Scale-out Data Deduplication Architecture

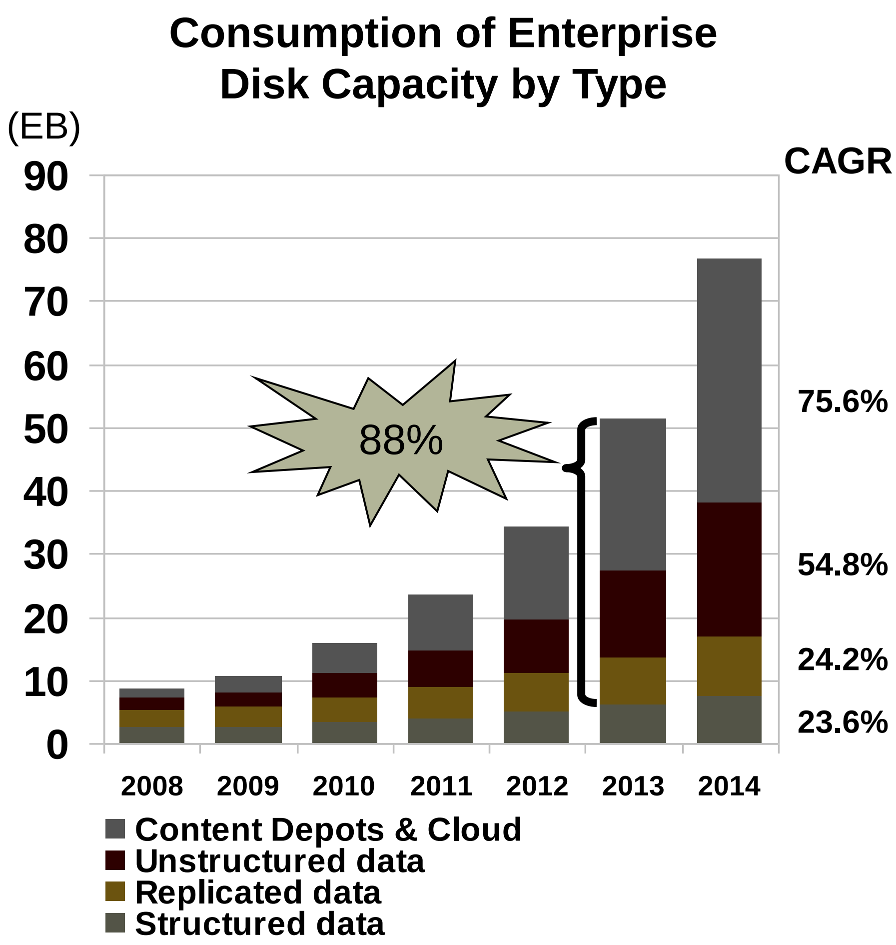
Gideon Senderov
Product Management & Technical Marketing
NEC Corporation of America

Outline

- Data Growth and Retention
- Deduplication Methods
- Legacy Architecture Limitations
- Adaptive Platform for Long-term Data
- Enhanced Scale-out Attributes
  - Scalability
  - Performance
  - Resiliency

