COMPUTERWORLD SNIA

SNW

Transforming the Information Infrastructure: Build, Manage, Optimize.

FALL 2011

Journey to a Picture Perfect Data Storage System

Chris Gladwin

CEO, Cleversafe

Client Profile

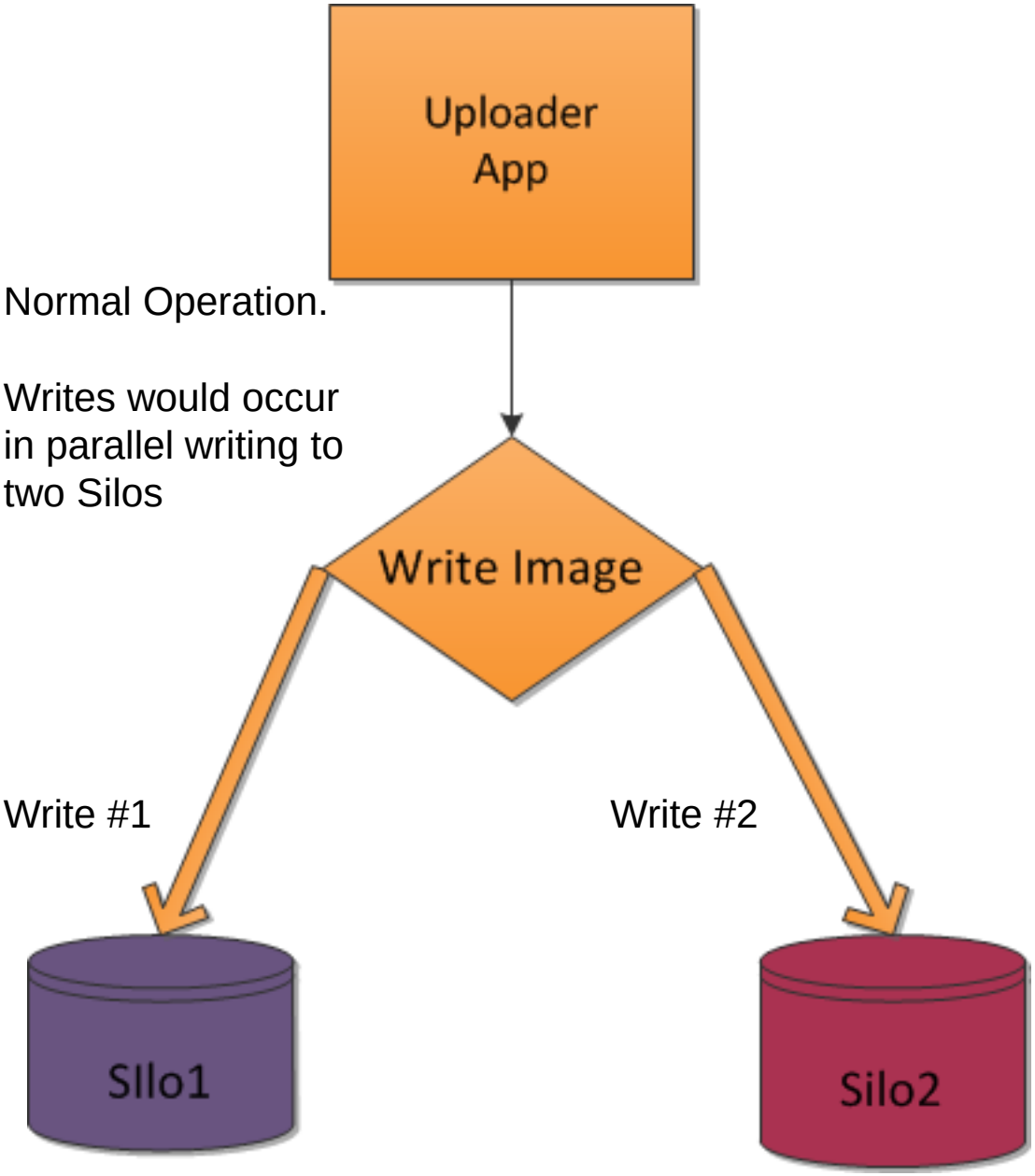
- Free UNLIMITED online photo storage – FOREVER
- 19 petabytes and growing at 40% annually
- 25% image size growth annually

Petabyte Scale Challenges

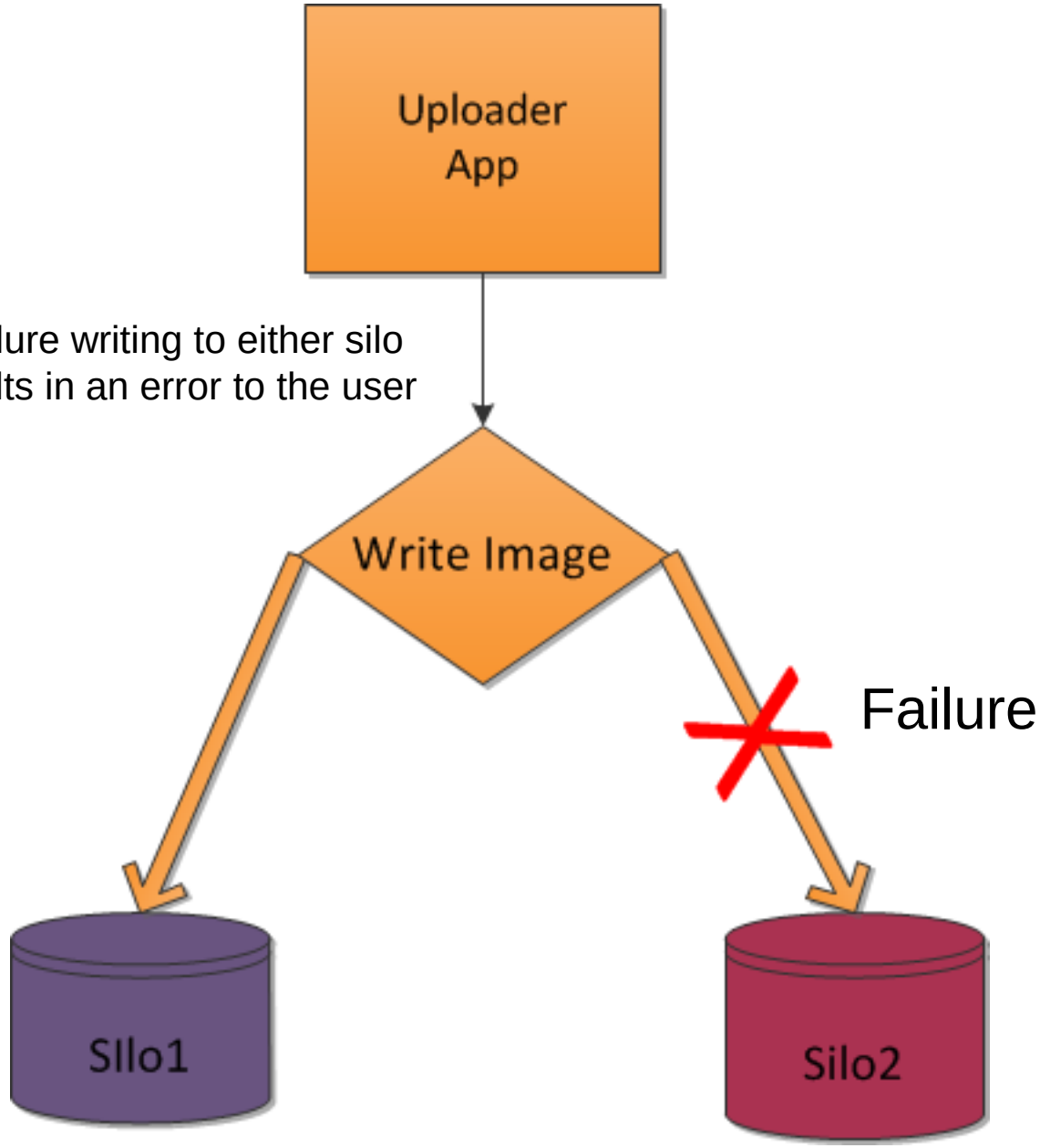
- When does RAID stop scaling?
  - Near catastrophe
    - 2 PB array failure, lost 172 drives
    - No data lost, but 3 days to calculate parity, 3 weeks to return to dual parity
  - Availability/integrity issues
    - Bad user experience/site downtime = revenue loss
    - System Admins constantly firefighting
    - Poor internal perception of Storage team
- Asked to drastically lower costs by 66%
  - Already managing 19PB with 3 System Admins
  - Already had industry leading pricing with multiple vendors

