

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

Patty Driever, IBM
Howard L. Johnson, Brocade

3 August 2010 (9:30am – 10:30am)
Session 7320
Room 104



SHARE in Boston

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 provided 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 fill the pipe?
 - What if you can't fill the pipe?
 - 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?
- 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?

What are they and how do they work?

BUFFER CREDITS

What is Buffer-to-Buffer Credit?

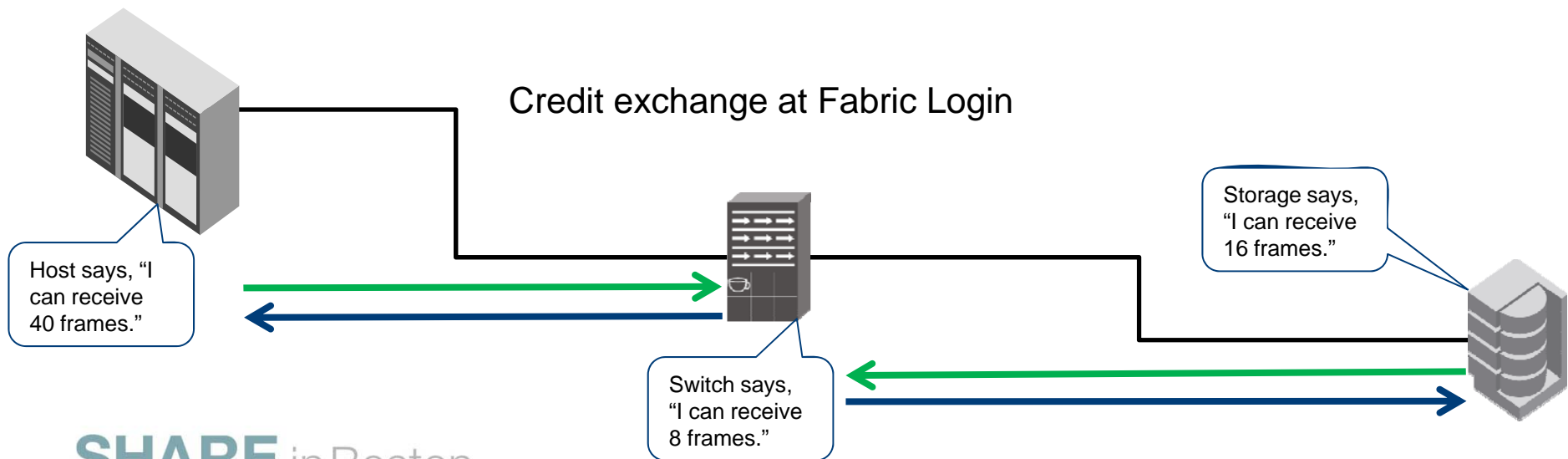
- The greater the BB Credit....
 - A. The faster frames can be sent
 - B. The farther apart the two ports can be
 - C. The larger the frames can be
 - D. None of the above

What is Buffer-to-Buffer Credit?

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

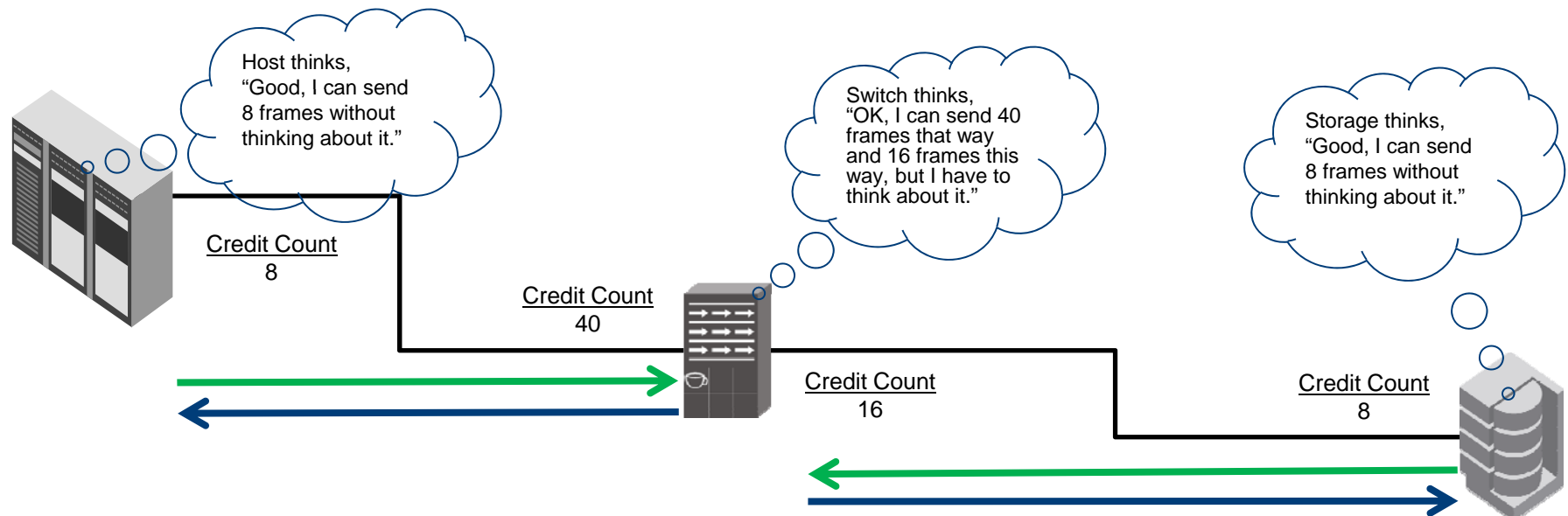
Flow Control

- 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



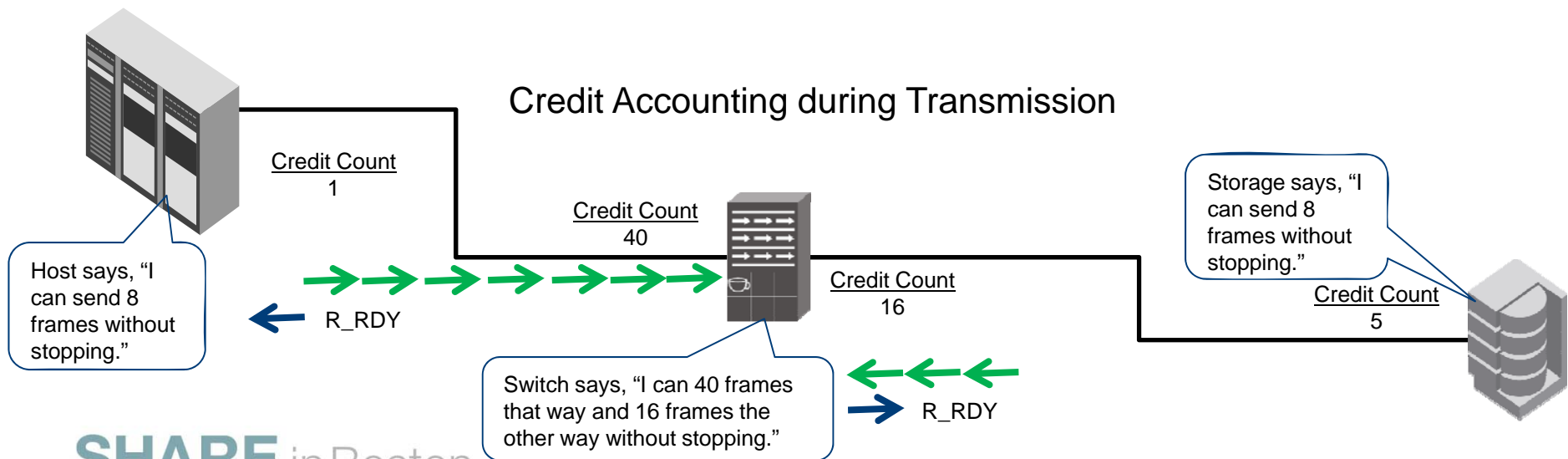
Buffer Credit

- 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



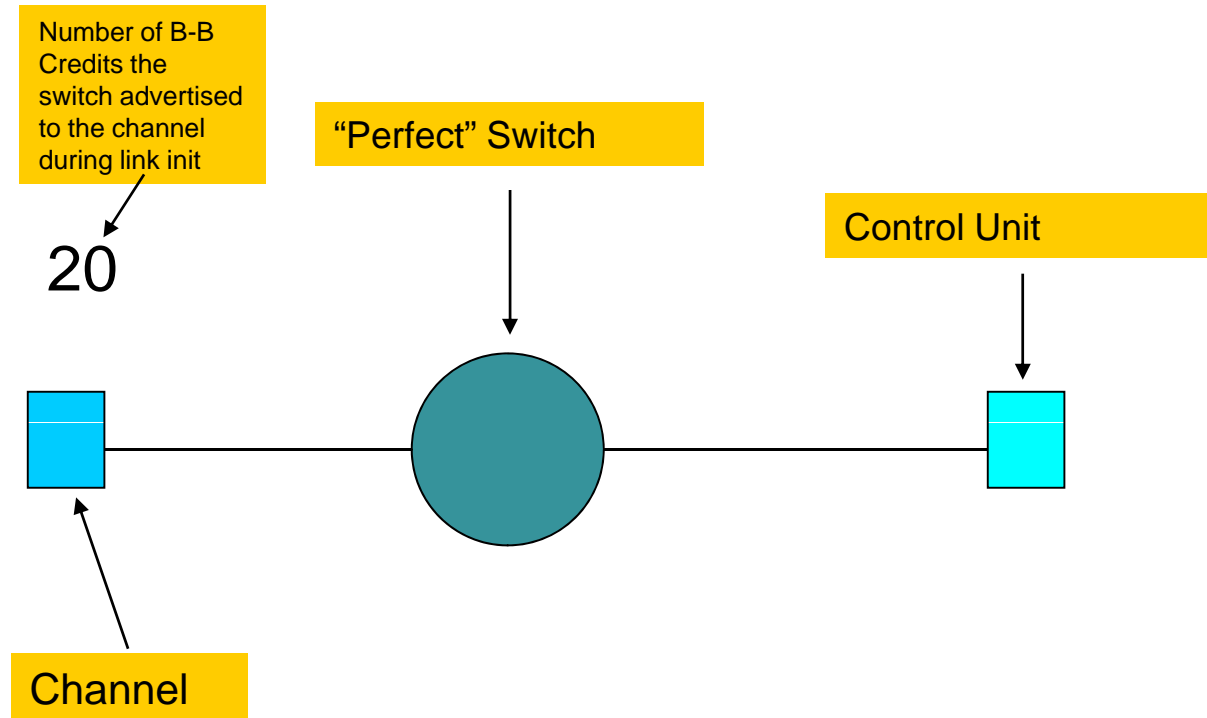
Urban Legend: Buffer Credits at Zero are a Problem

- 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

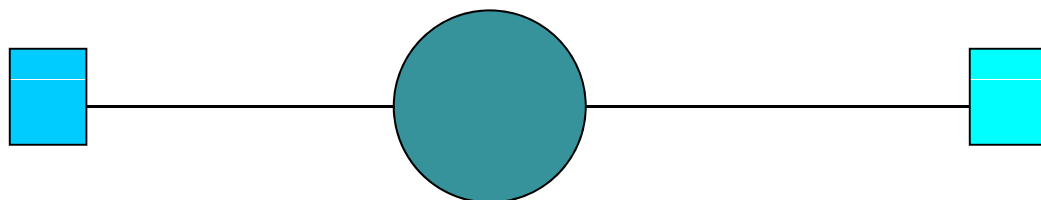
BUFFER CREDITS

Initial Conditions

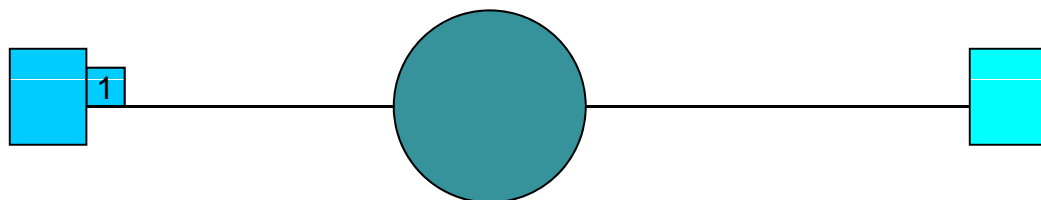


NOTE: In these animations, both the frames and R_RDY's are numbered. This is for illustrative purposes only. In reality, neither the frames nor the R_RDY's are numbered. The arrival of an R_RDY only informs the receiver that **A** frame has been forwarded, now **WHICH** frame has been forwarded.

