Designing Continuous Access Enterprise Virtual Array Solutions

Don Fraser
Disaster Tolerant Solutions Architect
Continuous Access EVA Product Engineering
hp Network & Server Storage

HP WORLD 2003
Solutions and Technology Conference & Expo

Agenda

- Supported configurations
- Understanding the requirements
- Physics of distance and Performance Estimation
- Best practice design for the storage

- questions taken throughout

Supported solutions

Today, each array pair is limited to

- Maximum of 64 copy sets
- Maximum of 64 DR group
- Maximum of 8 copy sets per DR group
- Minimum of 1 disk group per storage system; maximum of 16 (or number of drives divided by 8,which ever is lower)
- 128 HBA/FCA, 64 per fabric = 64 servers
- Maximum of 512 Vdisks per disk group. 1-GB to 2-TB copy set size
- Maximum of 16 storage systems in the SAN

Supported solutions

Today, each array pair is limited to (continued)

- One replication relationship per array pair
- Maximum of 7(Brocade)/3 (McDATA) switch hops between source and destination storage arrays
- All DR groups in a disk group must copy to the same disk group at the destination site
- Source and destination disk groups need not have the same geometry and must be of the same size
- Maximum of 8 snapshots or snapclones per DR group at the source or destination site and up to 7 per source vdisk

Supported solutions

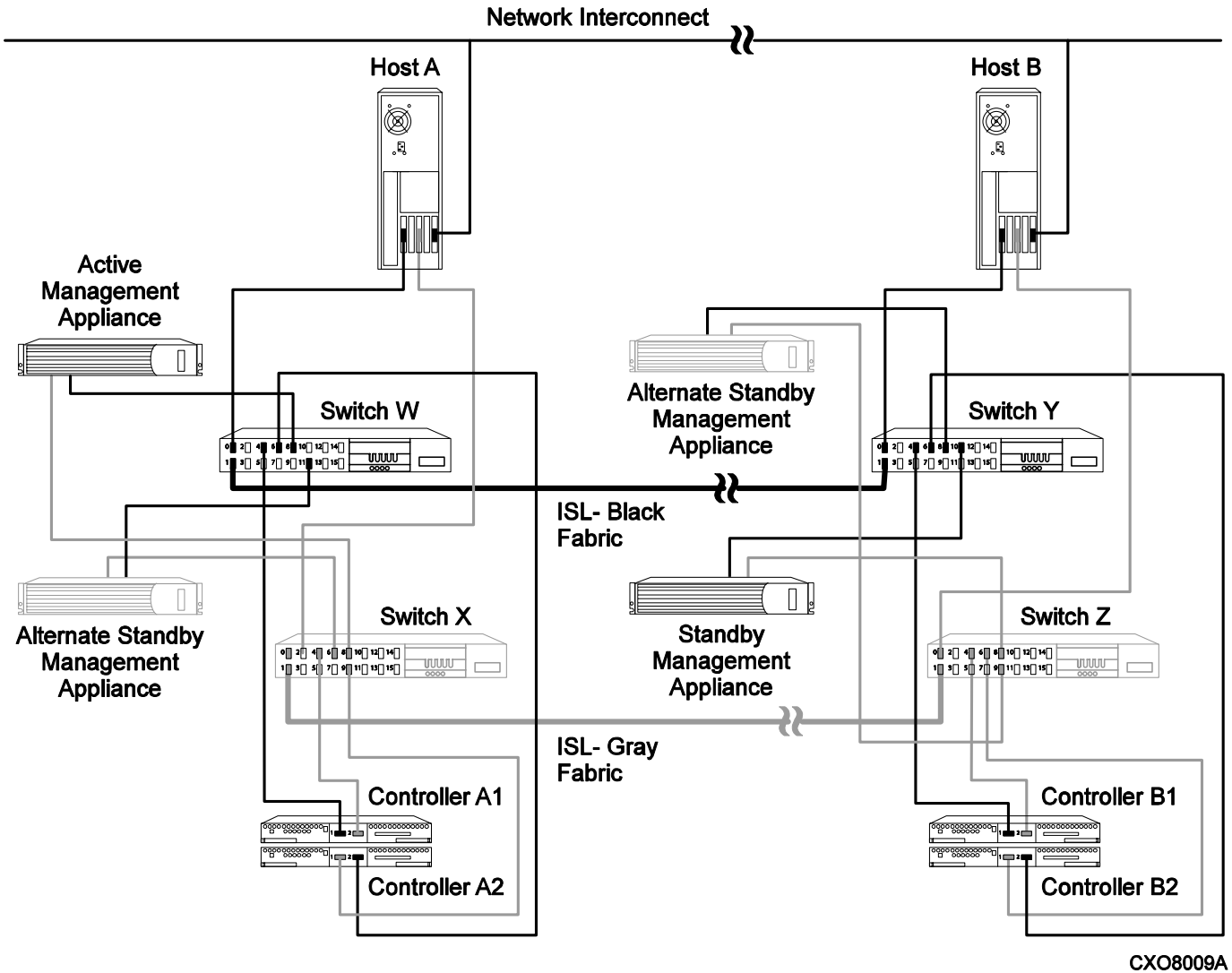
Today's Supported Operating Systems

- hp OpenVMS 7.2-2, 7.3-1
- hp Tru64 5.1a. 5.1b
- hp HP-UX 11.0, 11.11
- IBM AIX 4.3.3, 5.1
- Microsoft Windows: NT; 2000; 2003 (32 bit)
- Novell Netware 5.1, 6.0
- Red Hat AS 2.1
- SUN Solaris 2.6, 7, 8, 9
- SuSE SLES 7, 8

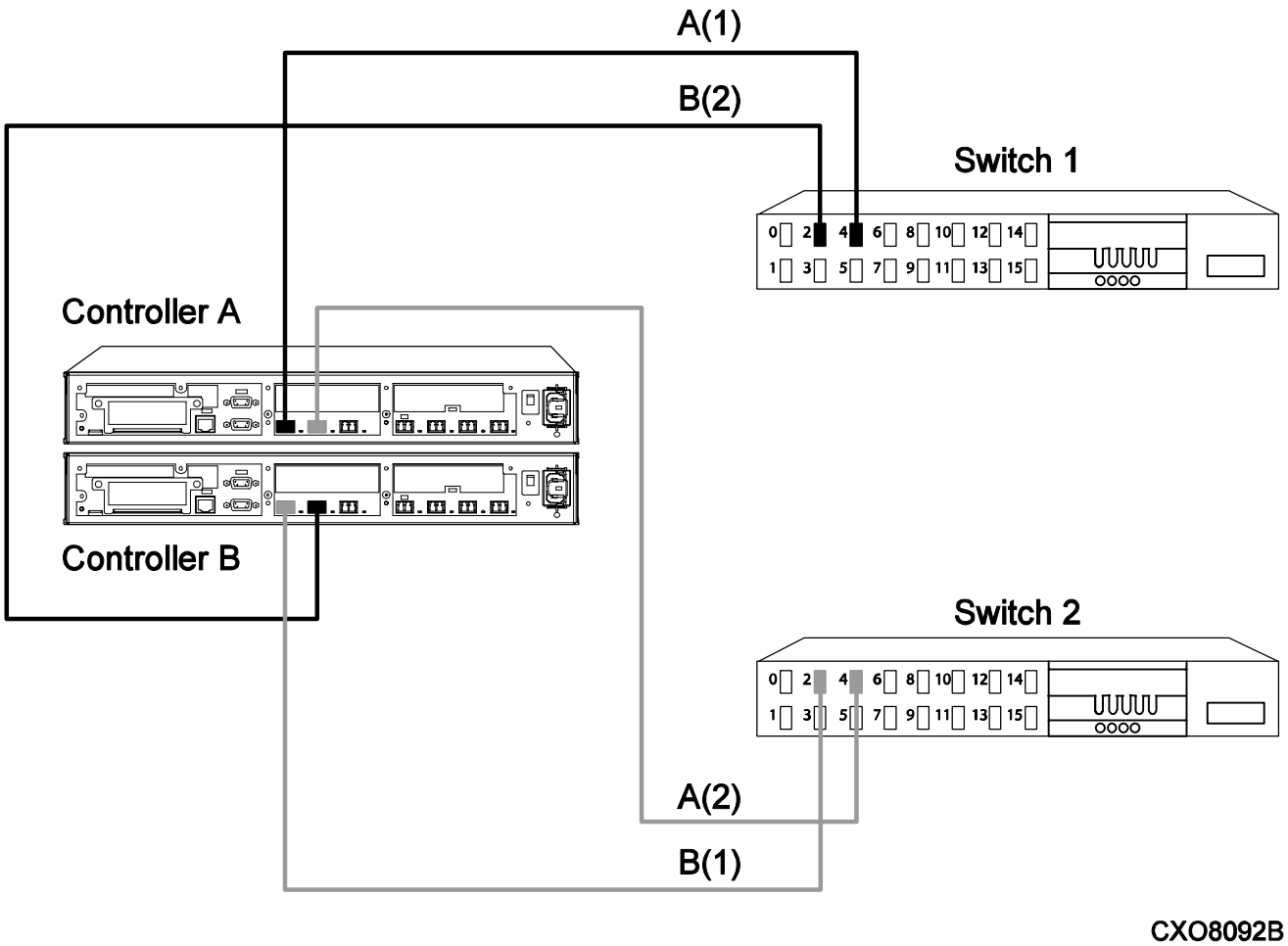
Supported solutions

- Basic solution
  - direct fiber
  - using WDM
  - using FCIP
  - stretched cluster
- Optional solutions
  - single HBA
  - single switch
  - single fabric
- Advanced solutions
  - open discussion

