Data Replication Technology

Starring John Wolfgang as Bob Kern

John Wolfgang	jwolfgan@us.ibm.com	520-799-4819
Bob Kern	bobkern@us.ibm.com	520-799-5465
Ken Boyd	kenboyd@us.ibm.com	520-799-2720
Ken Day	mycroft@us.ibm.com	520-799-4582

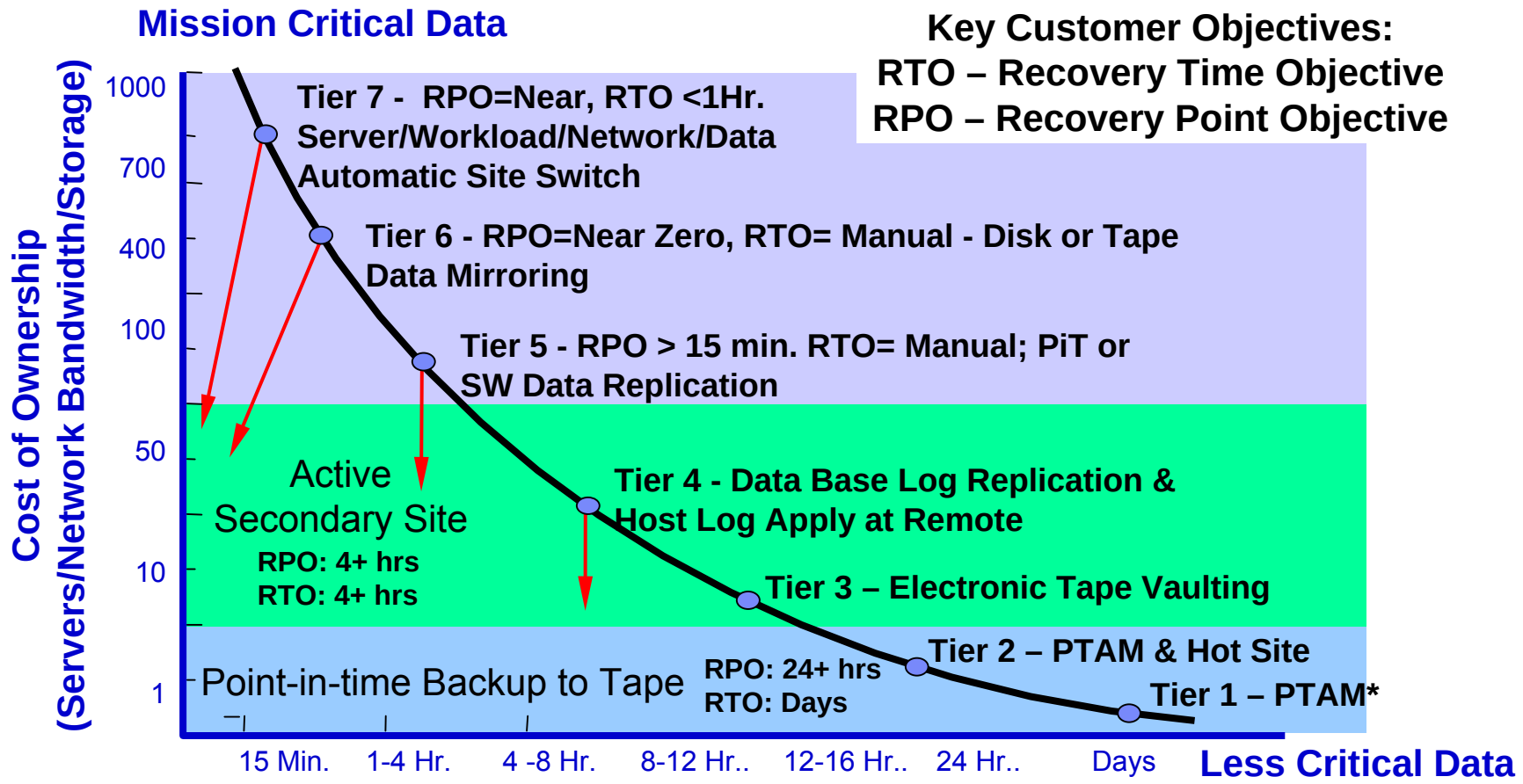
IBM Storage Systems
Tucson, AZ

NASA/IEEE MSST 2004
12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies
The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004

Agenda

- Disaster Recovery/Business Continuance Background
- Application Based vs Storage Based Data Replication
- Data Replication Technology
  - Local Subsystem Replication
  - Remote Replication (Synchronous & Asynchronous)
  - Cross Site Connectivity
- The Current Marketplace
- Emerging Data Replication Technology
  - Host Based
  - Switch Based
  - SAN Replication Appliance within Data Path
  - SAN Replication Appliance outside Data Path
  - Storage Subsystem Peer
- Key Questions for Any Solution
- Discussion

7 Tiers of Business Recovery Options



Time to Recover – How quickly is an application recovered after a disaster?

*PTAM – Pickup Truck Access Method

Lessons from 9/11

- Rolling disasters happen
- Distance is more important
- Redundancy may be smoke and mirrors
- Recovery needs a greater dependency upon automation and less on people
- Recovery site:
  - ★ Capacity (MIP's and GBs) site needs to be sized for the production environment that will run there
  - ★ Disasters may cause multiple companies to recover and that puts stress on the commercial business recovery services
  - ★ D/R Plan after successful recovery from disaster
- Rethinking of synchronous versus asynchronous

Value of Data Replication

Operational Efficiency

Data Mining
Content Distribution
Software Testing

Availability Improvements

Backup Window
Tape Backup
Data Migration
Archival

Disaster Recovery/ Business Continuity

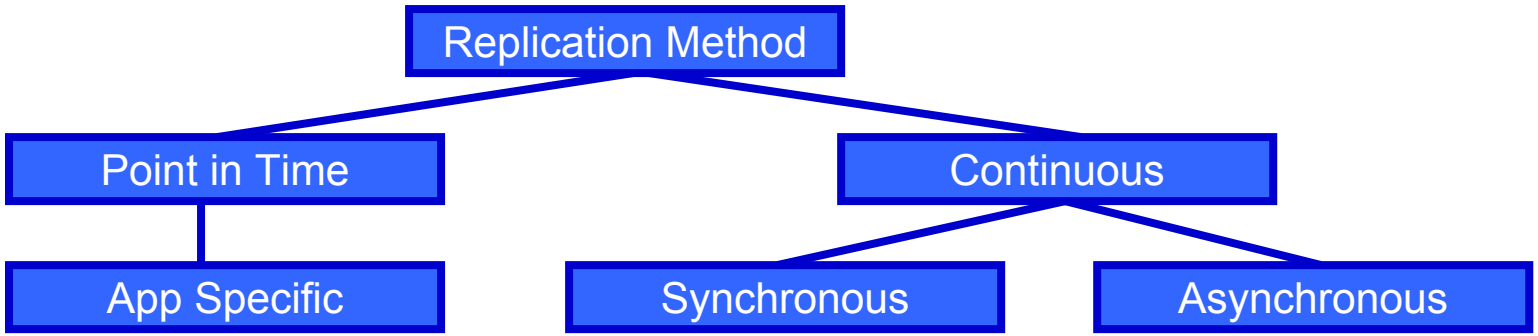
Minimize data loss
Minimize restart time
Increase distance
Enable automation

Data Replication Building Blocks

Flash Copy	PPRC	File Flash	Migration Copy	Synch Asynch	Tape Repl

Data Replication Building Blocks

- Data replication technologies provide non-disruptive ways to relocate, migrate and/or copy data.
- Replication can be performed by the host, in the SAN, in disk systems or tape systems.



File System
  AIX JFS2, GPFS, Stg Tank
Logical Volume Managers
  AIX LVM
  Veritas Application
Database Log Shipping
Disk System FlashCopy
Tape Volume Replication

Disk System
  ESS PPRC
  EMC SRDF
  HDS HRC
  FAStT
  DB/Messaging Transaction Replication
LVM
  AIX LVM, GeoRM,
  Veritas Volume Replicator

LVM
  GeoRM
  Veritas Volume Replicator
Disk System
  ESS PPRC-XD, XRC
  EMC SRDF
  HDS HRC
DB/Messaging Transaction Replication
Emerging New Players

Software Based vs Storage Based Data Replication

- Application/File/Transaction Based
  - ★ Specific to application file system/DB
  - ★ Generally less data is transferred -> lower Telco costs
  - ★ No coordination across applications, FS, DBs, etc
  - ★ Applications change - replication may need to change
  - ★ May forget “other" related data necessary for recovery
  - ★ With many transfers occurring in a corporation, it may be difficult to determine what is where in a disaster. RTO/RPO may not be repeatable, auditing may be difficult
  - ★ Many targets possible (ex. millions of PDAs or cell phones)
  - ★ Others

Software Based vs Storage Based Data Replication

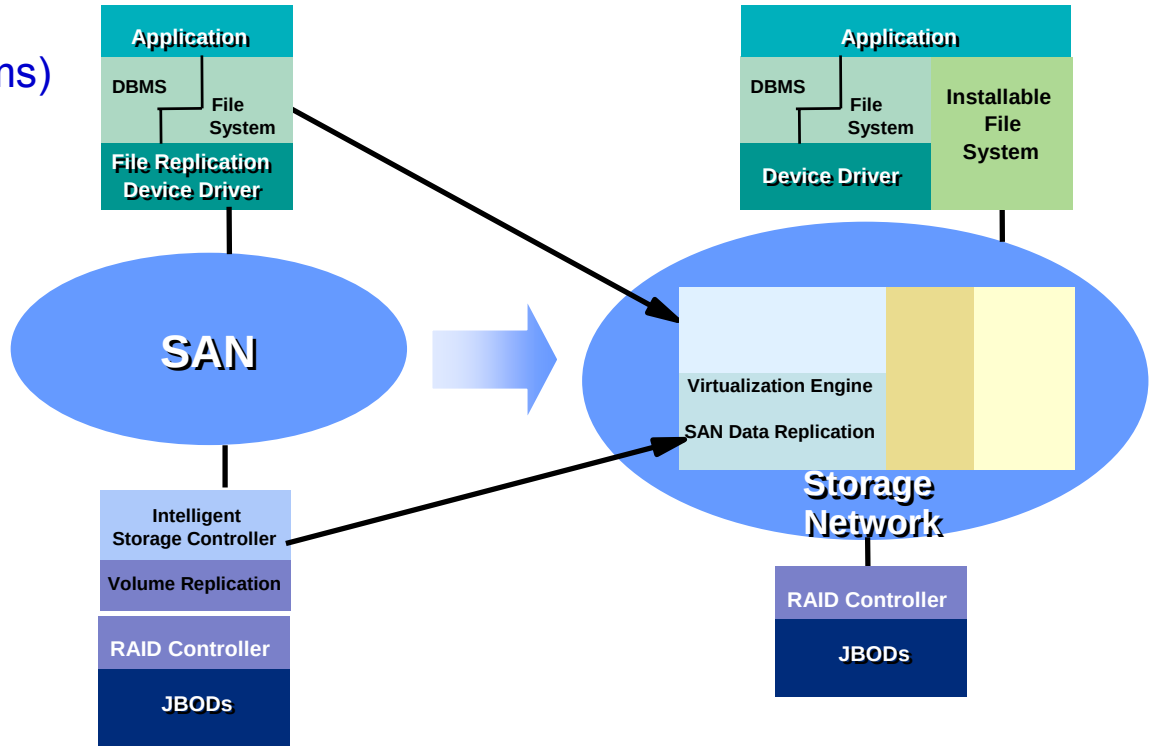
- Block/Record Based
  - ★ Independent of application, file systems, databases, etc
  - ★ Common technique for corporation. (managed by operations)
  - ★ Generally more data transferred -> higher Telco costs
  - ★ Consistency groups yield cross volume/storage subsystem data integrity/consistency
  - ★ Independent of application changes. Mirror all pools of storage
  - ★ Consistent repeatable RPO. RTO depends on server/data/workload/network
  - ★ Generally a handful of targets
  - ★ Specific to data replication technique (tied to specific architecture & devices that support it)
  - ★ Others

Data Consistency for Block Based Storage

- Only consider "*power fail*" consistency
- Typical Database transaction:
  - ★ Journal entry indicating database update which is about to occur
  - ★ Update database
  - ★ Journal entry indicating database update has occurred
- Host is very careful to do each of the transactions in order
  - ★ This provides power fail data consistency
- These transactions are likely done to different volumes on different control units
- Failure to be careful about transaction order results in loss of data consistency and data may become unusable
- In order to ensure data consistency at secondary site, dependent writes must be done in order
  - ★ Writes that are done in parallel are not dependent

Placement of Data Replication Function

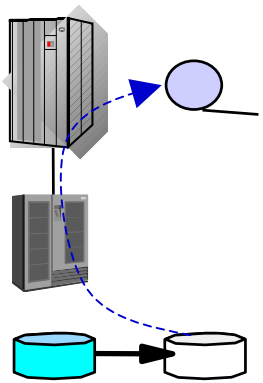
- Application/File/Transaction Based
  - ★Host Software Based
  - ★Application (ex. Mail systems)
  - ★LVM (Write/Write)
  - ★File system
  - ★Data Base
- Block/Record Based
  - ★Storage Subsystem Based
  - ★Host Based
  - ★Switch Based
  - ★SAN Appliance
  - ★In Data Path
  - ★Outside Data Path
  - ★Subsystem Peer