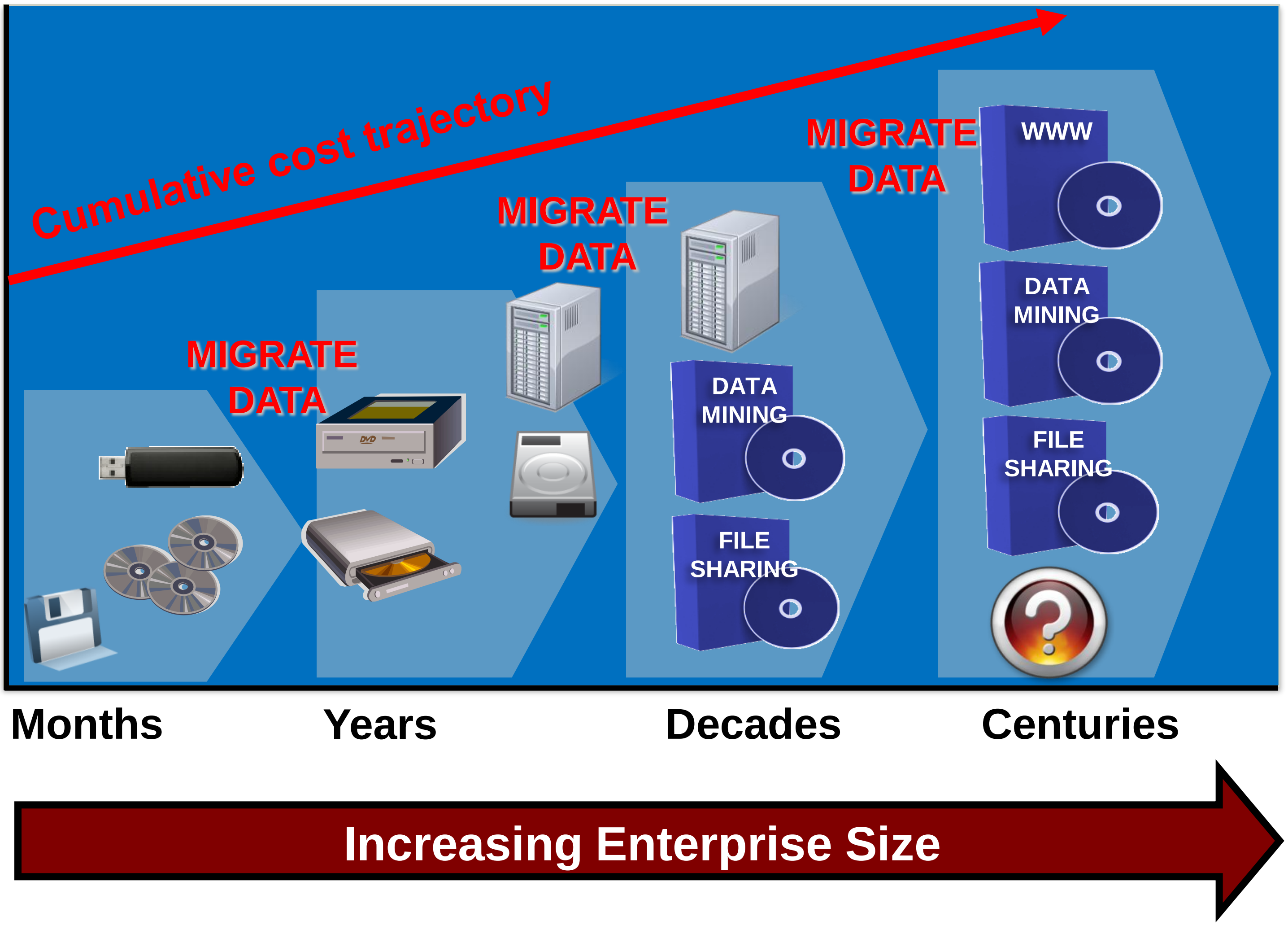
Long-term Data

- According to SNIA, 80% of files are dormant, unchanged.
- According to Contoural, average file shares grow 40% annually.
- Non-mission critical data filling costly Tier 1 storage



Source: IDC 2010

Data Outlasts Its Container



Data Migration Costs

- In 2011 there will be 1775 Exabytes of information
- Archive data outlasts a storage container, data migrate costs $1K – $1.8K per TB
- 15% of Storage Admin's time spent on data migration

Source: IDC, Gartner, Contoural Consulting

Deduplication Granularity

- File (SIS)
  - Data deduplication performed at a file or object granularity
- Sub-file (Block)
  - Data deduplication performed at a sub-file granularity
    - Fixed size chunk
    - Variable size chunk
- Sub-file variable size chunk deduplication benefits
  - Greater efficiency by deduplicating data at finer granularity
  - Automatic adjustments (“slide”) across inserted data