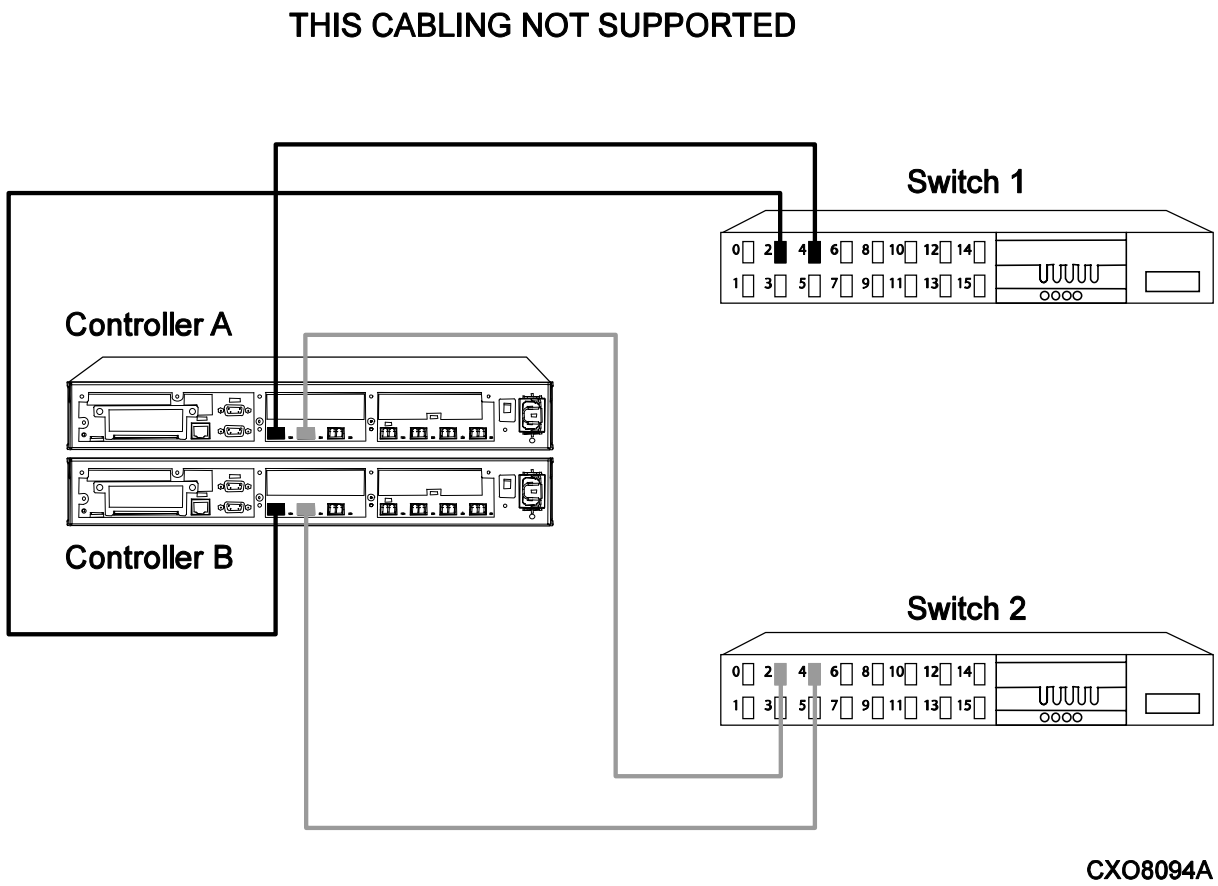
Basic Continuous Access over-fiber configuration



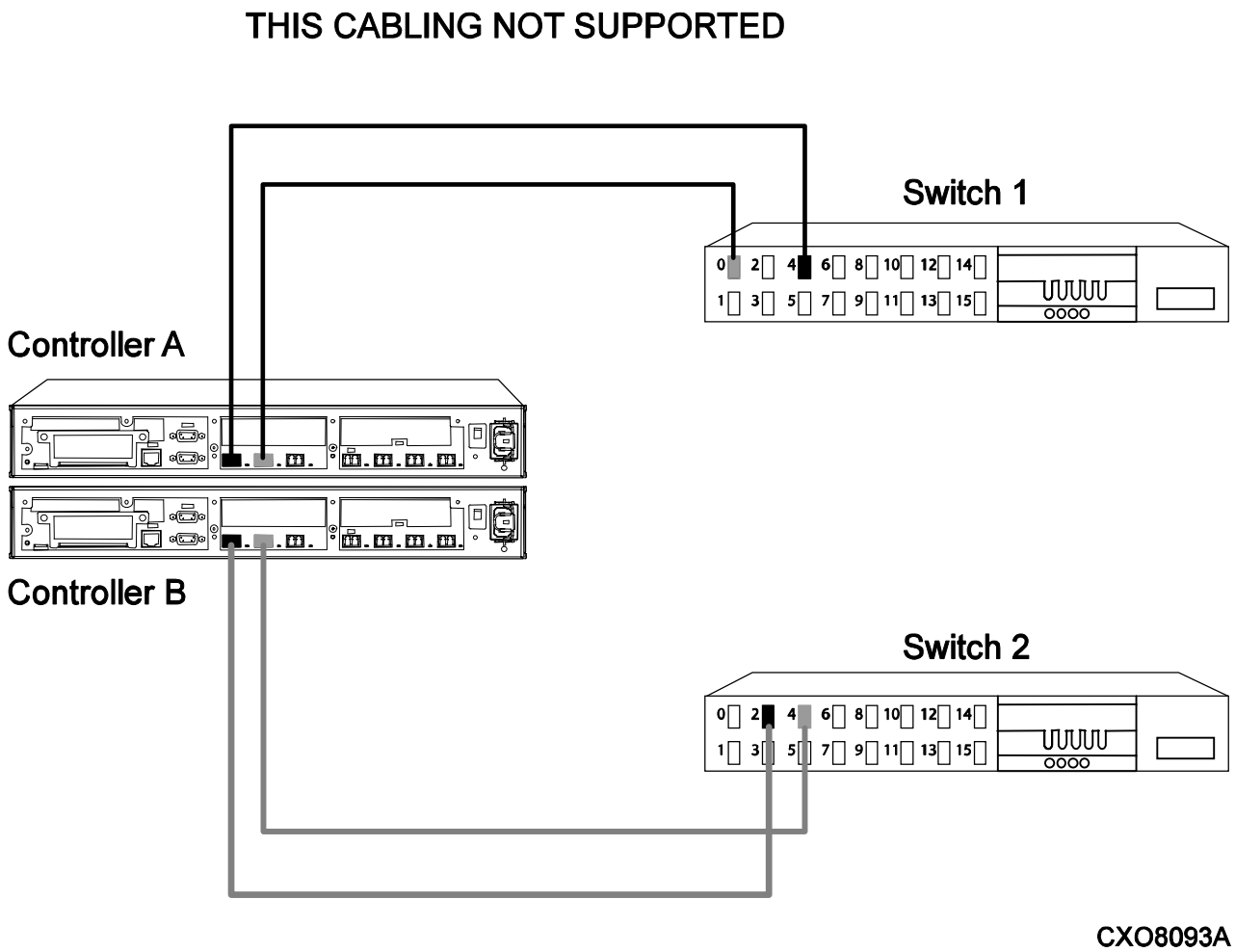
Supported cabling



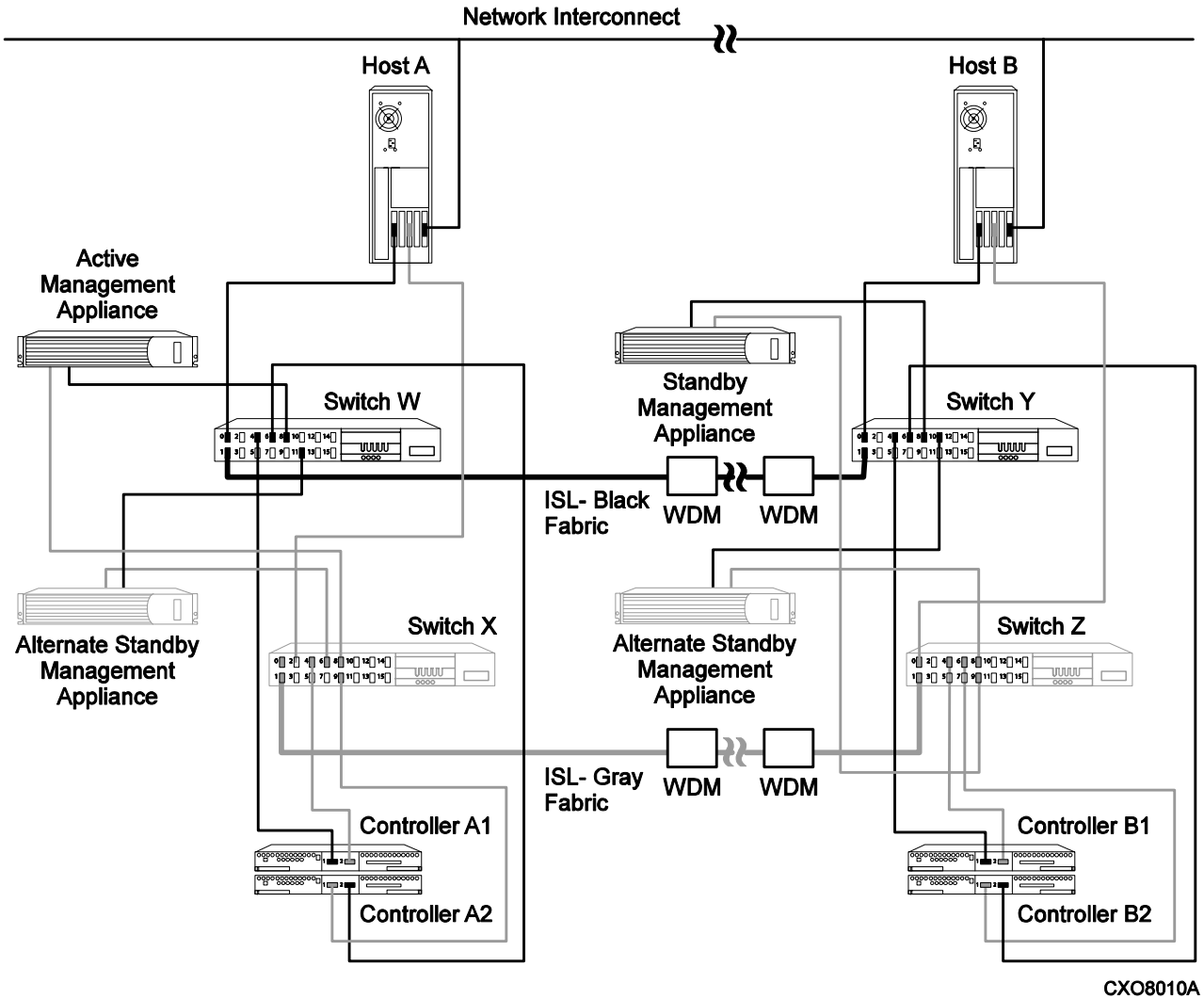
Example 1 of cabling that is not supported



