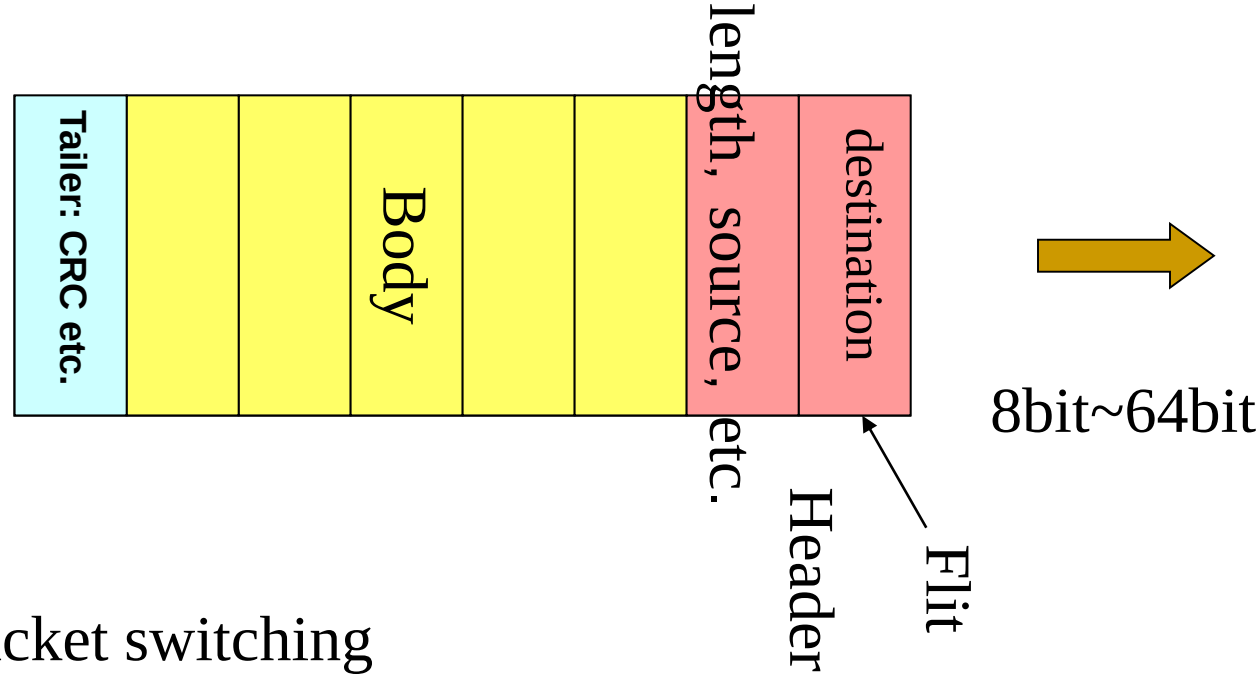
Techniques for packet transfer in parallel machines

AMANO, Hideharu

Textbook pp.166－185

Packet transfer



Packet switching

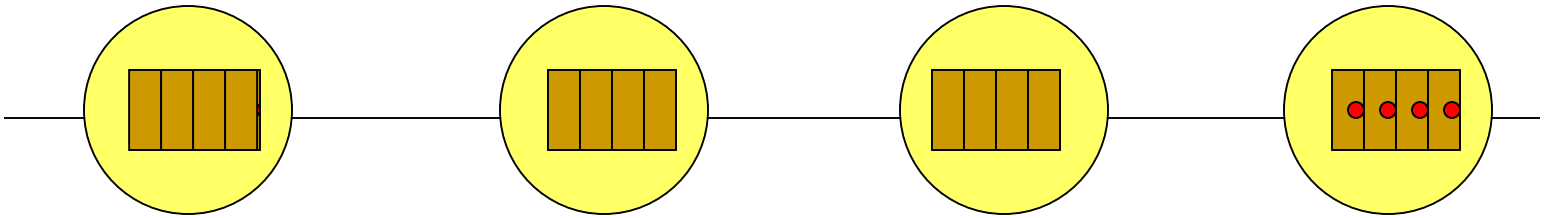
Circuit switching

Flit: Atomic unit for packet transfer
Flit width is not always link width.