Normal Operation.

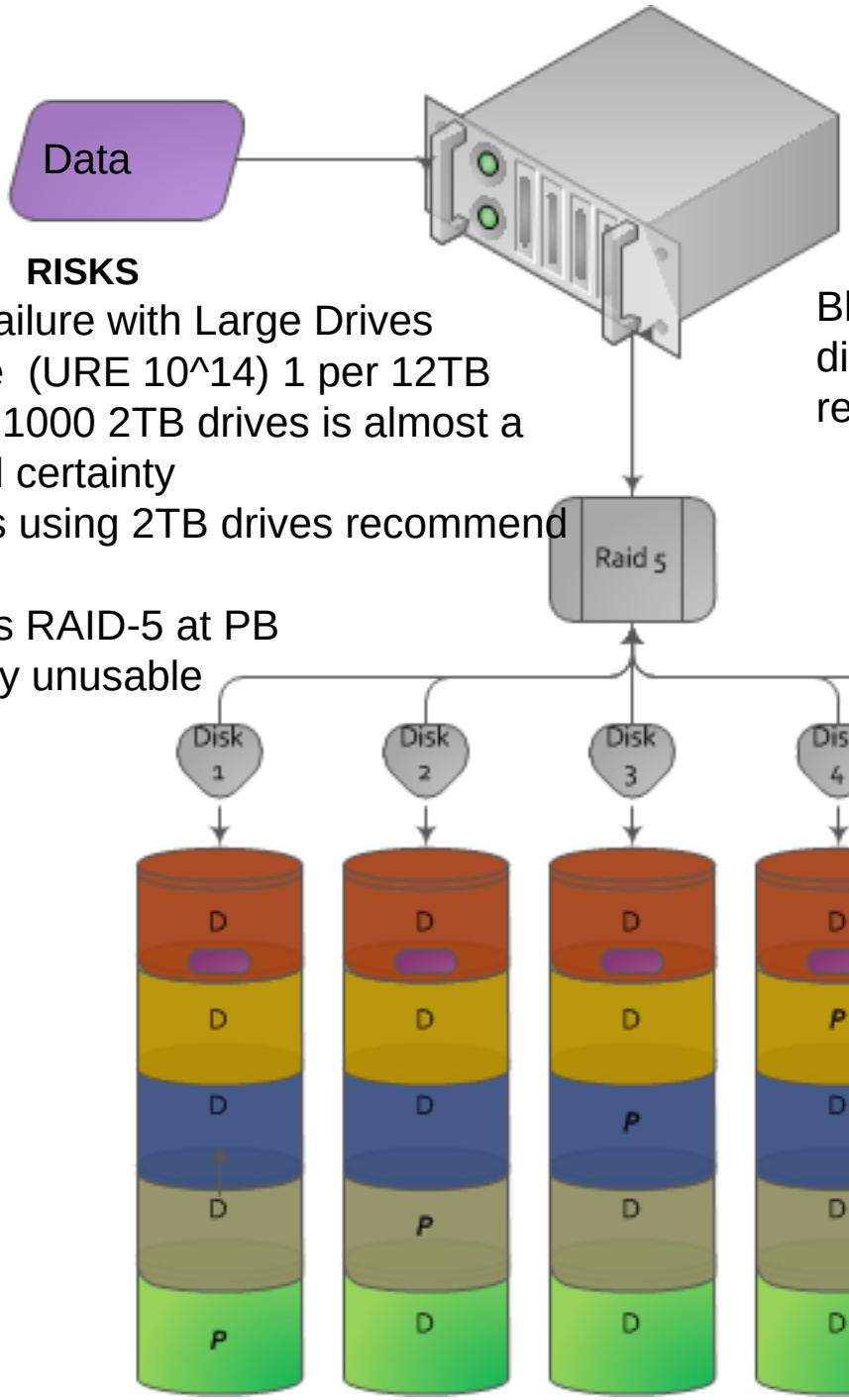
Writes would occur in parallel writing to two Silos