Example 2 of cabling that is not supported



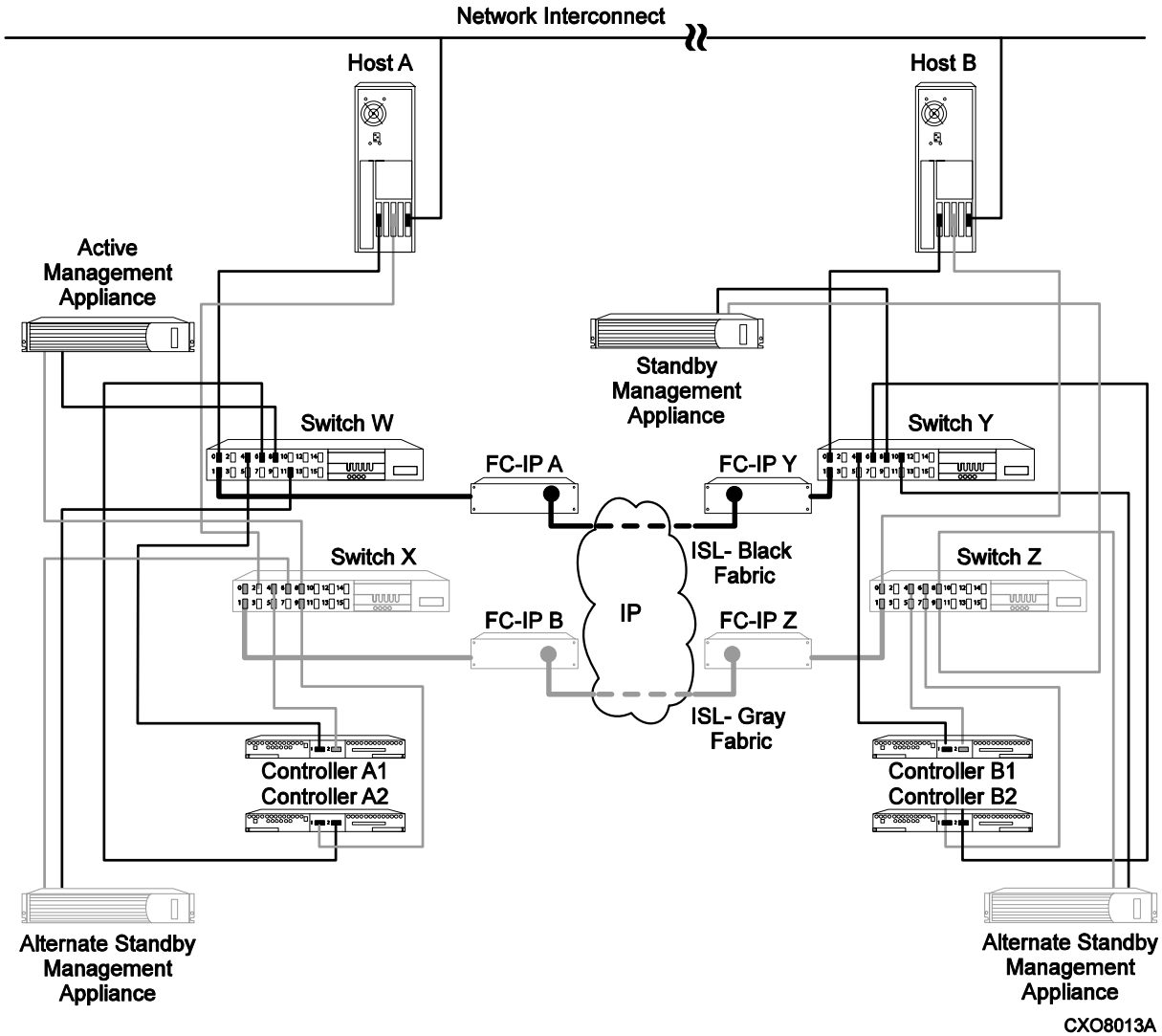
Continuous Access EVA-over-WDM configuration



Continuous Access EVA-over-WDM configuration (cont)

- All WDM extensions regardless of vendor, or type of technology are supported.

Continuous Access EVA-over-IP or SONET configuration



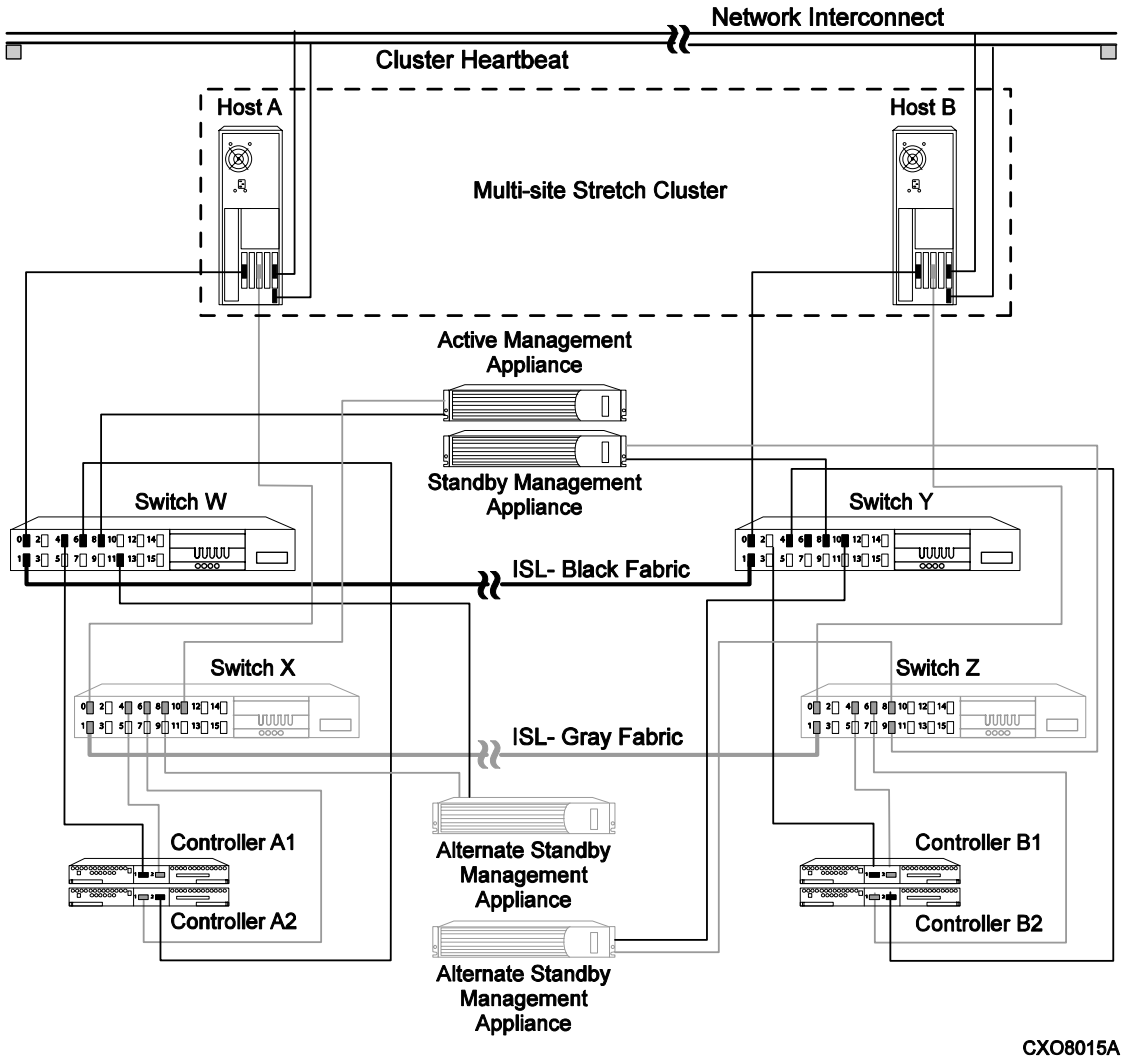
Continuous Access EVA-over-IP or SONET configuration (cont)