Packet transfer method

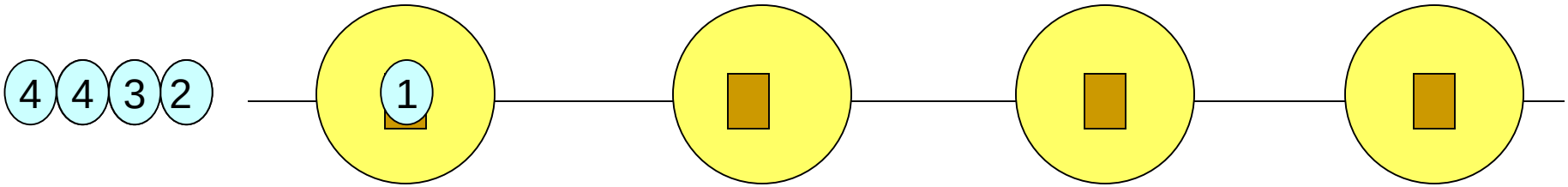
- Store and Forward
  - Entire packet is stored in the buffer of each node
  - TCP/IP protocol must use it
- Wormhole routing
  - Each flit can go forward as possible
  - If the head is blocked, entire packet is stopped.
- Virtual Cut Through
  - If the head is blocked, the rest of packet is stored into the buffer in the node.

Store and Forward



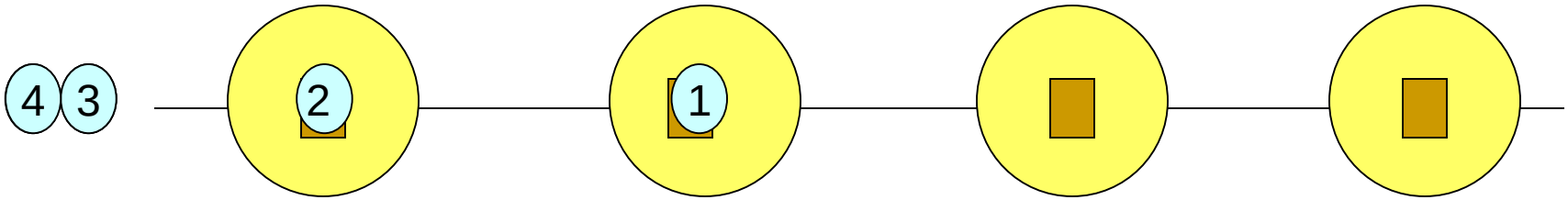
All flits of packet are stored into the buffer in the node.
Large latency D(h+b)
Large requirement of buffer
Re-transmission of faulty packets can be done by the software (TCP/IP uses this method)

Wormhole



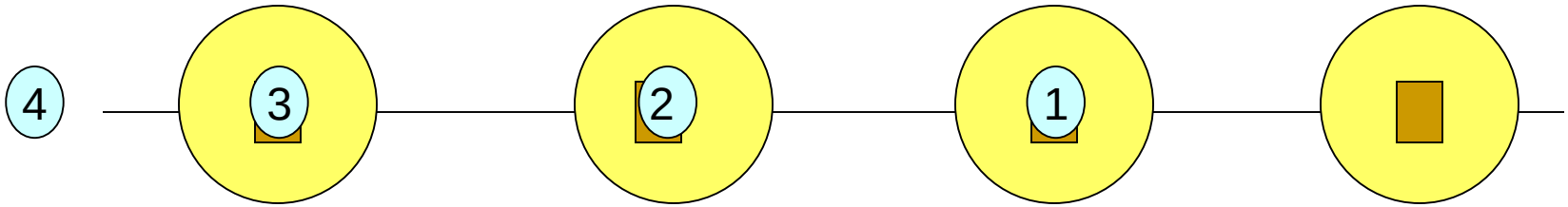
The head of the packet can go as possible
Small latency hD+b
Small buffer requirement
Hardware router is required.

