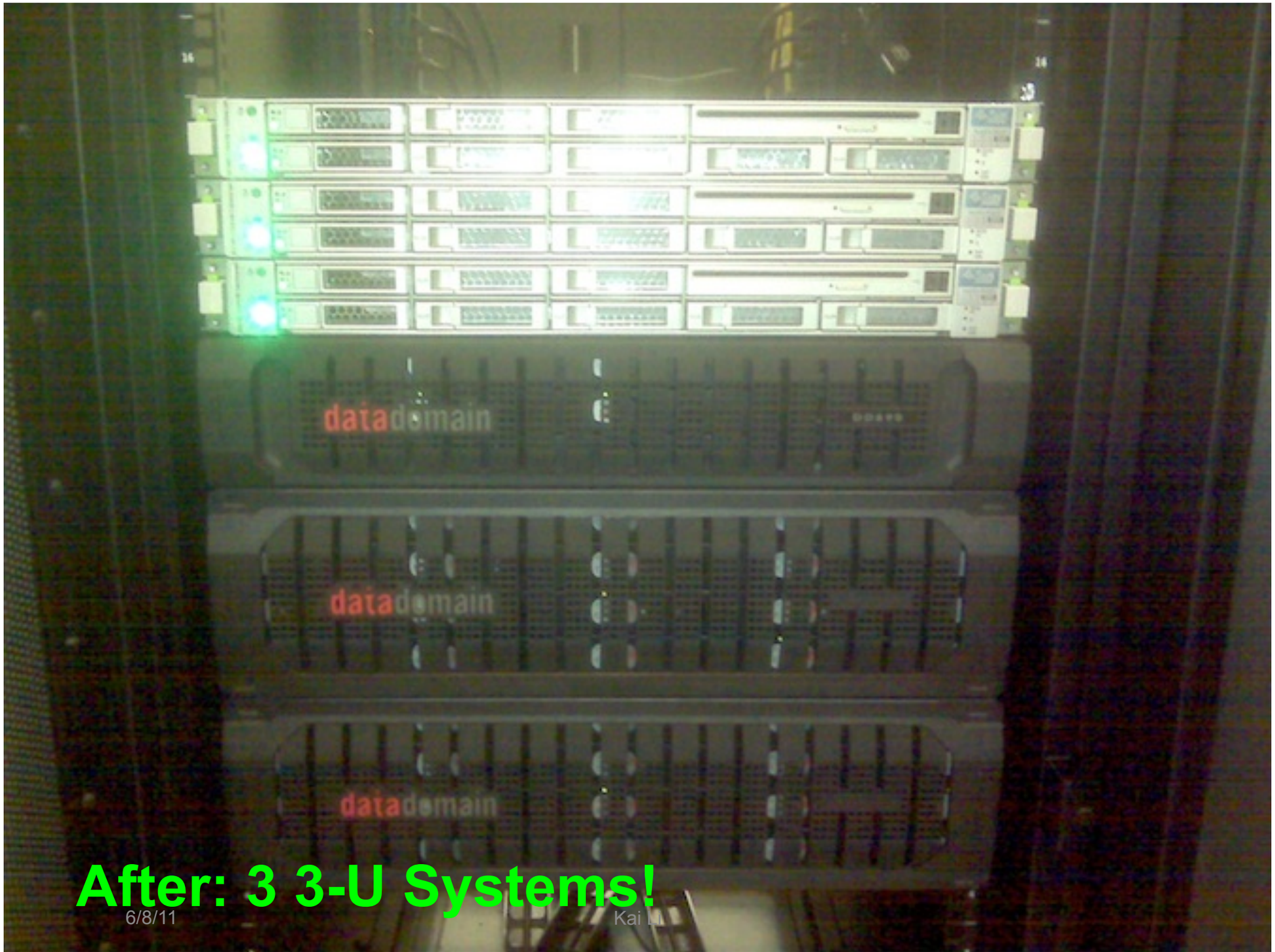
A Data Center
using **Deduplication Storage** Eco-System



Value Propositions:

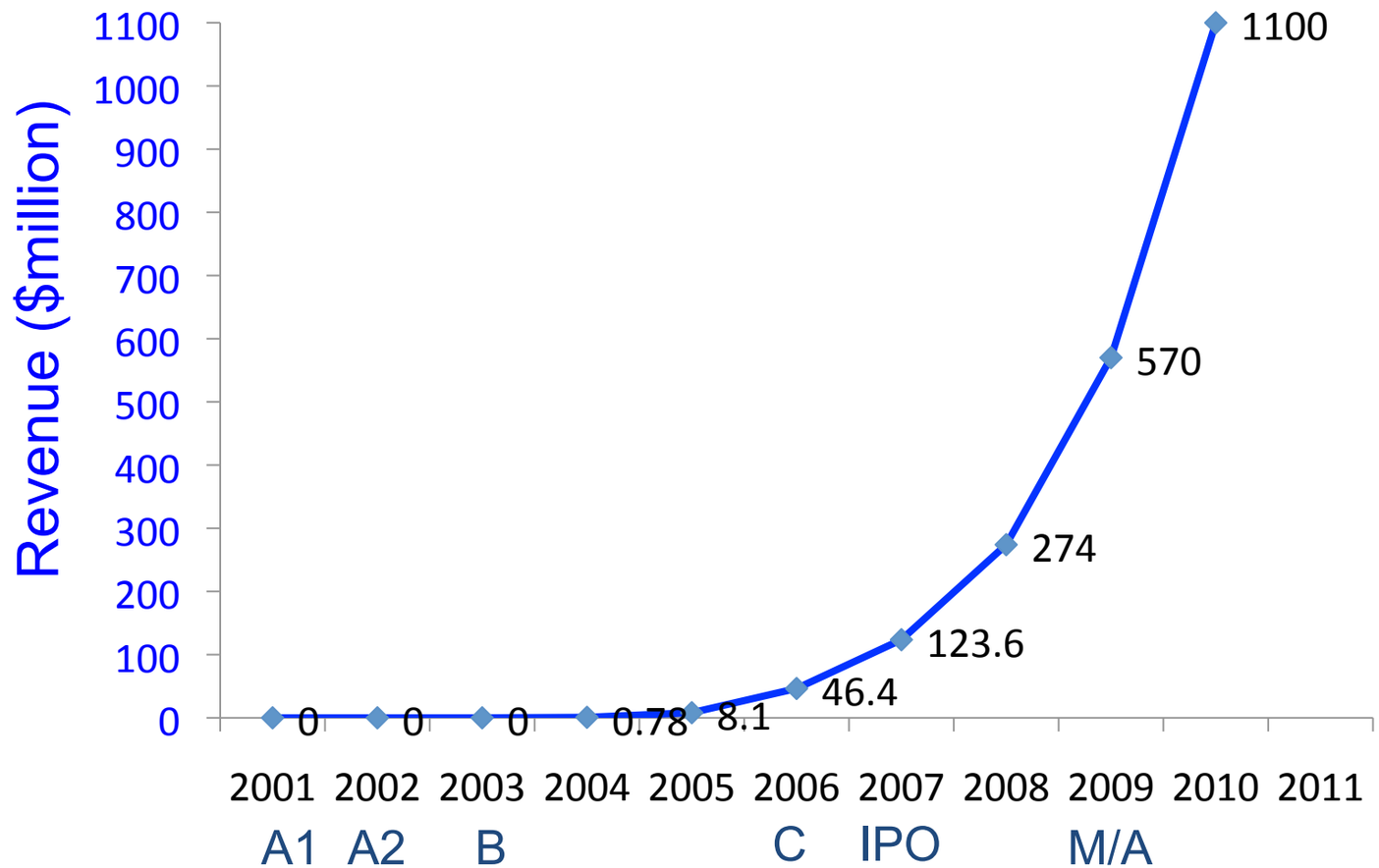
- Purchase: ~Tape libraries
- Space: 10-30X reduction
- WAN BW: 10-50X reduction
- Power: ~10X reduction

Before: 17 Tape Libraries

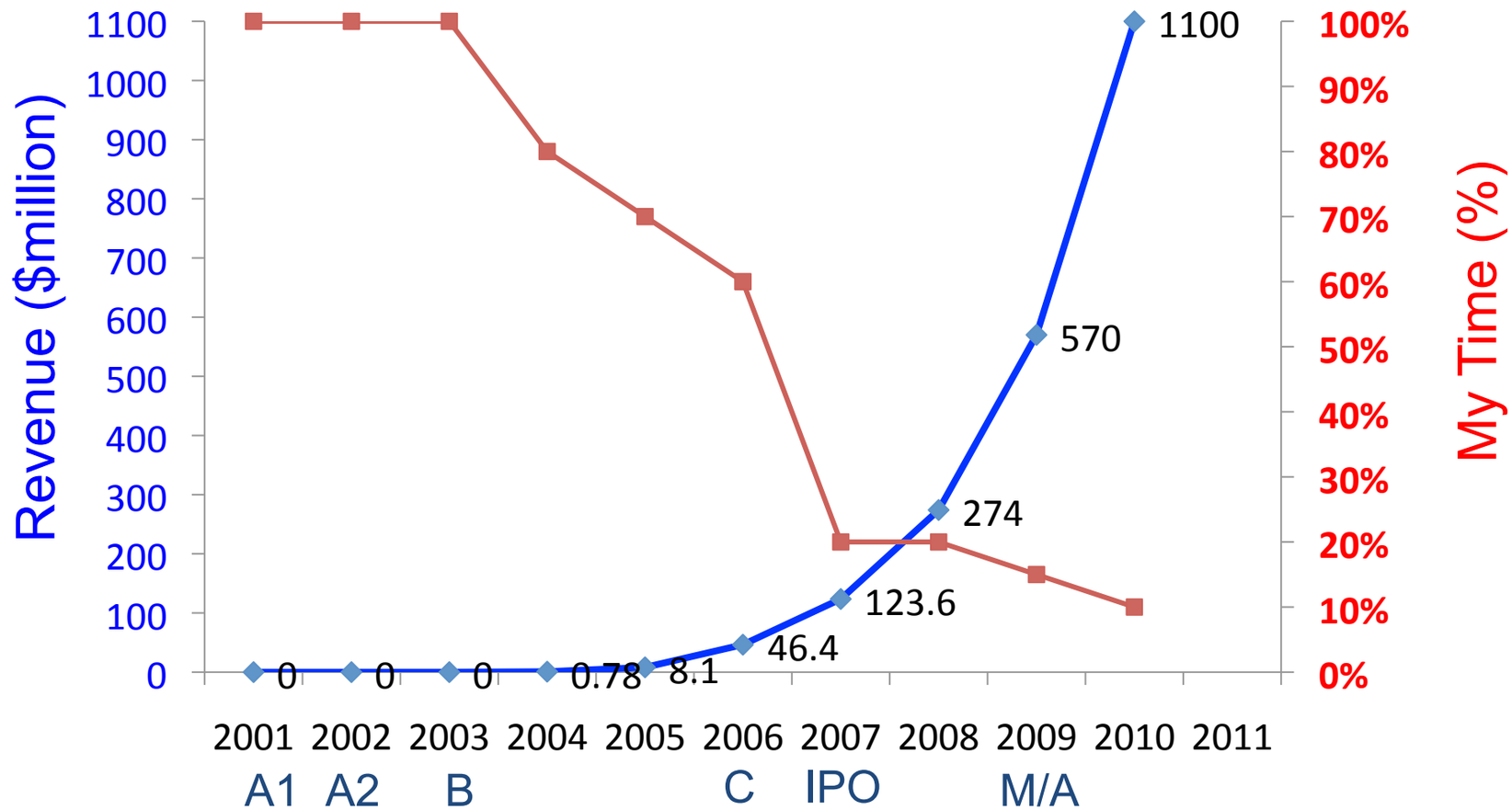


After: 3 3-U Systems!

Data Domain Product Revenue



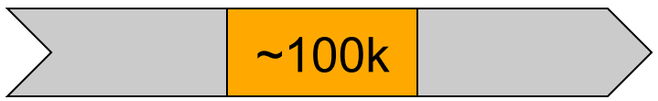
My Contributions?