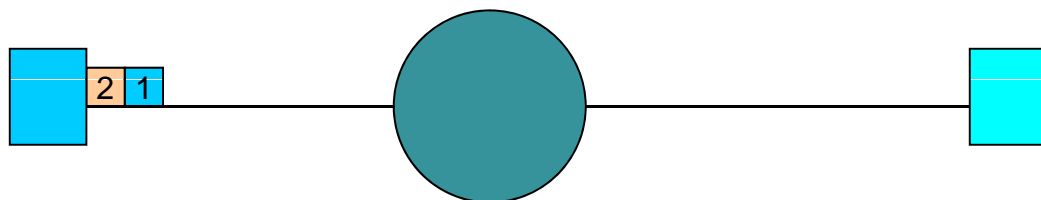
20



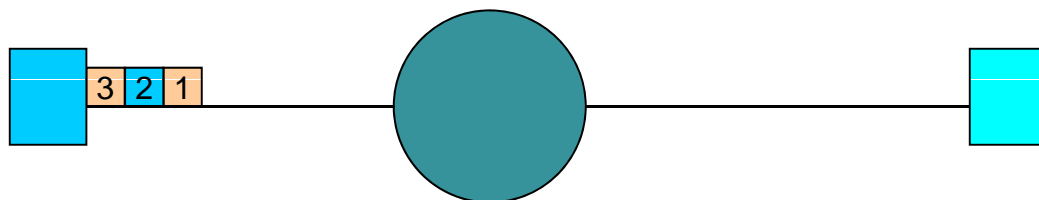
19



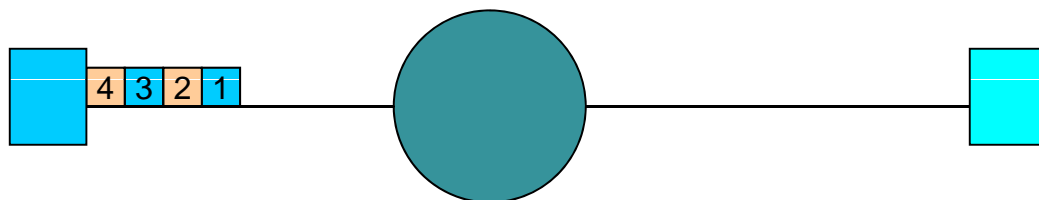
18



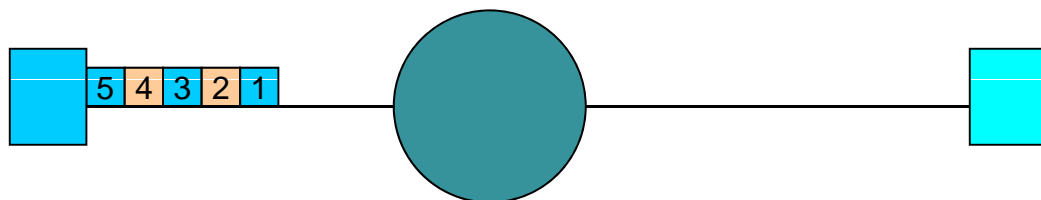
17



16

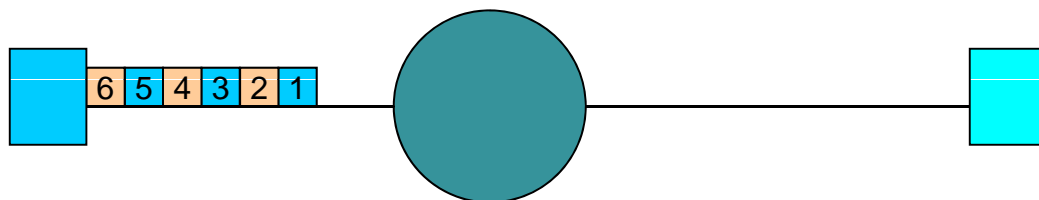


15



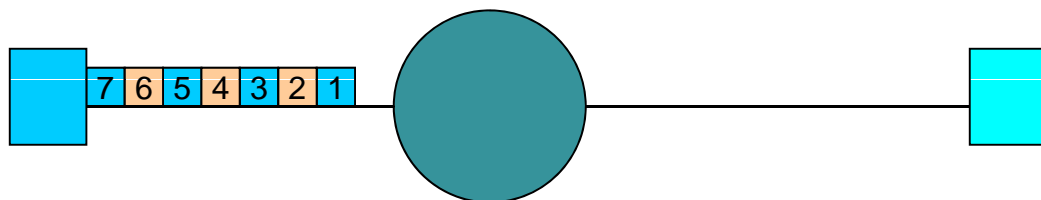
SHARE in Boston

14



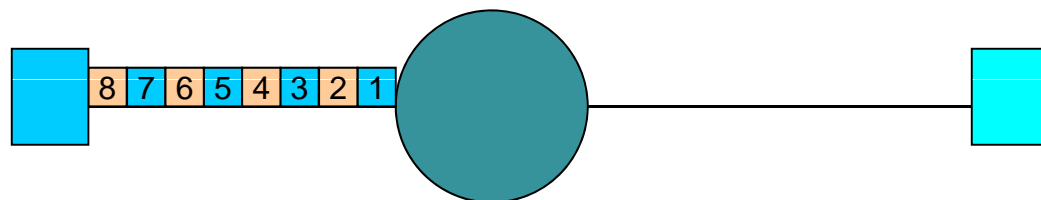
SHARE in Boston

13

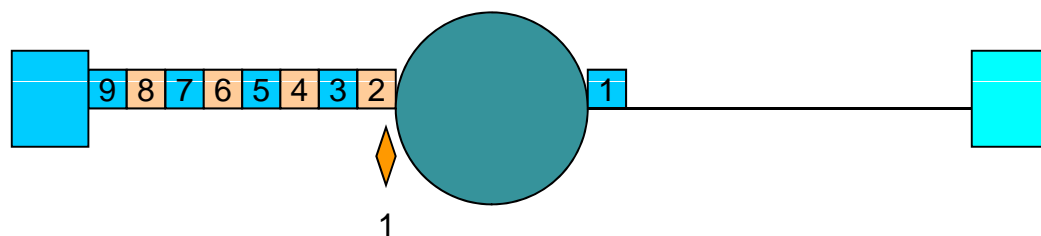


SHARE in Boston

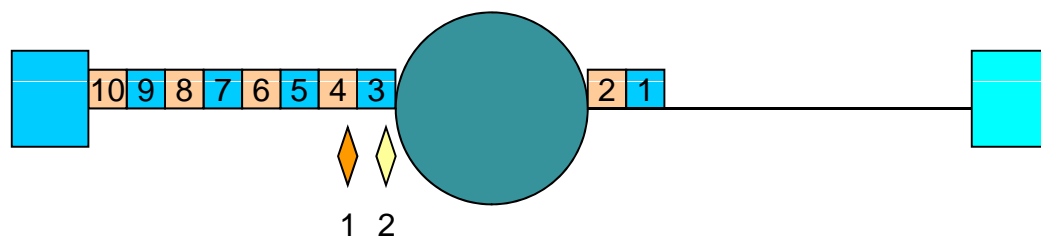
12



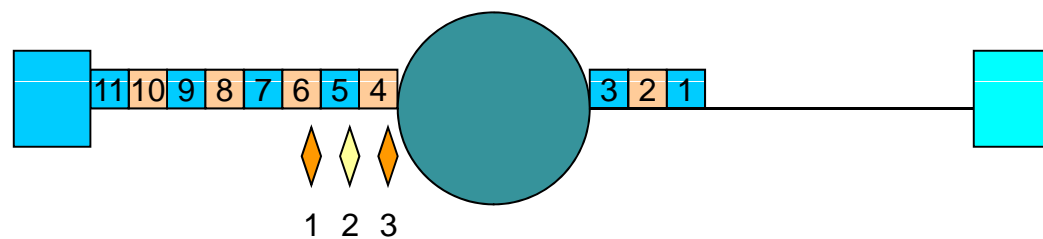
11



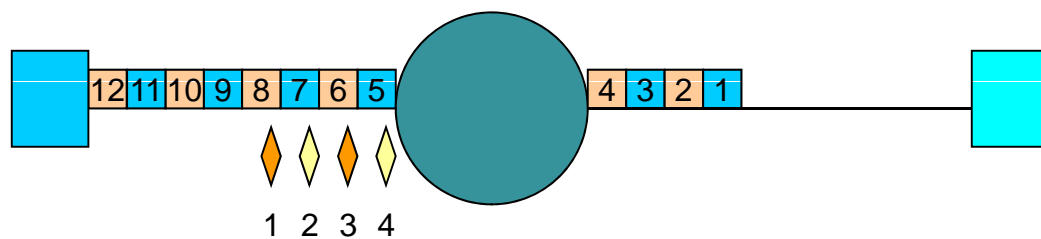
10



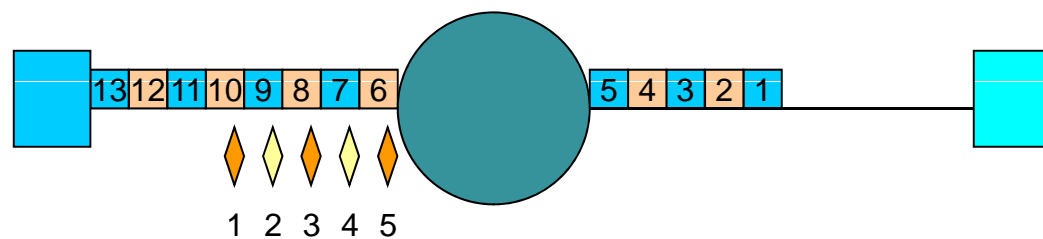
9



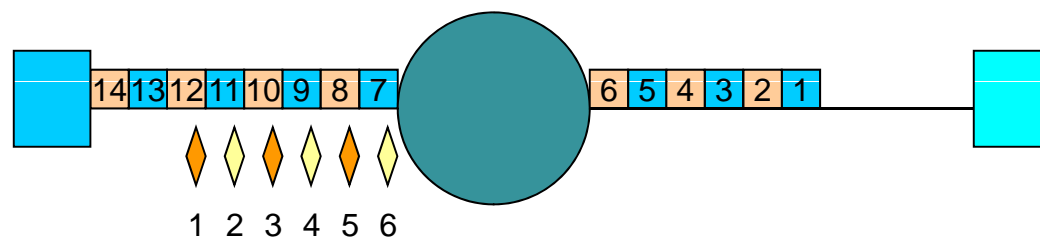
8



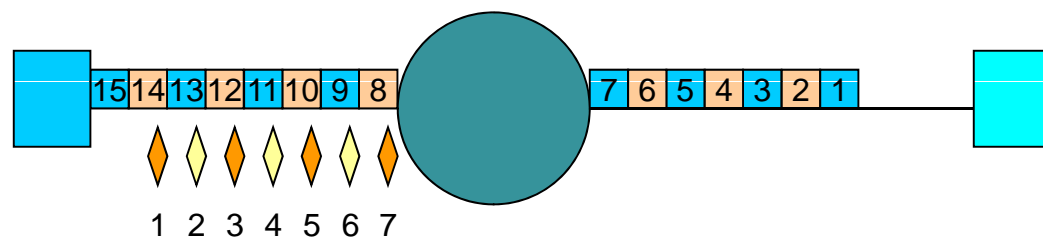
7



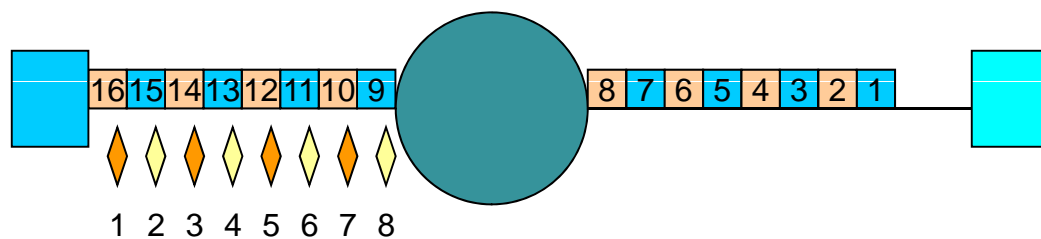
6



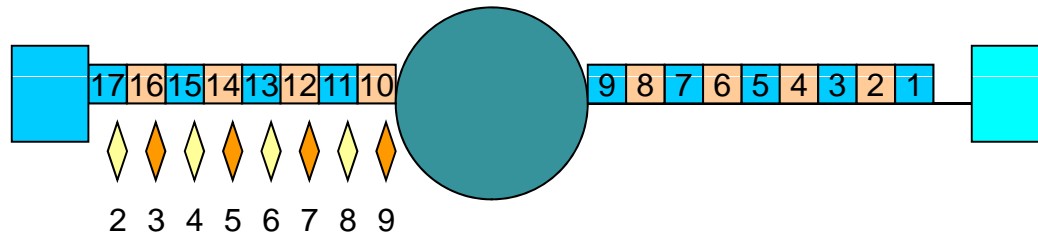
5



4

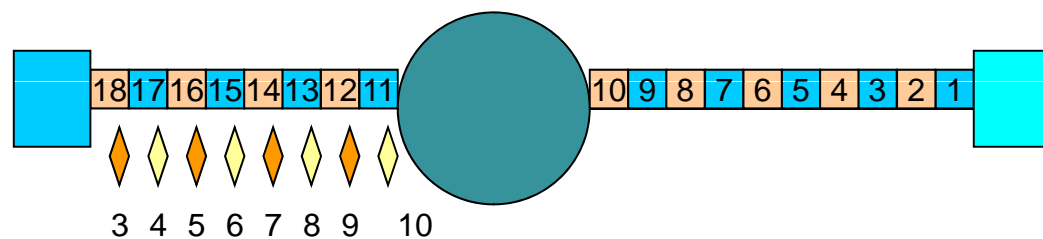


4

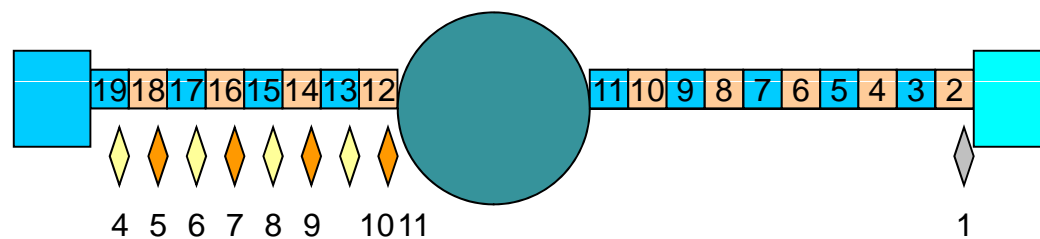




4



4



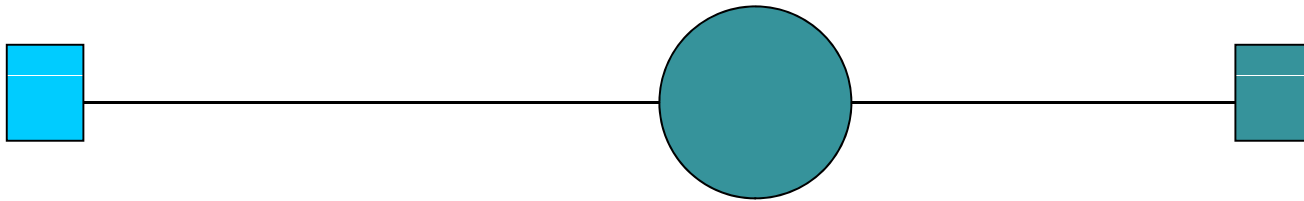
THIS PAGE INTENTIONALLY
LEFT BLANK

Example: A not so full pipe

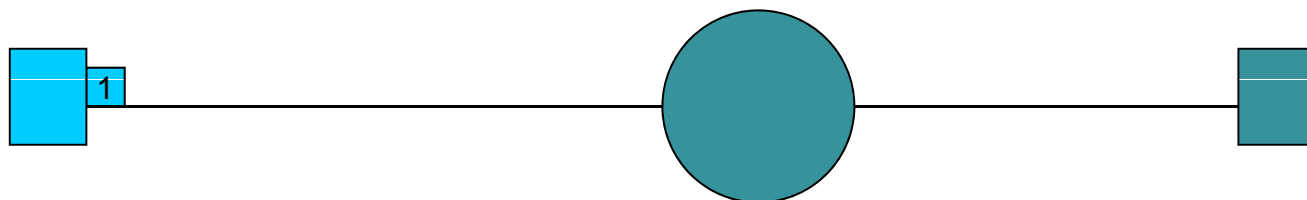
BUFFER CREDITS

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

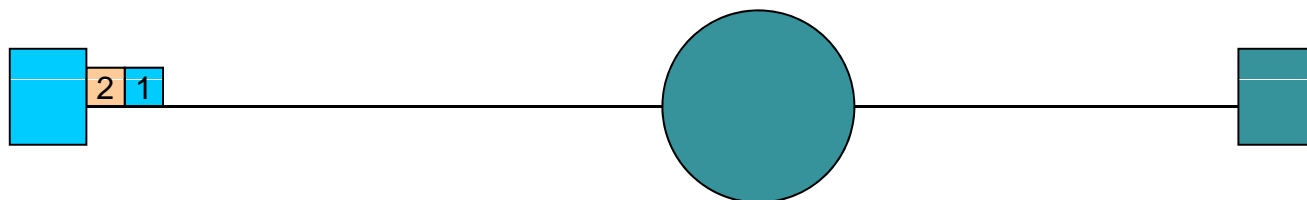
20



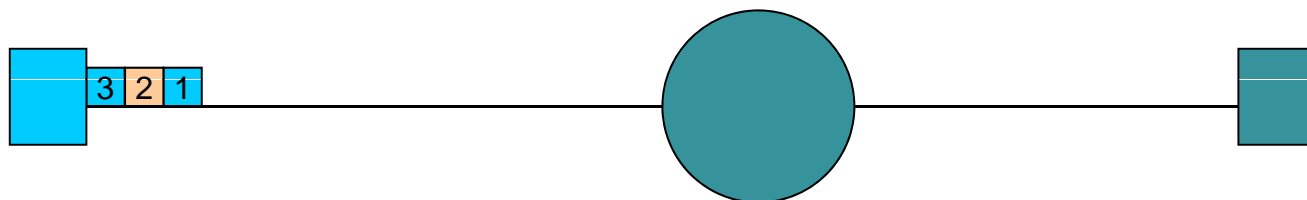
19



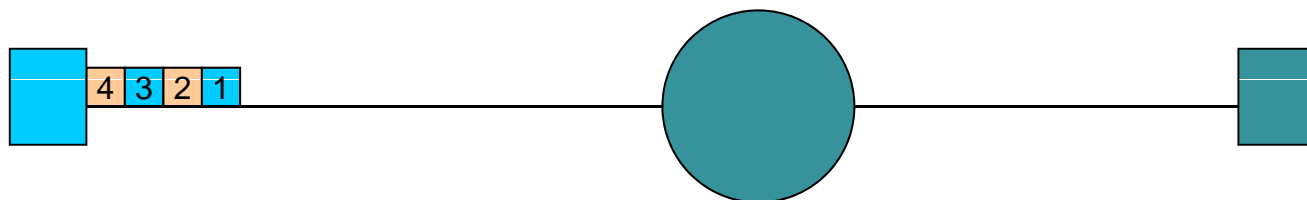
18



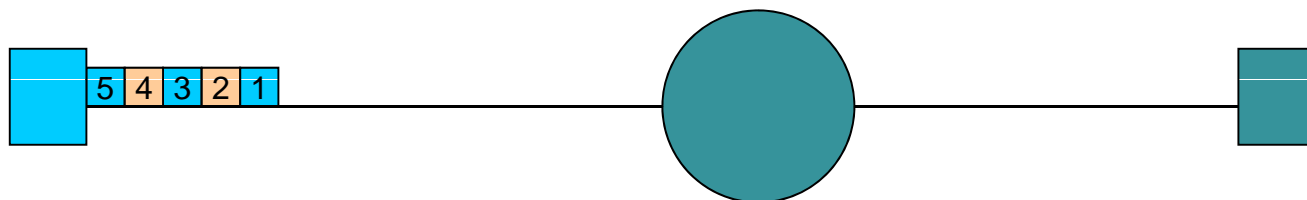
17



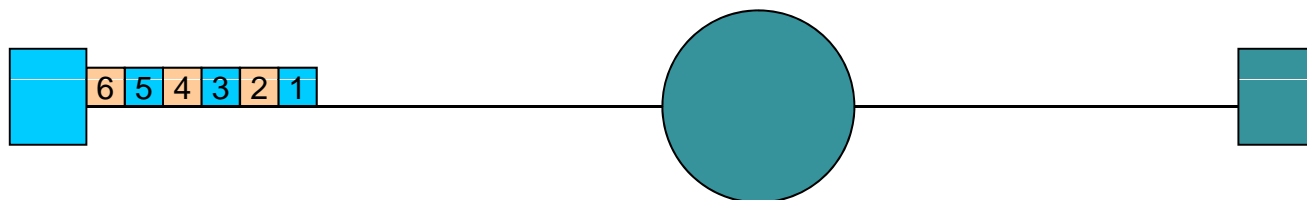
16



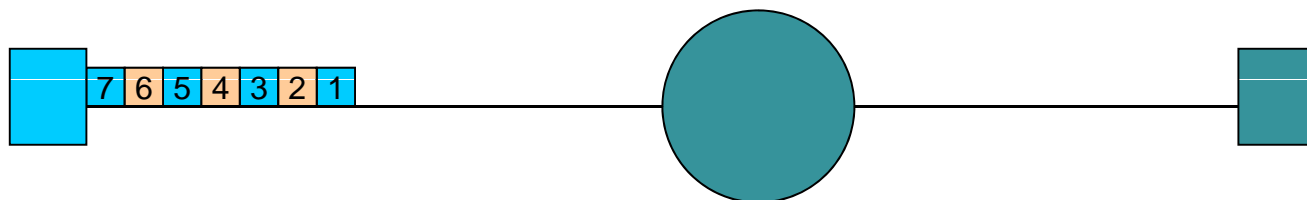
15



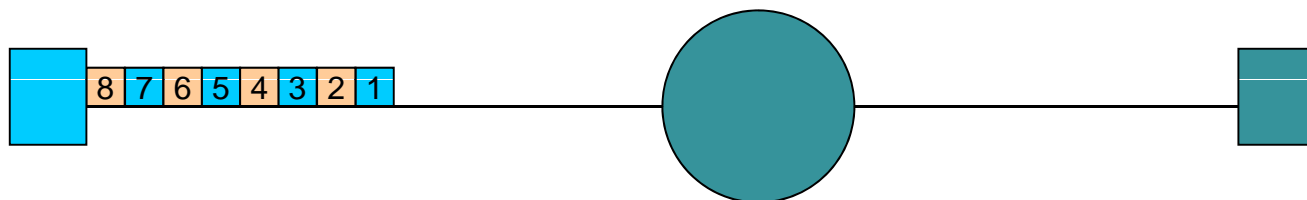
14



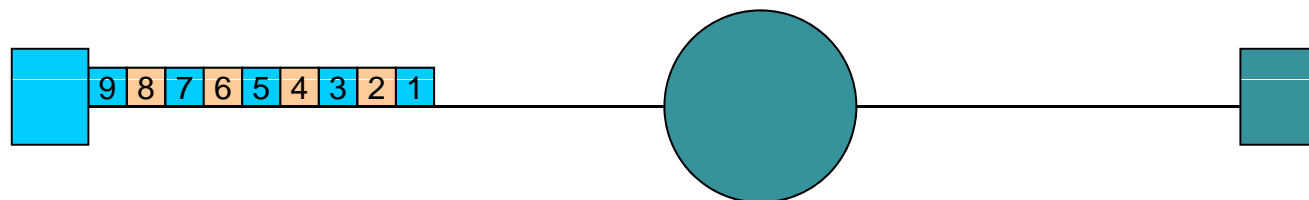
13



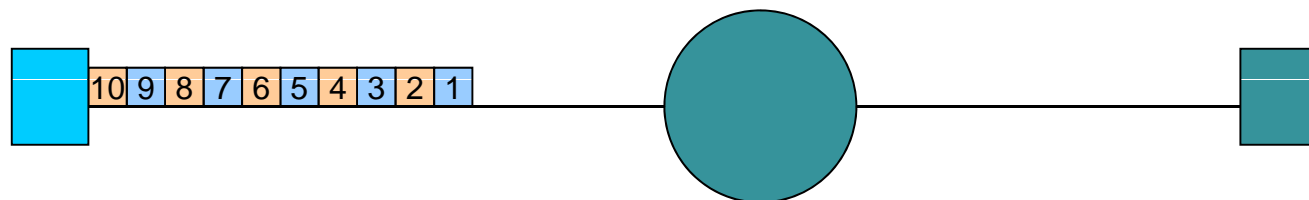
12



11

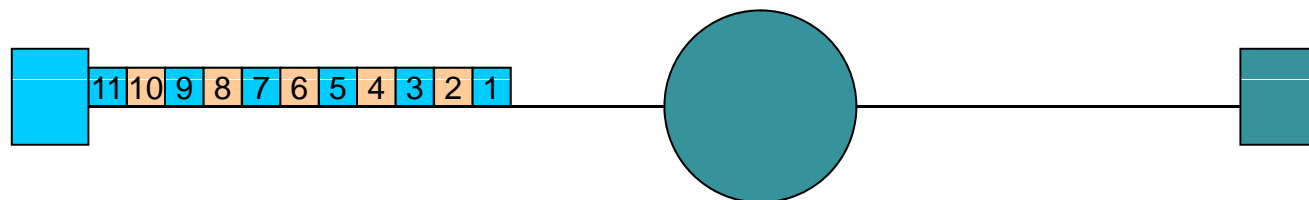


10

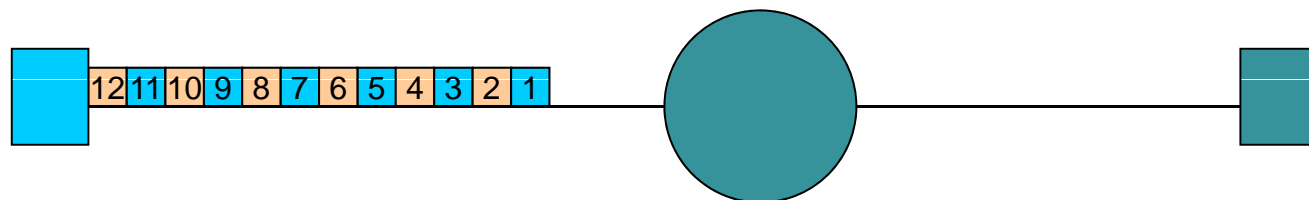


SHARE in Boston

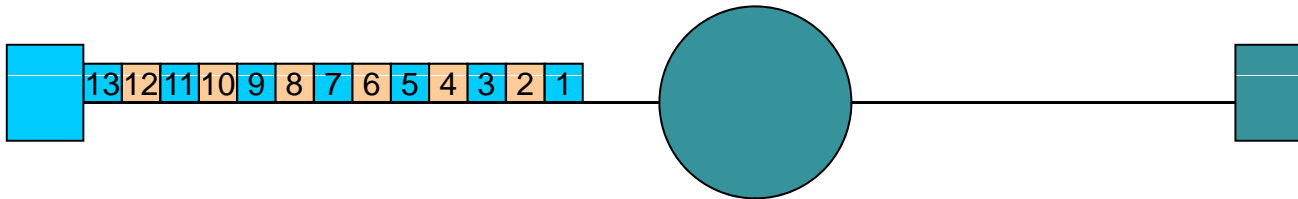
9



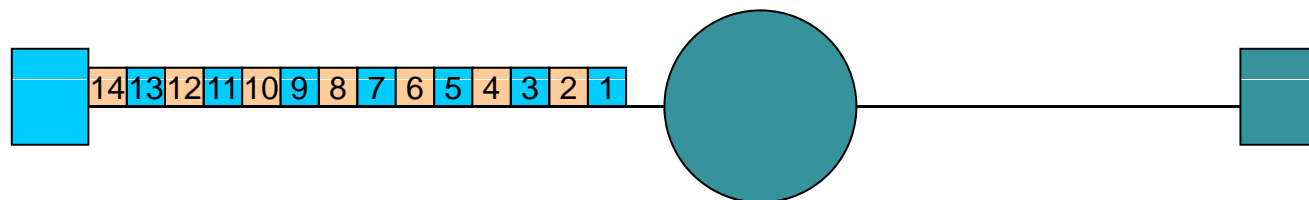
8



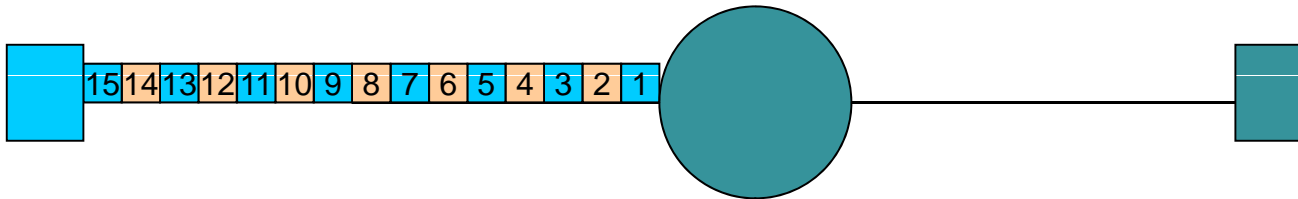
7



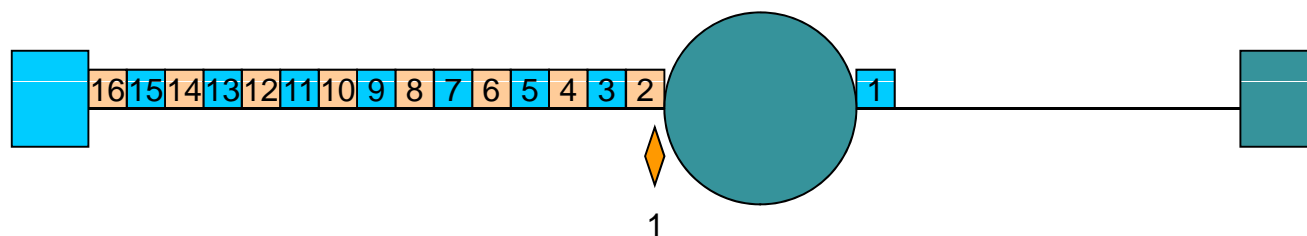
6



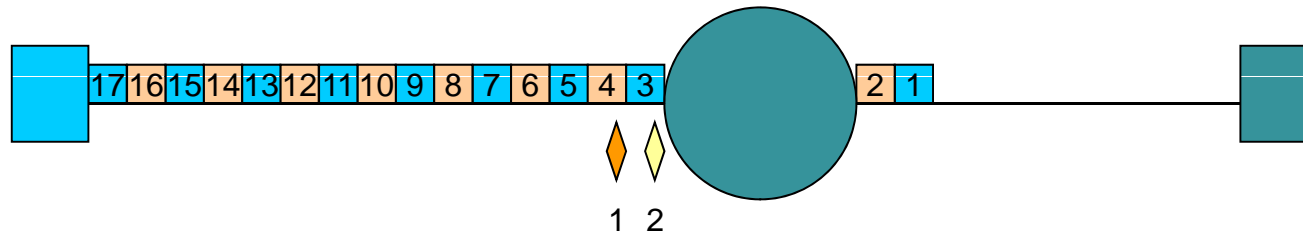
5



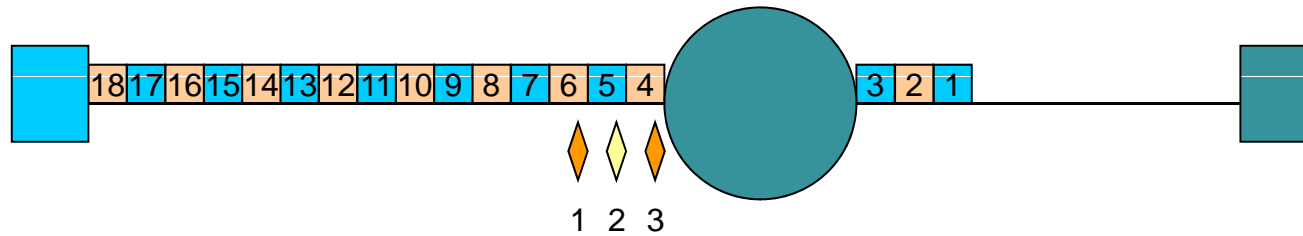
4



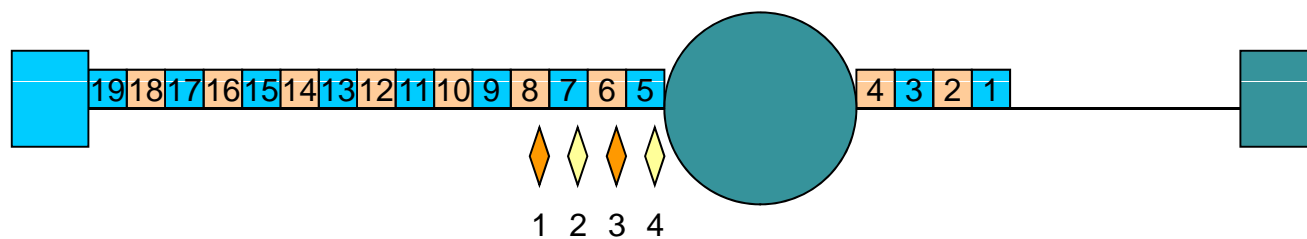
3



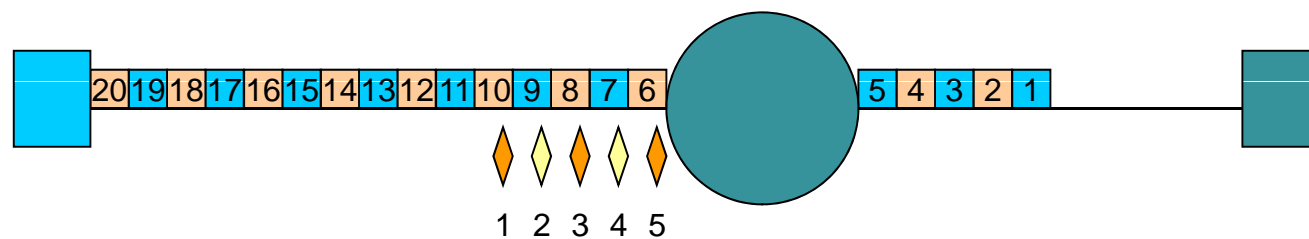