Inline vs. Post-process

- Inline
  - Data deduplication performed before writing the deduplicated data
- Post-Process
  - Data deduplication performed after the data to be deduplicated has been initially stored
- Inline deduplication benefits
  - Eliminate need for disk buffer for un-deduplicated data
  - Eliminate need for subsequent “idle time” for processing
  - Immediate replication and shorter RTO for DR

Generic vs. Application-aware

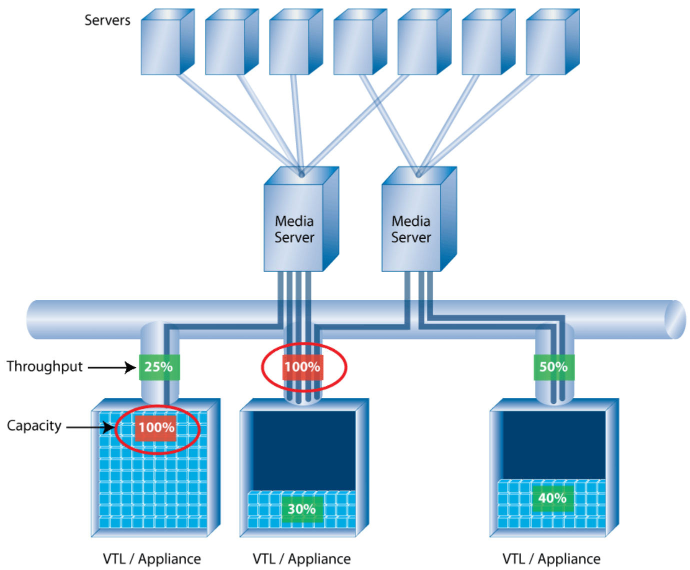
- Generic
  - Deduplication with same algorithm against all data streams
- Application-aware
  - Custom deduplication for specific data streams to account for metadata inserted by the corresponding applications
- Application-aware deduplication benefits
  - Greater efficiency by eliminating negative impact of application metadata on deduplication ratio
  - Cross-application deduplication for greater scope and higher deduplication efficiency

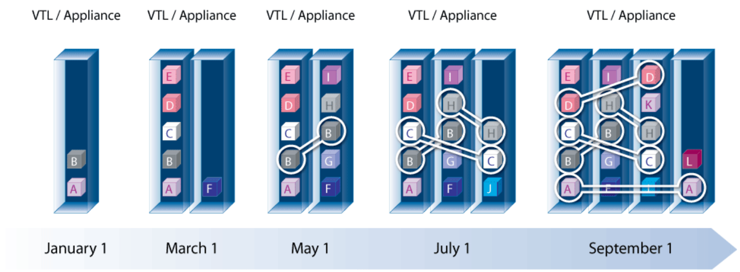
Local vs. Global

- Local
  - Deduplication across a single node or sub-node, requiring separate deduplication repositories for multiple nodes
- Global
  - Deduplication spanning multiple nodes with a single shared deduplication repository across all nodes
- Global deduplication benefits
  - Greater scalability of a single deduplication repository
  - Greater efficiency with a broader scope spanning multiple nodes