A failure writing to either silo results in an error to the user



RAID-5



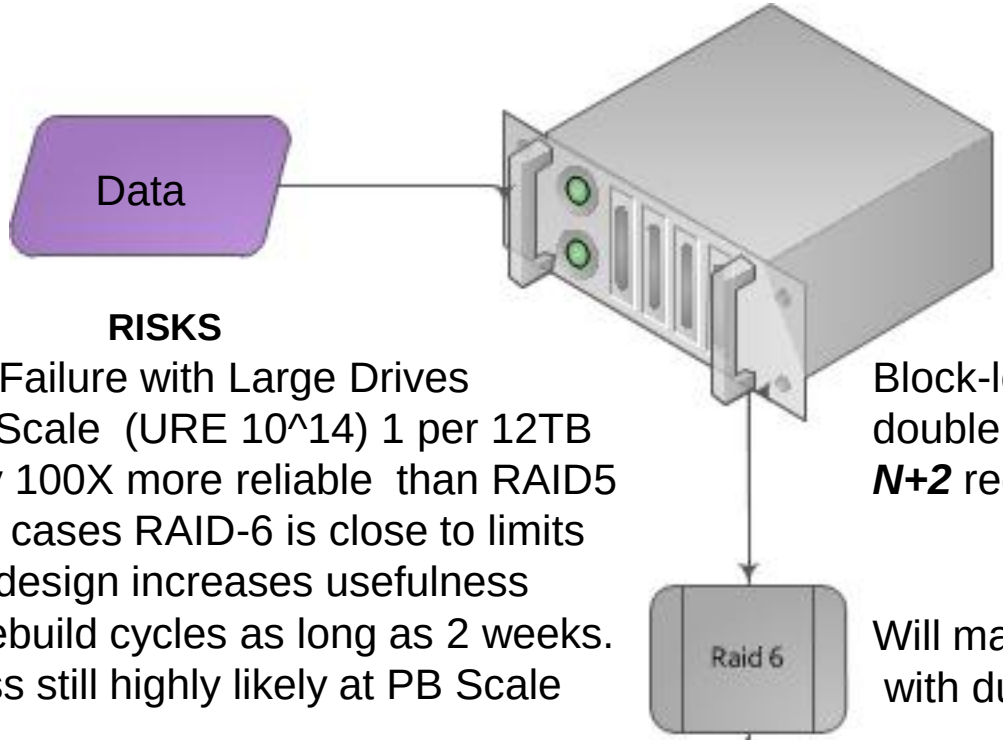
RISKS

- Secondary Failure with Large Drives
- BER at Scale (URE 10^14) 1 per 12TB
- Data Loss at 1000 2TB drives is almost a mathematical certainty
- Most vendors using 2TB drives recommend RAID-6
- In most cases RAID-5 at PB scale is already unusable

Block-level striping with distributed parity ***N+1*** redundancy

Will maintain data integrity with a single disk failure

RAID-6

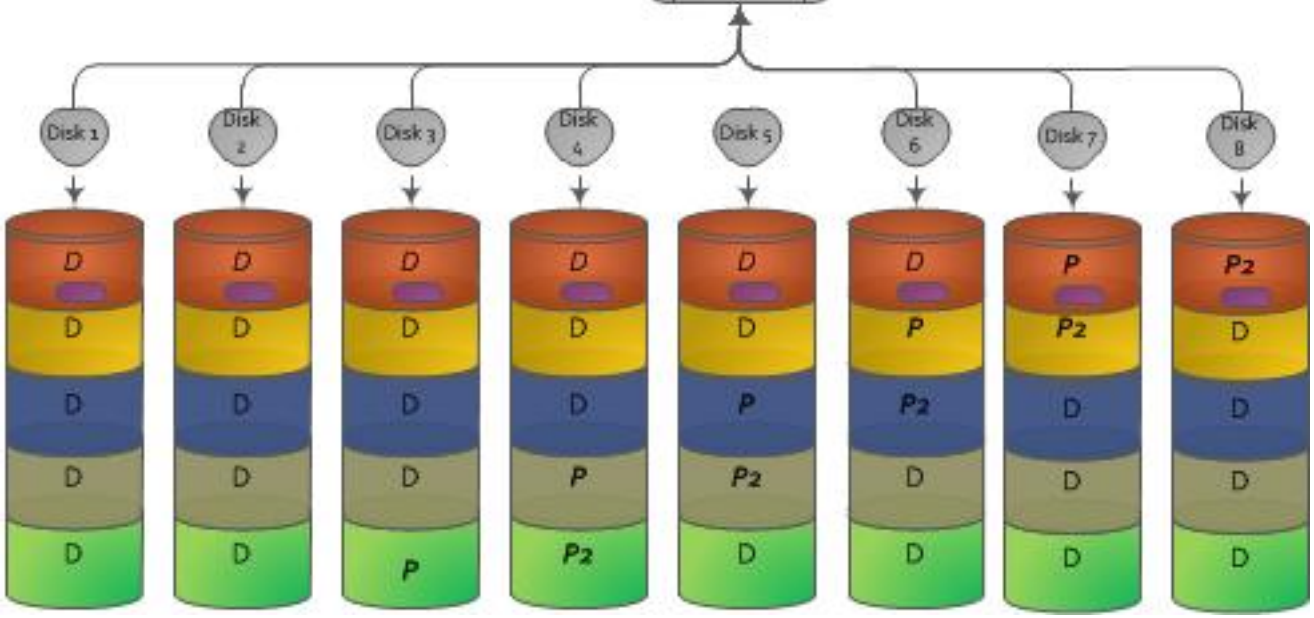


RISKS

- Tertiary Failure with Large Drives
- BER at Scale (URE 10^14) 1 per 12TB
- Roughly 100X more reliable than RAID5
- In many cases RAID-6 is close to limits
- Vendor design increases usefulness
- Long Rebuild cycles as long as 2 weeks.
- Data loss still highly likely at PB Scale