2



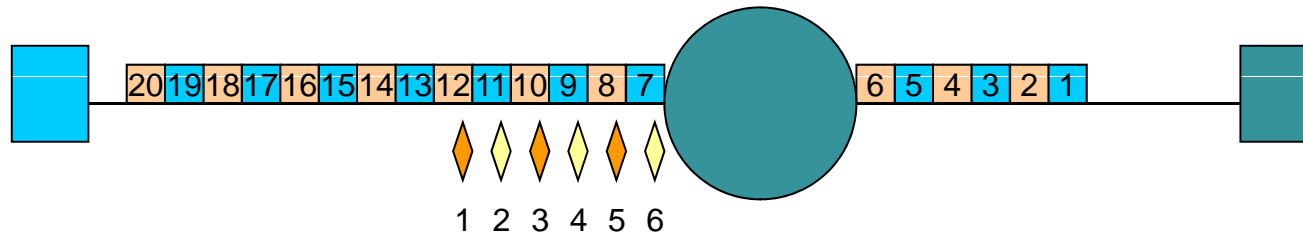
1



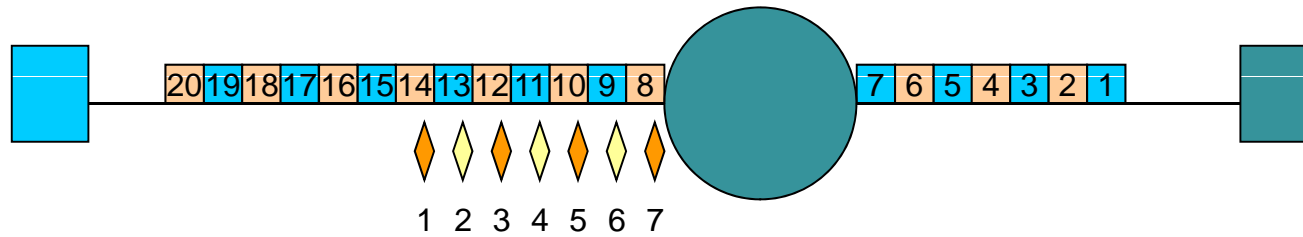
0



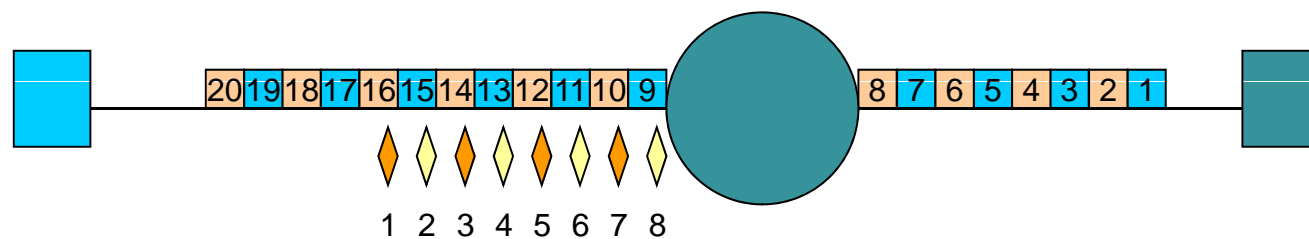
0



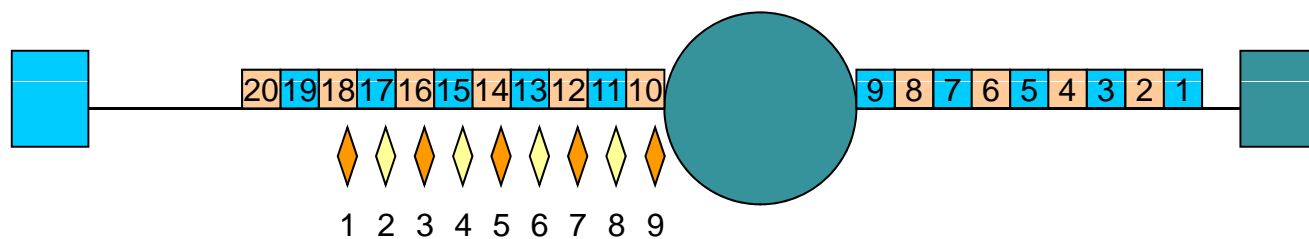
0



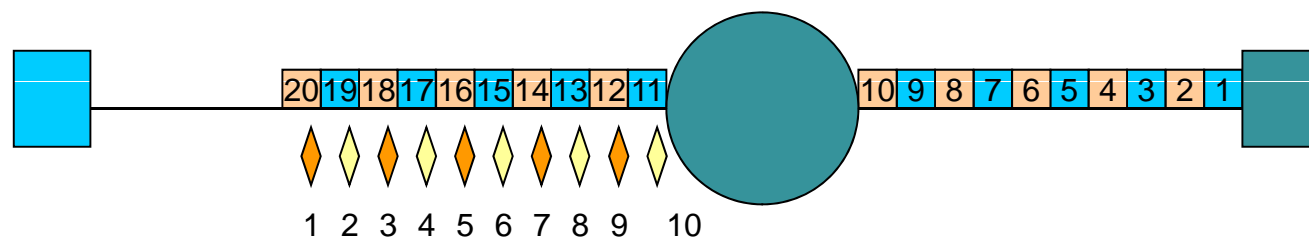
0



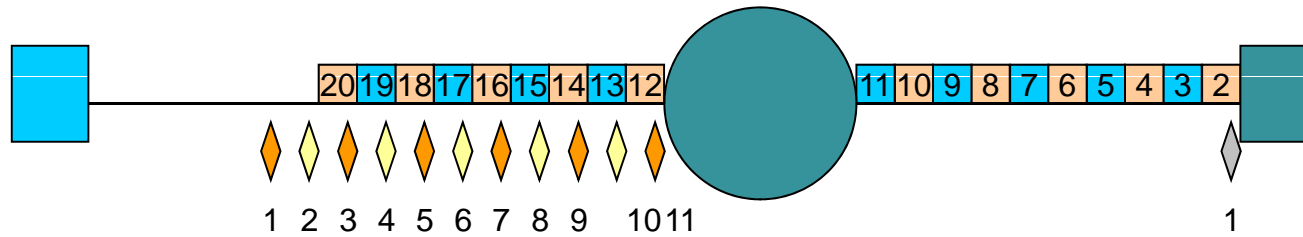
0



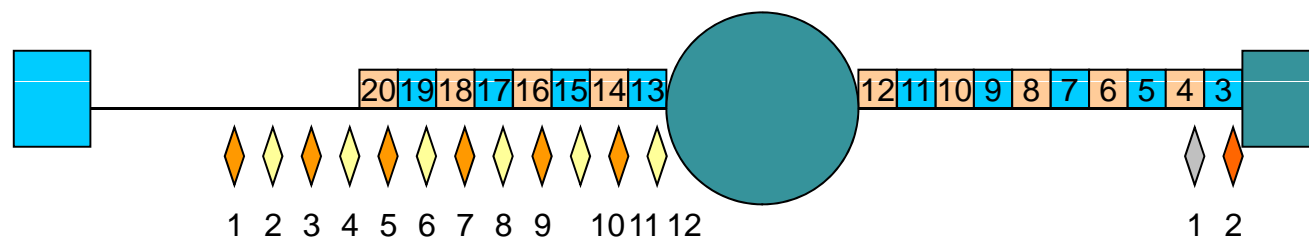
0



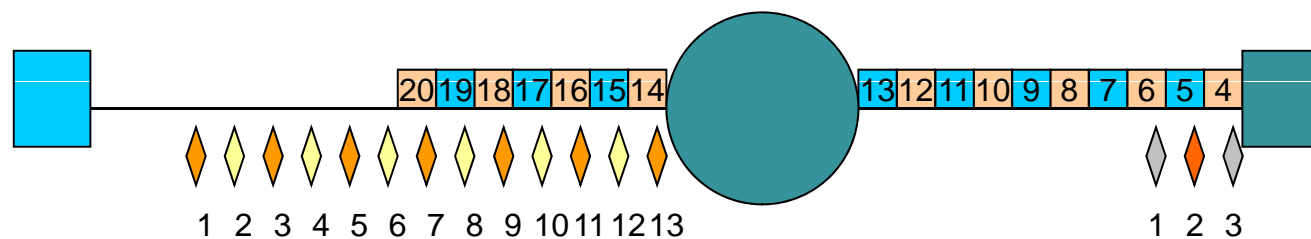
0



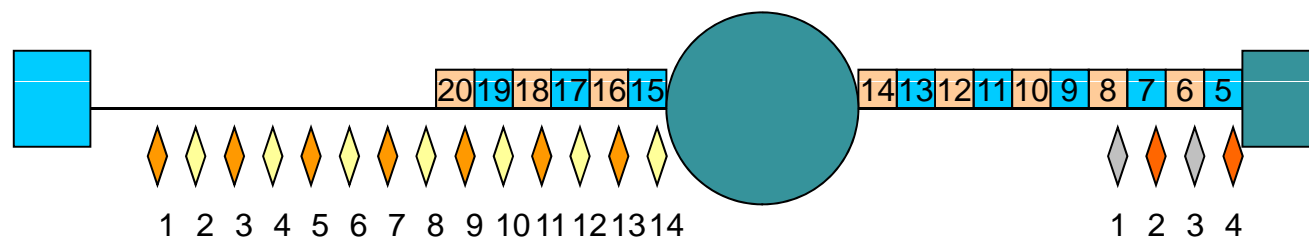
0



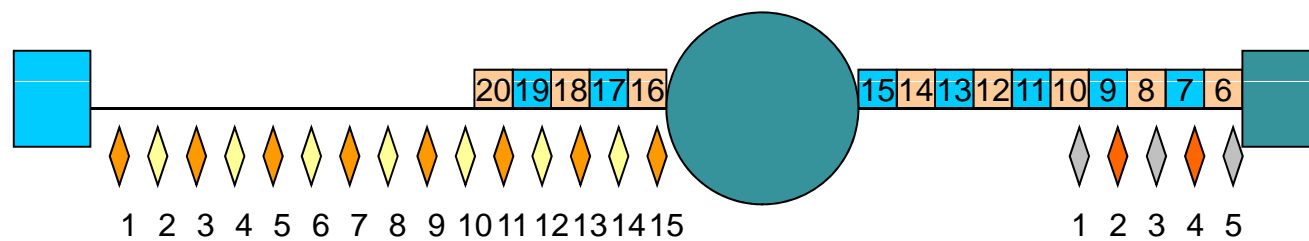
0



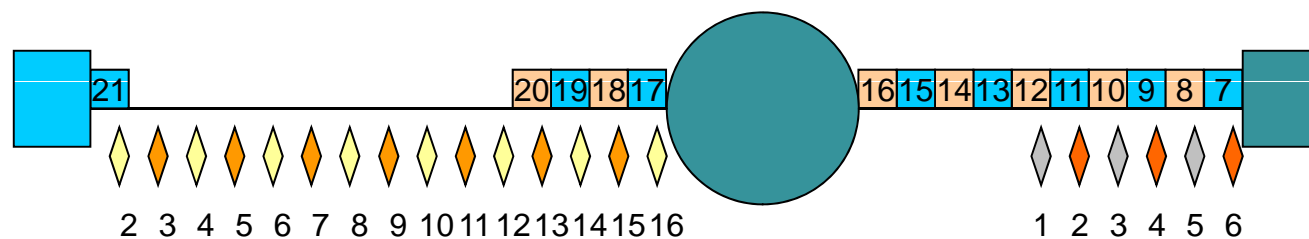
0



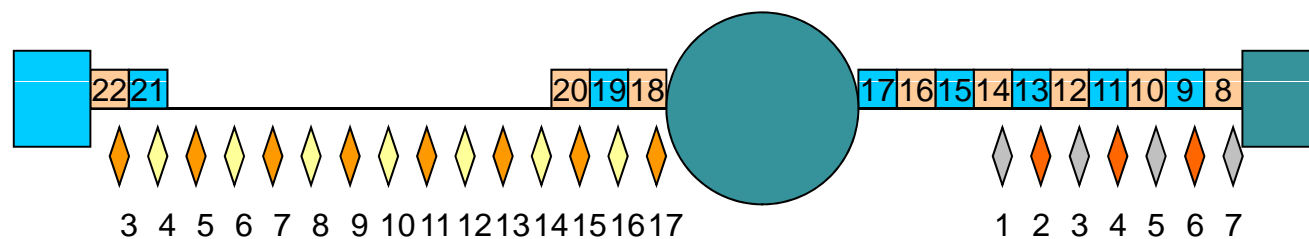
0



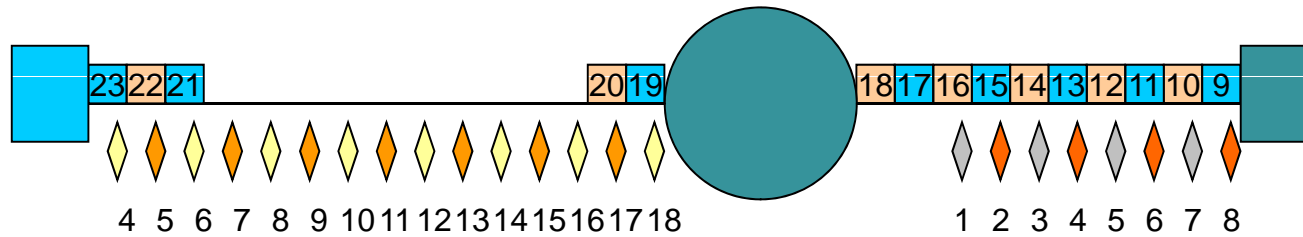
0



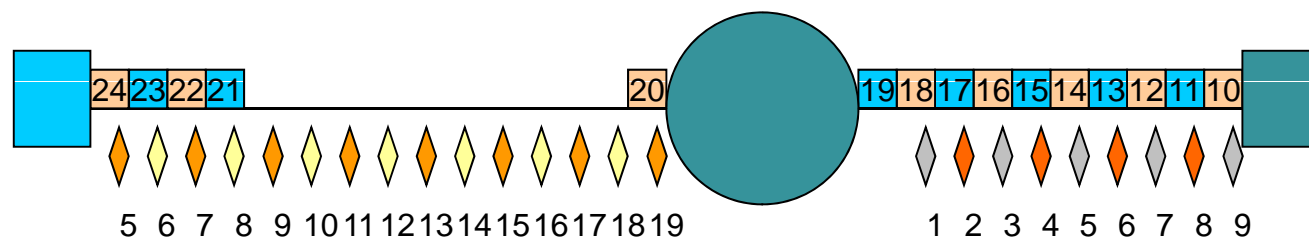
0



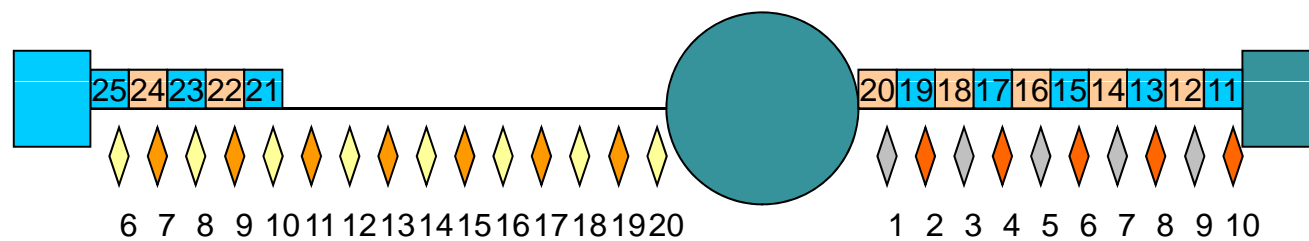
0



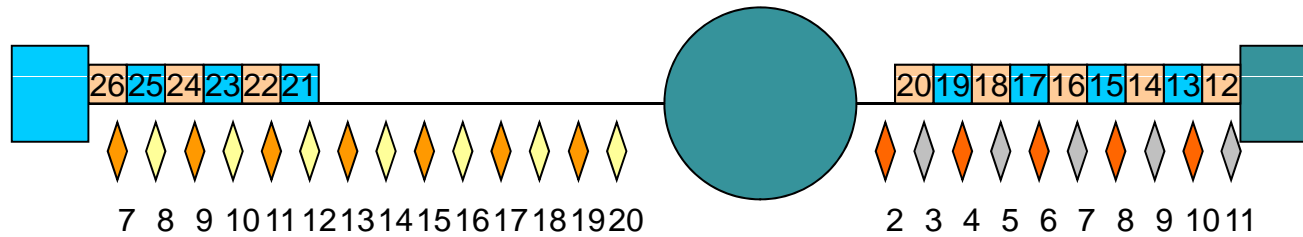
0



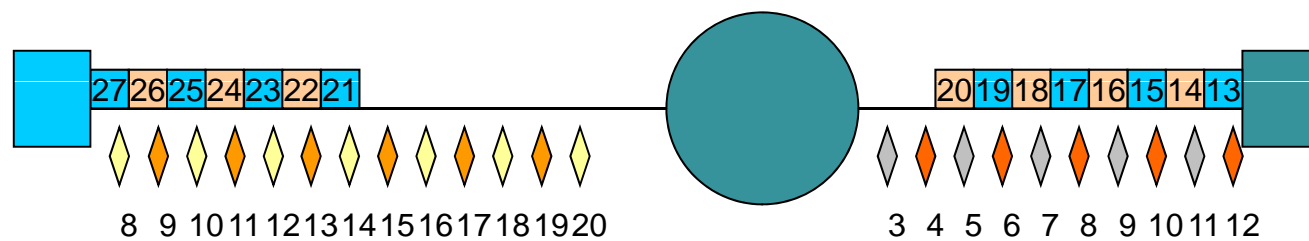
0



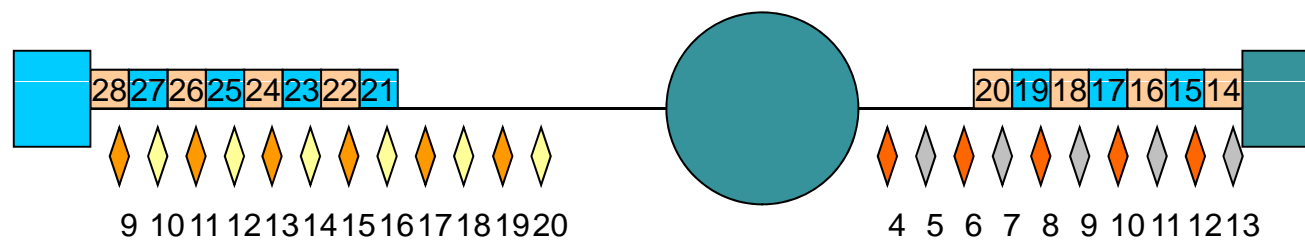
0



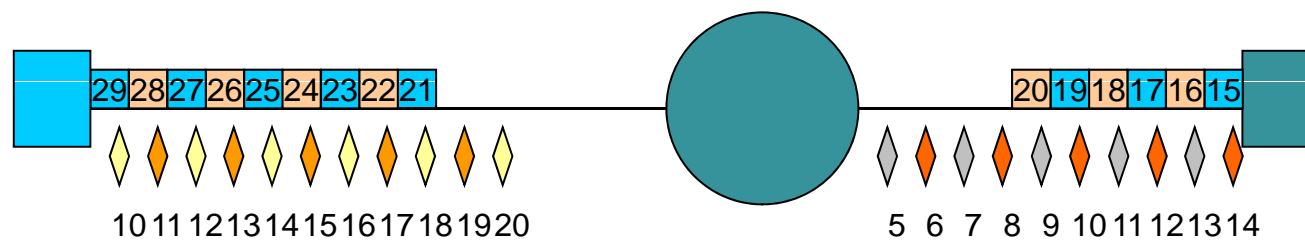
0



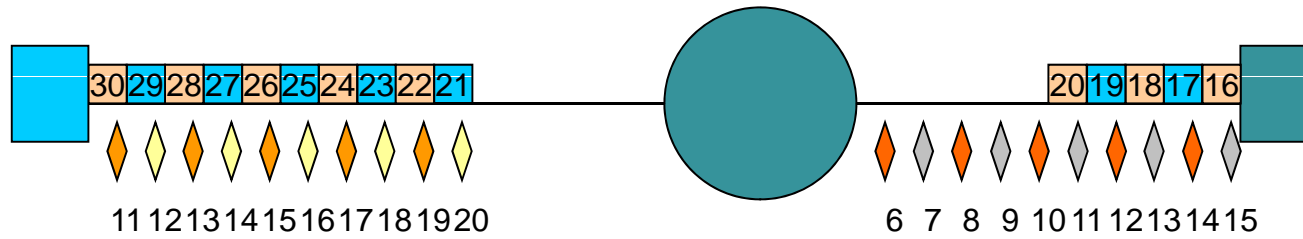
0



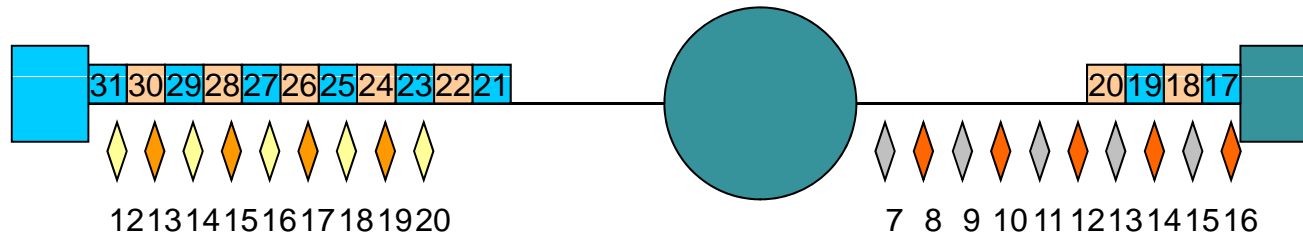
0



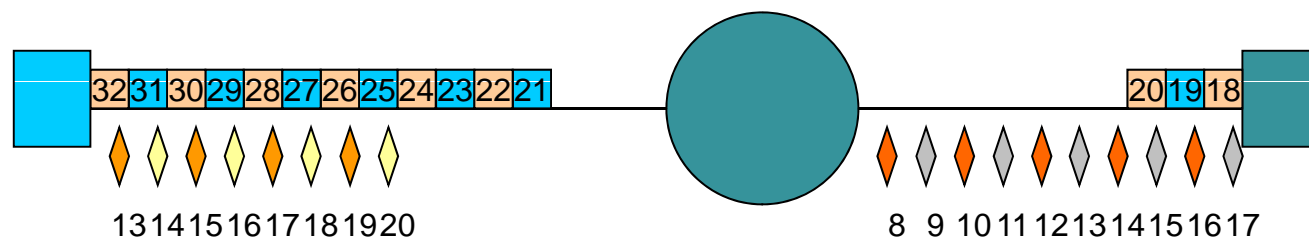
0



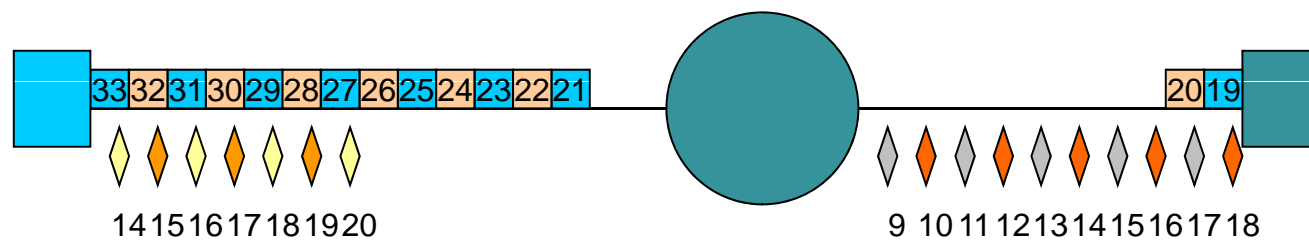
0



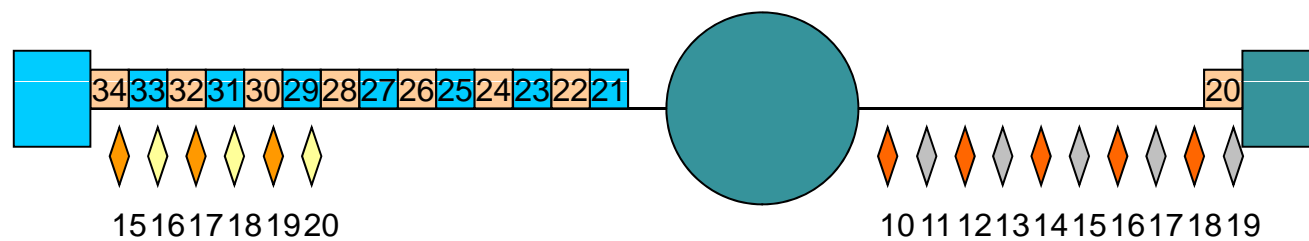
0



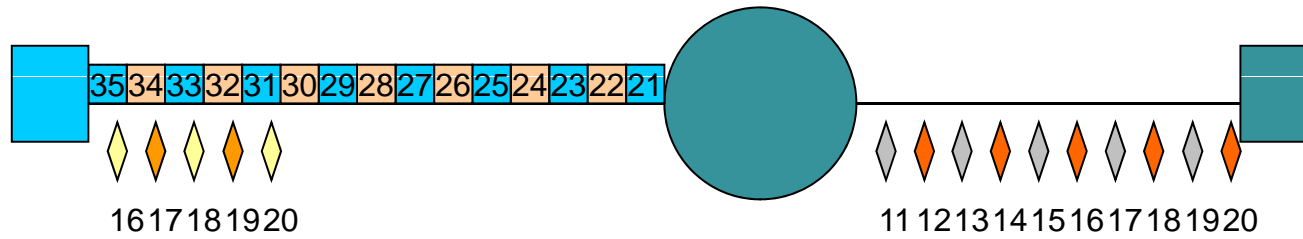
0



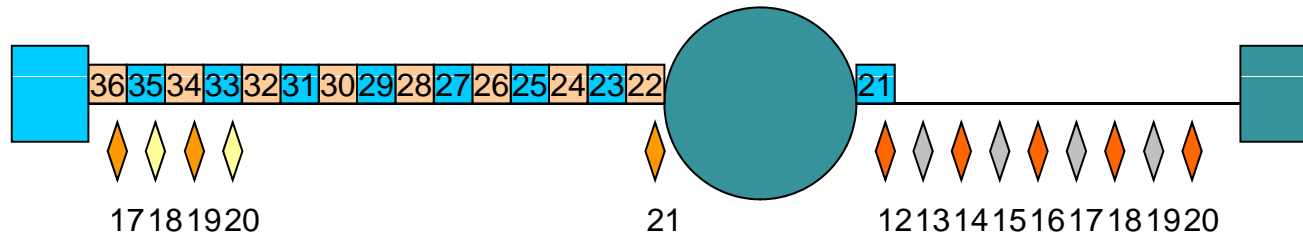
0



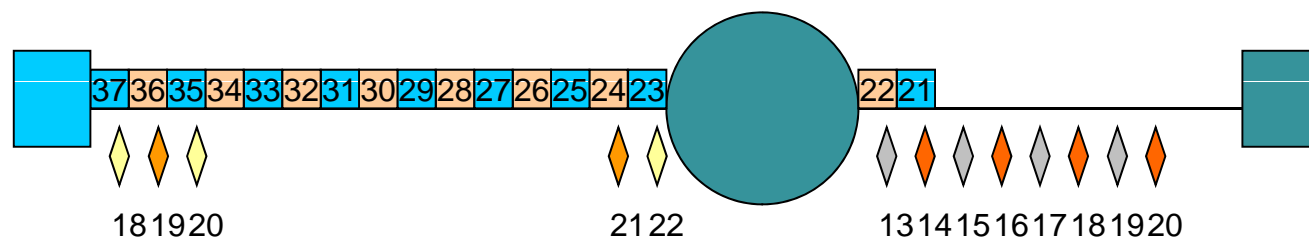
0



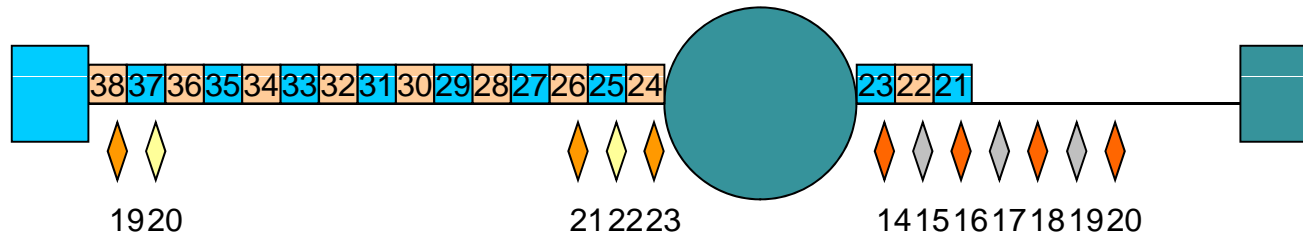
0



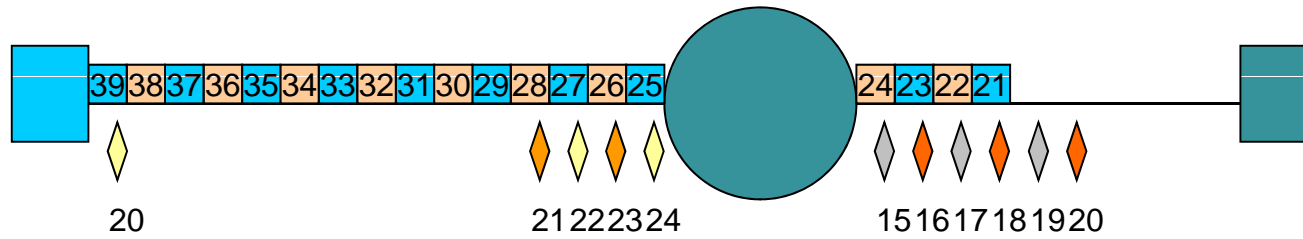
0



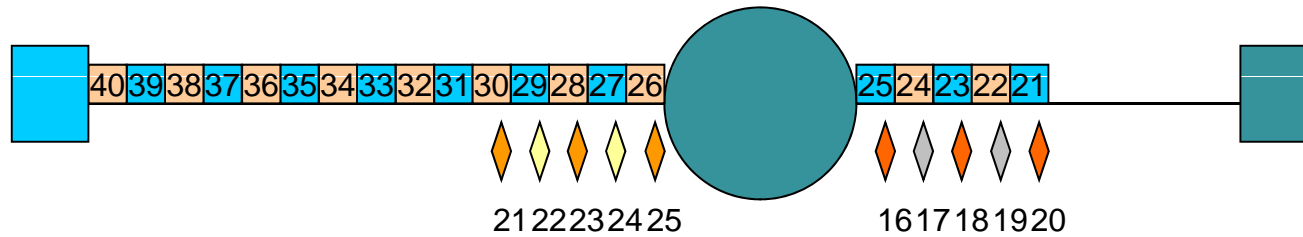
0



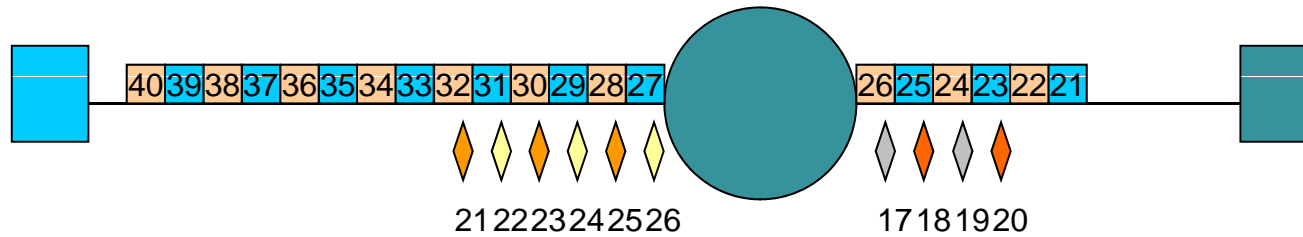
0



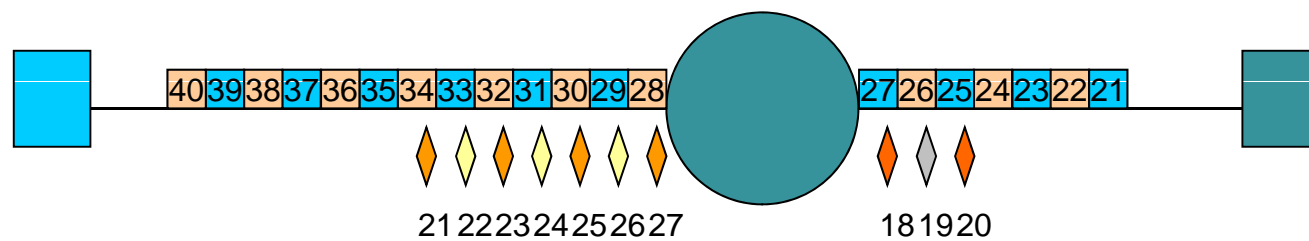
0



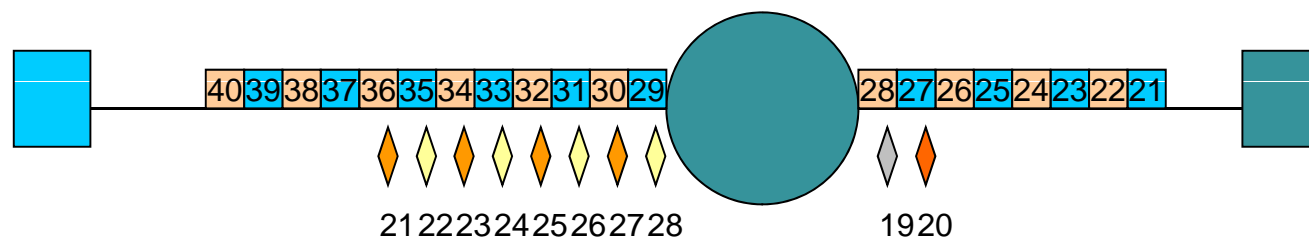
0



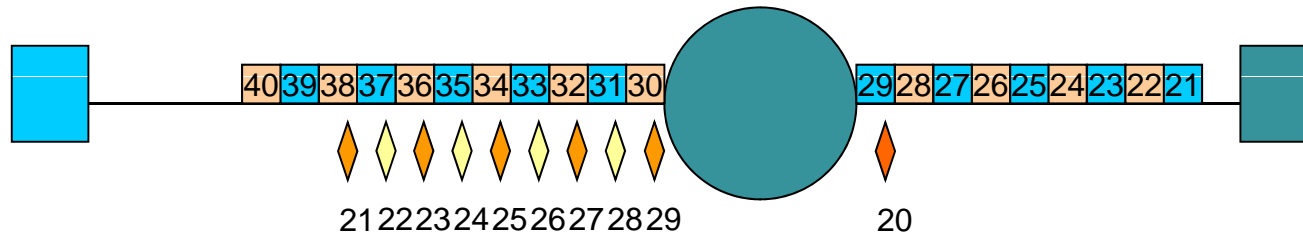
0



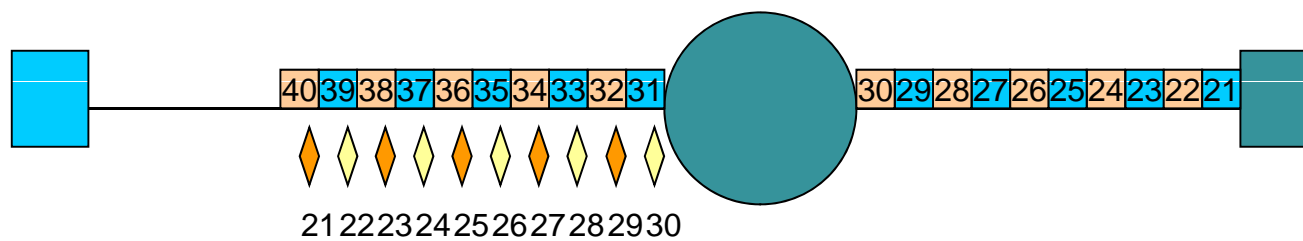
0



0

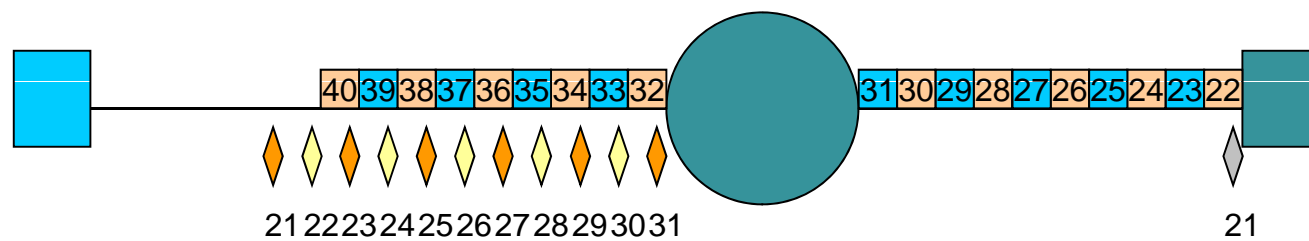


0

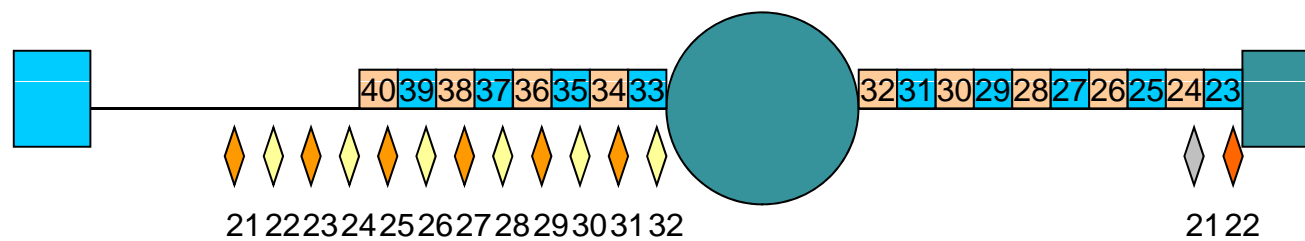




0



0



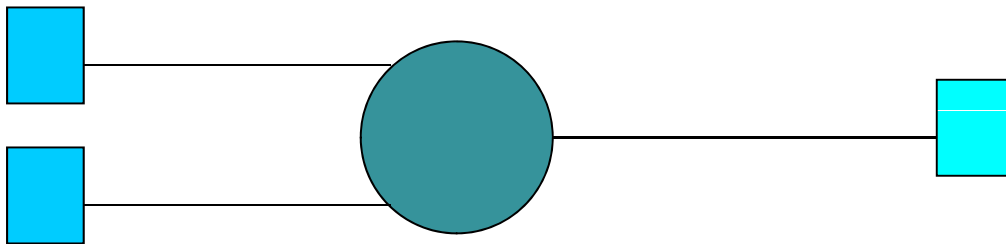
THIS PAGE INTENTIONALLY
LEFT BLANK

Example: Multiple Senders and One Receiver

BUFFER CREDITS

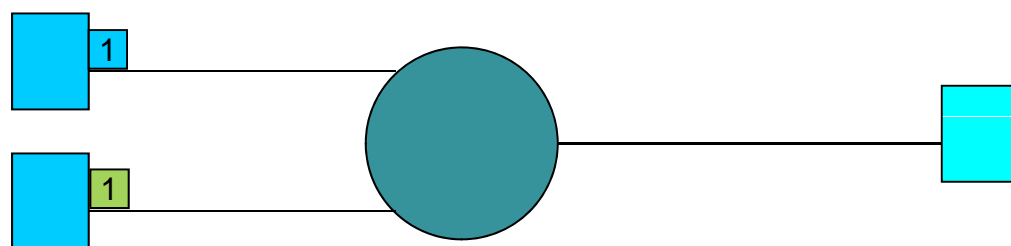
Suppose there are multiple senders to one receiver
Each sender attempts to send at 100% link speed

