Under the Hood of Ficon

Lou Ricci
IBM®
lricci@us.ibm.com

3 March 2011
Session 9030

Legal Stuff

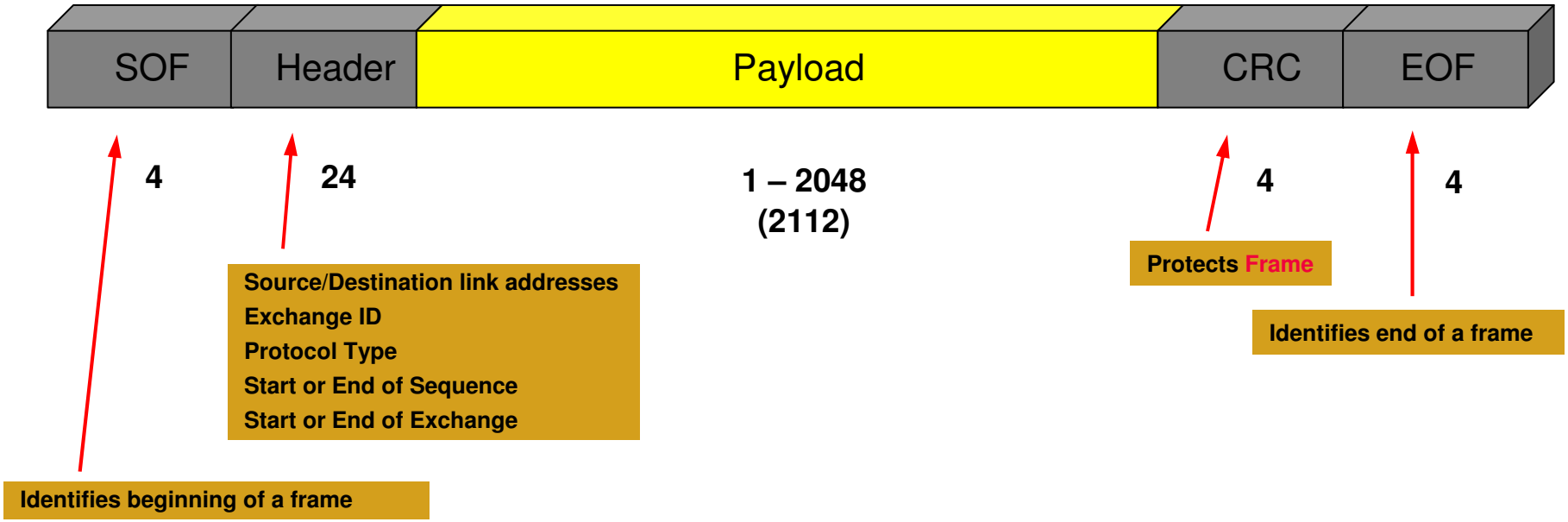
- Notice
  - IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing to: *IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.*
  - Any references in this information to non-IBM Web sites are provied for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.
- Trademarks
  - The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both: FICON® IBM® Redbooks™ System z10™ z/OS® zSeries® z10™
  - Other Company, product, or service names may be trademarks or service marks of others.

Agenda

- Frames, Sequences and Exchanges
- IDAW vs MIDAW
- Buffer to Buffer Credit
- Ficon Recovery

Fibre Channel Frame

The basic building block is the FRAME

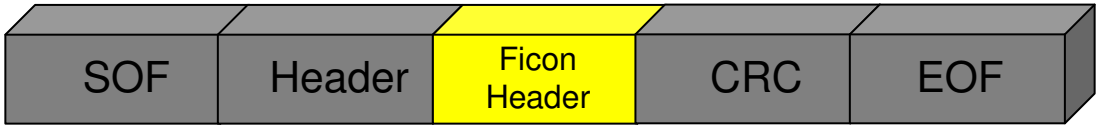


Sequences and IUs

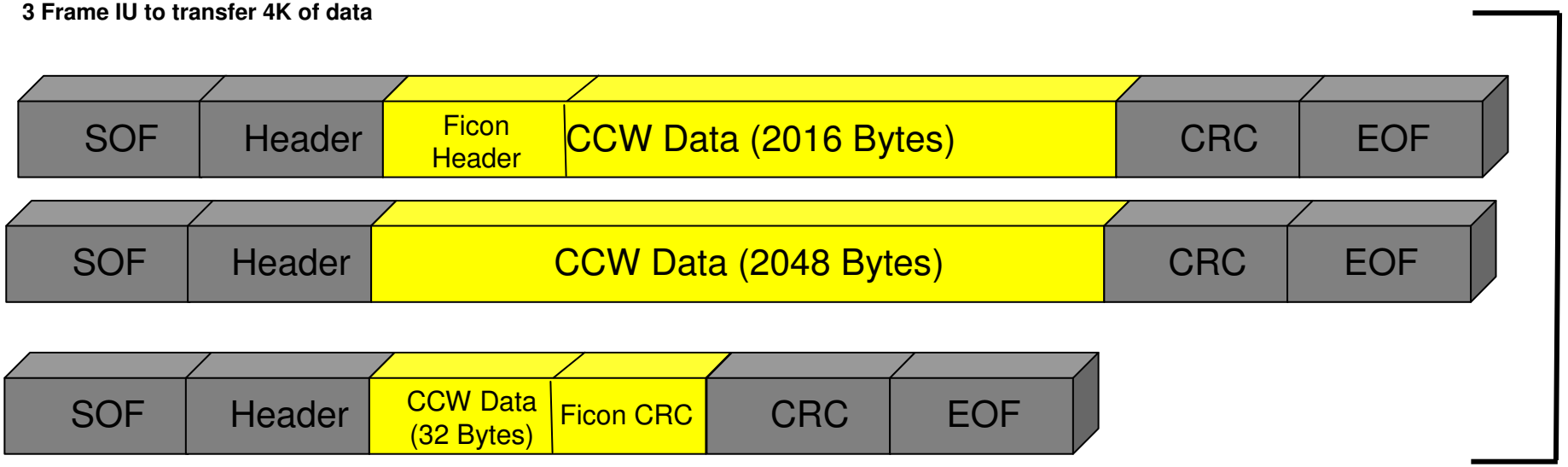
- Each Upper Layer Protocol (ULP) defines the contents and format of it's own Information Units (IUs)
  - Commands
  - Data
  - Status
  - Control
  - Etc
- Ficon IUs can be up to 8K (8192) in size
  - 8160 (8K-32) bytes of data
    - 32 bytes contain Ficon Header information
  - 4 frames are needed for the largest IU
- The collection of frame(s) that make up a IU are called a Sequence
  - A Sequence may be as small as a single Frame

Ficon IU Examples

1 Frame IU to transfer a Read CCW



3 Frame IU to transfer 4K of data



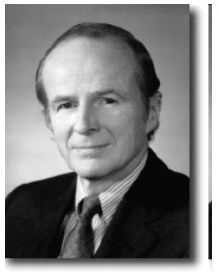
Exchanges

- Fibre Channel Architecture defines an Exchange as
  - “A mechanism for identifying and managing an operation between two ports“
- All IUs (a.k.a. Sequences) that make up a single I/O operation are part of an Exchange

Ficon Exchanges

- In Ficon, each concurrent I/O operation uses two Exchanges
  - One unidirectional Exchange for IUs from the Channel to the CU
  - A different unidirectional Exchange for IUs from the CU to the Channel
- The PAIR is commonly know as a “Ficon Exchange”

How many Exchanges do I need?



SHARE
Technology · Connections · Results

- Little's Law states:
  - *The number of "things" in a system can be determined by multiplying the average arrival rate of those "things" by the average time each "thing" stays in the system*

  - Applied to Ficon:
    - The average number of Exchanges active at any given time = Average I/O rate * Average response time

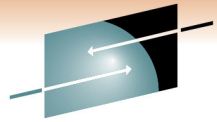
    - Example: 5000 Ficon I/Os / Second on a given channel with .4ms service time[1] needs 2 Active Exchanges (pairs) at any given time