- Supported Gateways;
  - CNT
    - 1001, 1101 FC to GbE IP
  - SAN Valley
    - SL 700 single channel FC to GbE IP
    - SL1000 dual channel FC to GbE IP
  - Akara
    - Optical Utility Services Platform (OUSP) 2000 family
    - supports STS-1 thru OC-48
  - Cisco
    - PA-FC-1G Fibre Channel Port Adapter
    - installable into the Cisco 7200 VXR and 7401ASR routers
    - supports 10/100/1000 Mbps Ethernet

Stretched clusters

- One cluster split over two sites
- Two halves are separated by more than
  - 300 m at 2 Gbps
  - 500 m at 1 Gbps
  - the limit of a short wave fiber connection
- Supported for Windows and Tru64
  - Windows to 100 km (62.5 miles)
  - Tru64 to 6 km

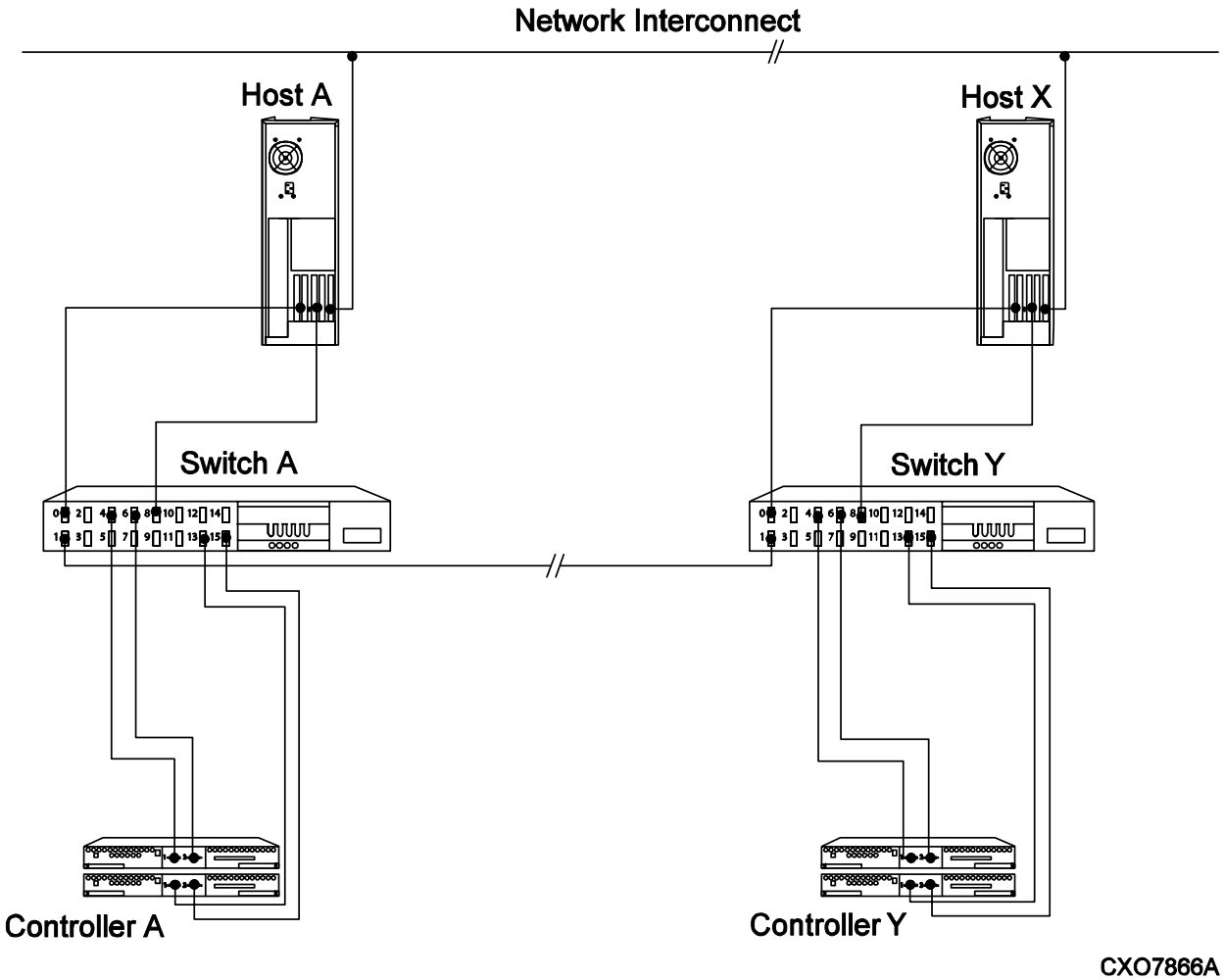
Continuous Access EVA stretched cluster configuration



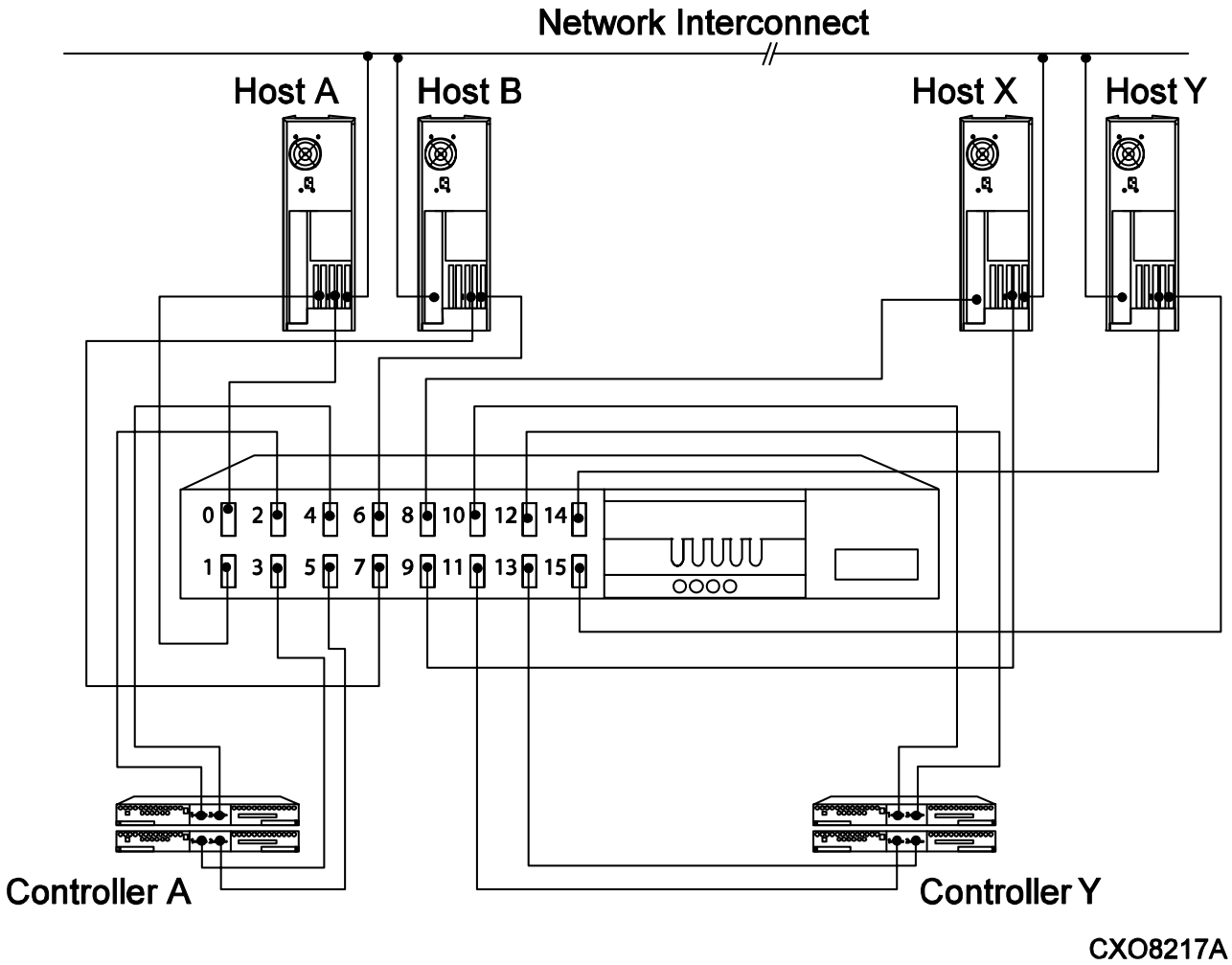
Optional configurations

- Single HBA
  - multi-path driver required
- Single fabric
- Single switch

Single-fabric configuration



Single-switch configuration



Advanced configurations

- Review of Today's Limits
  - 16 EVA (8 pairs of arrays)
  - 128 HBA per replicating array pair
    - 64 servers @ 2 HBA per server
    - need not be symmetric
  - bi-directional replication between array pairs

Advanced configurations

- Lab is looking at
  - 128 copy sets, DR groups
  - multi-relationship replication - different LUNs
    - same source array, different destination (A->B, A->C)
    - same destination array, different source (B->A, C->A)
    - cascaded (A -> B -> C)
  - Asynchronous replication