20



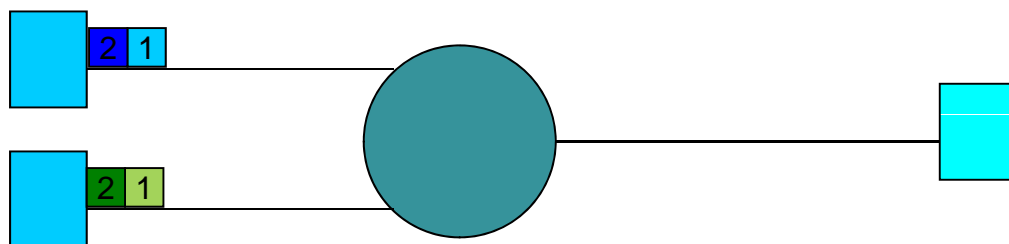
20

19



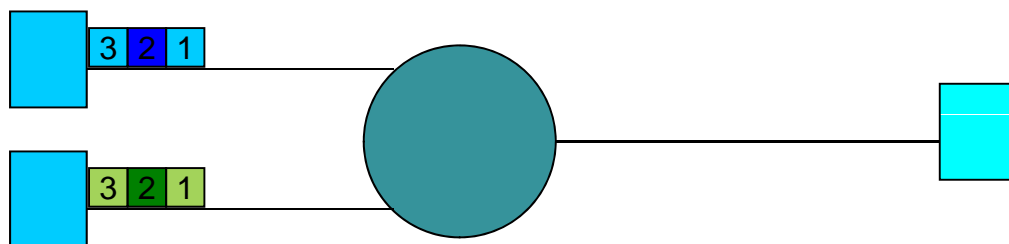
19

18



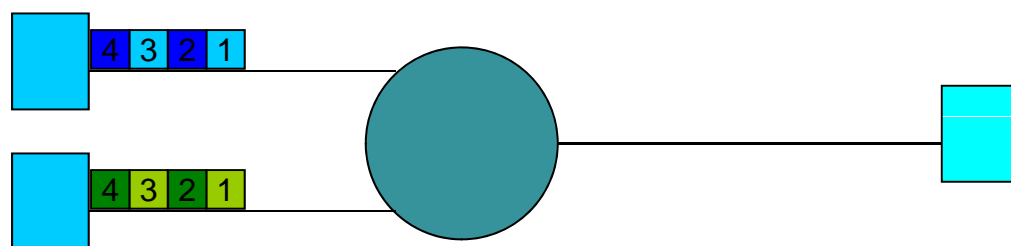
18

17



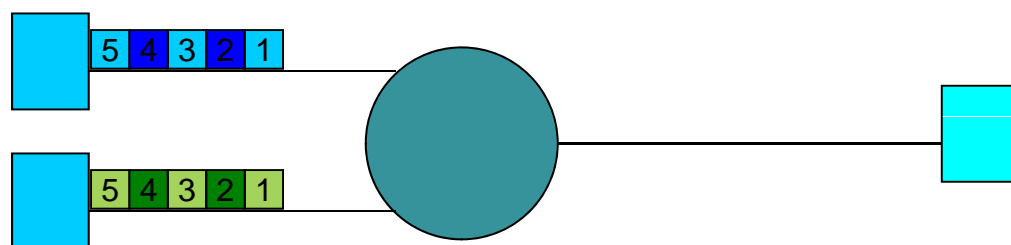
17

16



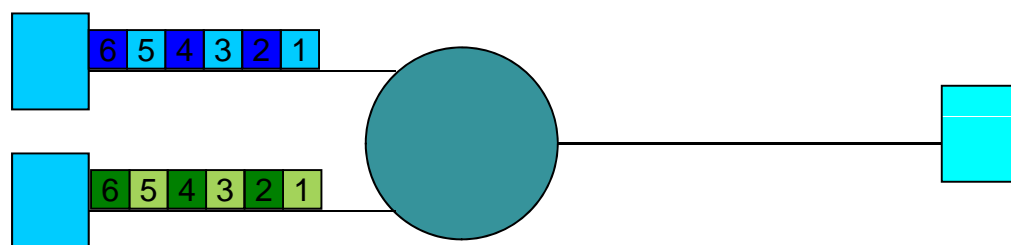
16

15



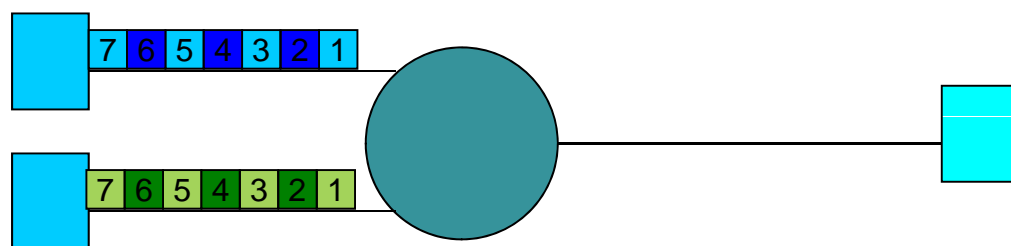
15

14



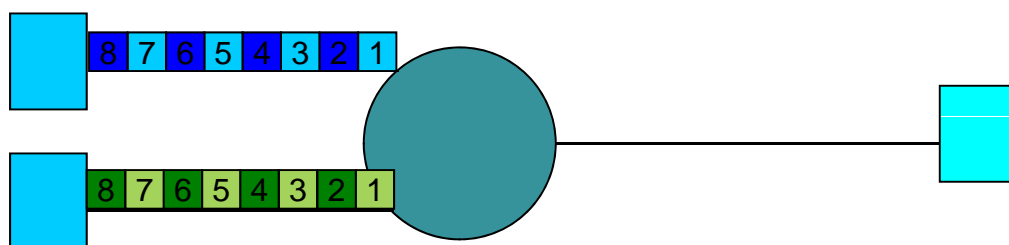
14

13

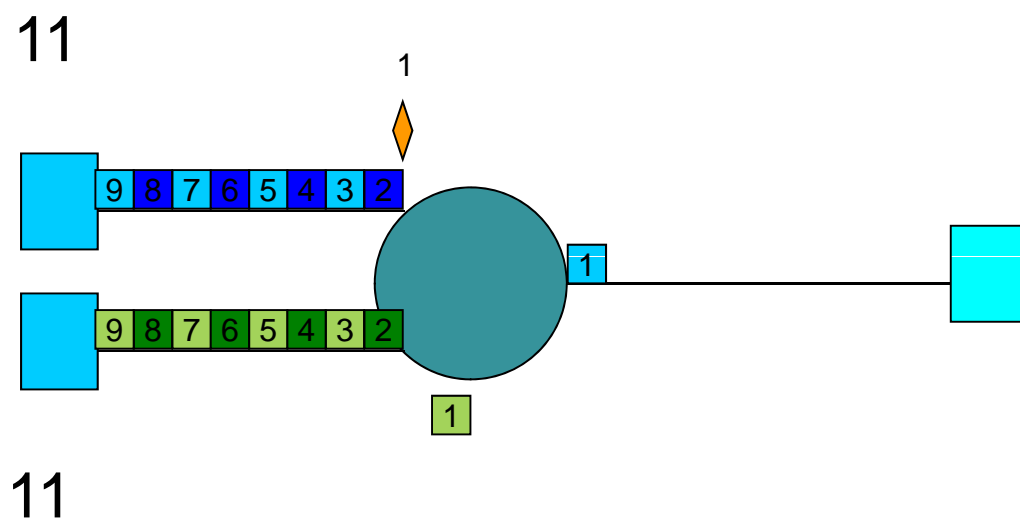


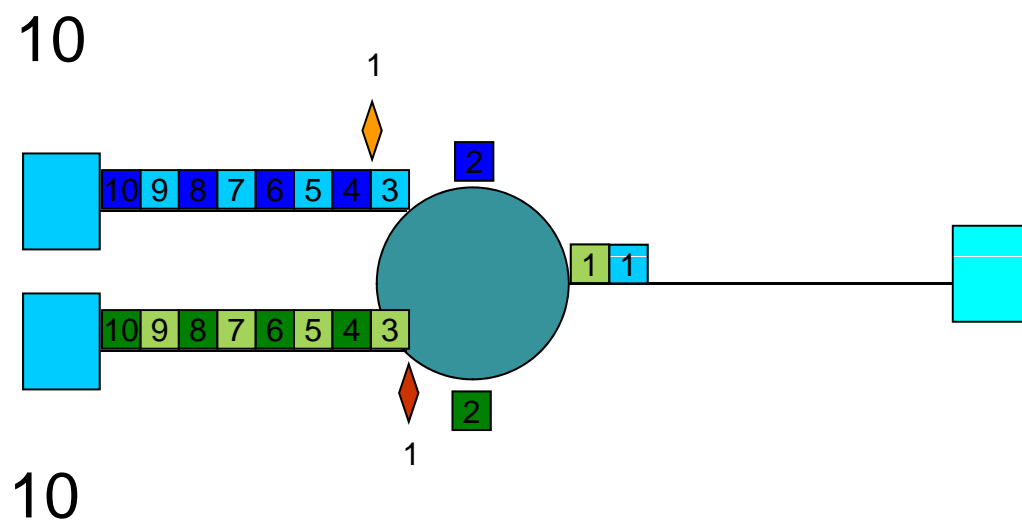
13

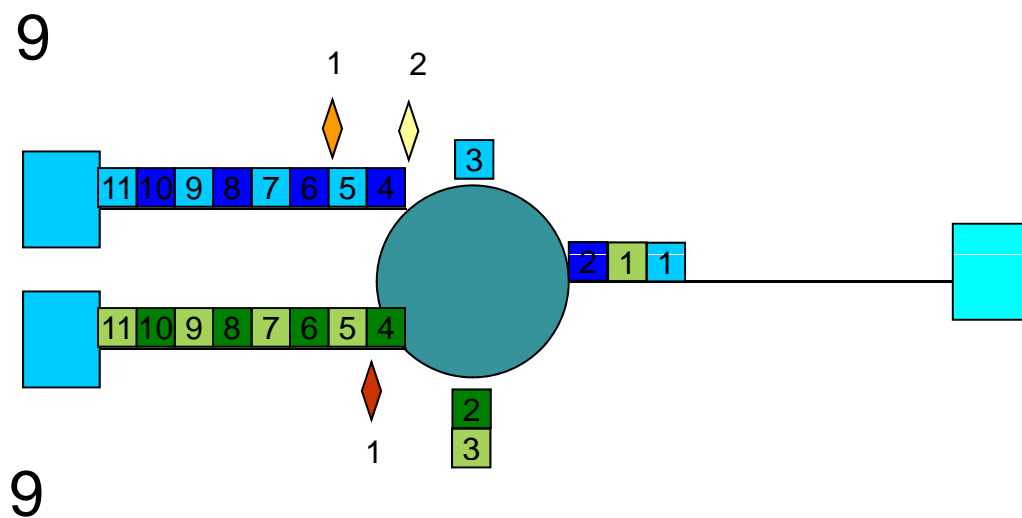
12

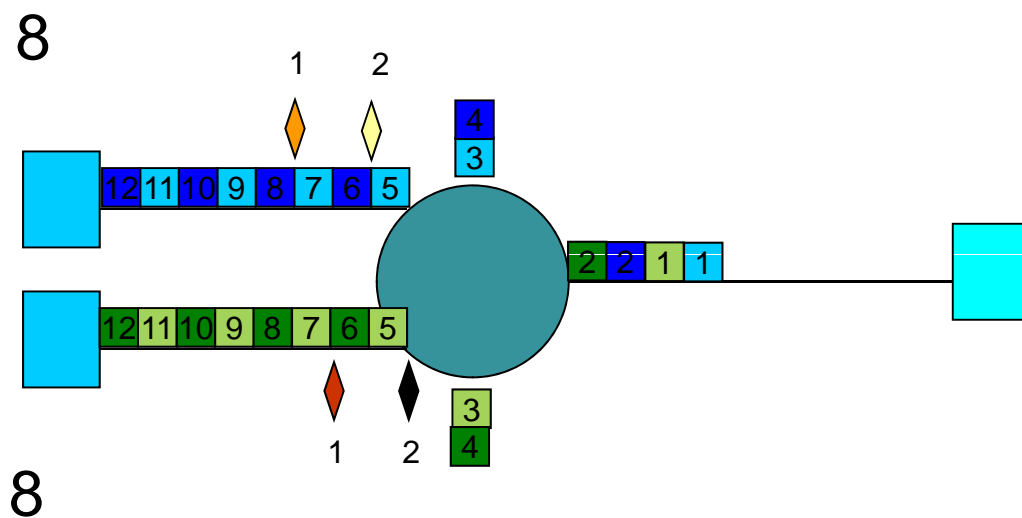


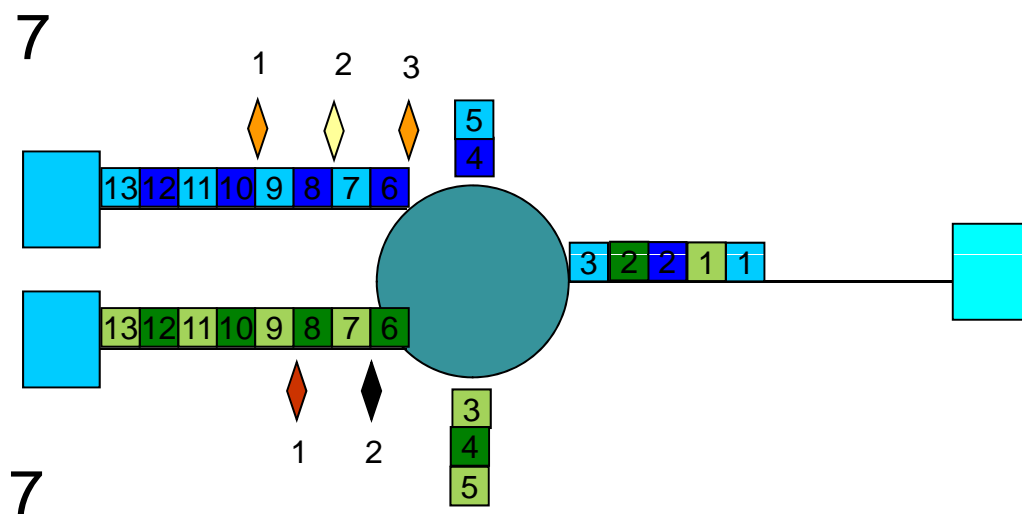
12

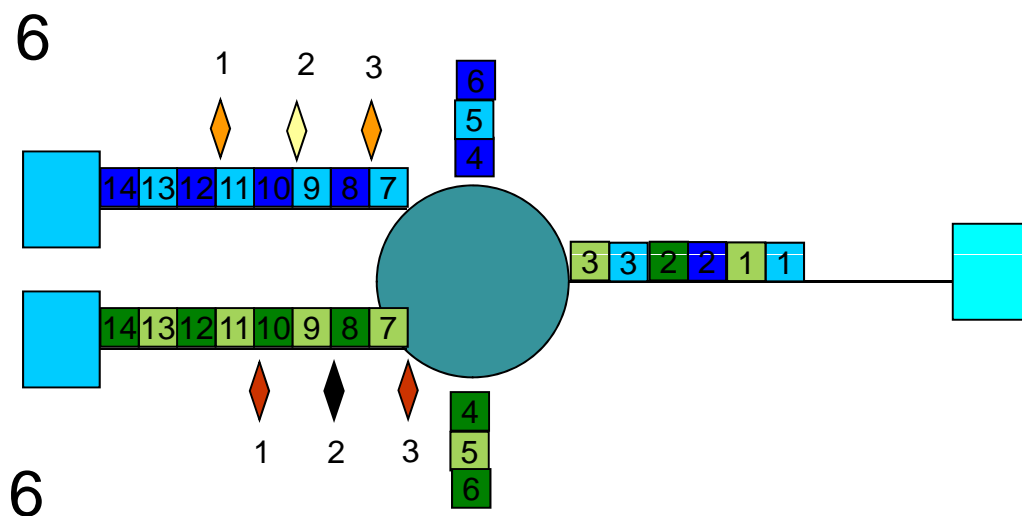


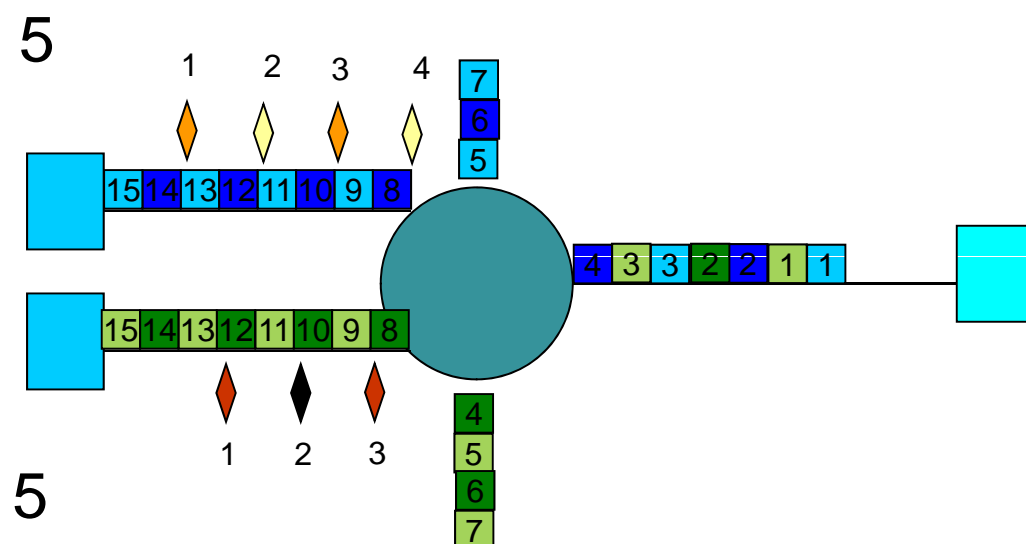


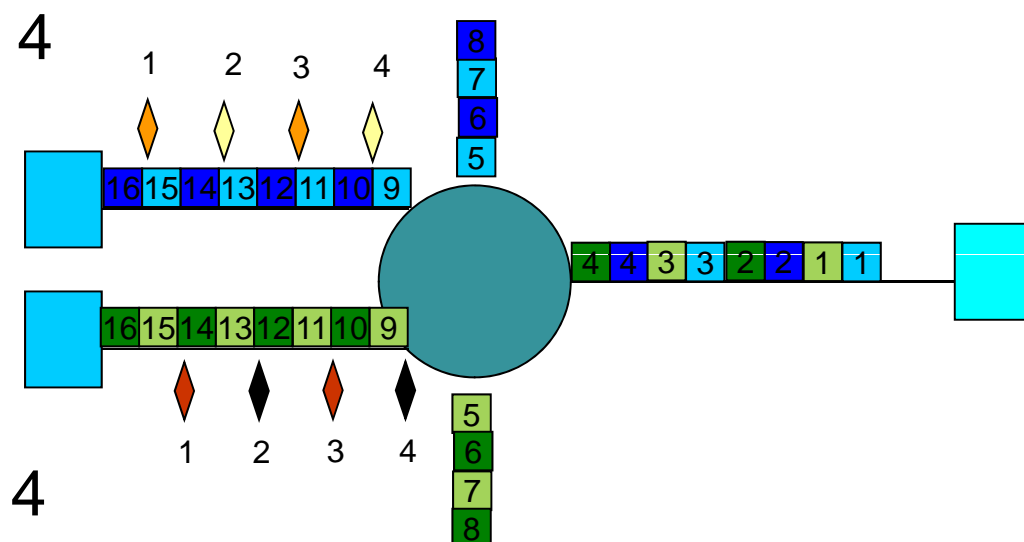


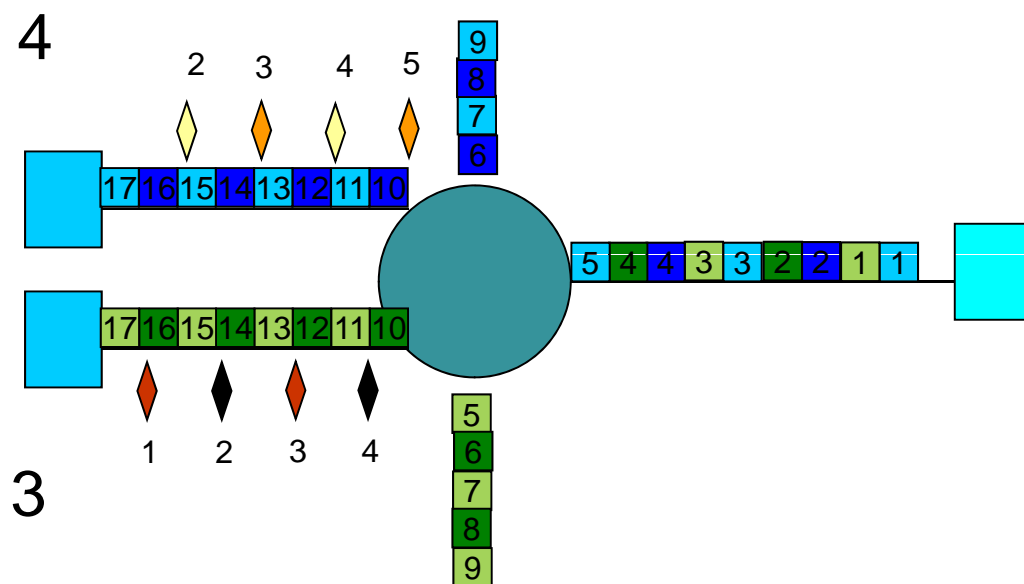


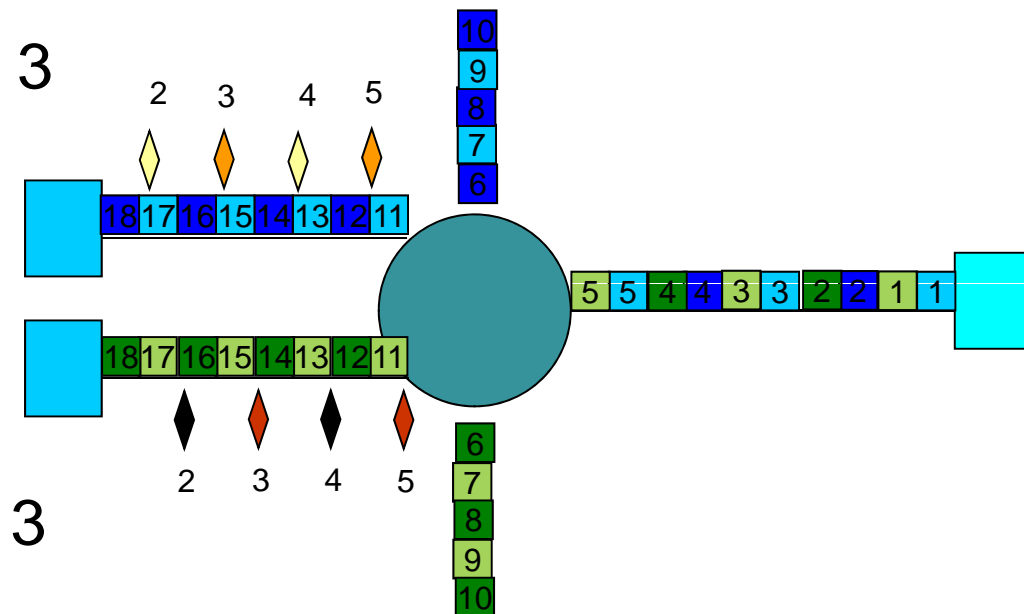


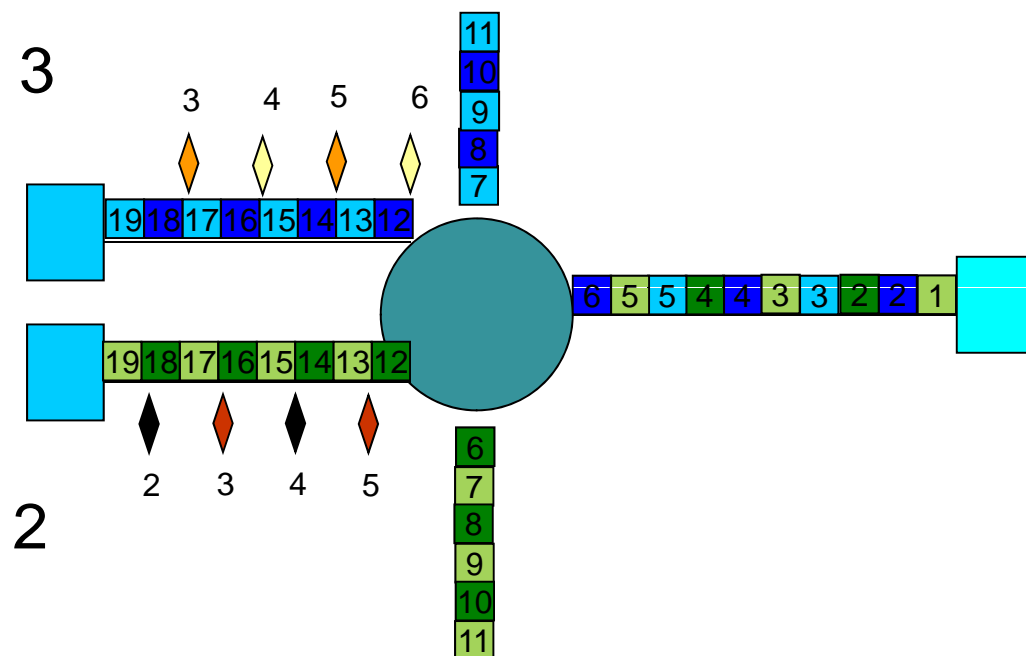


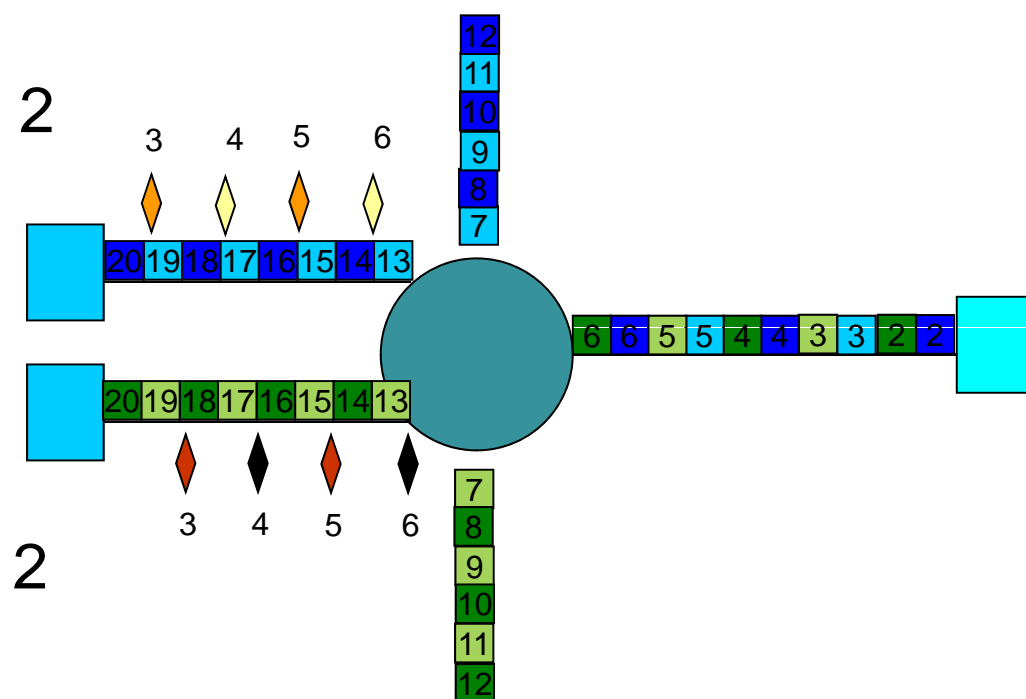