[1] The amount of time the I/O is active in the channel

Agenda

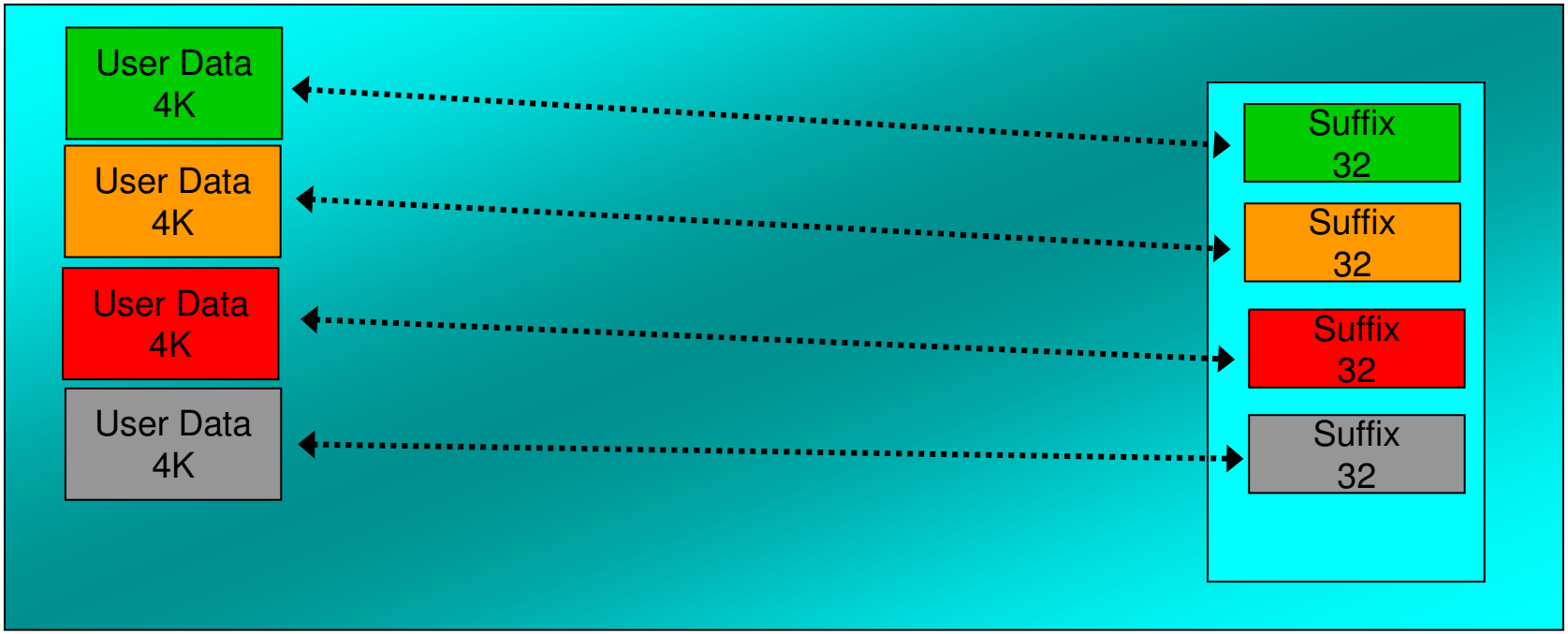
- Frames, Sequences and Exchanges
- IDAW vs MIDAW
- Buffer to Buffer Credit
- Ficon Recovery

What Problem Does MIDAW Solve?



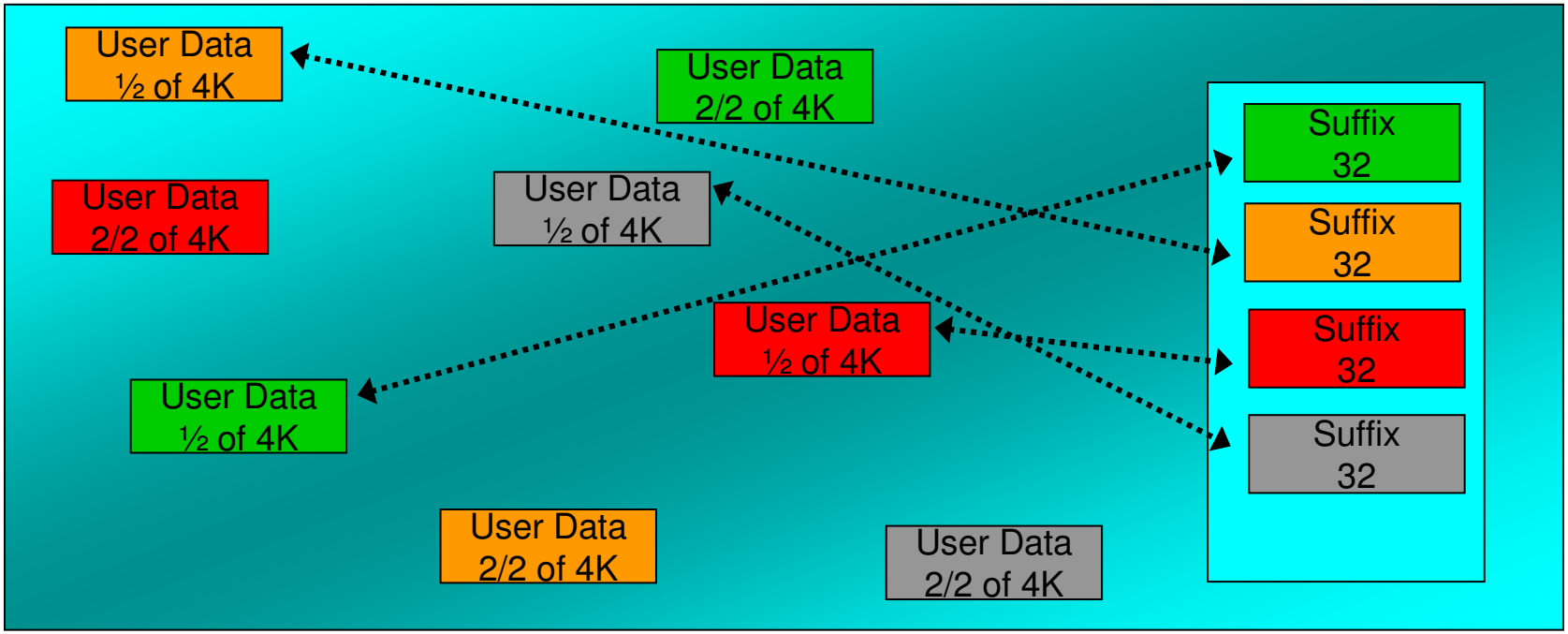
- Extended Format data sets have a small suffix appended to the data.
- Data and Suffix are in DISCONTIGUOUS virtual storage.
- Data and Suffix are in the same physical record on the DASD volume.
- This combination results in less than optimal channel, control unit, and link efficiencies.

EF Data in VIRTUAL Storage



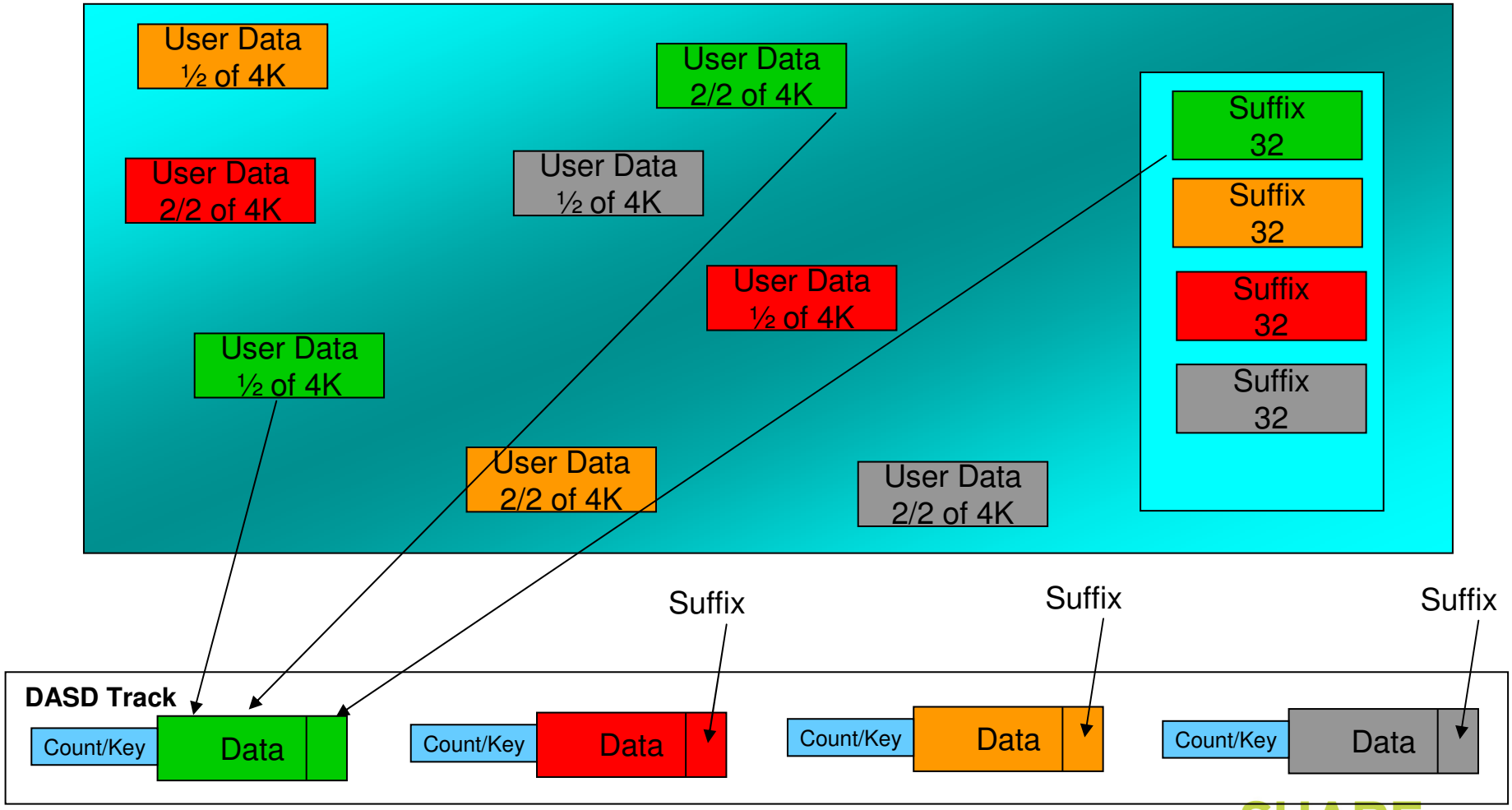
In virtual storage the User Data is contiguous