Local Subsystem Data Replication

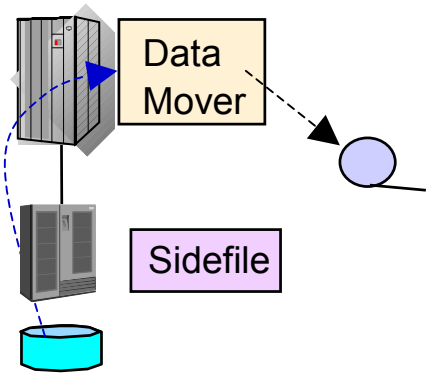
z-Series & Open Systems

FlashCopy

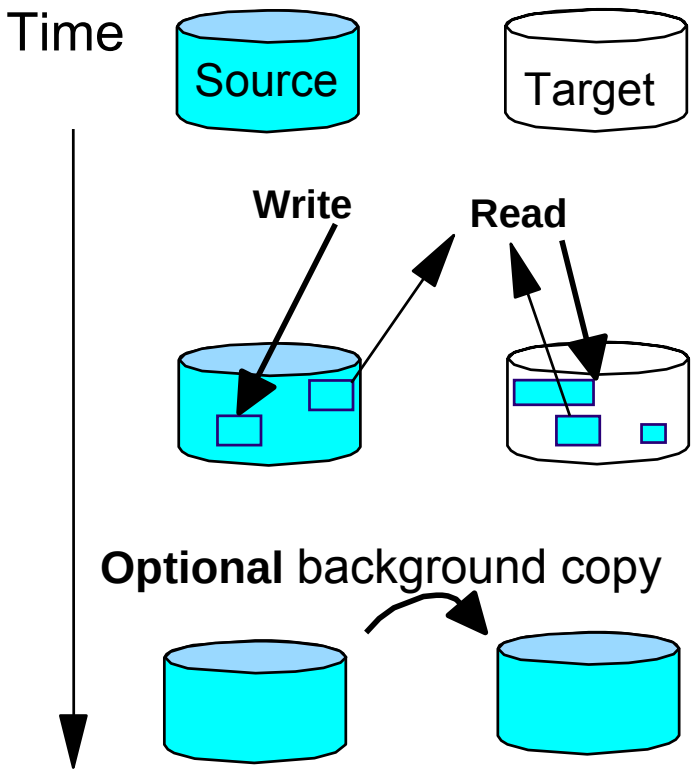


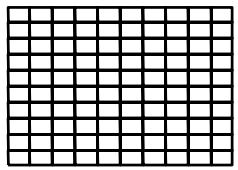
z-Series Only

Concurrent Copy



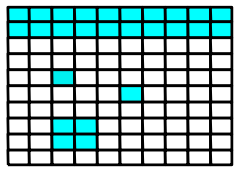
FlashCopy – Internal to Storage Subsystem





Copy data command issued
Copy is immediately available

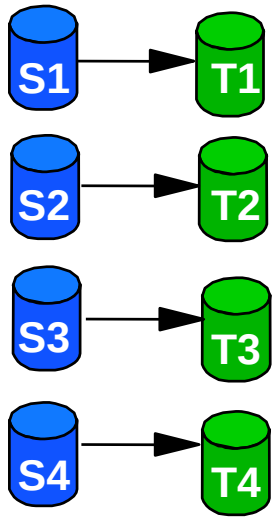
Read and write to both
source and copy possible



When copy is complete,
relationship between
source and target ends

Normal Operations -> No Background Copy
- Dump -> Tape
- FlashCopy before Batch

ESS PiT Consistent FlashCopy

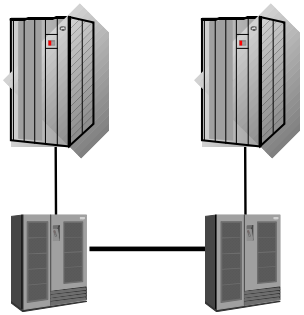


- FlashCopy S1 to T1
  - Writes cannot proceed on S1
  - Any writes occurring on S2-S4 are not dependent writes
- FlashCopy S2 to T2
  - Writes cannot proceed on S1 or S2
  - Any writes occurring on S3-S4 are not dependent writes
- FlashCopy S3 to T3 and S4 to T4
- T1-T4 contain a consistent copy
- Issue Consistency Group Created
  - Writes may proceed to S1-S4.

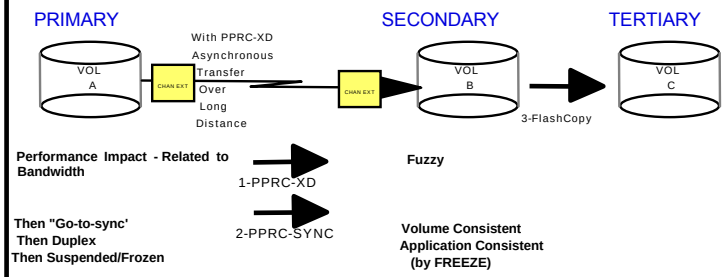
- Hold off initiation / completion of write I/O to the source volumes until FlashCopy establish is completed
- Select source and target volumes with consistency option
- Enables creation of a consistent point-in-time copy across multiple volumes with minimum host impact and no operator intervention required
- Source and target volumes are within one ESS
- Consistency Groups Can Overlap Multiple ESSs

Synchronous & Asynchronous Data Replication

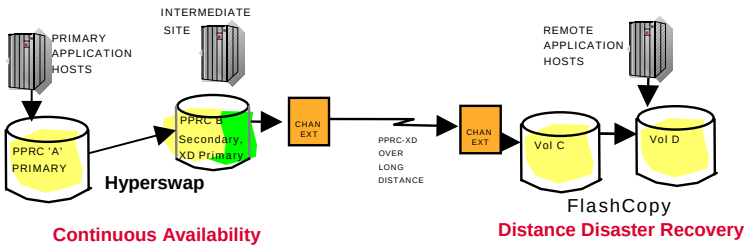
Synchronous PPRC



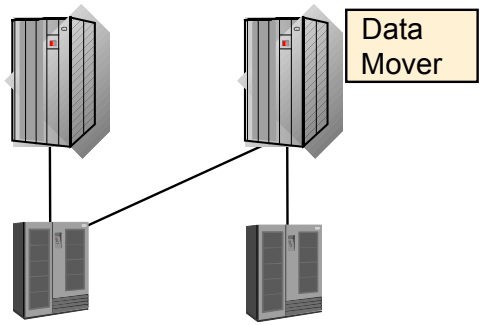
PPRC-XD



Asynchronous Cascading PPRC



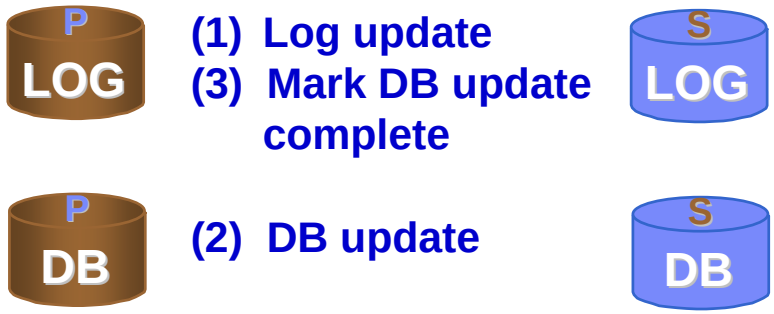
Asynchronous - XRC (zSeries Only)



Consistency Groups

Cross Volume/Subsystem Data Consistency/Data Integrity

- Important for PiT Copy Solutions
- Important for D/R (DB Restart instead of DB Recovery)
- Important for Integrity of the Data (ex. DB Logs & Data Volumes)
- Scope can be Disk Storage Subsystem(s) BUT
  Generally requires Global Systems Level Monitoring



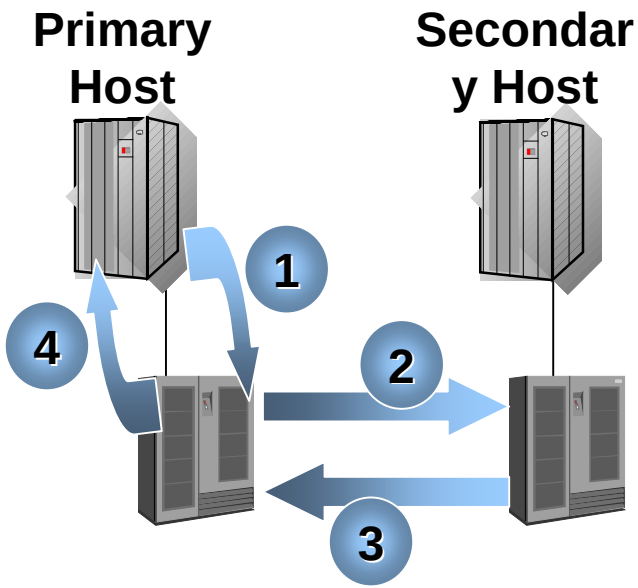
Three Approaches Used in Marketplace:

- Data Freeze Methodology
- Time Based Sequence (SYSPLEX Timer - Only Cross System Clock in Marketplace)
- Put all data requiring consistency on a single LUN

Peer to Peer Remote Copy - Synchronous

- Implemented by many storage vendors
- No data loss is goal (RPO= 0)
- Impact on application write I/Os - distance dependent
- Utilizes automation
  - To freeze secondary upon disaster
  - To provide cross-CU data consistency
- Integrated with DR solutions ex. GDPS (Server/Data/Workload/Network)
  - Freeze capability
  - Data consistency
  - Simplified and fast recovery
  - Automated reconfiguration
  - HyperSwap™ capable
- PPRC Features:
  - Distance to 400km
  - zSeries & Open Systems Support.
  - 1-8 paths per LSS/SSID pair. (ESS 800 - 2 paths w/FCP)
  - Peer-to-Peer Link Optimizations
  - Ability to FlashCopy PPRC Primary or Secondary.

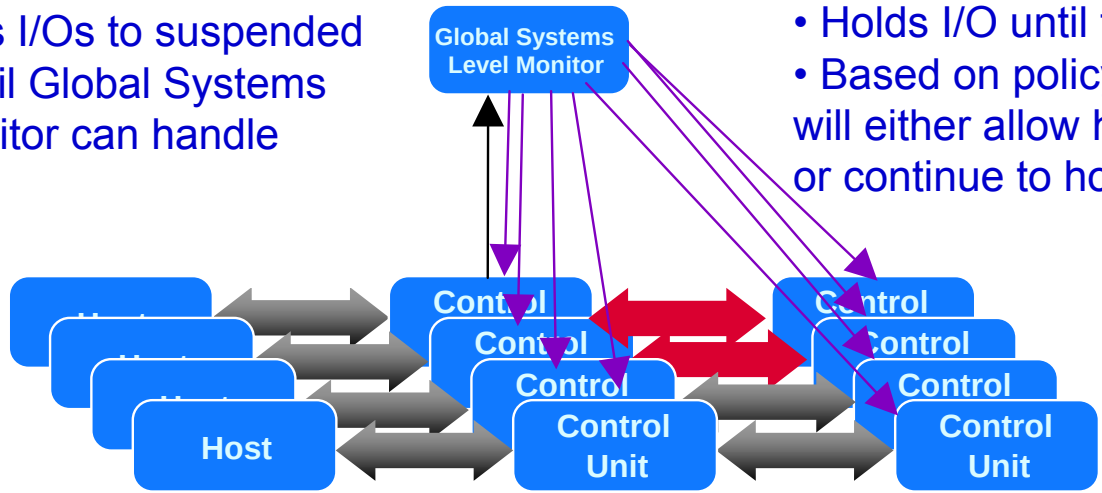
Note: It takes 20ms for light to travel 3000 km round trip



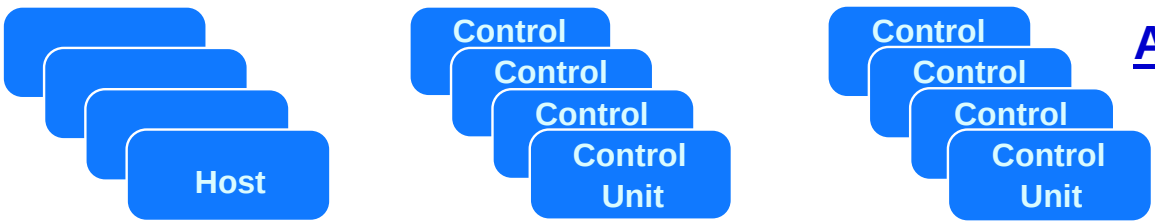
Data Freeze

- Notification of suspending event
- CU holds I/Os to suspended device until Global Systems Level Monitor can handle condition

- Global Systems Level Monitor issues freeze command to all CU's
- Suspends all PPRC pairs
- Holds I/O until told to continue
- Based on policy, Global systems Monitor will either allow host operation to resume or continue to hold them