Block-level striping with double distributed parity ***N+2*** redundancy

Will maintain data integrity with dual disk failures

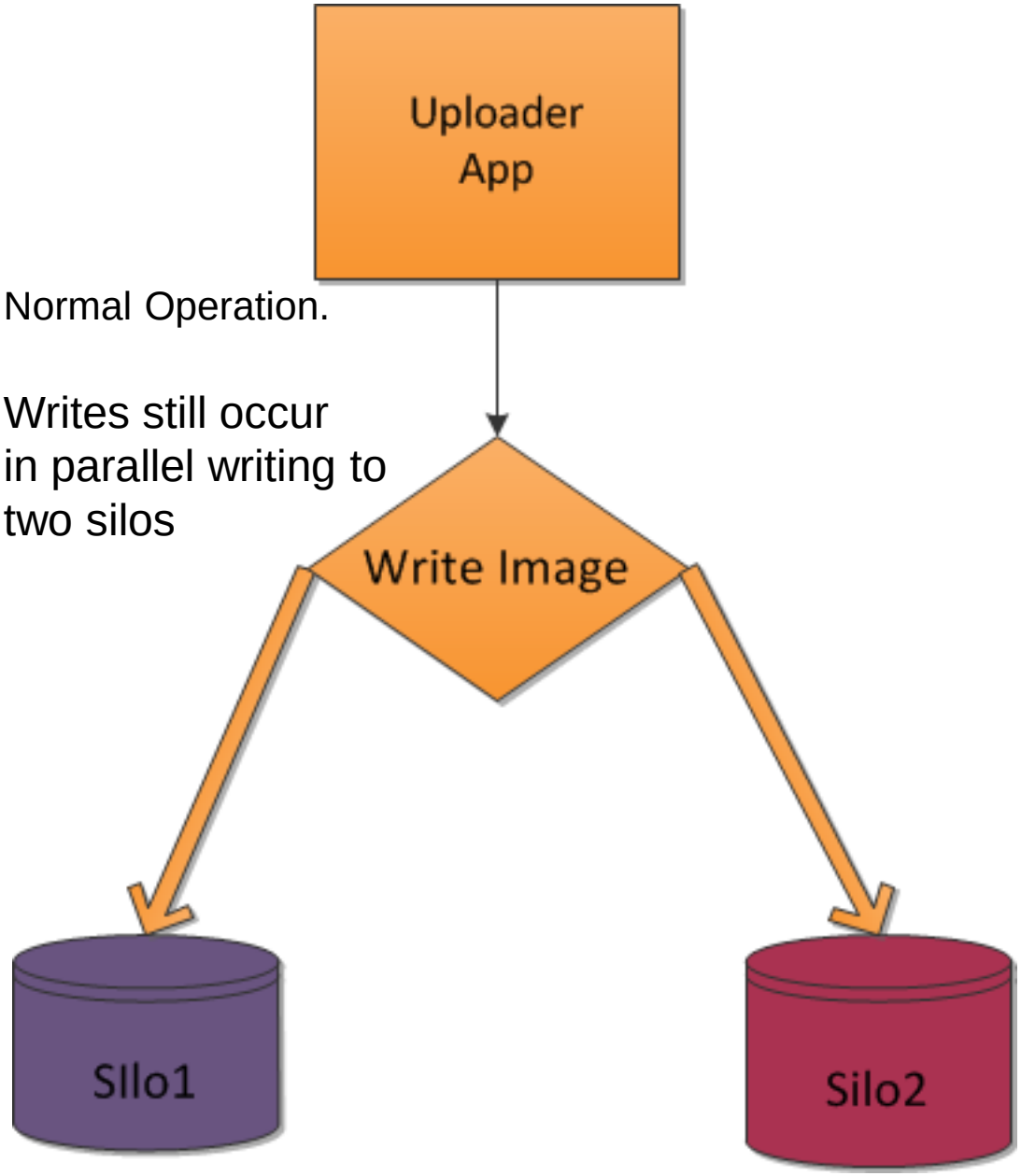


Objectives

- Graceful failure and recovery
  - Greater site availability
  - Hardware errors transparent to users
  - No executive notifications about storage-related outages
  - Eliminate the need for potentially dangerous site-impacting heroics from System Admins
  - No impact to the bottom line
- Future-proof storage
  - Architecture to support larger, slower drives
  - Sustain data integrity with long rebuild times
  - Lower cost
  - Commoditize storage
  - Roughly 100X integrity per extra bit of Parity

Normal Operation.

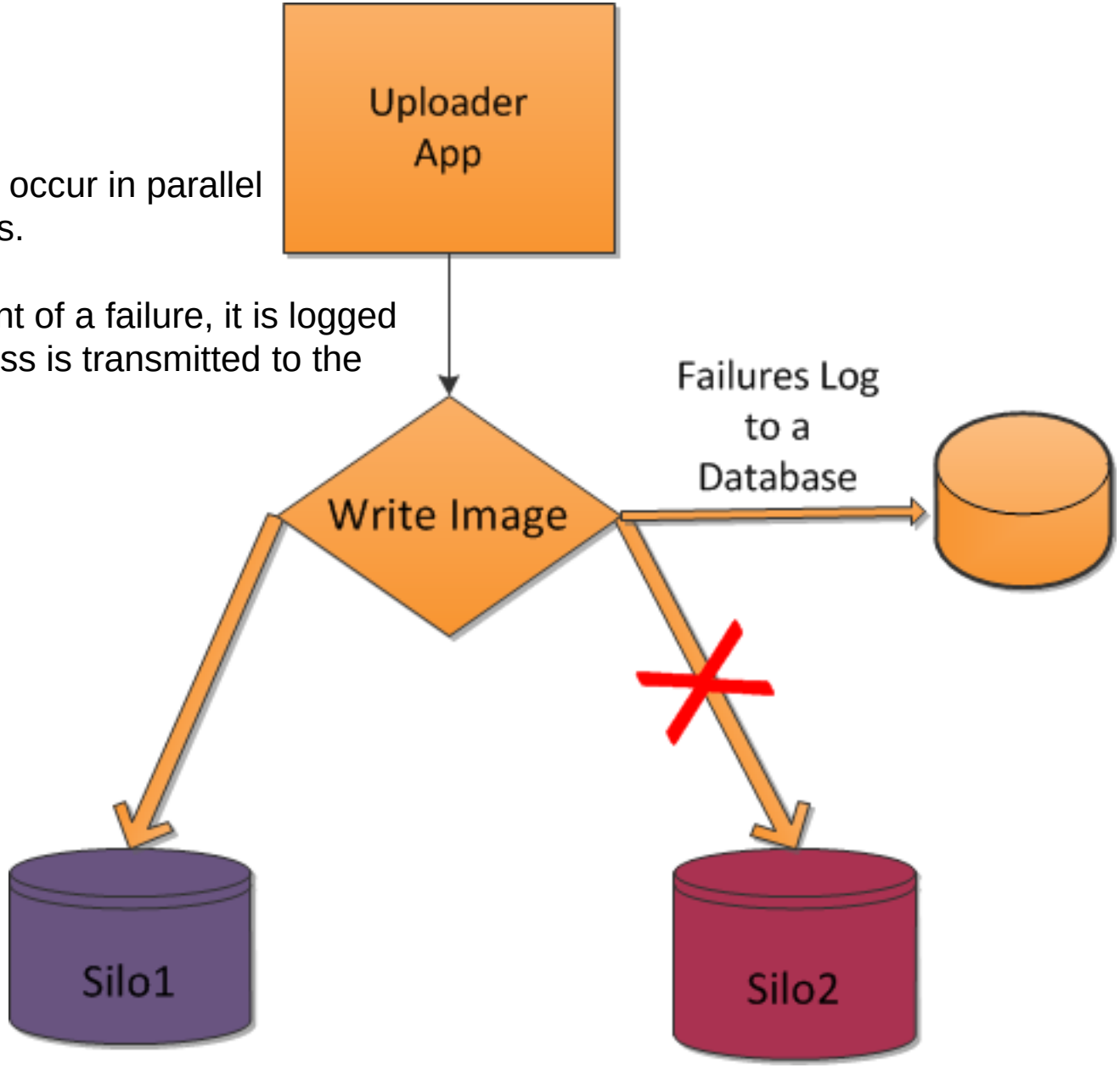
Writes still occur in parallel writing to two silos