Deduplication Storage

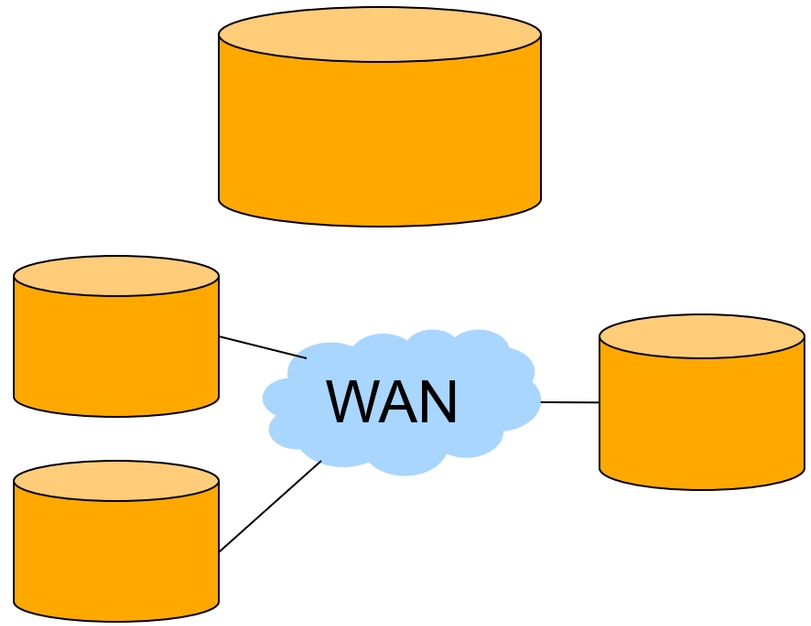
Deduplication: Find Redundancy in A Large Window

"Local" compression



Encode a sliding window [Ziv&Lempel77]

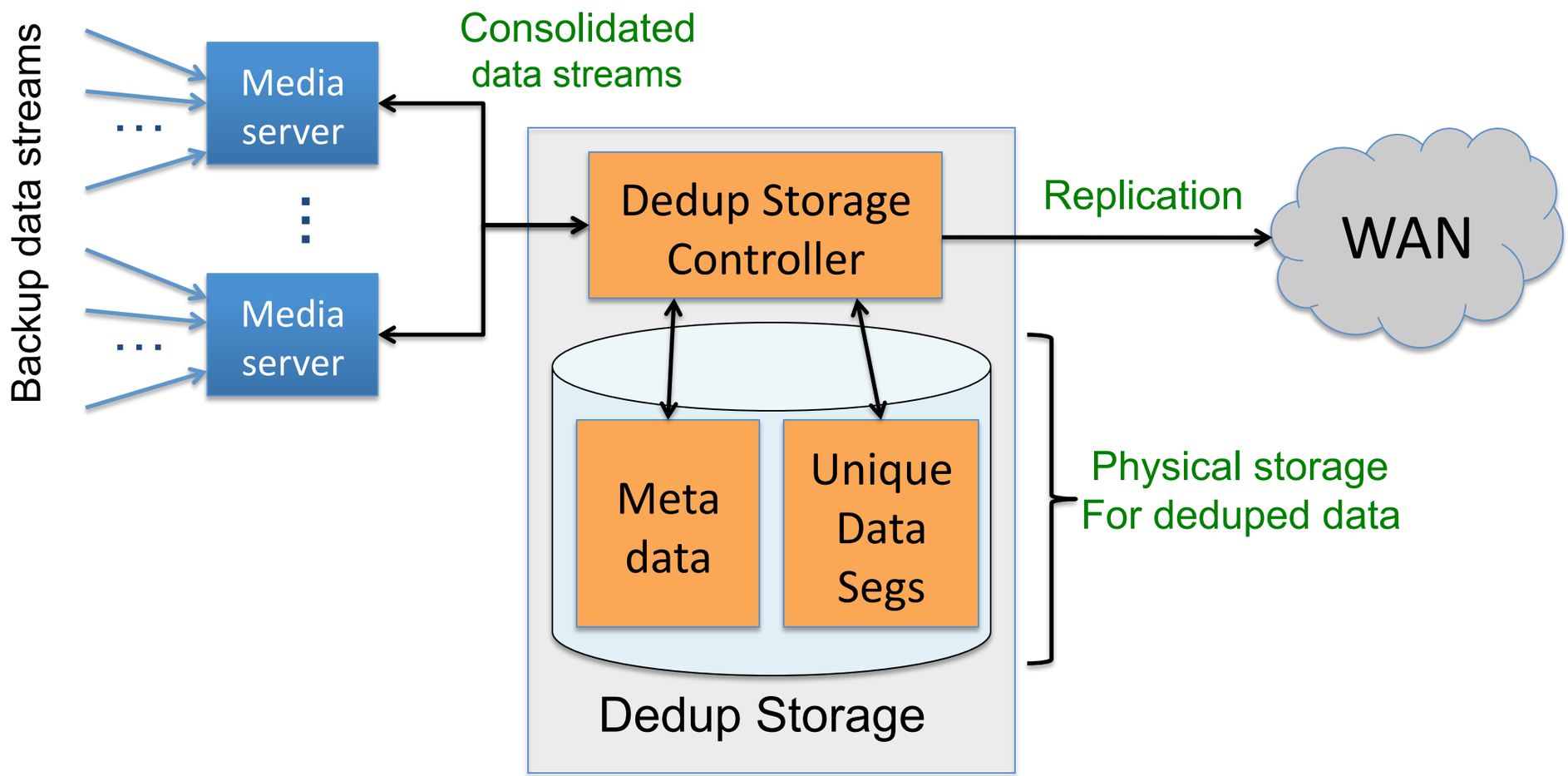
Deduplication



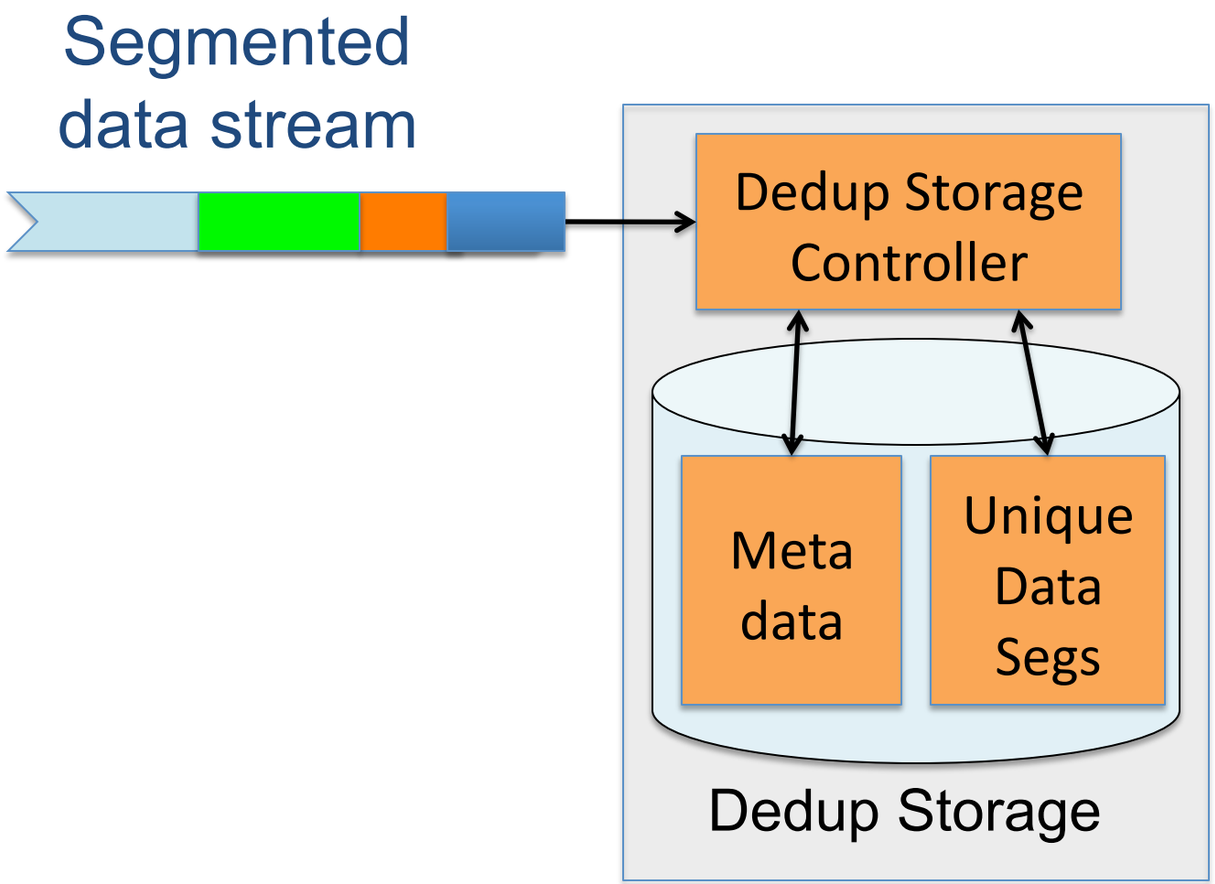
~2-3X compression

~10-30X compression

Dedup Storage System for Backups



How Does It Work?

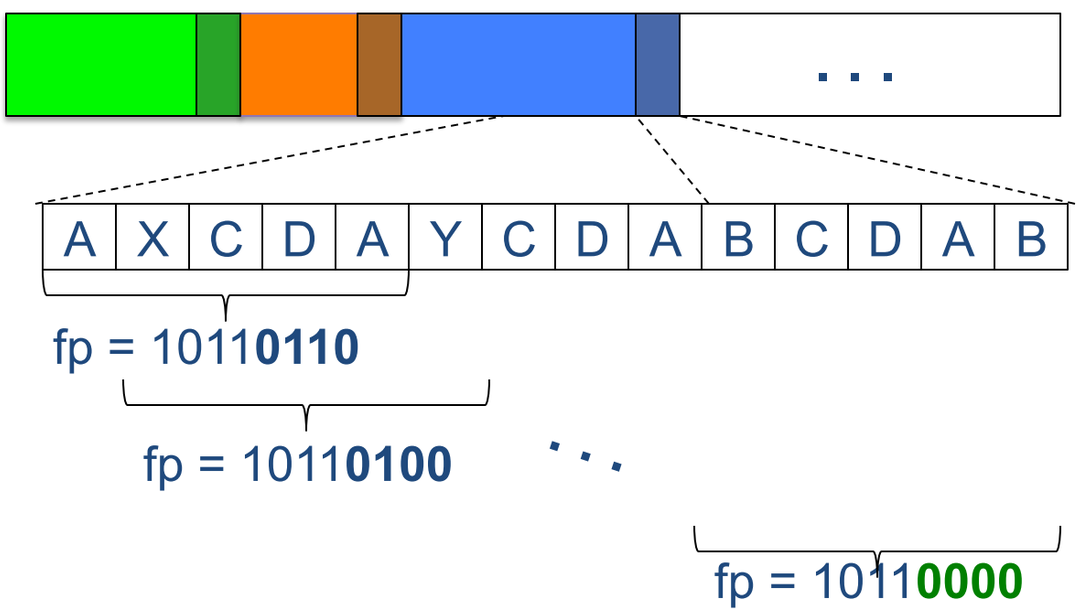


Fixed vs. Variable Segmentation

- Fixed size

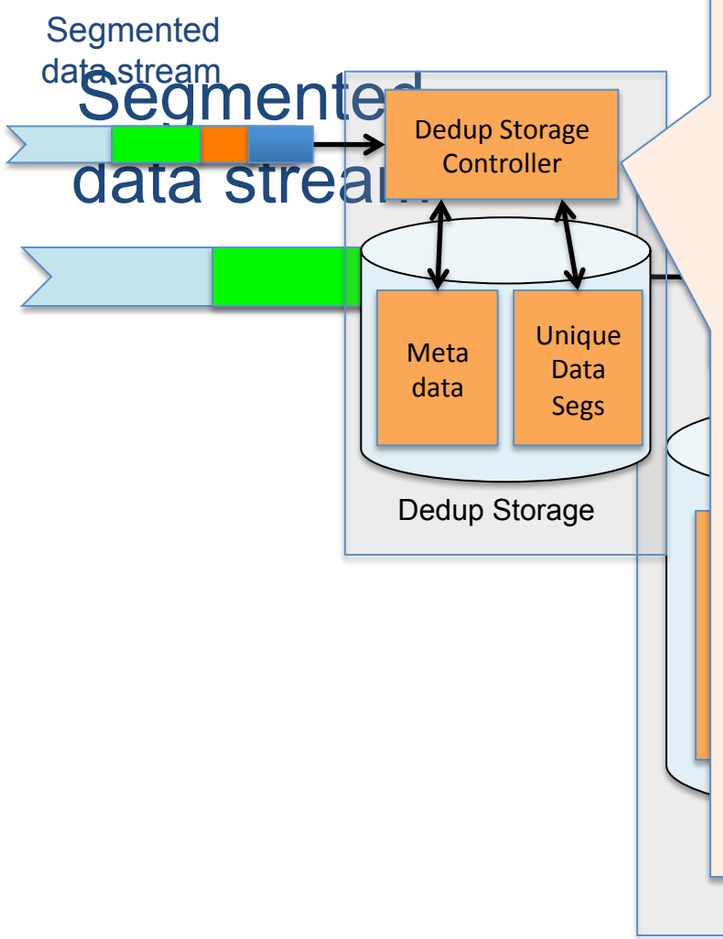
Cannot handle deletes, shifts

- Content-based, variable size



No problem w/ deletes, shifts [Manber93, Brin94]