Legacy Scalability Limitations

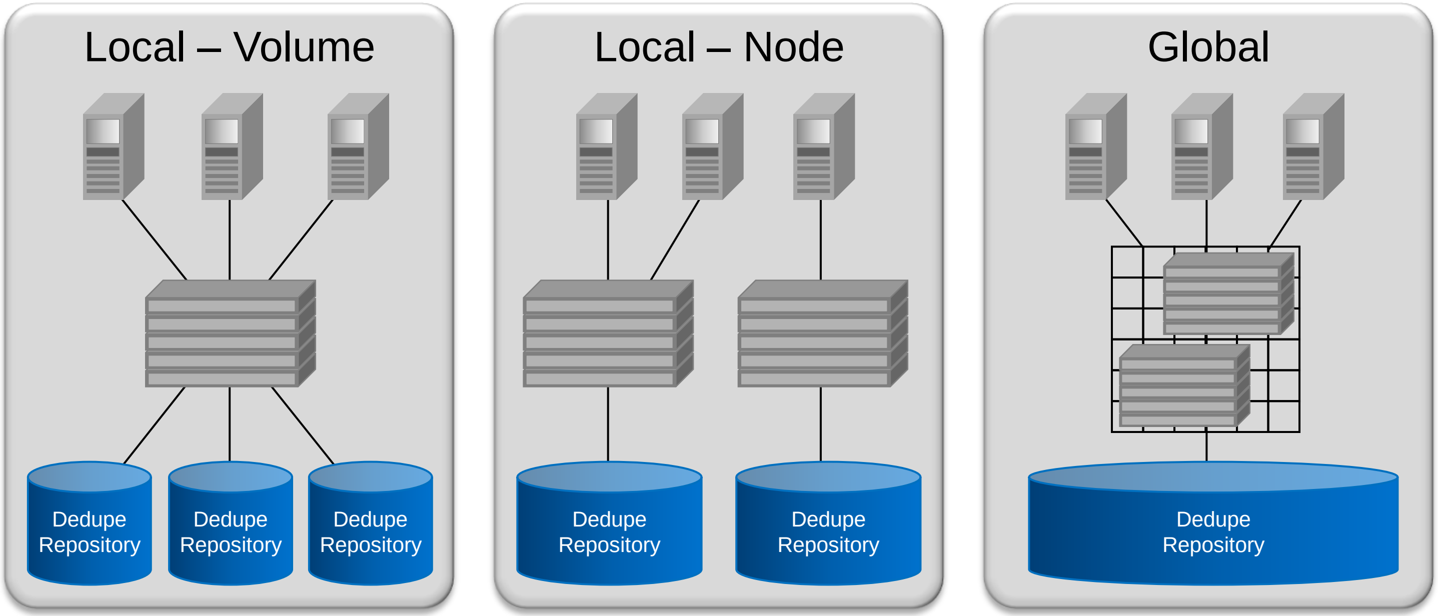
- Inadequate scalability of capacity & performance
  - Cannot scale performance to keep up with growth
  - Multiple products with different architectures
  - More siloed capacity to manage
- Limited deduplication scope
  - Limited scalability proliferates duplicate data across appliances
  - Lower deduplication ratio for large environments