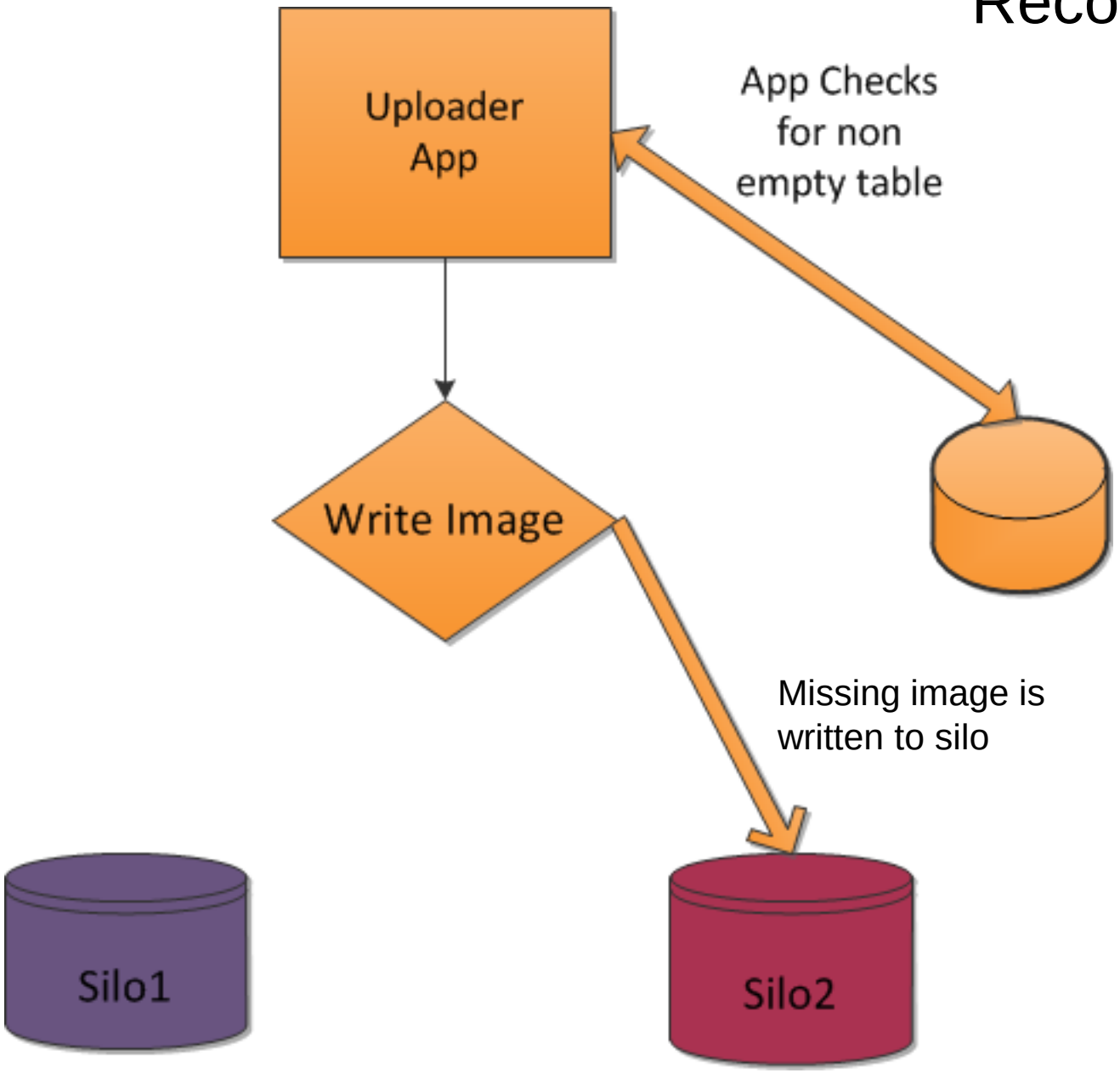
Graceful Failure

Writes still occur in parallel to two silos.

In the event of a failure, it is logged and success is transmitted to the user.