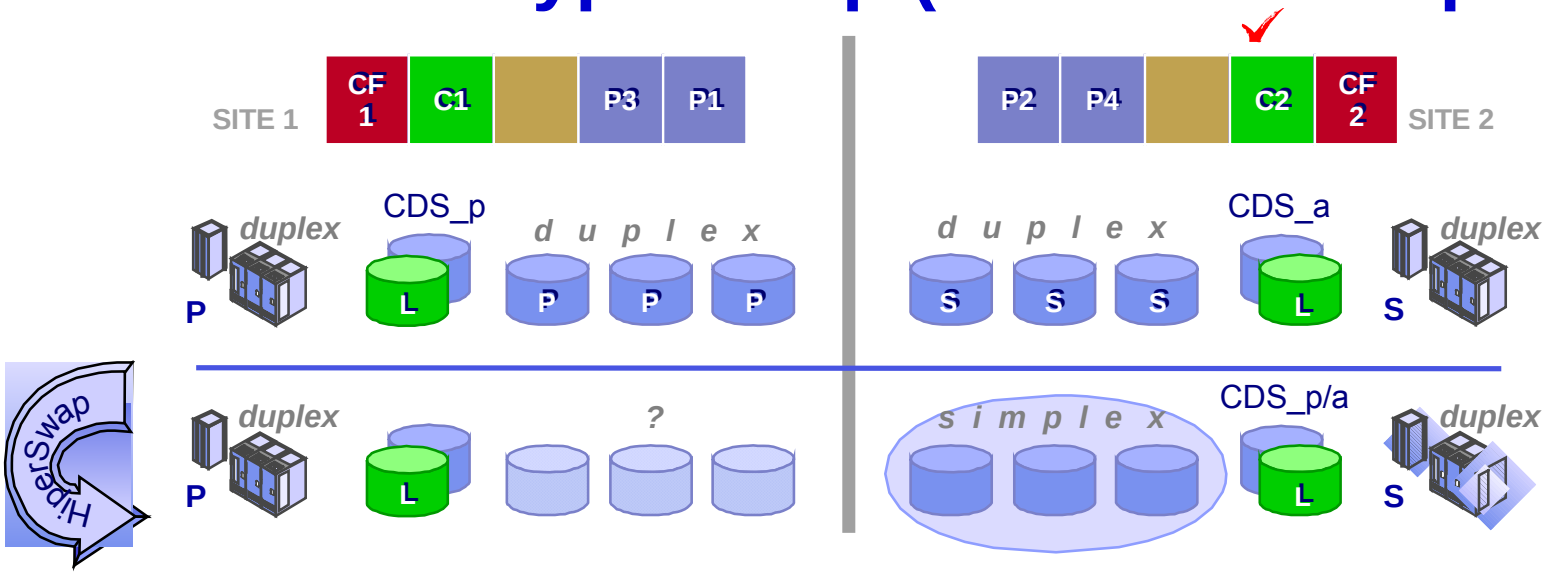
During



After, depending on policy

- Both sites are suspended but in sync
- Sites are not in-sync, however data at secondary site is consistent

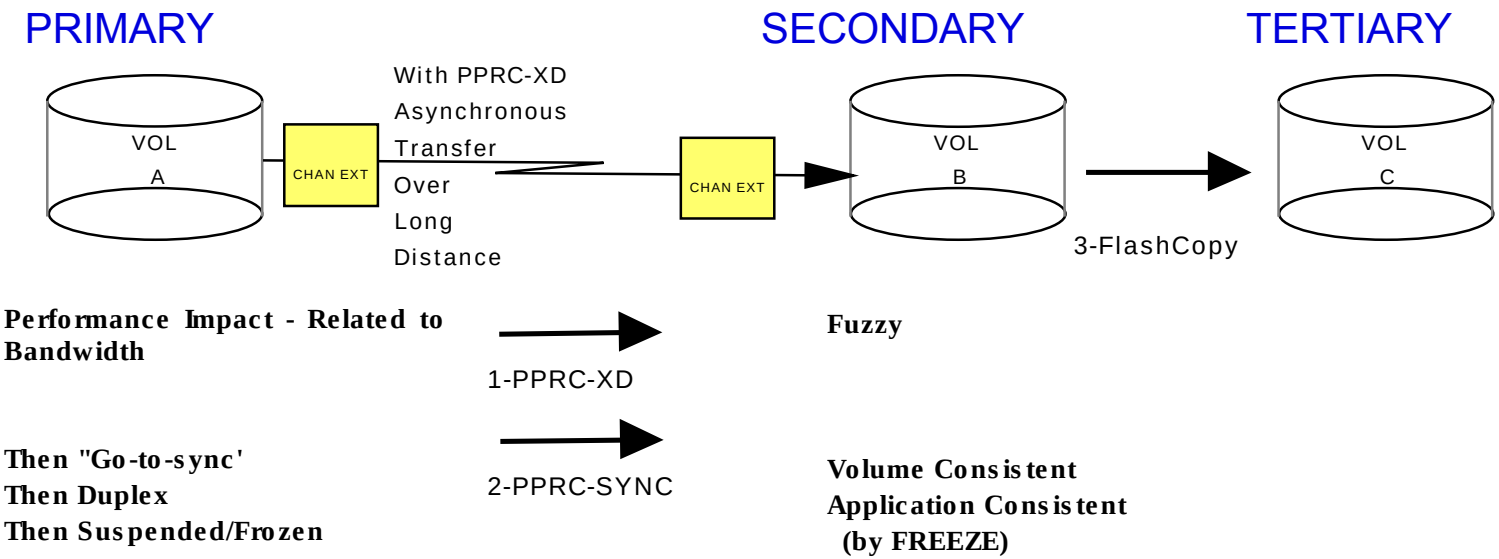
GDPS/PPRC - Hyperswap (Planned & Unplanned)



- GDPS/PPRC Hyperswap Planned & Unplanned Site Switches
- Procedure Step 1
  - Route exception condition (disk Subsystem I/O failure) to the active master system (must be a controlling system)
  - HyperSwap disk configuration (all disk subsystems)
- Procedure Step 2 - executed automatically via script on controlling system (C2)
  - Select secondary volumes (SYSRES, IODF)
  - Switch Coupled Data Set (switch to alternate CDS and spare in site-2)

All production systems remain active throughout the procedure

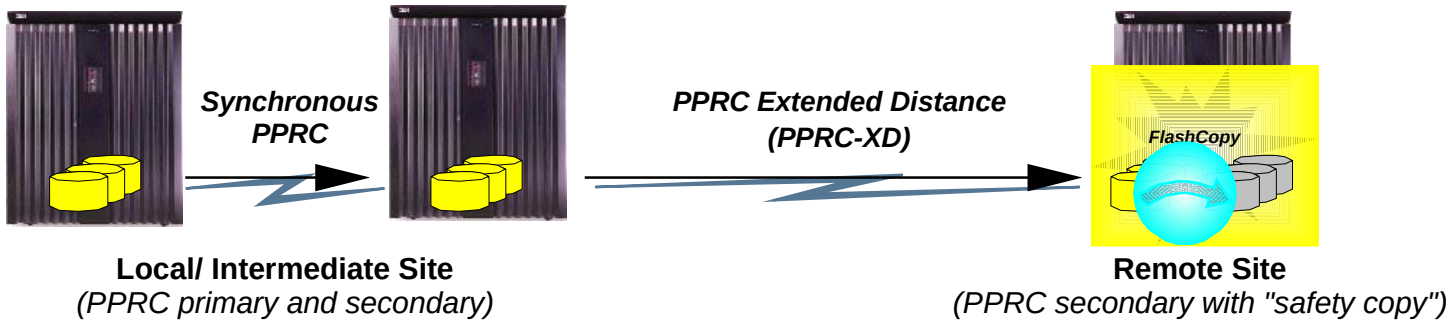
PPRC-XD (Extended Distance)

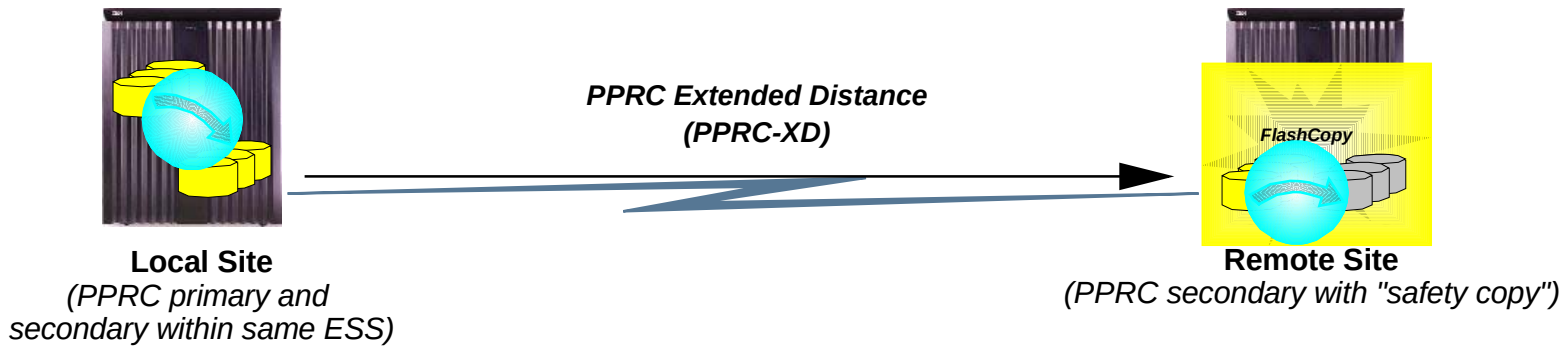


- Asynchronous transfer of updates over long distance through channel extenders
- Little performance impact on applications
- Creates a fuzzy secondary copy
- Can transition to Synchronous mode until full duplex to create PiT Consistency
  - FREEZE to get application level consistency
- FlashCopy onto tertiary to save consistent checkpoint
- Oracle Redo Logs/SAP Hot Standby/Quiesce DB->FlashCopy Resume
- Channel Extenders Compress & Batch PPRC-XD Updates yielding High Bandwidth Utilization
- Test- 256 PPRC XD Pairs, 6000 writes/second, 1200 miles, 2 OC30 lines caught up in 8 seconds

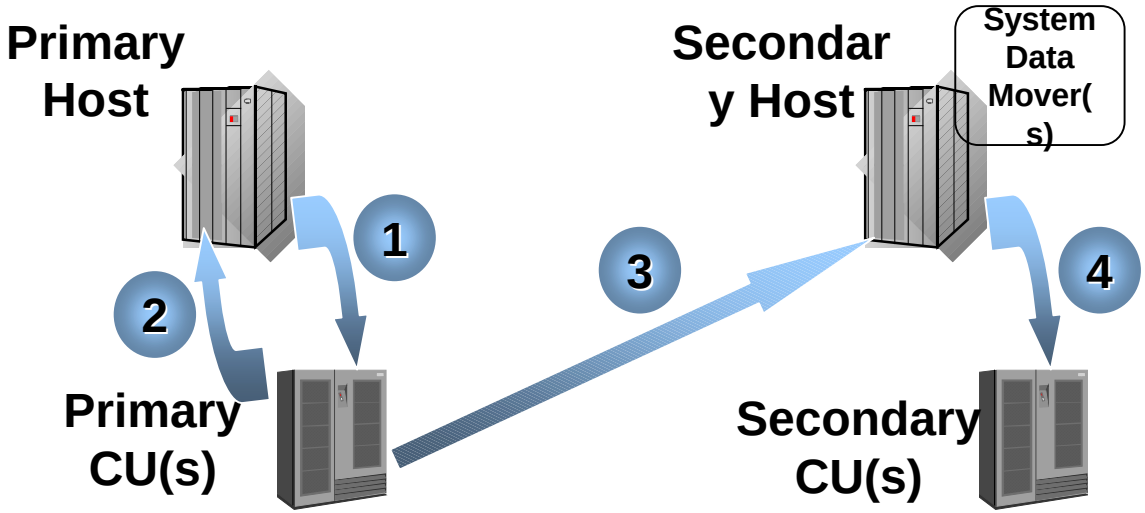
Asynchronous Cascading PPRC

- **Three-site and two-site configuration options**
  - **Flexible configuration possibilities**
    - Better application resiliency, at metro or long distances
    - Made simpler: no operational change between the two configurations
    - Match TCO (Total Cost of Ownership) to desired Tier of Recovery