EF Data in REAL Storage

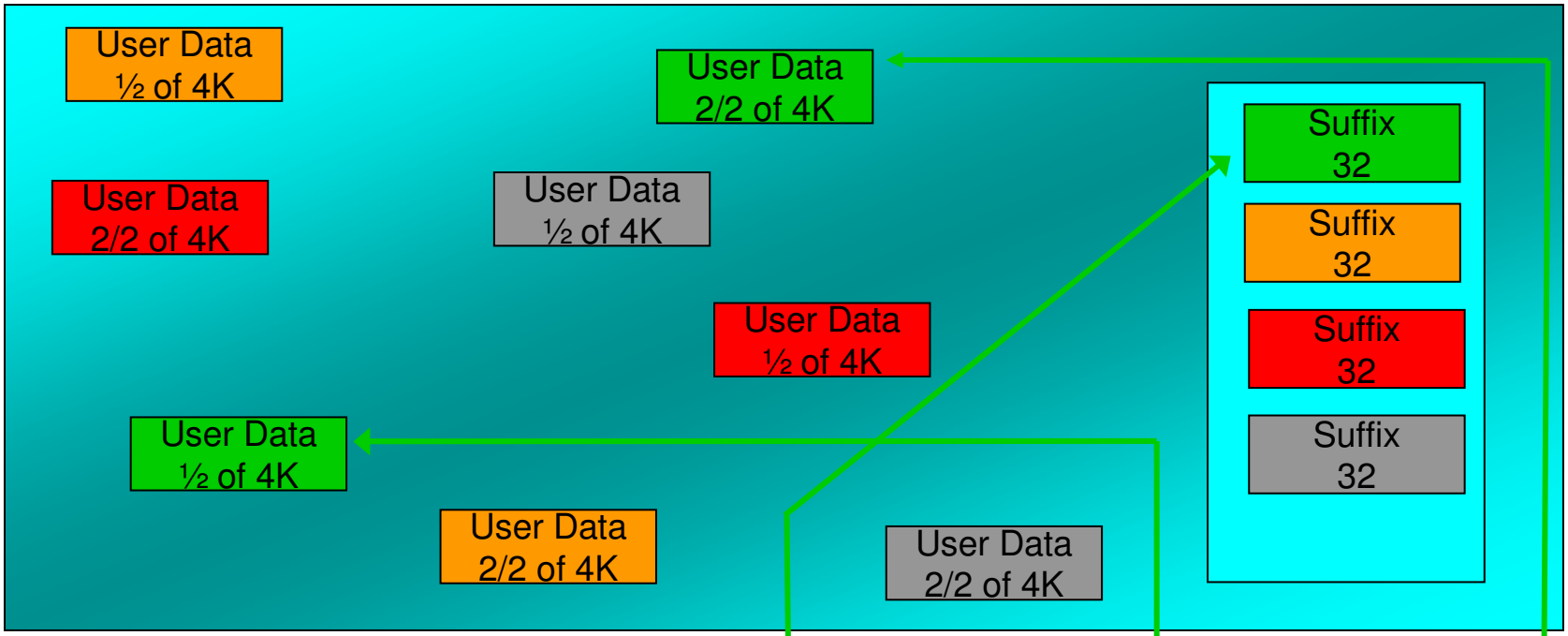


In REAL storage the User Data is scattered about

EF Data from REAL Storage to DASD volume



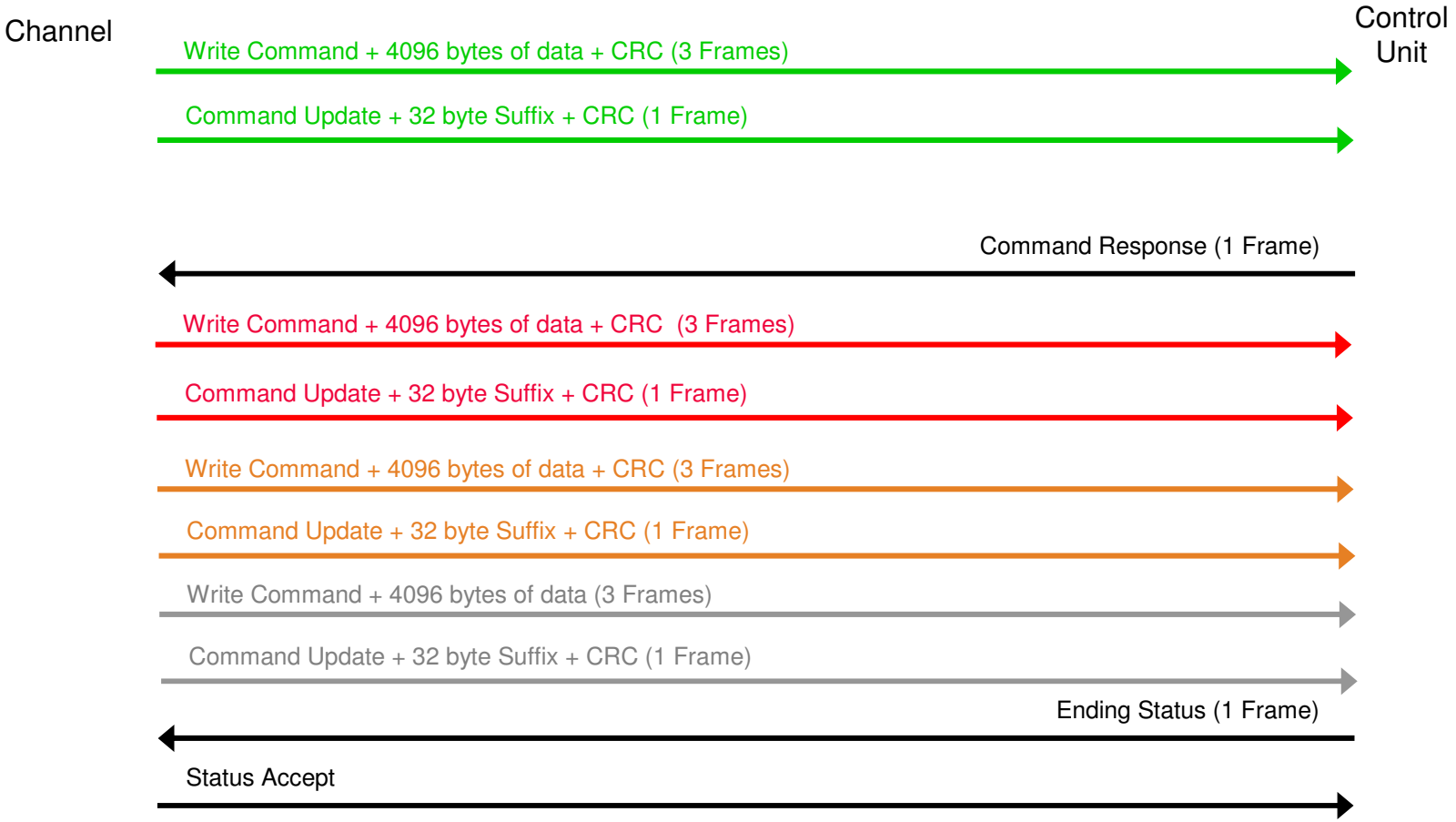
How It's Done With IDAW



Write	CD,IDA	1000	00465E00
XX	CC	20	00175400
Write	CD,IDA	1000	00465E08
XX	CC	20	00175420
Write	CD,IDA	1000	00465E10
XX	CC	20	00175440
Write	CD,IDA	1000	00465E18
XX		20	00175460