Wormhole



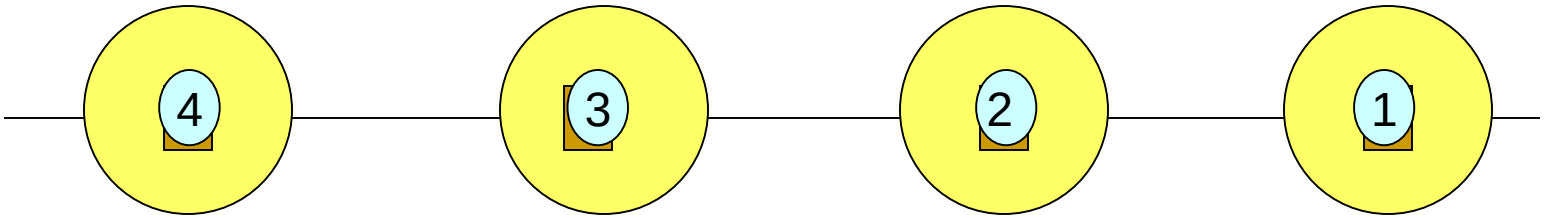
The head of the packet can go as possible
Small latency hD+b
Small buffer requirement
Hardware router is required.

Wormhole



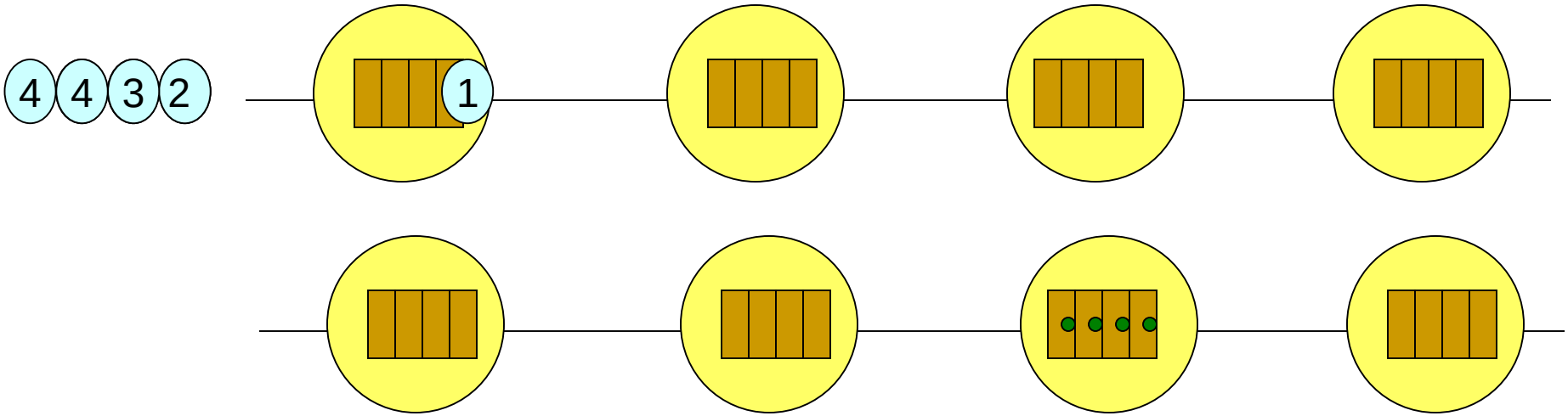
The head of the packet can go as possible
Small latency hD+b
Small buffer requirement
Hardware router is required.

