



Data integrity considerations

Robert Snively

November 5, 2007

T11/-7-691v0

What is a data integrity failure?

In a storage subsystem, a data integrity failure is any occurrence of an event where data transmitted to the storage subsystem was subsequently recovered incorrectly from the storage subsystem with no indication either during the storage or recovery of the data that an error occurred.



What is a data integrity failure, really?

This document investigates only those data integrity failures associated with the Fibre Channel and FCoE transmission paths.

It does NOT consider other possible sources of data integrity failures, including:

- Failure of the operating system and drivers to correctly structure the sequence and content of storage (SCSI or FICON) commands
- Failure of the Initiator to correctly collect, encapsulate, and transmit the storage commands and data to the transmission path.
- Failure of the target (including the storage device) to correctly interpret the storage commands, correctly store and preserve the data, and correctly read the data back.

Principal data integrity exposures are in the ENode to FCF paths because FCF paths are relatively regular and well controlled.



What is a data integrity failure, really?

Data integrity failures include only those failures for which NO NOTIFICATION of the failure is provided.

- If notification is provided, it is not a data integrity failure, but a retryable and recoverable transmission error.
- Denial of Service issues are considered in some obvious cases in this study.

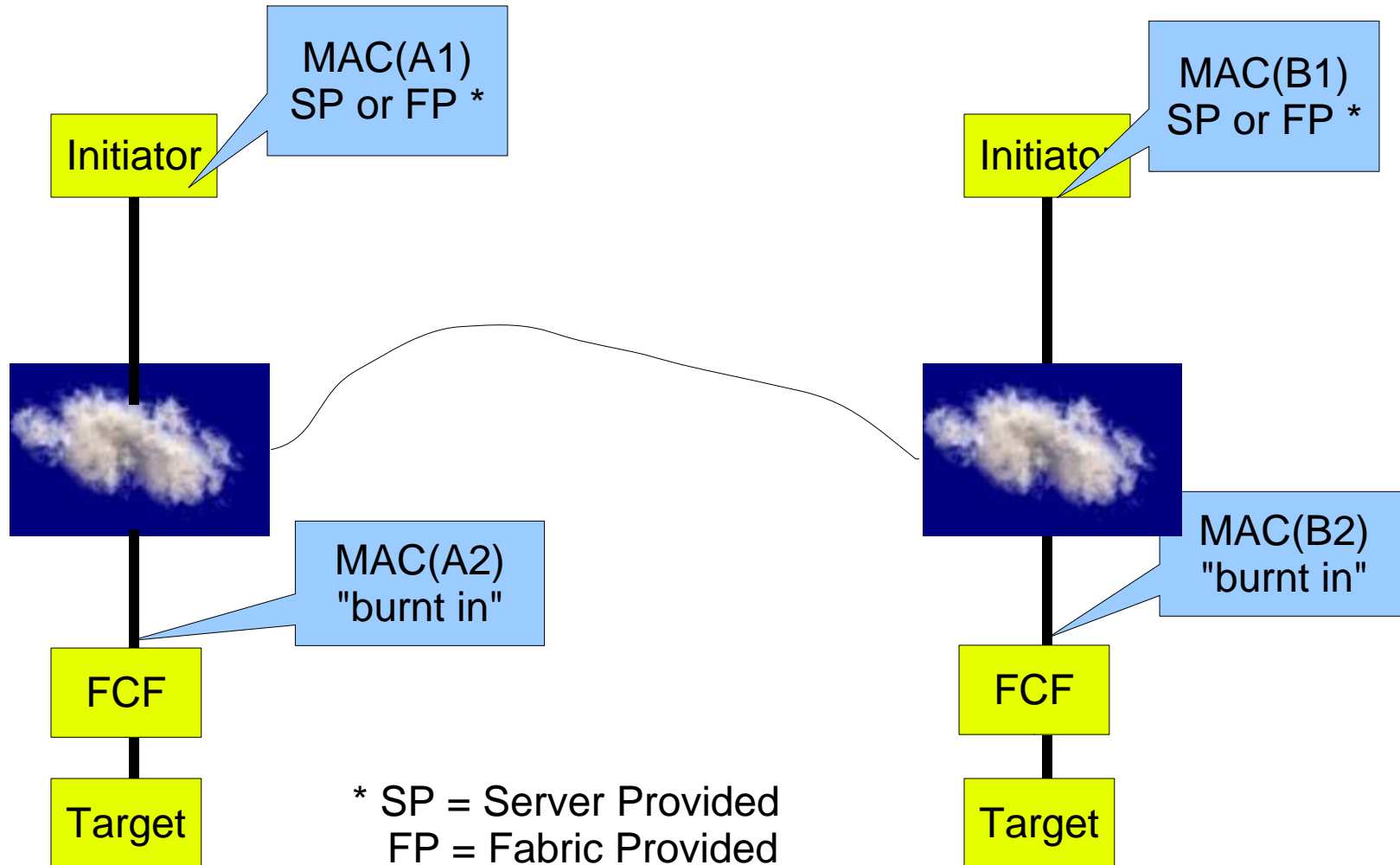
It is assumed that all components of the system are standards compliant. Incorrect data caused by non-compliant devices is a design failure. Incorrect data caused by malicious devices is a security failure.