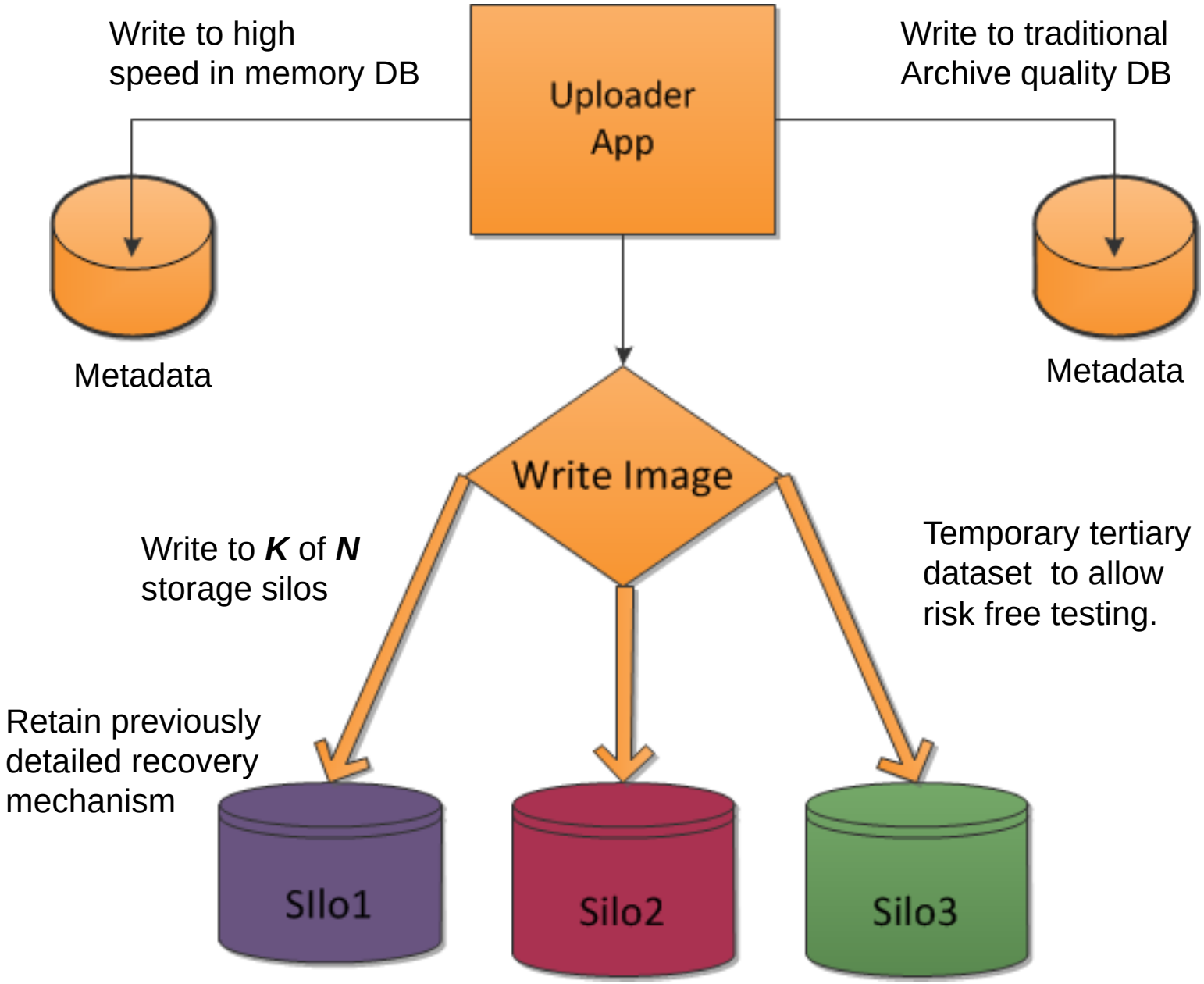
Recovery


Uploader App
App Checks for non empty table
Write Image
Missing image is written to silo
Silo1
Silo2

Prepare for new Technology

The Y Adapter

- Storage flexibility
- Maintain existing archiving standards
- Enable web scale metadata repositories
- Enable new storage technologies


Write to high
speed in memory DB
Uploader
App
Write to traditional
Archive quality DB
Metadata
Metadata
Write Image
Write to *K* of *N*
storage silos
Temporary tertiary
dataset to allow
risk free testing.
Retain previously
detailed recovery
mechanism
SIlo1
Silo2
Silo3

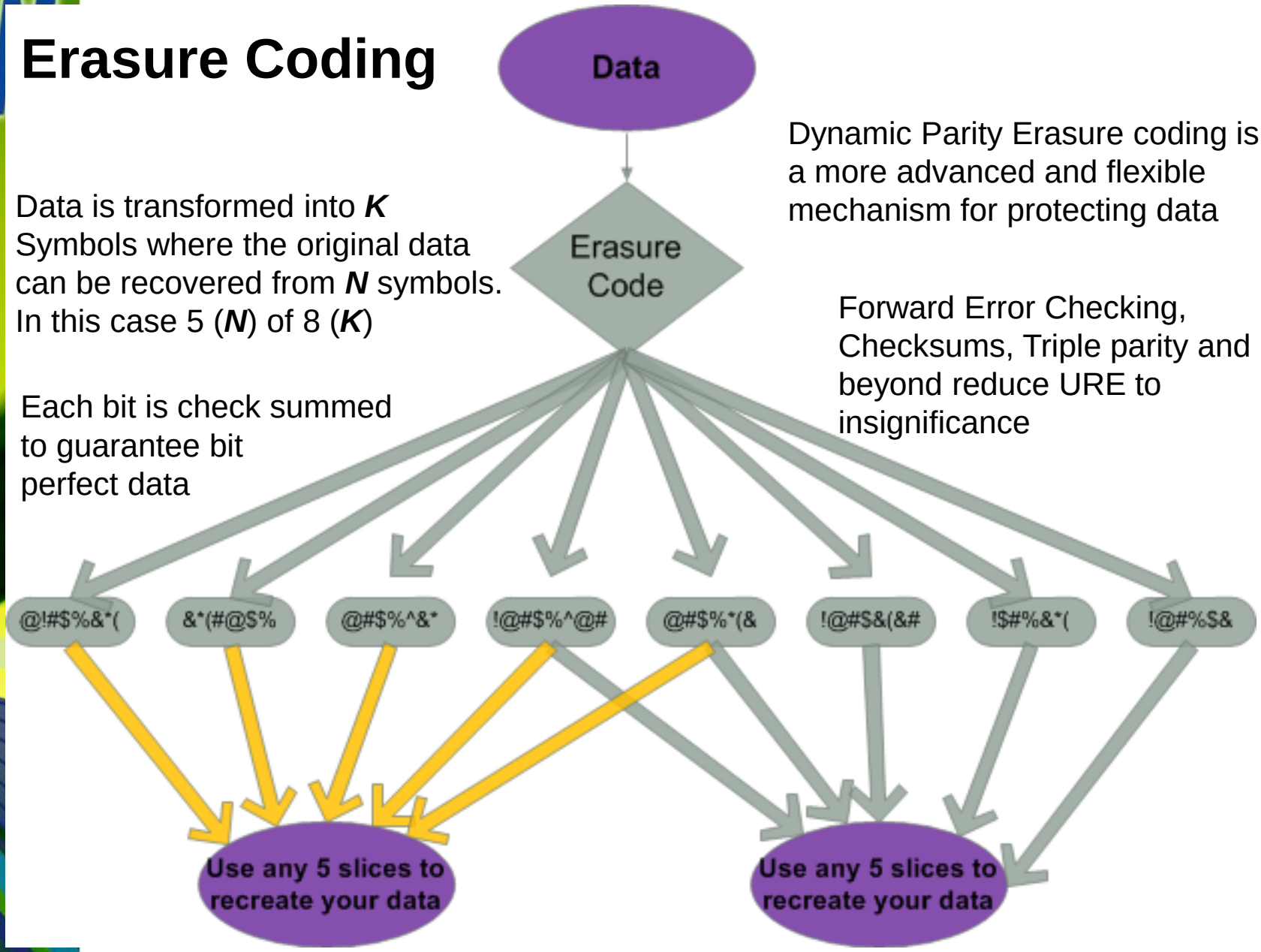
Erasure Coding

Data is transformed into ***K*** Symbols where the original data can be recovered from ***N*** symbols. In this case 5 (***N***) of 8 (***K***)