03EC4000
03C5B800
02FE9000
0C78D000
0E015800
097F8000
0D06E800
0BF43000

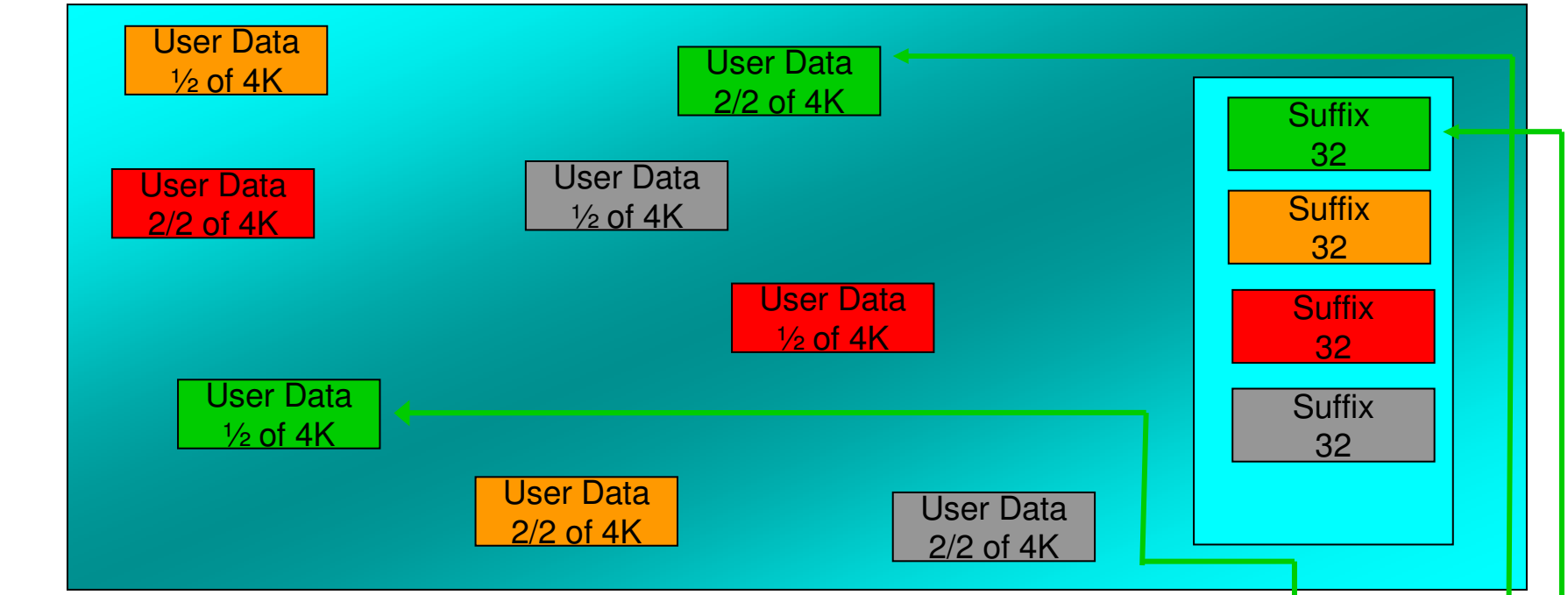
Sequence View Of the FC Link with IDAW



What is MIDAW?

- Modified InDirect Address Word
- A new method of gathering and scattering data into & from non-contiguous system z storage locations during an I/O operation.
- Designed to improve performance of certain applications
  - DB2 sequential workloads that use Media Manager to process small records with Extended Format data sets
- Reduces the number of CCWs in a channel program for these workloads

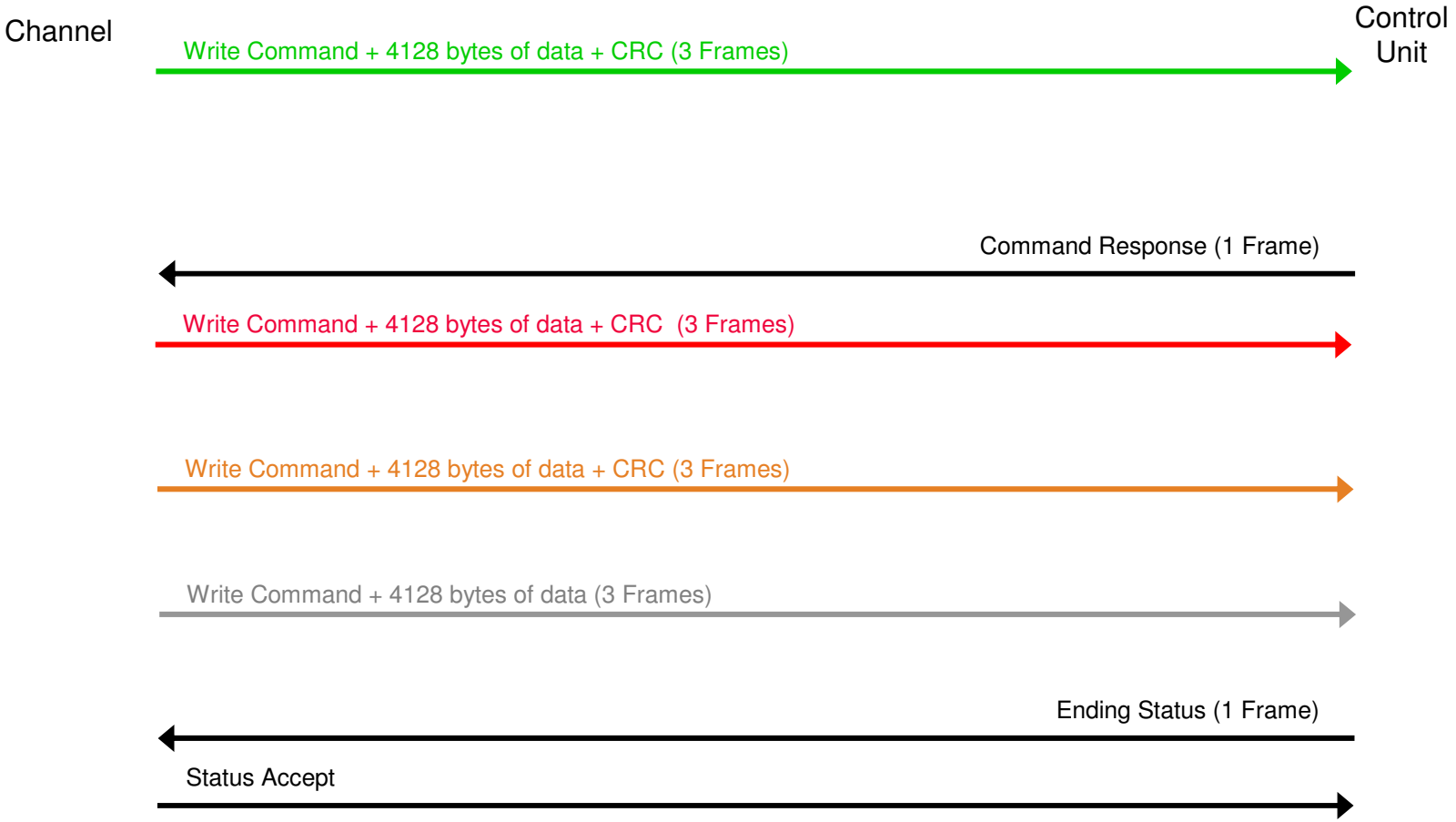
How It's Done With MIDAW



Write	CC,MIDA	1020	465E00
Write	CC,MIDA	1020	465E18
Write	CC,MIDA	1020	465E30
Write	MIDA	1020	465E448