Wormhole



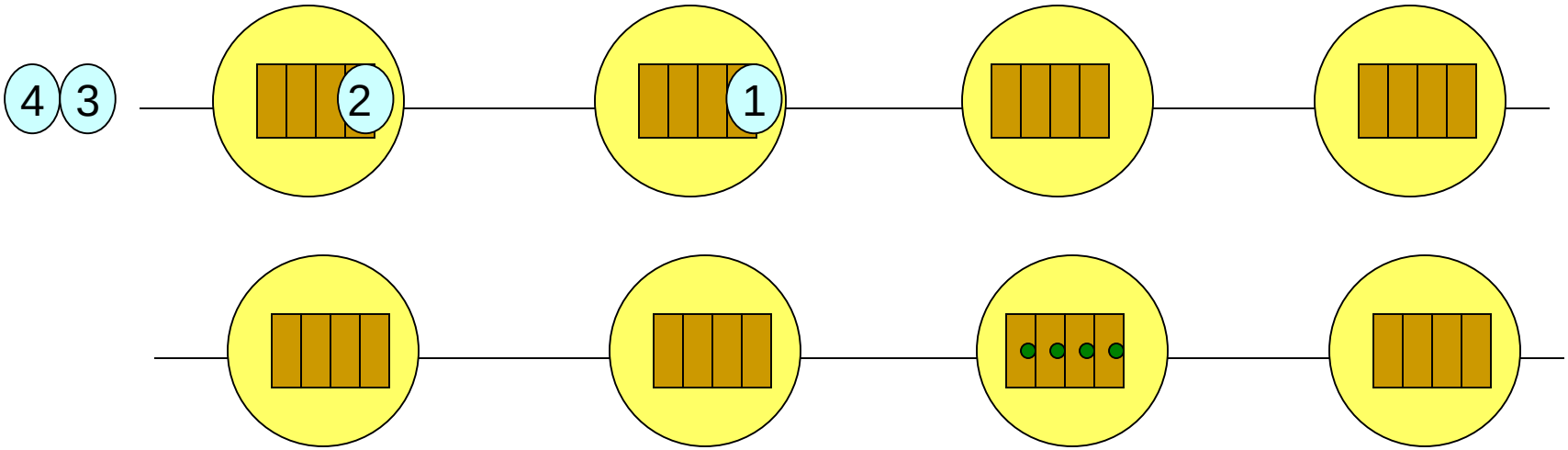
The head of the packet can go as possible
Small latency hD+b
Small buffer requirement
Hardware router is required.

Virtual Cut Through



If blocked, the rest of packet is stored in the buffer
The same latency as Wormhole
The same buffer requirement as Store and Forward
Hardware router is required.

Virtual Cut Through



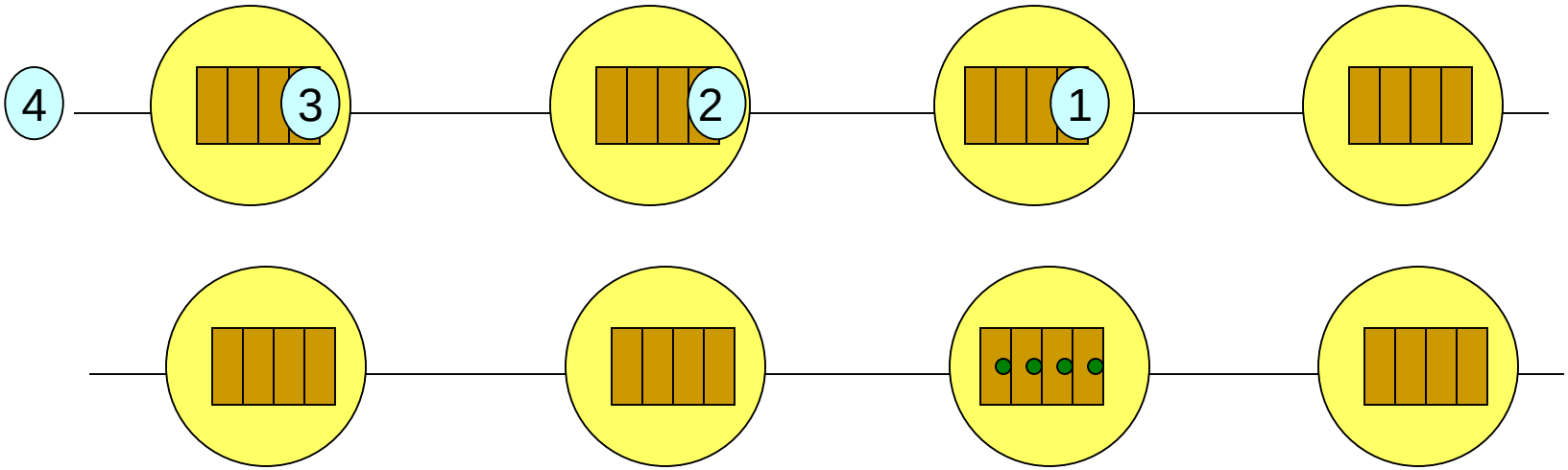
If blocked, the rest of packet is stored in the buffer

