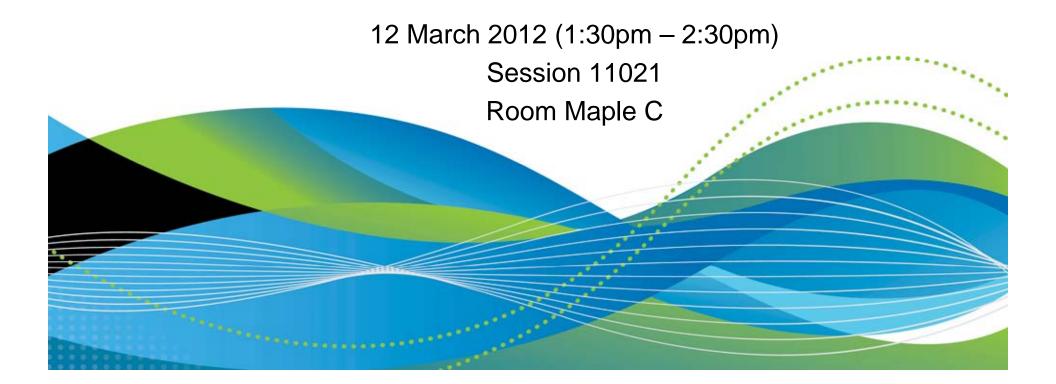




Buffer-to-Buffer Credits, Exchanges, and Urban Legends

Howard L. Johnson, Brocade





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.





Abstract

• Performance in a FICON network is influenced by the underlying flow control mechanisms of Fibre Channel. In this session, we examine how Buffer-to-Buffer credits flow from the channel to the control unit. We also look at how exchanges are used in FICON applications and how they change with the introduction of zHPF. During both examinations, we explore the role of the FICON Director in managing Buffer-to-Buffer credits and exchanges over a cascaded network. Throughout the session, we debunk the various FICON "Urban Legends" featuring credits and exchanges. Take the opportunity to learn from two of the FICON industry's leading experts in channel and fabric development and join our session.





Agenda

- Buffer Credits
 - What are they and how do they work?
 - How do you <u>fill the pipe</u>?
 - What if you <u>can't fill the pipe</u>?
 - What's wrong with multiple senders and one receiver?
 - What happens when the pipes are different sizes?
 - What's it like in the real world?
 - How do cascades Directors work?
- Exchanges
 - What are they and how do they work?
 - What's an Exchange?
 - How many exchanges are needed?
 - Can they be "reused?"
 - Can you have too many exchanges?
- Error Sensitivity
 - Is FICON more sensitive to errors than FCP?
 - How sensitive are FICON frames to loss or corruption?
 - What recovery actions are taken?
 - What are the differences with FCP?





What are they and how do they work?

BUFFER CREDITS





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?

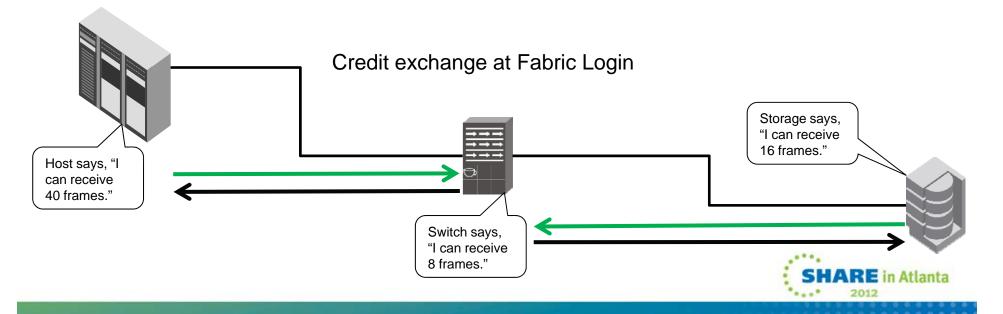
- The greater the BB Credit....
 - A.
 - B. The farther apart the two ports can be
 - C.
 - D.





Flow Control

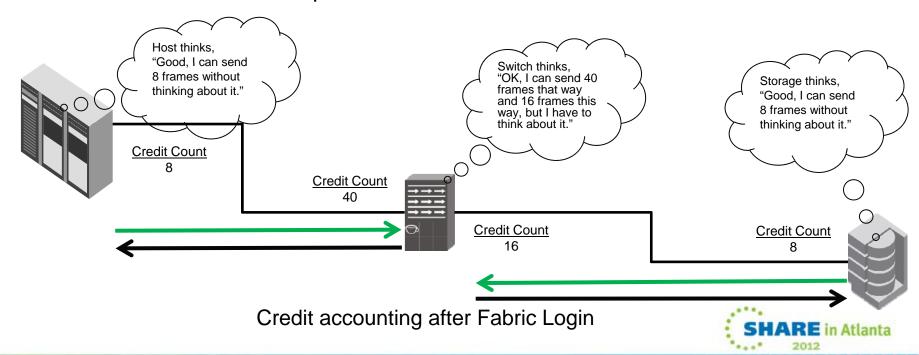
- Related to the devices' ability to receive and process frames
- Manages when frames are coming faster than they can be processed
- Dropped frames occur when frames are arriving too fast to be processed
- Frames can only be transmitted when the receiver is ready
- Credit establishment communicates the number of frames a device can receive at a time
- The credit value is exchanged at login
- Transmission stops when credit runs out
- The receiver indicates when it is ready to receive more frames





Buffer Credit

- At initialization, the two ports establish credit
 - Each buffer credit corresponds to a frame (regardless of size)
- Each side can support different values
 - Credit Count
- If a port doesn't have credit, it can't send a frame
 - Credit Count has reached zero
- Mechanism limits frame drops

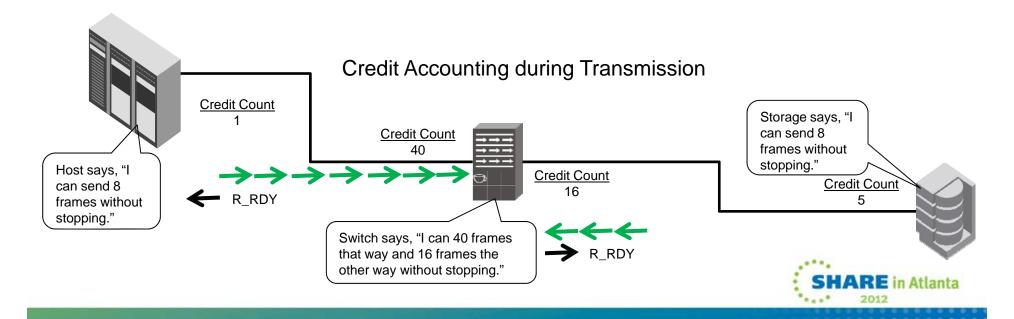




Receiver Ready (R_RDY)

- R RDY
 - Used for link level flow control
 - Called buffer-to-buffer credit (BB Credit)
- R_RDY is not a frame
 - It is a "primitive" so it doesn't consume a buffer

- Frame transmission
 - BB Credit is decremented
 - Once for each frame transmitted
 - When BB Credit = 0
 - Transmission stops
- Frame received
 - R_RDY is sent
 - Causes transmitter to increment BB Credit



Urban Legend: Buffer Credits at Zero are a Problem



- Buffer credit determines DISTANCE
 - The distance two nodes can be apart and still maintain full link frame rate
- Buffer credit is the number of FRAME buffers
 - A port provides for it's NEAREST neighbor for RECEIVING frames
 - Does NOT have to be symmetrical
- Buffer credit is a FRAME count.
 - Not a data SIZE
 - A 1 byte frame consumes 1 buffer credit
 - A 2K byte frame consumes 1 buffer credit
- Number of credits needed is determined by:
 - Raw Link Speed
 - Speed of light thru a fiber
 - Distance between two adjacent nodes





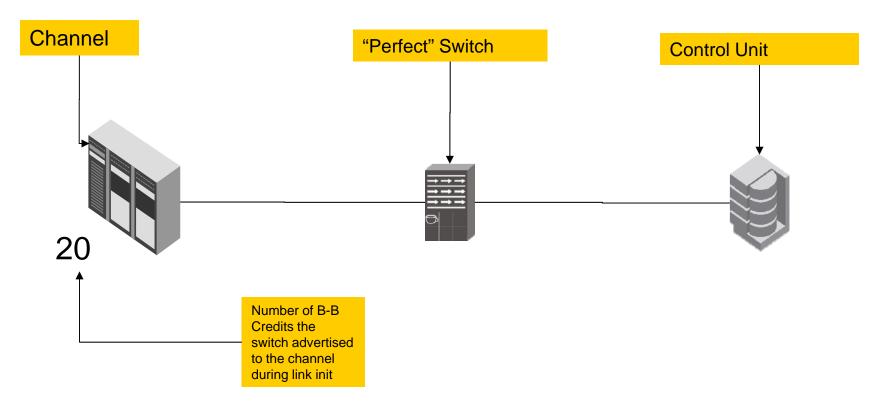
Example: A full pipe

BUFFER CREDITS



Initial Conditions

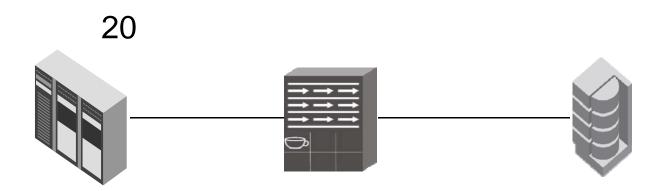




NOTE: In these animations, both the frames and R_RDY's are numbered. This is for illustrative purposes only. In reality, neither the frames nor the R_RDY's are numbered. The arrival of an R_RDY only informs the receiver that A frame has been forwarded, now WHICH frame has been forwarded.

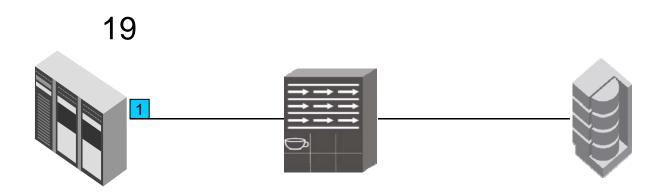






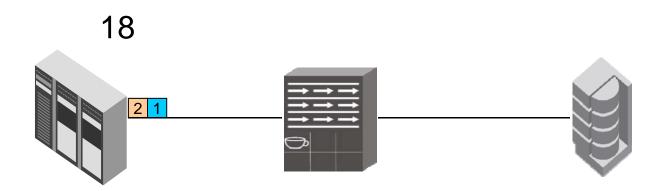






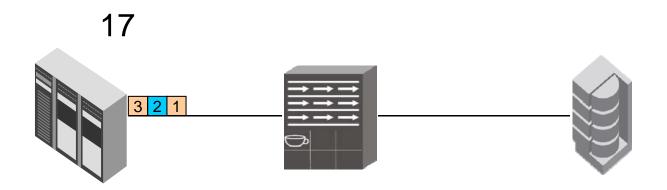






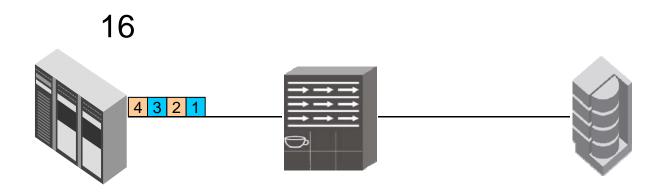






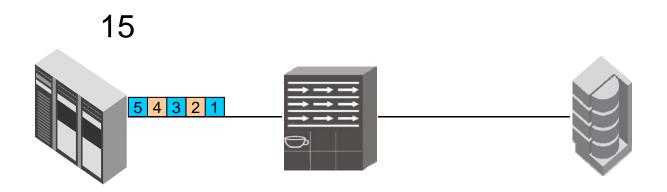






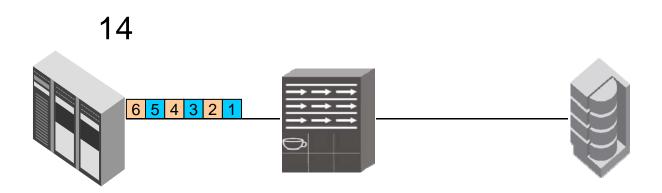






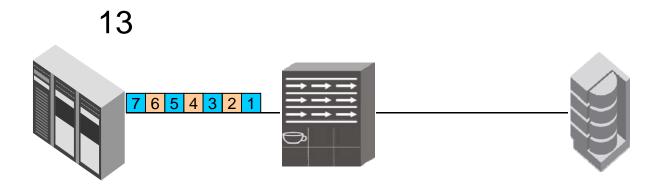






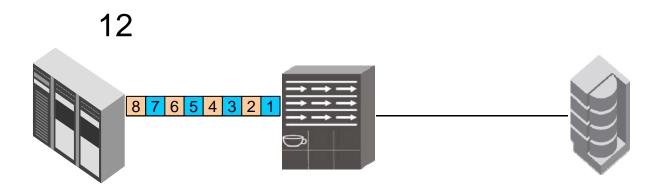






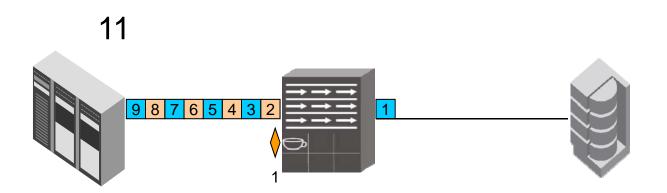






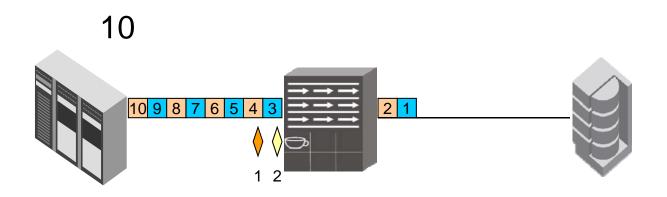






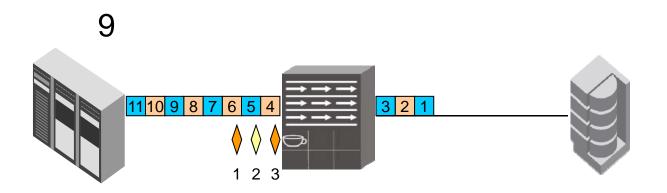






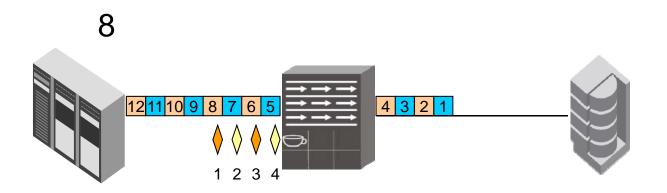






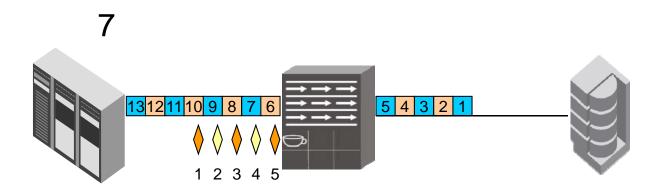






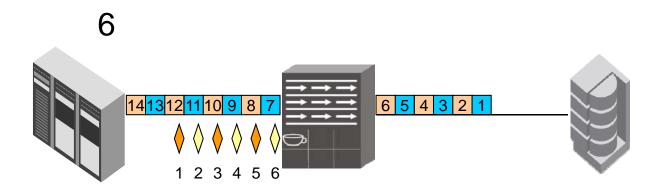






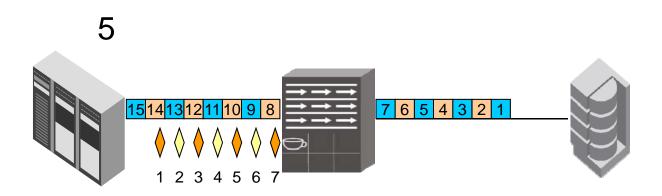






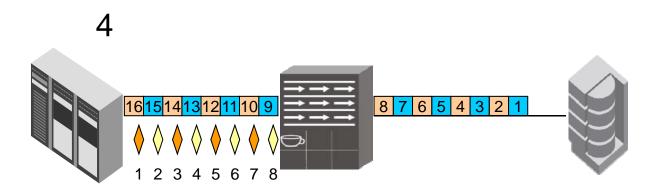






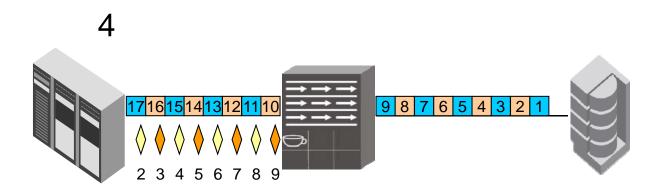








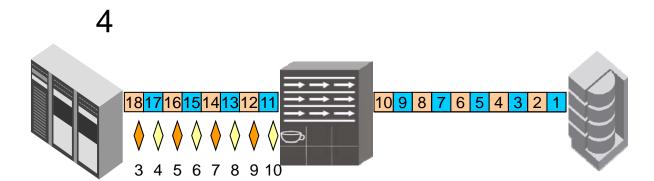






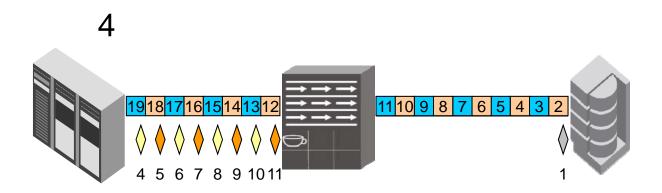
















THIS PAGE INTENTIONALLY LEFT BLANK





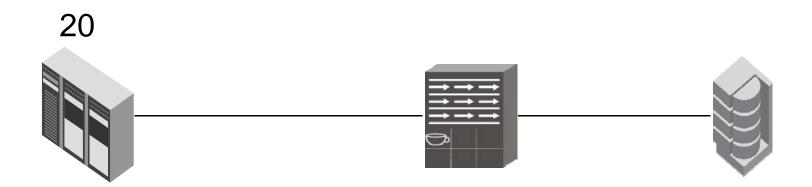
Example: A not so full pipe

BUFFER CREDITS



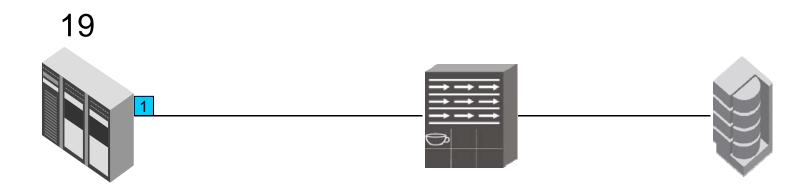


Suppose the switch is too far away from the channel for the B-B credit it advertised to the channel