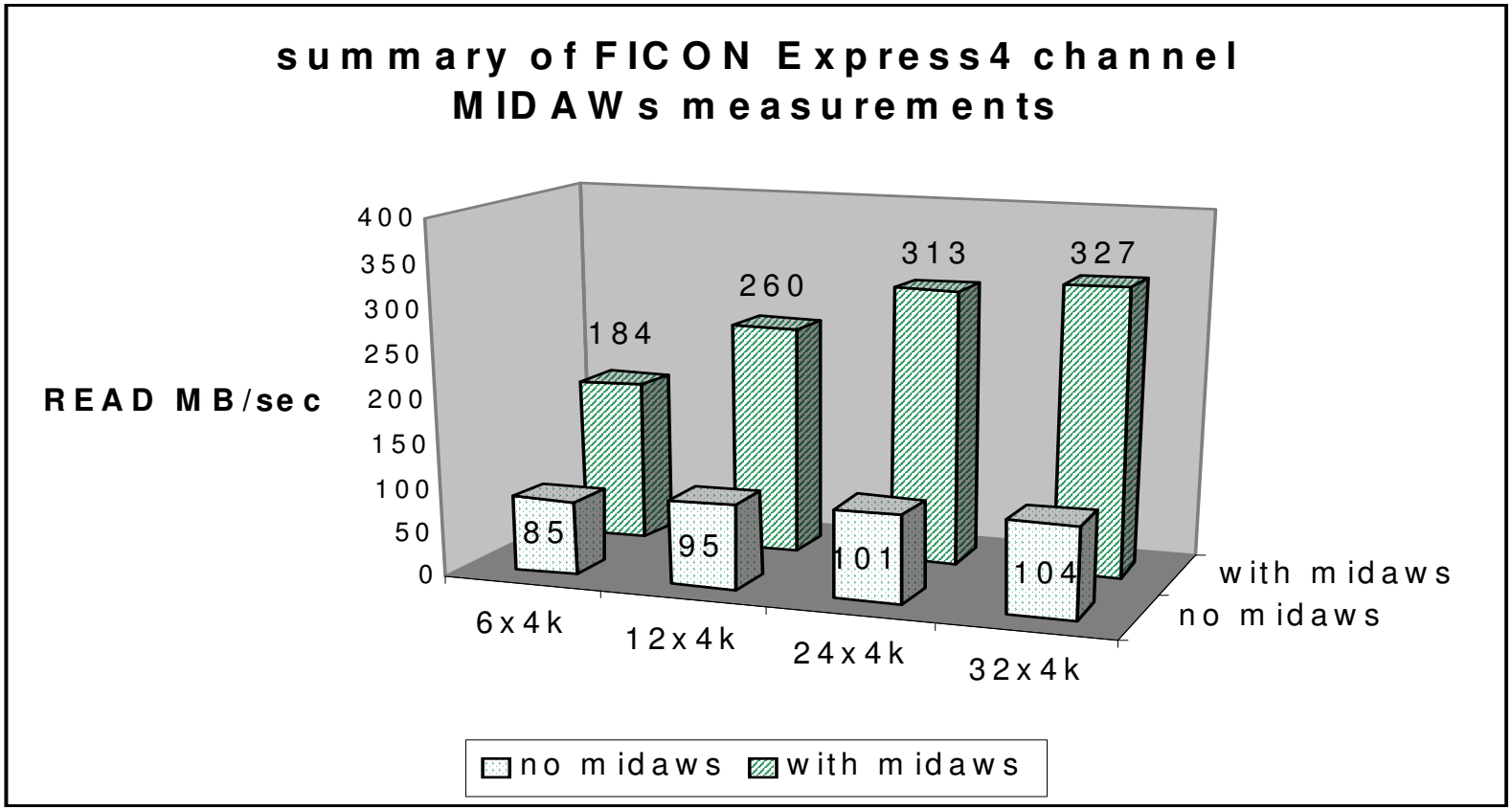
03EC4000	800
03C5B800	800
00175400	20
02FE9000	800
0C78D000	800
00175420	20

Sequence View Of the FC Link with MIDAW



IDAW vs MIDAW comparison

	Channel to CU		CU to Channel	
	IDAW	MIDAW	IDAW	MIDAW
Frames	17	13	2	2
Sequences	9	5	2	2



* This performance data was measured in a controlled environment running an I/O driver program under z/OS. The actual throughput or performance 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.

See also: http://www.redbooks.ibm.com/redpapers/pdfs/redp4201.pdf

Agenda

- Frames, Sequences and Exchanges
- IDAW vs MIDAW
- **Buffer to Buffer Credit**
- **Ficon Recovery**

What is Buffer-to-Buffer Credit?

- The greater the BB Credit….
  - A. The faster frames can be sent
  - B. The farther apart the two ports can be
  - C. The larger the frames can be
  - D. None of the above

What is Buffer-to-Buffer Credit?

- The greater the BB Credit….
  - A.
  - B. The farther apart the two ports can be
  - C.
  - D.

What is Buffer-to-Buffer Credit?

- BB Credit is the number of FRAME buffers a port provides for it's NEAREST neighbor

- BB Credit does NOT have to be symmetrical

What is Buffer-to-Buffer Credit?

- BB credit value determines the DISTANCE two nodes can be apart and still maintain full link frame rate
- BB credit is on a FRAME basis, not frame SIZE basis
  - A 1 byte frame consumes 1 credit
  - A 2K byte frame consumes 1 credit
- Number of credits needed determined by:
  - Raw Link Speed
  - Speed of light thru a fiber
  - Distance between two adjacent nodes
  - Average frame size

What is Buffer-to-Buffer Credit?

- Each time a frame is sent, the sender decrements its available credit count by 1
- Each time a frame receiver clears a frame buffer it sends a “R_RDY”
  - Special 4 byte character – NOT a frame !
- Reception of a R_RDY causes the available credit count to be incremented by 1

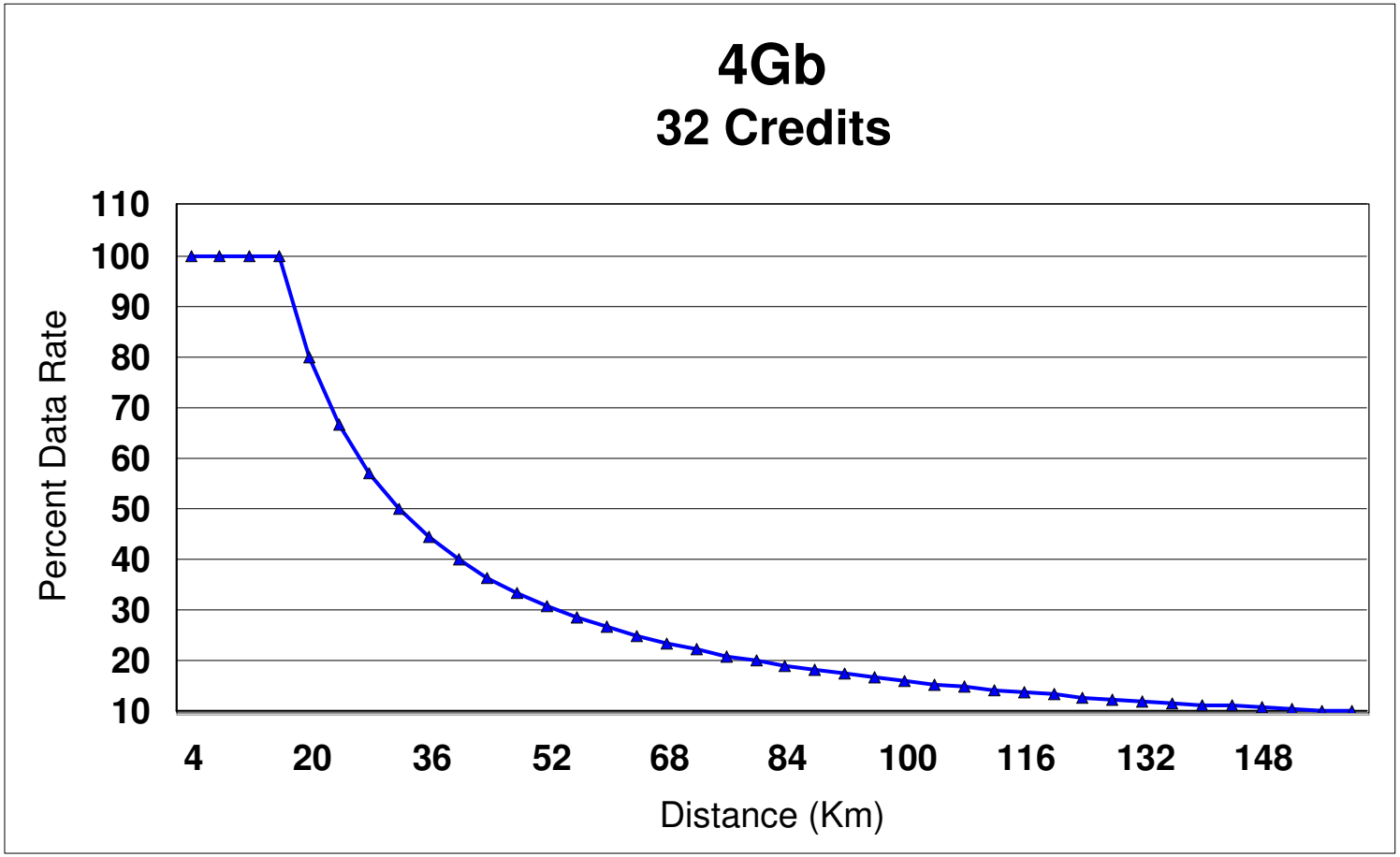
BB-Credit Animations

How much credit do I need?