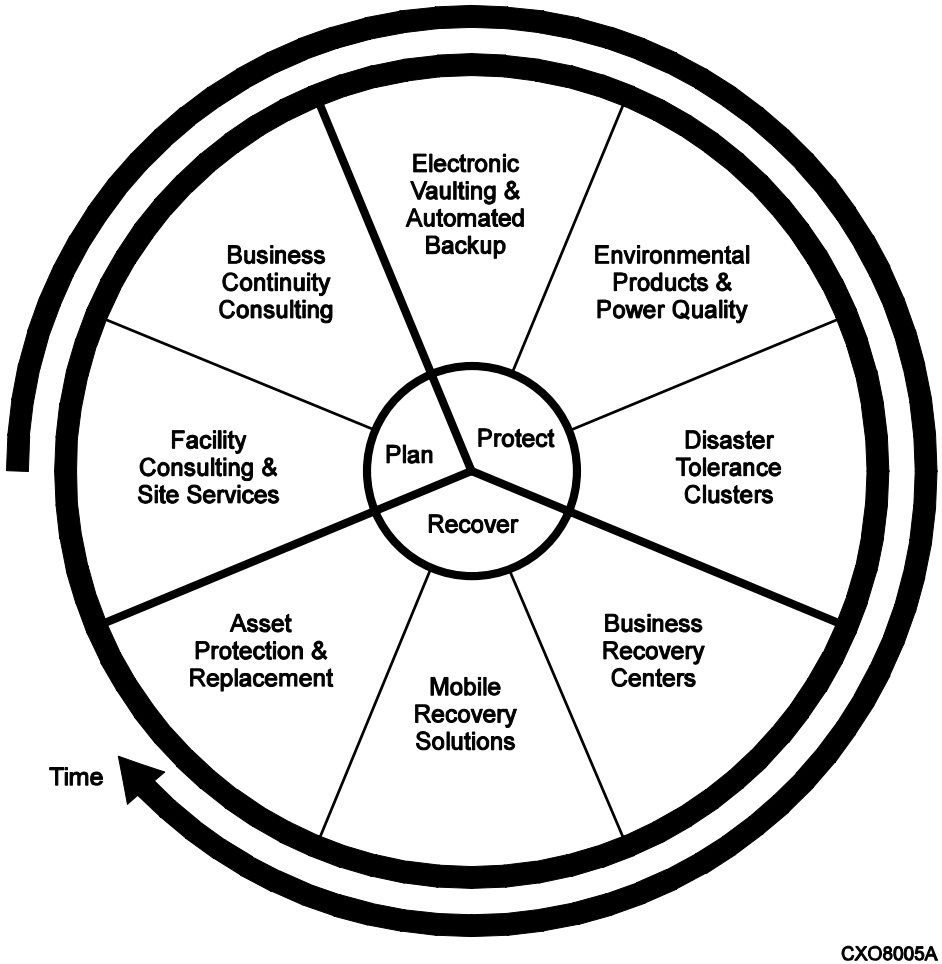
Agenda

- Supported configurations
- **Understanding the requirements**
- Physics of distance and performance estimation
- Best practice design for the storage

Know the requirements

- Continuous Access EVA is
  - about protecting the data
  - about being part of a plan
- Continuous Access EVA is not
  - about high availability per se

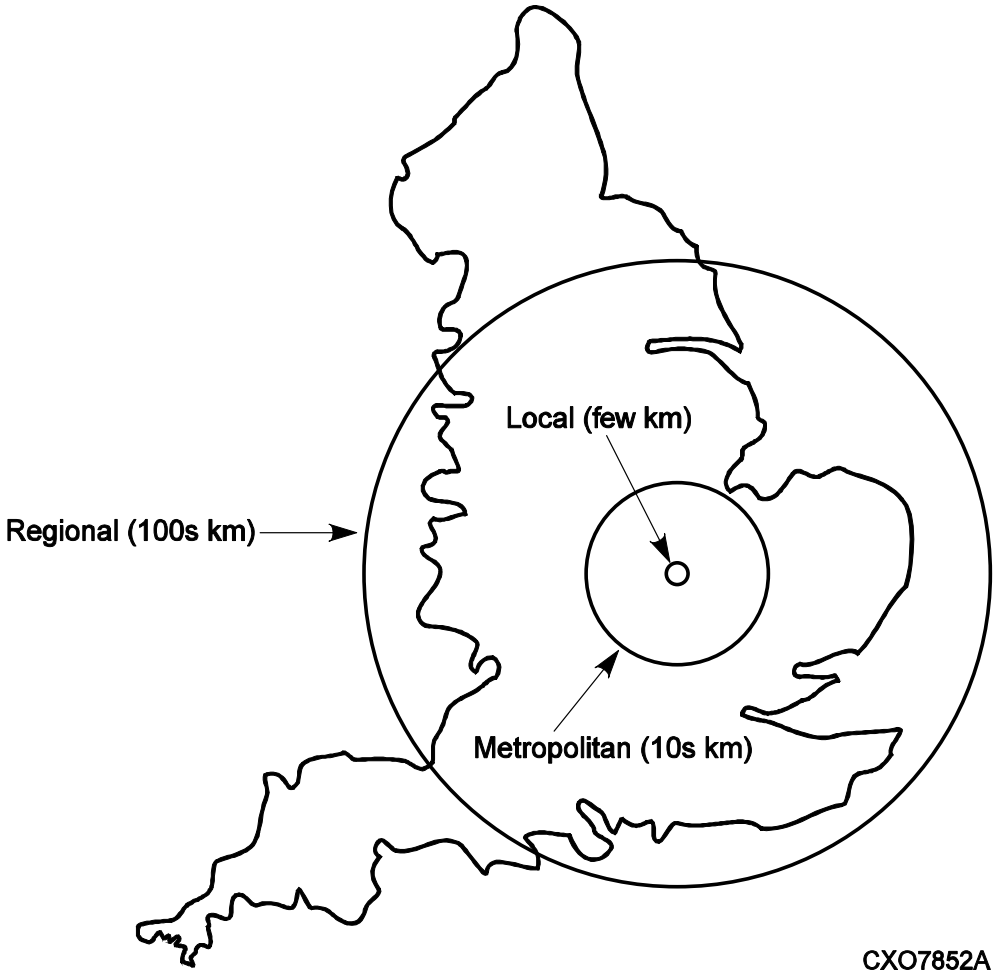
Business continuity model



What are the threats?

- Local
  - Less than10 km/10 miles
  - Such as a fire
- Metro
  - To 10s of km or 10s of miles
  - Small flood or localized storm
- Regional
  - To 100s of km or 100s of miles
  - Major flood or large storm