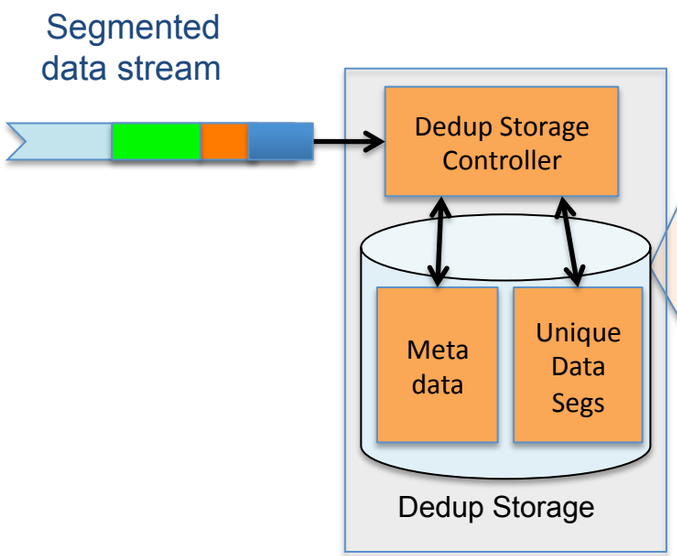
More Details



For each segment

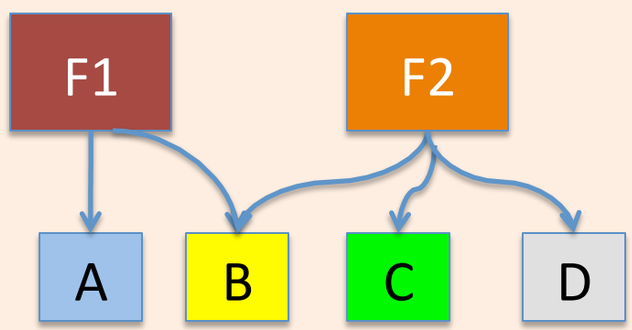
- Compute a strong fingerprint
- Use an index to lookup
  - If unique
    - put fingerprint into index
    - apply local compression to segment
    - store segment
    - store meta data
  - If duplicate
    - store meta data

More Details Continued

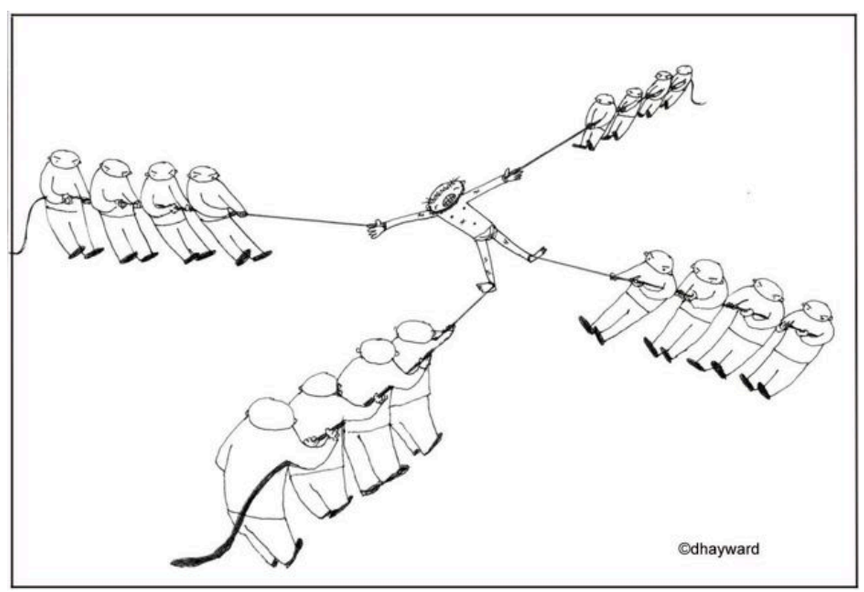


Concurrent Garbage Collection

- A segment may be shared by multiple files
- Backups need to be deleted sometime
- GC cannot interfere with backups



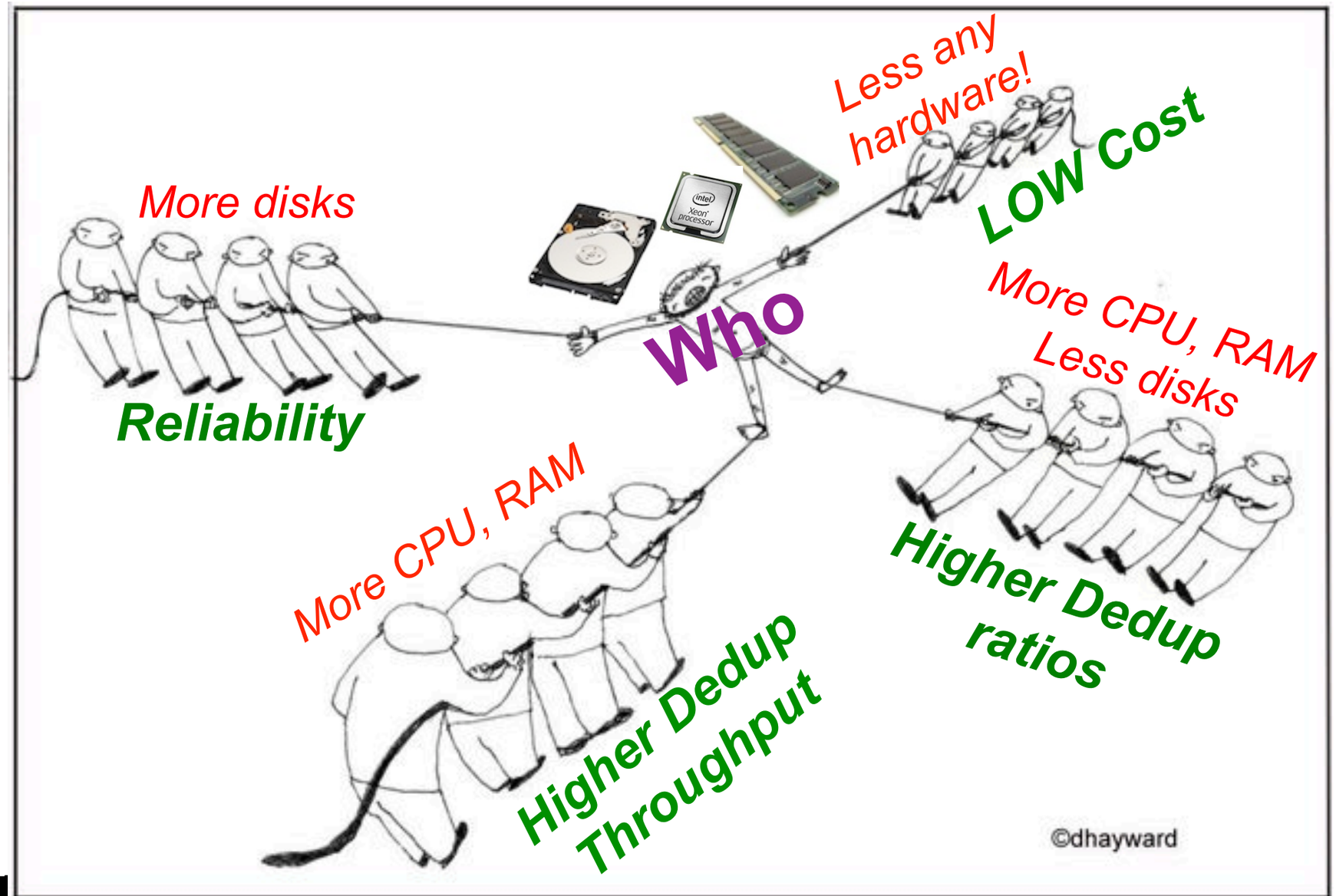
Design Challenges



Challenges for Backup Storage

- Facts
  - Data doubles every 18 months
  - 24 hours / day
- Requirements
  - Complete backups within the “backup window time”
  - Fast recovery from local or remote backups
  - No new budget
- Challenges
  - Low cost
  - High compression ratio
  - Increase deduplication throughput and capacity
  - High availability and data integrity

Forces in Dedup Storage Design



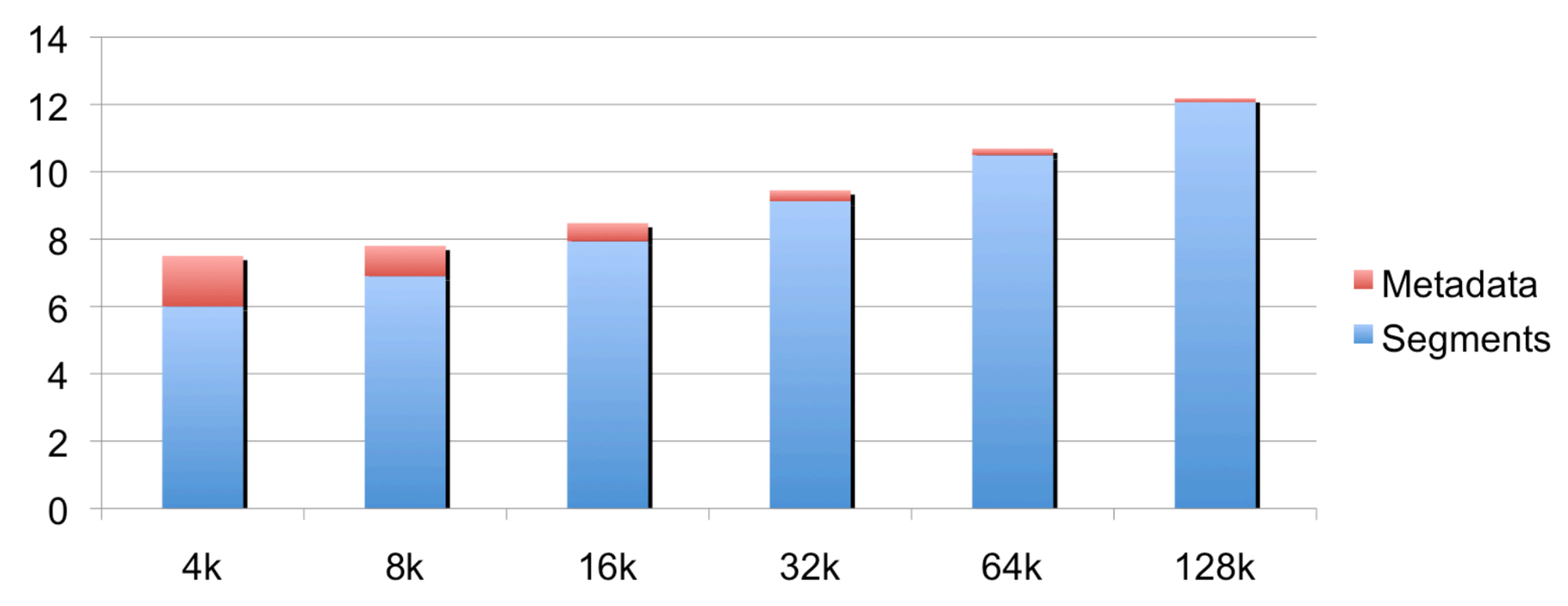
Reliability and Data Integrity

- Much more seriously here than primary storage
  - Lots of redundancy has been removed (30x!)
  - Last stop of data protection
- Data Domain's approach
  - Data is stored in a log of self-describing containers
  - Append only to avoid overwrites
  - Verifying containers and files (1st time and all the time)
  - Meta data reconstruction from containers
  - NVRAM logging for crash recovery and fault isolation
  - Self-correction from software RAID-6