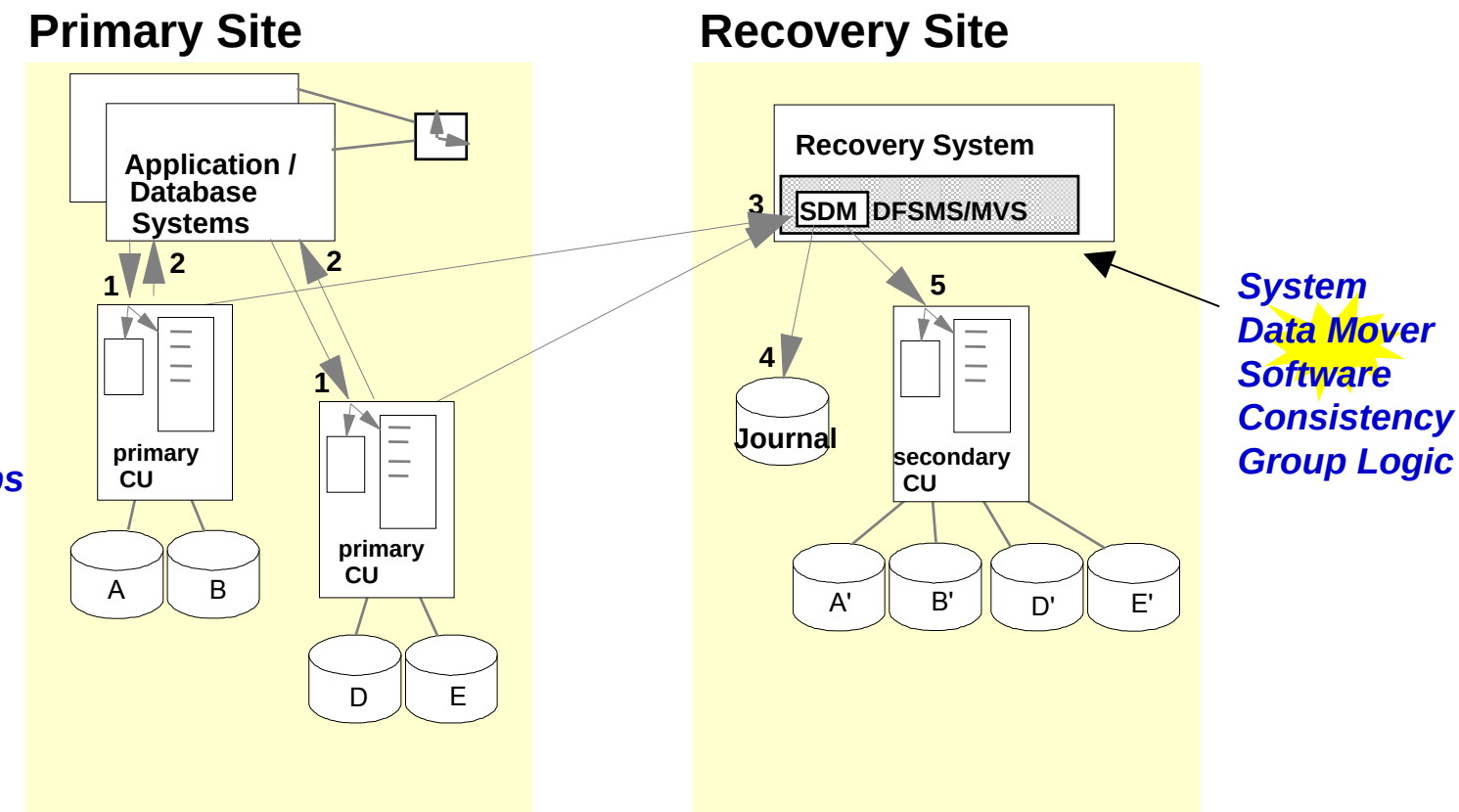


Extended Remote Copy – Asynchronous (zSeries Only)

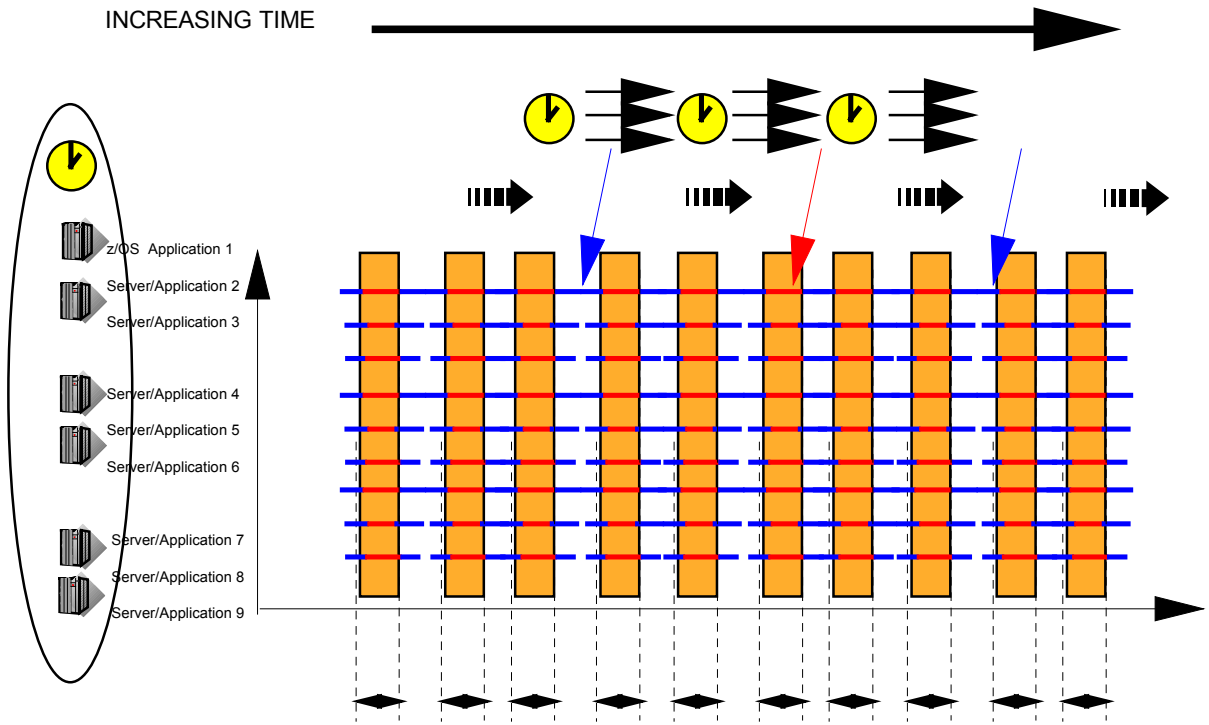


- XRC - open, non-proprietary implementation on several vendors
- Multiple reader offload support
- Dynamic balancing of application write bandwidth vs SDM read performance
- Minimal impact to primary application I/O
- Offload from utility device (different from application I/O)
- Unplanned outage support (Suspend/Resume)
- Host Mips required to form time based consistency groups

Time Based Consistency Groups (zSeries Only)



Time Based Consistency Groups

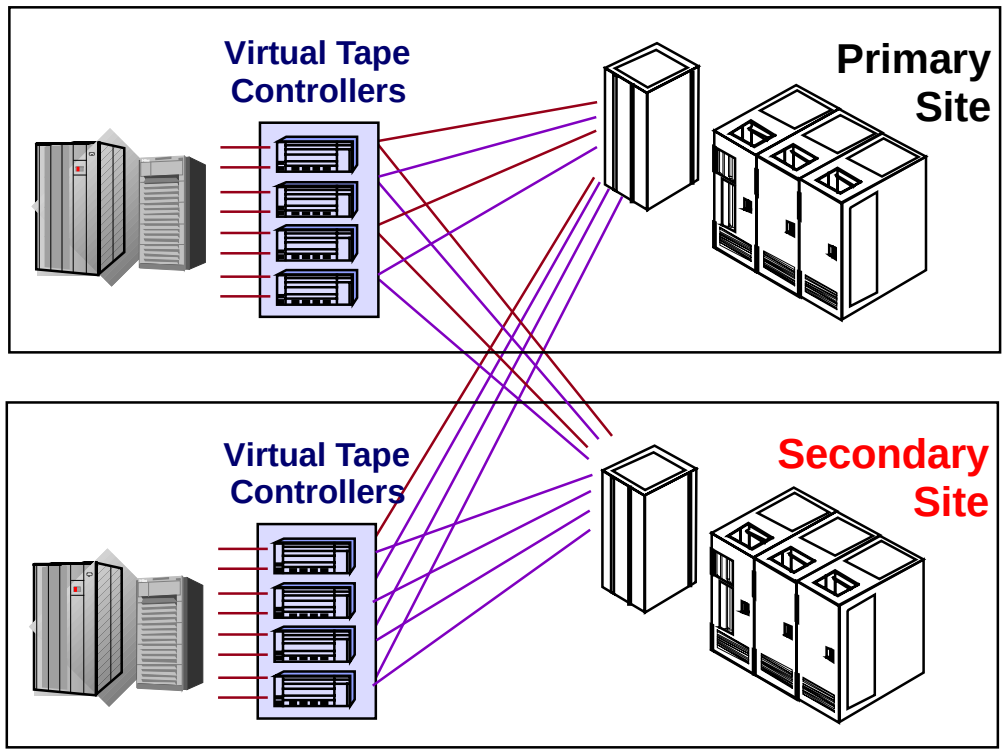


XRC SDM uses common Sysplex Timer timestamp to sort the incoming data, form consistency groups 10's of times per second across large numbers of volumes, disk frames, and z/OS® images.

XRC is able to back out in-flight incomplete write sequences because in event of outage, the XRC SDM does not write out incomplete data, thus what is on the disk is the most recent complete Consistency Group.

Peer to Peer Virtual Tape Server

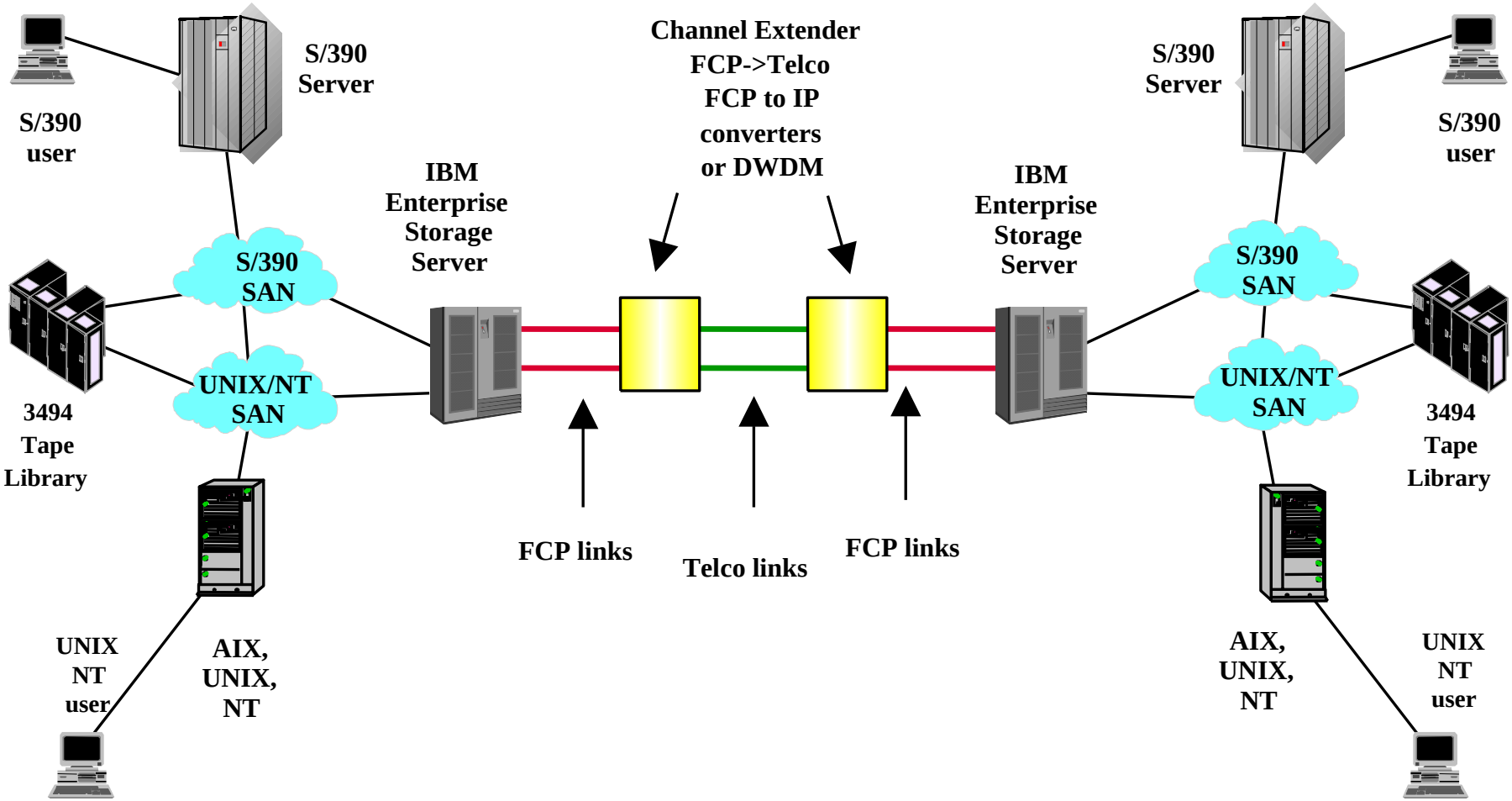
- Primary / secondary VTS
  - ★ Primary performs host I/O
  - ★ Secondary receives and stores copies
- Use for
  - ★ Maintenance
  - ★ Planned failover
  - ★ Unplanned failover



Considerations for Cross-Site Connectivity

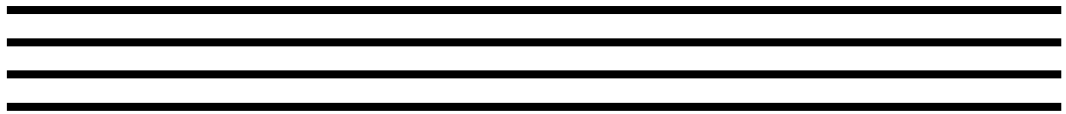
- How much bandwidth is required?
- What is available?
- What is supported for the required workloads?
- What does it cost?
- What is the distance?