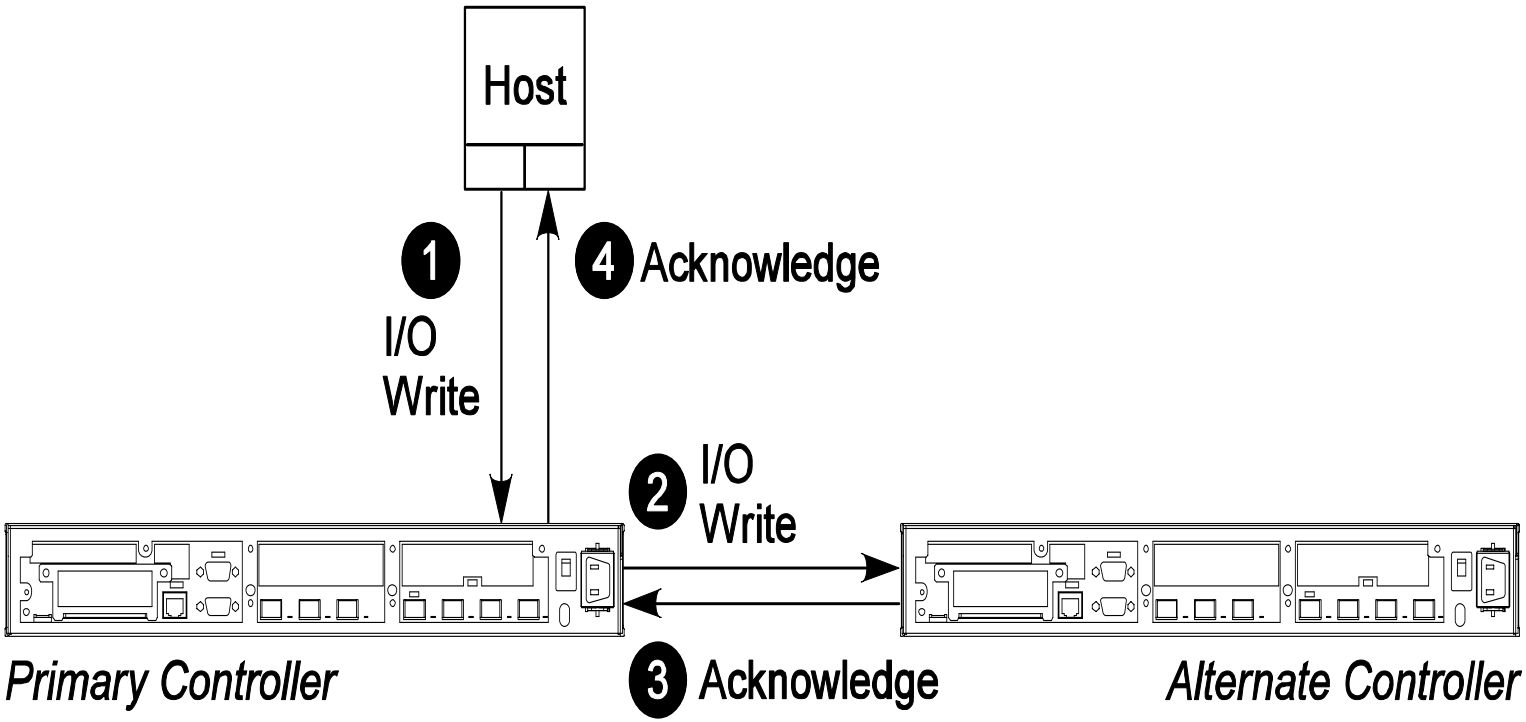
Threat radius



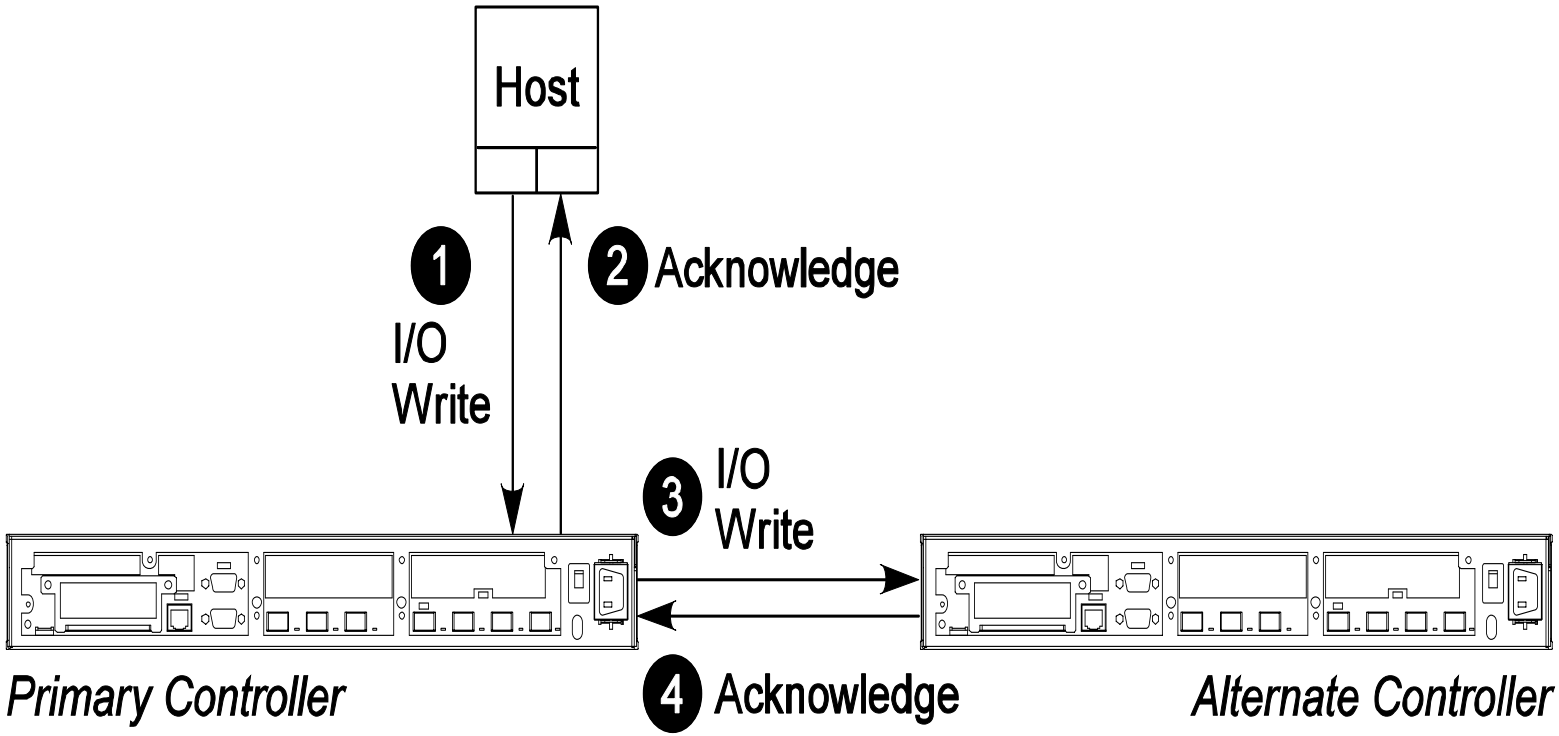
Types of replication

- Synchronous
  - Complete replication on both arrays before acknowledging as complete back to host
- Asynchronous
  - Complete replication on destination array after acknowledging as complete back to host

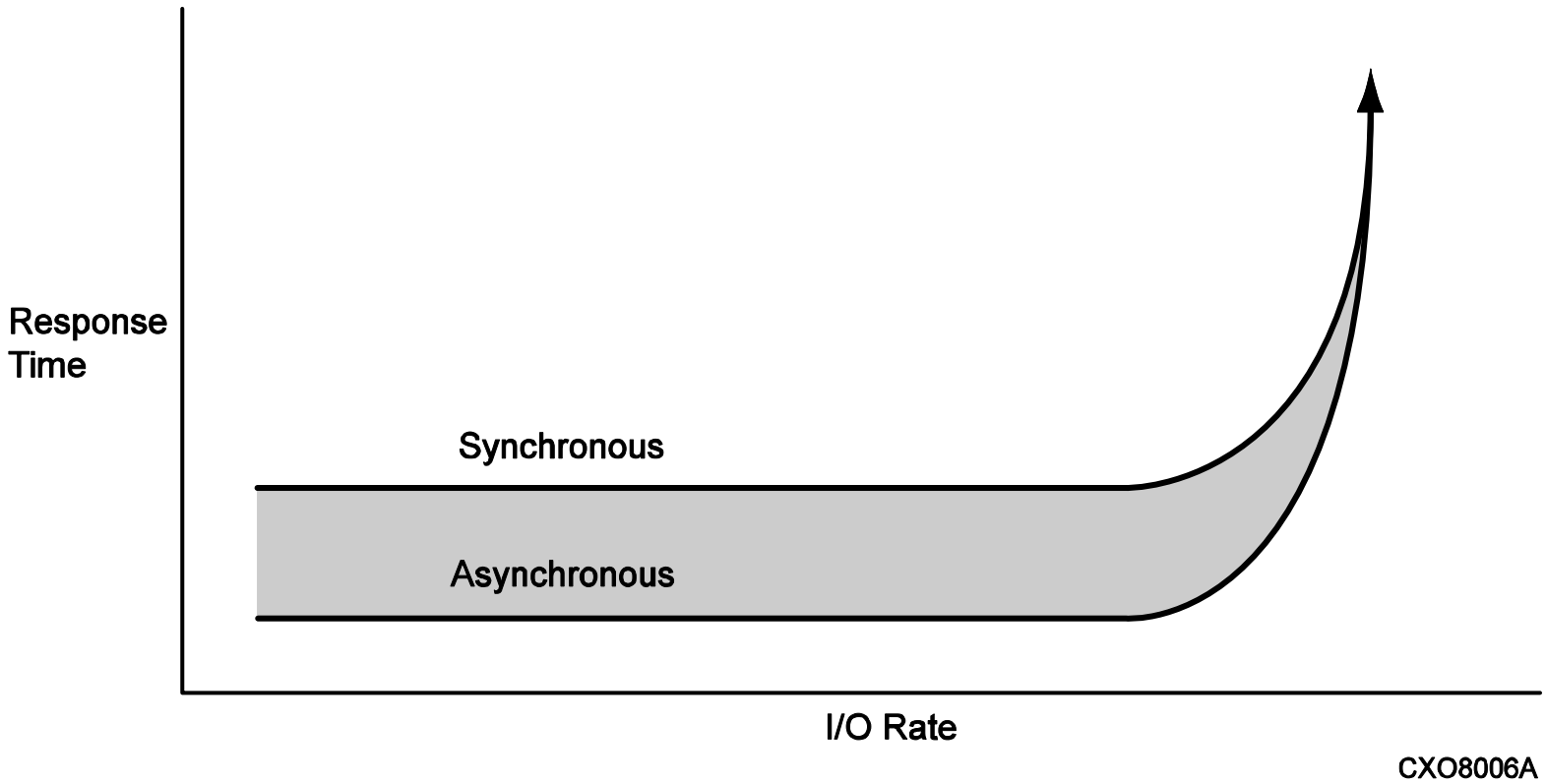
Synchronous replication



Asynchronous replication



Asynchronous versus synchronous replication saturation



Types of replication

- With synchronous replication remote copy contains completed writes
- With asynchronous replication remote copy is a bit behind and may not contain all completed writes from an application perspective
  - Number of outstanding writes is dependant on delay

Agenda

- Supported configurations
- Understanding the requirements
- **Physics of distance and Performance Estimation**
- Best practice design for the storage

Physics of Distance and Performance Estimation

- Physics of distance
  - understanding why it takes so long
- Performance Estimator
  - creating an educated guess
  - based on
    - size of writes
    - distance between sites
    - link bandwidth
  - first a single write stream
  - then impact of multiple streams

Physics of Distance

Or why does it take so long?

- Speed of light is $3 * 10^8$ m per second in vacuum
  - in wire, 1 nano-second is 30 cm (12 inches)
- Speed of light is $2 * 10^8$ m per second in most fiber
  - in fiber, 1 nano-second is 20 cm (8 inches)
  - or 5 microseconds (μSec) per kilometer

Physics of Distance

Because, over distance it does!