High Deduplication Ratio

- High deduplication factor ➔ hardware cost
  - Smaller segments achieve higher compression ratios
  - Smaller segments imply higher ratio of fingerprint index size to physical disk storage
- Data Domain's approach
  - Understand the sweet spots of segment sizes
  - Multiple local compression algorithms

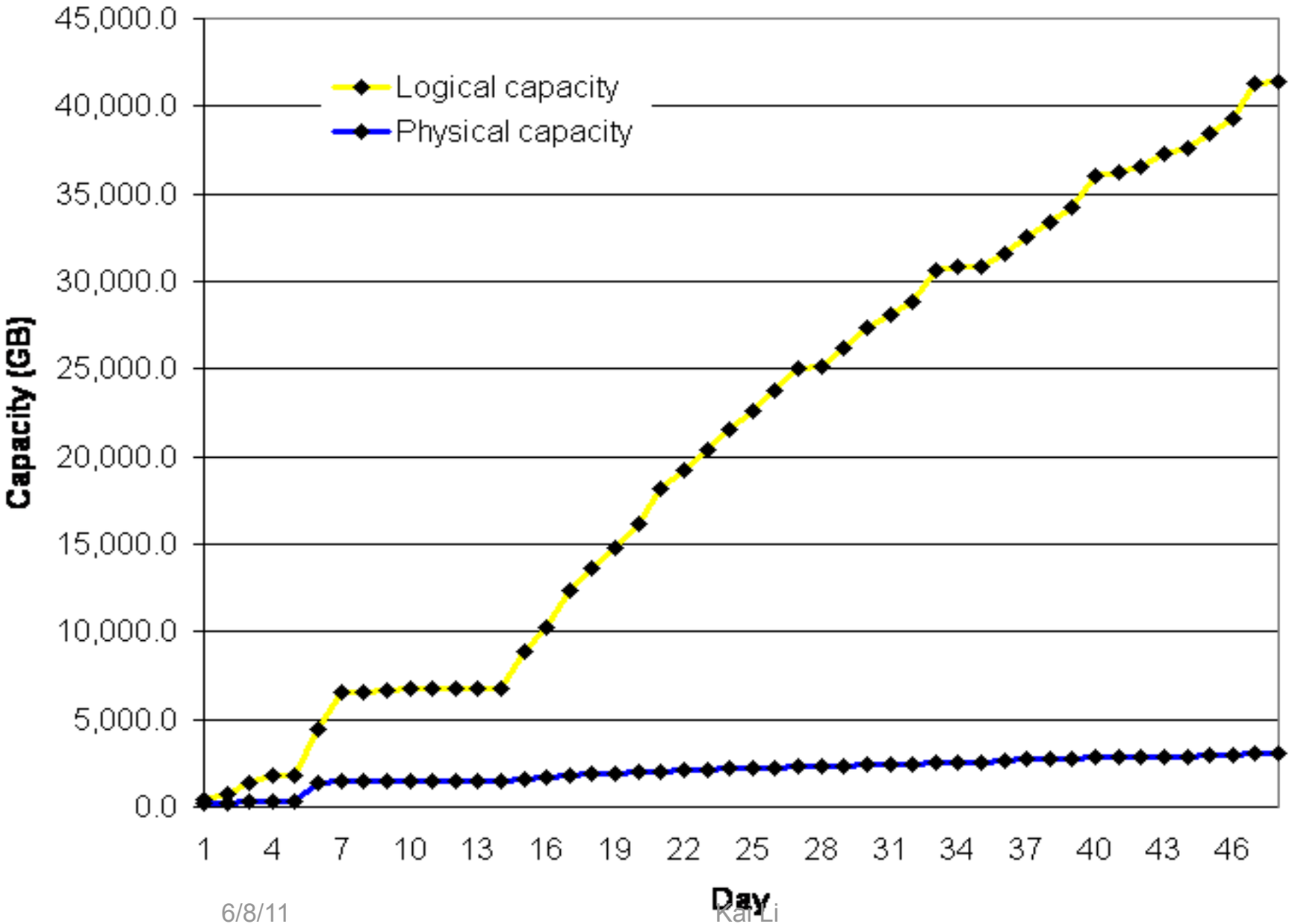
Segment Sizes for Backup Data



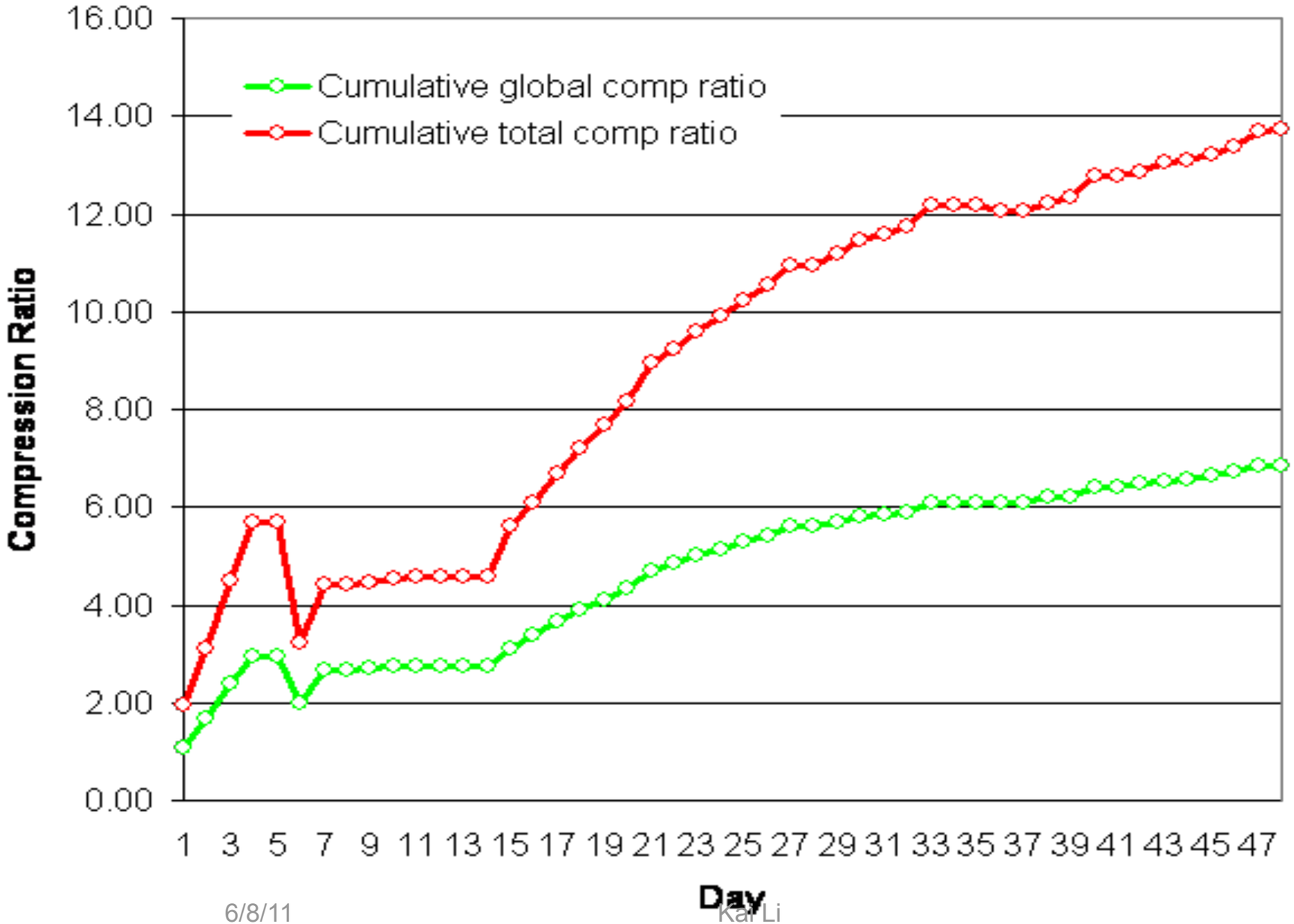
Rule of thumb: 2X segment size will

- increase space for unique segments by 15%
- decrease metadata by about 50%
- deduce disk I/Os for writes and reads