- If a data verification action required by the standards is not performed, additional sources of error naturally exist.

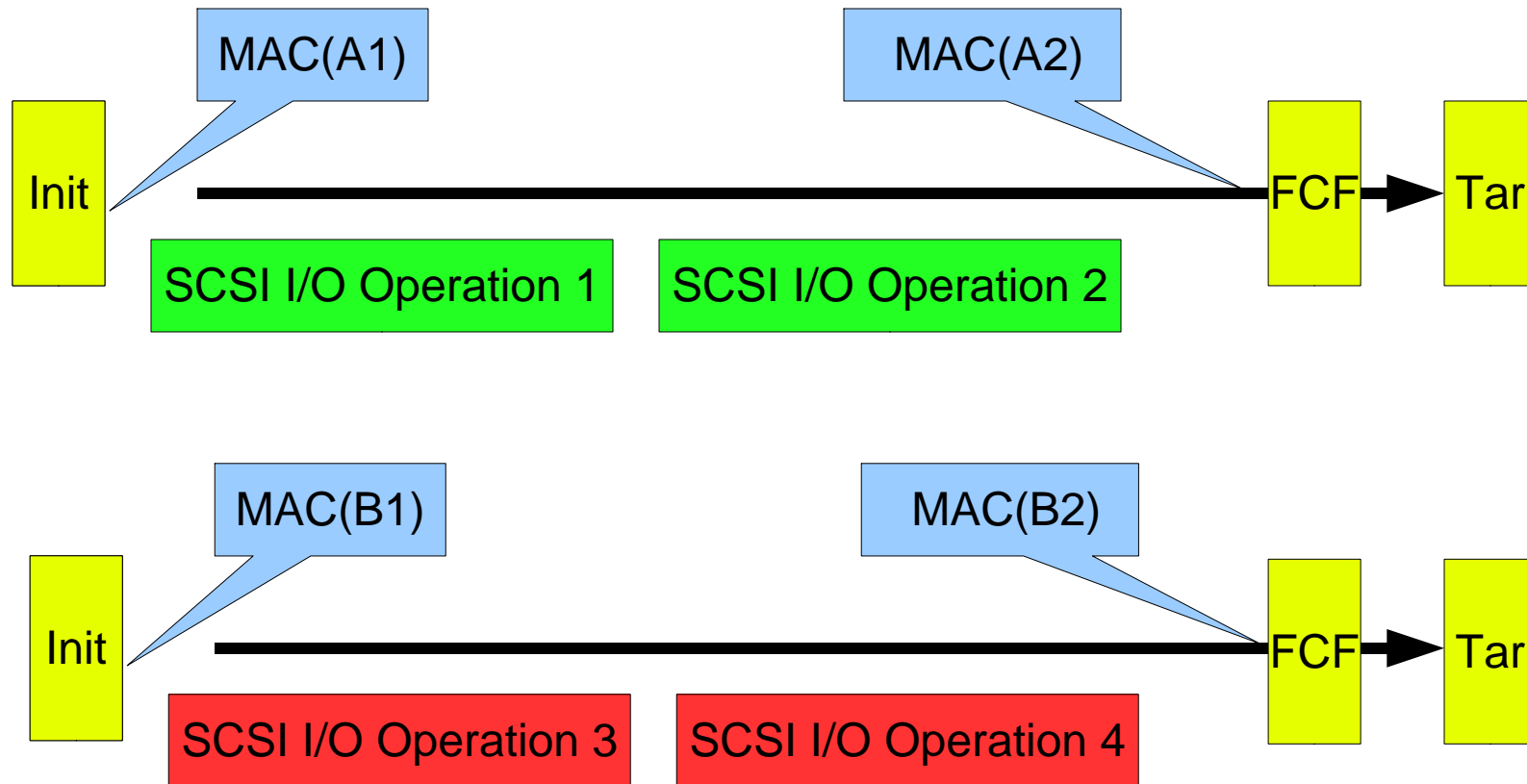
Note that sufficient properly identified errors may limit service (Denial of Service), but will not cause data integrity exposures.



Consider two systems



Consider two streams of FCP data



Note an oddity

Frames from ENodes to FCF ports cannot be misdirected. They will always arrive at the correct "burnt-in"* FCF MAC.

Frames from FCF ports to ENodes may be misdirected. They may arrive at a replicated ENode MAC.

Note that to date, the principal concerns have been Initiator ENode to FCF ports. Target ENodes have traditionally used burnt in addresses, requiring FCF MACs to support both behaviors, but largely securing the ports from the problems described in these pages.

- * Burnt-in MACs are established during manufacture and are world-wide unique.
- Server provided ENode MACs are either burnt-in or provided by a system-wide unique registration process.
- Fabric provided ENode MACs are mapped from the FC_MAP and FC_ID parameters provided by the FCF during discovery and login.