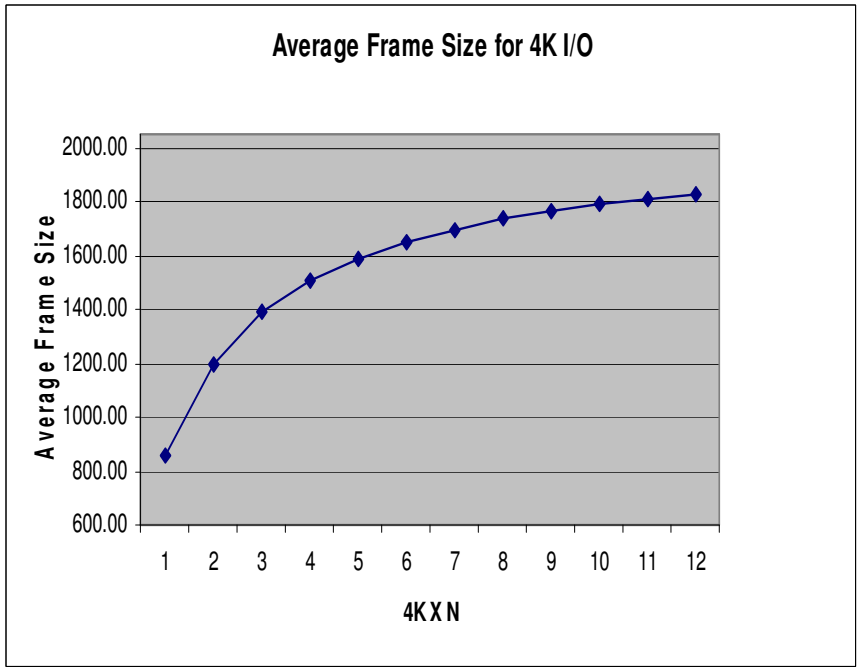
- Good “Rule of thumb”

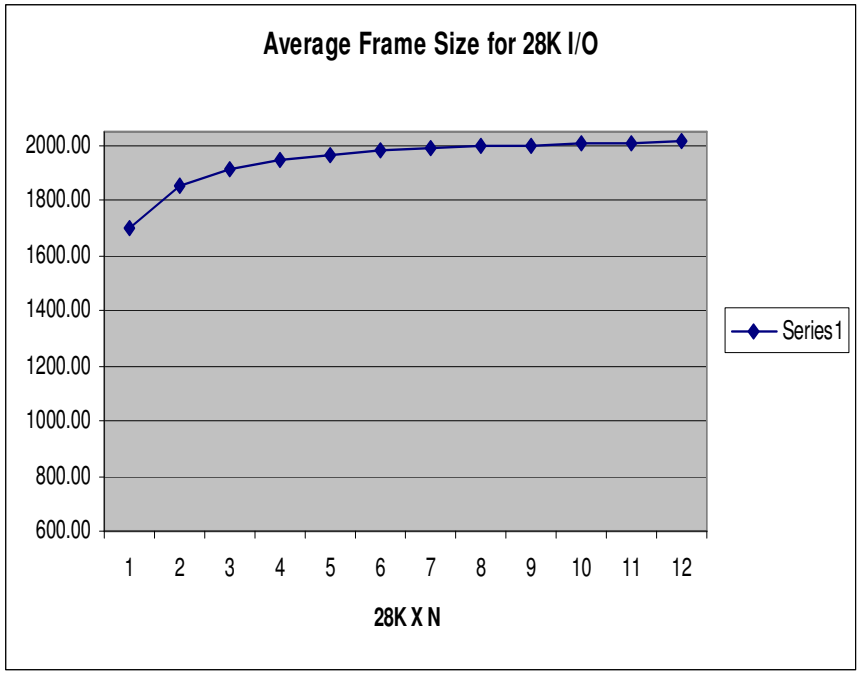
$$\text{Number of credits needed} = 1 + \frac{\text{Link speed in Gb/s} * \text{Distance in Km}}{\text{Average Frame Size in Kb}}$$

How does credit affect Bandwidth?



Average Frame Size vs Block Size





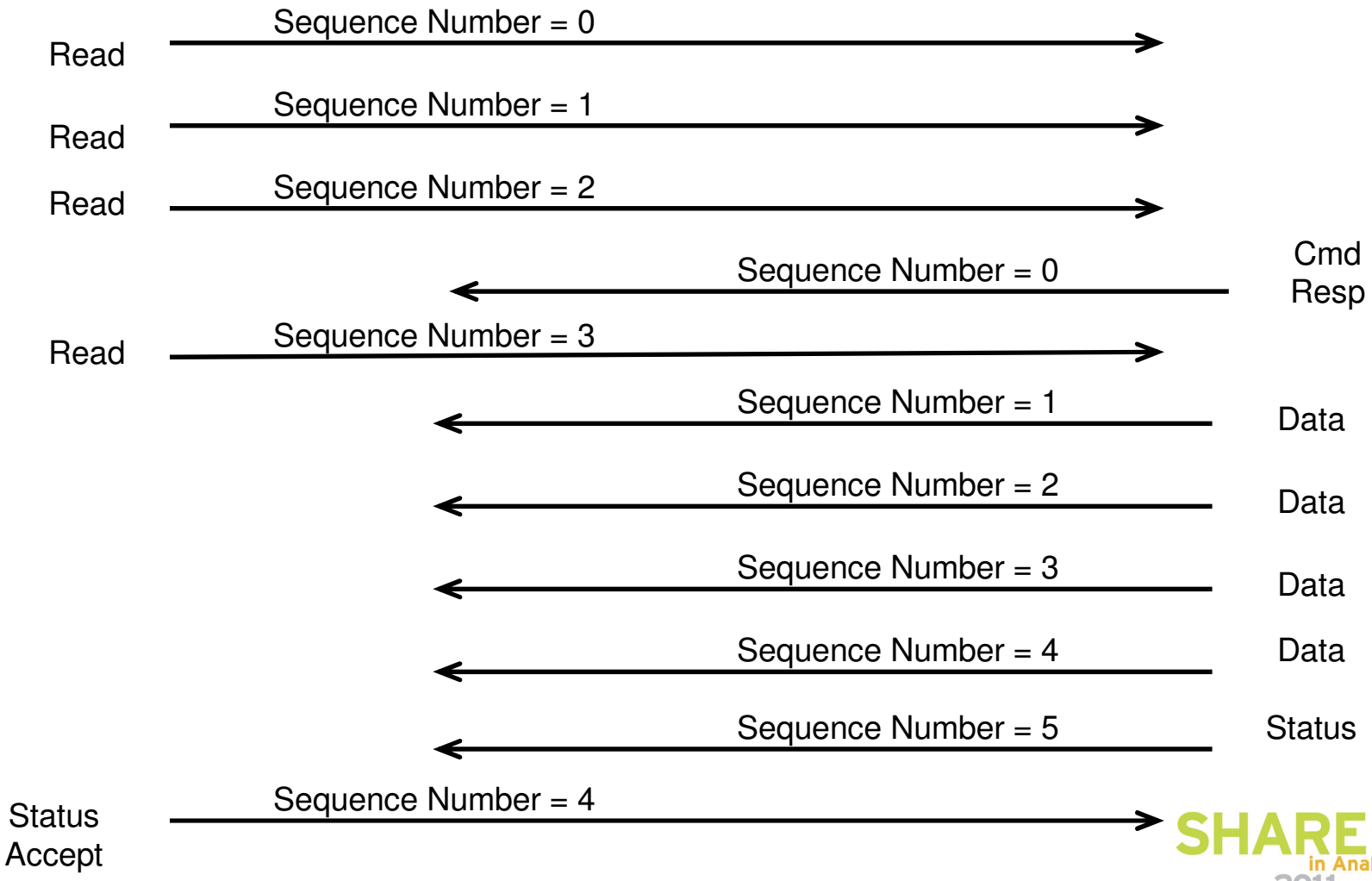
Agenda

- Frames, Sequences and Exchanges
- IDAW vs MIDAW
- Buffer to Buffer Credit
- **Ficon Recovery**