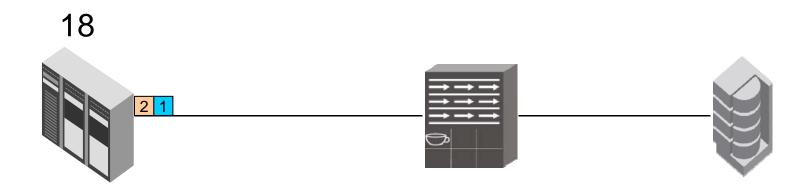






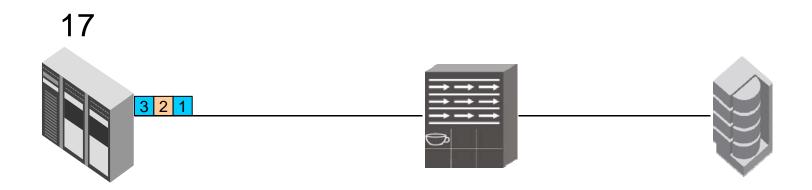






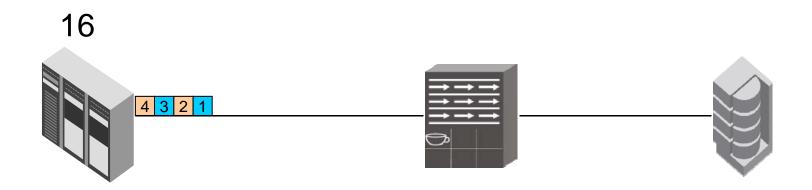






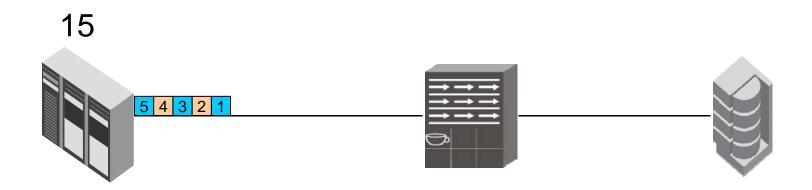






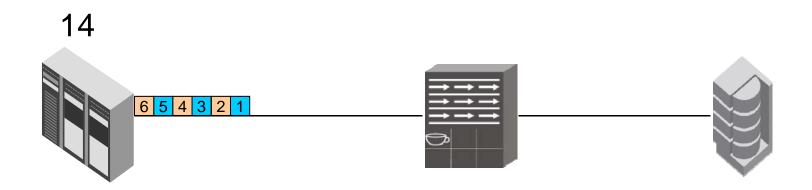






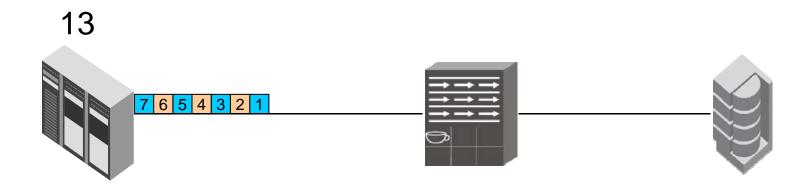






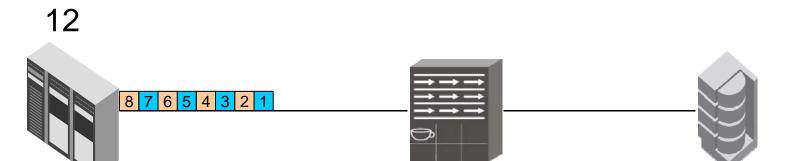






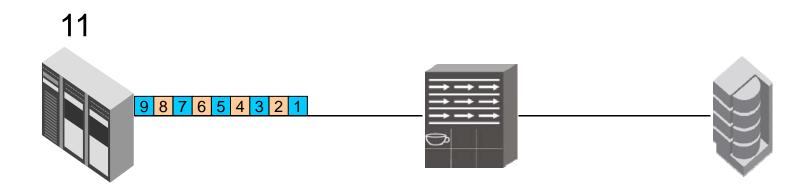






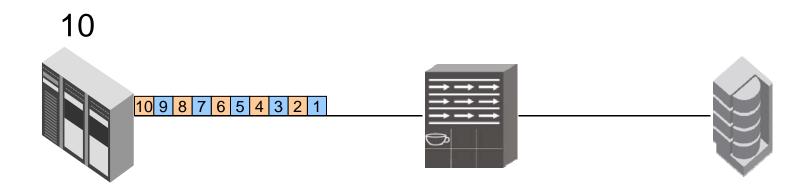






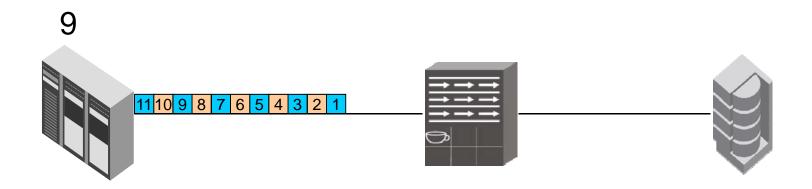






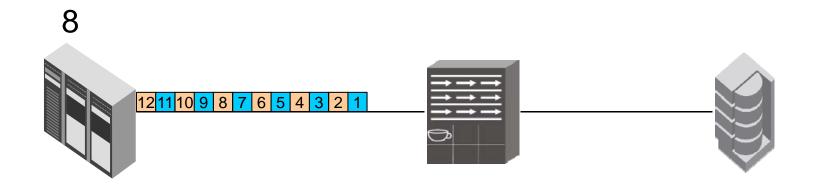






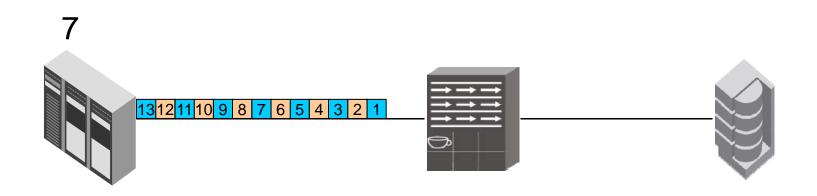






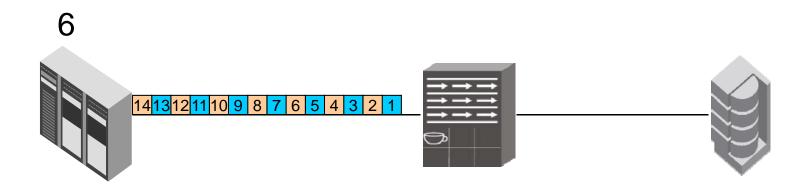






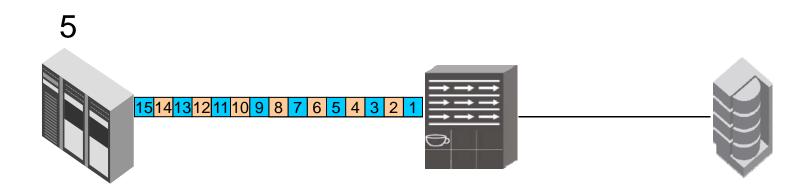






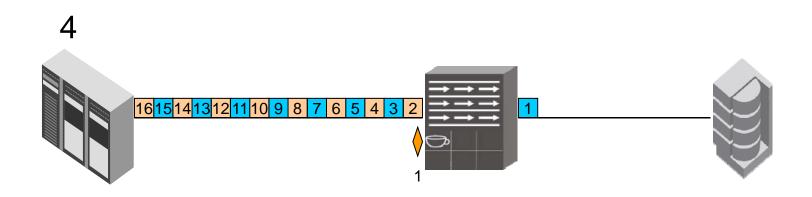






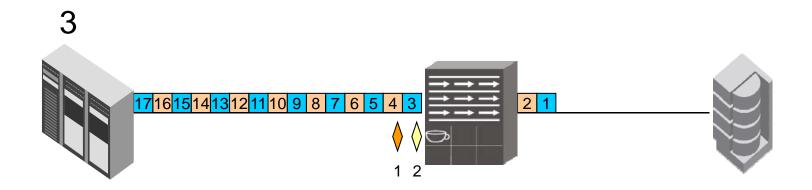








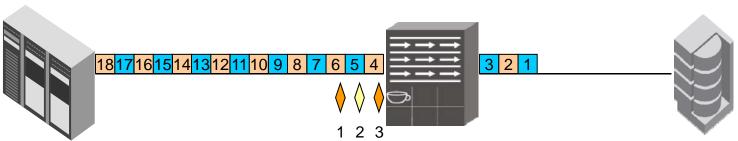






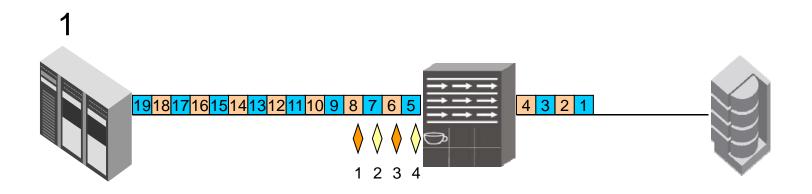






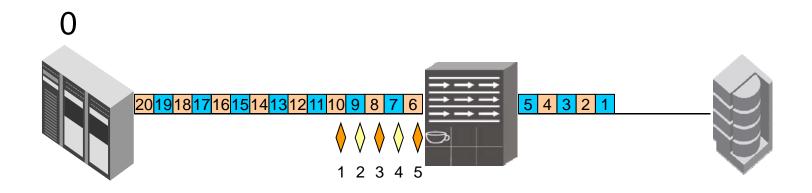






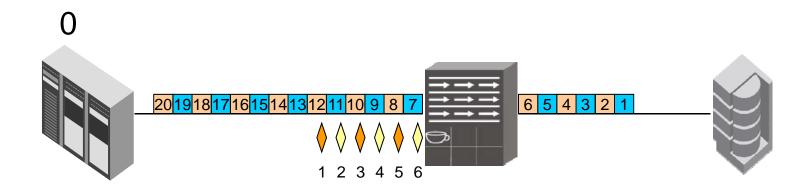






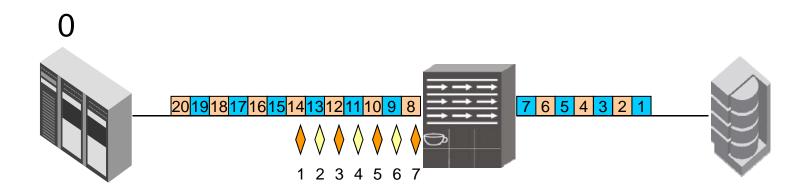






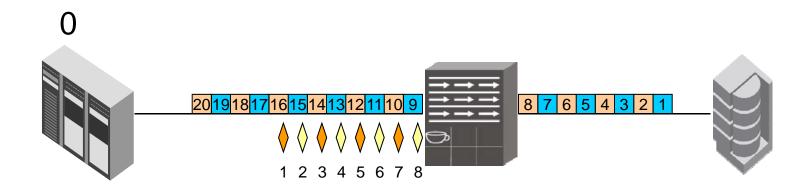






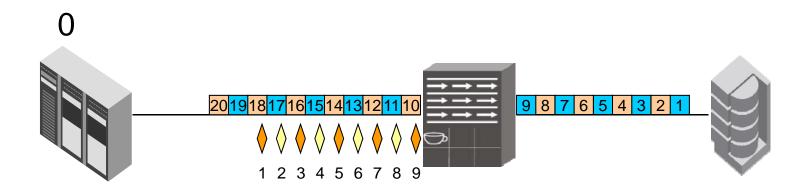






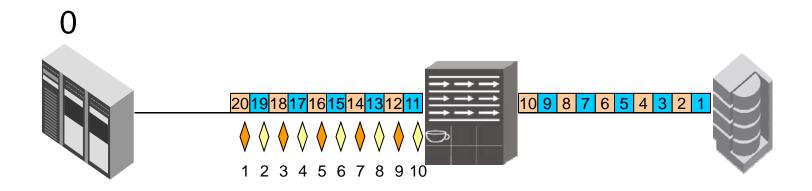






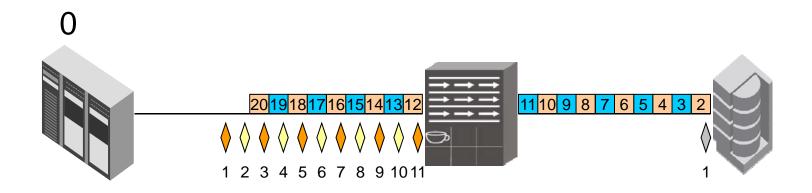






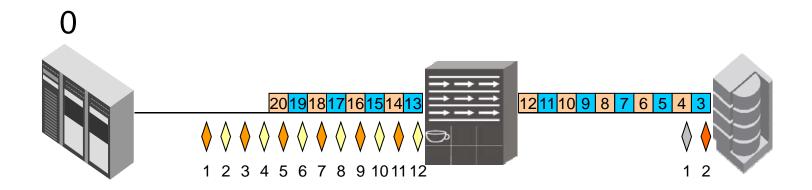






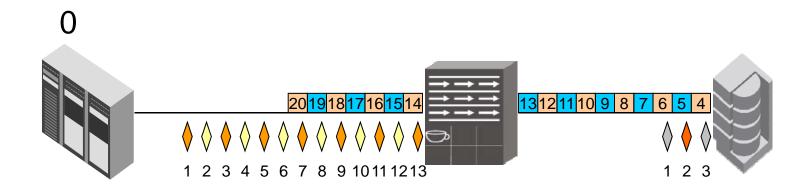






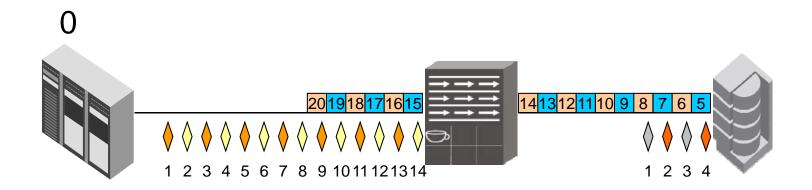






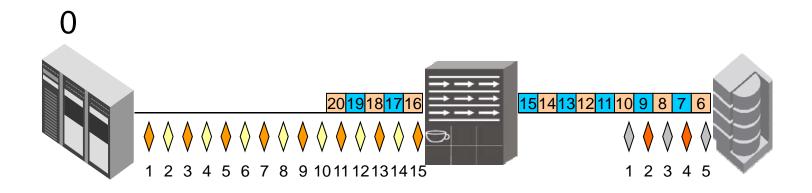






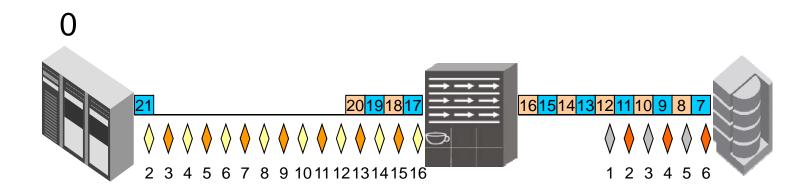






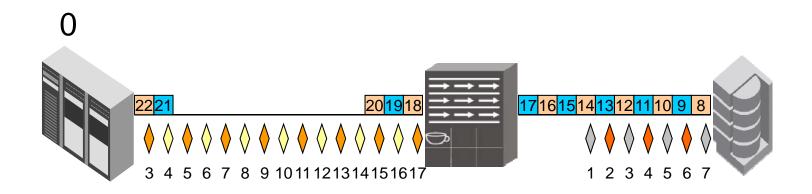