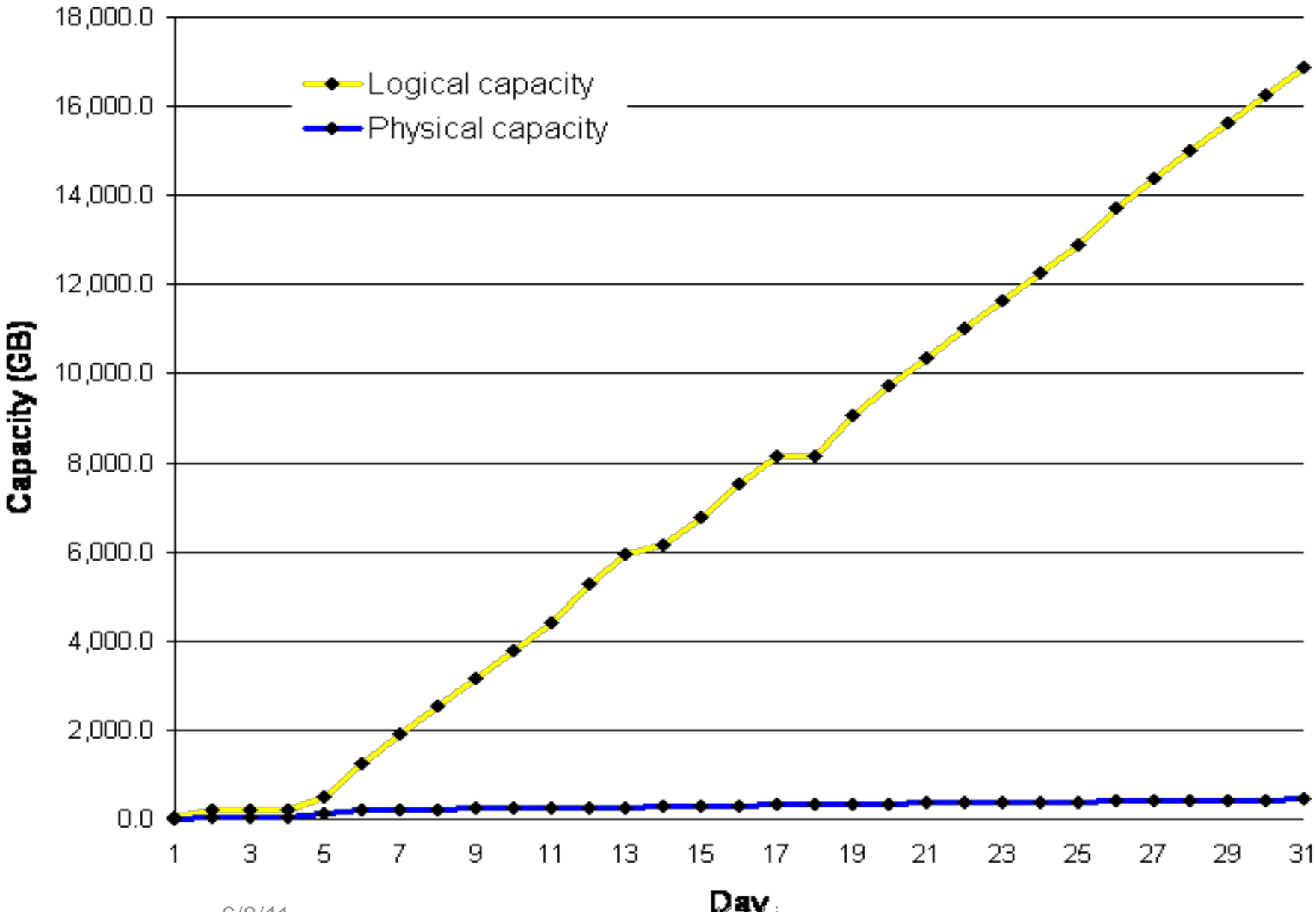
Real World Example at Datacenter A



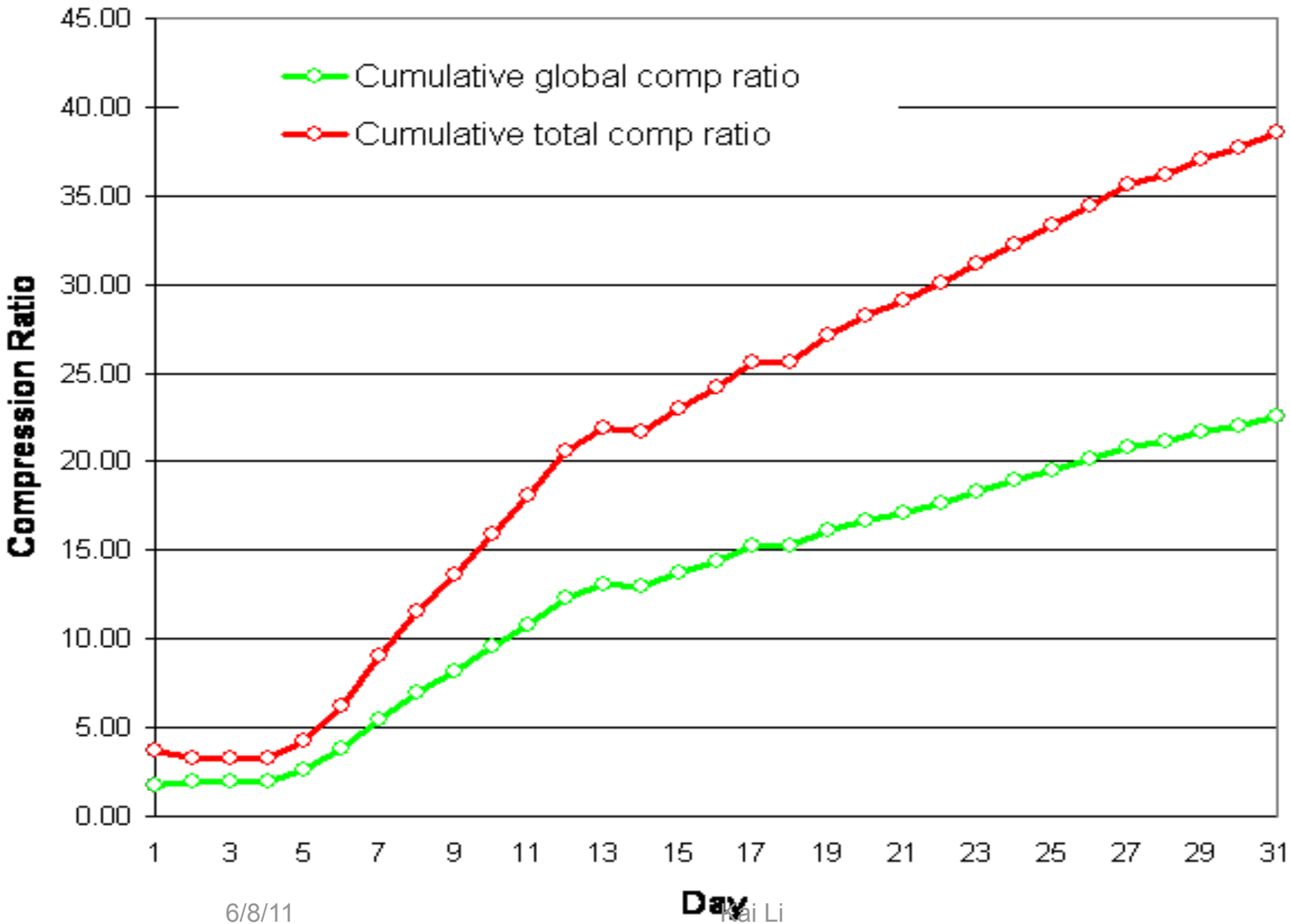
Real World Compression at Datacenter A



Real World Example at Datacenter B



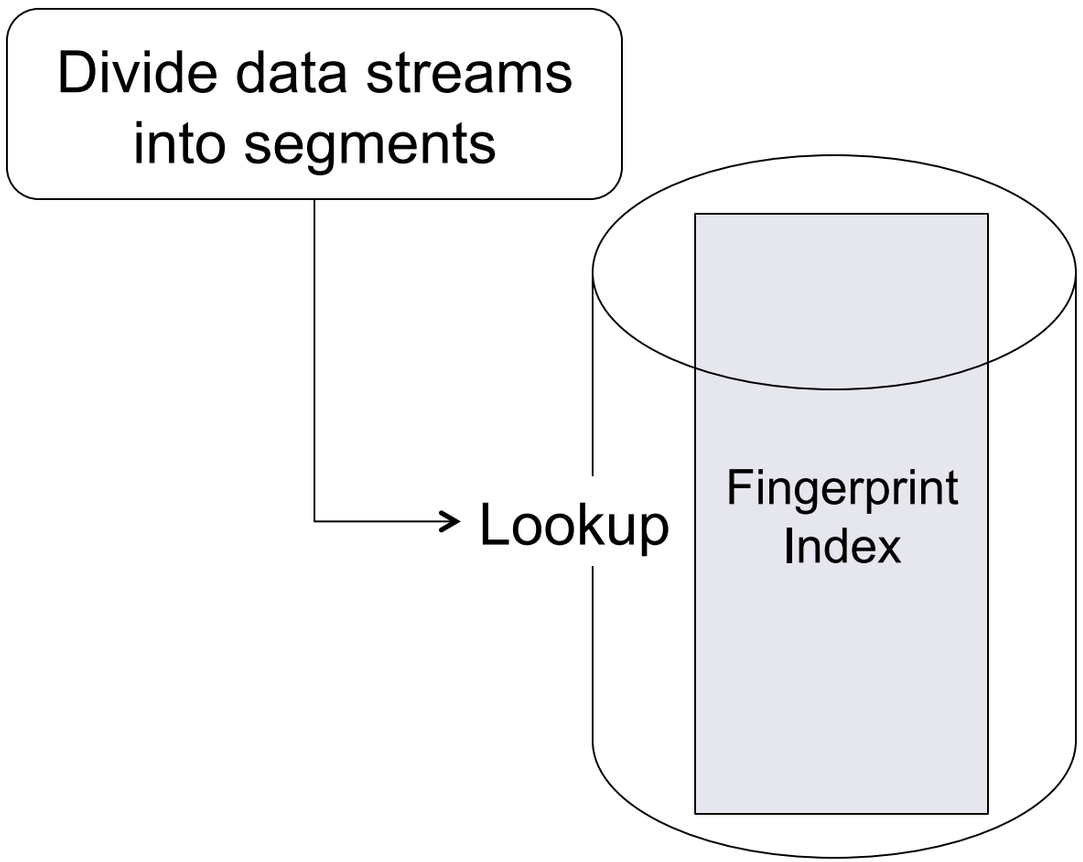
Real World Compression at Datacenter B



High Deduplication Throughput

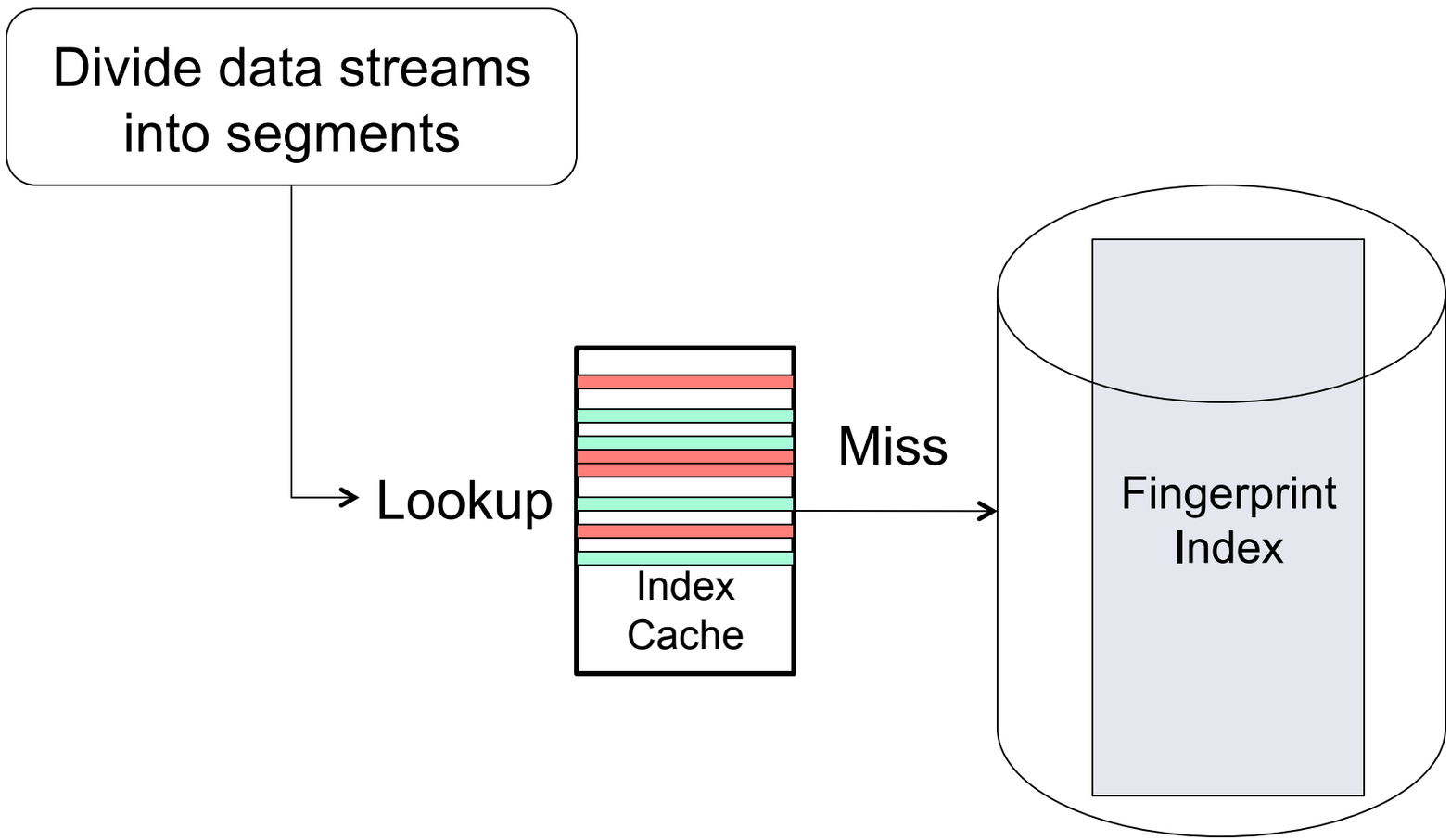
- Need to double every 18 months or faster
  - Data grows fast
  - Backup window time is fixed
  - Complete backups within the backup window time
- Data Domain's approach
  - **A sophisticated cache for index**
  - Several techniques to reduce memory and CPU requirements
  - Bet on multicore processors

Why Challenging?