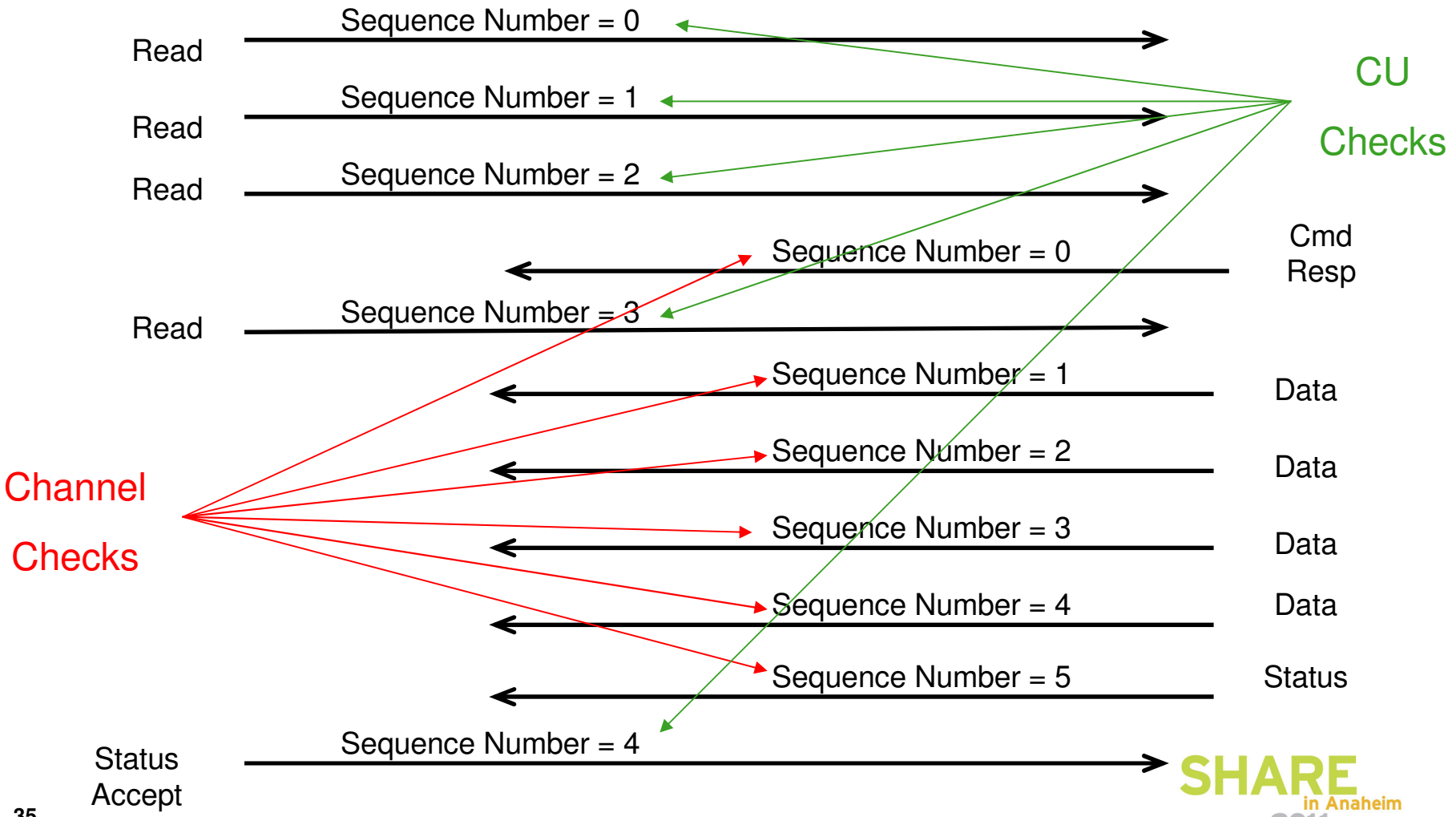
Error Detection and Recovery

- Device Level
  - Error scope is one single I/O operation
- Link Level
  - Error scope is all active I/O on the affected link
    - Channel to Switch Link ➔ All I/O active on the channel
    - Switch to CU Link ➔ All I/O on that one destination link
- Channel Internal Errors