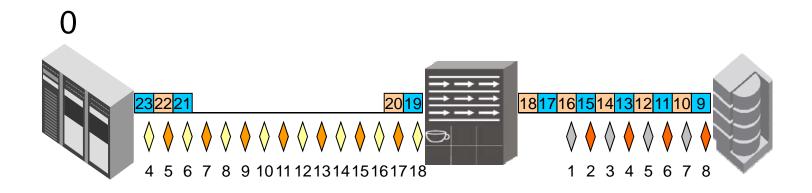






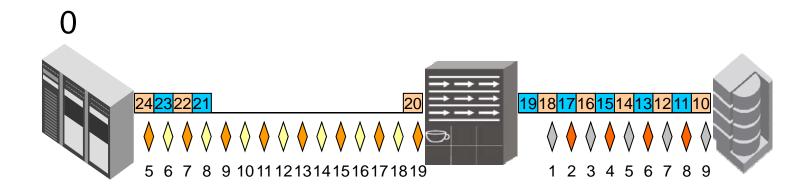






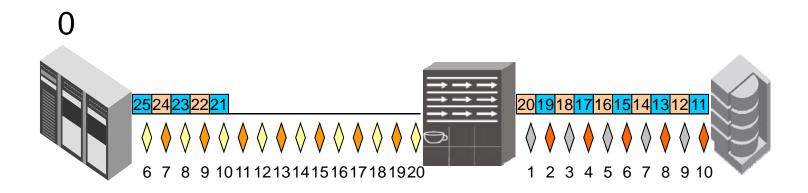






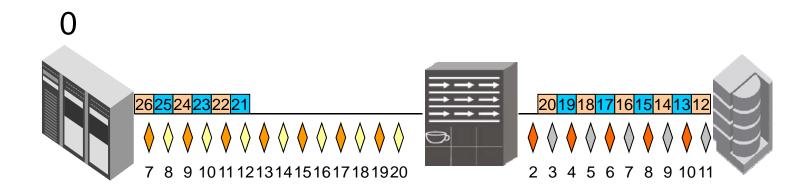






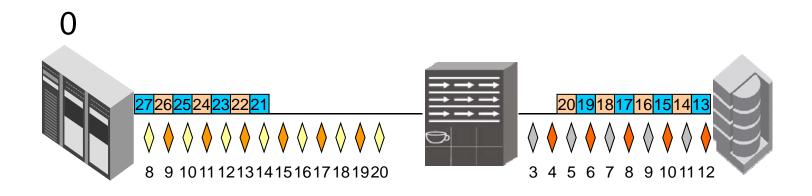






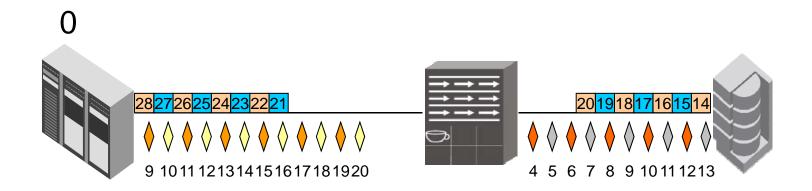






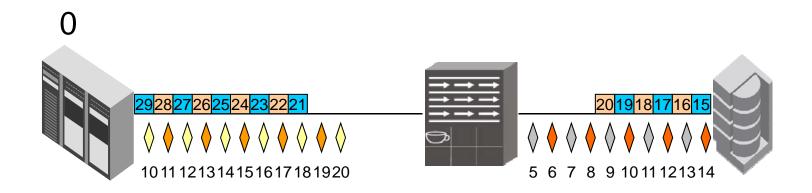






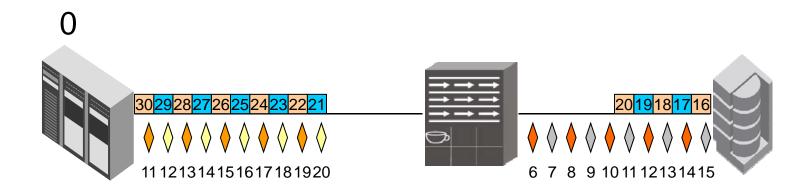






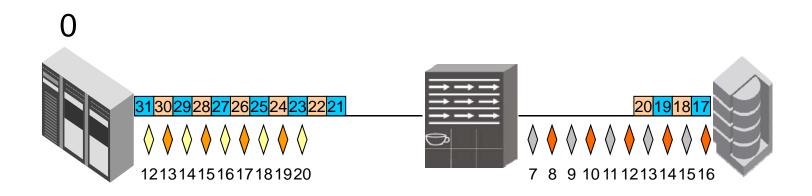






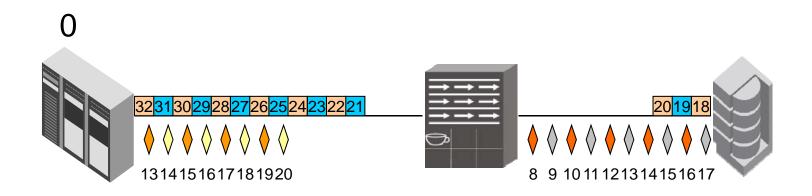






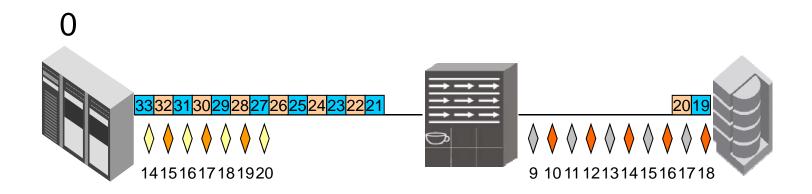






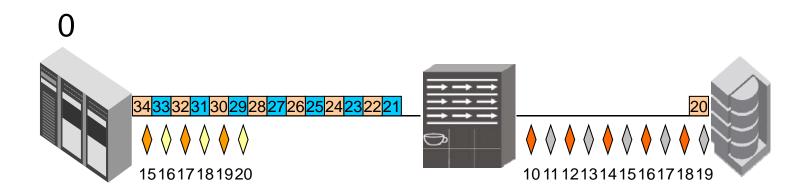






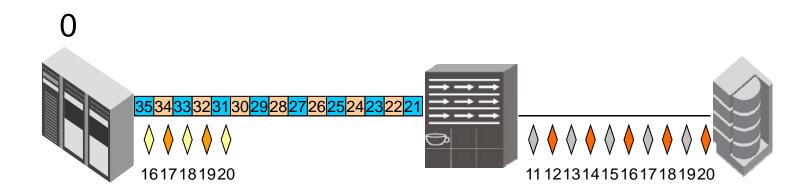






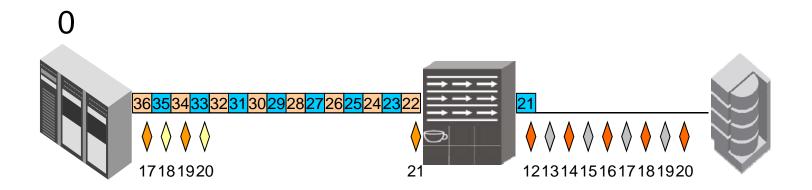






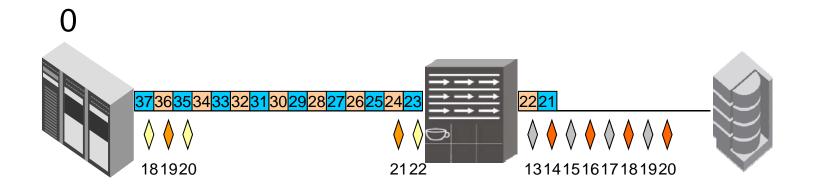






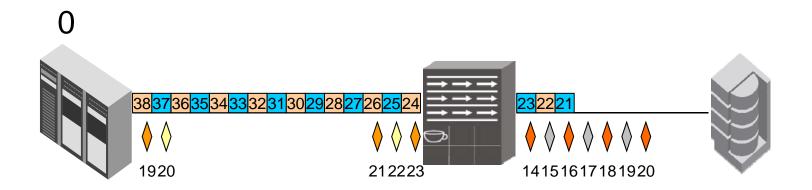






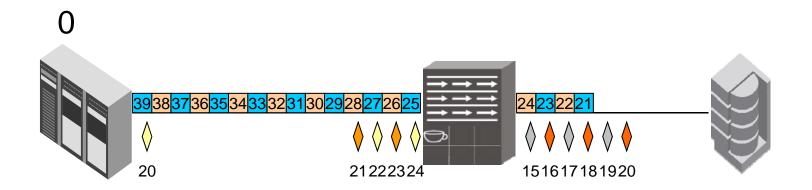






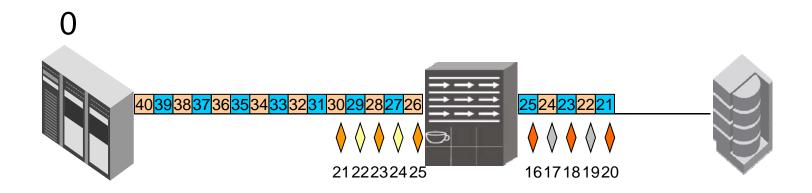






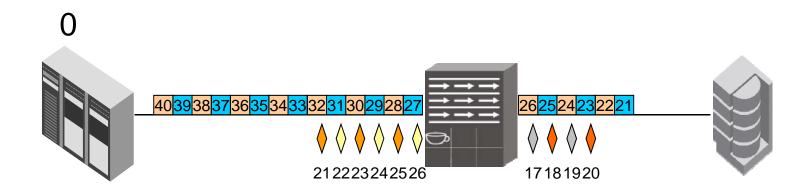






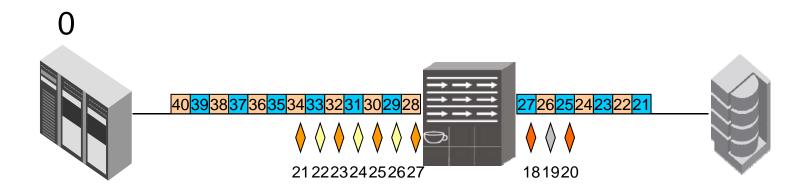






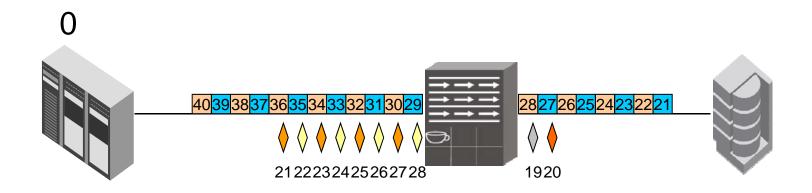






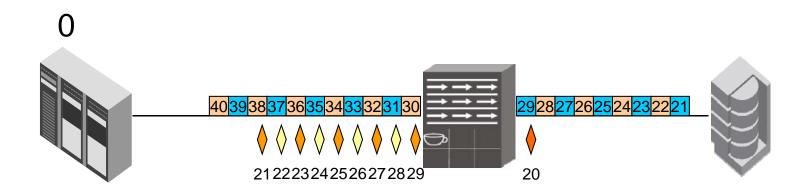






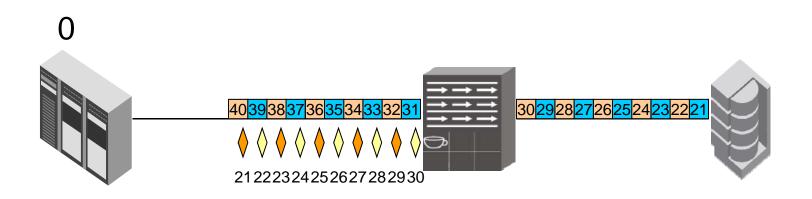








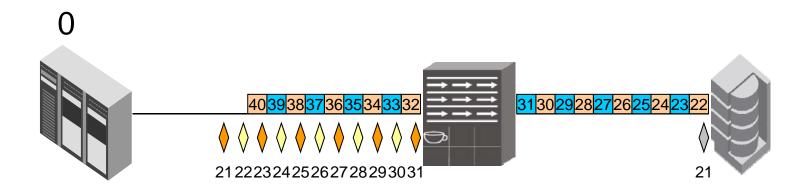






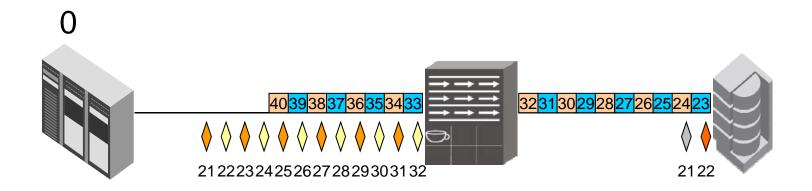
















THIS PAGE INTENTIONALLY LEFT BLANK





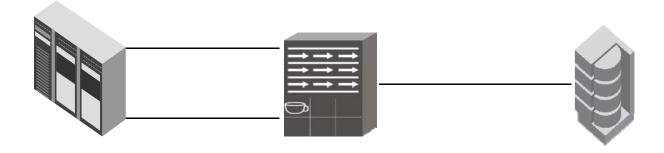
Example: Multiple Senders and One Receiver

BUFFER CREDITS





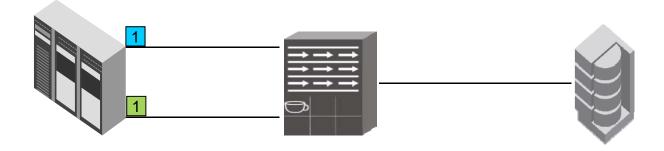
Suppose there are multiple senders to one receiver Each sender attempts to send at 100% link speed