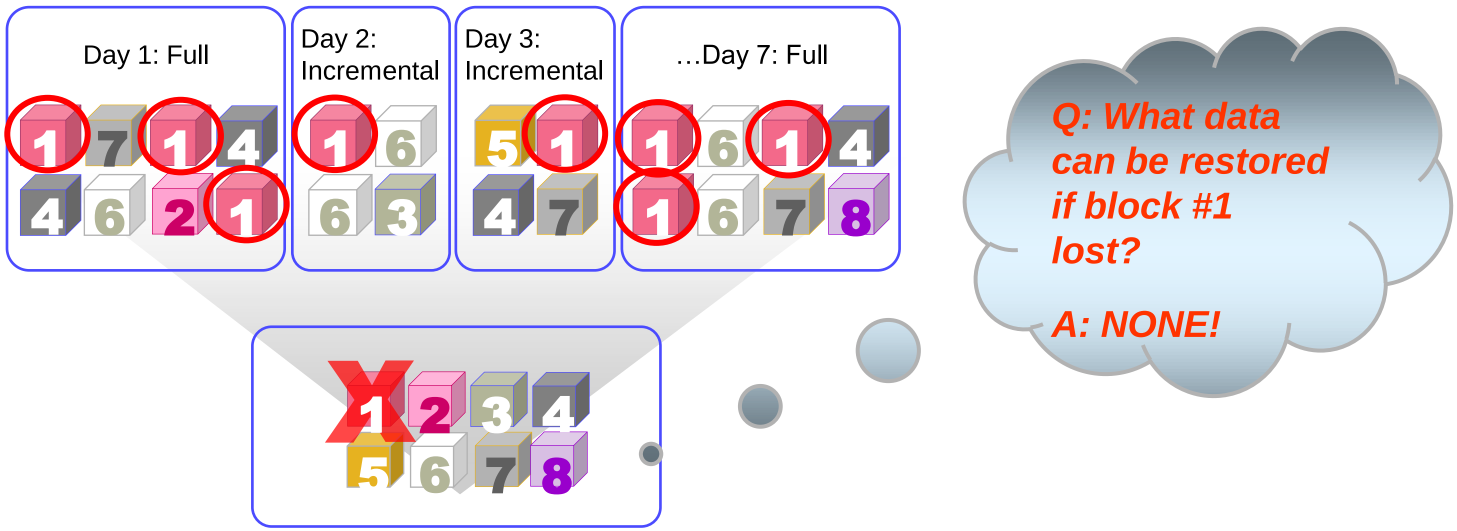


Local vs. Global Scope

- Single deduplication repository across entire solution
- Data deduplication across ALL data from ALL nodes
  - Cross-node dedupe for greater efficiency
  - Cross-application dedupe with application-awareness



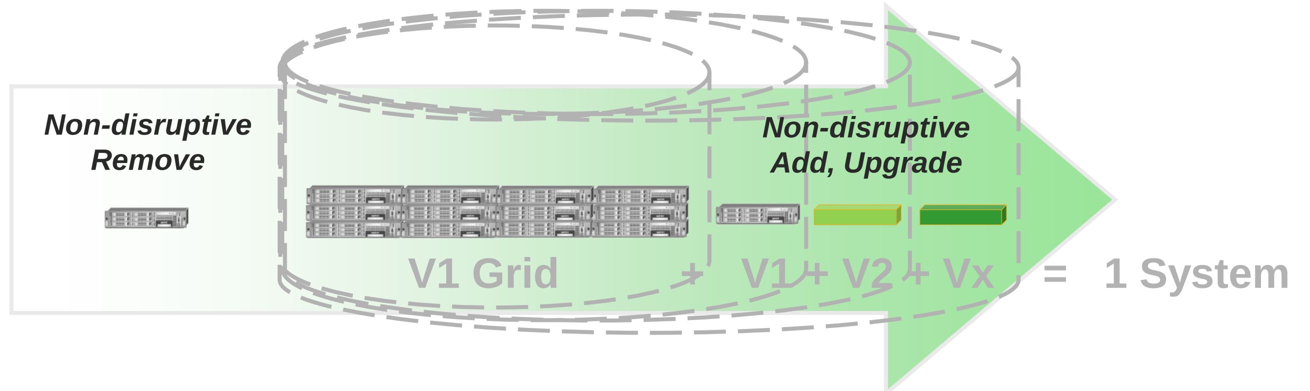
Legacy Resiliency Limitations



Traditional RAID is not sufficient for deduped data

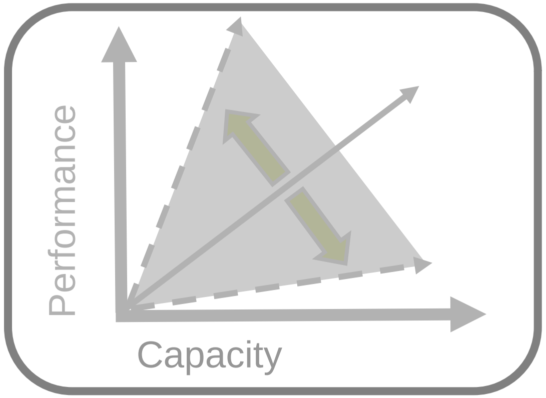
Adaptive Scale-out Architecture

One system becomes Greener, Faster, Denser



- Concurrently support multiple HW generations
- Non-disruptive resource scaling
- Zero provisioning, dynamic self-management
- Automatically balance system workload
- In-place technology refresh – No migration