The same latency as Wormhole

The same buffer requirement as Store and Forward

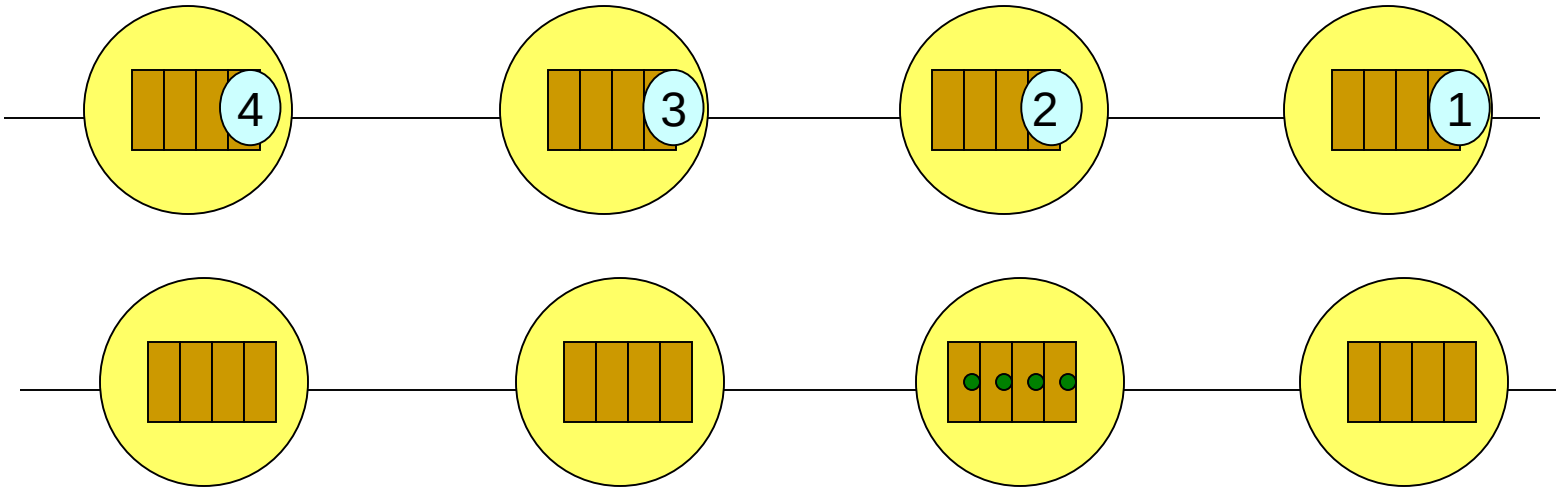
Hardware router is required.

Virtual Cut Through



If blocked, the rest of packet is stored in the buffer
The same latency as Wormhole
The same buffer requirement as Store and Forward
Hardware router is required.