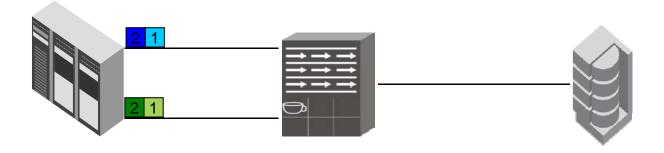








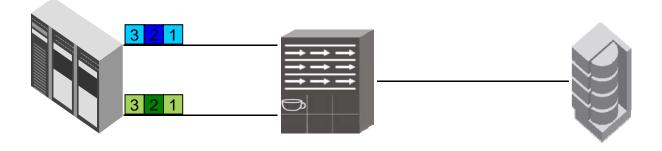






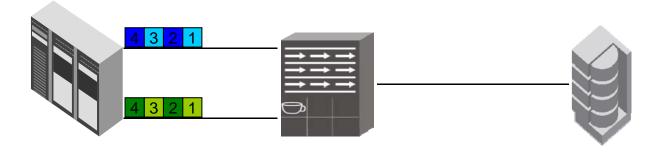








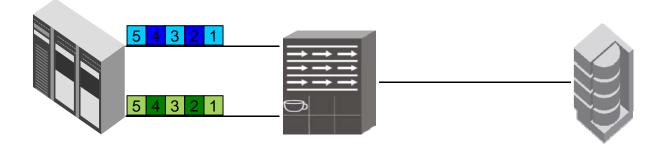






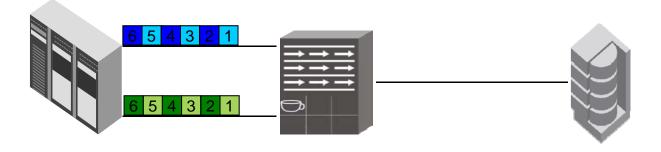






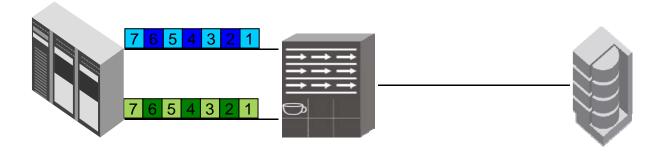






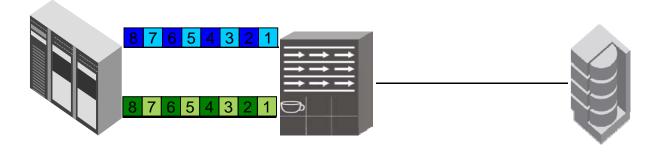






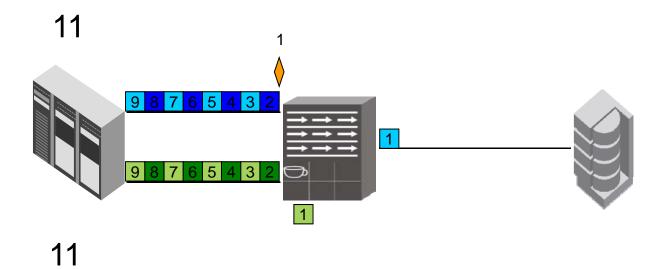






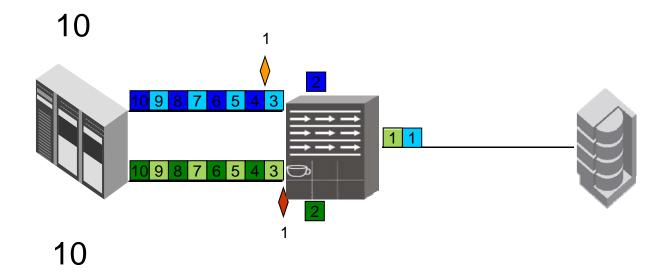






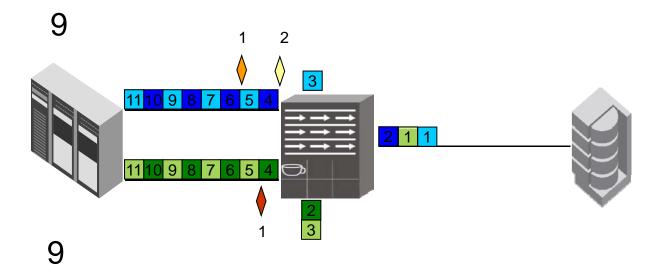






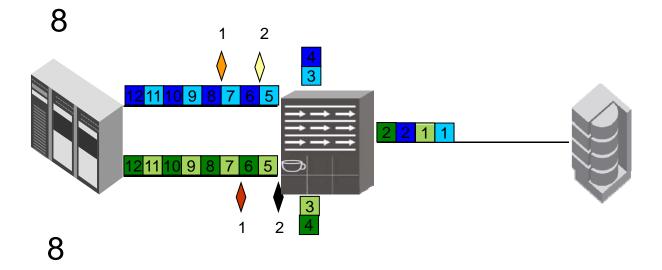






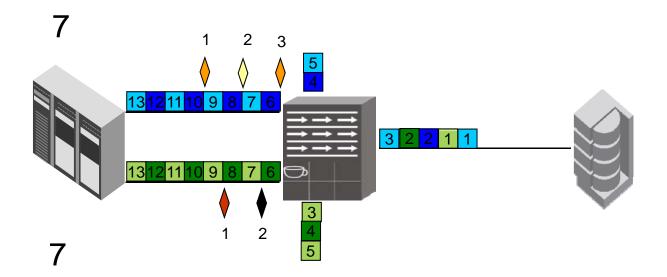






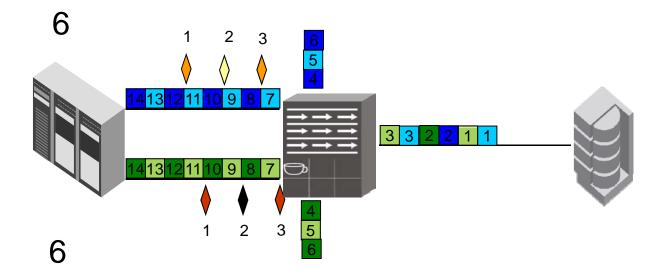






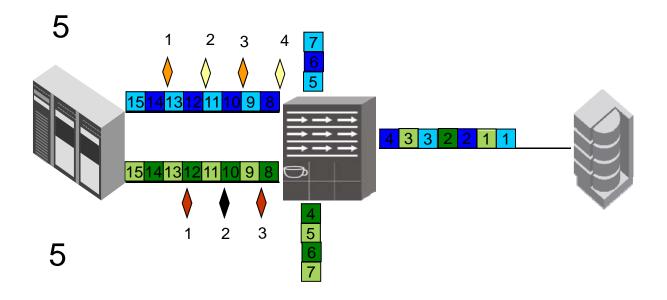






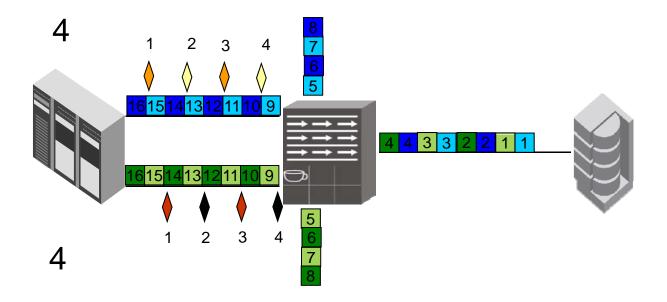






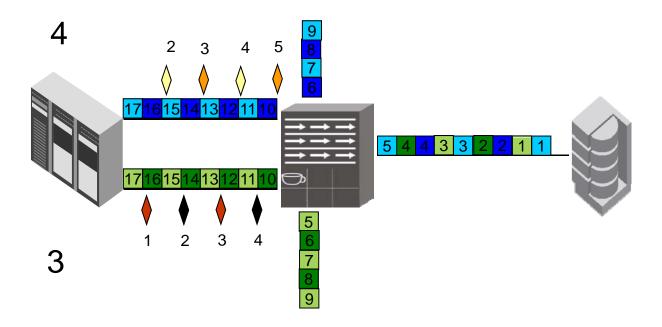






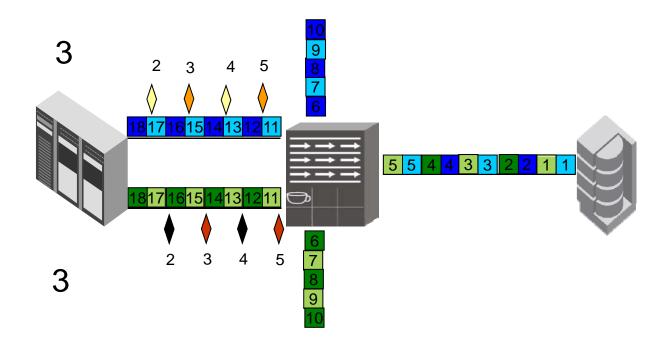






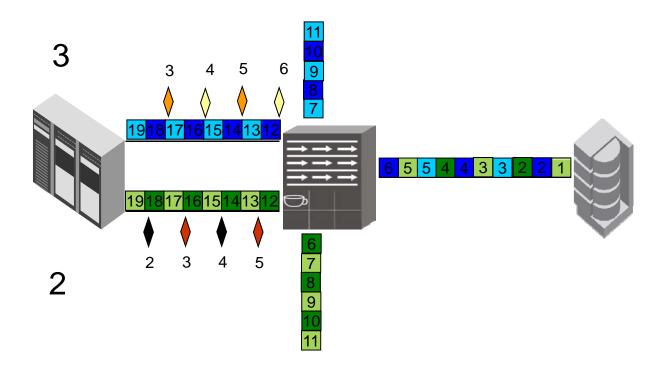






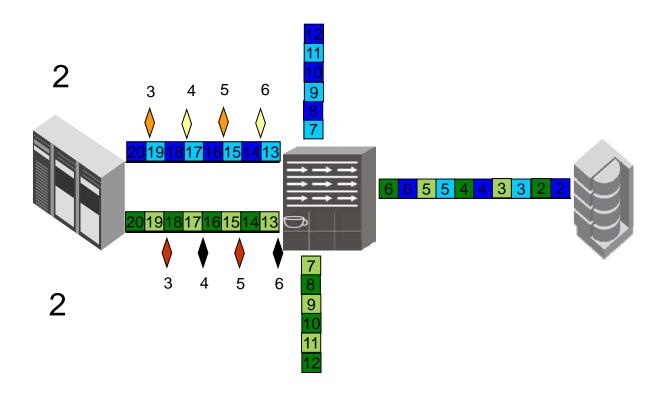






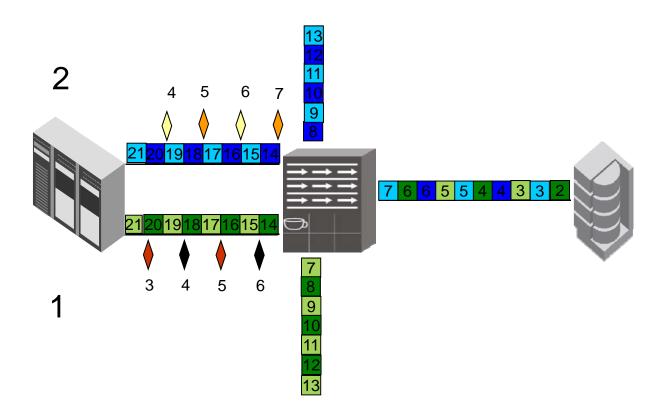






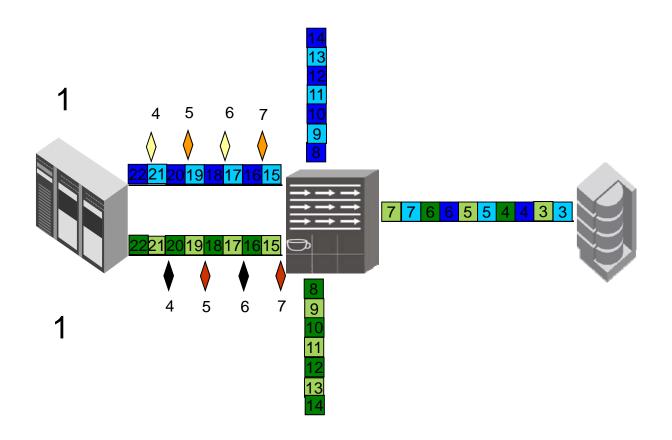






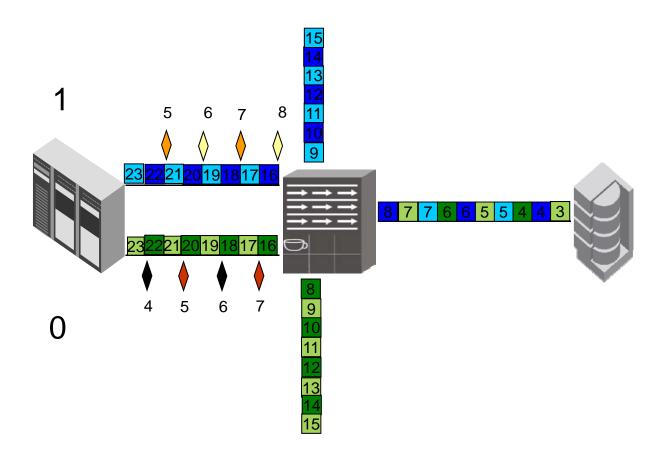






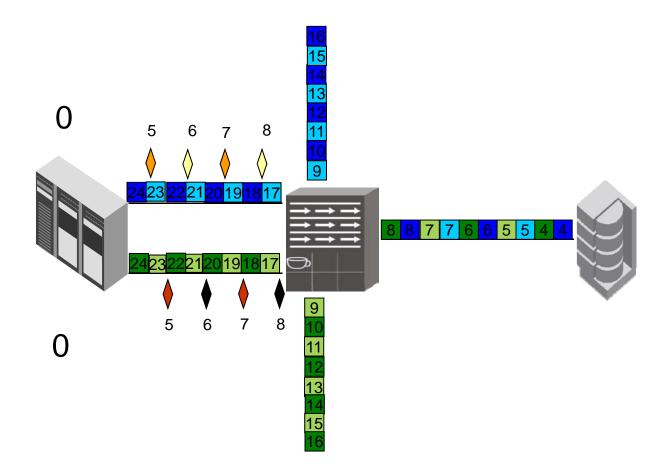






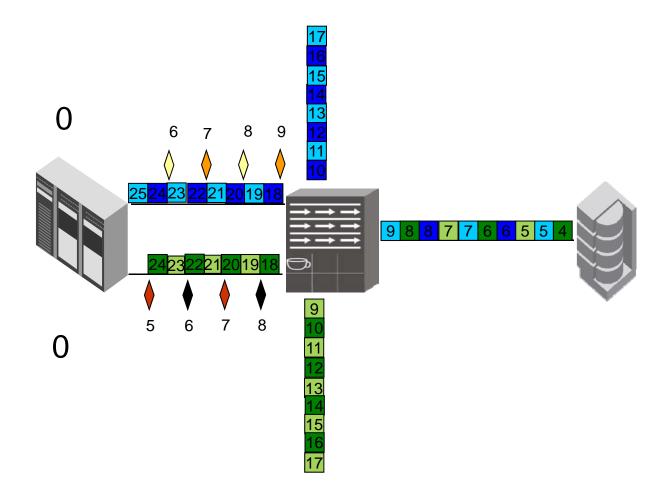






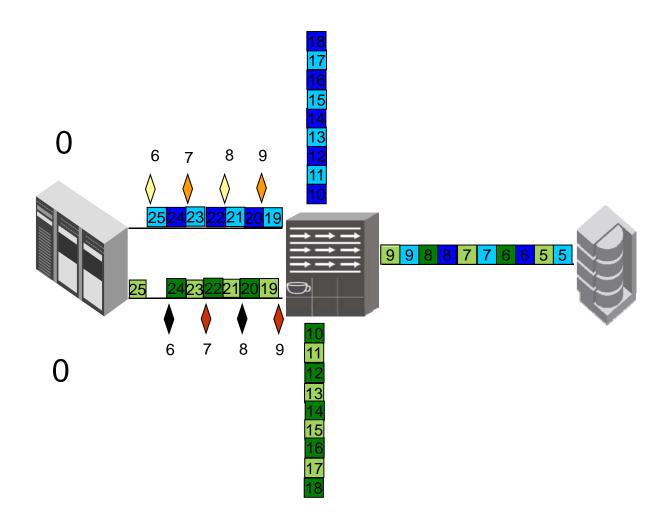






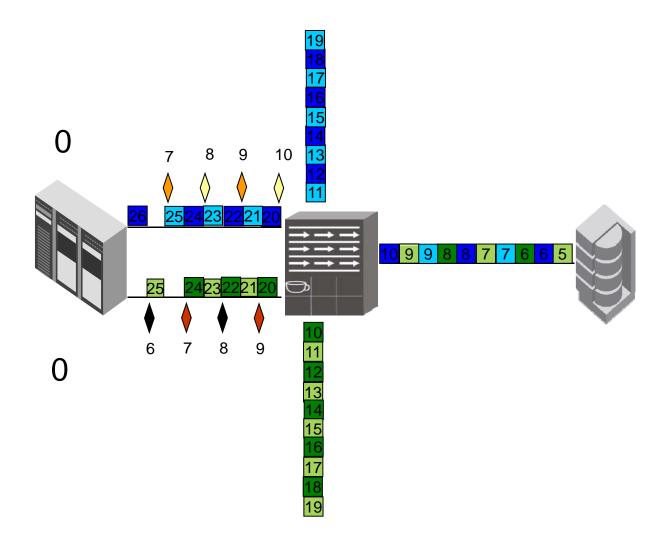