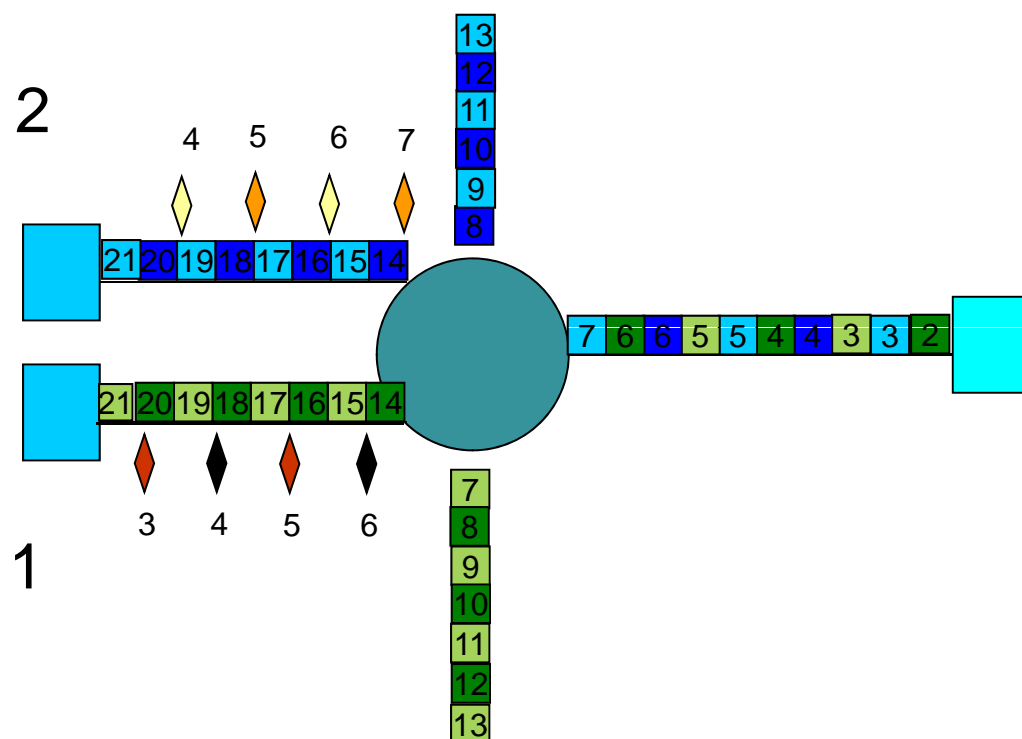


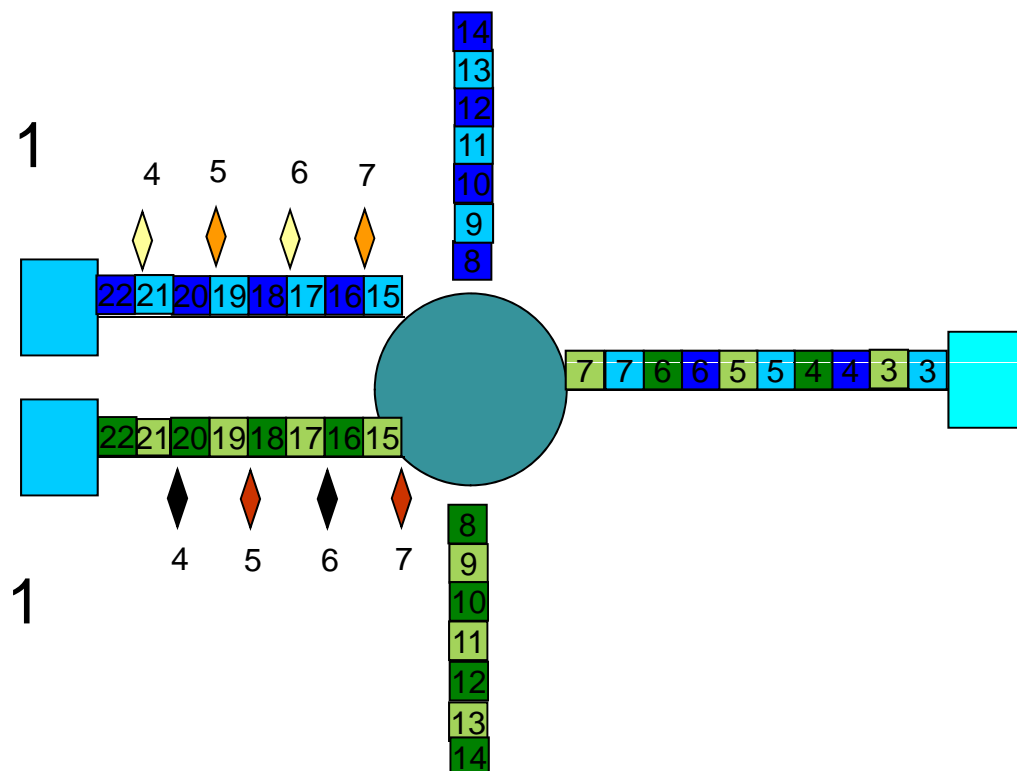


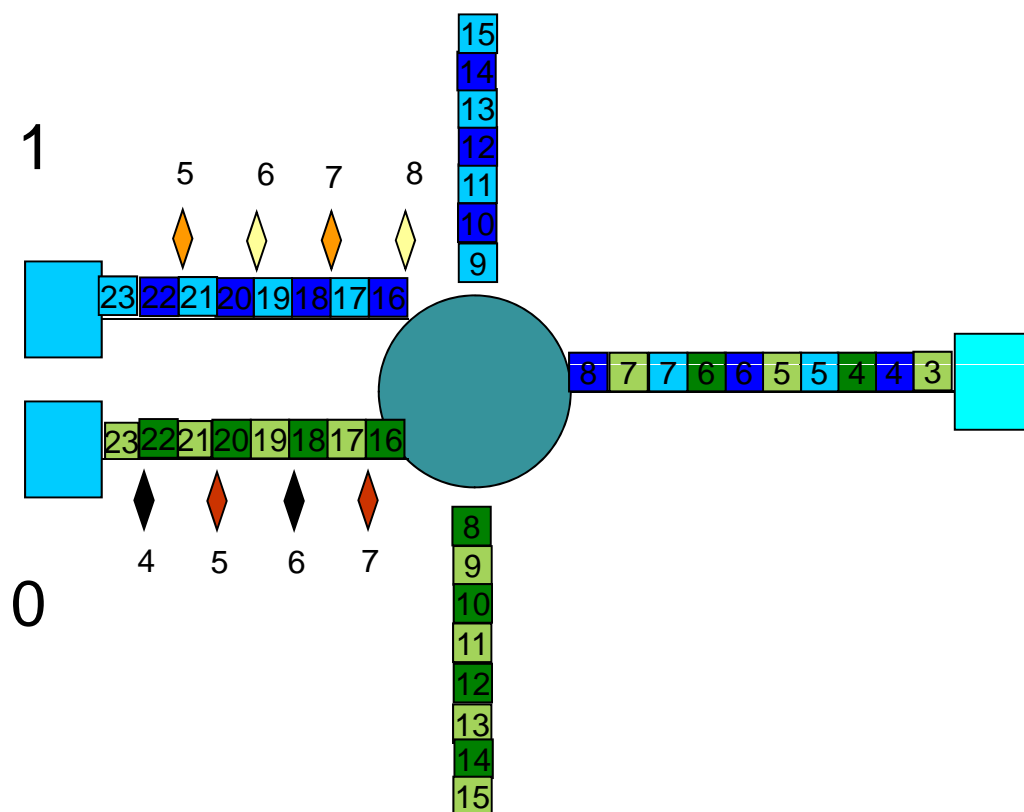


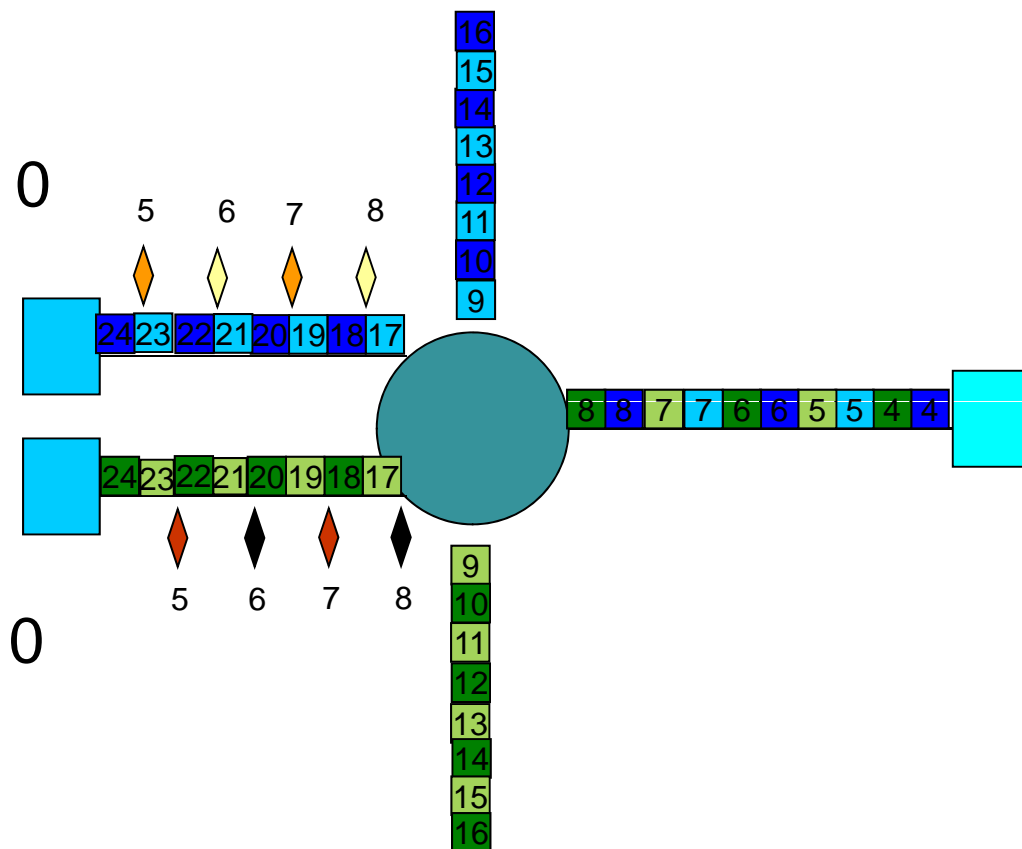


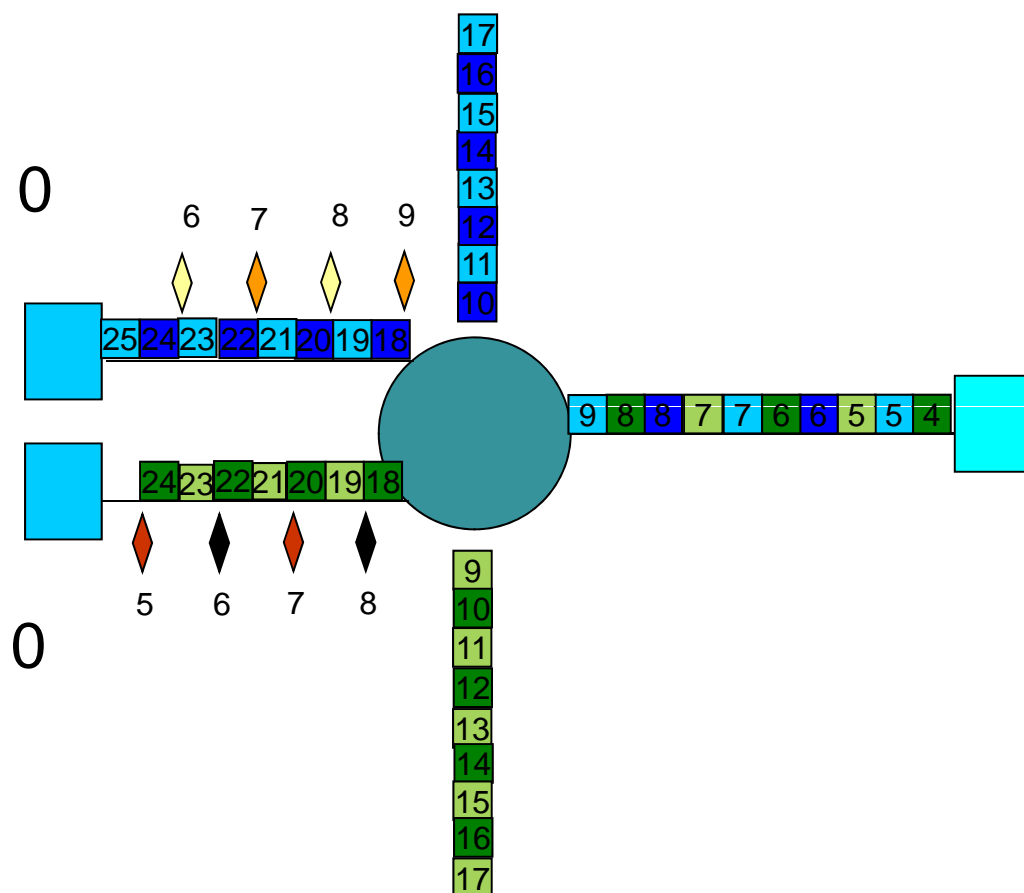


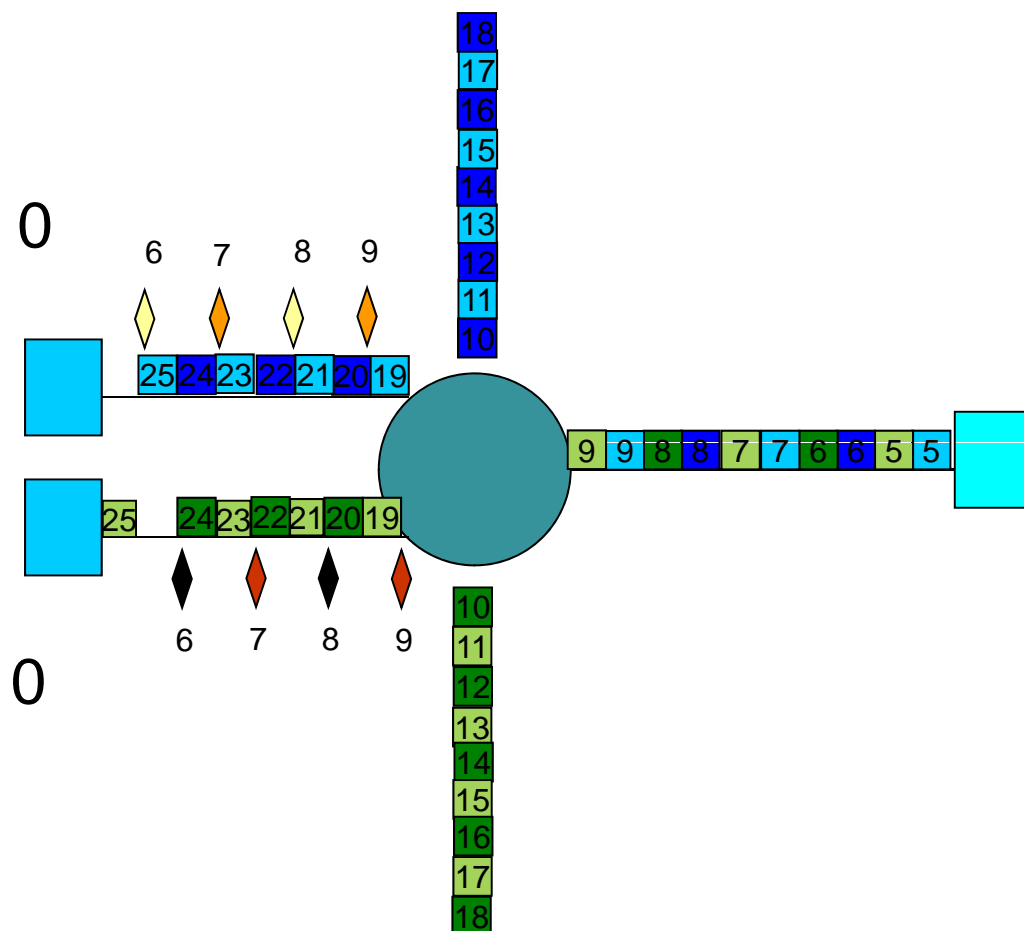


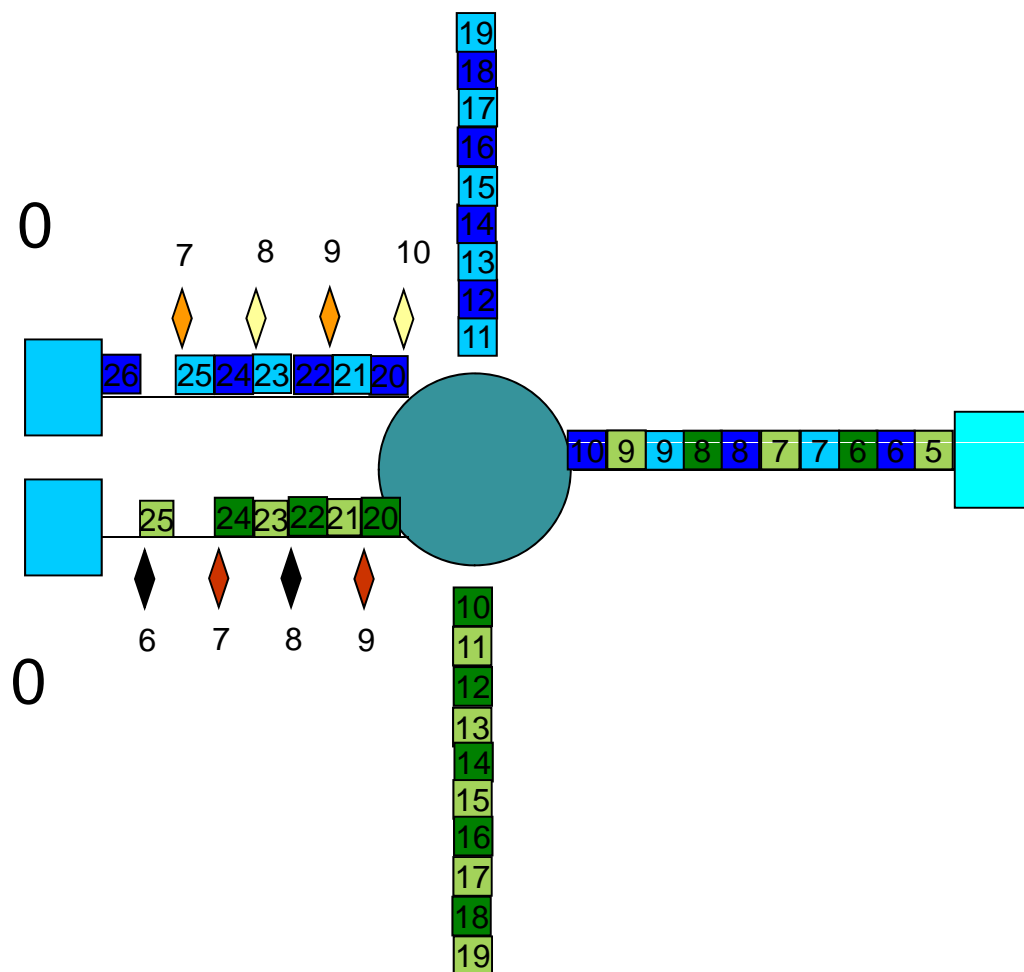


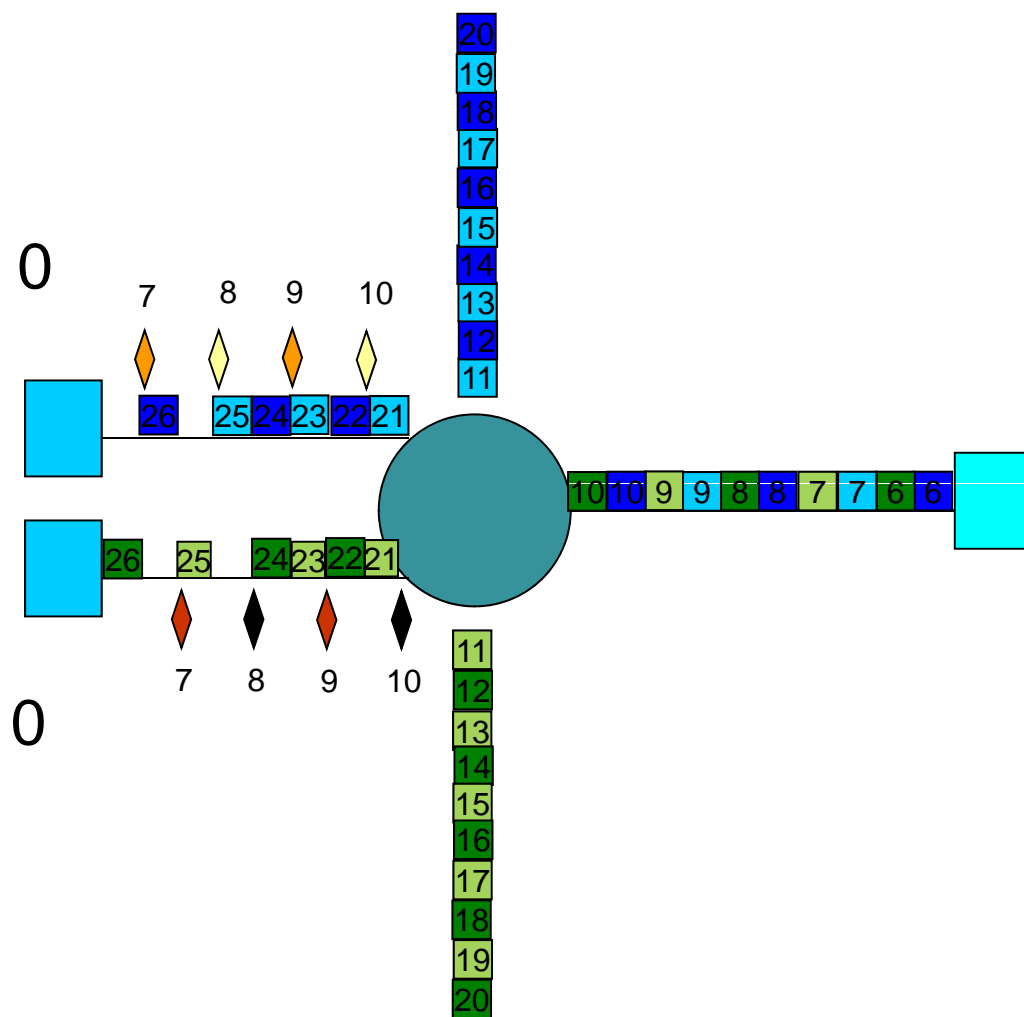


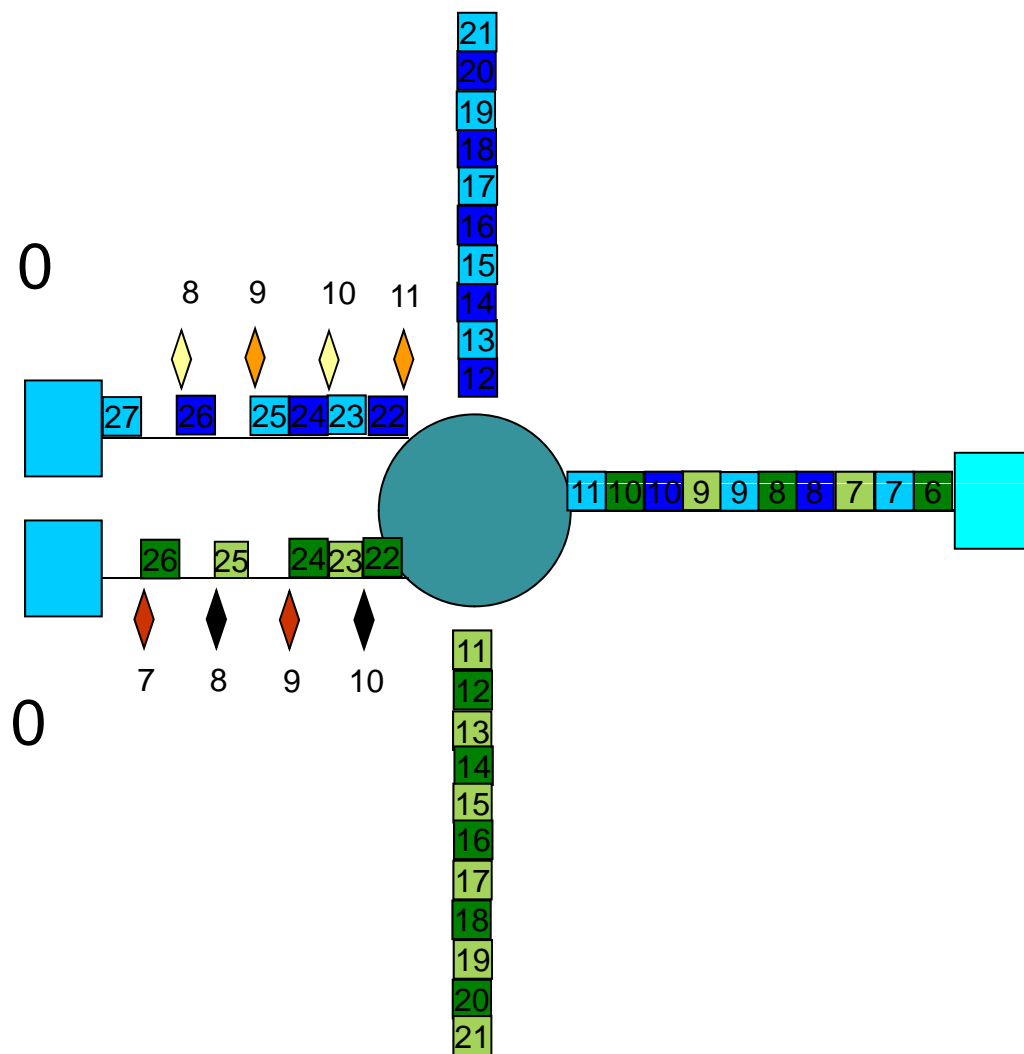


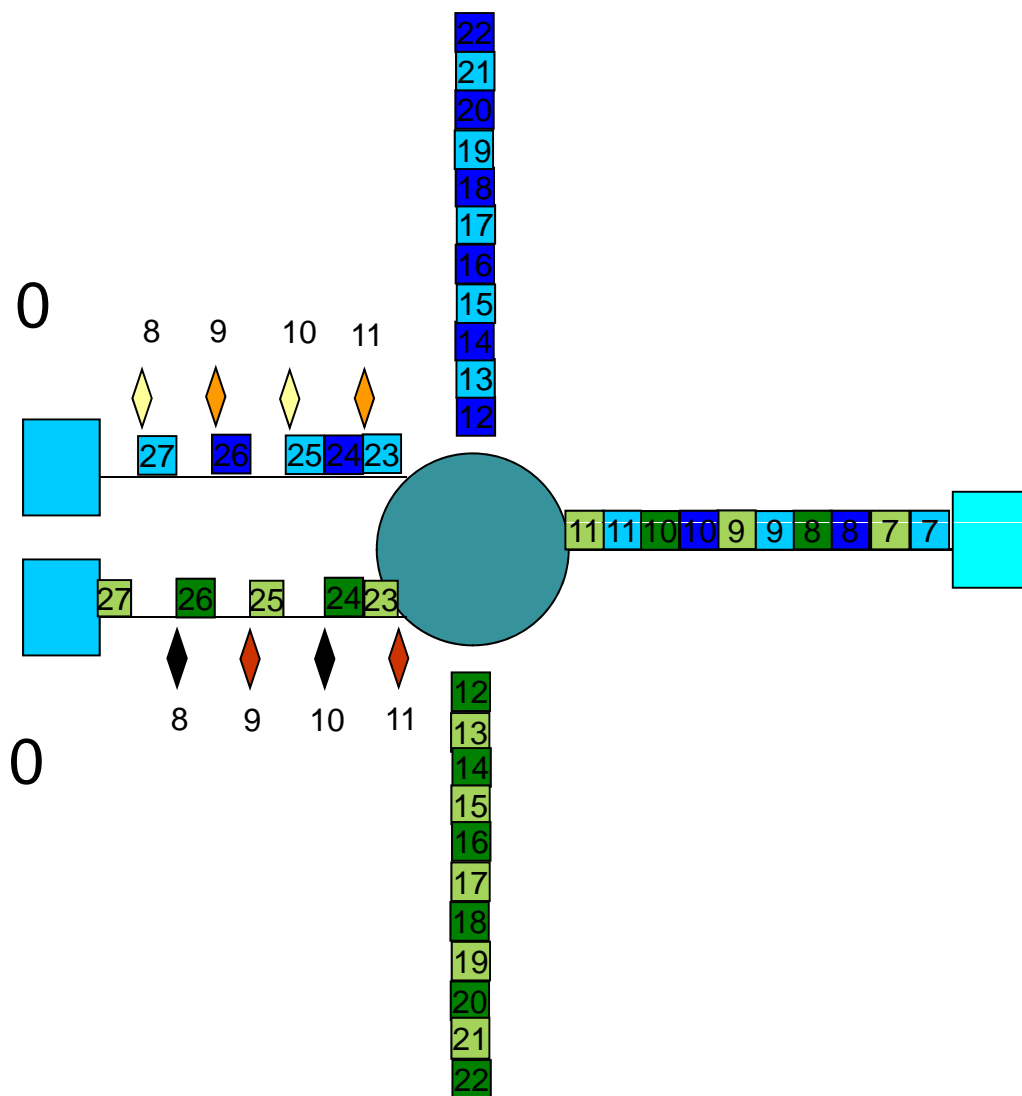


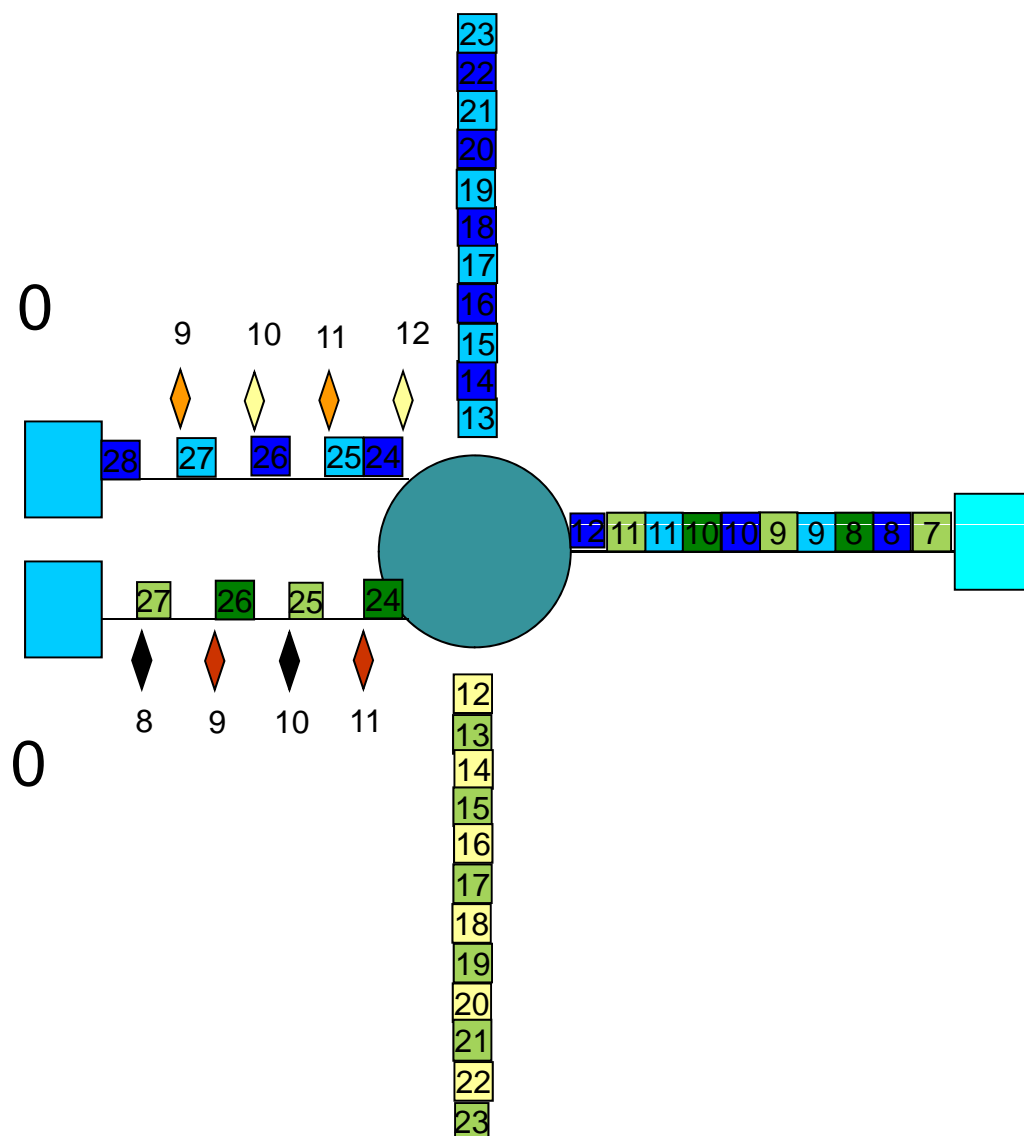


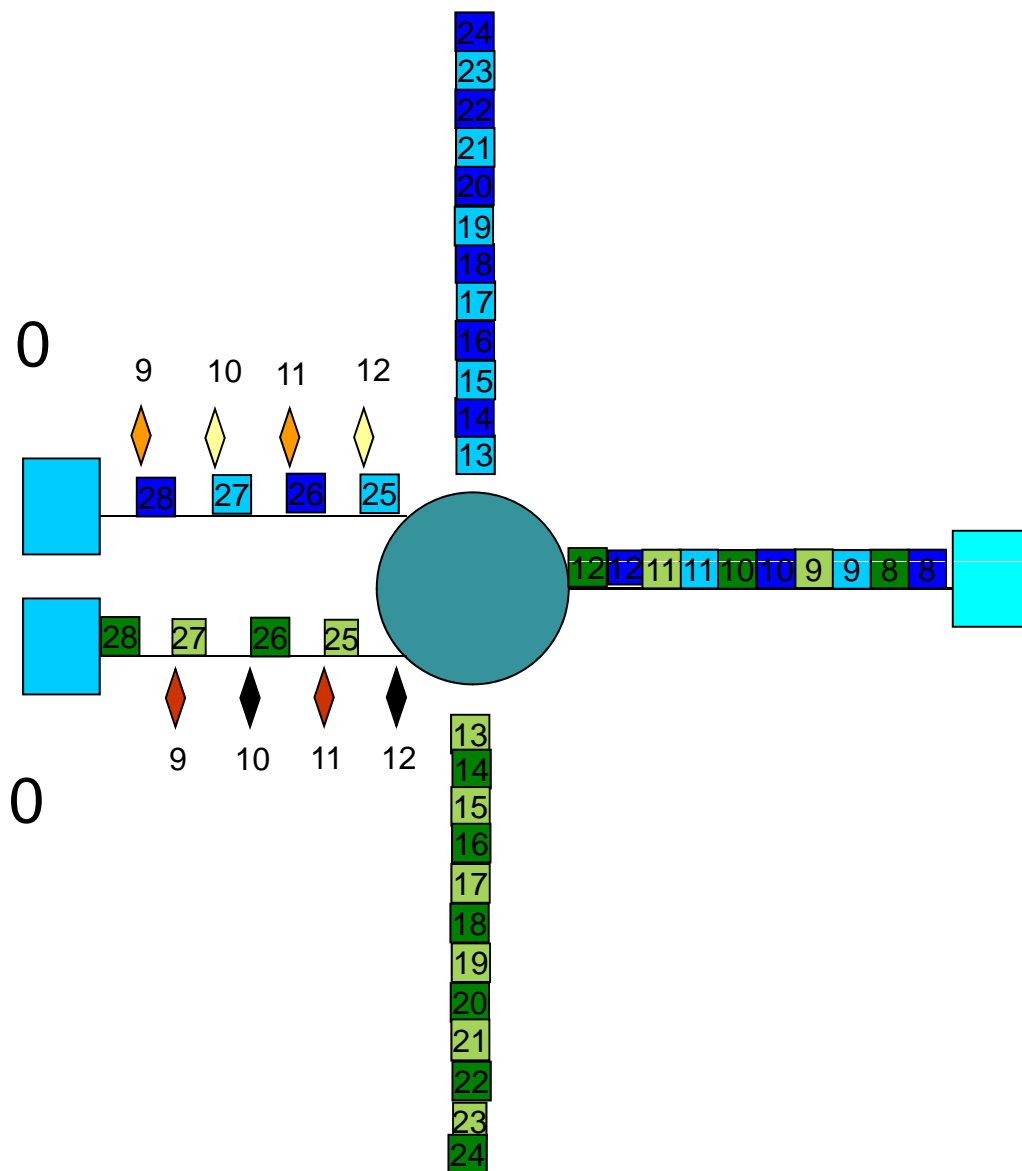


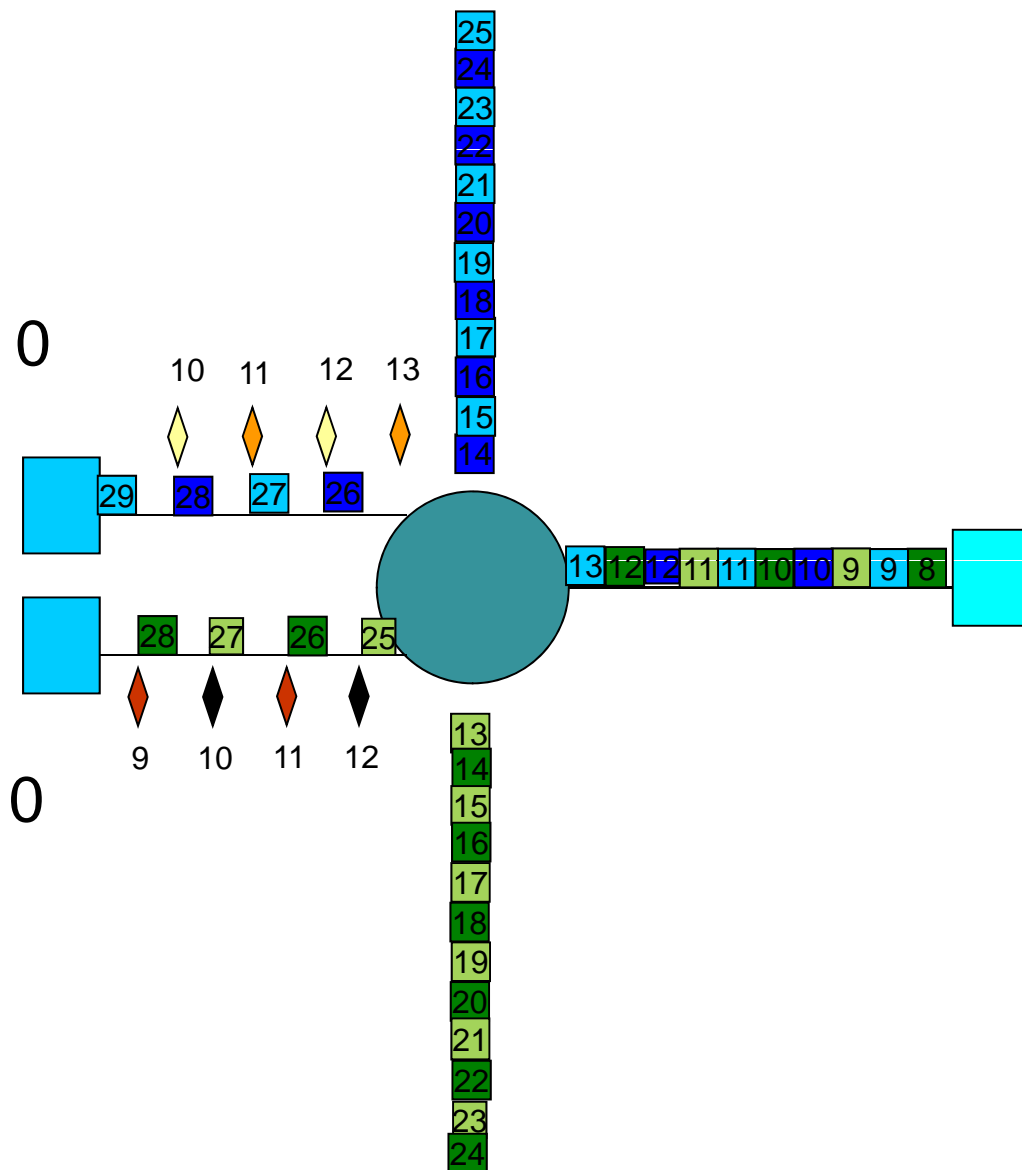


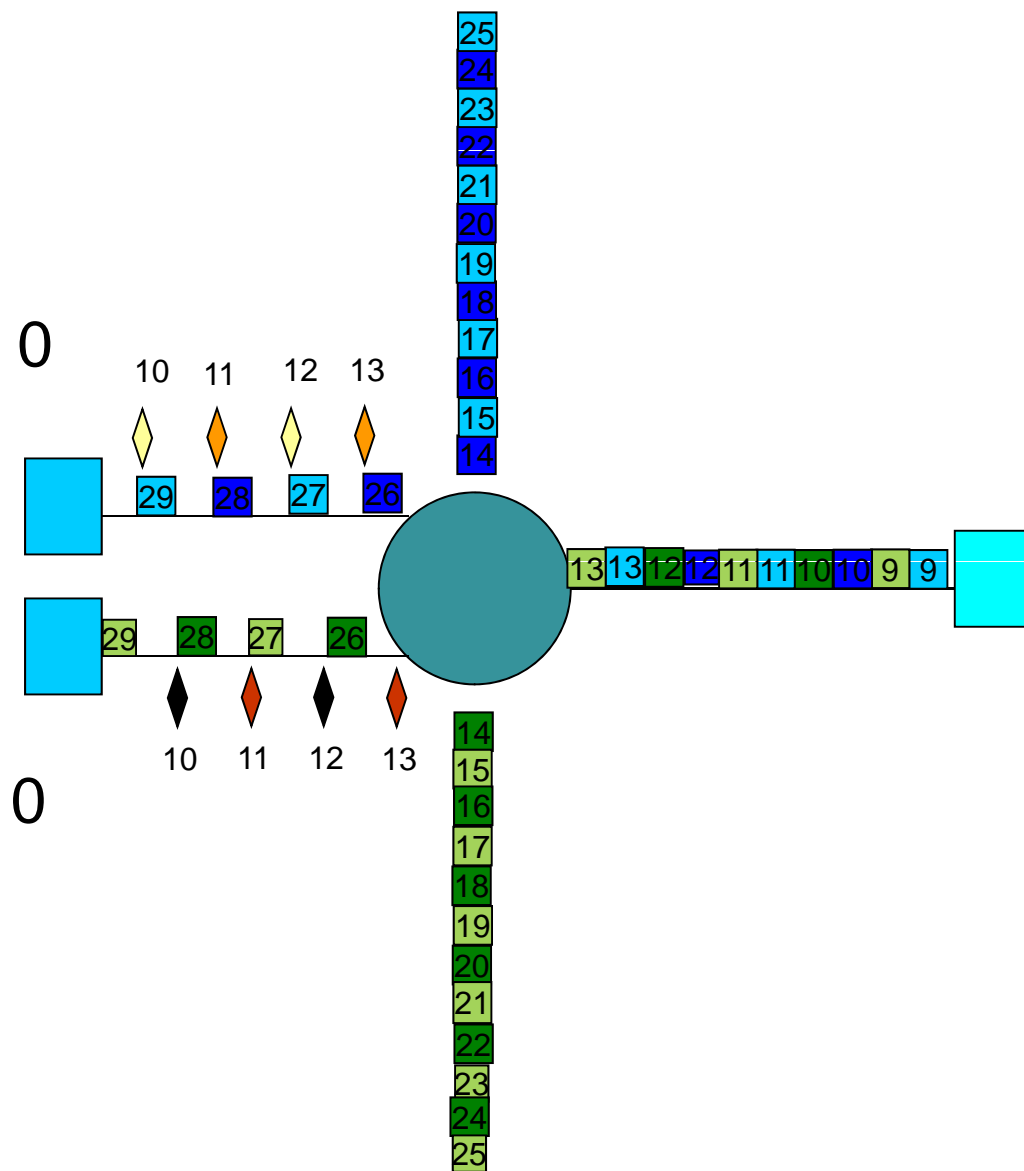


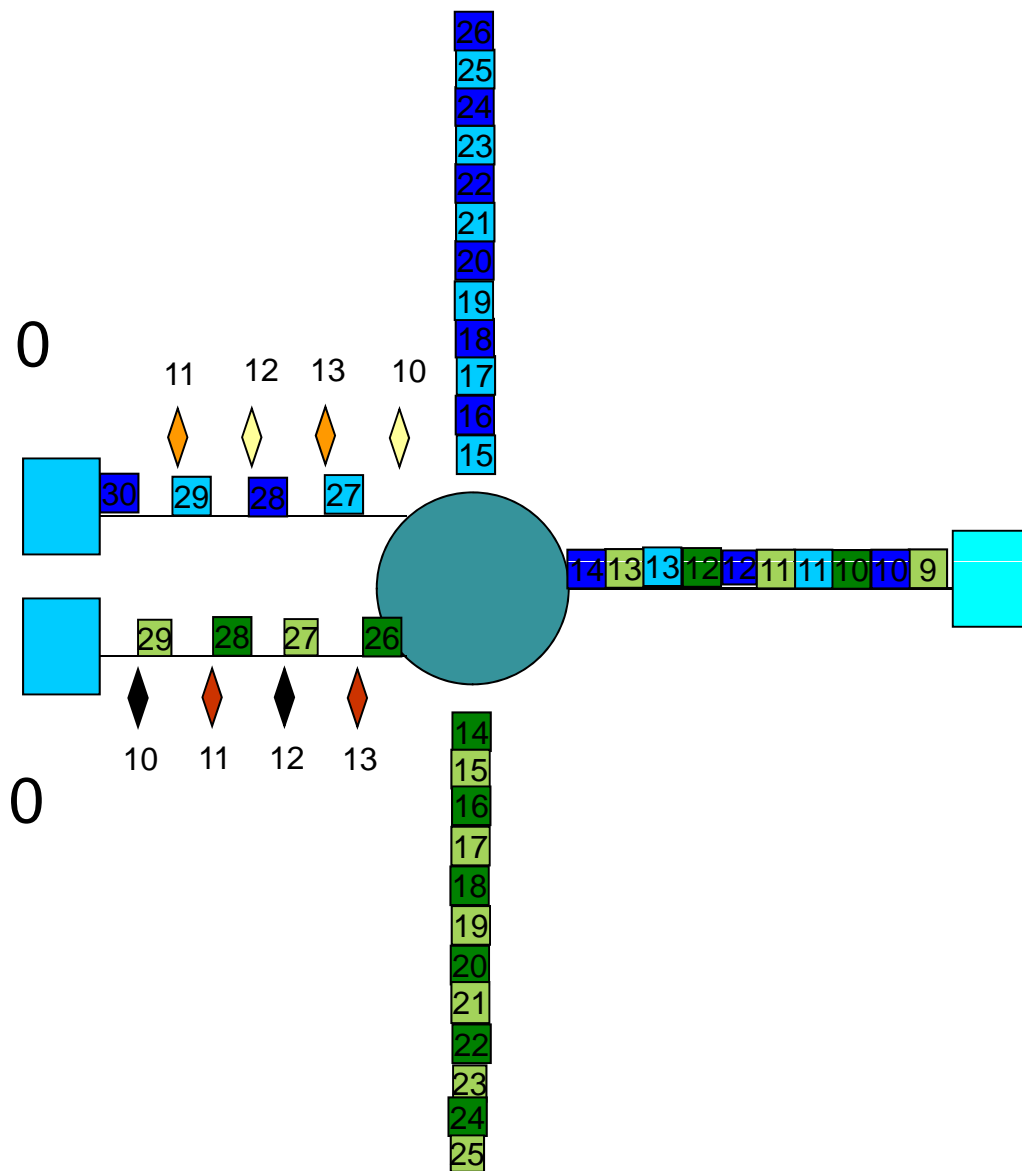












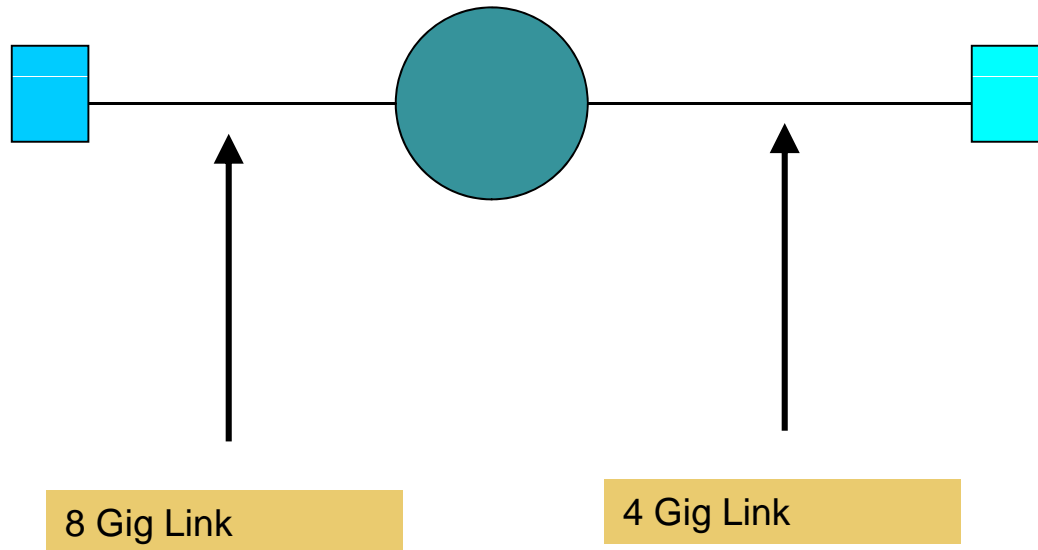
THIS PAGE INTENTIONALLY
LEFT BLANK

Example: Different sized pipes

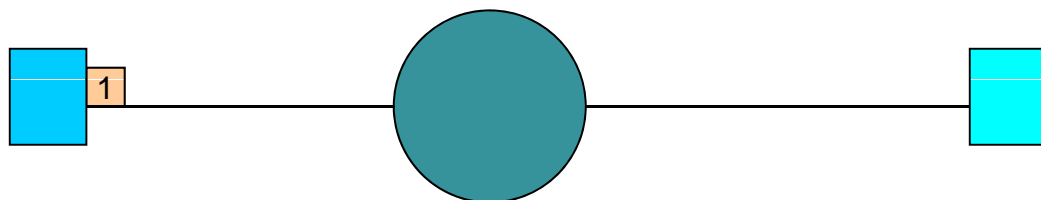
BUFFER CREDITS

Fat Pipe / Skinny Pipe

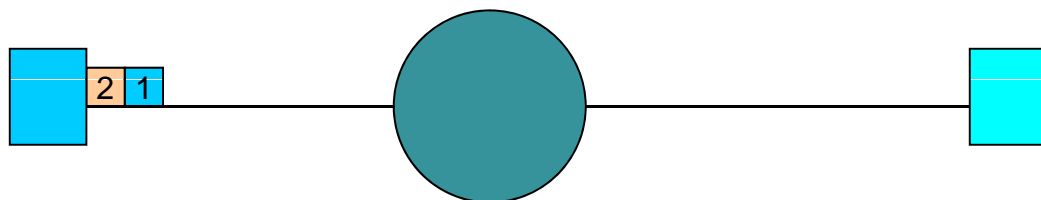
20