- SCSI read/write and HSG80 replication is two round trips
  - each 100 km adds [5 μSec / Km *100 km * 4 trips] = 2 mSec
  - reduces performance similar to slower drive
    - 15k RPM -> 2 mSec average seek time
    - 7200 RPM -> 4 mSec average seek time
- Continuous Access EVA replication is one round trip
  - each 100 km adds 1 mSec
  - performance at 200 km is similar to using slower drives

Performance Estimator

Creating an educated guess

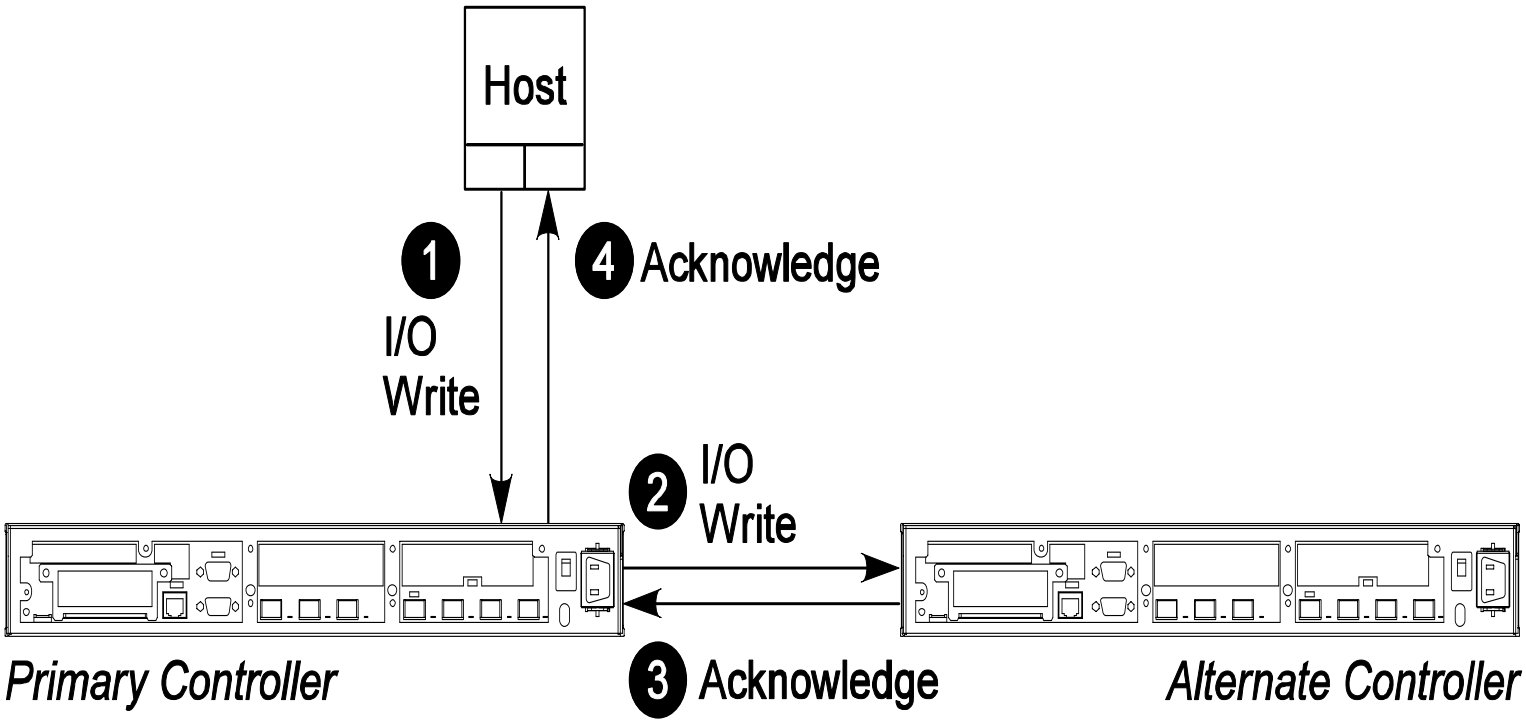
- Based on time to complete a single synchronous replication write for a given bandwidth of link
- Add impact of distance
- Calculate how many single writes per second
- Multiply by number of parallel streams
- Reduce by expected utilization

Performance Estimator

Time to complete a single synchronous replication at zero separation distance – 4 step process

1. Host writes to first array's cache
2. Replicates the write, and sends it to second array
3. When in cache, second array acknowledges back to first array
4. First array acknowledges back to host issuing write.

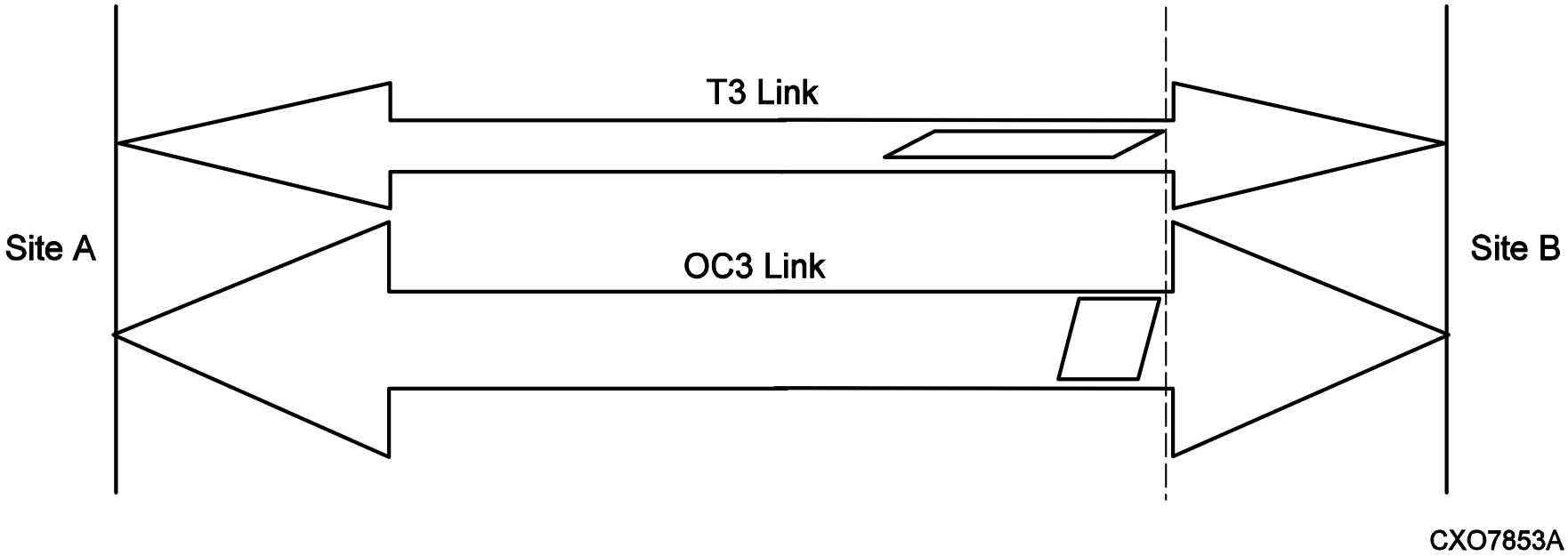
Performance Estimator



CXO8222A

Performance Estimator

- **Effects of distance does not depend on intersite link bandwidth**



- **Delay is 5 nanosecond per kilometer per trip**

Performance Estimator

Calculate number of single writes per second

- Invert response time in mSec per Write
- Yields peak writes per second
- For a single I/O stream
  - a single, simple application

Using the Performance Estimator

hp StorageWorks Continuous Access EVA Replication Performance Estimator - V1, VCS V3
For a Single I/O Stream using direct connect fiber, WDM, or Fiber Channel over IP
Estimates based on One Outstanding Synchronous Replication of Local Writes

Enter One Way Intersite Latency: 2.0 ms — 400 km or 249 miles

Enter Size of a Write Data Packet (KiloBytes): 8 KB — 256 KB max.

	2 Gbps fiber	1 Gbps fiber	FCIP	FCIP
Link Bandwidth in megabits / sec	2000 Mbps	1000 Mbps	100.0 Mbps	10.0 Mbps
Packet Load/ Unload Time:	0.21 ms	0.27 ms	1.01 ms	7.18 ms
ms per I/O:	4.56 ms	4.66 ms	5.38 ms	13.85 ms
I/Os per Second:	219.5 I/O/sec	214.4 I/O/sec	185.9 I/O/sec	72.2 I/O/sec
or	6.32 GB/h	6.17 GB/h	5.35 GB/h	2.08 GB/h
Approximately:	17.56 Mbps	17.15 Mbps	14.87 Mbps	5.78 Mbps
% Bandwidth	0.88%	1.72%	14.87%	57.77%

Performance Estimator

- Estimator only tells you how many writes per second are available using a single I/O stream – a first guess

- Now look at how to increase that for multi-stream I/O or multi-threaded applications

Performance Estimator

- Start with knowing controller limits
  - each replication path has 32, 8KB buffers
  - writes up to 8K use one buffer
  - writes over 8K use more than one buffer
  - port 1 to port 1 is one path, port 2 to port 2 another
- Also understand that link utilization will limit bandwidth
  - don't over subscribe
  - limits are: average 40%; peak 45%
- Finally the controller can only do so much over time.
  - current understanding is x6 at zero distance
  - increases as distance allows for more outstanding writes