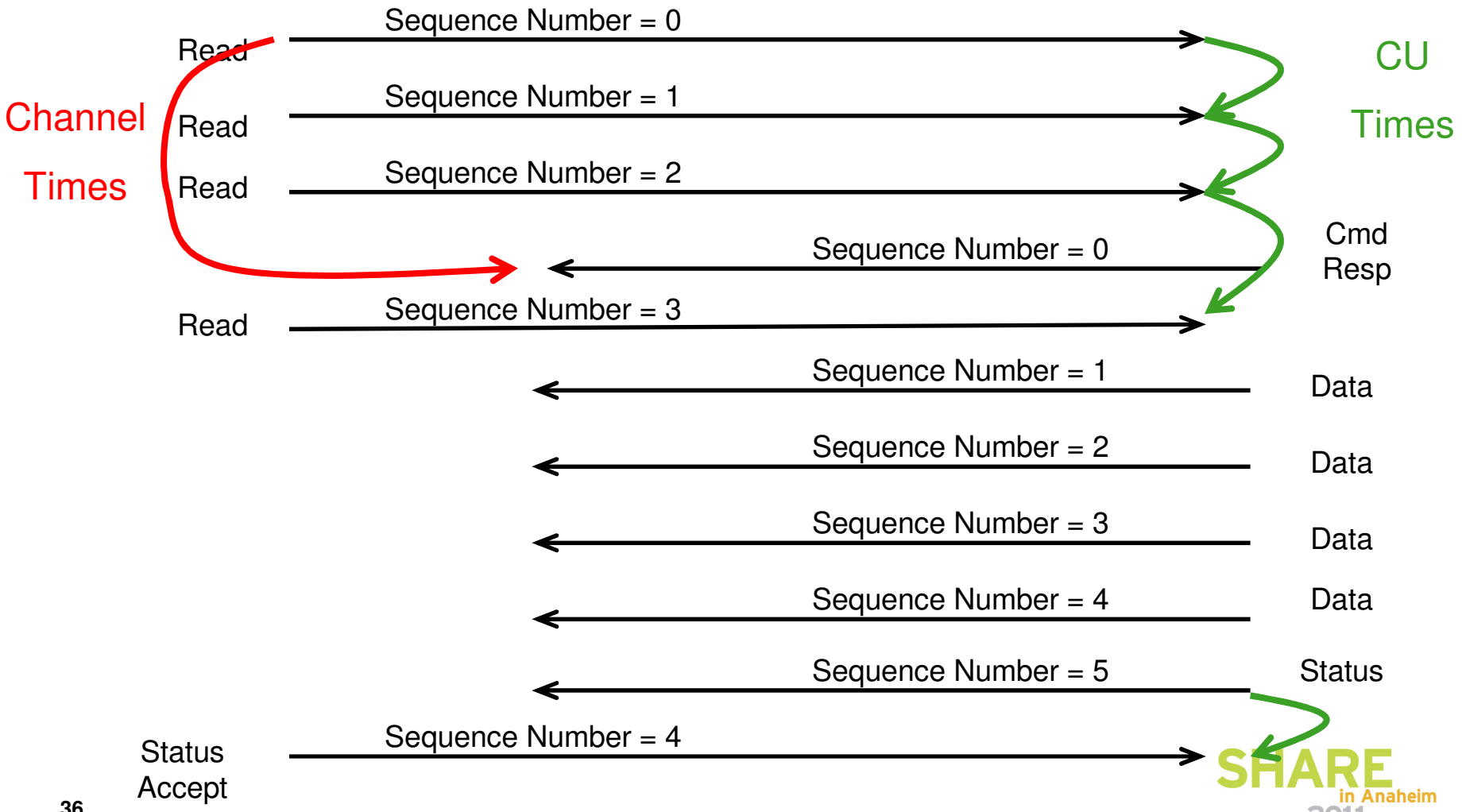
Device Level Errors



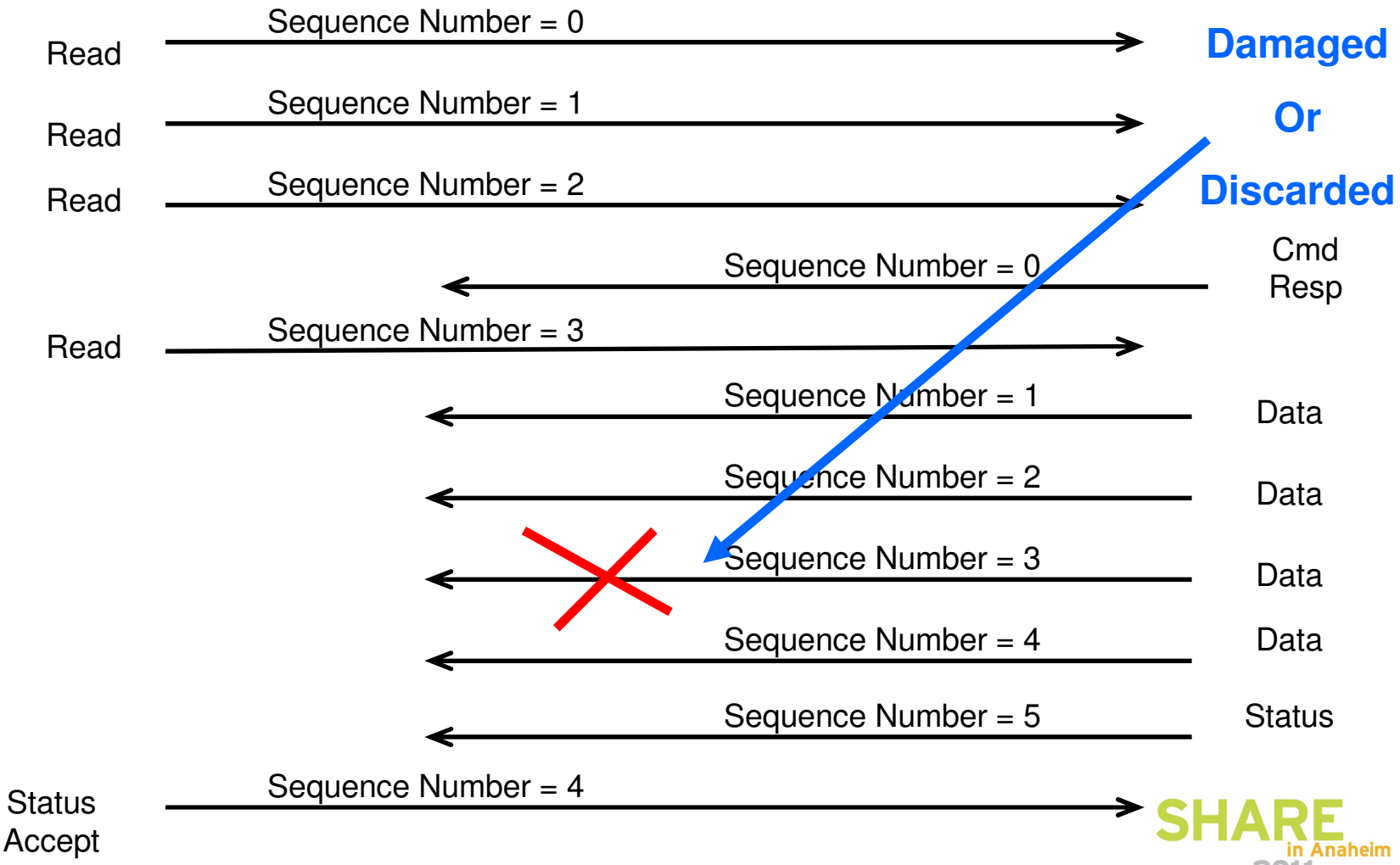
Device Level Errors



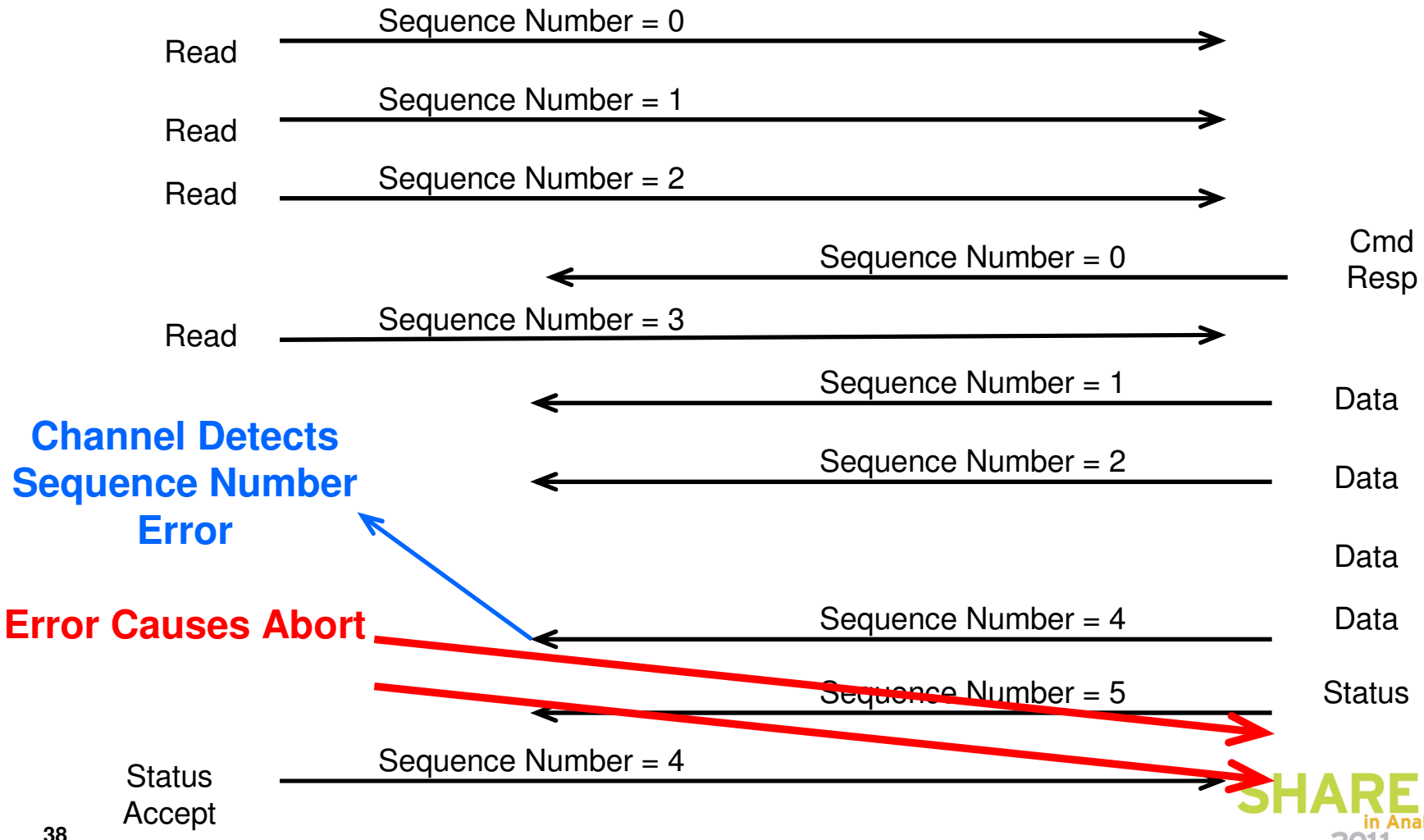
Device Level Errors



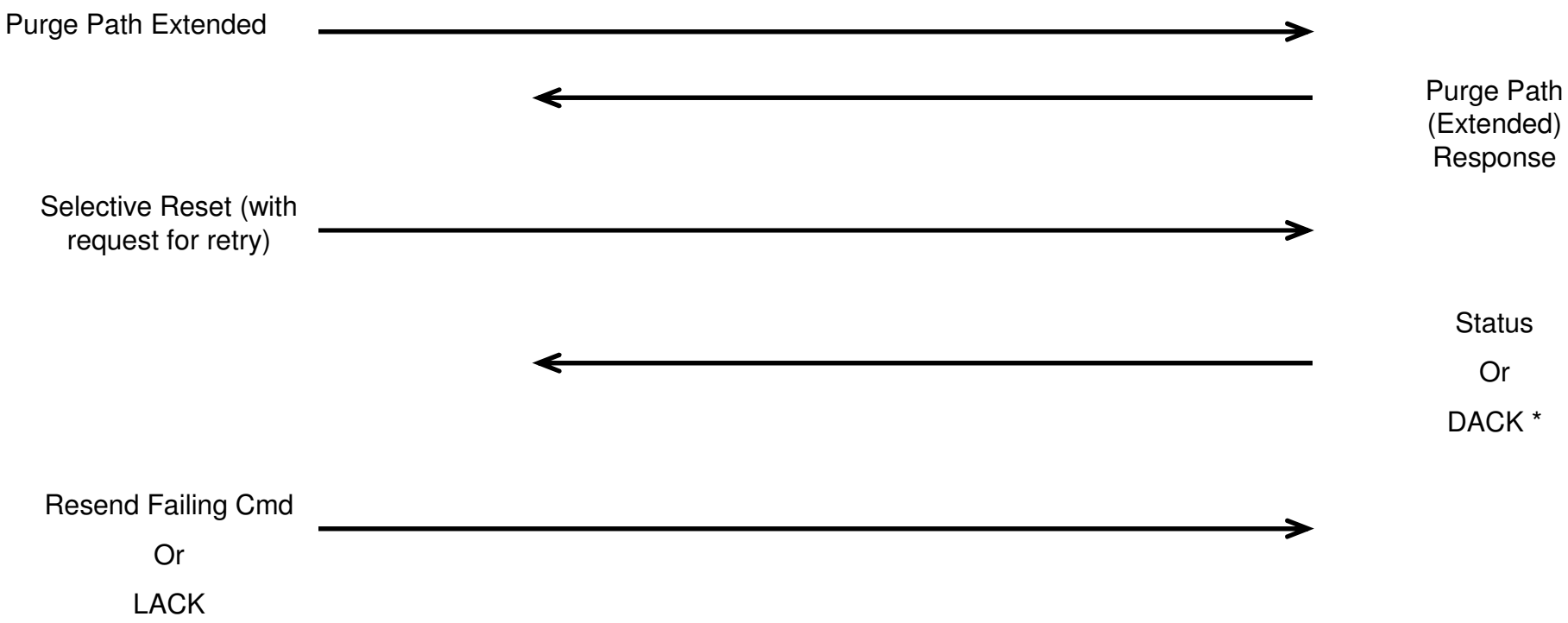
Device Level Errors



Device Level Errors



Device Level Errors



* DACK will result in an IFCC (IOS 050i) interrupt to software

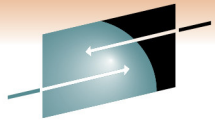
Link Recovery

- Three initiators:
  - Link Between Channel and Switch ‘Fails’
  - State Change from Switch
  - Timeout
- For Remote Links, Channel Attempts a ‘Ping’
  - If channel gets a response – all is ok
  - If ‘no one home’ response, repeat Ping in 4 seconds
  - Still ‘no one home’ response – declare link dead
- 5 Link fails in 5 Minutes ➔ Flapping Link Threshold

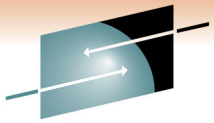
Internal Errors

- Firmware detected
  - “Shouldn’t be able to get here from there”
- Hardware detected
  - Parity Error
  - Cross Check
- SAP Detected
  - Timeouts
  - Out of context messages

Internal Error Recovery



- Channel Hardware Reset
- Firmware Reloaded
- Channel is Re-Initialized
- Any in-progress Operations Terminated



- Any Additional Questions?

Thank You For Your Time And Attention