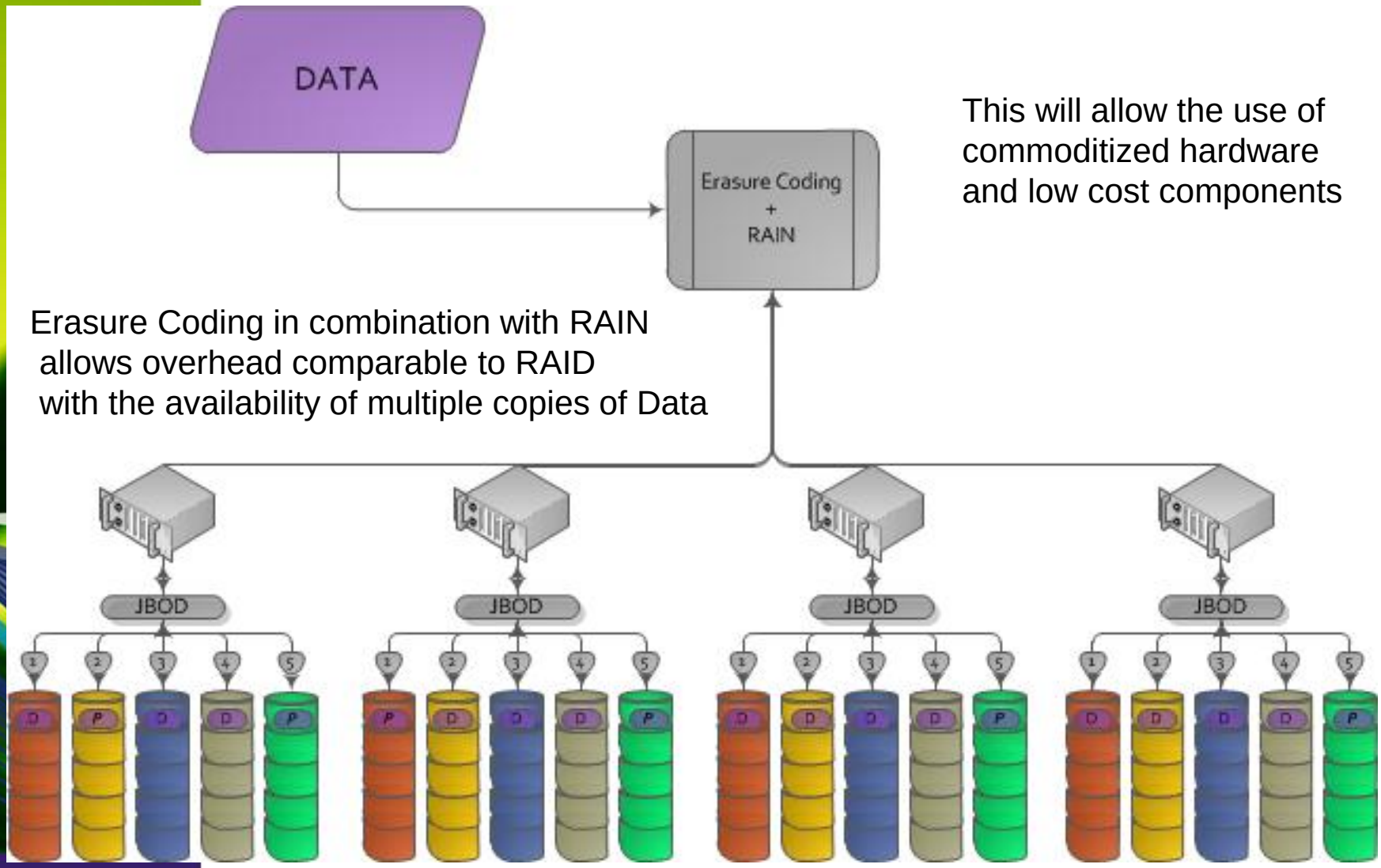
Each bit is check summed to guarantee bit perfect data

Dynamic Parity Erasure coding is a more advanced and flexible mechanism for protecting data

Forward Error Checking, Checksums, Triple parity and beyond reduce URE to insignificance



RAIN
Redundant Array of Inexpensive Nodes



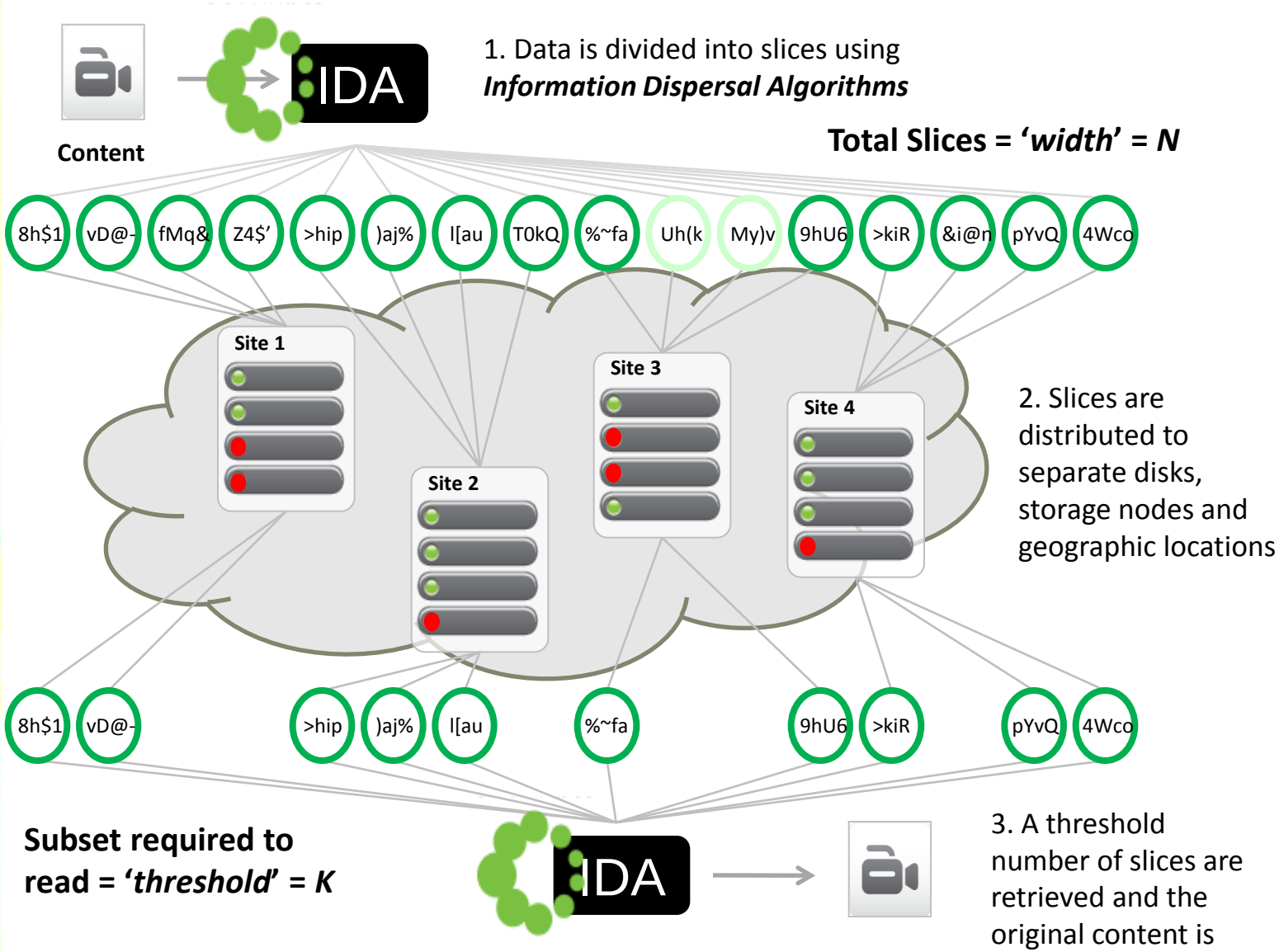
This will allow the use of commoditized hardware and low cost components