Virtual Cut Through



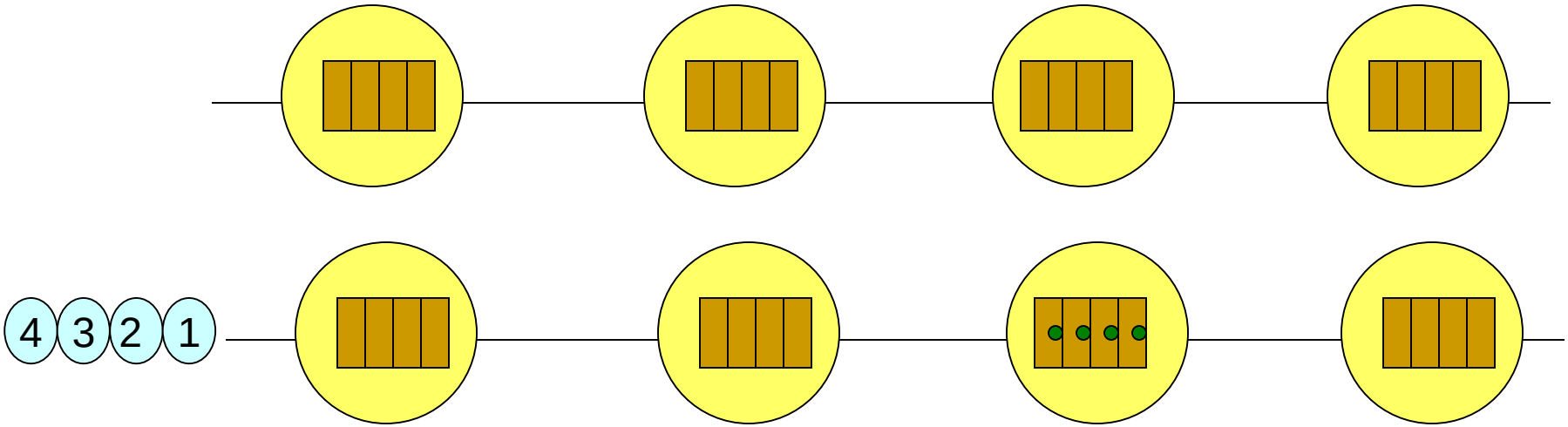
If blocked, the rest of packet is stored in the buffer

The same latency as Wormhole

The same buffer requirement as Store and Forward

Hardware router is required.

Virtual Cut Through



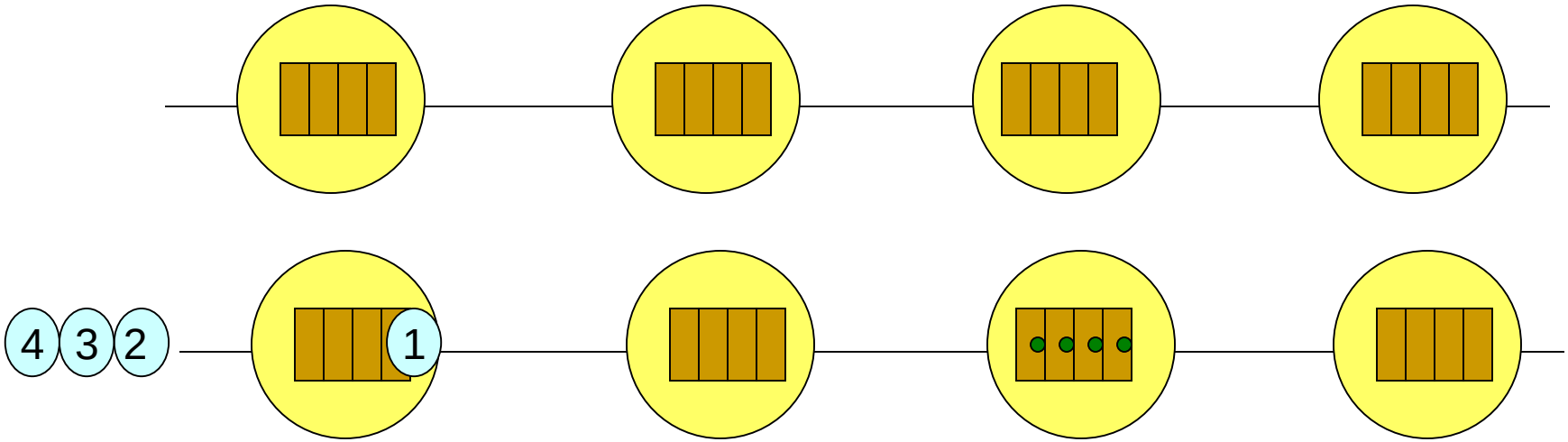
If blocked, the rest of packet is stored in the buffer