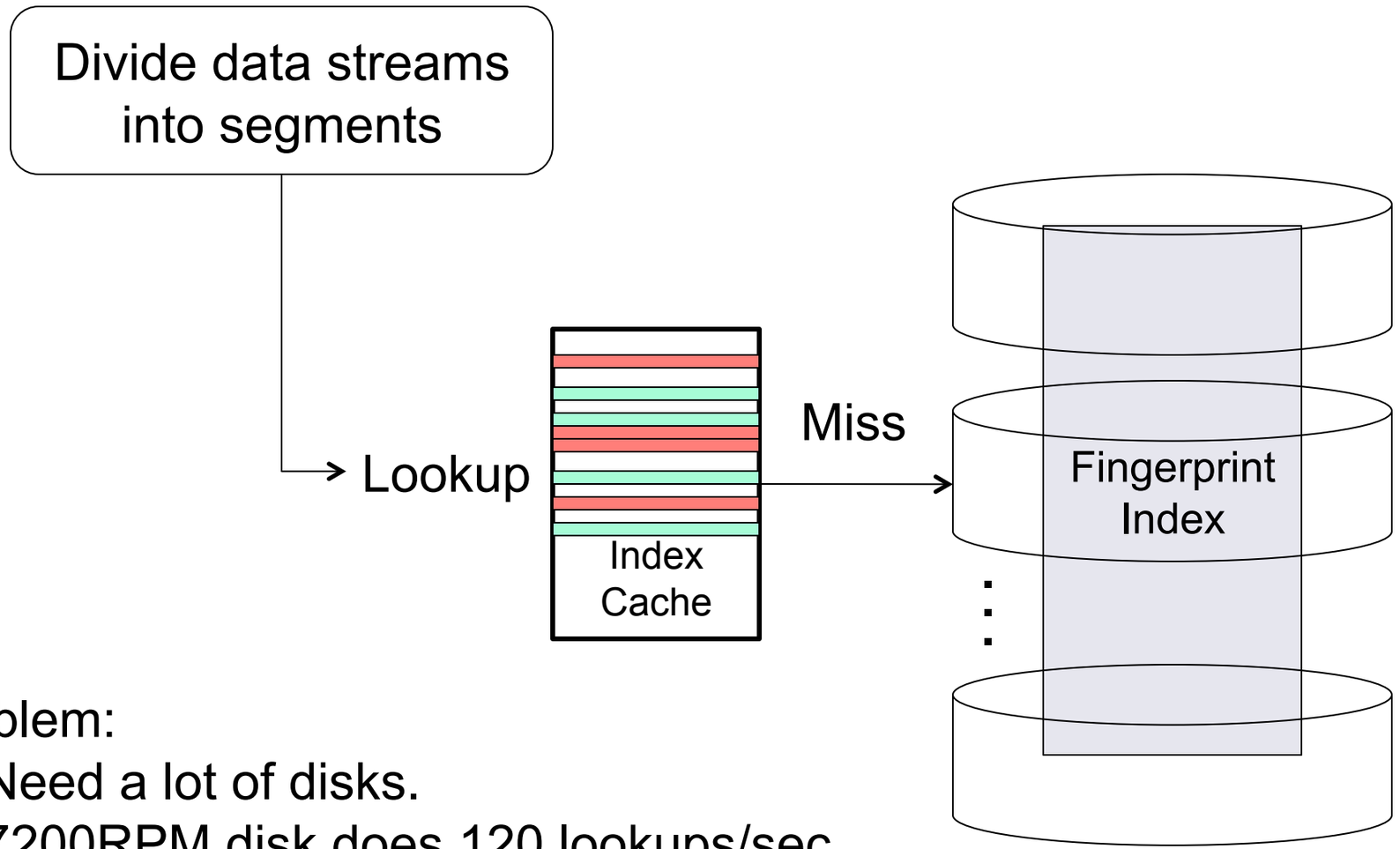
Index size for 80TB w/ 8KB segments
= (80TB/8KB) * 20B
= **200GB!**

Caching?



Problem: **No locality**.

Parallel Index Need Many Disks [Venti02]



Problem:

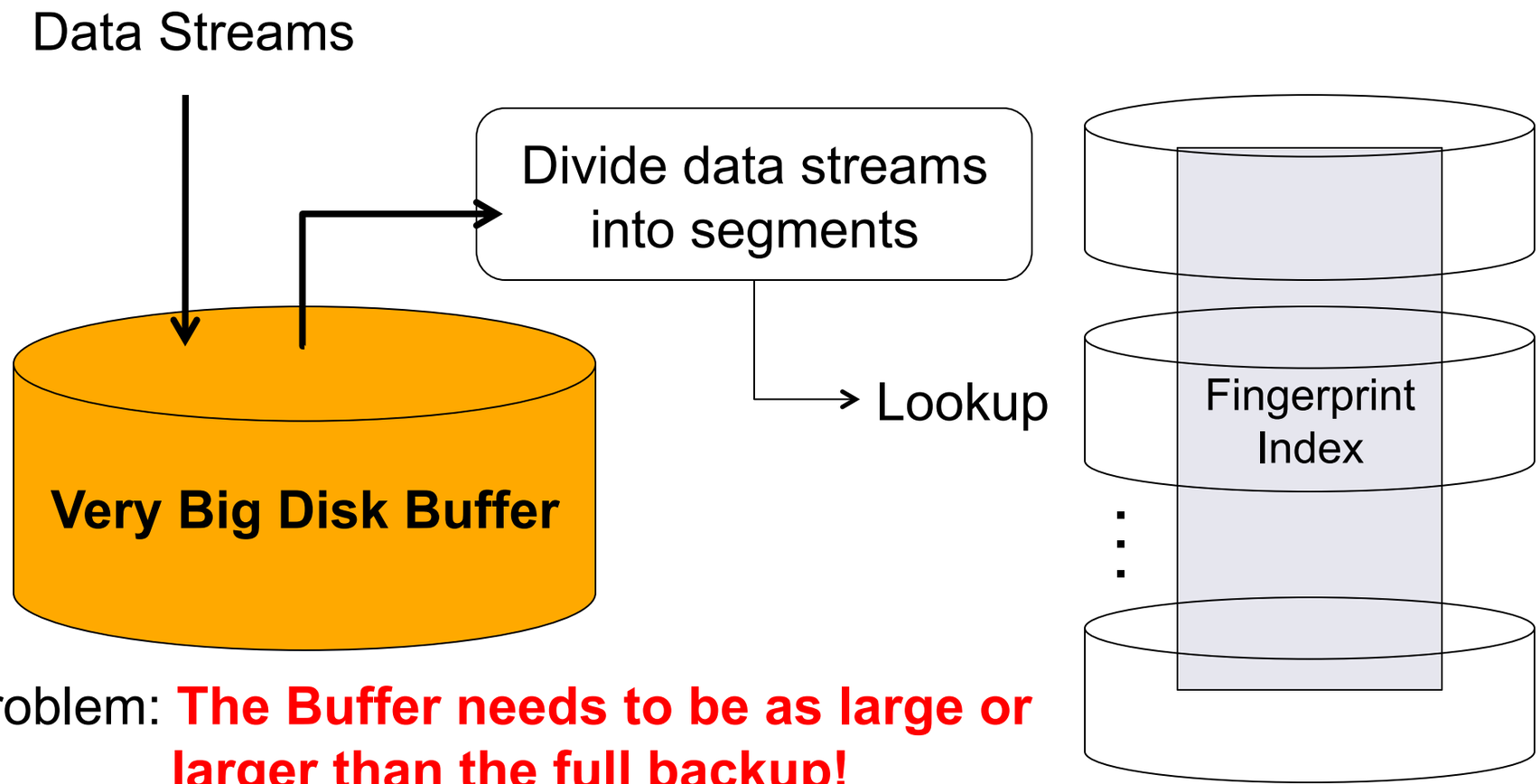
Need a lot of disks.

7200RPM disk does 120 lookups/sec.

1MB/sec with 8KB segment per disk

1GB/sec needs 1,000 disks!

Staging Needs More Disks



Problem: **The Buffer needs to be as large or larger than the full backup!**

Big delay and may still never catch up

A Combination of Techniques

- Layout data on disk with “duplicate locality”
- A sophisticated cache for the fingerprint index
  - Summary data structure for new data
  - “locality-preserved caching” for old data
- Parallelized software systems to leverage multicore processors

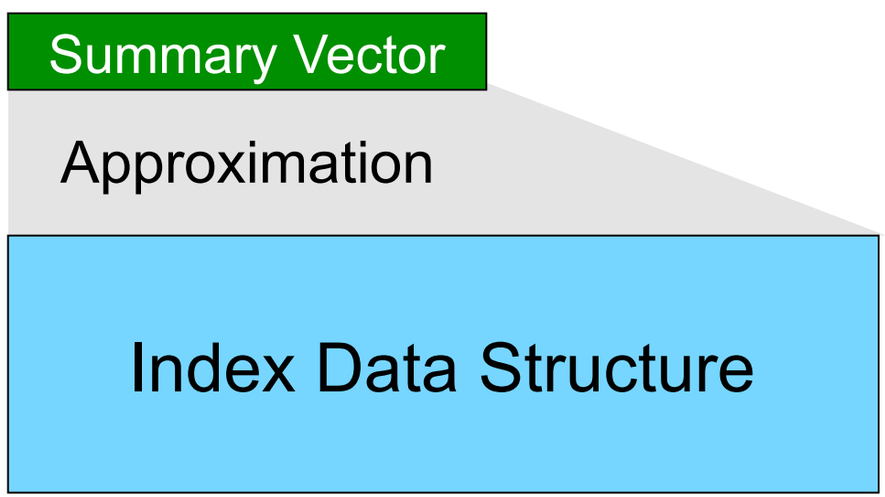
Benjamin Zhu, Kai Li and Hugo Patterson. Avoiding the Disk Bottleneck in the Data Domain Deduplication File System. In Proceedings of The 6th USENIX Conference on File and Storage Technologies (FAST’08). February 2008