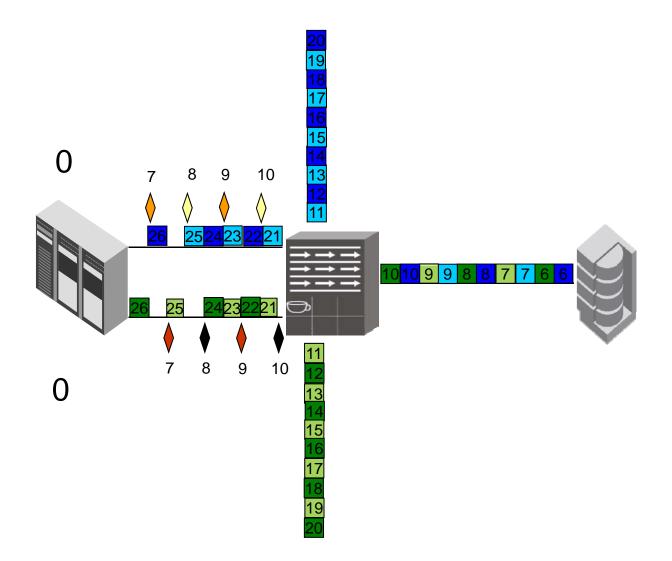






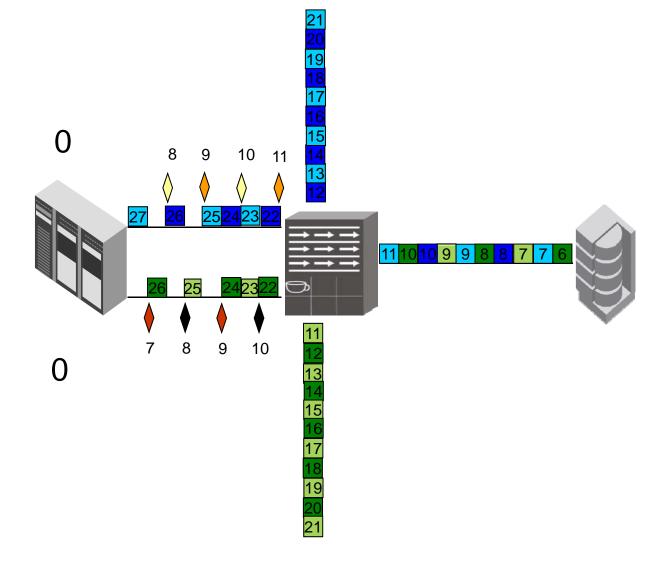




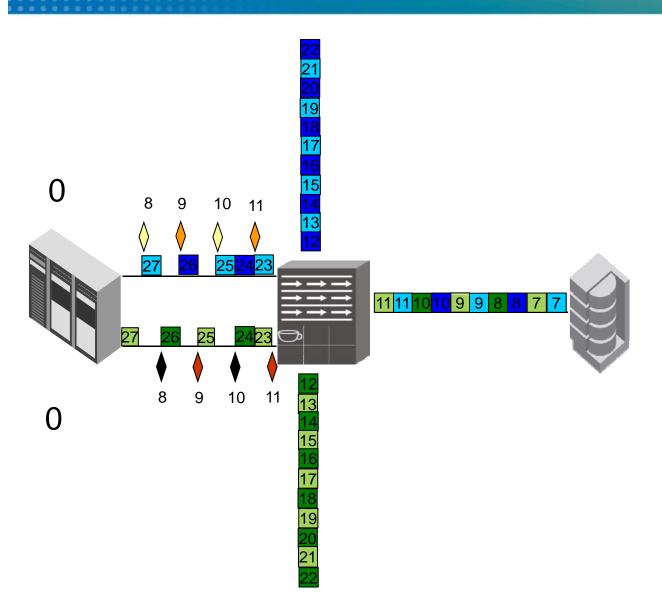






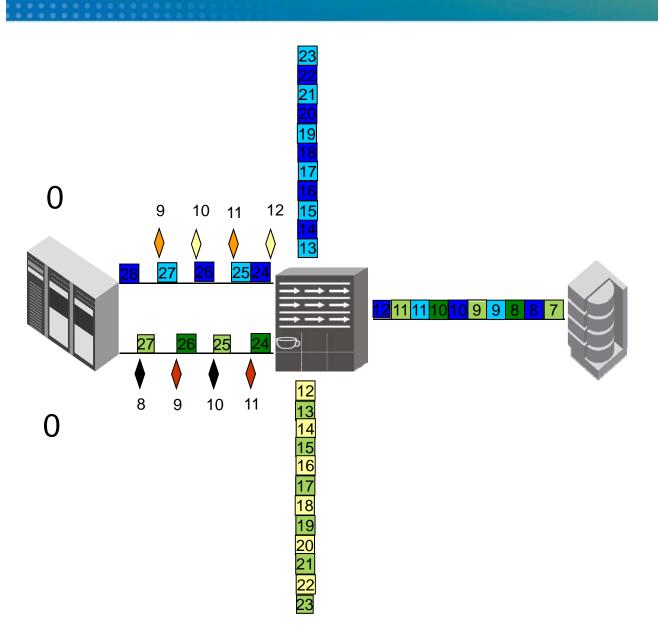






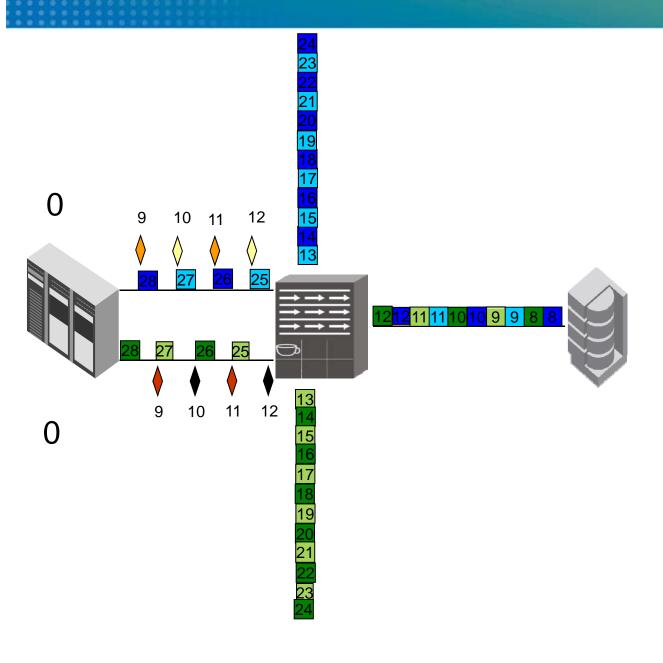






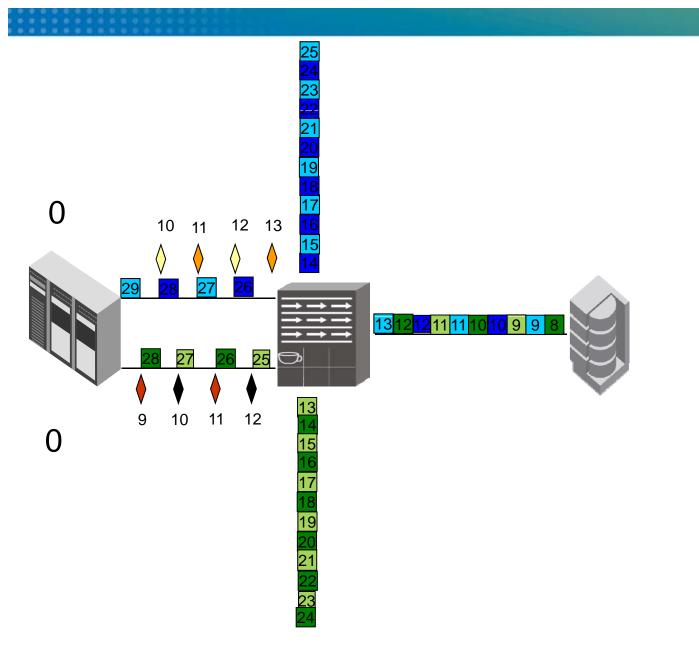






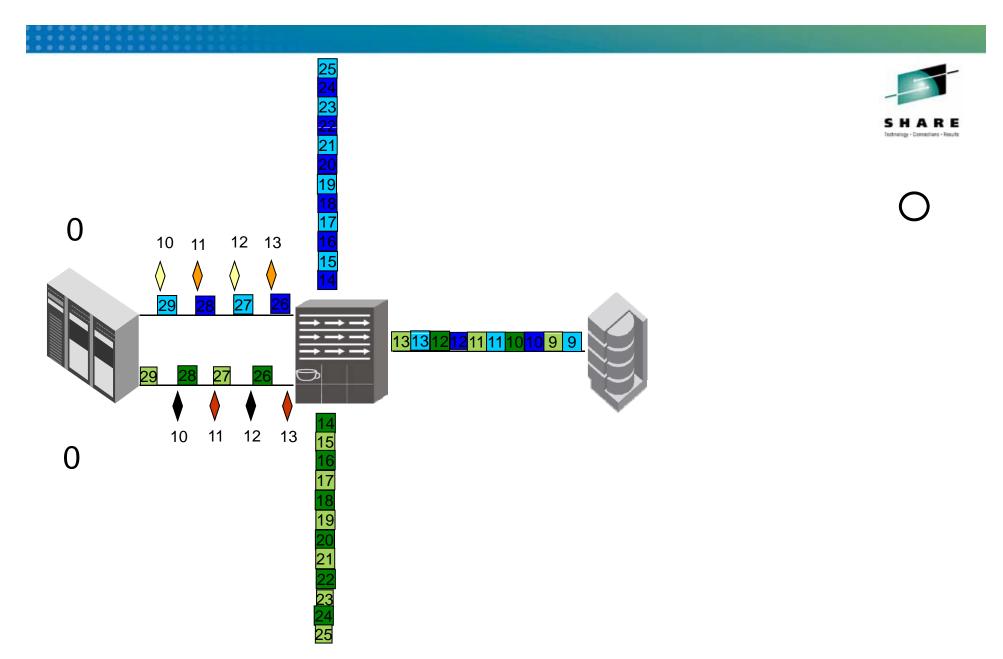




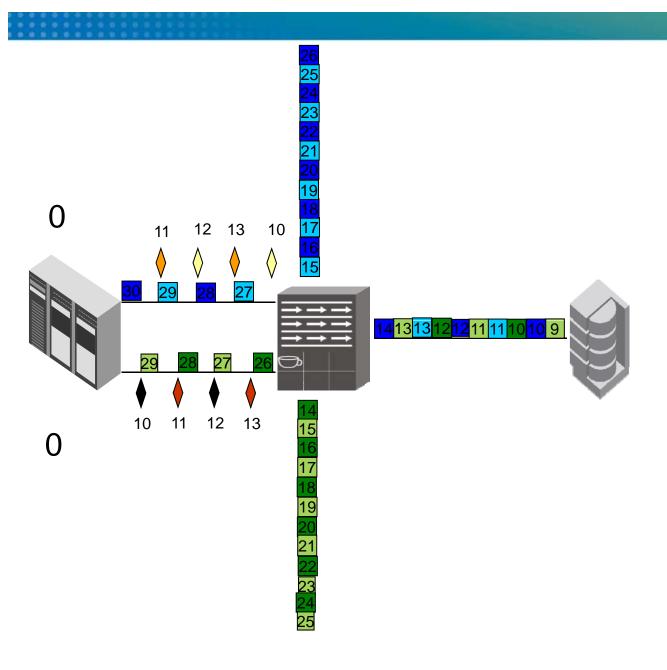


















THIS PAGE INTENTIONALLY LEFT BLANK





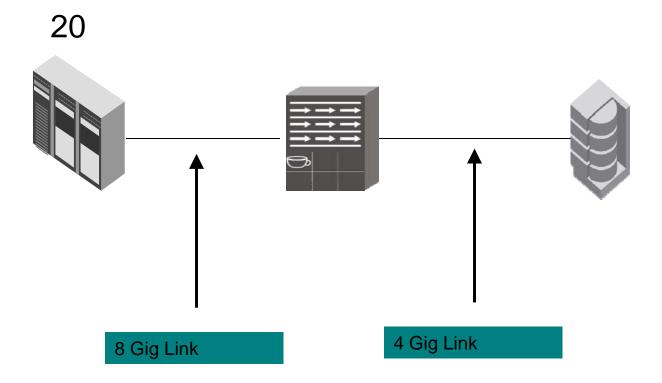
Example: Different sized pipes

BUFFER CREDITS



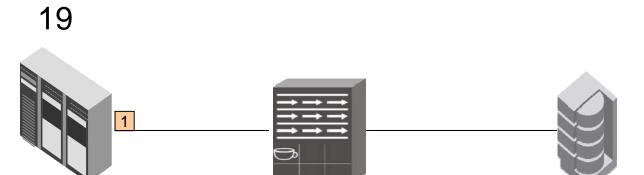


Fat Pipe / Skinny Pipe





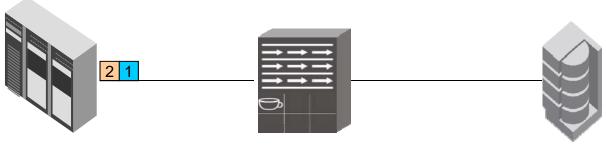






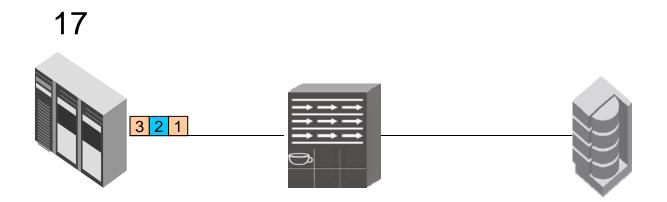








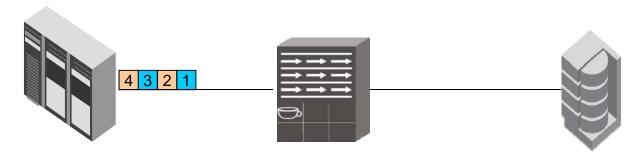






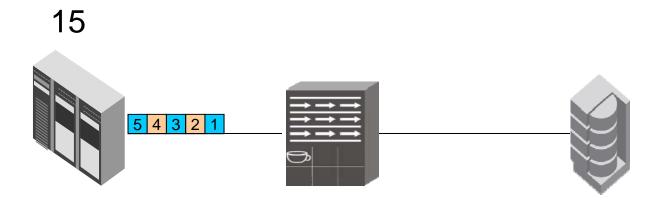






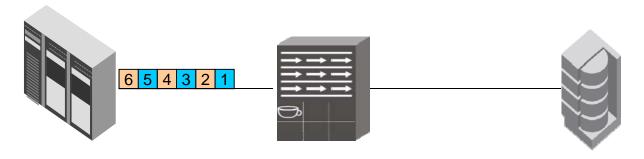








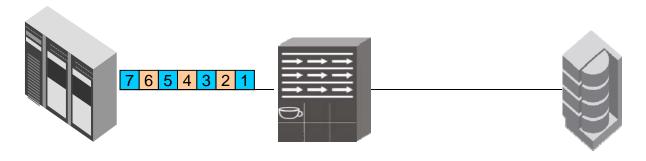






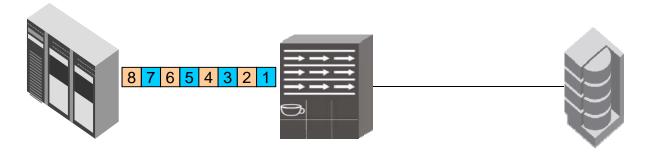






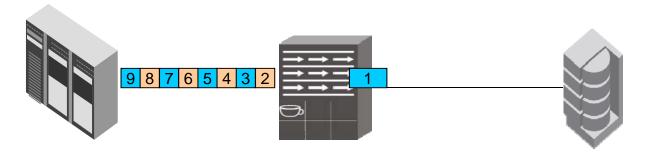






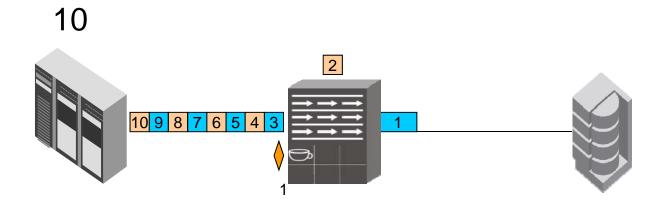






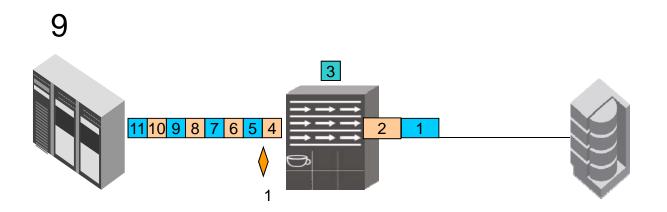






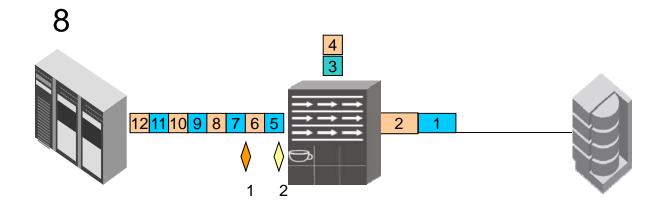






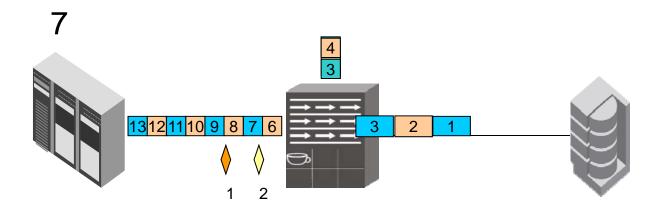






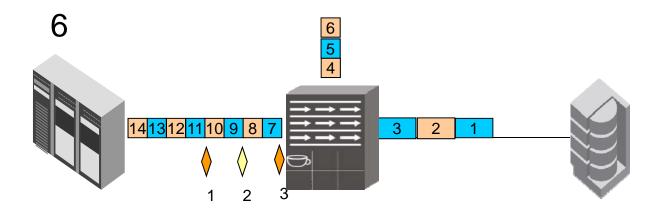






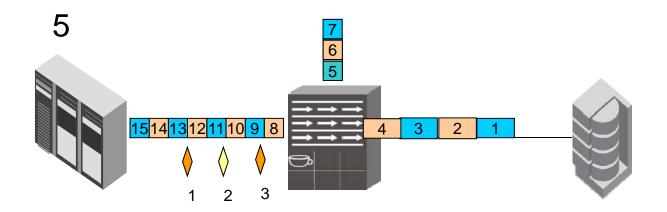