Calculation assumptions

Results of analysis depend on size, configuration, and number of systems.

- Assume two complete and independent computer complexes, configured identically with identical computing, networking and storage hardware and with identical application distribution.
- Assume 500 host ENodes per complex. All nodes under the same Virtual Center, if applicable.
- Assume 125 FCF ports per system
- Assume 125 target SAN ports per system
- Assumes each ENode port links with 4 FCF ports
- Assume reads and writes roughly equal in number and size.
- Assume average read or write is 16k Bytes, 8 frames.



FC Requirements to establish stream of data

Successful Discovery of MAC Addresses

- Establishes FCoE valid paths

Successful FLOGI to Initiator

- Establishes Originator FC_ID

Successful FLOGI to Target

- Establishes Responder FC_ID

Successful PLOGI

- Binds Originator and Responder FC_ID

Successful PRLI

- Binds Initiator and Target relationship, Originator to Responder.



Probability of duplicate MAC addresses

Identified failures require at least one duplicate MAC addresses

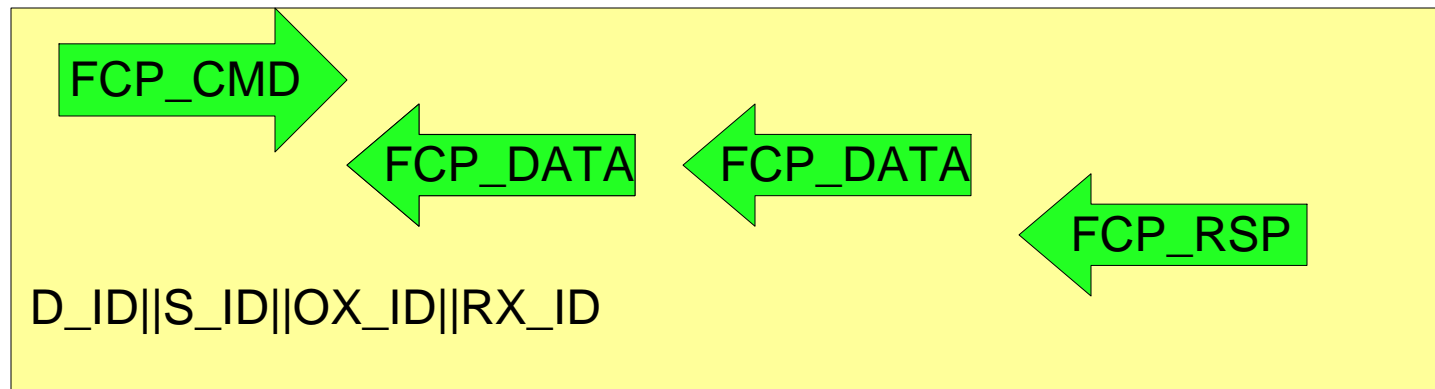
Subsequent probability estimates assume:

- Failure requires ENode MAC address to be duplicated:
 - If all ENode MAC addresses "burnt in" $P=0$
 - If all ENode MACs use VMWare model of MAC guaranteed unique by Virtual Center in each complex. $P=0.015$
 - This is NOT a Birthday-Bounds problem.
 - Simple sampling without replacement in a 24-bit space.
 - $\sim = \text{MACs selected}(\text{MACs assigned}/\text{total MACs})$
 - $\sim = 500(500/16,777,216)$
 - If all ENode MAC addresses Mapped $P \sim = 1$
 - Assumes 2 x 500 mapped ports joining
 - No mapping administration
 - FC_MAP and assigned FC_IDs same for both complex.

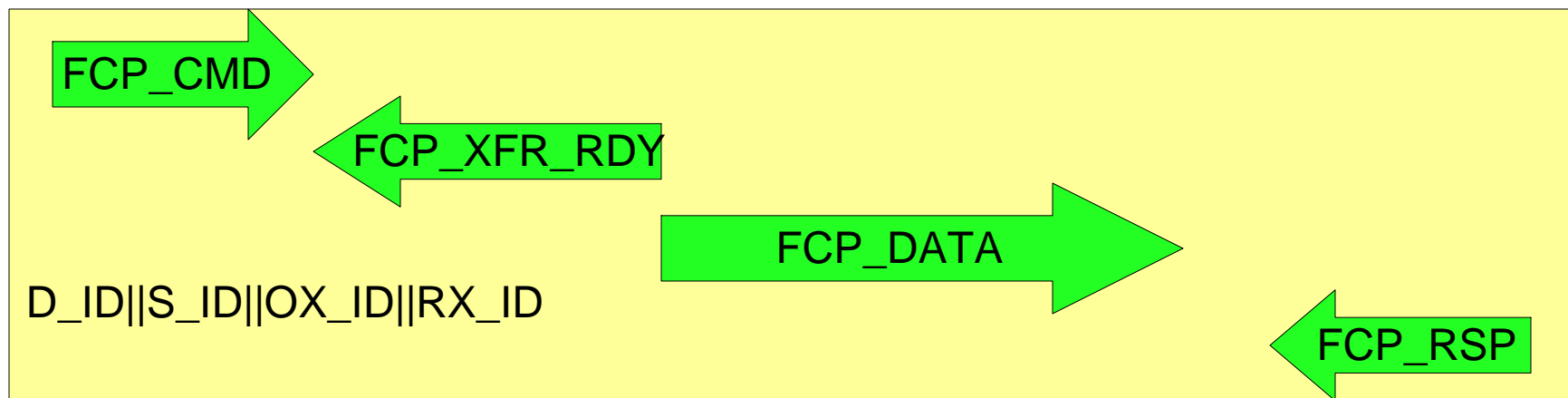


Each stream consists of I/O Operations:

Reads



Writes

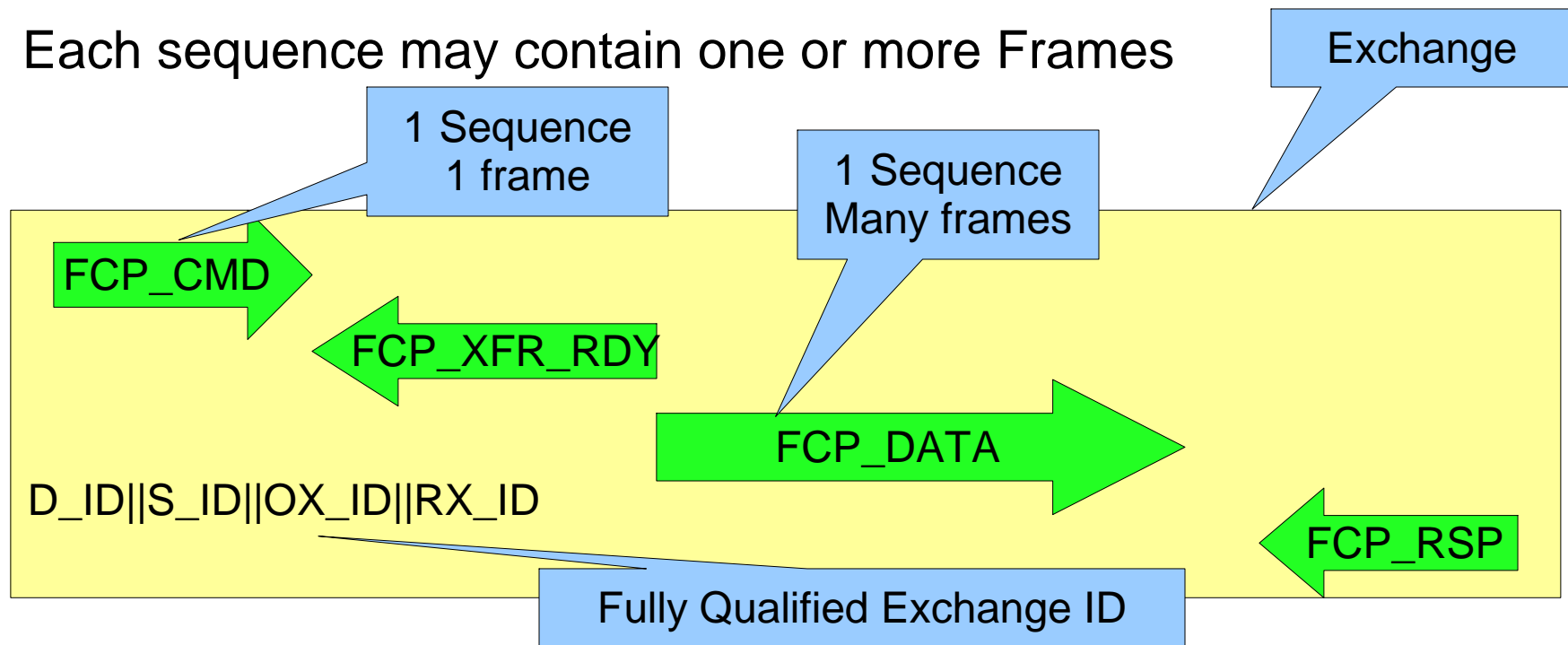


Each I/O Operation in FCP consists of

A single Exchange, transmitting two or more Information Units

Each Information Unit corresponds to a single FC sequence

Each sequence may contain one or more Frames



Possible data integrity failure sources

Case 1: Combining of data flows such that a duplicate FCP_CMD IU occurs.

- Causes one or more subsequent data operations to be performed incorrectly.

Case 2: Combining of data flows such that a frame or sequence in an FCP_DATA IU in one flow is replaced by a similar frame or sequence from another flow.