Data Replication over OC3/OC30/OC48/ATM/IP

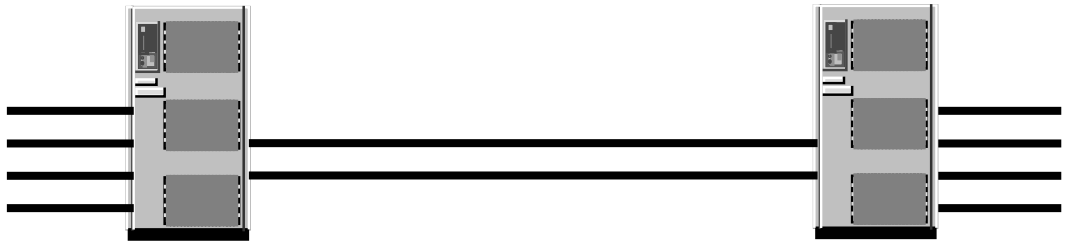


Types of Connectivity – Cross-Site Connection

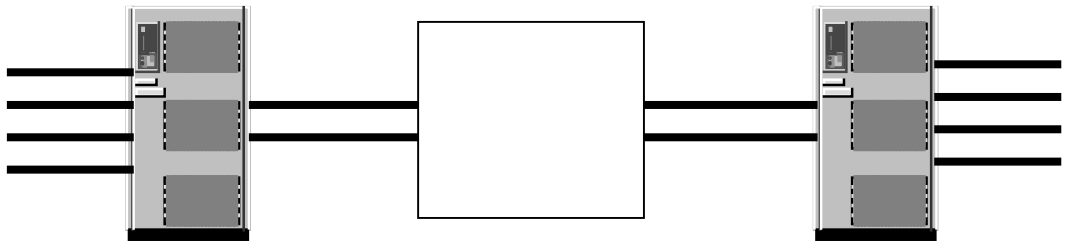
Dark Fiber



WDM/TDM



Channel Extenders (non-optical transport)



Optical Cross-Site Connectivity

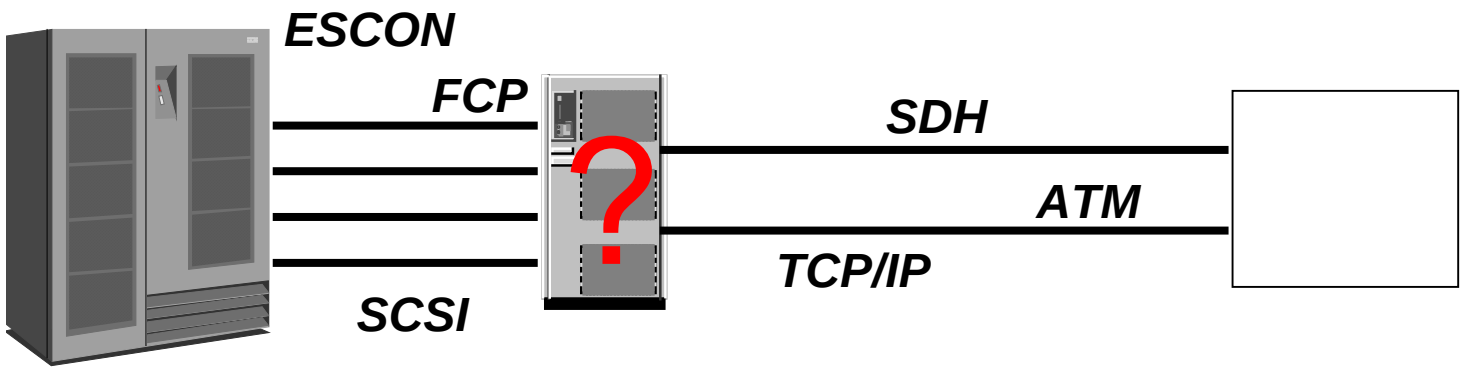
- **Distances driven primarily by optical considerations**
  - ★ Host/device optical signals have limited distance
  - ★ Switches/directors might have enhanced distance capability
  - ★ WDM also provides multiplexing and may allow optical redrive
- **Other considerations**
  - ★ Channel protocol runs end to end
  - ★ Protocol may suffer from droop beyond a certain distance
  - ★ Channel or switch/director provides buffering capability
  - ★ Link throughput will reduce if distance exceed buffer limits

Types of Non-Optical Channel Extender

- **Frame buffering channel extender**
  - ★ Channel protocol runs end to end
    - Protocol will suffer from droop beyond a certain distance
  - ★ Channel extender may provide buffering/compression/retransmission
- **Emulation channel extender**
  - ★ Channel protocol runs separately in each site
    - Channel extender emulates devices / host
  - ★ Channel extender may provide buffering/compression/retransmission

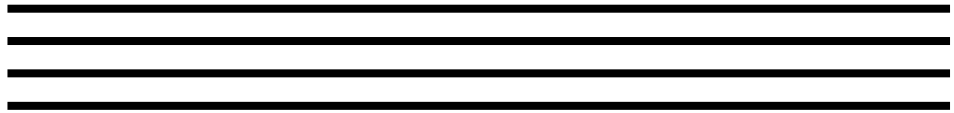
Non-Optical Transports for Storage Protocols - Overheads

- **Overheads exist in both Storage and Network protocols**
  - ★ 149Mb available on 155Mb SDH link
  - ★ 135Mb available on 155Mb ATM link
  - ★ 941.482 Mbps available for TCP on GigE without jumbo frames
  - ★ FC-2 payload is maximum of 95% of frame size
- **Some channel extenders may reduce storage protocol overheads**
  - ★ Emulation can strip data from the protocol and repackage at the other side

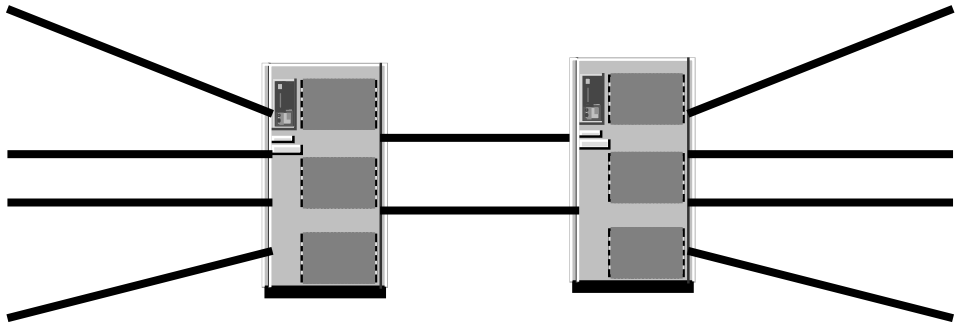


Types of Connectivity – Device Viewpoint

Direct Connection



Via Switches/Directors



- **Connection via switches/directors**
  - ★ Switch/director capabilities may allow for longer unrepeated distances
  - ★ Sharing of cross-site links or device ports may be possible

Non-Optical Transports for Storage Protocols – Network Characteristics

- **Latency**
  - ★ Key issue with most storage extension is latency of network
  - ★ Latency is not always advertised especially on backup routes
- **Resilience**
  - ★ Resilience can mean different things to different people
  - ★ Whether the storage service can run is the key item
- **Bandwidth**
  - ★ Different protocol channel extenders can handle variance differently
  - ★ Bandwidth and useable bandwidth are two different things

Business Continuity Problem

- Synchronous solutions do not work at distance
- Asynchronous solutions have data loss and potential problems managing consistency, particularly across different storage platforms
- Maximizing use of long distance link is critical for many customers
  - ★ Smaller customers may want to purchase extended links which meet maximum transfer requirements for a shift, not their 15 second peak
- Being able to test, fail forward, and fail back is critical

Need to give customers new solutions!

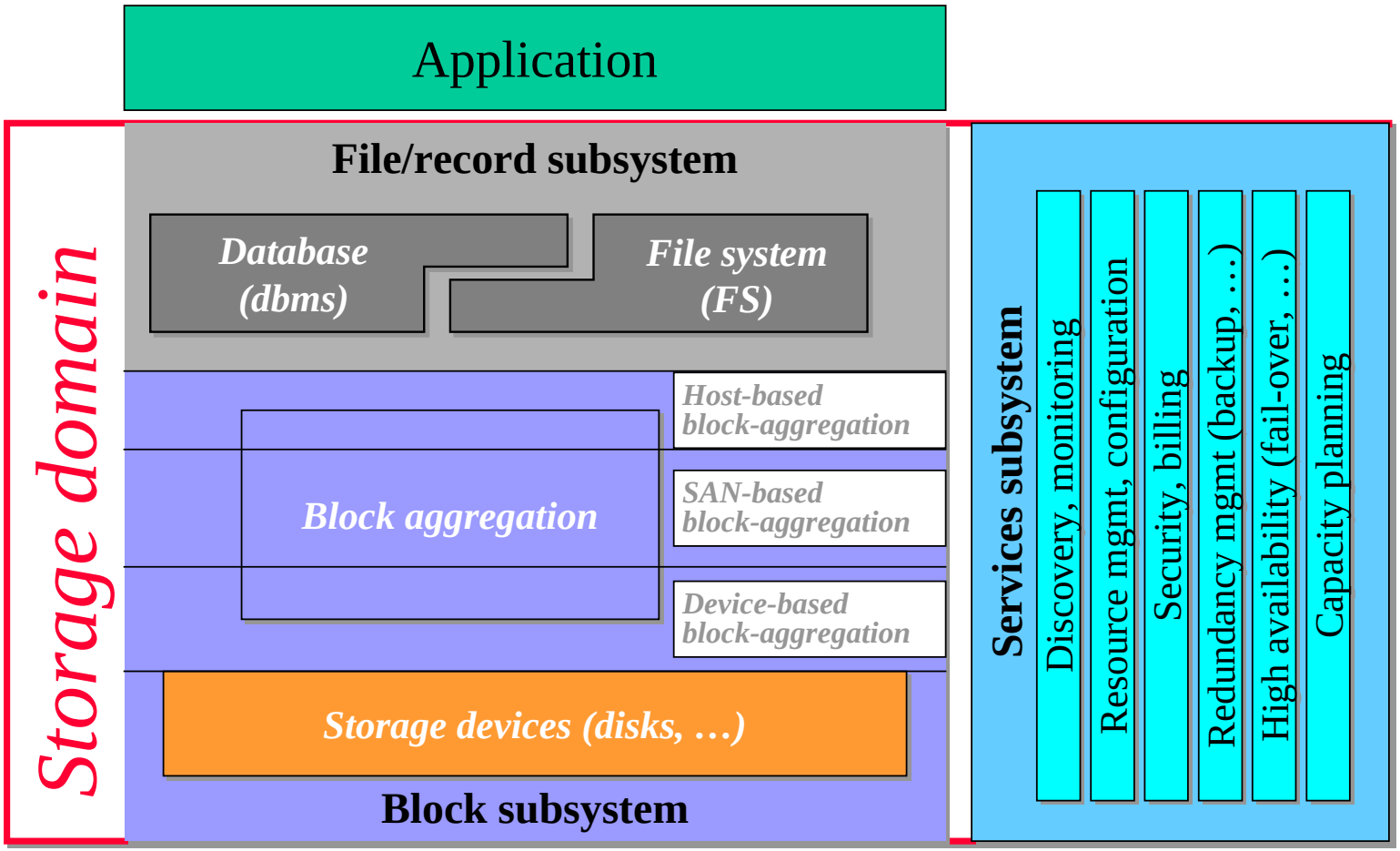
Marketplace Objectives

- Reduce TCO for data replication
- Storage vendor interoperability for data replication
  - ★ Any to any high-end, mid-range, low-end
- Reduce costs for producing data replication functions

Marketplace Observations

- **Could Drive NEW Data Replication Scenarios & Management Flexibility**
  - ★ Low cost solution to move data from local or regional offices to main data center
  - ★ Data migration/movement solution for distributed data consolidation efforts (simple install, simple day to day remote operation)
  - ★ Inter-operability across storage vendors disks yields customer choice & preserves current investments

The SNIA Shared Storage Model



SNIA SMI-S Standards being extended for Copy Services

Marketplace Observations

- Market Opportunity - Switch, Channel Extender, Software, Storage Vendors
- Several new startup companies
- May be combined with emerging virtualization products
- Technology not yet "proven" in the marketplace
  - ★ Cross volume/cross subsystem data integrity/data consistency issues can be a problem
- No interoperability with existing solutions
- Generally these companies do not participate in SNIA Copy Services Standards work