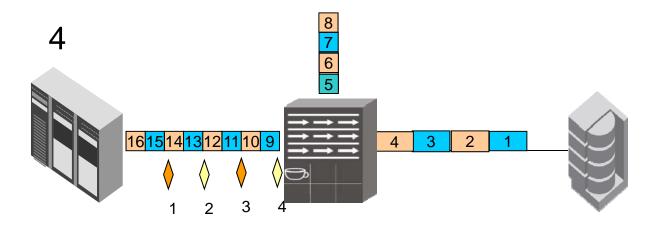






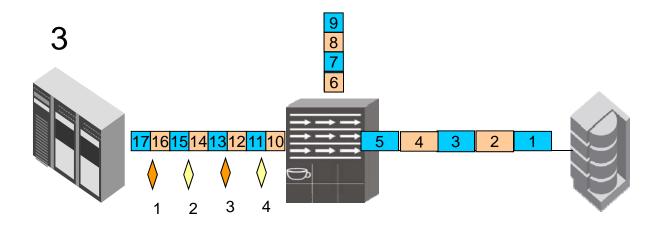






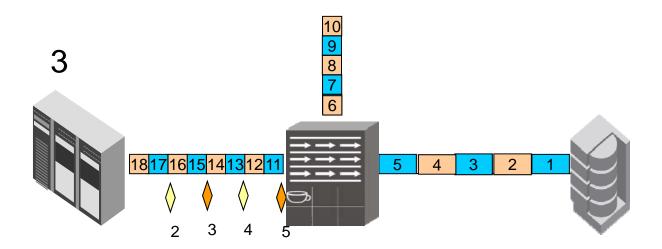






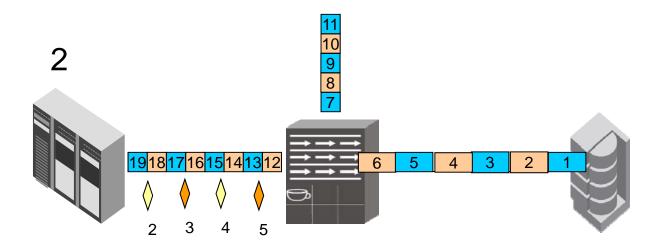






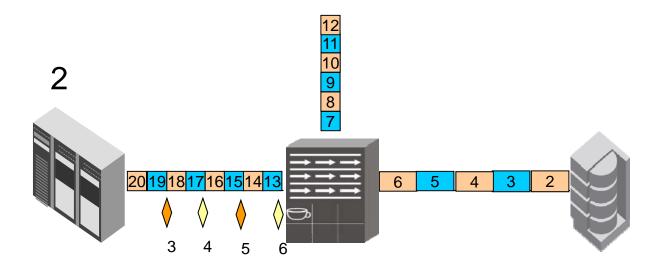






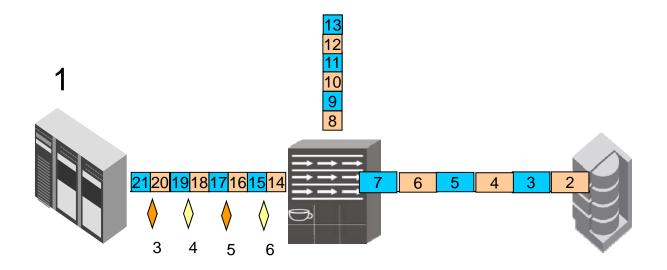






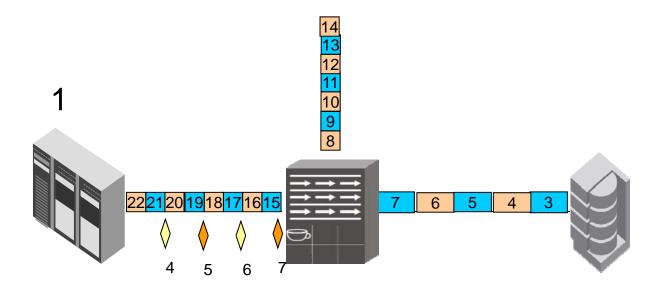






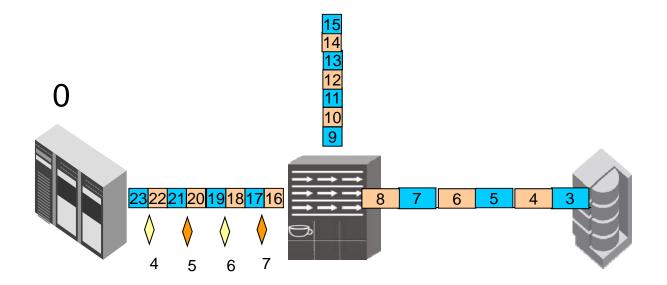






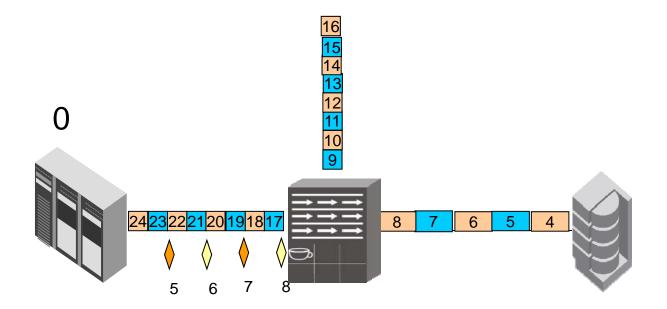






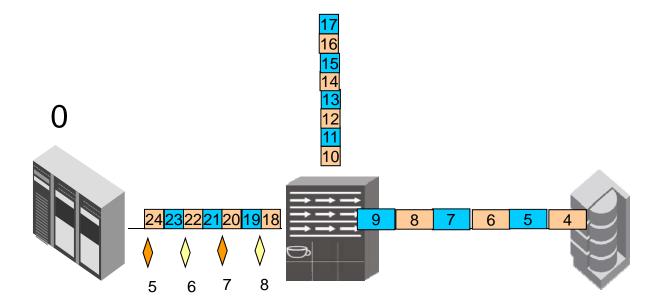






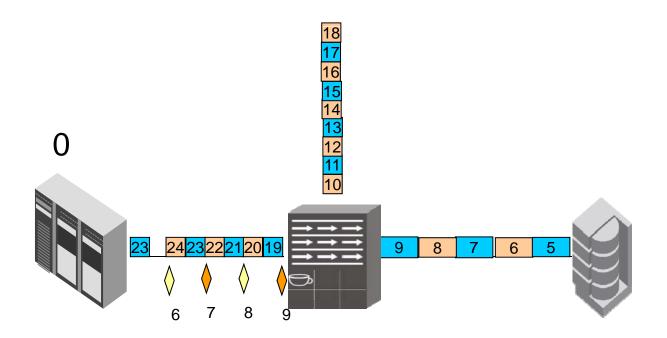






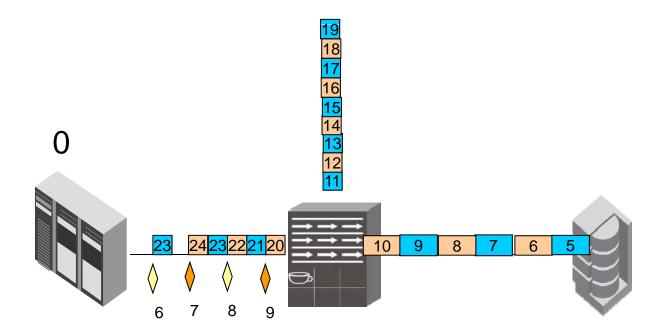






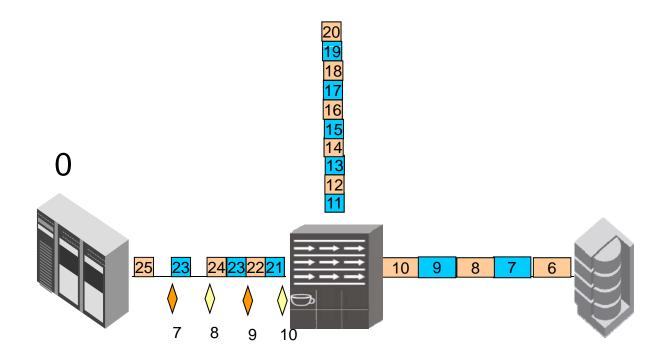






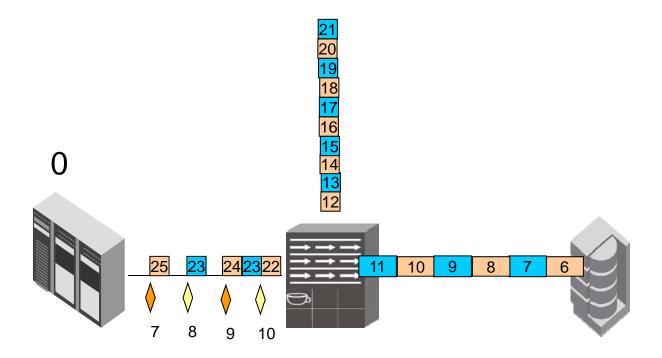




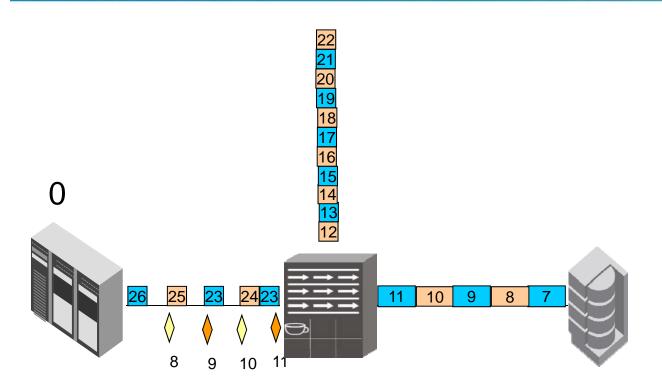






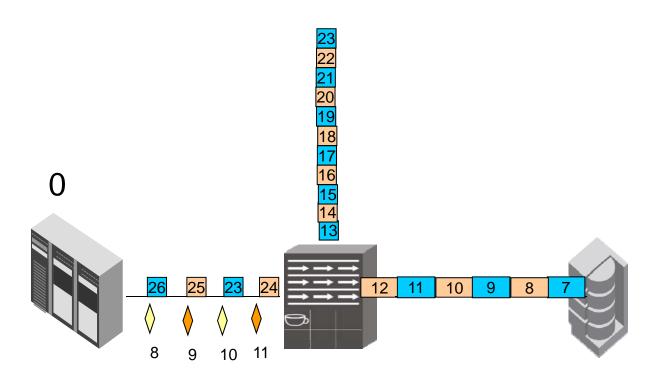






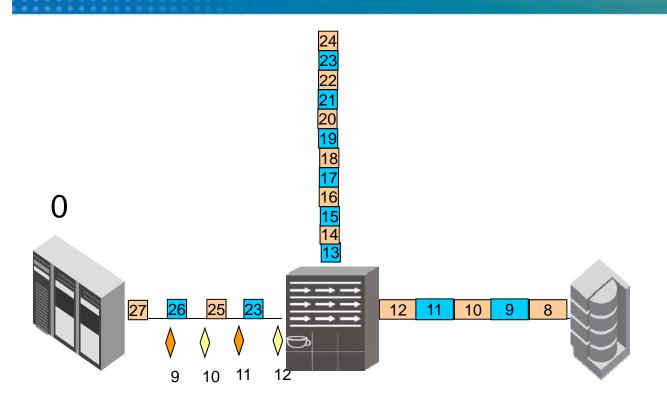






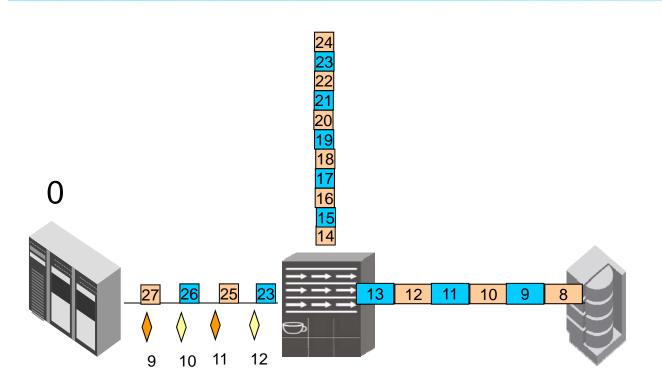






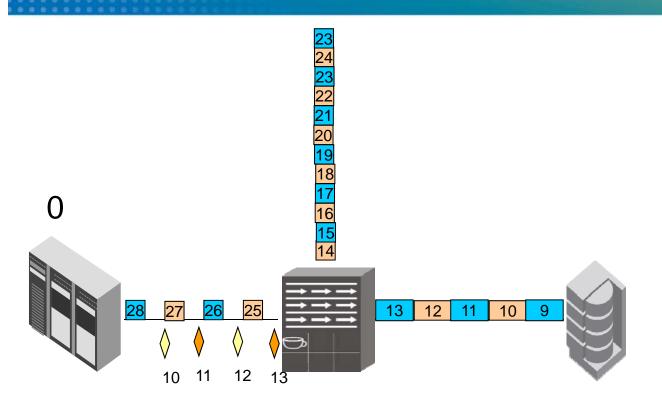






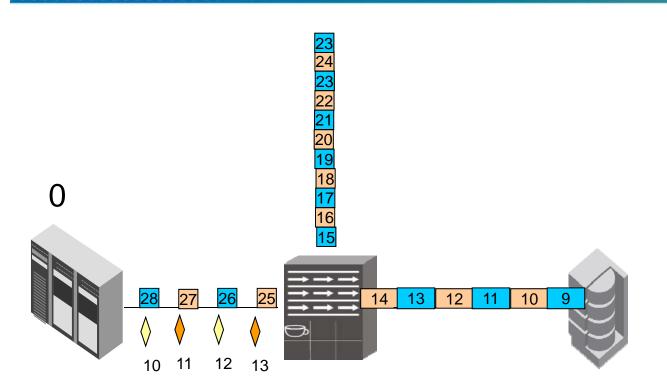








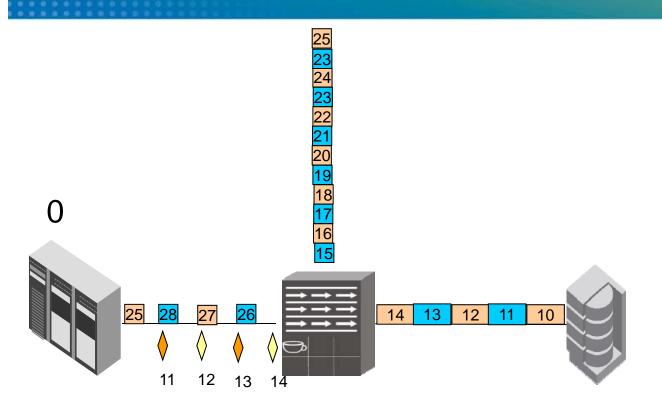


















THIS PAGE INTENTIONALLY LEFT BLANK





Example: Real Life?

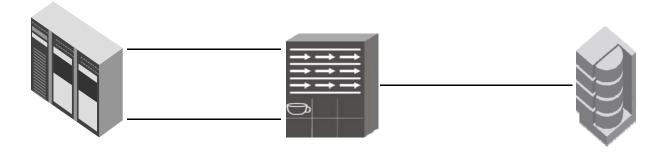
BUFFER CREDITS





More real life example: Two senders sending at 30% - 50% link rate to one receiver