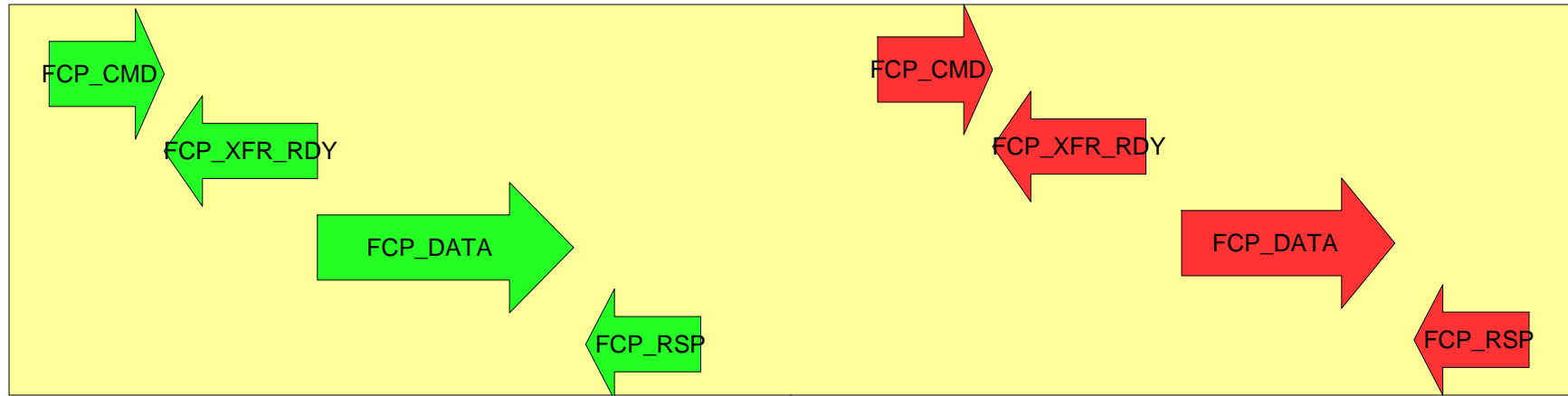
- Causes one or more blocks of data to be replaced with the wrong data.

Case 3: Combining of data flows such that a duplicate FCP_RSP IU occurs.

- Truncates data
- Provides incorrect response terminating data operation.



Timing conditions necessary for combined streams to interact in FCoE space, Case 1:



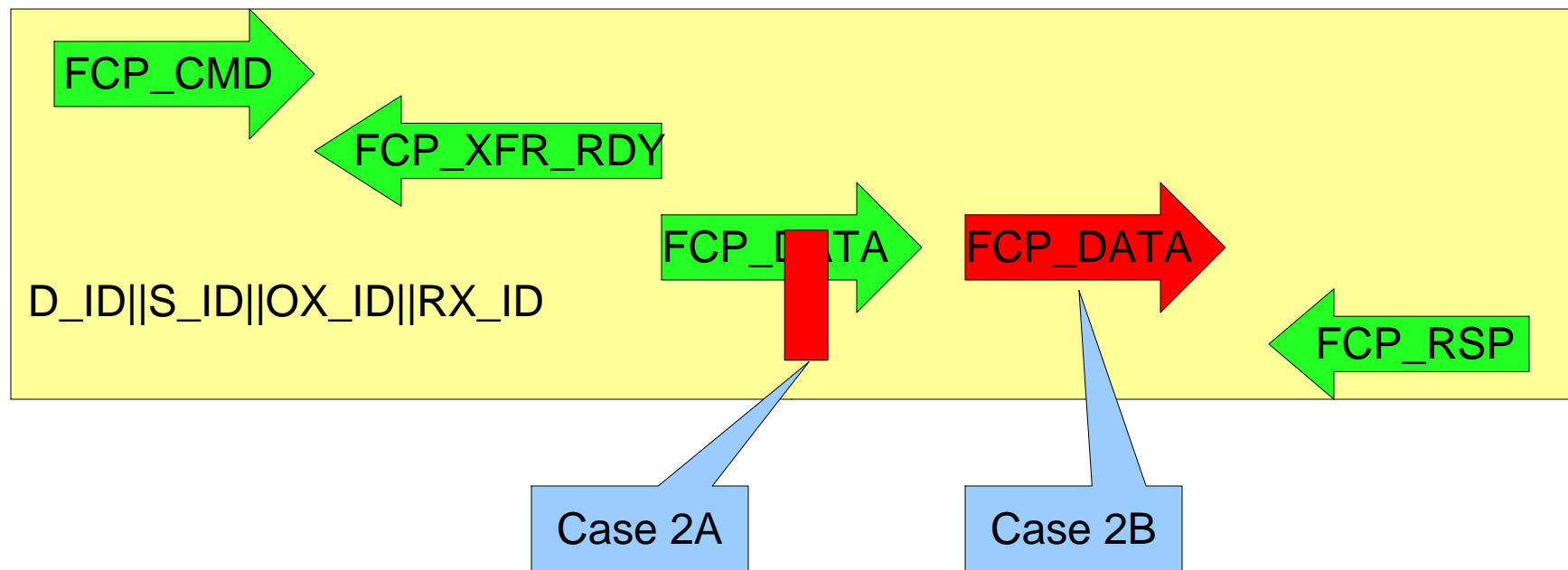
Conditions necessary for combined streams to interact in FCoE space, Case 1:

Summary: Probability of Command arriving at incorrect target is essentially zero. See backup slides B.