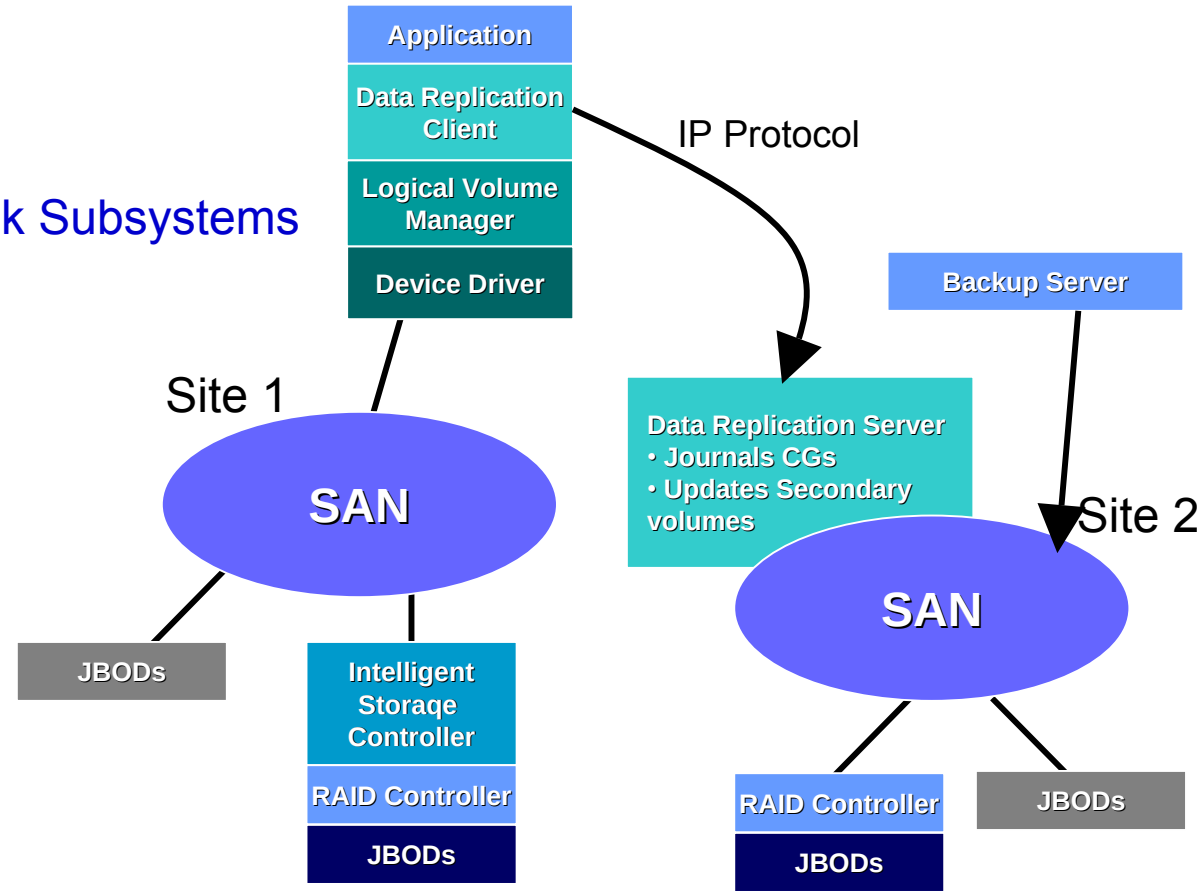
Emerging Data Replication Architectures

- Host Based
- Switch Based
- SAN Replication Appliance within Data Path
- SAN Replication Appliance outside Data Path
- Storage Subsystem Peer

Host Based Data Replication

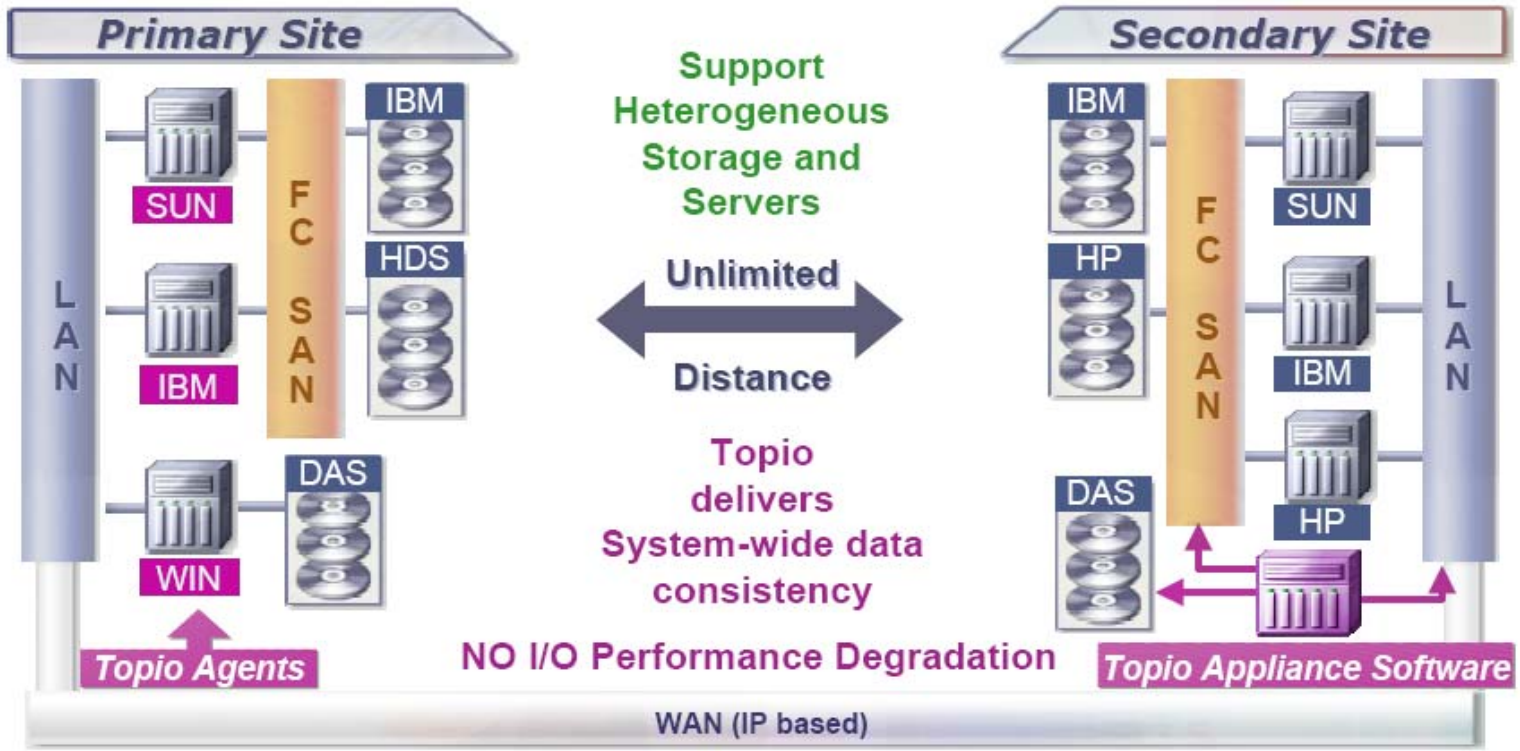
- Server Dependent
- Host Mips - typically 3-5%
- In Data Path
- Interoperability Between Disk Subsystems

- File/DB Subsystem Based
- Application Based
- DR Client Above LVM
- DR Client in Device Driver



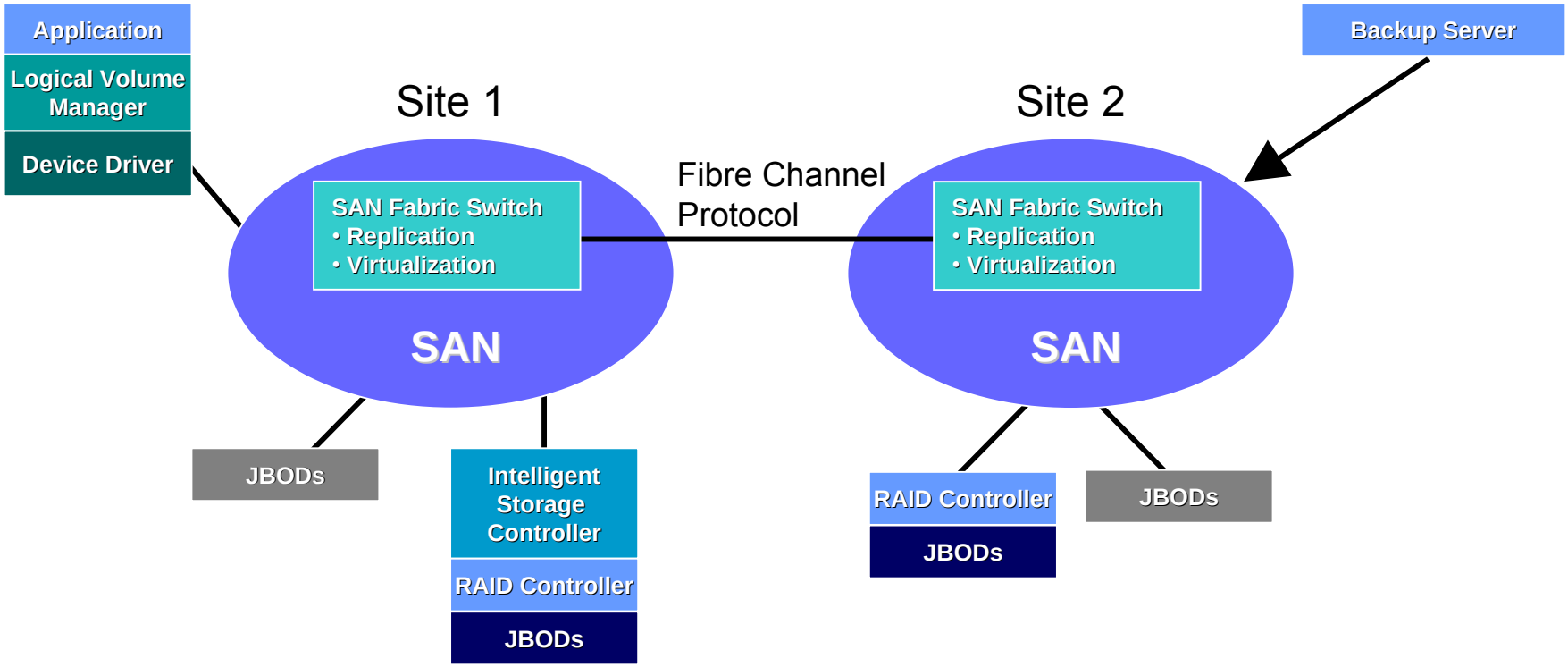
Topio

- Topio agents installed in all primary hosts
- All writes are also transferred to a single Topio appliance at secondary site
- The Topio appliance applies the data to the proper location



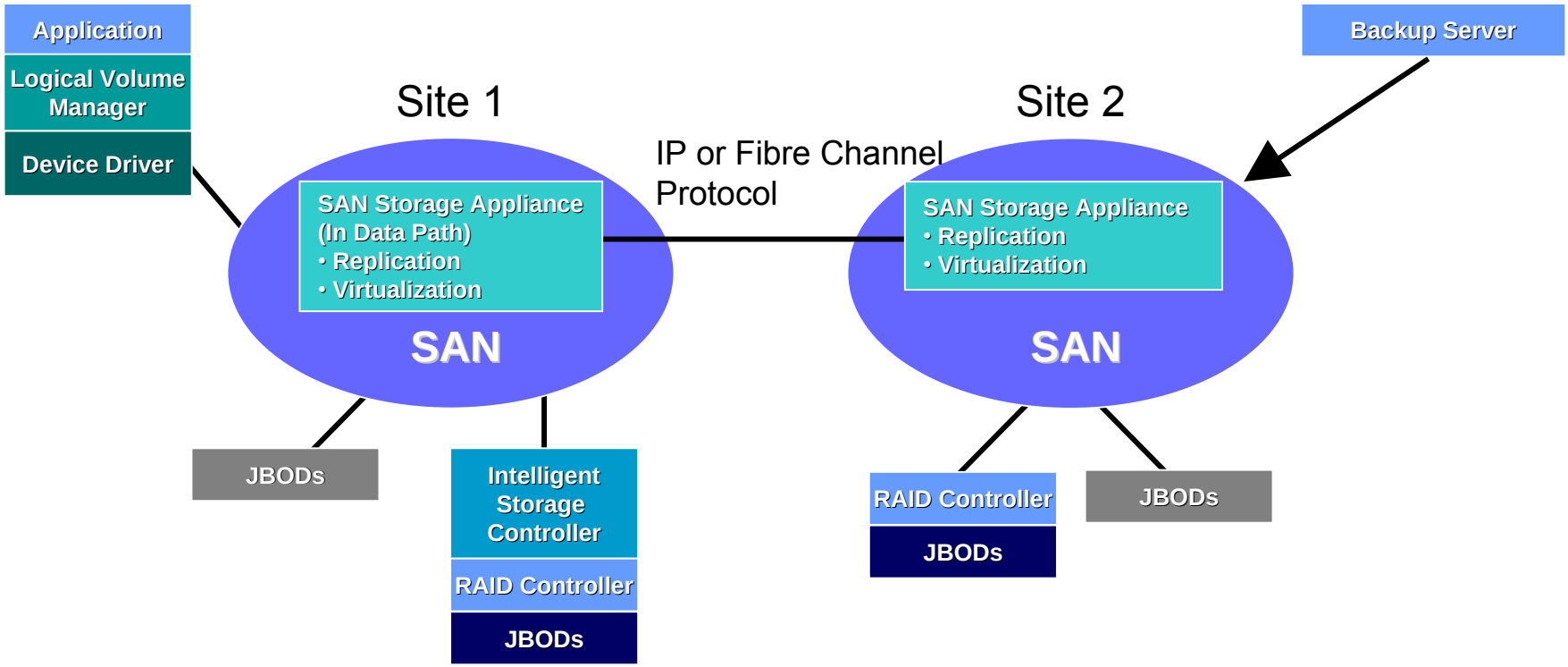
Switch Based Data Replication

- In Data Path
- Multi-Switch Function Management
- Within Existing Enterprise Box
- Interoperability across Disk Subsystems