Independent Linear Scalability



Uniquely Scale
Performance and Capacity
to Meet Current and
Future Needs

- Customized performance and capacity per application
- Dynamically adjust to changing data growth

Enhanced Scale-out Attributes

Criteria	Attributes
Scope	GLOBAL DEDUPE
Performance	SCALABLE DEDUPE
Protection	SAFE DEDUPE

Scope

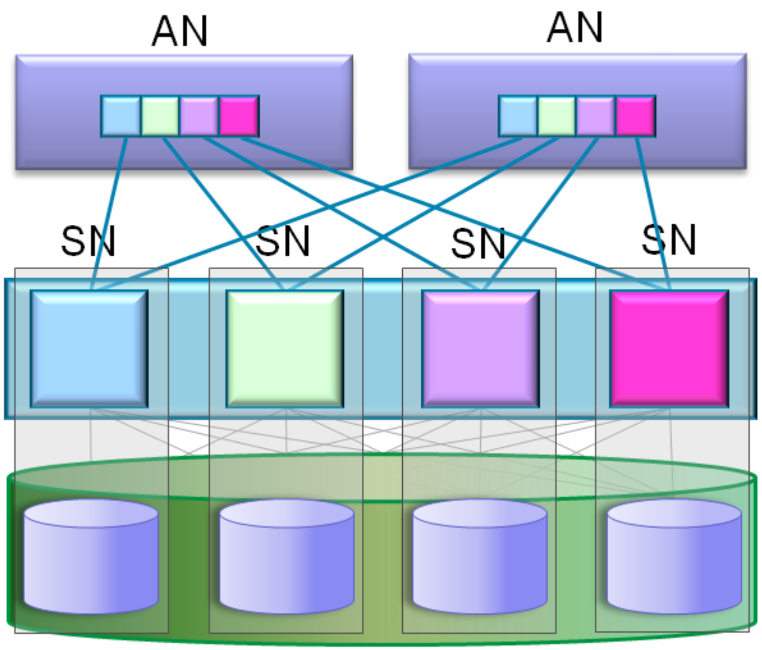
- Multiple applications and data sources across a single scalable deduplication repository
- Broader scope leads to greater efficiencies
  - Higher dedupe ratios
  - Improved capacity utilization
  - Easier management
  - Longer retention periods
  - Lower costs
- Single scalable system vs. multiple disparate silos
  - Cross-application deduplication for greater efficiency
  - Global deduplication for entire environment

Performance

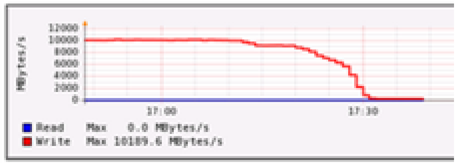
- Linear performance scalability to keep up with data growth
- Scalability
  - Linear scalability vs. high degradation
  - Independent performance scalability
  - Beyond the physical boundaries of the system
- Inline dedupe vs. post-process
  - Keeping up with the workload
- Effects of deduplication rates
  - Higher performance with higher dedupe