- FCP_CMD IU must travel to incorrect but valid FCF MAC from valid MAC address
- FCP_CMD IU must have a valid FC D_ID to be routed and a valid S_ID to be accepted by the remote target.
- FCP_CMD IU must be directed to a valid LUN
- Incorrect data stream assignment must be symmetrically valid for entire period of command operation.
- Permissions on zoning must be valid
- Permissions on LUN must be valid



Timing conditions necessary for combined streams to interact in FCoE space, Case 2:



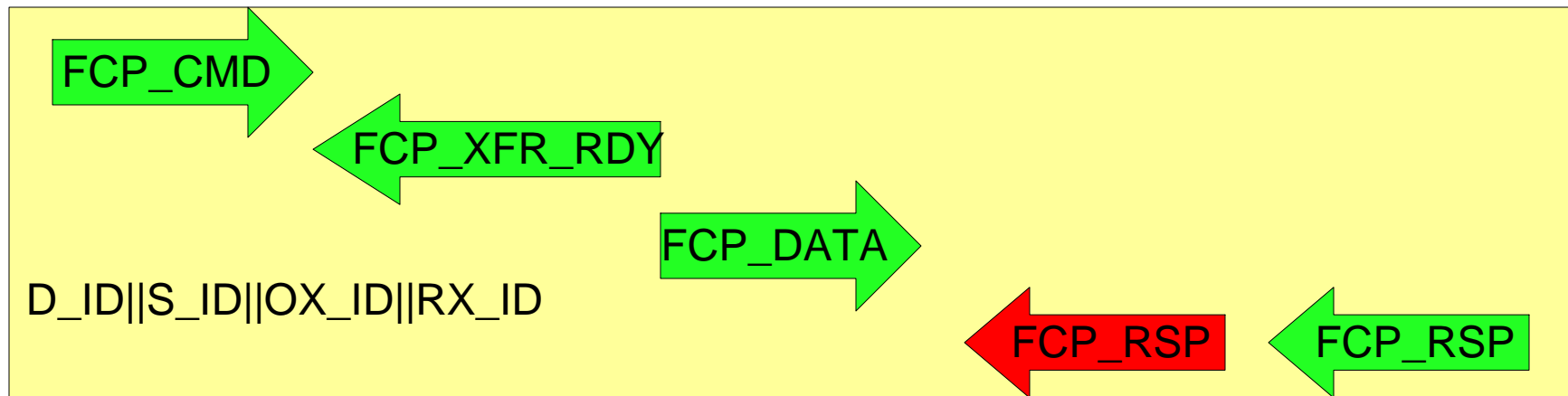
Conditions necessary for combined streams to interact in FCoE space, Case 2:

Summary: Probability of successfully inserting incorrect data frames with no SA checking is a maximum of 5.5×10^{-14} , with high rate of detected errors. With SA checking, probability is zero. See backup slides C.

- FCP_Data must travel to valid MAC from valid MAC address
- FCP_Data must have same FC_ID
- FCP_Data must have same fully qualified exchange identifier
- FCP_Data must have consistent sequence identifier
- FCP_Data must have consistent sequence count
- FCP_Data must have consistent Parameter Field
- Replaced FCP_Data must not be presented



Timing conditions necessary for combined streams to interact in FCoE space, Case 3:



Conditions necessary for combined streams to interact in FCoE space, Case 3:

Summary: Probability of combined streams providing false FCP_RSP is about 8.8×10^{-17} , with high rate of detected errors. See Backup Slides D.

- FCP_Data must travel to valid MAC from valid MAC address
- FCP_RSP must have same FC_ID
- FCP_RSP must have same fully qualified exchange identifier
- FCP_RSP must have consistent sequence identifier
- FCP_RSP must have consistent sequence count
- FCP_RSP must occur at proper time in protocol
- Replaced FCP_RSP must not be presented



Conclusions: The Bad

With careful attention to implement worst practices, data integrity failure rates may be generated with about 10^{-14} probability per data frame. At 10 Gb/s and full utilization, that is about one failure per 5.4 years per complex.

Worst practices:

- Use of Fabric Provisioned MAC addresses with no FC-MAP administration.
- Use of multiple VMWare environments with separate Virtual Centers within physical proximity close enough that their FCoE layer 2 Ethernet may accidentally be connected.



Conclusions: The Good

With careful attention to implement best practices, data integrity failure rates are essentially 0.

Best practices:

- Use Server-Provided MAC addresses.
- Use MAC Source Address Checking at all ENodes. This should be typical of most implementations.
- Use Continuously Increasing Sequence Count.
- Use a single Virtual Center for all systems within physical proximity close enough that their FCoE layer 2 Ethernet may accidentally be connected.
- Assure physical security procedures for administration.
- Use parameter field (FCP data offset) checking at all ENodes.
- Protect Ethernet switches from learning attacks with ACL.