The same latency as Wormhole

The same buffer requirement as Store and Forward

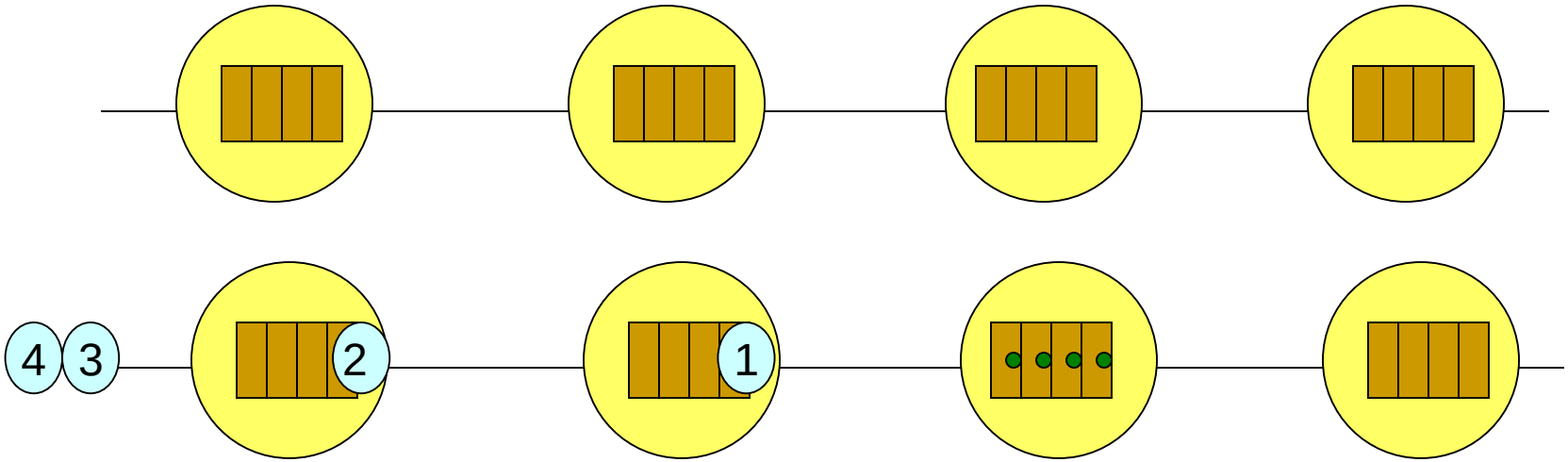
Hardware router is required.

Virtual Cut Through



If blocked, the rest of packet is stored in the buffer
The same latency as Wormhole
The same buffer requirement as Store and Forward
Hardware router is required.