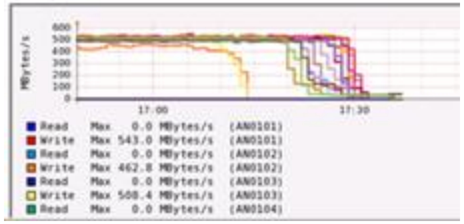
Scale-out Dedupe Architecture

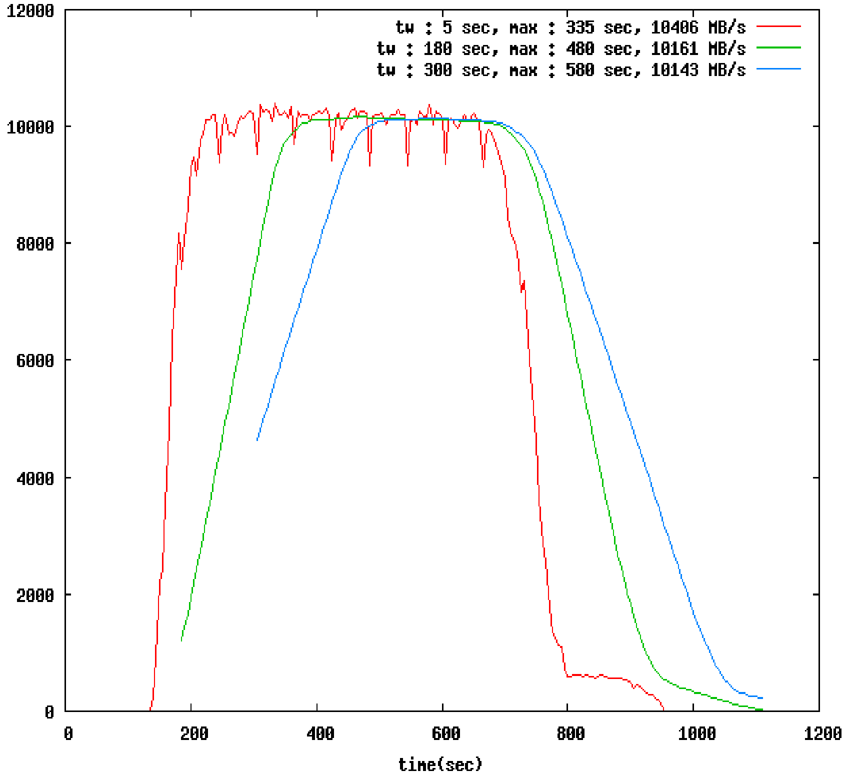
- Multi-node architecture
  - Inline data routing
  - Processing and memory scales with capacity
- Distributed hash table
- Linear performance scalability
- Global deduplication across ALL nodes



Linear Performance Scalability





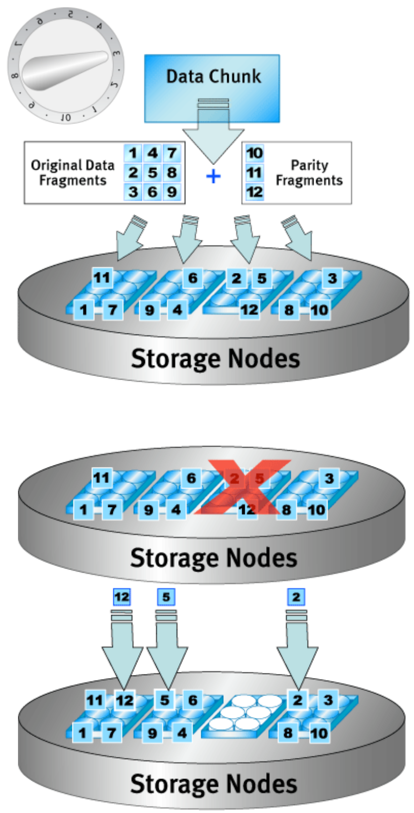


Resiliency

- Enhanced data resiliency, especially with fewer copies due to deduplication
- The greater the dedupe ratio, the greater the exposure
- Maximum protection against multiple failures
  - Component failures
  - Appliance or node failures
- Upgradeability and technology refresh
  - In-place technology refresh vs. forklift upgrade
  - Online versus downtime
  - Configurable resiliency for different applications
  - Investment protection