BROCADE



Backup Slides on statistics

A:

Helpful statistical formulas (Wikipedia)

The birthday problem can be generalised as follows: given n random integers drawn from a discrete uniform distribution with range $[1, d]$, what is the probability $p(n; d)$ that at least two numbers are the same? A reasonable approximation is given by:

$$p(n; d) \approx 1 - e^{-\frac{n(n-1)}{2d}}$$



B:

Probabilities for combined streams to interact in FCoE space, Case 1: (A)

a) FCP_CMD IU must travel to valid MAC from valid MAC address

- Discovery process defines valid links, Enode to FCF
- Because all FCF nodes are "burnt in" and therefore not duplicated, the command may ONLY arrive at the correct address.
- The probability of this condition being met is $P(a) = 0$.



B:

Probabilities for combined streams to interact in FCoE space, Case 1: (B)

b) FCP_CMD IU must have same or a valid FC D_ID and S_ID

- FC_IDs assumed to be assigned with common domain names and area names. The probability of this condition being met is $P(b) = 1$.

c) FCP_CMD IU must be directed to a valid LUN

- LUN assignment is configuration dependent. The probability of this condition being met is $0 \leq P(c) \leq 1$.



B:

Probabilities for combined streams to interact in FCoE space, Case 1: (C)

d) Incorrect data stream assignment must be symmetrically valid for entire period of command operation.

- Back traffic through FCF will have the proper replicated MAC. It will be delivered to whichever location was learned most recently.
- Probability of failure depends on link utilization and command latency. Commands that fail time out and are detected.
- Probability uncertain $0 \leq P(d) \leq 1$

e) Permissions on zoning must be valid

- Zoning is determined by WWN, but enforced by FC_ID. Identical systems will have identical FC_ID zoning. $P(e) = 1$

f) Permissions on LUN must be valid

- Simple systems will not use LUN zoning or reservations. $P(f) = 1$



B:

Result, Case 1:

The probabilities are independent to a first approximation. The effective probability of a failure is the product of the probability of each failure:

$$P(\text{failure, Case A}) = P(a) P(b) P(c) P(d) P(e) P(f) = 0$$

- FCP_CMD IU must travel to incorrect but valid FCF MAC from valid MAC address
 - $P(a) = 0$
- FCP_CMD IU must have a valid FC D_ID to be routed and a valid S_ID to be accepted by the remote target.
 - $P(b) = 1$
- FCP_CMD IU must be directed to a valid LUN
 - $P(c) = \text{configuration dependent}$
- Incorrect data stream assignment must be symmetrically valid for entire period of command operation.
 - $P(d) = \text{utilization dependent}$
- Permissions on zoning must be valid
 - $P(e) \leq 1$
- Permissions on LUN must be valid
 - $P(f) = 1$



C:

Probabilities for combined streams to interact in FCoE space, Case 2 (A):

a) FCP_Data must travel to valid MAC from valid MAC address

- Write transfer is routed to FCF. Probability of mixing data to burned in FCF address is zero.
- Read data is routed to ENode. Data may be routed to either replicated ENode with undeterminable probability. Consider probability to be 0 for Server provisioned MAC, about 0.015 for VMWare provisioned MAC, about 1 for Fabric provisioned MAC unless properly administered. Probability reduced to zero if SA is tested for validity. 50% of data transfers are reads, reducing worst case probability to 50%.
- $P(a) \leq 0.5$ (0 with SA checking)

b) FCP_Data IU must have same FC D_ID

- FC_IDs assumed to be assigned with common domain names and area names. The probability of this condition being met is $P(b) = 1$.



C:

Probabilities for combined streams to interact in FCoE space, Case 2 (B):

- c) FCP_Data must have same fully qualified exchange identifier
- Fully qualified exchange identifier most variable part is OX_ID. RX_ID may be constant, D_ID and S_ID likely to be similar.
 - OX_ID is 16-bit number. Probability of match $P(c) = 1/(2^{16}) = 15 \times 10^{-6}$
- d) FCP_Data must have consistent sequence identifier
- SEQ_ID is an 8-bit number unique between a D_ID and S_ID. Probability of match $P(d) = 1/(2^8), = 3.9 \times 10^{-3}$



C:

Probabilities for combined streams to interact in FCoE space, Case 2 (C):

e) FCP_Data must have consistent sequence count

- SEQ_CNT is a 16-bit field, assumed to be continuously increasing, though other cases are possible.
- Probability of matching $P(e) = 1/(2^{16})$, = 15×10^{-6}

f) FCP_Data must have consistent Parameter Field

- Parameter field contains information displacement, in our example one of 8 values (one for each data frame).
- Probability of matching $P(f) = 1/8$, = 12.5×10^{-2}



C:

Probabilities for combined streams to interact in FCoE space, Case 2 (D):

g) Replaced FCP_Data must not be presented

- Assumes that two data streams are combined by a learning attack. Assumes that identical frames are somehow interleaved.
- Two error notifications will occur with probability = 1.
 - The device from which the frames were diverted will either time out and post an error or detect missing frames and post an error.
 - The device to which the frames were diverted will post an error associated with frame ordering because it received duplicate frames.
- If two data streams are combined by a dual learning attack that causes identical streams to be perfectly swapped, no error will be detected. The probability of such an event depends on the data stream, but undetected occurrence is close to zero because of the above tests and because of the uncertainty of timing of the switch changes.
- P(g) 0 or very low, traffic dependent.