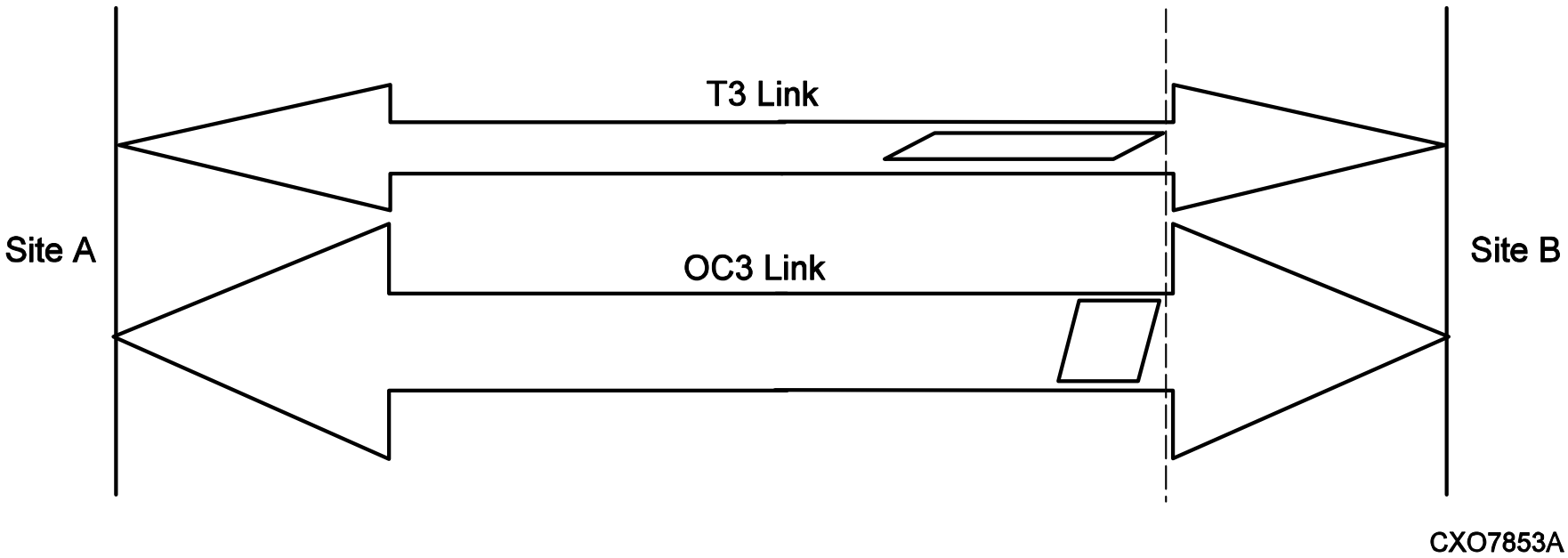
Performance Estimator

Reduce by expected utilization

- Use 100 % only if dedicated environment and able to issue writes as soon as previous completes
- Use 70% as theoretical peak, 50% as practical peak
- Remember Ethernet
  - a full one only uses 50% of capacity on average
  - and 70% utilization is not seen

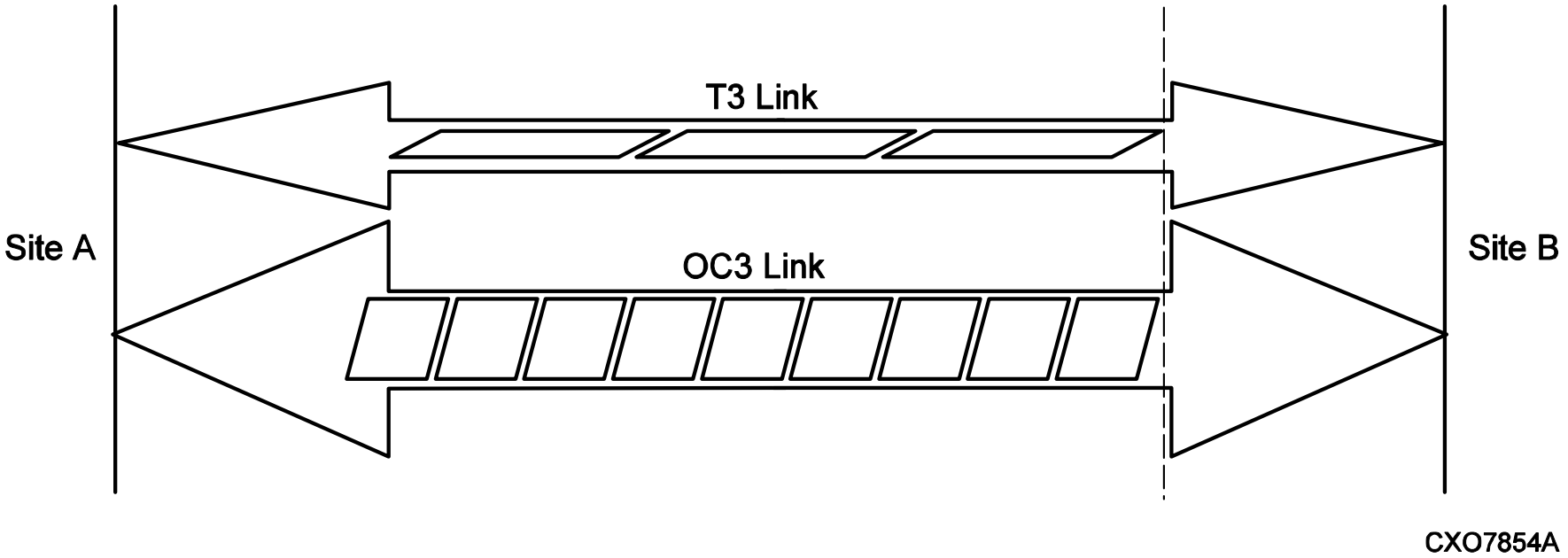
Performance Estimator

First estimate is based on number of single synchronous writes per second



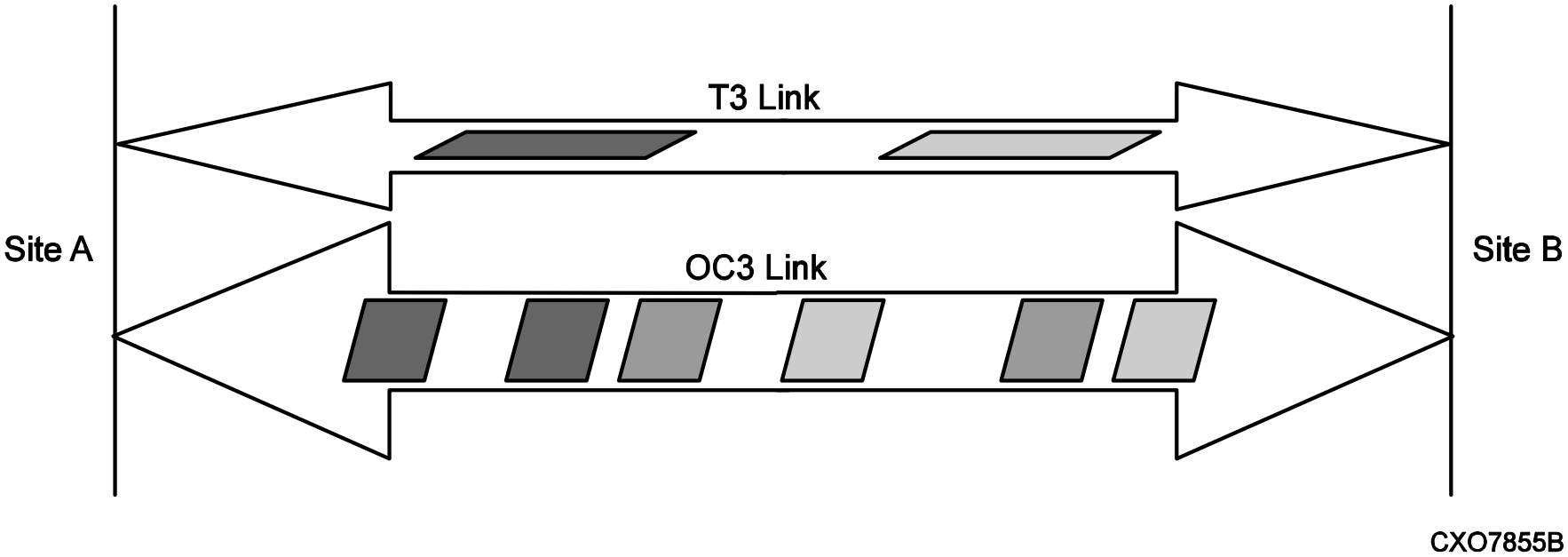
Performance Estimator

Then estimate effect of streaming multiple I/Os for a single application at 100 % expected utilization



Performance Estimator

Finally consider effect of streaming multiple I/Os for a single application at 70 % expected utilization



Performance Estimator

In review, its about creating a good guess!

- Know the size of the write and the distance (delay)
- Figure out how fast that one could be repeated
- Understand impact of multiple I/O in the pipe
  - 70 % peak
  - 50% expected
  - 40% average, 45% peak for planning purposes
- Understand impact of limits
  - link bandwidth
  - controller bandwidth

Agenda

- Supported configurations
- Understanding the requirements
- Physics of distance
- **Best practice design for the storage**

Best practice design

- Vraid types
- number of shelves
- number of drives
- number of disk groups

Best practice design

- Support all Vraid types, but not all are recommended
  - Best practice for Vraid 5 says 8 or more drive enclosures
  - Vraid 0 recommended only for scratch use
  - Use Vraid 1 anywhere

Best practice design

- The more drive enclosures the better
  - Vraid 5 will survive loss of enclosure if 8 or more
  - Vraid 0 will not survive loss of enclosure
  - Vraid 1 will survive loss of enclosure if 2 or more

- If using a 2C2D or 2C6D
  - Use only Vraid 1

Best practice design

- Number of drives should be 2 X drive enclosures
  - And 8X if using Vraid 5
- The number of disks in the disk group determines the maximum I/O rates
  - All Vdisks striped across all members of the disk group
  - Primarily for reads at 150 reads per second
  - Writes are restricted by replication overhead and not per drive response time
- Examples
  - 10 disks at 150 reads/second/disk yields a peak read rate of 1500 reads per second.
  - 50 disks yields a peak of 7500 reads per second
  - Assuming a random read pattern across the Vdisk

Best practice design

- Minimize number of disk groups
  - Because it improves performance (more disks per group)
  - Consistent with number of failure domains
  - For example:
    - Put data in one large disk group for capacity and read performance
    - Put transaction logs in another built for write performance
- If bi-directional Continuous Access EVA, then make disk groups symmetrical across both arrays

For more information

- Documents mentioned in this presentation are available from the Continuous Access EVA web site

  http://h18006.www1.hp.com/products/storage/software/conaccesseva/index.html

  - And then click on "technical documentation". As a starting point, see the CA EVA Design Guide.

Interex, Encompass and HP bring you a powerful new HP World.