Erasure Coding in combination with RAIN allows overhead comparable to RAID with the availability of multiple copies of Data

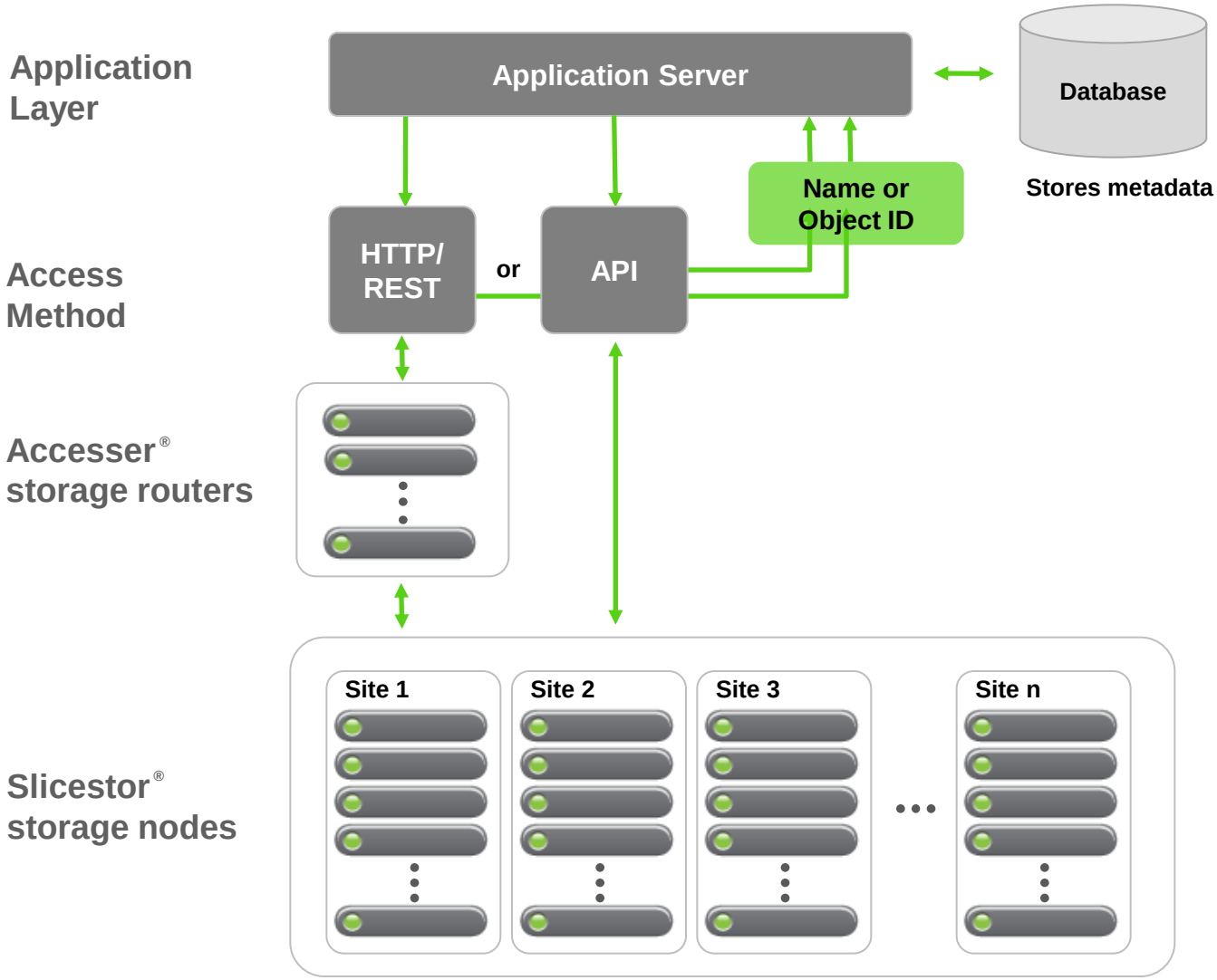
End to End Approach

- Identified over 40 potential strategic partners
- Developed extensive scoring metrics
  - Beyond $/IOP(S) and $/Gb
- Asked lots of questions
  - How does the application use the storage? How does the user use the application?
  - Large block? Small block? Random? Sequential? B2B? B2C? Peaky traffic? Weekly, Monthly, Seasonal patterns? Transactional?
  - Integrity --Does it scale? For how long? Can you account for 8TB drives? 10TB?
  - Reliability / Availability / Resiliency
    - Can you upgrade in place without fear?
    - How long before you lose data?
    - 24x7x365 Data Availability?

How Dispersed Storage Technology Works



Object Integration



Storage Brick

- Internal bandwidth of up to 600MB/s
- Limited by the network bandwidth
- Motherboard with long life support from SuperMicro
- 24 GB Memory, 45 Disks, 4U, 90TB
- Redundant, hot swap power supplies
- Locking SATA cables
- More expensive, still extremely low cost
- Enterprise ready

Why Cleversafe

- Highly motivated
- Perfect fit technology
- Smart engineering
- Strong management
- Well known VC
- Product Influence
- Entrepreneurial Spirit
- Rapid development
- Setting the standard
- Less Politics
- Empowered employees
- Prideful
- Big risk, big reward
- Innovation

Objectives Met

- ✓ Graceful failure and recovery
- ✓ Future proof storage
- ✓ Lowered costs

Key Takeaways

- Migrating metadata types and from traditional file systems isn’t as hard as you think it is.
- Flexibility in the application architecture allows you to test and implement new technology without risk.
- Flexible Storage design allows for separate tiers of storage using the same technology across the board from SSD to 10TB drives
- Intelligent software allows you to truly commoditize storage at scale.
- Take big risks intelligently.

THANK YOU!

Chris Gladwin
CEO, Cleversafe

COMPUTERWORLD SNIA

SNW

Transforming the Information Infrastructure: Build, Manage, Optimize.

FALL 2011