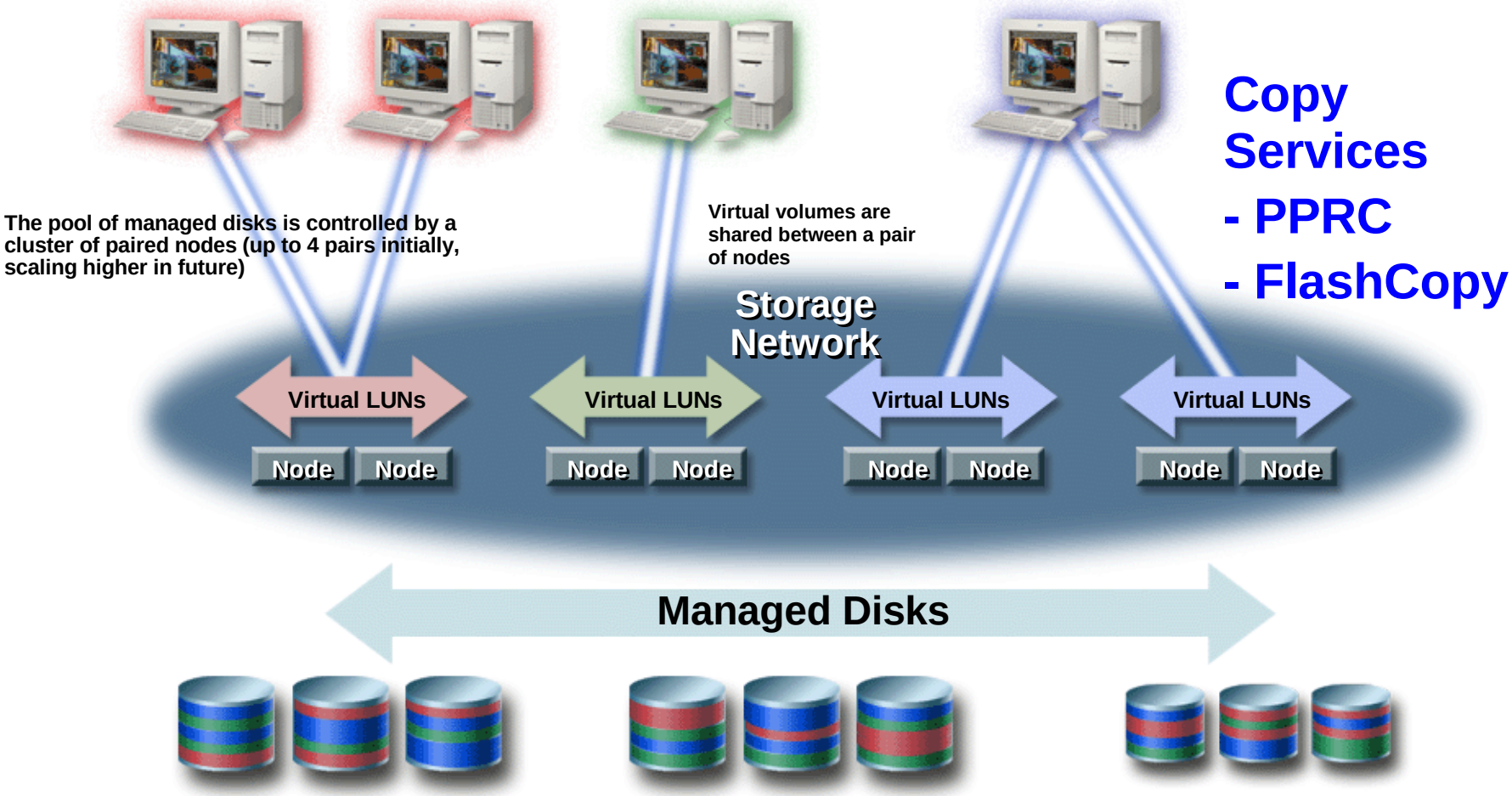
SAN Storage Appliance (Within Data Path)

- In Data Path
- New Box to Manage in Enterprise
- Interoperability across Disk Subsystems



IBM's Virtualization Engine

Redundant, modular, scalable, complete solution



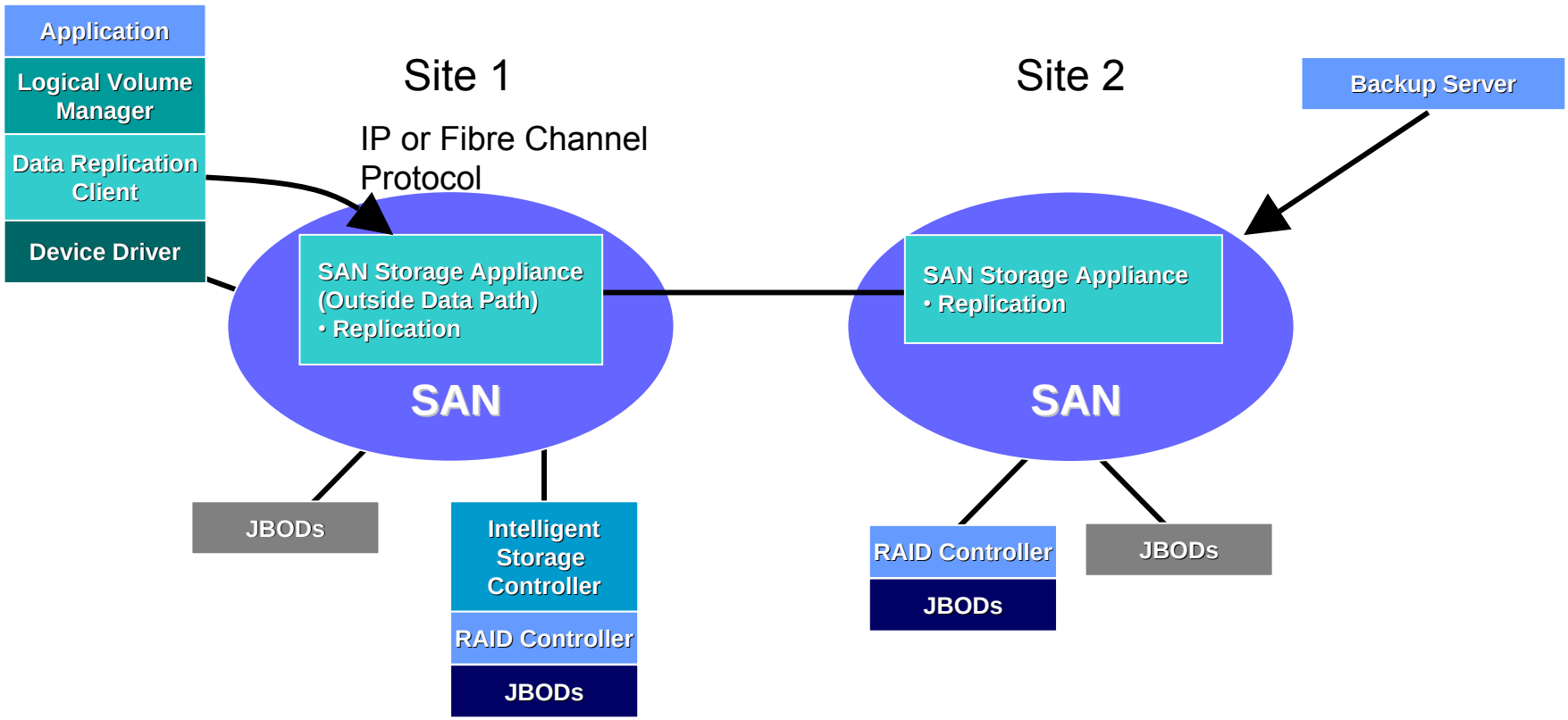
Copy Services

- PPRC

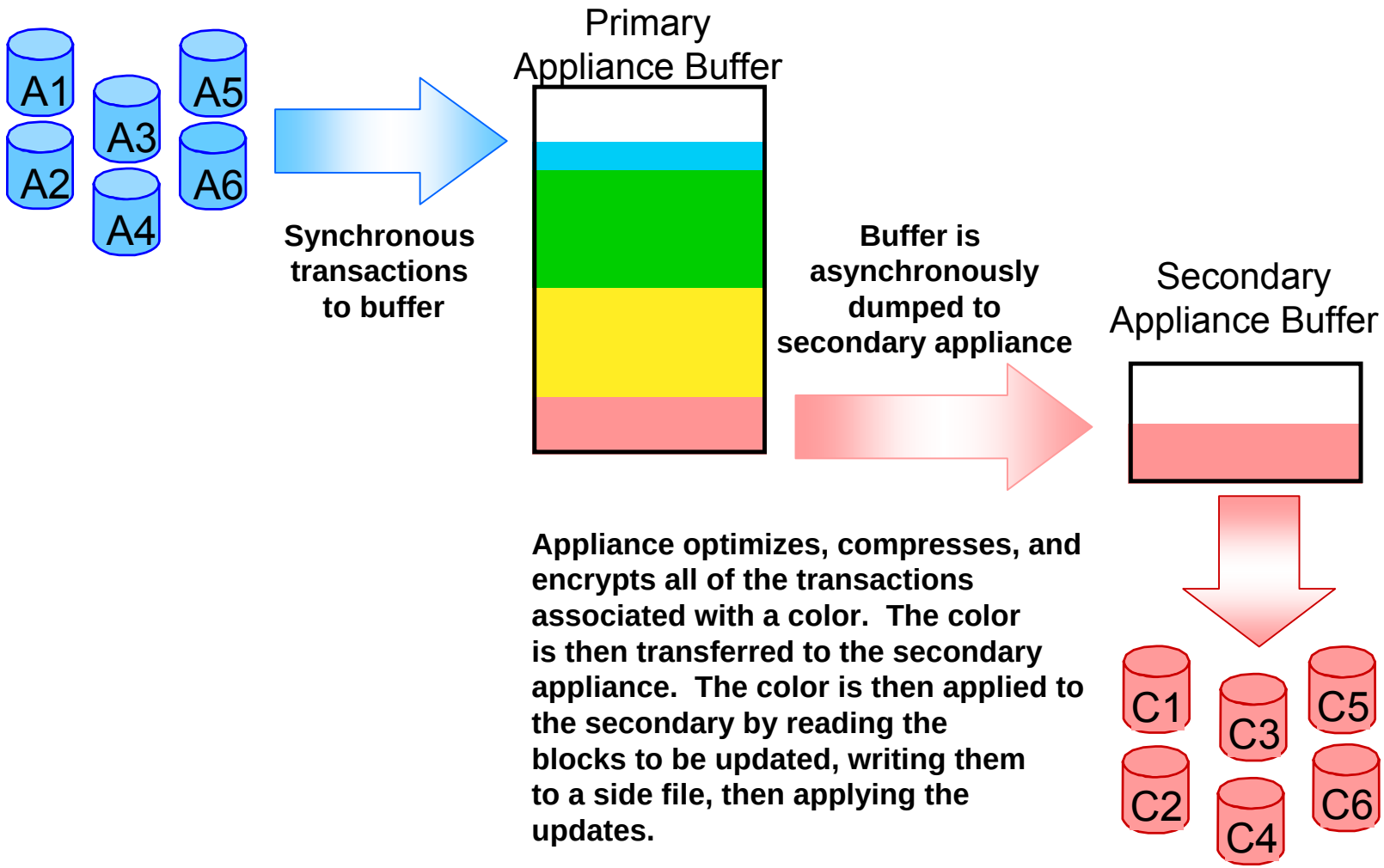
- FlashCopy

SAN Storage Appliance (Outside Data Path)

- Outside Data Path
- New Box to Manage in Enterprise
- Host Client Code Required
- Interoperability across Disk Solutions