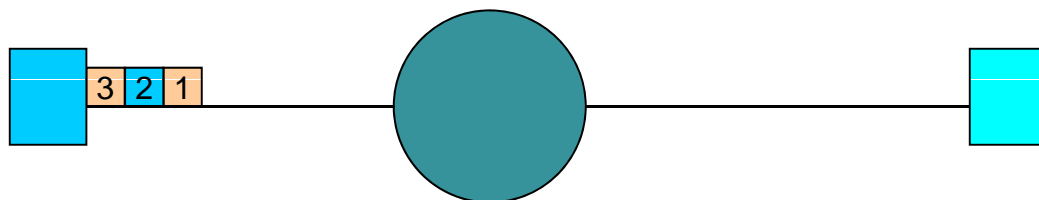
19



18

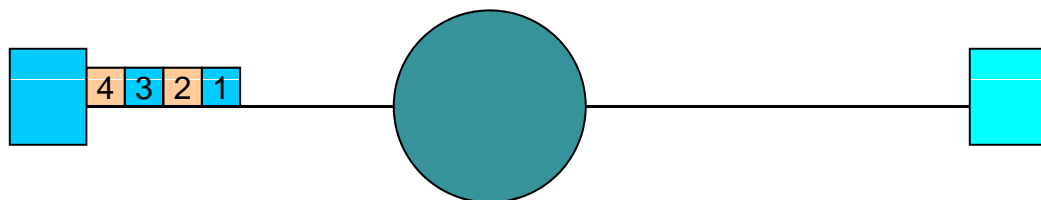


17

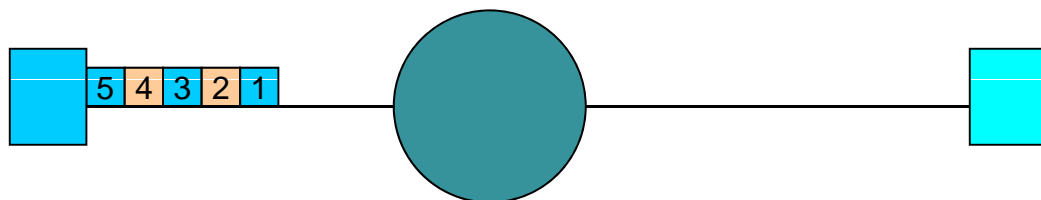


SHARE in Boston

16

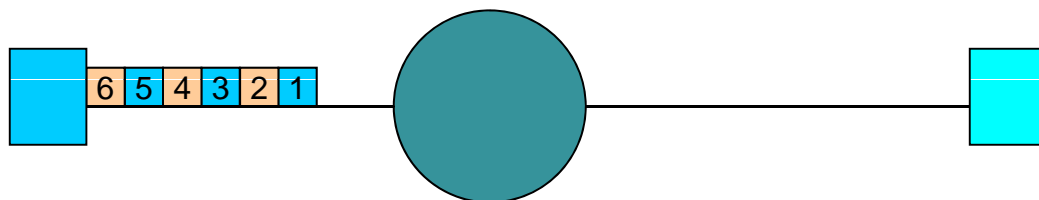


15



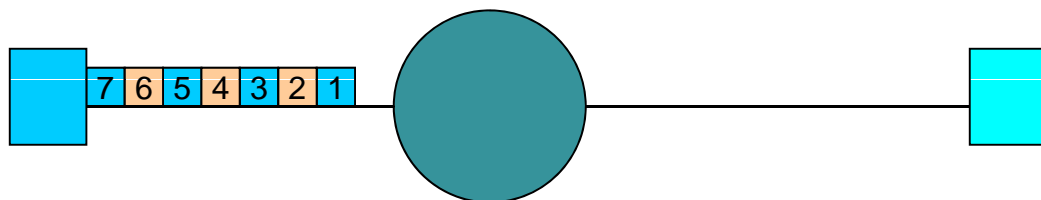
SHARE in Boston

14



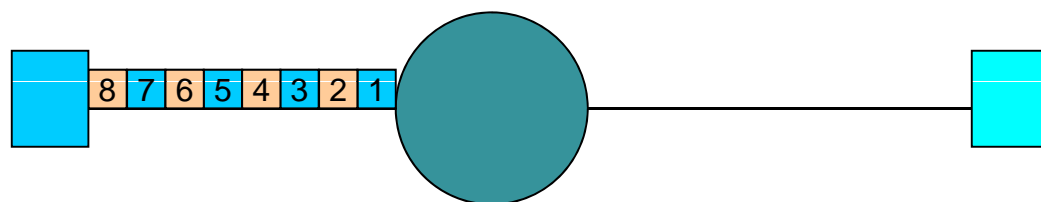
SHARE in Boston

13

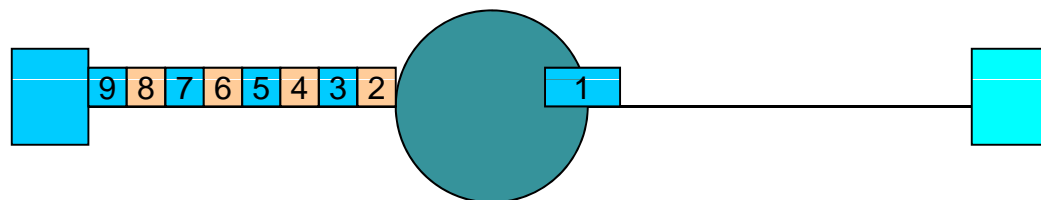


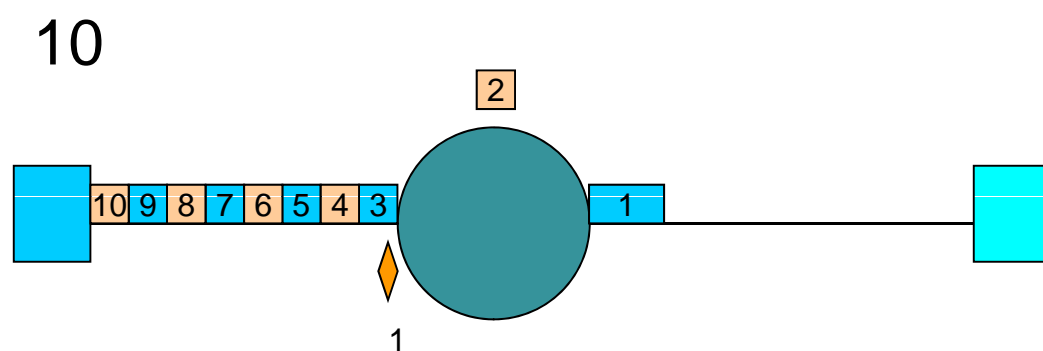
SHARE in Boston

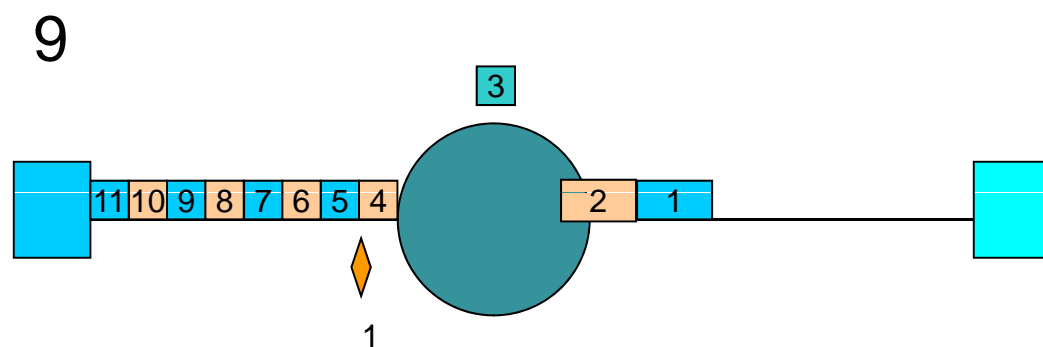
12

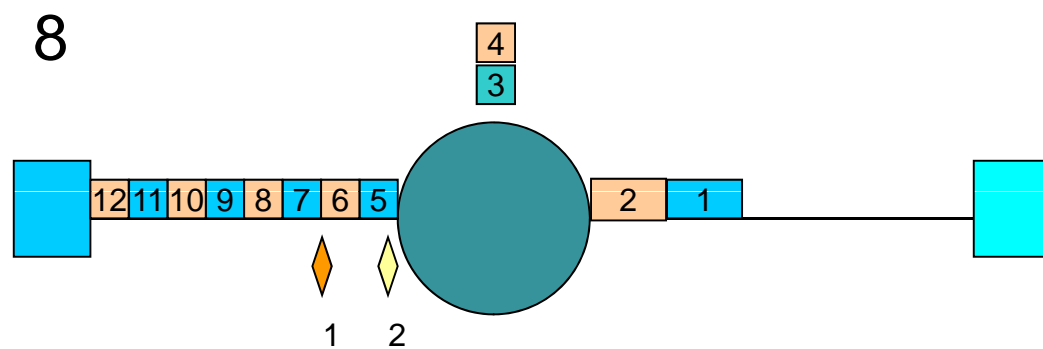


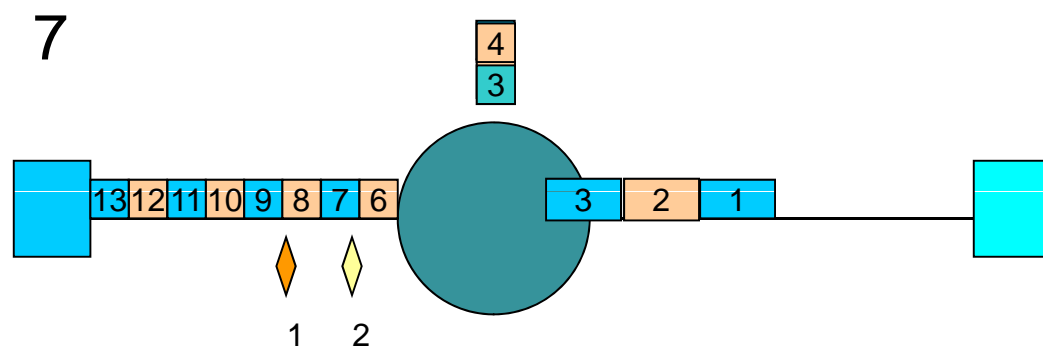
11

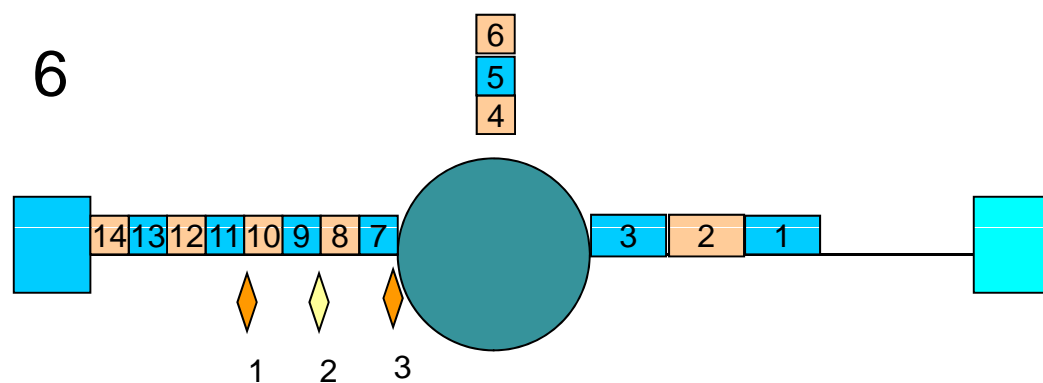


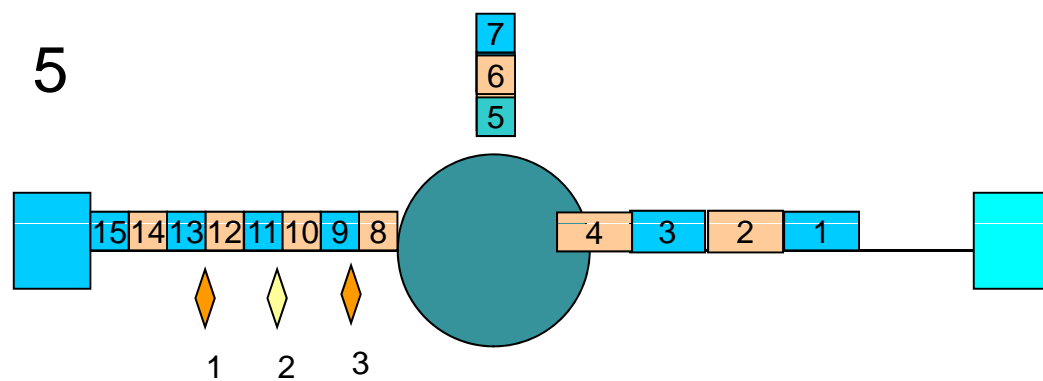


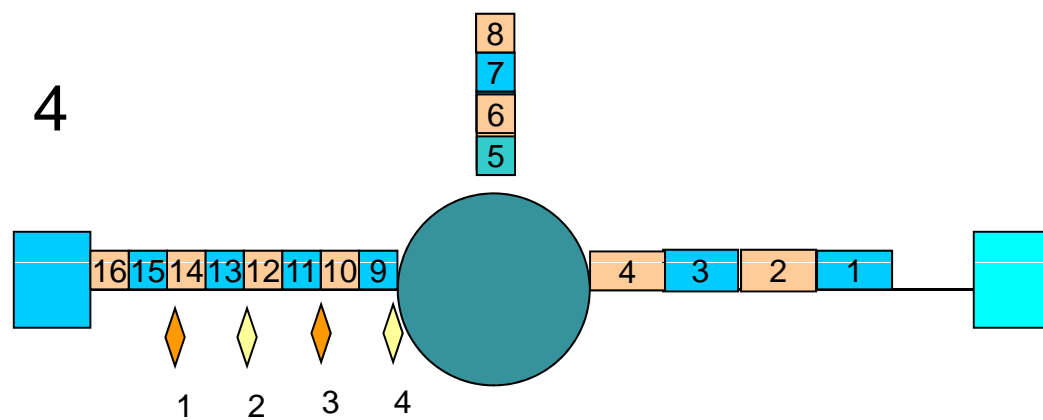


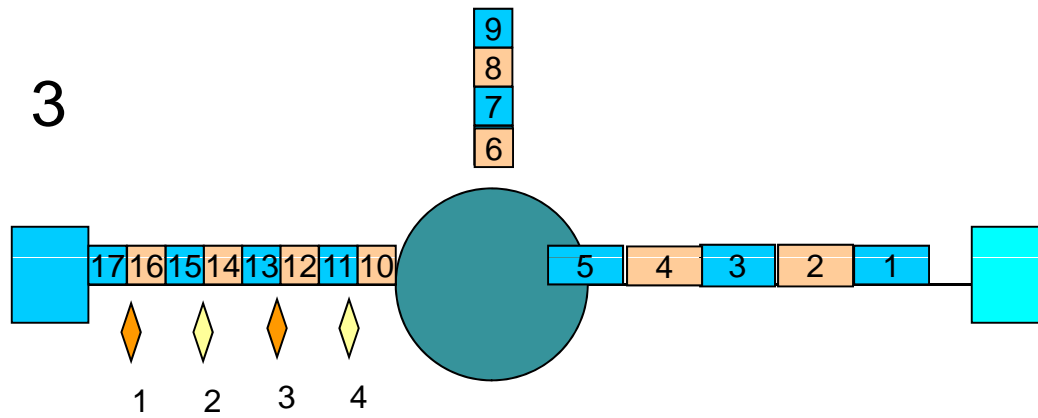


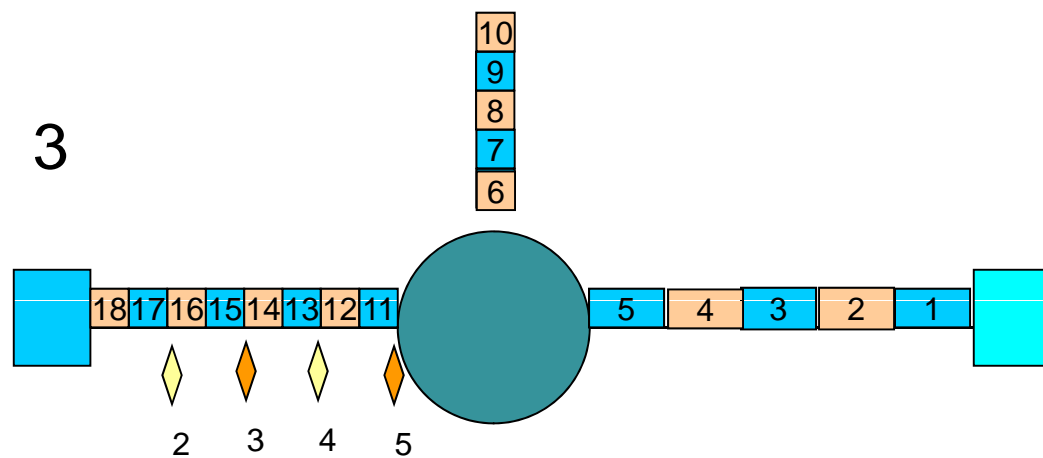


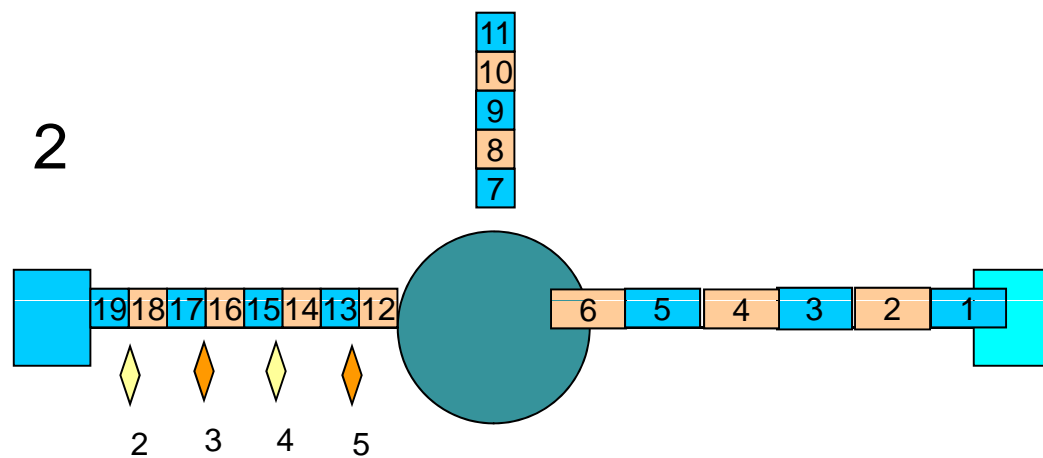


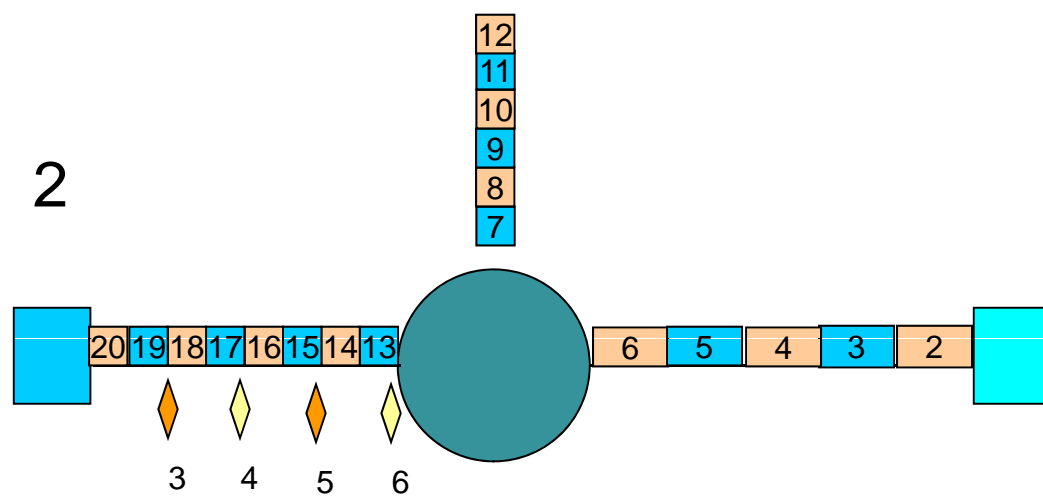


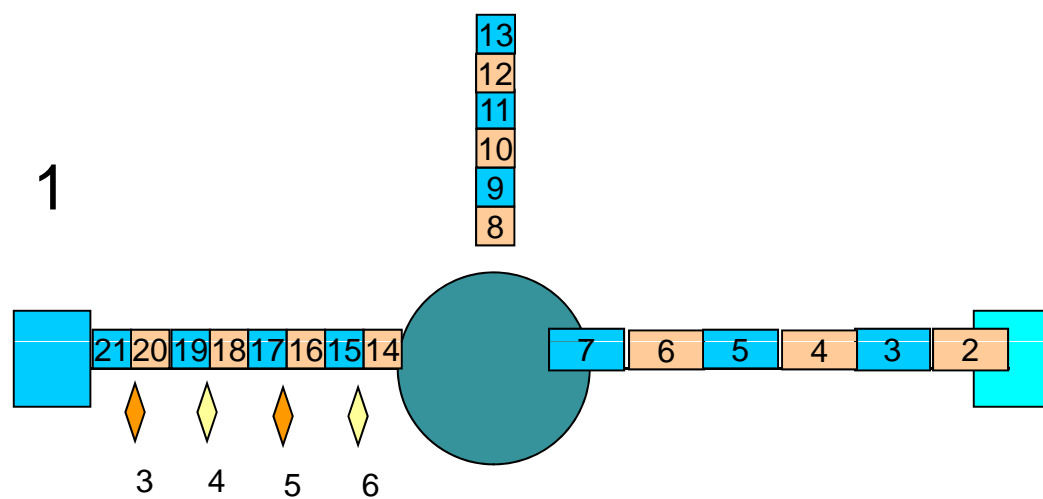