C:

Result, Case 2:

The probabilities are independent to a first approximation. The effective probability of a failure is the product of the probability of each failure: $P(\text{failure, Case A}) = P(a) P(b) P(c) P(d) P(e) P(f) P(g) \leq 5.5 \times 10^{-14}$

- FCP_Data must travel to valid MAC from valid MAC address

Fabric Provisioned MAC, SA not tested, $P(a) = .5$

If SA tested, $P(a) = 0$

- FCP_Data must have same FC_ID
 - $P(b) = 1$
- FCP_Data must have same fully qualified exchange identifier
 - $P(c) = 15 \times 10^{-6}$
- FCP_Data must have consistent sequence identifier
 - $P(d) = 3.9 \times 10^{-3}$
- FCP_Data must have consistent sequence count
 - $P(e) = 15 \times 10^{-6}$
- FCP_Data must have consistent Parameter Field
 - $P(f) = 12.5 \times 10^{-2}$
- Replaced FCP_Data must not be presented
 - $P(g) = 0$ over time

Maximum limit of probability is 5.5×10^{-14} , approaching zero over time.



D:

Probabilities for combined streams to interact in FCoE space, Case 3: (A)

- a) FCP_RSP must travel to valid MAC from valid MAC address
- FCP_RSP is routed to ENode. Data may be routed to either replicated ENode with undeterminable probability. Consider probability to be 0 for Server provisioned MAC, about 0.007 for VMWare provisioned MAC, about 1 for Fabric provisioned MAC unless properly administered. Probability reduced to zero if SA is tested for validity.
 - $P(a) \leq 0.5$ (0 with SA checking)
- b) FCP_RSP must have same FC_ID
- FC_IDs are likely to be replicated in identical systems. The probability of this condition being met is $P(b) = 1$.



D:

Probabilities for combined streams to interact in FCoE space, Case 3: (B)

- c) FCP_RSP must have same fully qualified exchange identifier
- Fully qualified exchange identifier most variable part is OX_ID. RX_ID may be constant, D_ID and S_ID likely to be similar.
 - OX_ID is 16-bit number. Probability of match $P(c) = 1/(2^{16}) = 15 \times 10^{-6}$
- d) FCP_RSP must have consistent sequence identifier
- SEQ_ID is an 8-bit number unique between a D_ID and S_ID. There is no value to compare it with, so $P(d) = 1$.
- e) FCP_RSP must have consistent sequence count
- SEQ_CNT is a 16-bit field, assumed to be continuously increasing, though other cases are possible.
 - Probability of matching $P(e) = 1/(2^{16}), = 15 \times 10^{-6}$



D:

Probabilities for combined streams to interact in FCoE space, Case 3: (B)

- f) FCP_RSP must occur at proper time in protocol
- Only failing time is after last data transfer and before transmission of correct FCP_RSP. All other cases will post a protocol error associated with incomplete data transfer. Time is about 1 microsecond out of 10 msec. $P(f) = 10^{-4}$
 - The device from which the FCP_RSP was diverted will time out and post an error, posting an error with probability = 1.
- g) Replaced FCP_RSP must not be presented
- Replaced FCP_RSP is discarded because it is received after the end of the Exchange. If FCP_CONF is requested, an additional FCP_RSP will post a protocol error with probability = 1.



D:

Result, Case 3:

The probabilities are independent to a first approximation. Effective probability of failure is the product of the probability of each: $P(\text{failure, Case 3}) = P(a) P(b) P(c) P(d) P(e) P(f) P(g) = 0$

FCP_RSP must travel to valid MAC from valid MAC address

- Fabric Provisioned MAC, SA not tested, $P(a) = 1$
- If SA tested, $P(a) = 0$

FCP_RSP must have same FC_ID

- $P(b) = 1$

FCP_RSP must have same fully qualified exchange identifier

- $P(c) = 15 \times 10^{-6}$

FCP_RSP must have consistent sequence identifier

- $P(d) = 3.9 \times 10^{-3}$

FCP_RSP must have consistent sequence count

- $P(e) = 15 \times 10^{-6}$

FCP_RSP must occur at the right time ($\frac{1}{2}$ will indicate failures anyway)

- $P(f) = 10^{-4}$

Replaced FCP_RSP must not be presented

- $P(g) = 1$ if no FCP_CONF requested.

Maximum limit of probability is 8.8×10^{-17} , with high error rate symptoms.



E:

Probability of failure to failures per time:

A 10^{-14} probability per data frame at 10 Gb/s, is roughly one failure per 5.4 years.

- 10 Gb/s = 1.25 GB/s
- Frame length = 2116 bytes plus gaps
- Time/frame = 2116 Bytes per frame / 1.25 GB per sec
= 1.692 microseconds.
- Time for 10^{+14} frames = $1.692 * 10^{+8}$ seconds = 5.4 years