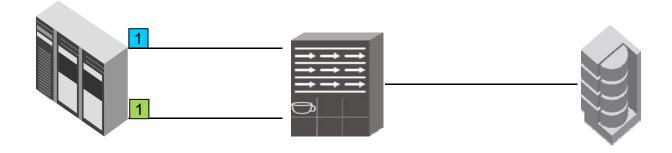






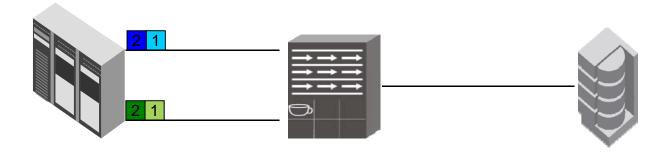








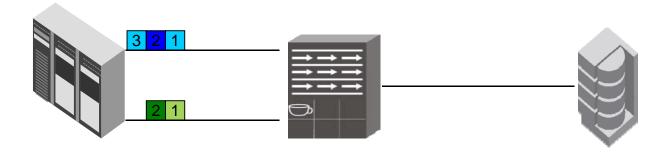






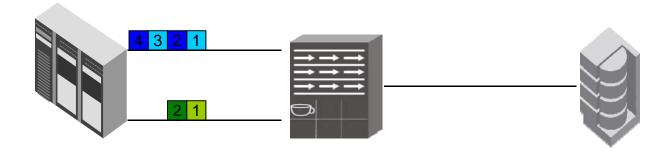






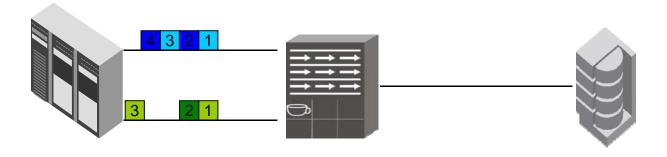






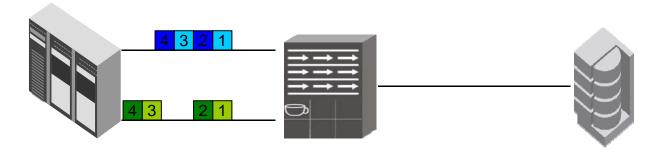






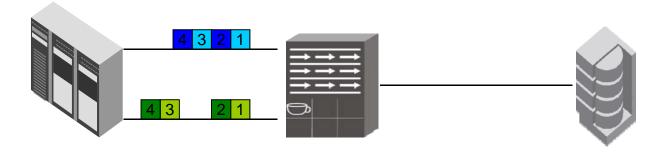






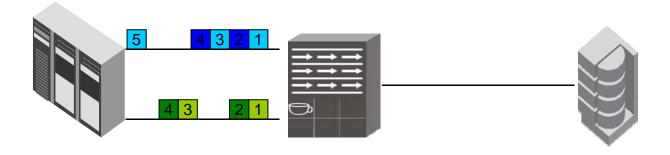






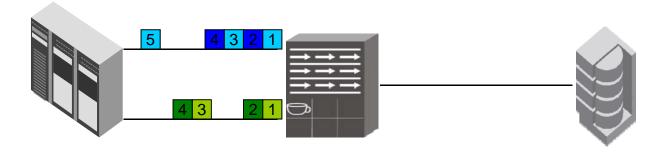






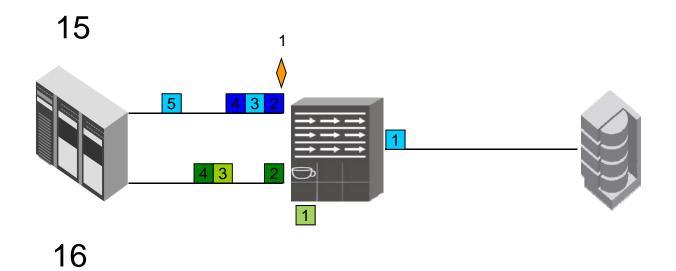






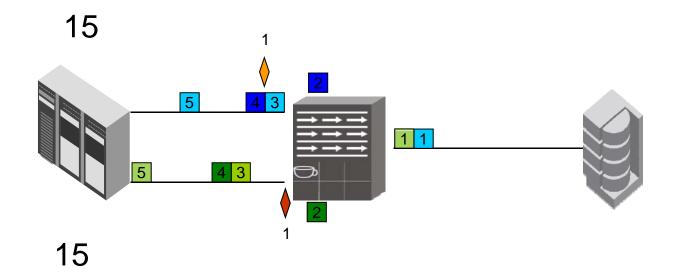






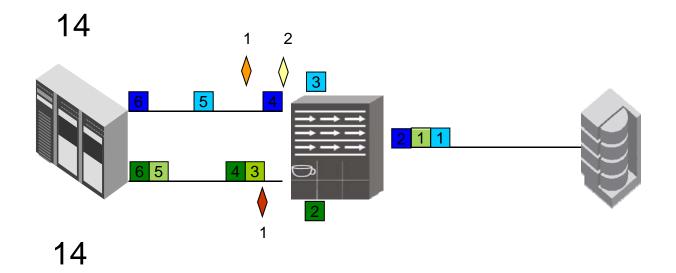






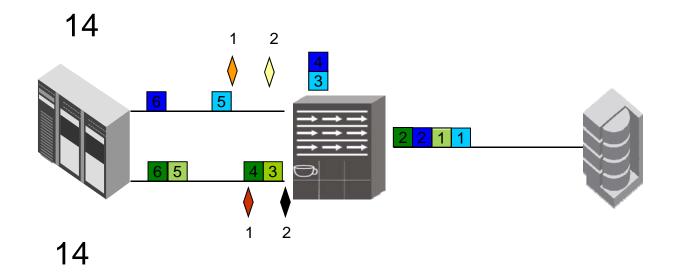






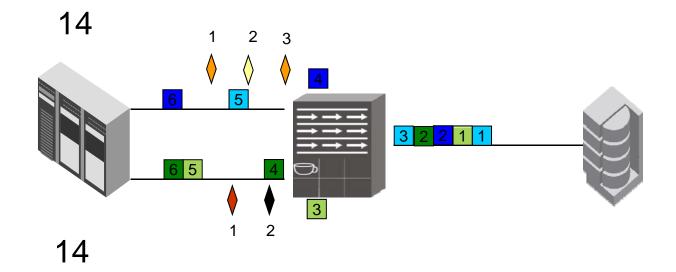






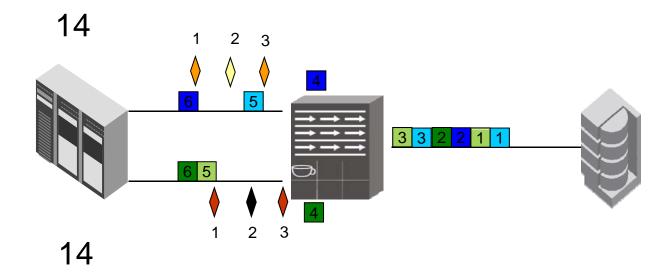






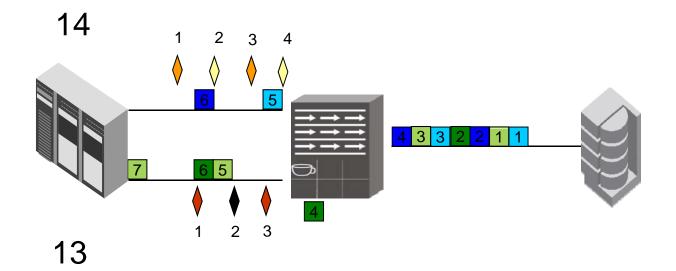






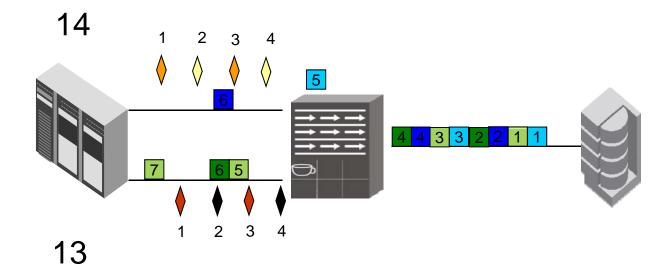






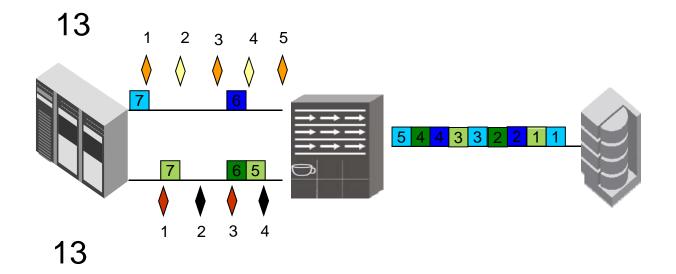






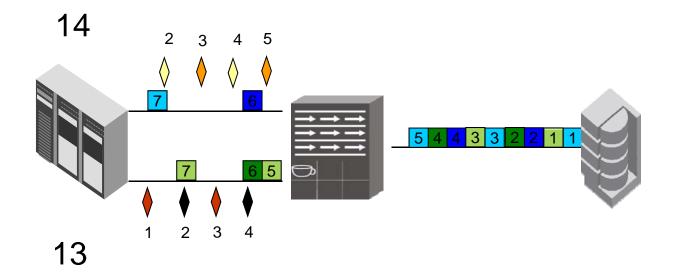






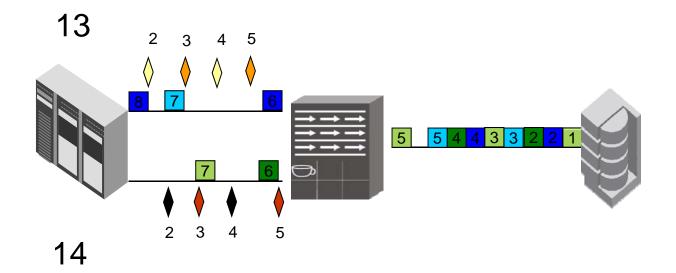






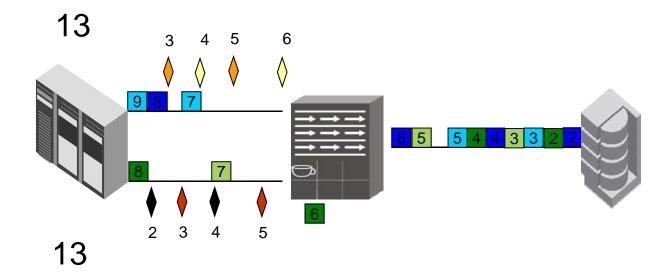






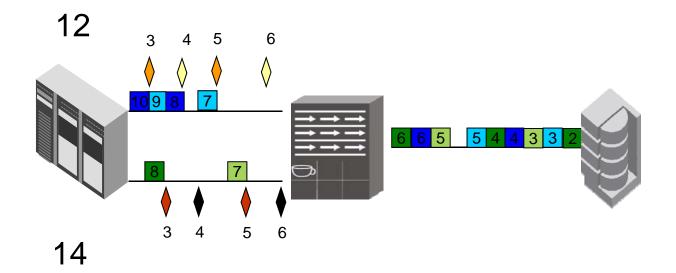






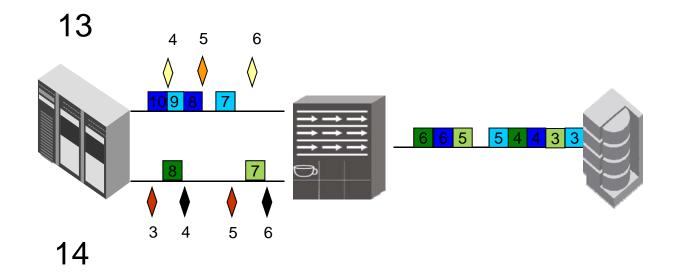






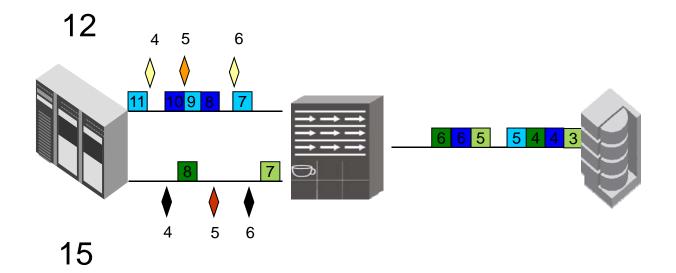






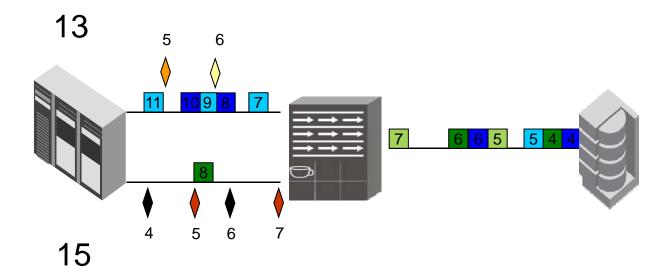






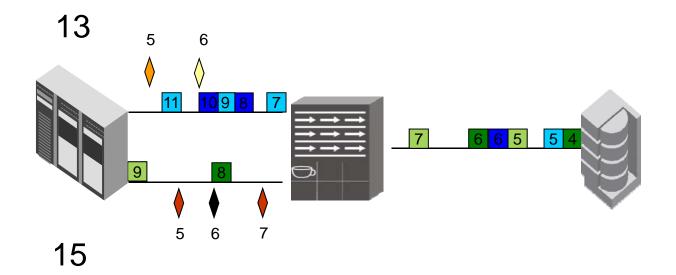






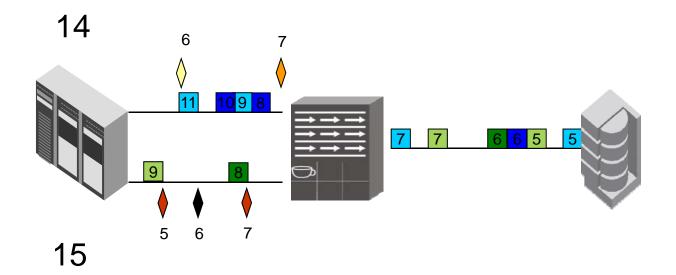






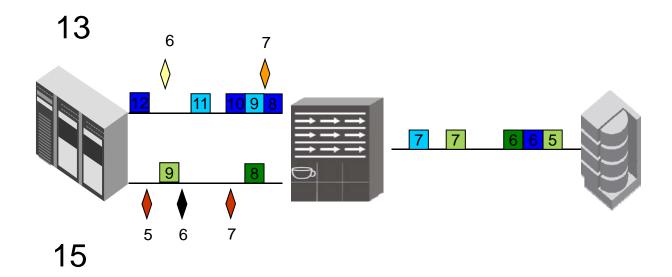






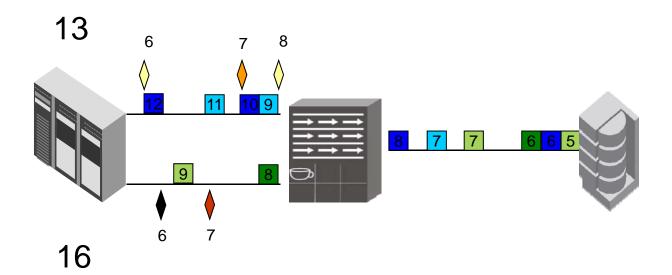






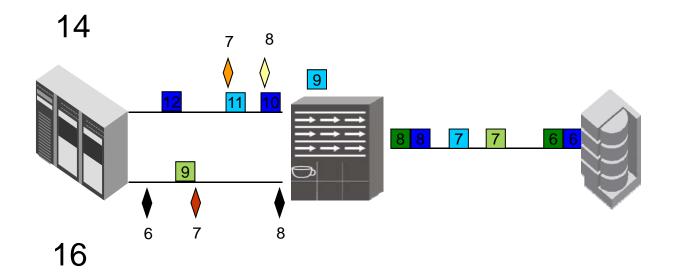
















THIS PAGE INTENTIONALLY LEFT BLANK



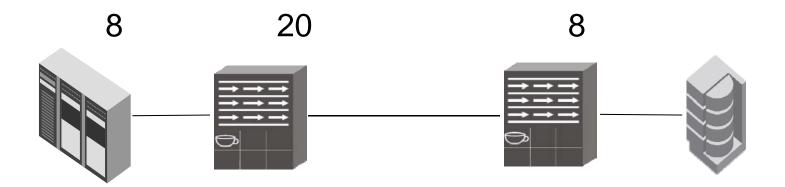


Example: Cascaded Directors

BUFFER CREDITS

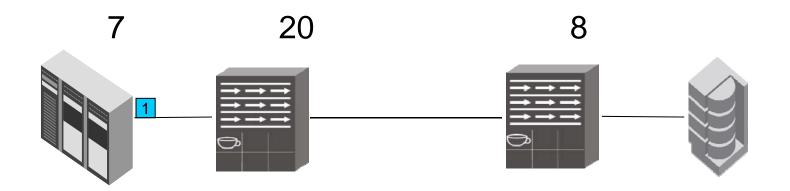






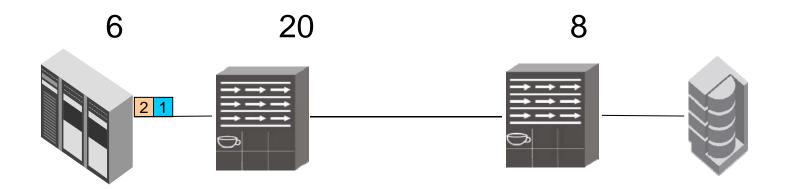






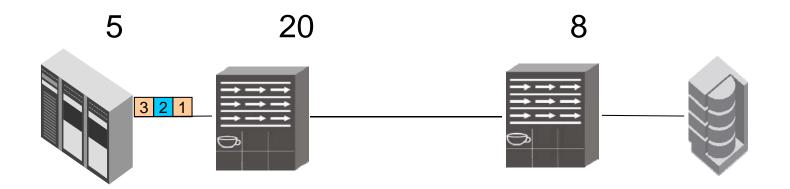






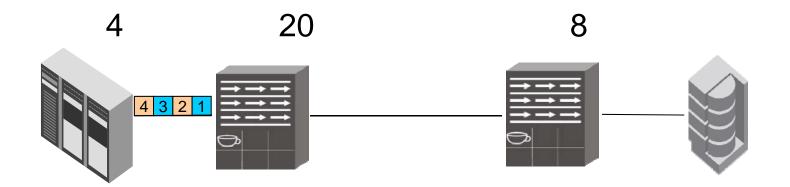






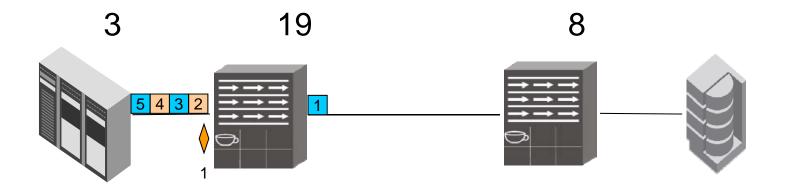






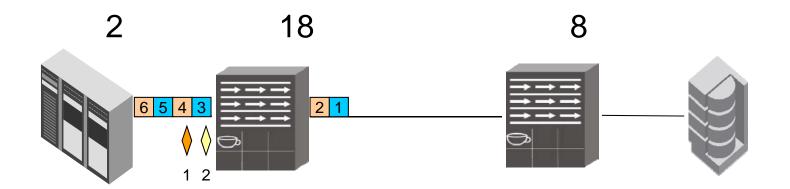






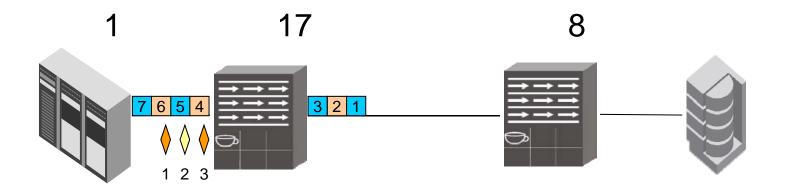






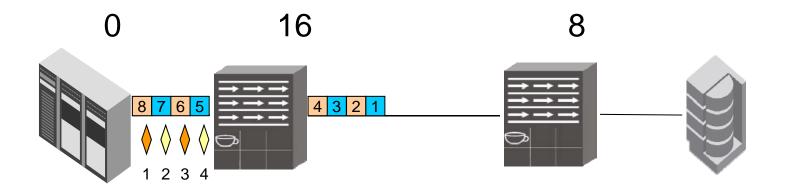






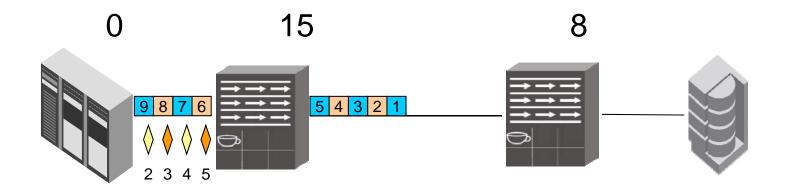






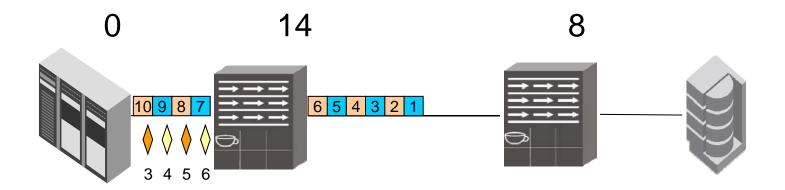






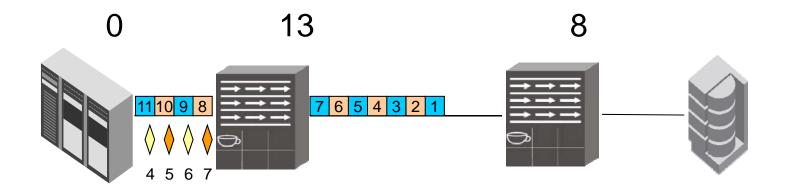






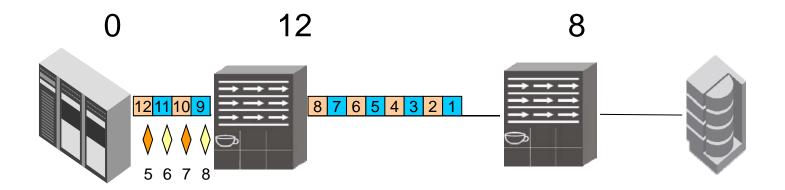






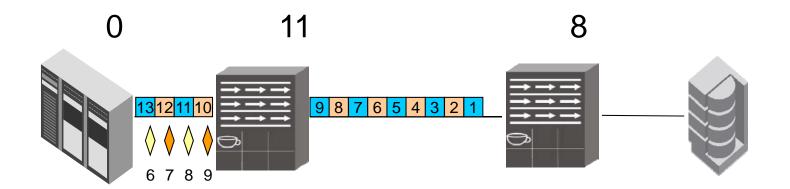






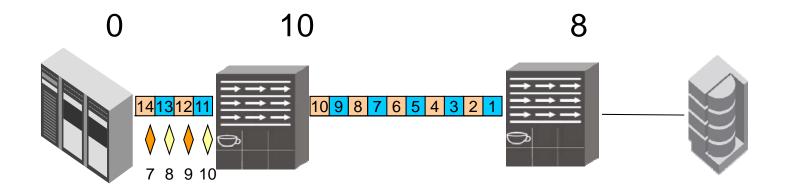






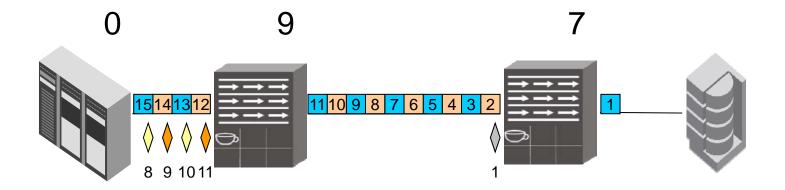






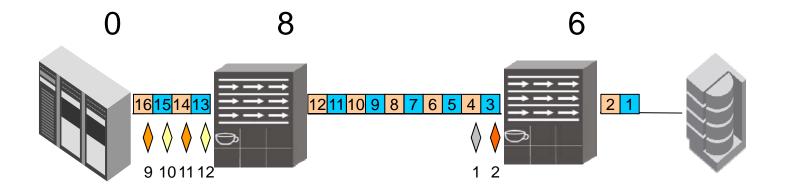






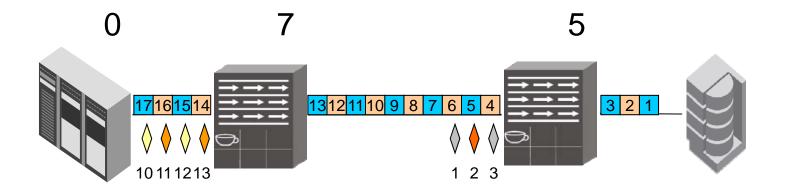






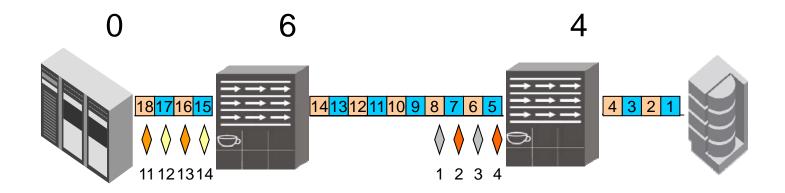






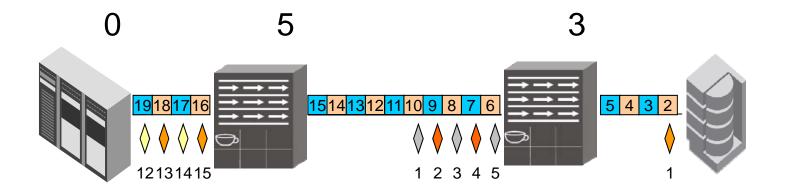






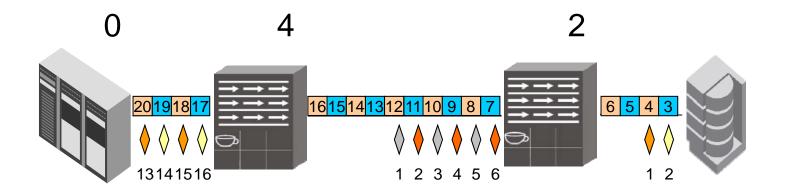






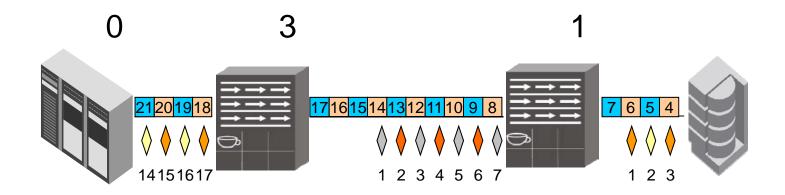






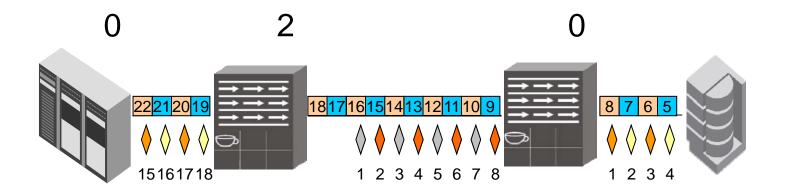






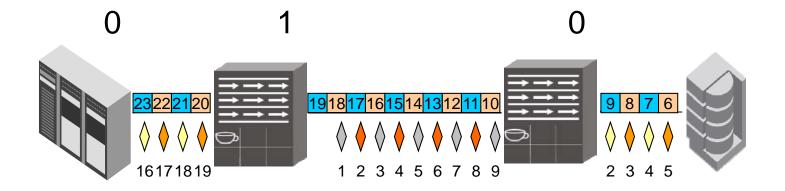








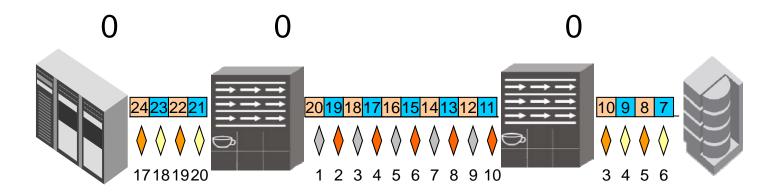






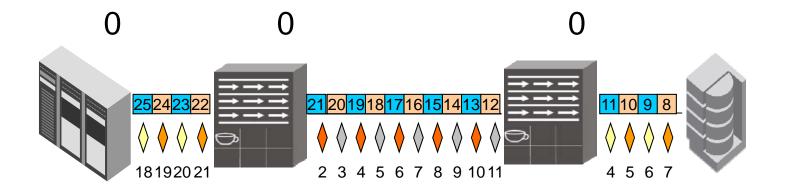
















THIS PAGE INTENTIONALLY LEFT BLANK





Other Neat Stuff

BUFFER CREDITS





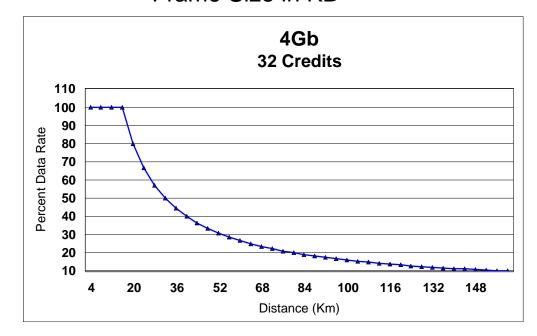
How much credit do I need?

Good "Rule of thumb"

Number of credits needed = 1 + <u>Link speed in Gb/s * Distance in km</u>
Frame Size in KB

Example: 20 km at 1 Gb/s $1 + 1 \cdot 20 = 11$

Example: 10 km at 4 Gb/s 1 + 4 * 10 = 21







How "long" is a frame?

- Traveling at the speed of light, a frame can be very long
 - At 1G, the length of a frame is about 4-kilometers.
 - At 2G, the length of a frame is about 2-kilometers.
 - At 4G, the length of a frame is about 1-kilometer.
 - At 8G, the length of a frame is about 500-meters.
 - At 16G, the length of a frame is about 200-meters.





How "fast" is a frame?

- Speed of light in fibre
 - 200,000 km/second
 - 5-microseconds/km
- Transmission Rates in fibre
 - At 1G, a frame is sent in about 20-microseconds.
 - At 2G, a frame is sent in about 10-microseconds.
 - At 4G, a frame is sent in about 5-microseconds.
 - At 8G, a frame is sent in about 2.5-microseconds.
 - At 16G, a frame is sent in about 1-microsecond.





EXCHANGES





The parts of a transmission

Frame

- Building block of an Fibre Channel connection
- Contains the information to be transmitted
- The address of the source and destination
- Control information

Sequence

- A set of one or more related Frames
- Transmitted unidirectionally from one port to an other

Exchange

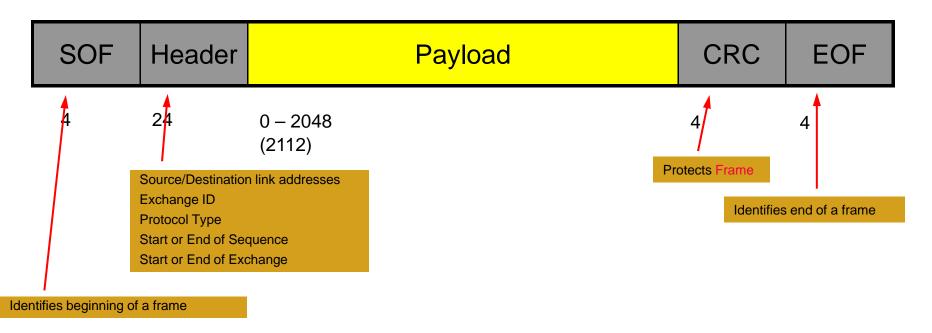
- An Exchange is one or more nonconcurrent sequences
- A single operation
- May be unidirectional or bidirectional





Fibre Channel Frame

The basic building block is the FRAME







Ficon IU Examples

1 Frame IU to transfer a Read CCW

SOF Header Ficon CRC	EOF
----------------------	-----

3 Frame IU to transfer 4K of data

SOF	Header	Ficon Header	CCW Data (2016 bytes)	CRC	EOF
SOF	Header		CCW Data (2048 bytes)	CRC	EOF
SOF	Header	CCW Data (32 bytes)	Ficon CRC EOF		



Urban Legend: FICON uses fewer Exchanges Than FCP



- 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
- 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"





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





How many Exchanges do I need?

- 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 time¹ needs 2 Active Exchanges (pairs) at any given time



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



ERROR SENSITIVITY



Urban Legend: FICON is more sensitive than FCP



- Is Ficon More Sensitive to Errors than FCP?
 - Reasons for Link Errors are the same
- Is a Ficon frame more likely to get lost, damaged or corrupted than FCP?
 - The probability is the same
 - Frames are frames
- When a Ficon frame gets lost, damaged or corrupted, is the recovery action different from FCP?
 - Both protocols retry when errors occur
 - FCP by the Device Driver
 - Ficon by IOS/ERP





So What are the Differences?

- z Operating Systems tend to provide more detailed messages
- Ficon does provide additional debug data and actions
 - RNID
 - Link Error Status Blocks
 - Extensive State Change Processing

Table 89 - Link Error Status Block format for RLS command

Bits Word	31	00
0	Link Failure Count	
1	Loss-of-Synchronization Count	
2	Loss-of-Signal Count	
3	Primitive Sequence Protocol Error	