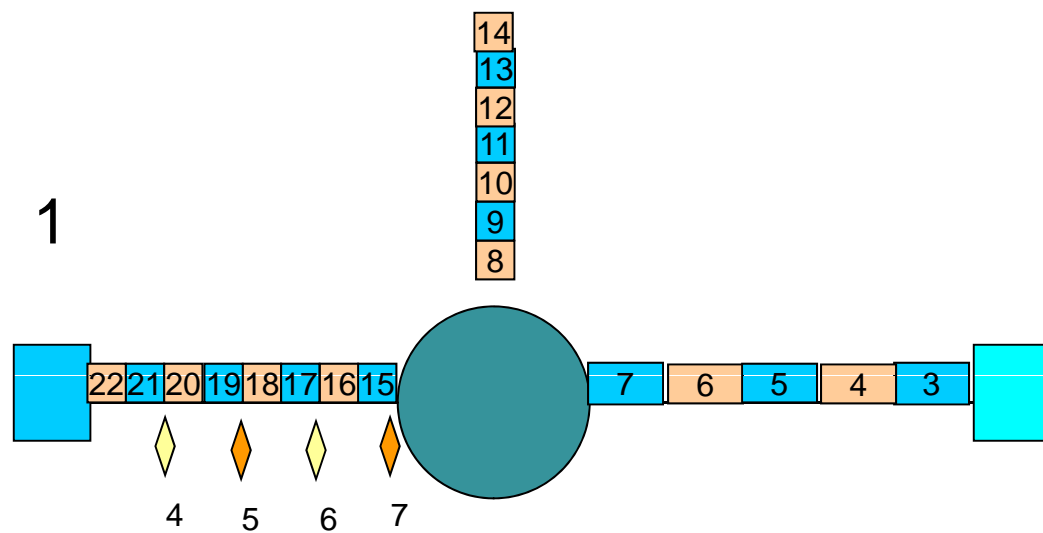


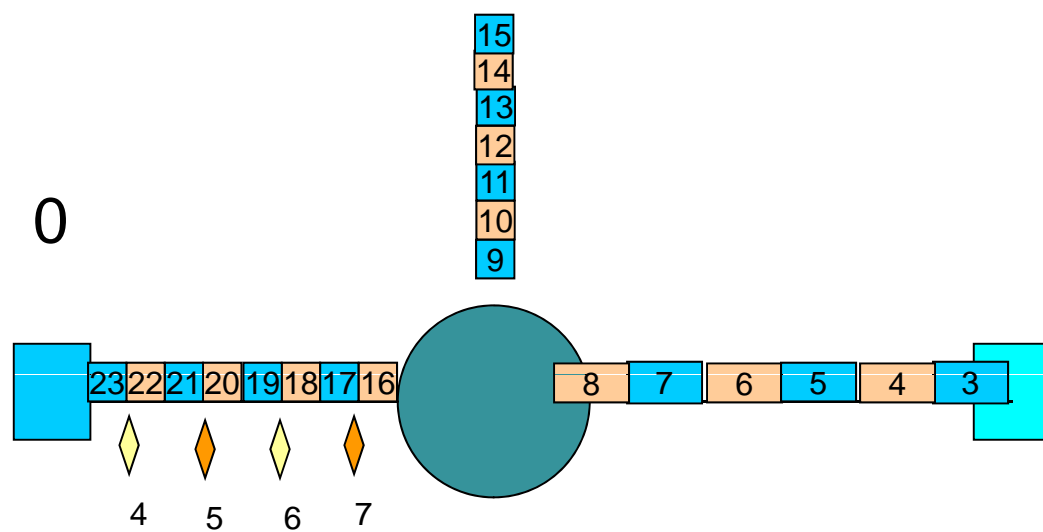


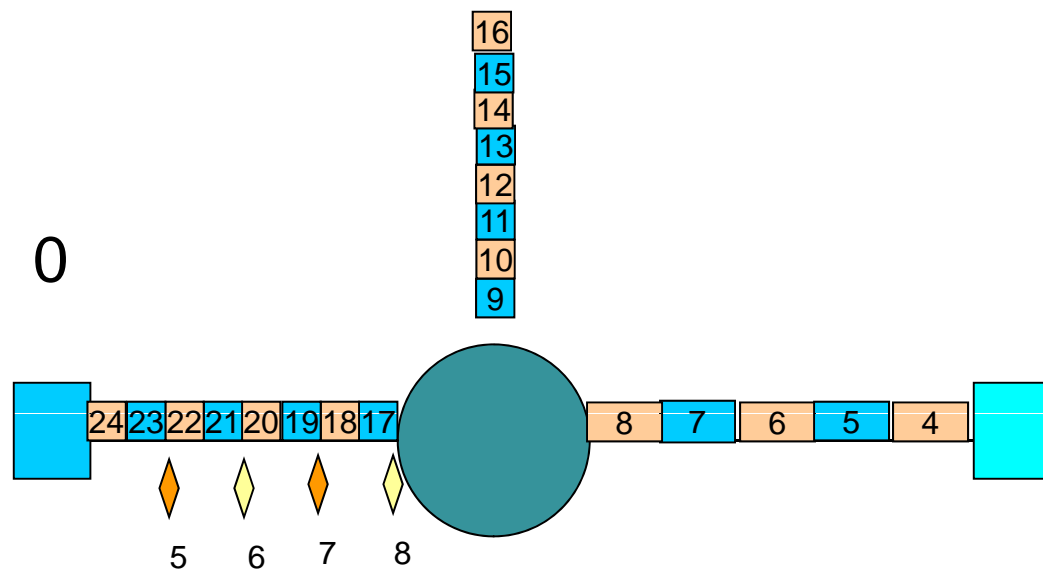


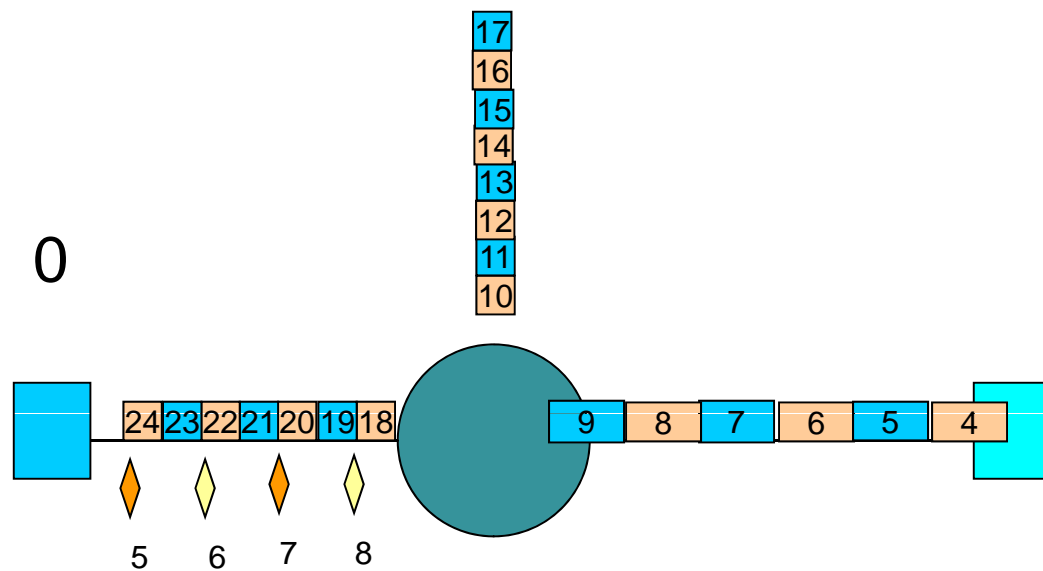


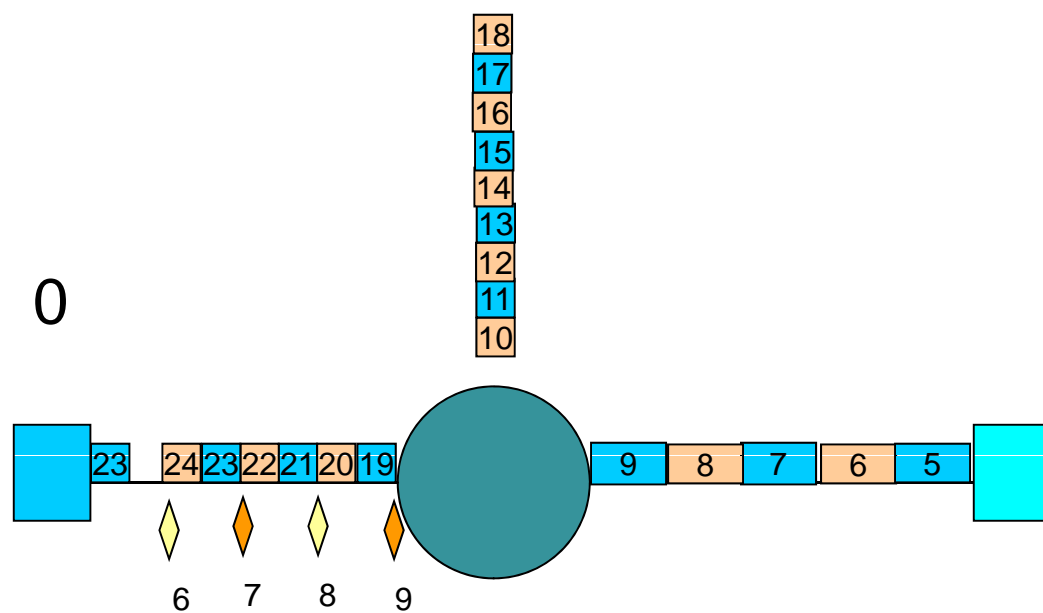


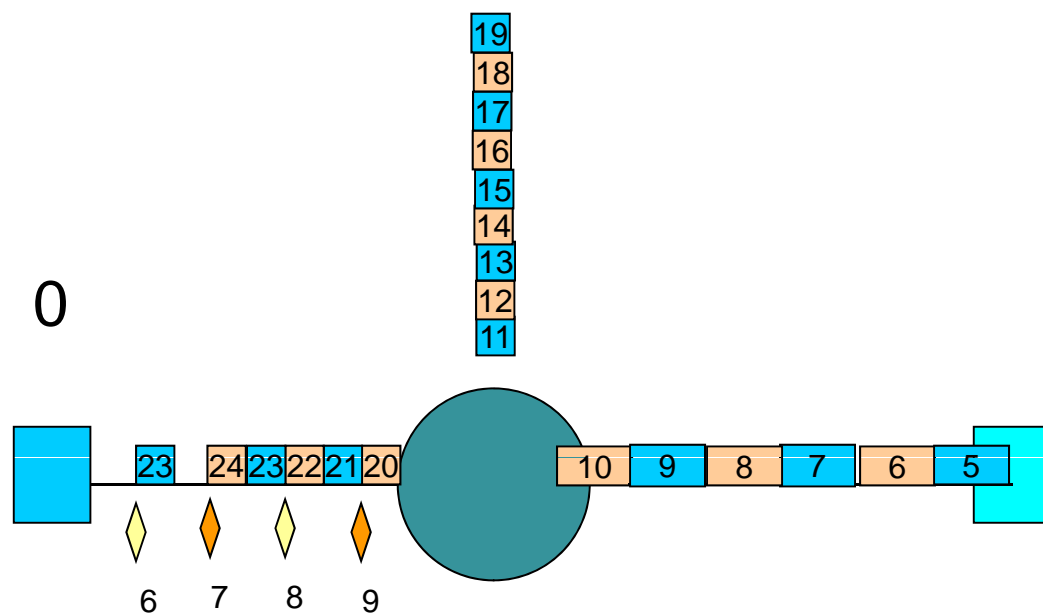


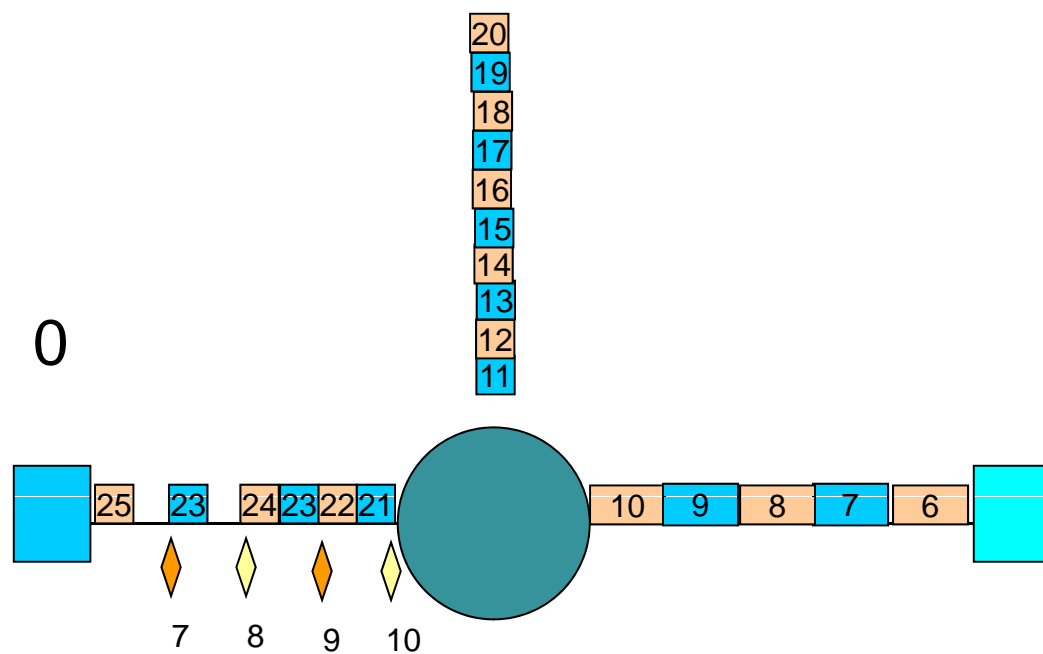


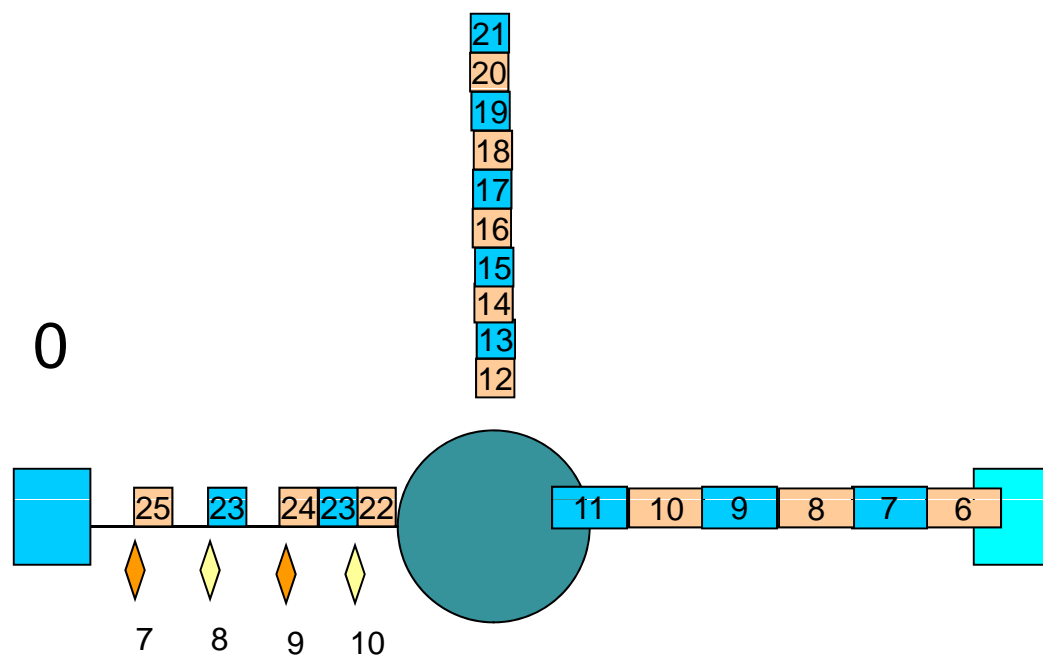


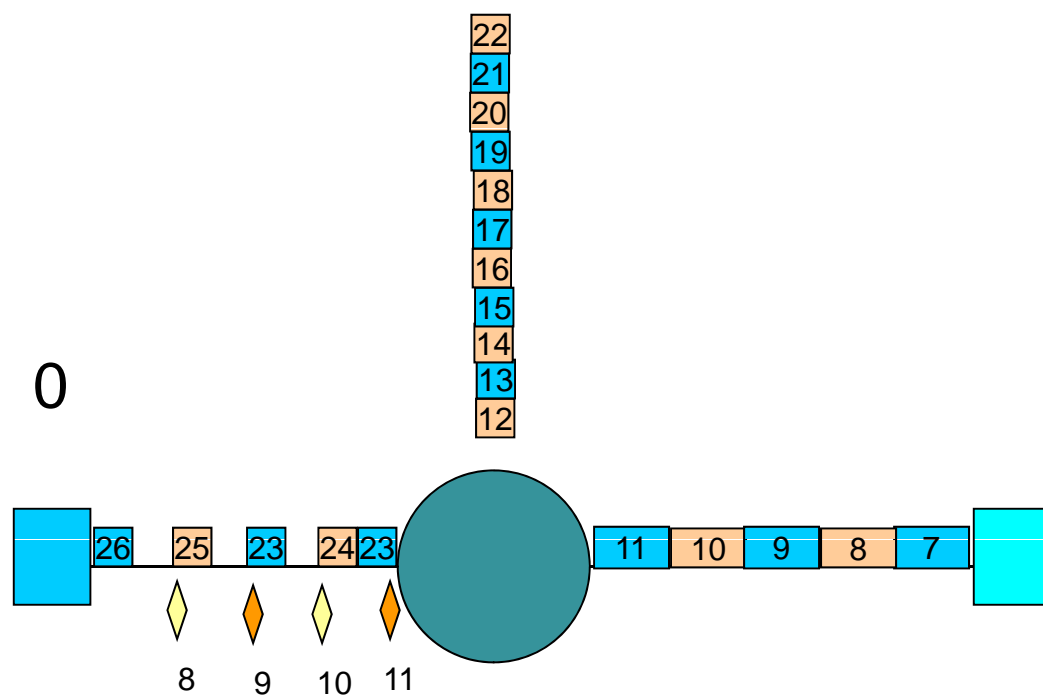


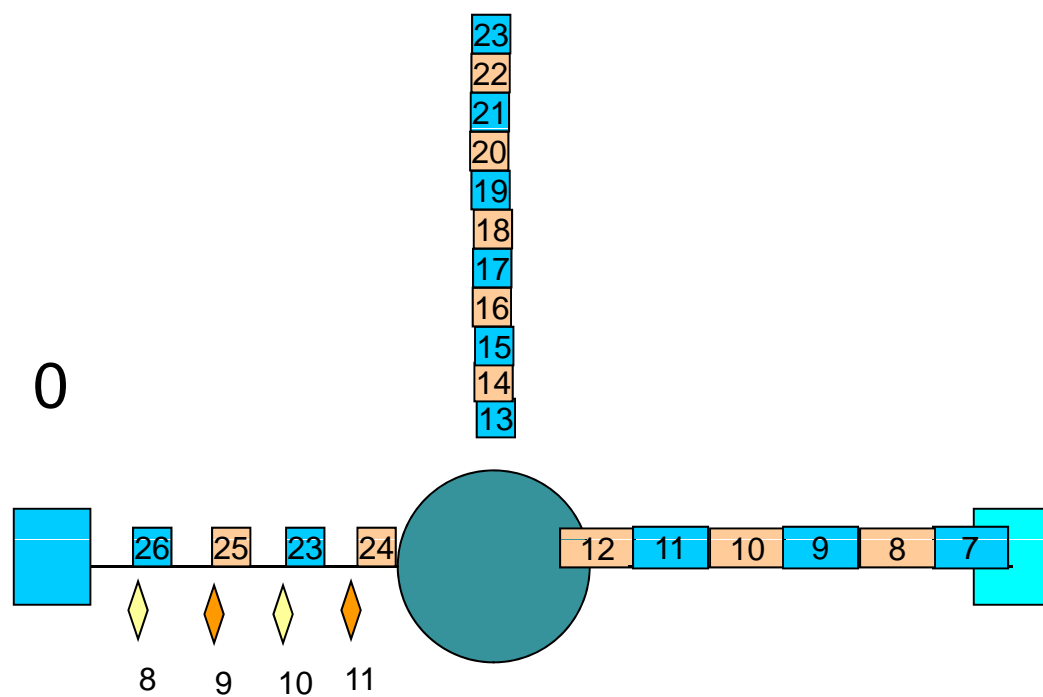


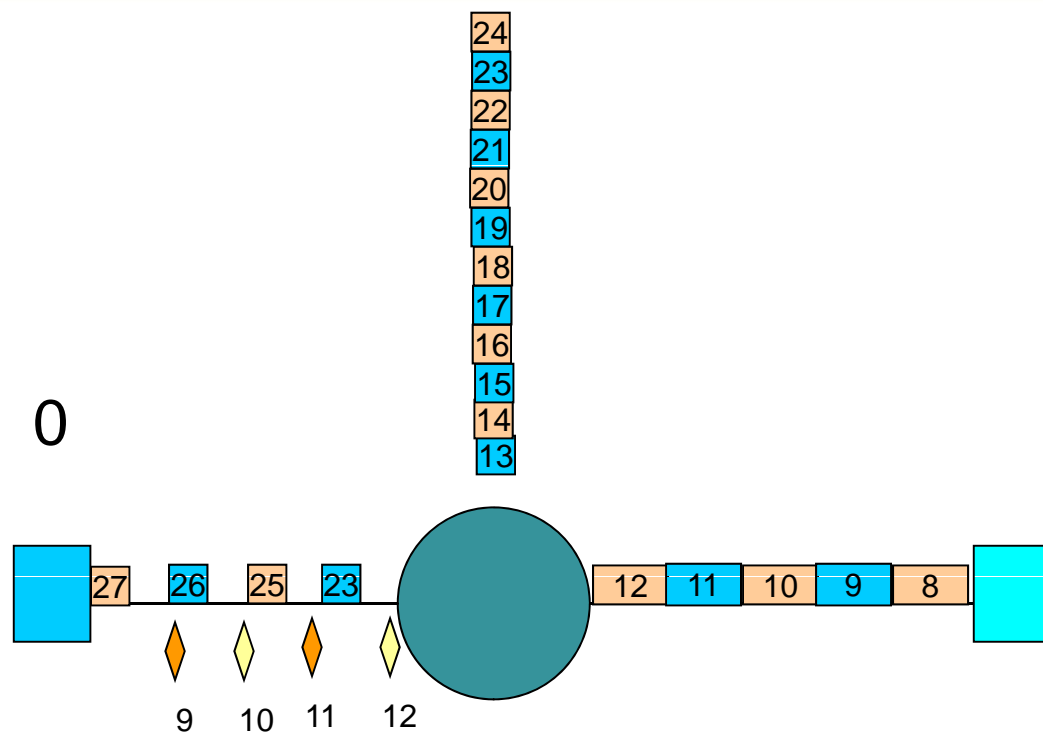


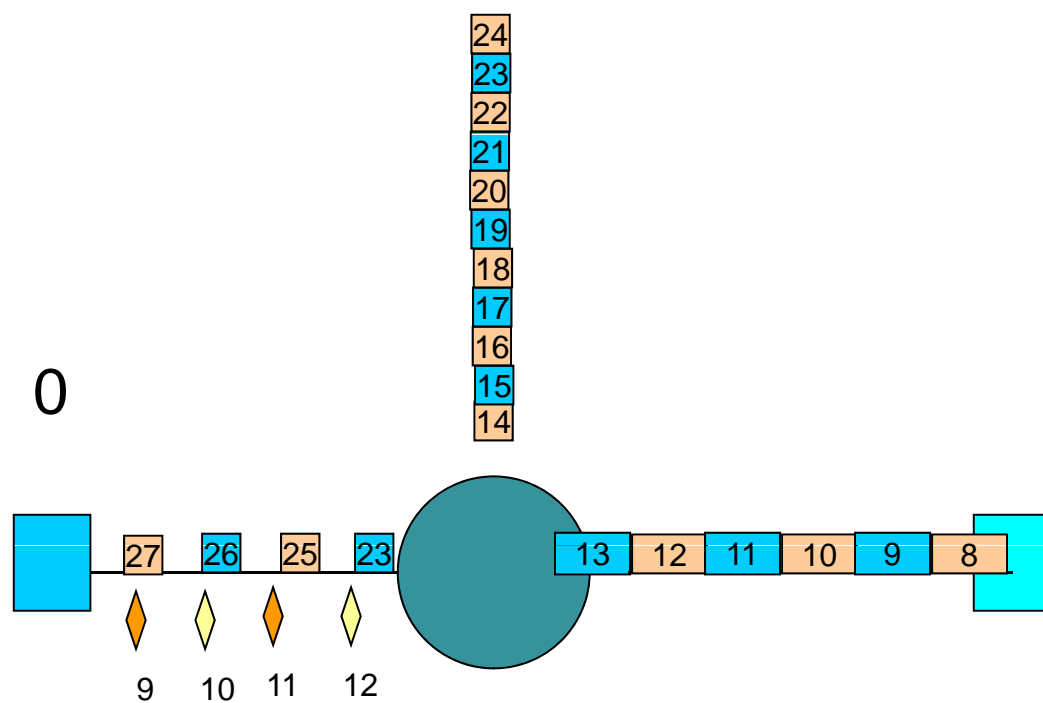




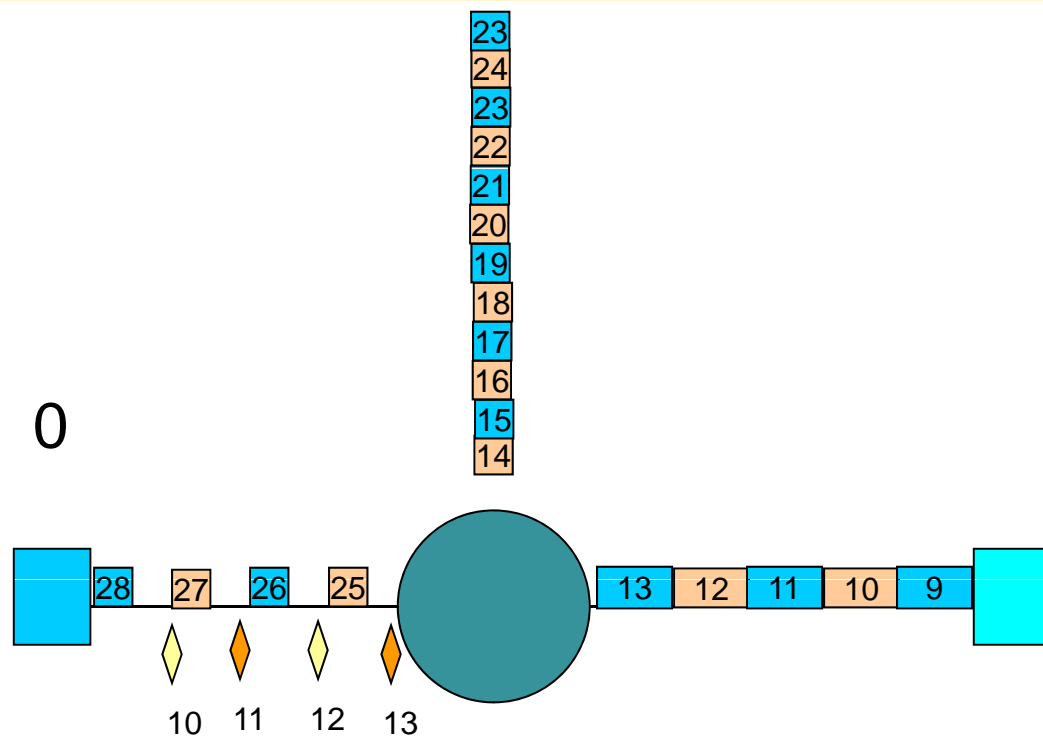


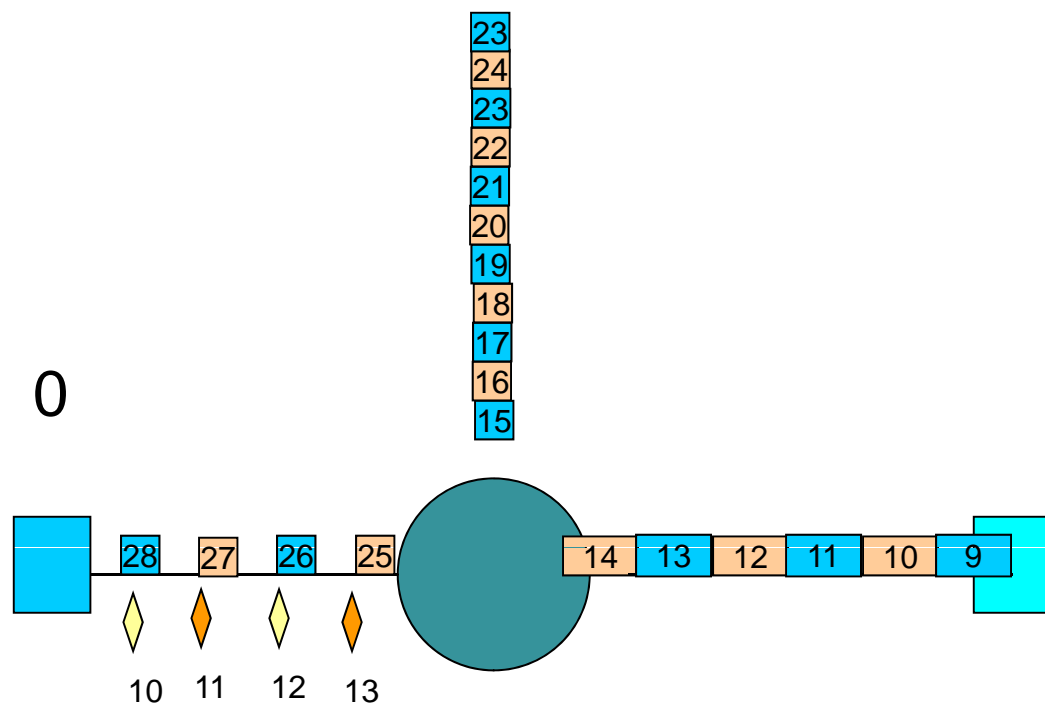




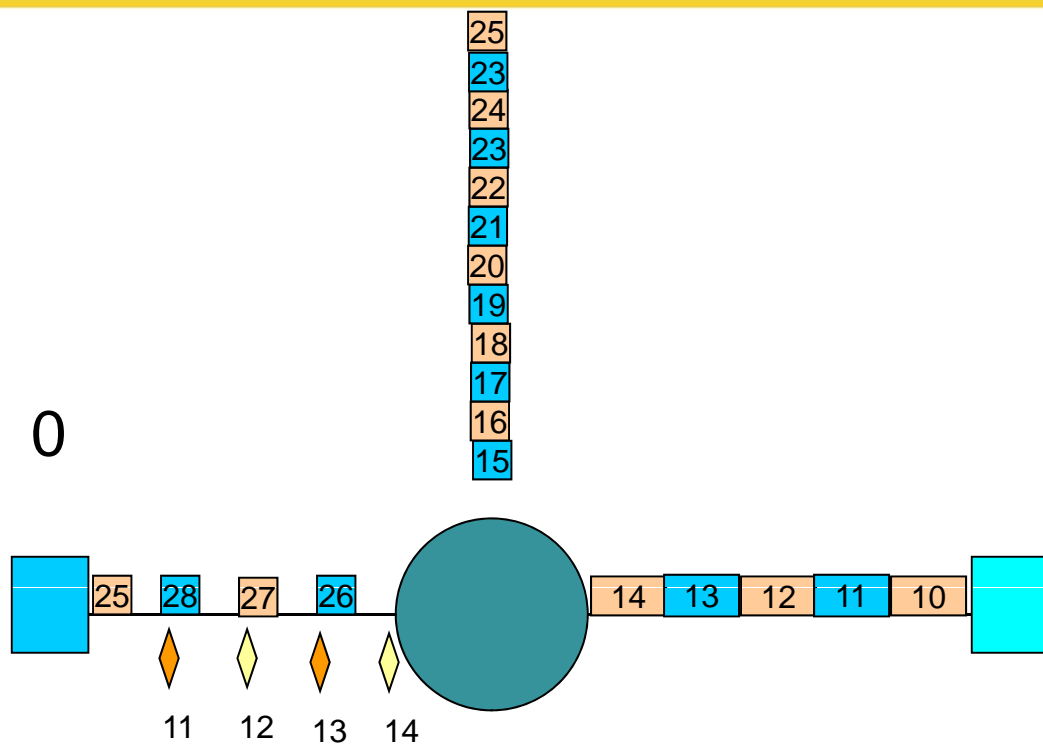


0





0



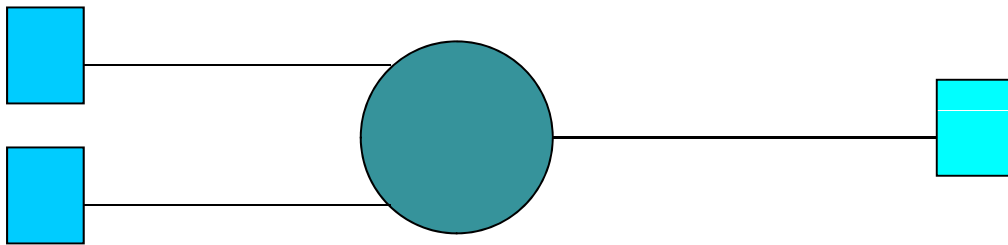
THIS PAGE INTENTIONALLY
LEFT BLANK

Example: Real Life?

BUFFER CREDITS

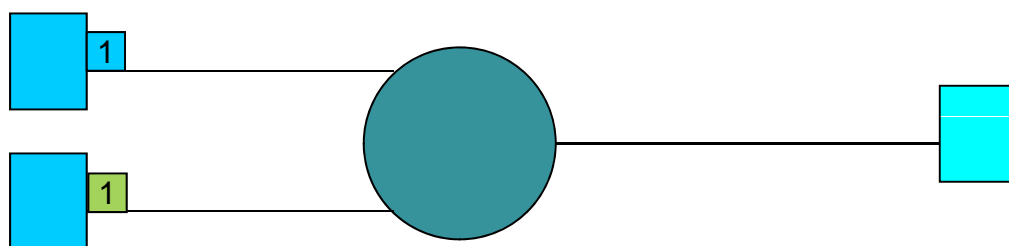
More real life example: Two senders sending at 30% - 50% link rate to one receiver