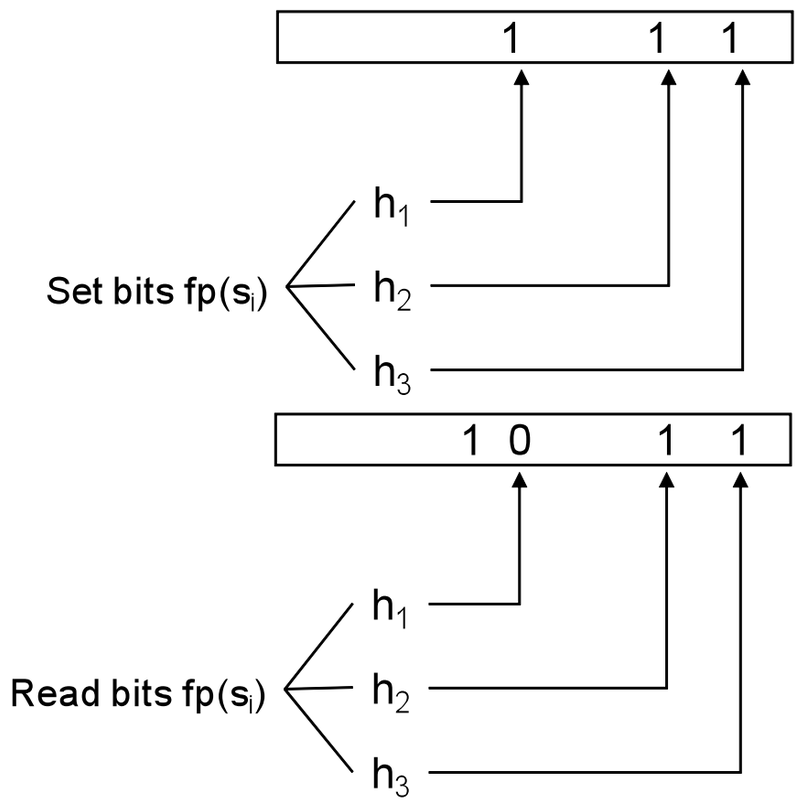
Summary Vector

Goal: Use minimal memory to test for new data

⇒ Summarize what segments have been stored, with Bloom filter (Bloom'70) in RAM

⇒ If Summary Vector says no, it's new segment

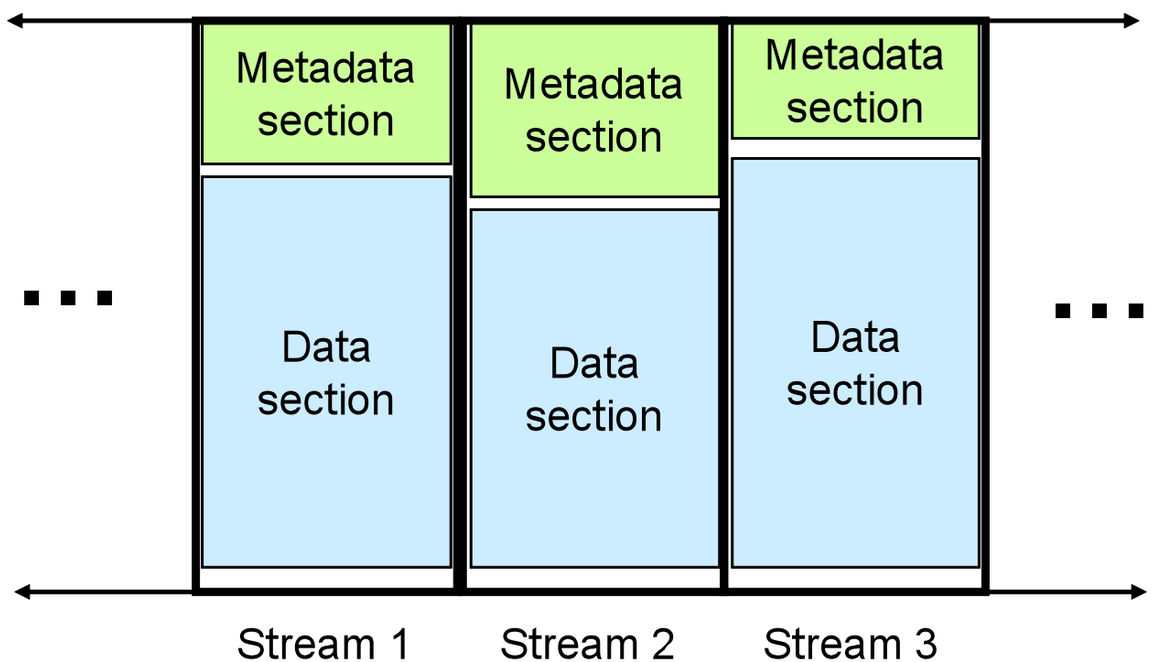




Stream Informed Segment Layout

Goal: Capture “duplicate locality” on disk

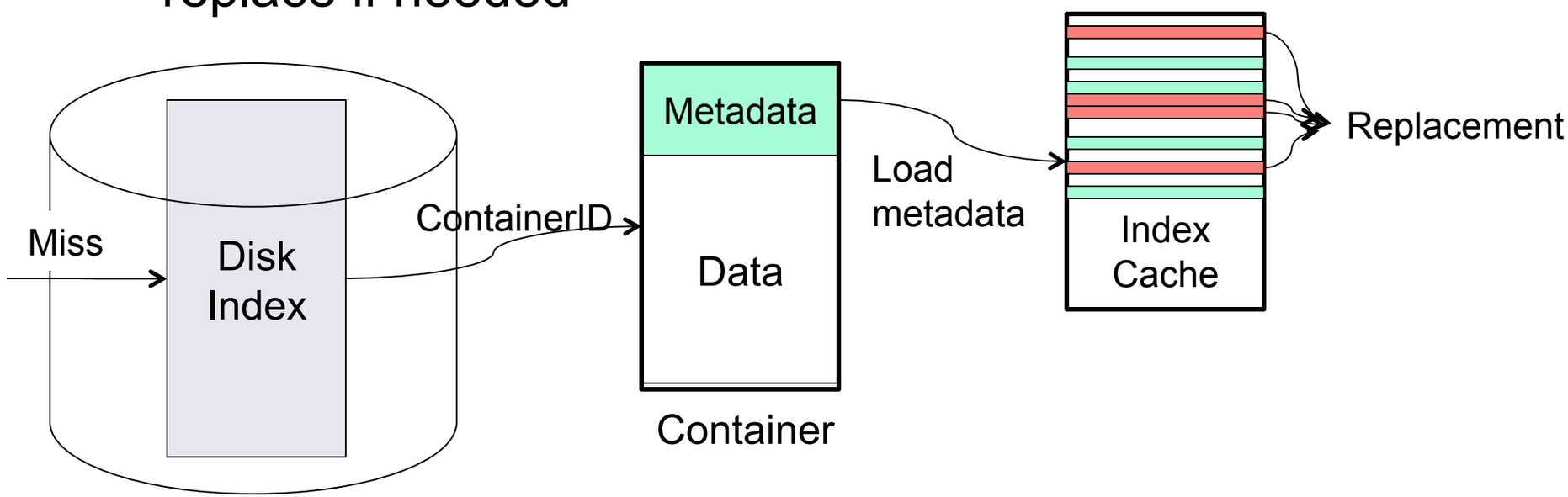
- Segments from the same stream are stored in the same “containers”
- Metadata (index data) are also in the containers



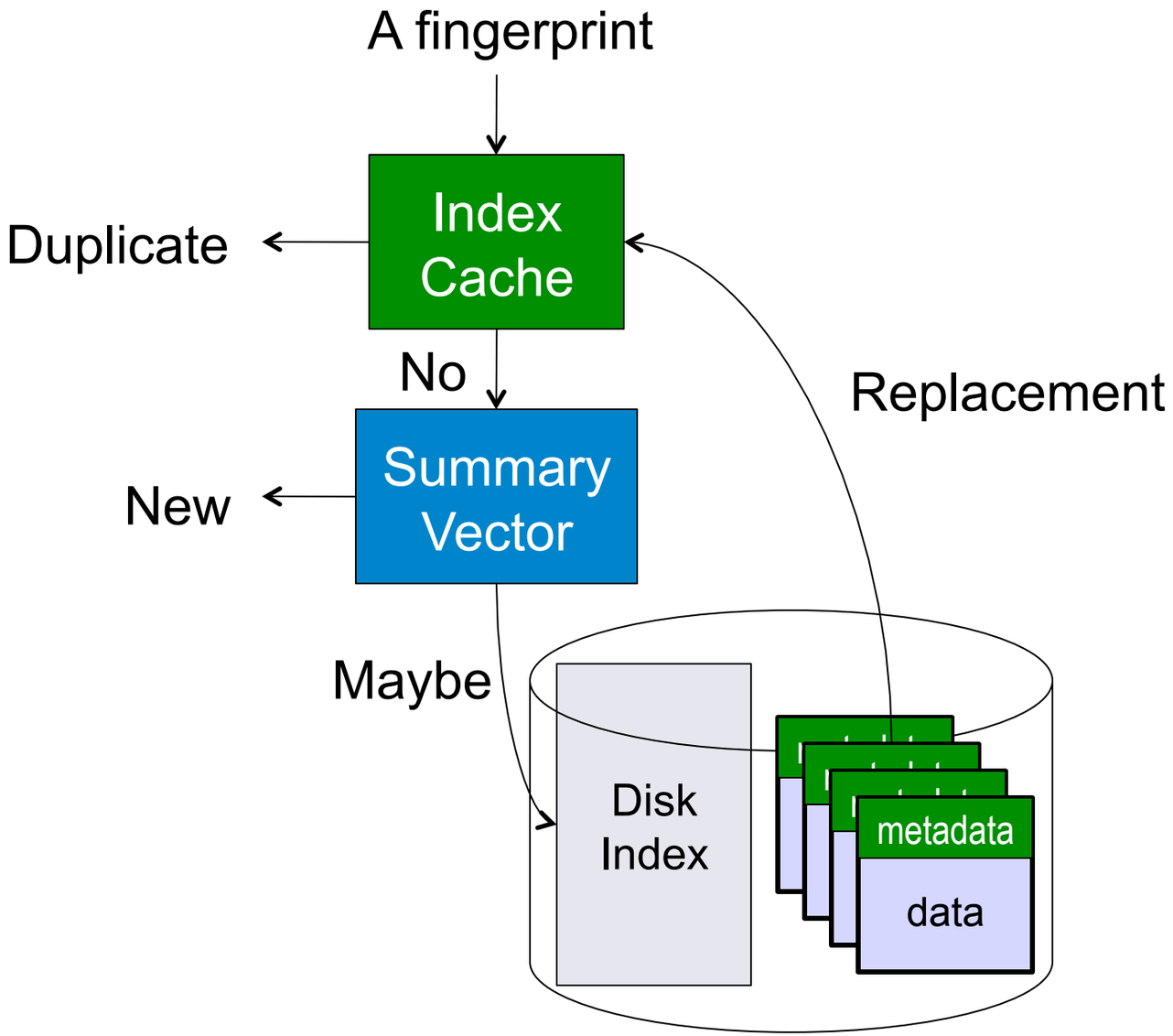
Locality Preserved Caching (LPC)

Goal: Maintain “duplicate locality” in the cache

- Disk Index has all <fingerprint, containerID> pairs
- Index Cache caches a subset of such pairs
- On a miss, lookup Disk Index to find containerID
- Load the metadata of a container into Index Cache, replace if needed