Erasure-coded Resiliency

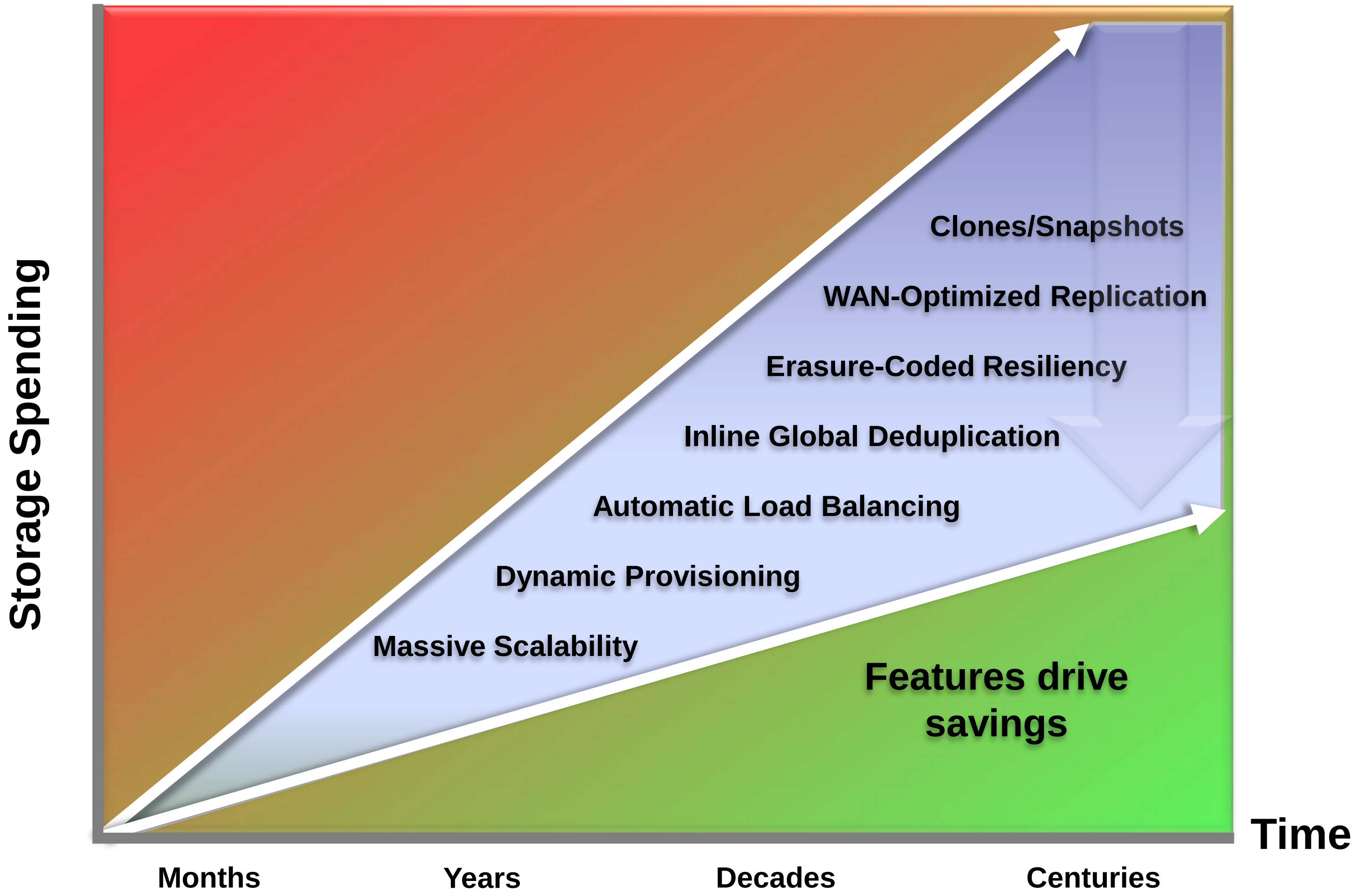
- “User dialable” disk/node protection
  - Intermix of multiple resiliency levels
  - Dynamically allocated protection
- Greater protection with less overhead
  - No idle spare drives
- Faster healing while maximizing I/O performance
  - Only data is reconstructed
  - Reconstruction across multiple spindles/processors



Long-term Evolution

- Functionality
  - Application-specific requirements (Security, Resiliency)
  - Regulatory compliance requirements
- Performance
  - Linear scalability throughput HW/SW enhancements
  - Multiple I/O profiles
- Attributes
  - Flexibility
  - Efficiency
  - Granularity

Long-term Data Protection



Driving Innovation Through the Information Infrastructure

SPRING 2011