20



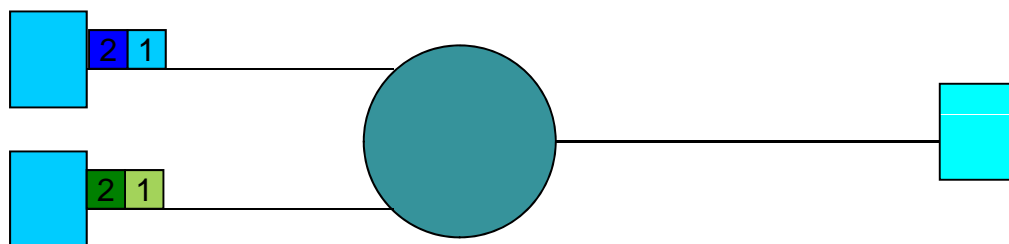
20

19



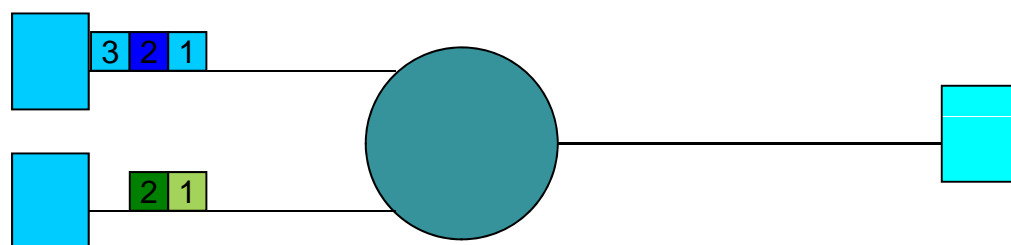
19

18



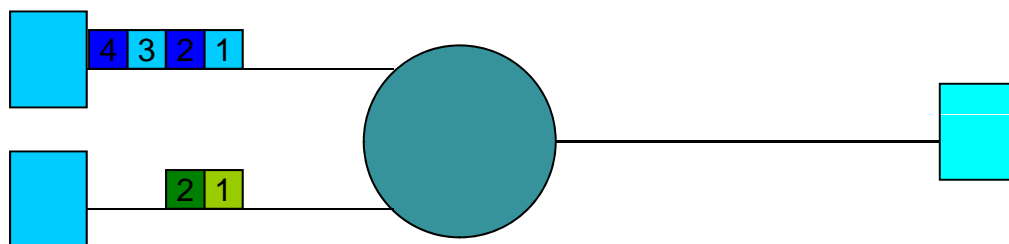
18

17



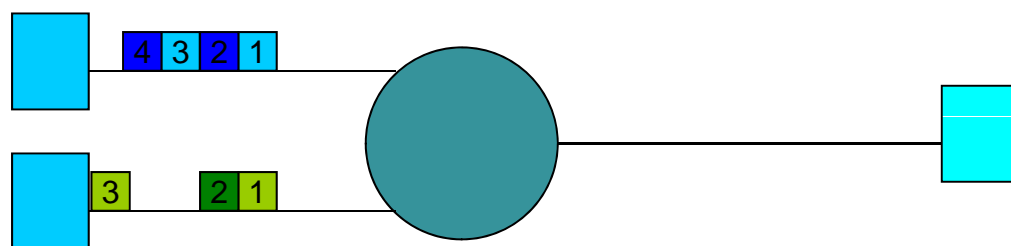
18

16



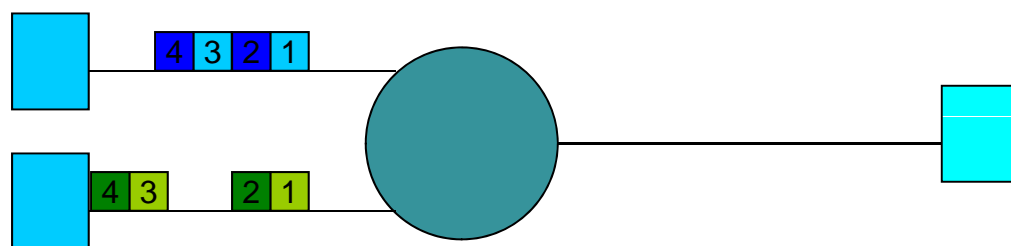
18

16



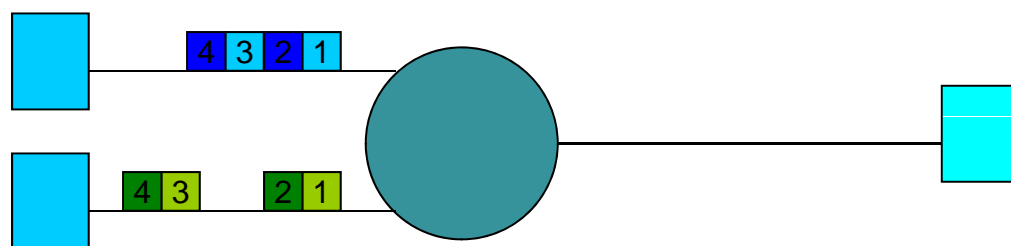
17

16



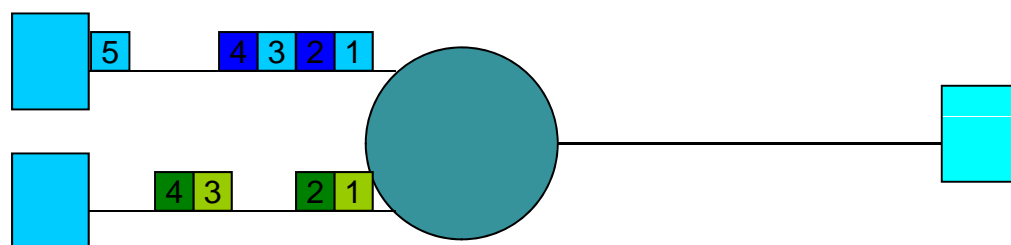
16

16



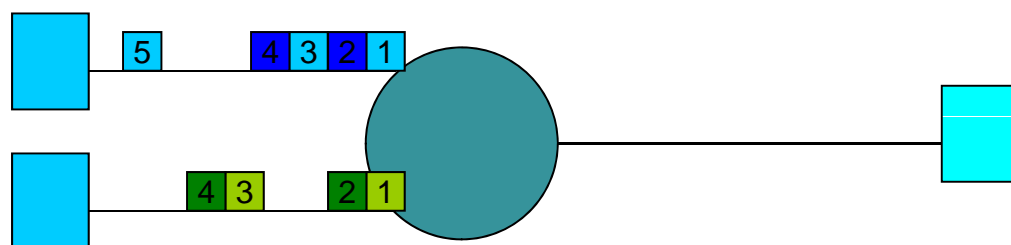
16

15

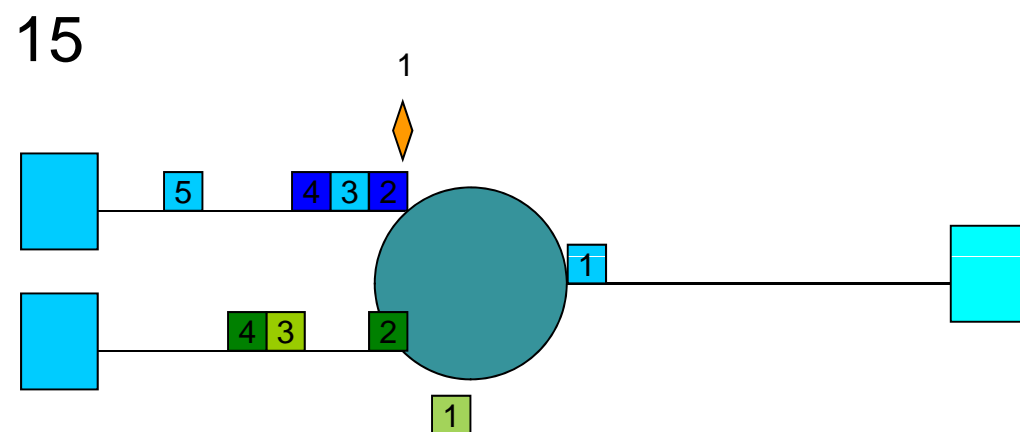


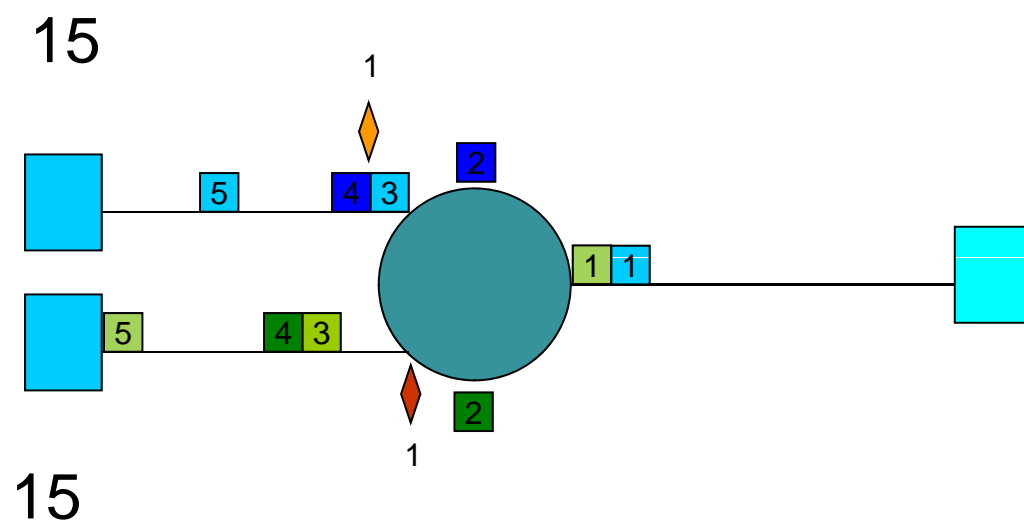
16

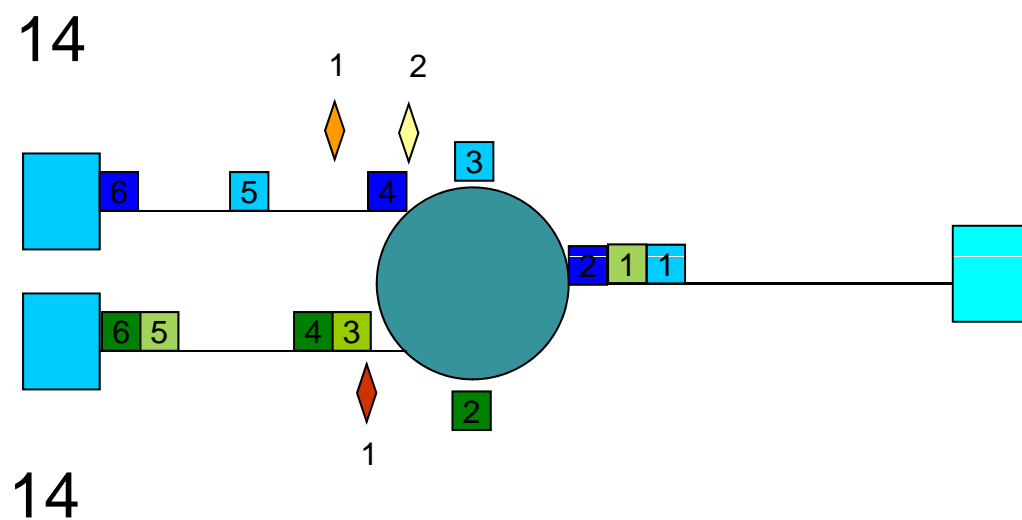
15

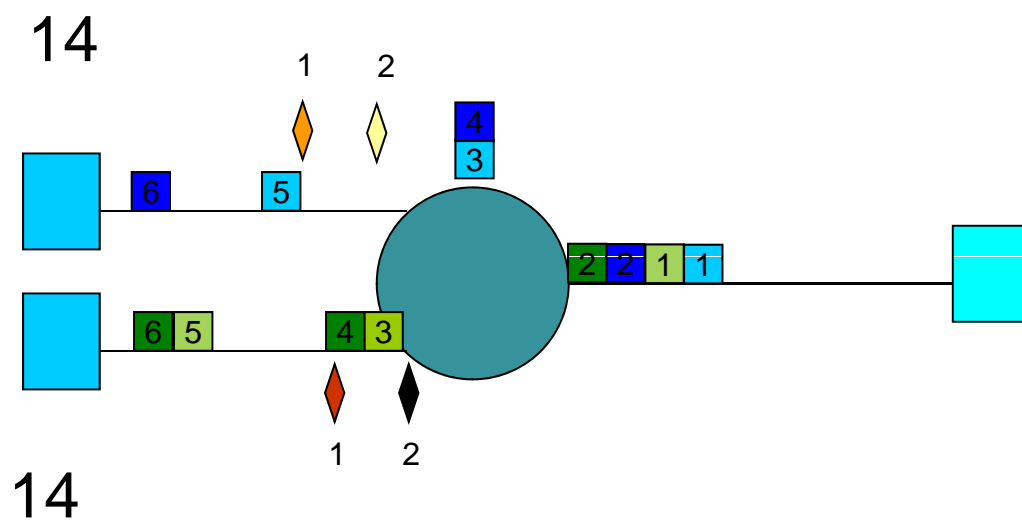


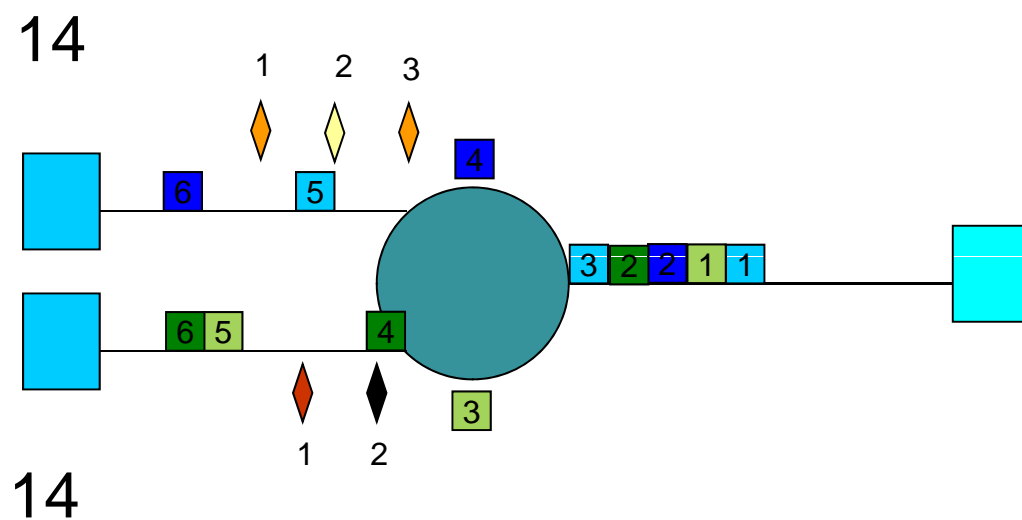
16

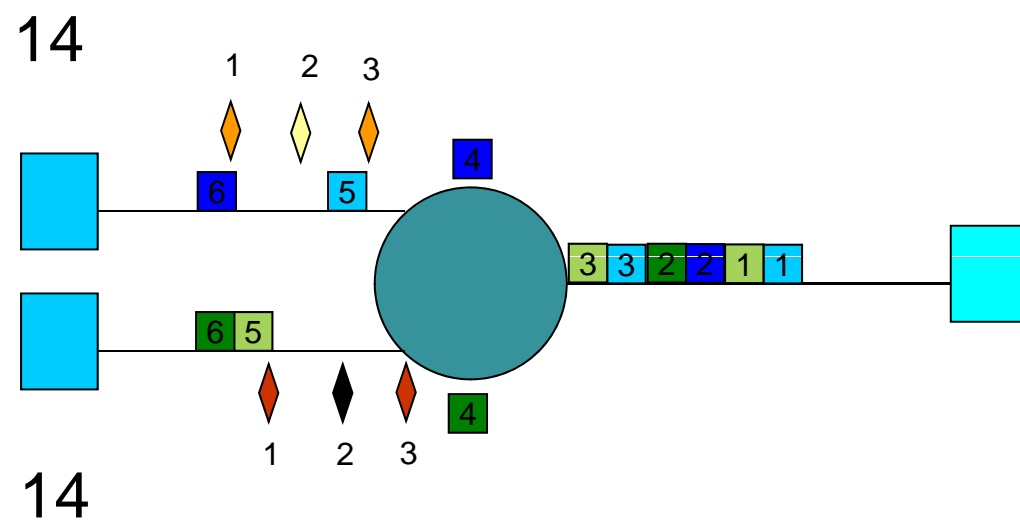


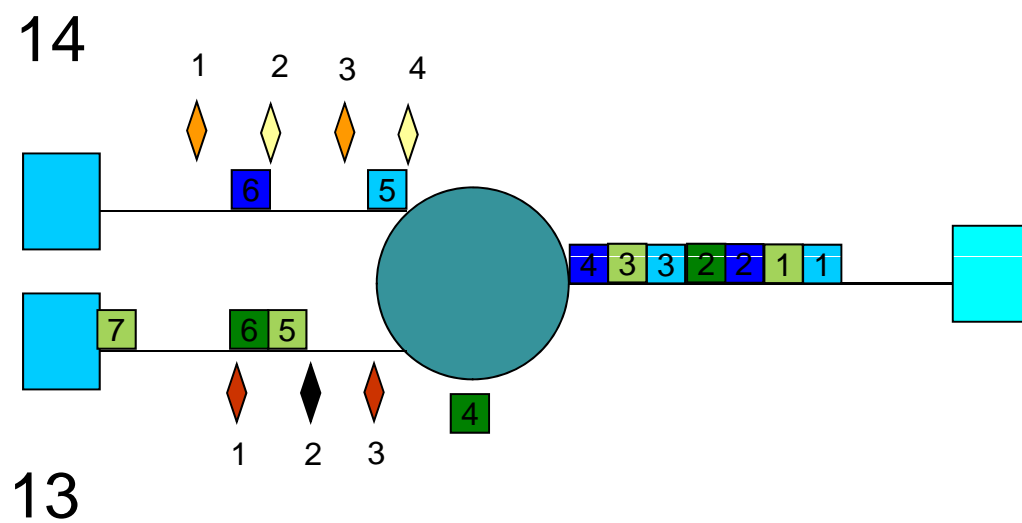


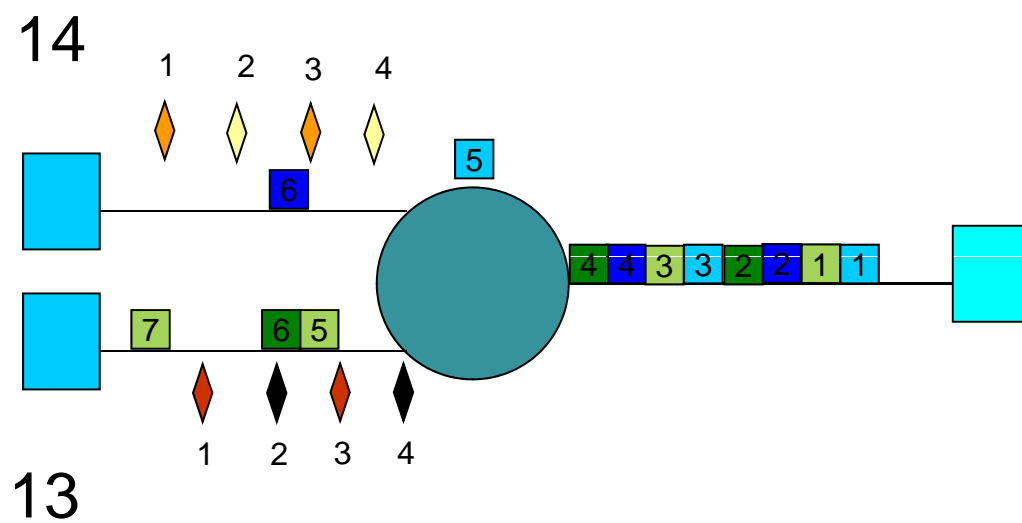


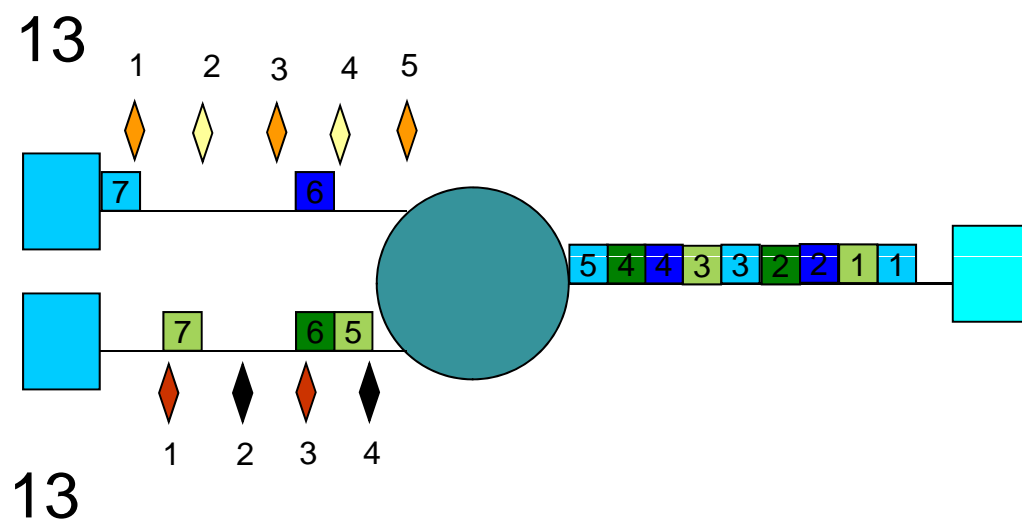


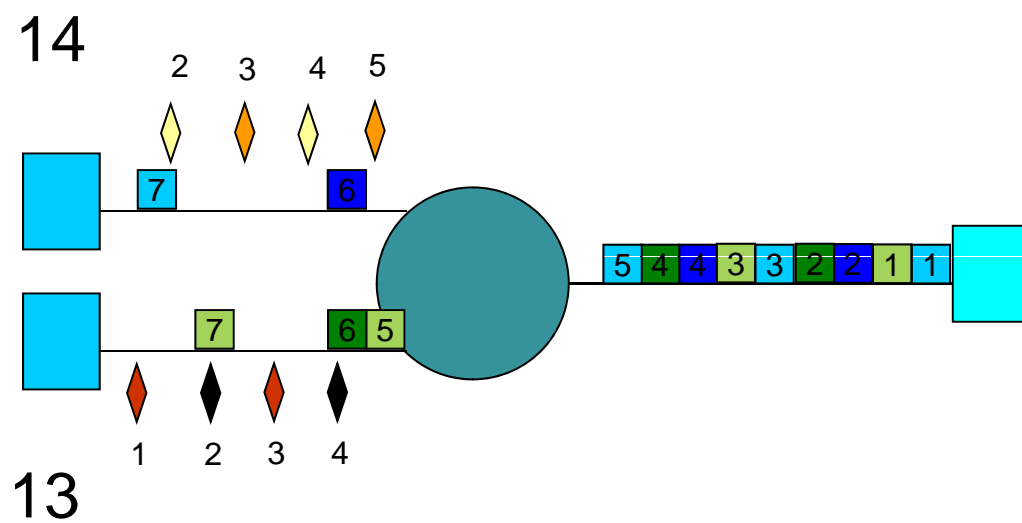


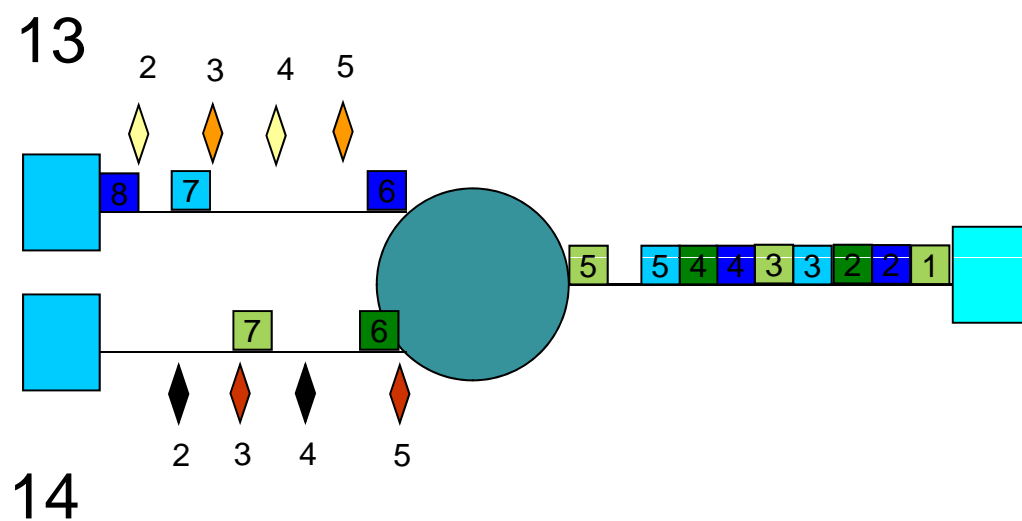


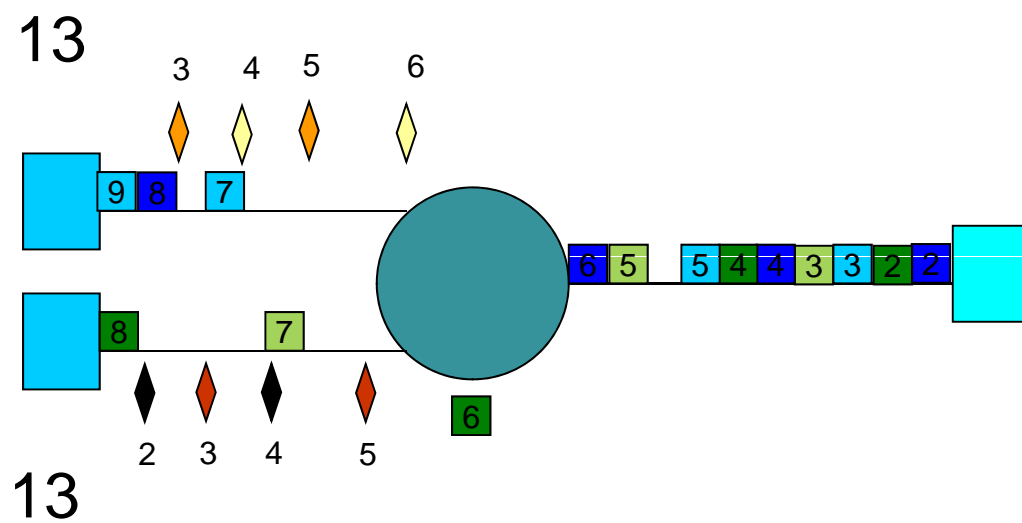


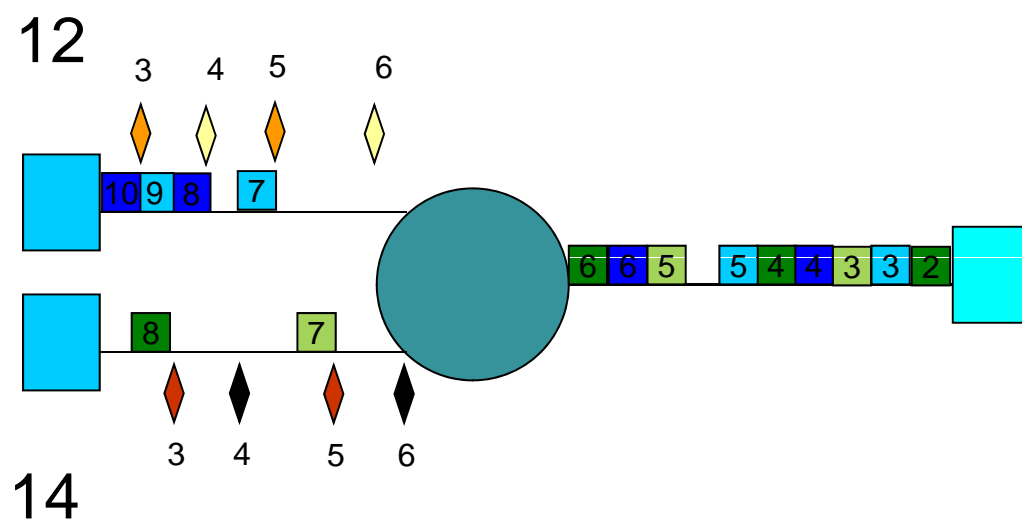


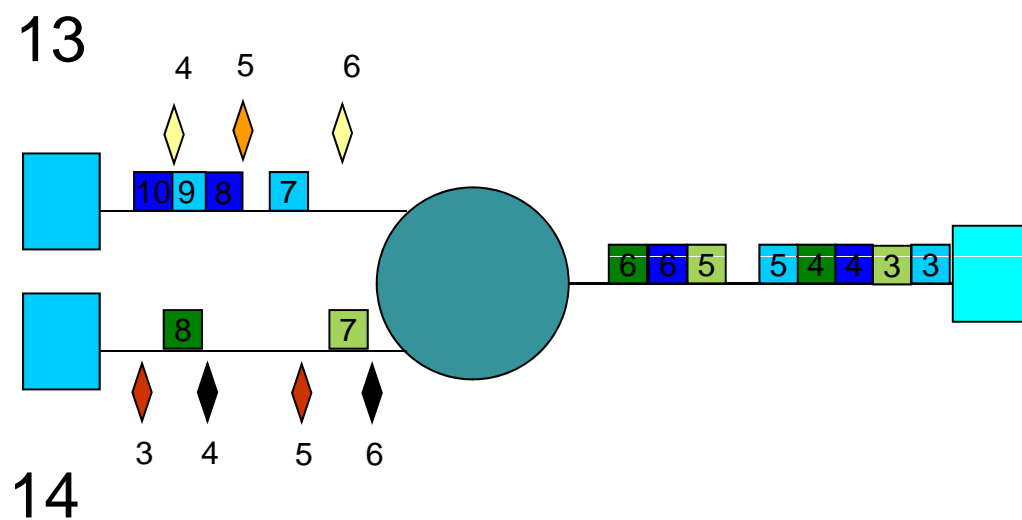


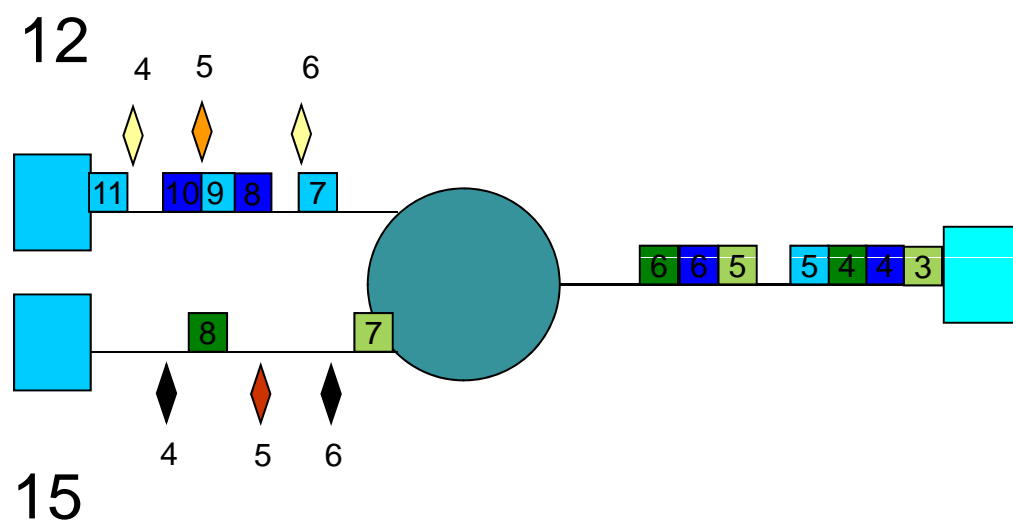


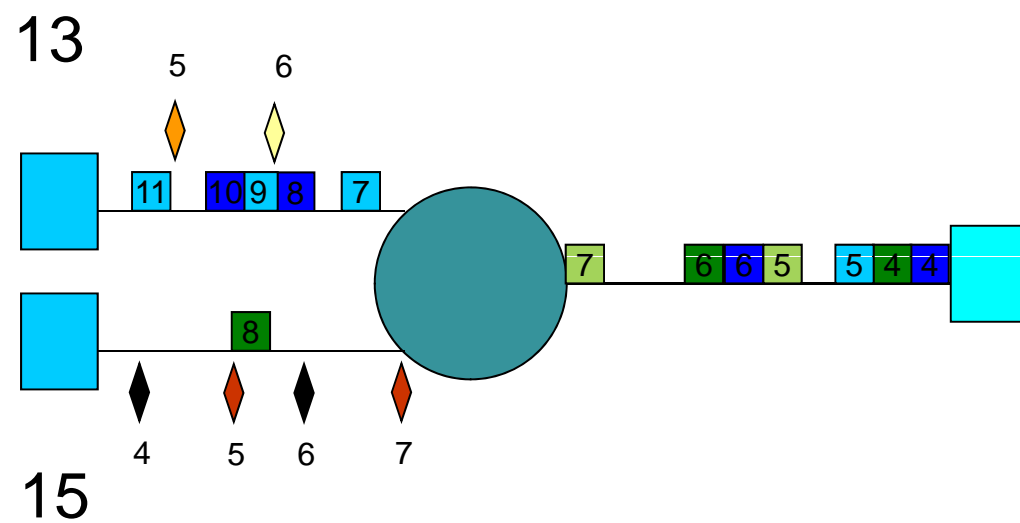


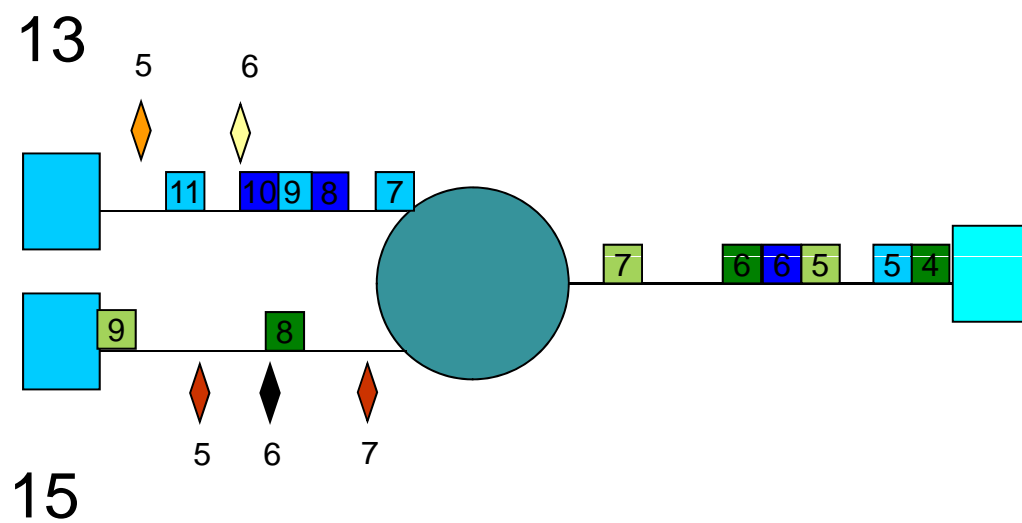


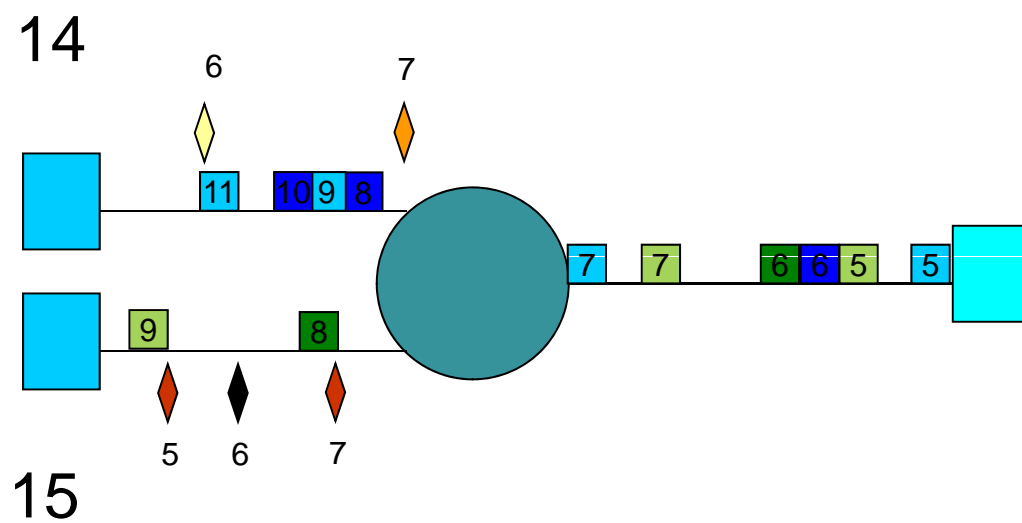


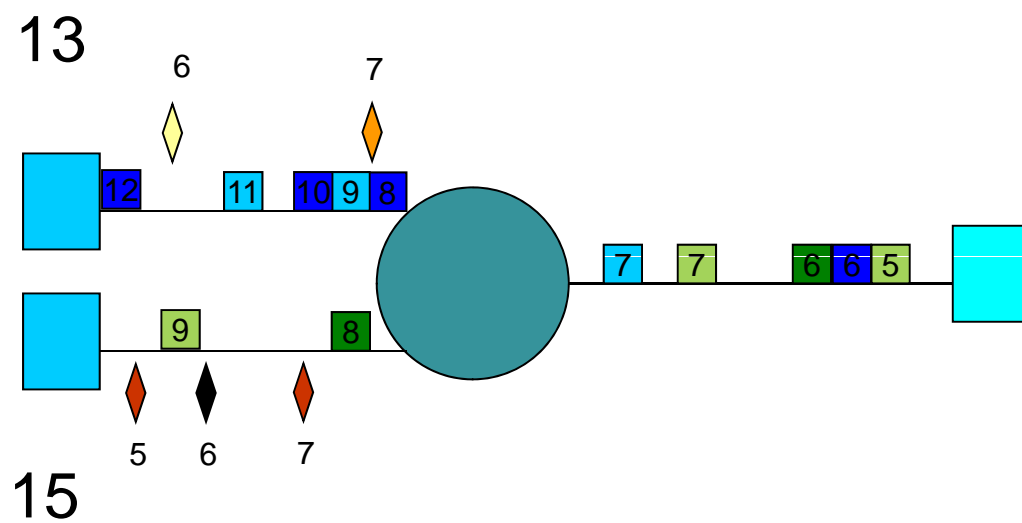


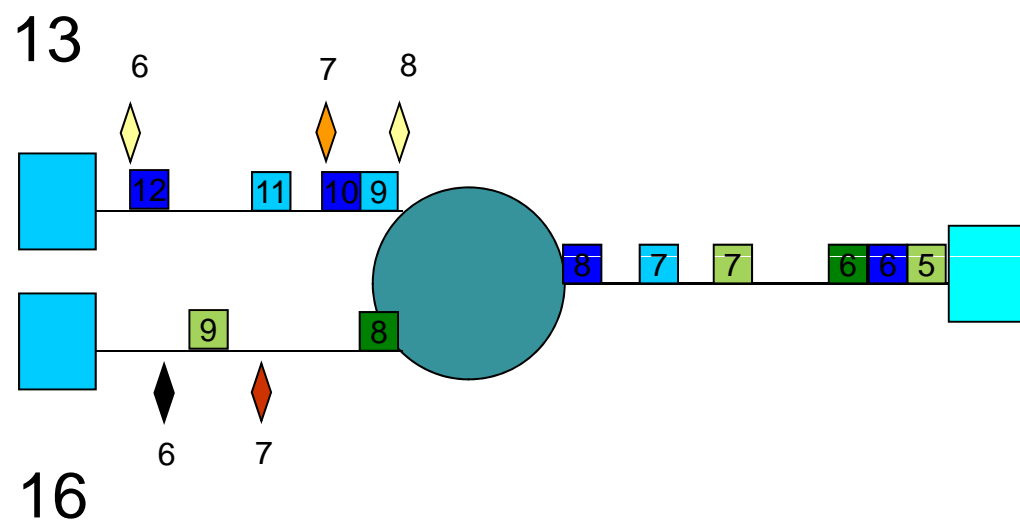


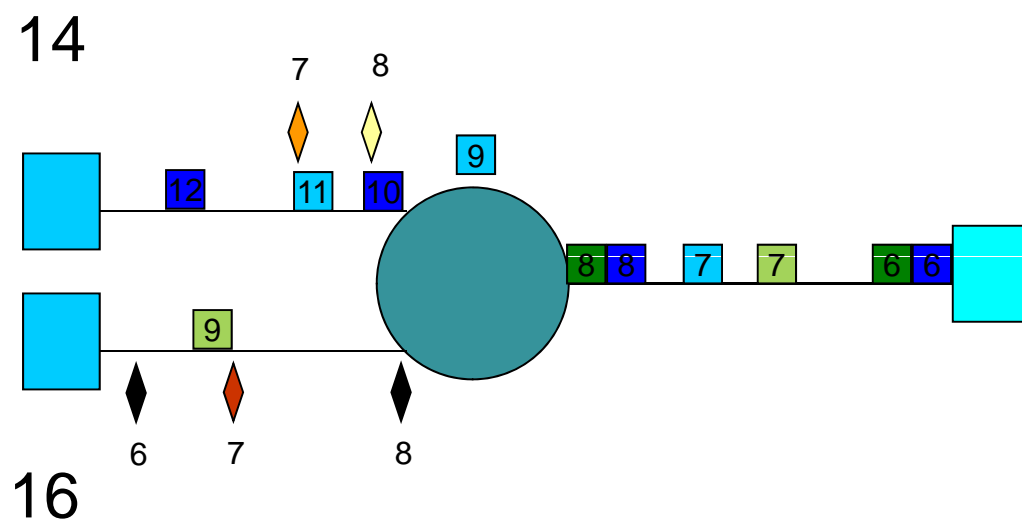












Other Neat Stuff

BUFFER CREDITS

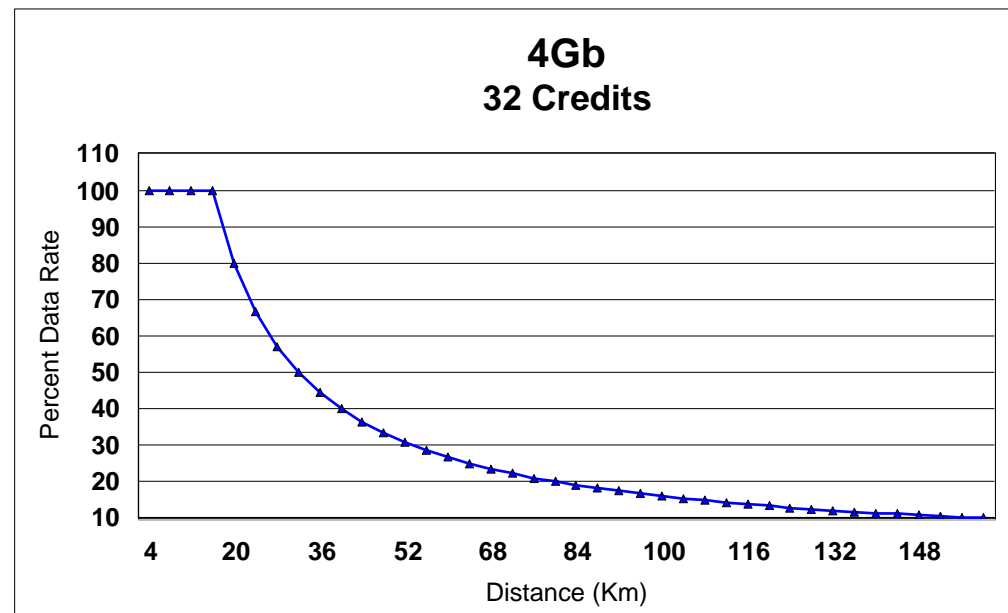
How much credit do I need?

- Good “Rule of thumb”

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

Example: 20 Km at 1 Gb/s
 $1 + \frac{1 * 20}{2} = 11$

Example: 10 Km at 4 Gb/s
 $1 + \frac{4 * 10}{2} = 21$



How “long” is a frame?

- Traveling at the speed of light, a frame can be very long
 - At 1G a frame is about 4Km in length
 - At 2G a frame is about 2Km in length
 - At 4G a frame is about 1Km in length
 - At 8G a frame is about 0.5Km in length

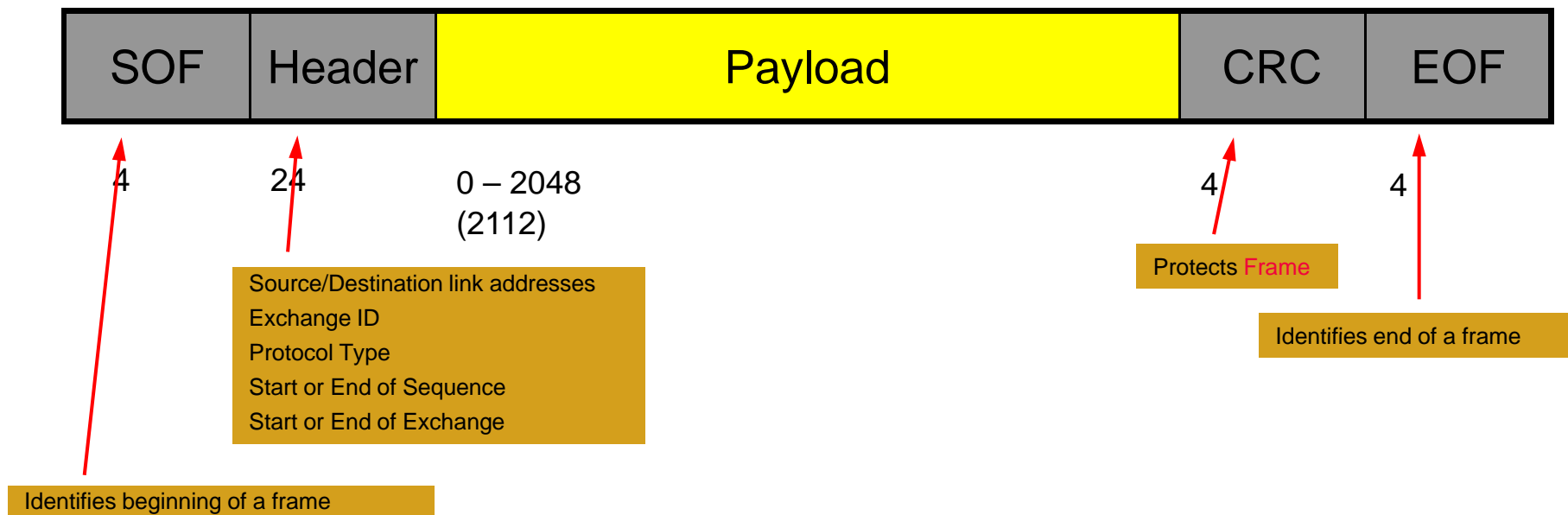
EXCHANGES

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**

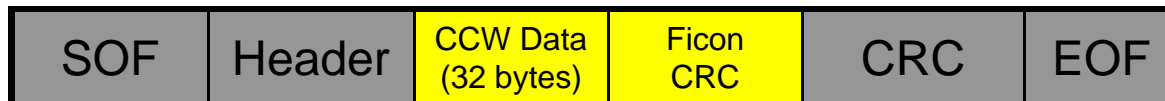
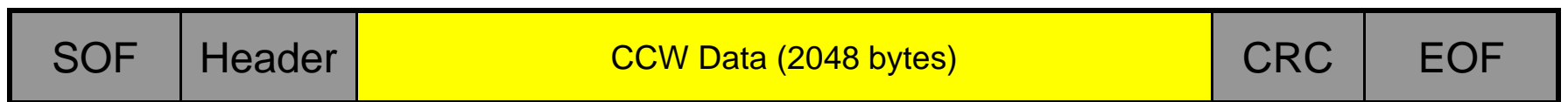
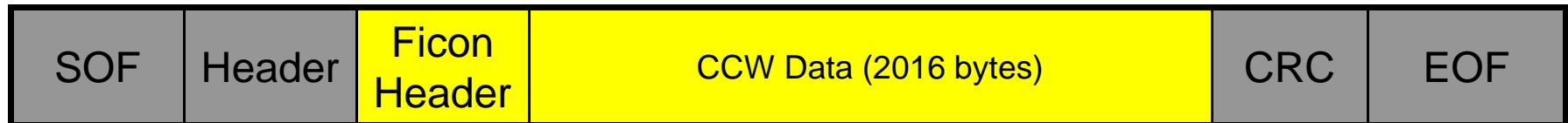


Ficon IU Examples

1 Frame IU to transfer a Read CCW



3 Frame IU to transfer 4K of data



Urban Legend: FICON uses fewer 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**
- 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 its 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

Summary

- Buffer Credits
 - Distance
 - Flow Control
- Exchanges
 - Unidirectional
 - Bidirectional

SHARE

Boston

August 2010

THANK YOU!

REFERENCES

Speaker Biography

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

Speaker Biography

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

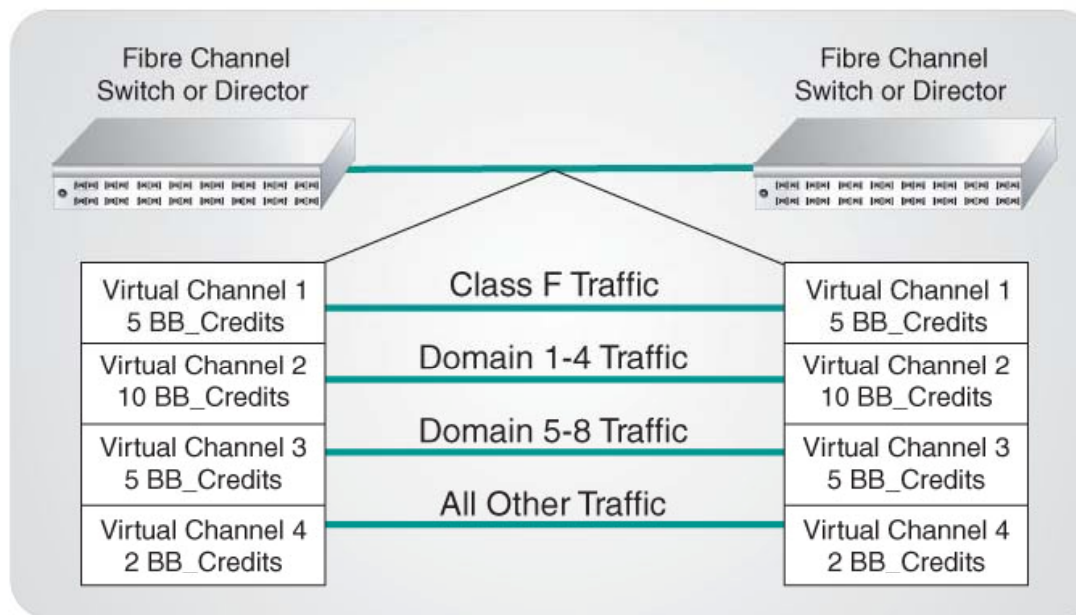
BONUS SLIDES

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**