Forming Point-In-Time Consistency

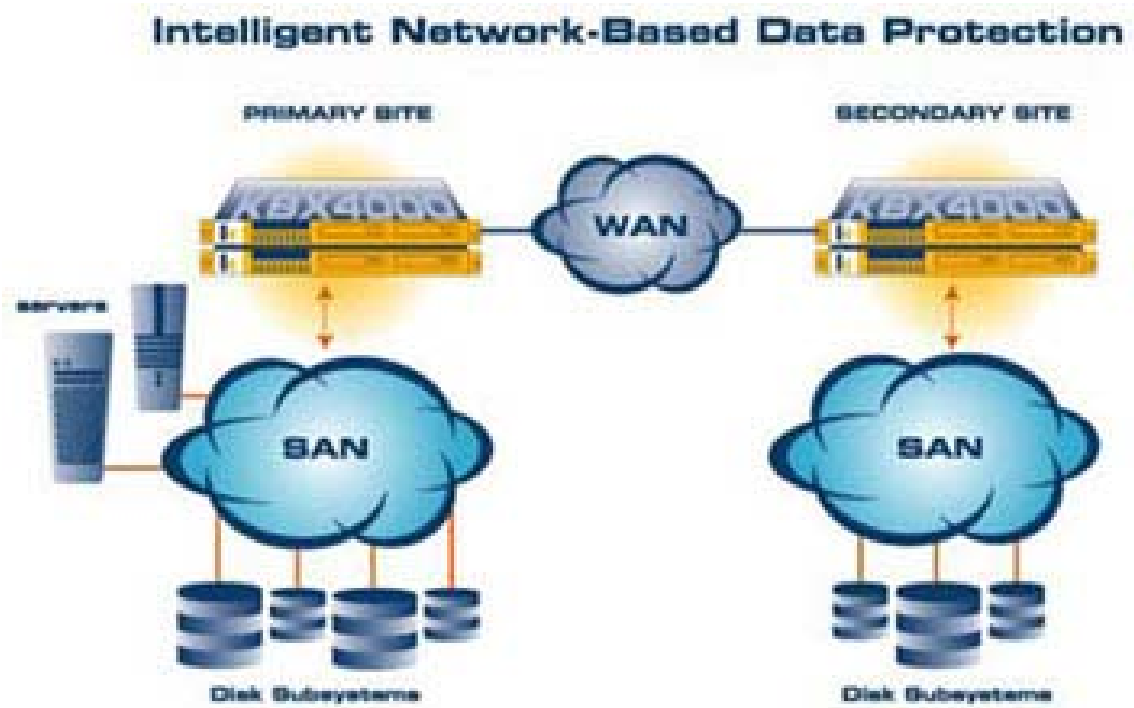


Point-in-Time Consistency Groups

- A consistency group is a group of I/Os which represent a consistent "point-in-time" view of the data
- Appliance Optimizations:
  - ★ Optimize transactions in a consistency group
    - Eliminate blocks which have been multiply written
    - Form large blocks for efficient transmission over extended link
  - ★ Compress/Encrypt data between appliances
  - ★ Apply consistency groups without the use of flash technology (keep multiple versions)
    - Includes things such as: Beginning of Day, Beginning of Hour, last 10 consistency groups, etc

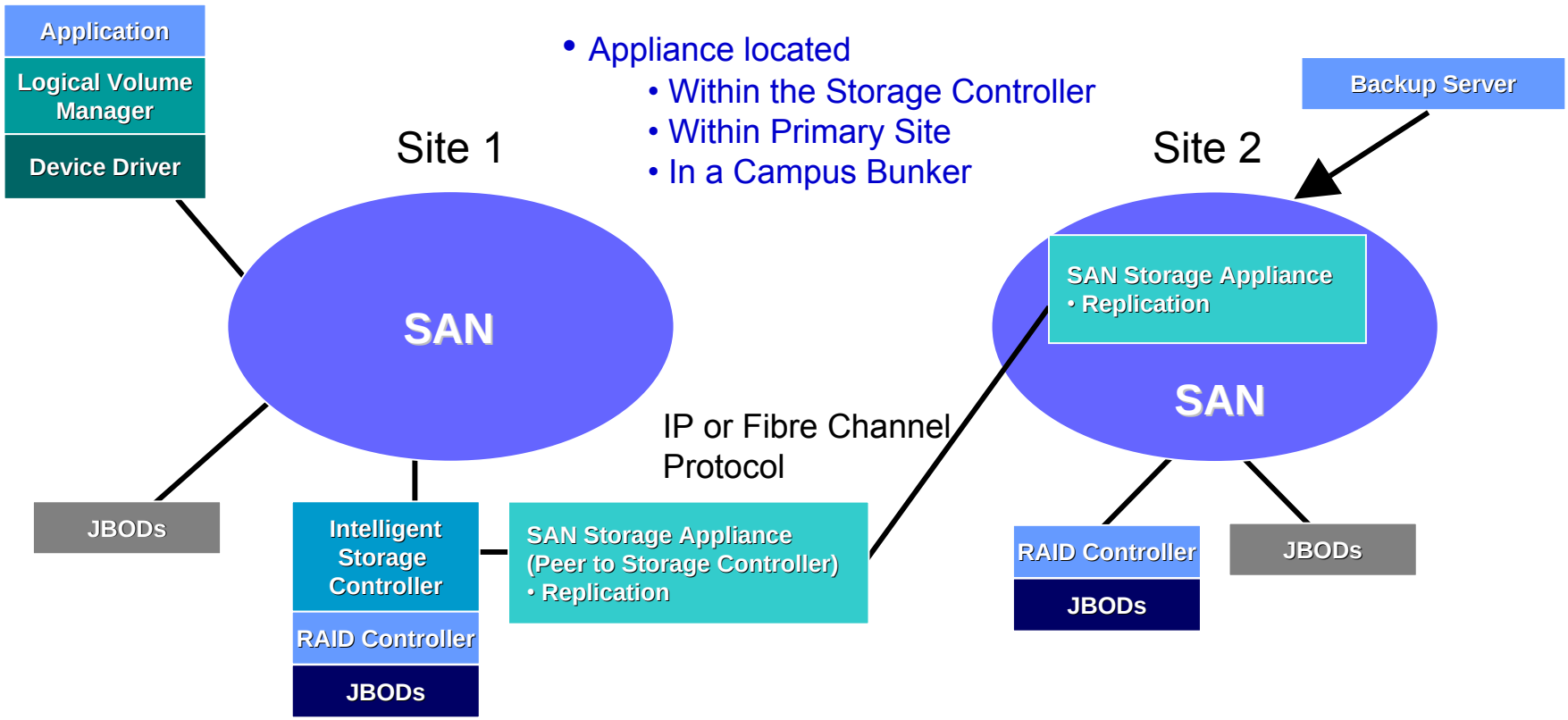
Kashya

- Device driver installed in all hosts
- Each write to a replicated LUN is first sent to the appliance
- After successfully received by the appliance, write is sent to CU
- Consistent sets of data applied to secondary site
- Ability to roll state of secondary site backwards and forwards

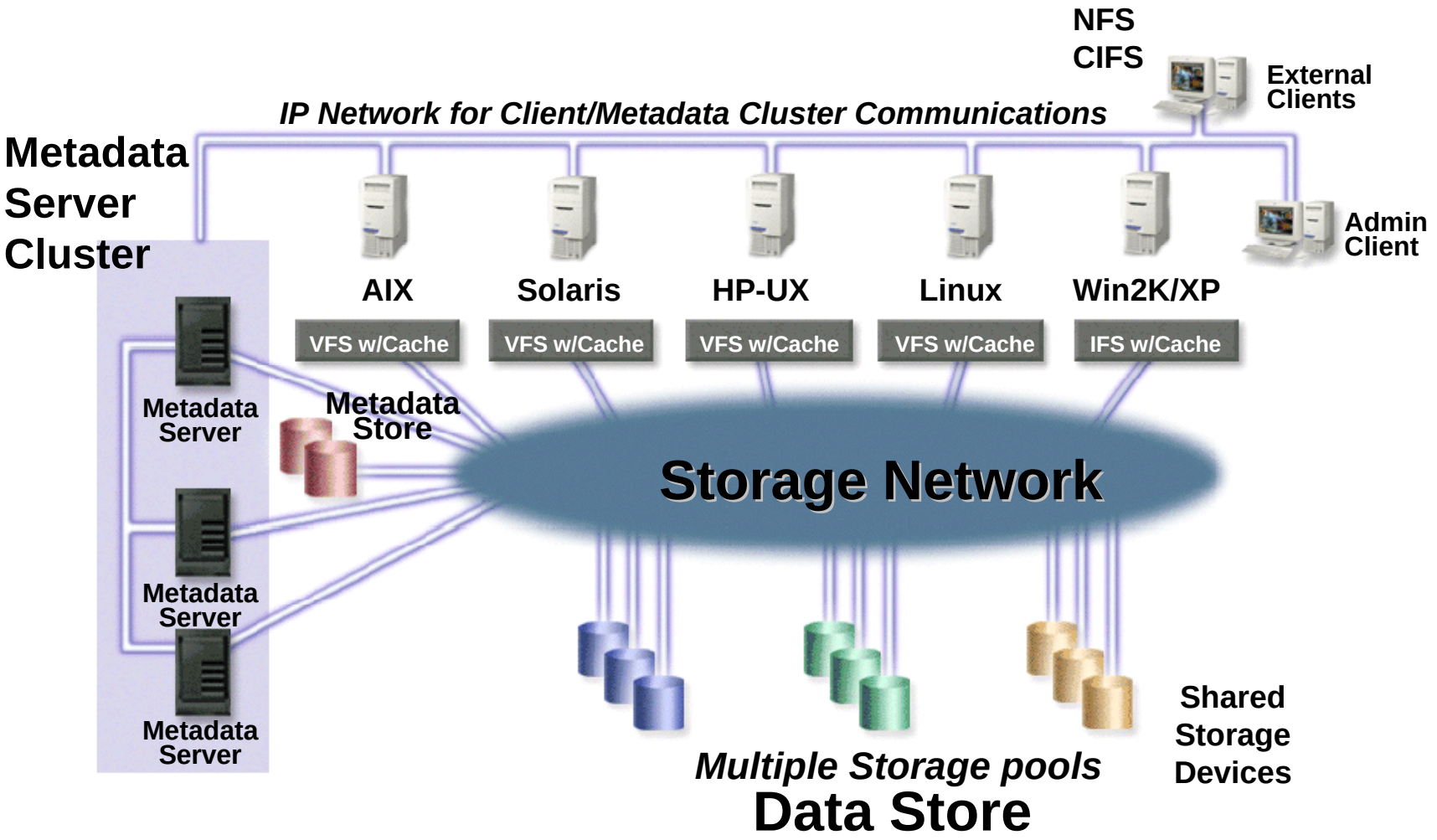


Storage Appliance (Peer to Storage Controller)

- Outside of Data Path
- No Host Client Code Required
- Requires Storage Subsystem Support



IBM's SAN File System



Key Questions for Any Solution

- How does the solution provide cross volume/cross subsystem data integrity/data consistency ?
- What is the impact to the primary application I/O ?
- What happens if data replication fails or slows down ?
- Interoperability with other data replication solutions ?
- Cost of installing & maintaining solution ?
- Do solutions within data path provide “concurrent maintenance” ?
- What flexibility does the solution provide ?
- If I failover, how do I failback ?
- If I use different “types" of disk subsystems, in a failover can I maintain my QoS to my users ?
- Others ...

Discussion

- How has 9/11 affected your DR plans, if at all?
- In your businesses, what do you feel is more important
  - ★ Long distance separation of data sites?
  - ★ Ensuring RPO of 0?
  - ★ Has this changed at all in the past few years?
- How difficult is it to manage your storage infrastructure?
- Do you have resources (hardware, people, time) to practice your DR plans?
  - ★ Do you actually practice?
  - ★ Would you like to?

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

AIX®
IBM®
IBM logo
Database 2
DFSMS/MVS®
Enterprise Storage Server®
ESCON®
eServer®
FICON®
GDPS®
Geographically Dispersed Parallel Sysplex®
HACMP/6000®
HyperSwap®
Enter

MVS/ESA
NetView®
OS/390®
Parallel Sysplex®
Remote Copy Management Facility®
S/390
System/390®
Sysplex Timer
zSeries®

The following are trademarks or registered trademarks of other companies.

Lotus, Notes, and Domino are trademarks or registered trademerks of Lotus Development Corporation
Tivoli is a trademark of Tivoli Systems Inc.
Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries
UNIX is a registered trademark of The Open Group in the United States and other countries.
Microsoft, Windows and Windows NT are reegistered trademarks of Microsoft Corporation.
SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

IBM considers a product “Year 2000 ready” if the product, when used in accordance with its associated documentation, is capable of correctly processing, providing and/or receiving date data within and between the 20th and 21st centuries, provided that all products (for example, hardware, software and firmware) used with the product properly exchange accurate date data with it. Any statements concerning the Year 2000 readiness of any IBM products contained in this presentation are Year 2000 Readiness Disclosures, subject to the Year 2000 Information and Readiness Disclosure Act of 1998.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Data Replication Technology

John Wolfgang

John Wolfgang	jwolfgan@us.ibm.com	520-799-4819
Bob Kern	bobkern@us.ibm.com	520-799-5465
Ken Boyd	kenboyd@us.ibm.com	520-799-2720
Ken Day	mycroft@us.ibm.com	520-799-4582

IBM Storage Systems
Tucson, AZ

NASA/IEEE MSST 2004
12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies
The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004