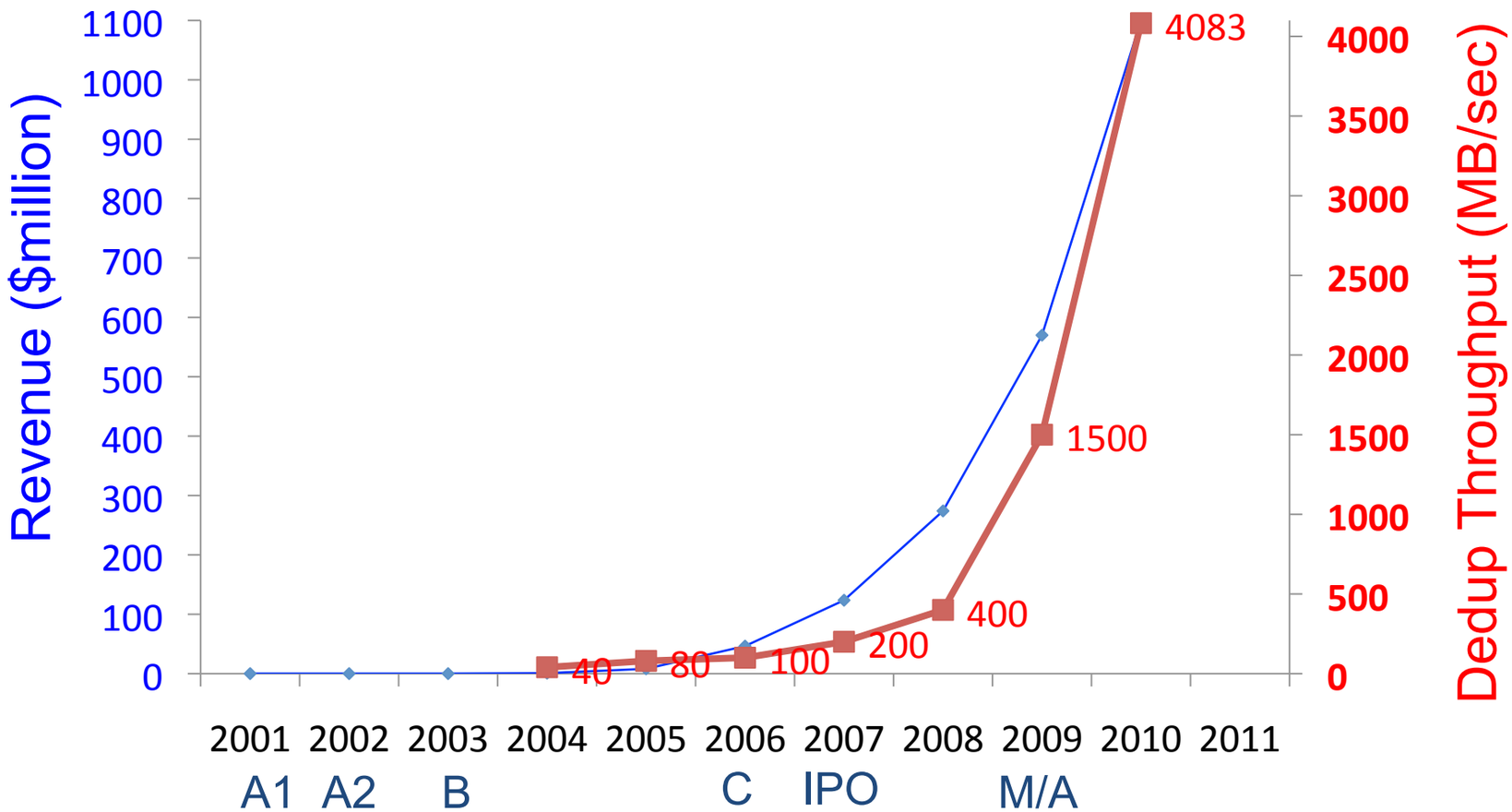
Putting Them Together



Disk I/O Reduction Results

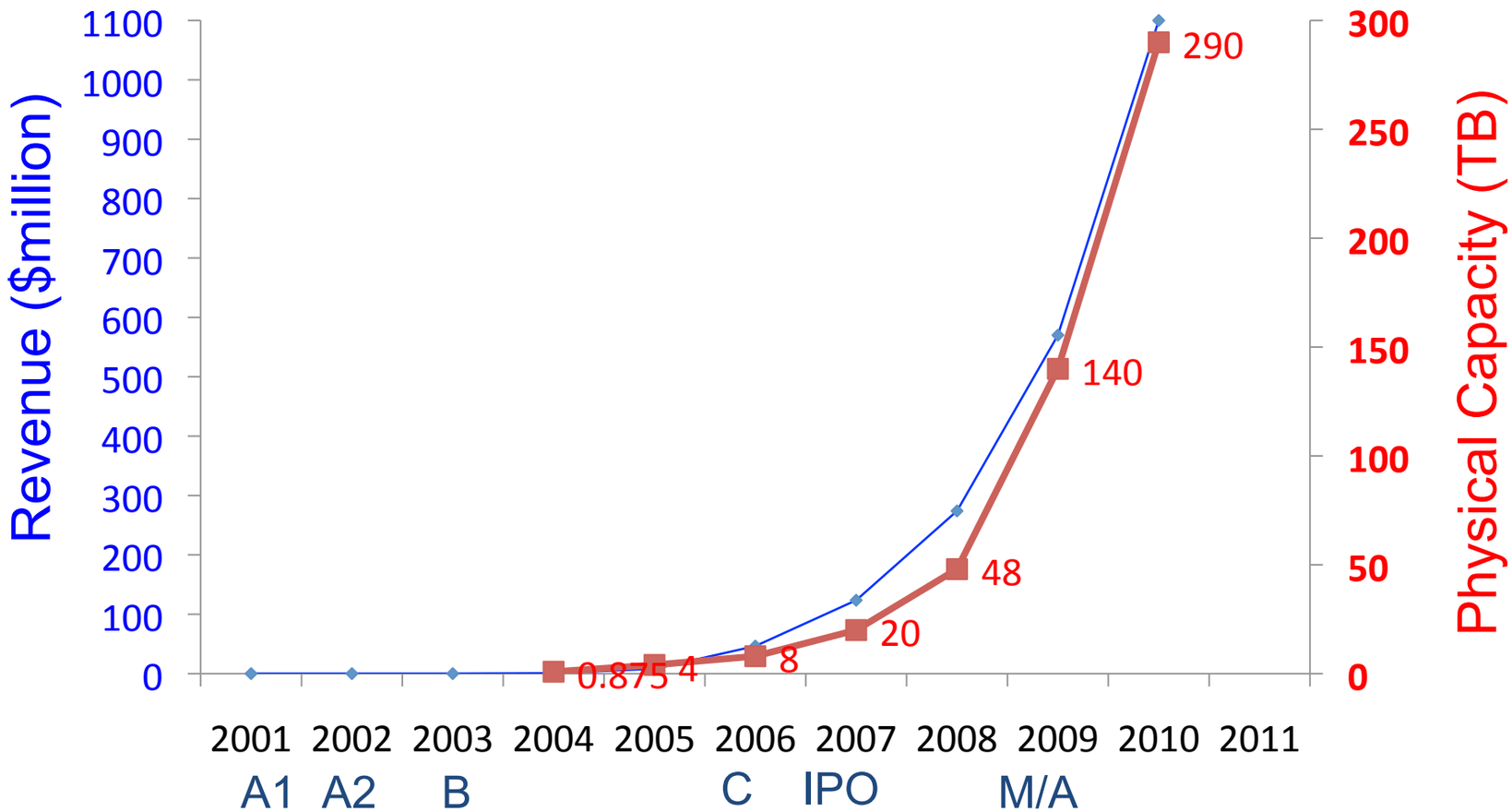
	Exchange data (2.56TB) *135-daily full backups*		Engineering data (2.39TB) *100-day daily inc, weekly full*	
	# disk I/Os	% of total	# disk I/Os	% of total
No summary, No SISL/LPC	328,613,503	100.00%	318,236,712	100.00%
Summary only	274,364,788	83.49%	259,135,171	81.43%
SISL/LPC only	57,725,844	17.57%	60,358,875	18.97%
Summary & SISL/LPC	**3,477,129**	**1.06%**	**1,257,316**	**0.40%**

Revenue vs. Dedup Throughput



Dedup throughput improved by ~100X in 6 years

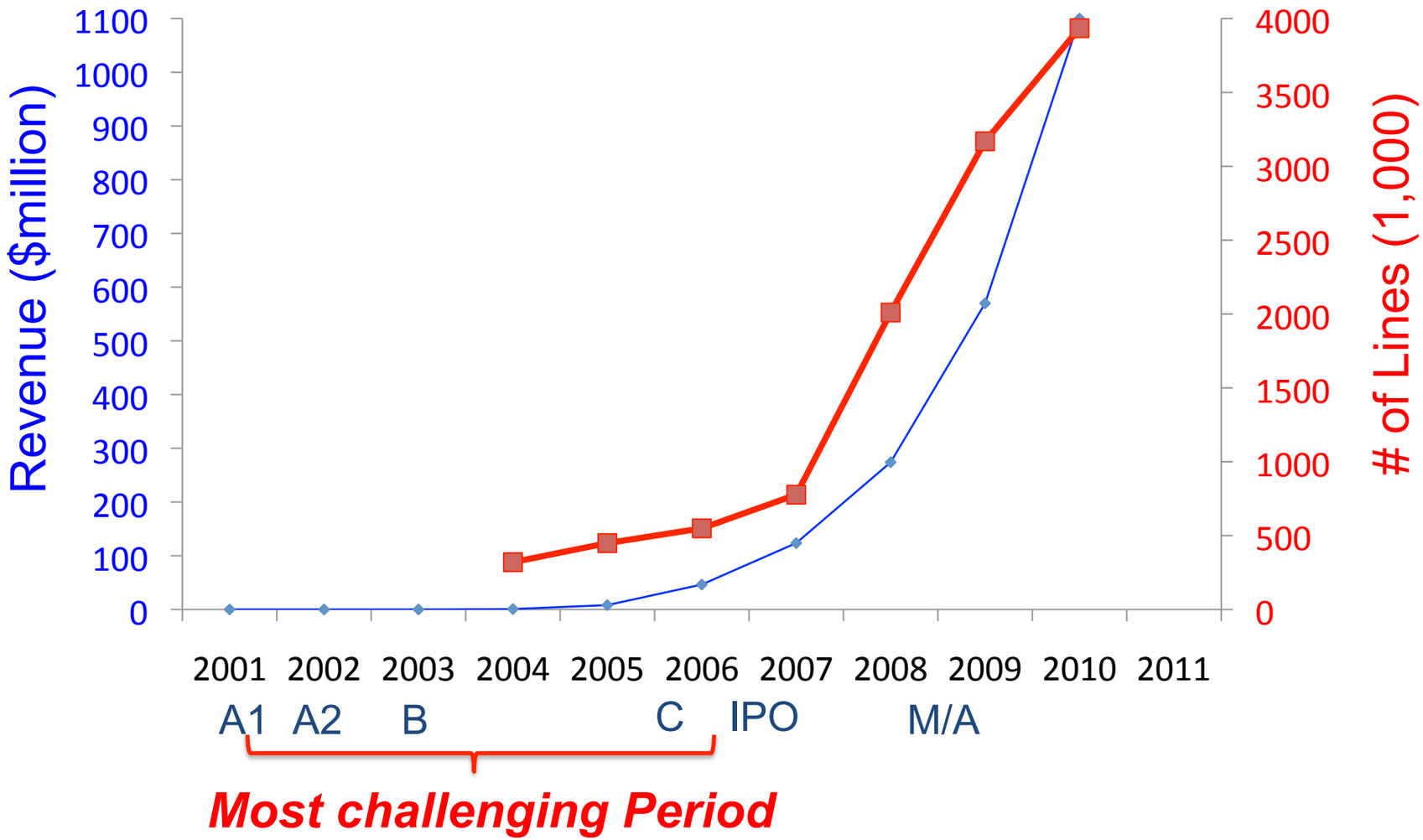
Revenue vs. Physical Capacity



Usable physical space increased by ~330X in 6 years

Engineering Challenges

Revenue vs. Software Complexity



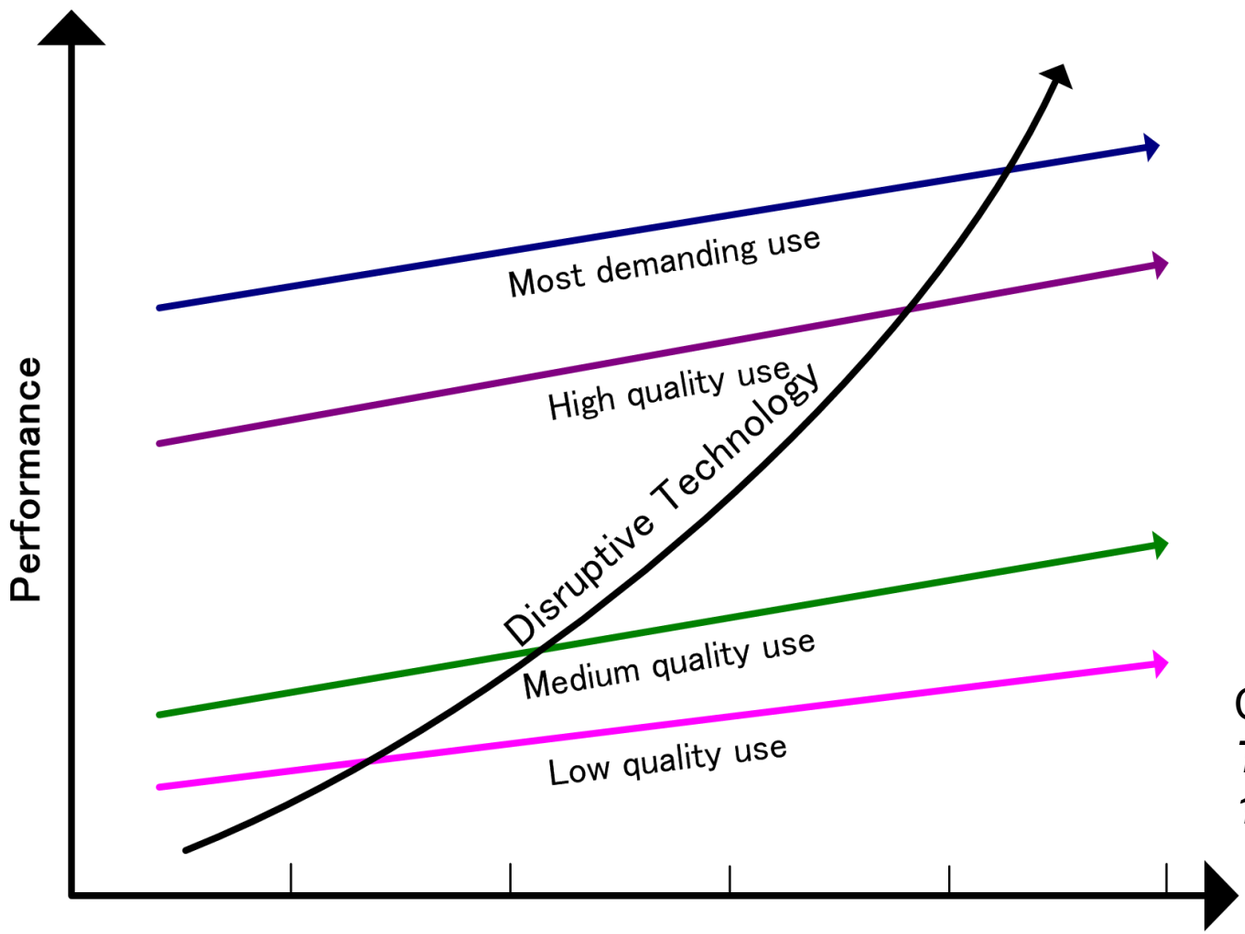
Real Challenge Is Roadmap

- First product definition
  - Customers are suspicious about new storage products
  - Must provide enough value propositions
  - Must balance engineering efforts
- Following releases
  - Provide right value propositions
  - Competitive in market place
  - Deliver quality release on time

Disruptive Starts at Low Quality

Disrupt an existing market

- Improve a service in ways that the market does not expect



Clayton M. Christensen,
The Innovator's Dilemma.
1997

What's Special about Storage Product

- Market entry barrier is higher for data centers
- There is a magic number for mileage
- Many investors are impatient

Future…

Future Impact and Challenges

- Nearline storage
  - Handle many large and small files
  - Relatively low latency for small I/Os
- Archival storage
  - Need locking, shredding, long term retention, …
  - Further improve compression ratios
- Primary storage
  - Reduce the cost of Flash without sacrificing performance
  - Rethink the storage eco-system for data centers
- Cloud storage
  - What are the right building blocks

Thank You