4	Invalid Transmission Word	
5	Invalid CRC Count	

Source: FC-FS-3 INCITS/T11 Draft Standard v0.92

See www.t11.org





Thank you

SUMMARY





Summary

- Buffer Credits
 - Distance
 - Flow Control
- Exchanges
 - Unidirectional
 - Bidirectional
- Error Sensitivity
 - Recovery
 - Reporting





SHARE, Atlanta, March 2012

Buffer-to-Buffer Credits, Exchanges, and Urban Legends Session 11021

THANK YOU!





REFERENCES





Speaker Biography

- Howard L. Johnson
 - BROCADE
 - Technology Architect, FICON
 - 27 years technical development and management
- Contact Information
 - howard.johnson@brocade.com





Speaker Biography

- Lou Ricci
 - IBM
 - 32-years
 - 24-years in channel development
 - An inventor of FICON
 - FICON Firmware Team Leader
- Contact Information
 - <u>Iricci@us.ibm.com</u>





Speaker Biography

- Patty Driever
 - IBM
 - System z I/O and Networking Technologist
- Contact Information
 - pgd@us.ibm.com





BONUS SLIDES





Speed of Light

- 299 792 458 meters per second
 - In a vacuum
 - National Institute of Standards and Technology
 - http://www.nist.gov/pml/wmd/metric/length.cfm
- 300 000 000 meters per second
 - Common approximation
 - Wikipedia
 - http://en.wikipedia.org/wiki/Speed_of_light
- 1.44 1.46
 - Refractive index of 1300 nanometer fiber
 - Encyclopedia of Laser Physics and Technology
 - http://www.rp-photonics.com/refractive_index.html
- 1.5
 - Common approximation
 - Wikipedia
 - http://en.wikipedia.org/wiki/Optical_fiber#Materials
- 200 000 000 meters per second
 - Calculated speed of light in single-mode fiber
 - V = C/N
 - 200 000 000 = 300 000 000 / 1.5





Calculating Buffer Credits

- Axioms
 - Fibre Channel Frames are held in Buffers
 - Receiver memory is measured in frame buffers
 - · Transmitter counts frames buffers used by counting frames sent
 - Objective is to maintain full link bandwidth
 - Enough Receiver memory is required to receive all of the frames that can be transmitted in the time it takes to return the "Receiver Ready" acknowledgement
 - Includes Frames "in flight"
 - Frames transmitted but not yet received (i.e. frames on the link)
 - Includes time to respond to frames received
 - R_RDY acknowledgement
 - Increments Transmitter frame count
- Definitions
 - Buffers are a function of distance
 - Distance == Rate * Time
 - D == RT
 - Buffers == f(Distance)
 - B == f(D)
 - B == f(RT)
 - · Rate is the Average Frame Transmission Rate
 - R
 - Time is a function of the speed of light in a fiber and distance
 - Time == Distance / Rate
 - Rate is the speed of light in a fiber
 - T == D / 200,000,000 meters/second
 - B == f(R * (D / 2 * 108 meters/second))





Average Frame Transmission Rate

- Frame Transmission Rate
 - Number of frames transmitted per second
 - Next frame is transmitted as soon as the previous frame completes
 - While transmission count is non-zero
 - Number of frames transmitted depends on the amount of data per frame
- Average Frame Transmission Rate
 - Effective Data Transmission Rate / Average Frame Density
 - R == eX / aF
- Effective Data Transmission Rate
 - Scale * Baud Rate * Encoding Factor
 - eX == S * bR * E
 - eX == S * (1.0625 * 10⁸ bytes/second)
- Average Frame Density
 - Frame Header is 36-bytes
 - Frame Payload is 0 2112-bytes
 - aF == 36-bytes + Average Payload bytes
 - aF == 36 + A bytes/frame
- Average Frame Transmission Rate
 - R == eX / aF
 - R == (S * (1.0625 * 10⁸ bytes/second)) / ((36 + A) bytes/frame)
 - R is expressed in frames / second





Effective Data Transmission Rate

- Effective Data Transmission Rate
 - Scale * Baud Rate * Encoding Factor
 - eX == S * bR * E
- Scale
 - · Fibre Channel baud rate scalars
 - 1, 2, 4, 8 Gbs
- Baud Rate
 - Fibre Channel baud rate
 - The baud rate is 1.0625 * 109
 - bR == S * (1.0625 * 109 bits/second)
- Encoding Factor
 - · Fibre Channel encoding factors
 - For 1, 2, 4, and 8 Gbs
 - 8-data bits / 10-transmission bits
 - 8b10b
- Data bits per second
 - eX == S * (1.0625 * 109 bits/second) * 8-data bits/10-transmission bits
 - eX == S * (8.5 * 108 bits/second)
- Data bytes per second
 - eX == S * (8.5 * 108 bits/second) * 1-byte/8-bits
 - eX == S * (1.0625 * 108 bytes/second)
- Arithmetic
 - eX == S * (1.0625 * 10⁹ bits/second) * (8-bits/10-bits) * (1-byte/8-bits)
 - eX == S * (1.0625 * 108 bytes/second)





Buffer Credit Calculation

- Function
 - B == f(R * T)
 - B == R frames/second * T seconds
- Rate
 - Rate == Effective Data Transmission Rate / Average Frame Density
 - R == eX / aF
 - $R == (S * (1.0625 * 10^8 \text{ bytes/second})) / ((36 + A) \text{ bytes/frame})$
- If the latency of the fibre is known
 - B == (S * (1.0625 * 10⁸ bytes/second)) / ((36 + A) bytes/frame) * T seconds
 - B is the number of buffers expressed as frames
 - S is the scale (1, 2, 4, or 8)
 - · A is the average bytes per frame
 - T round-trip time
- If the distance of the fibre is known
 - T == D / 200,000,000 meters/second
 - B == R frames/second * (D / 200,000,000 meters/second)
 - B == (S * (1.0625 * 10⁸ bytes/second)) / ((36 + A) bytes/frame) * (D / (2 * 10⁸ meters/second)))
 - B is the number of buffers expressed as frames
 - S is the scale (1, 2, 4, or 8)
 - A is the average bytes per frame
 - D round-trip distance in meters





End to End Credit

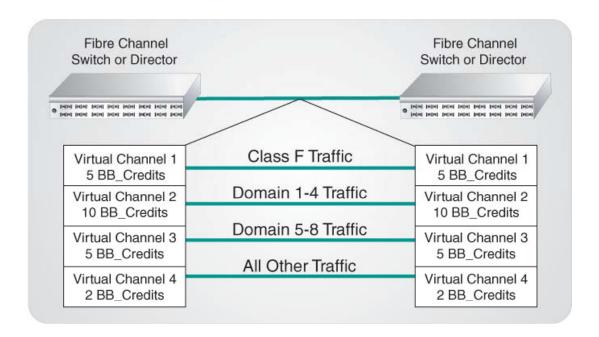
- Device to Device Flow Control
 - Between source and destination
 - Not the links
 - Similar to buffer-to-buffer flow control
 - At N_Port Login
 - Report available receive buffers (EE_Credit)
 - Transmitter counts buffers transmitted (EE_Credit_CNT)
 - Receiver acknowledges frame (ACK)
 - ACK 1 (a single data frame in a sequence) most common
 - ACK n (several (N) consecutive data frames in a sequence)
 - ACK 0 (all data frames in a sequence) not used





Virtual Channels

- Technology to allocate BB_Credits to particular data flows
 - Class F traffic has one data flow
 - Assigned with Zoning by using special Zone names







THIS SLIDE INTENTIONALLY LEFT BLANK