Virtual Cut Through



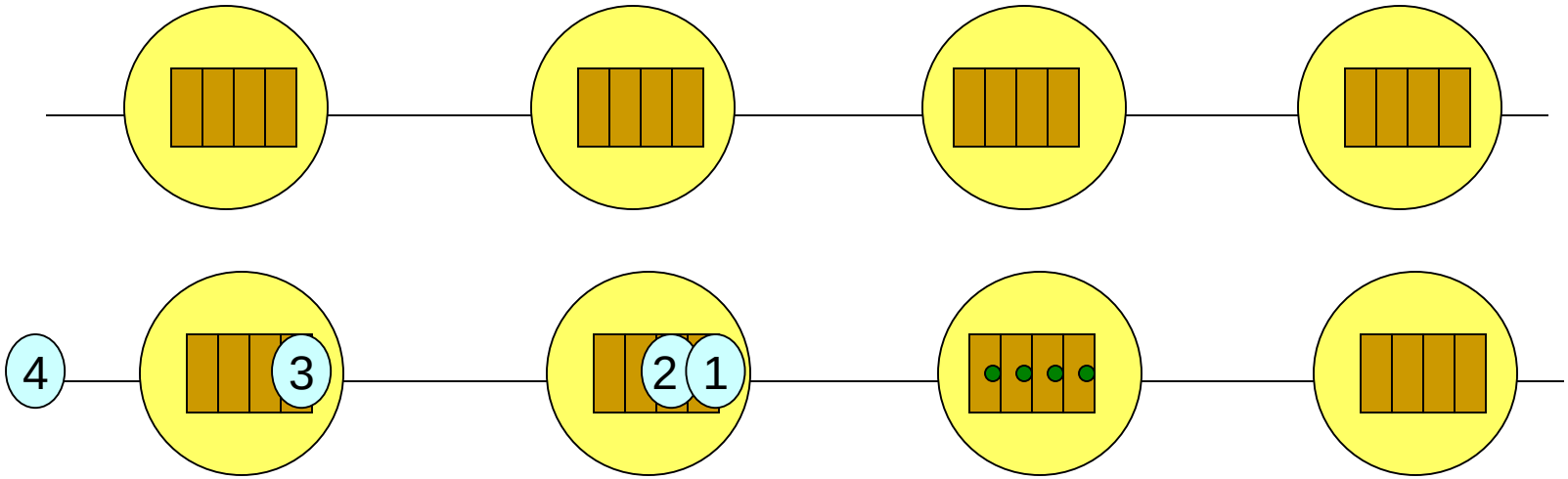
If blocked, the rest of packet is stored in the buffer

The same latency as Wormhole

The same buffer requirement as Store and Forward

Hardware router is required.

Virtual Cut Through



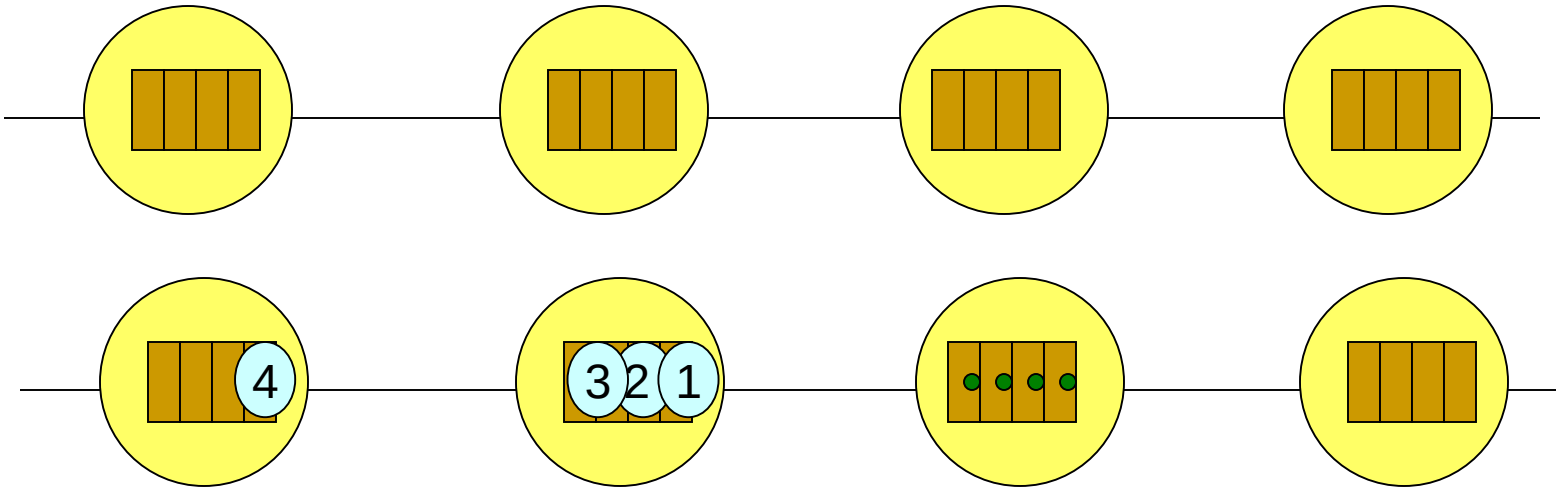
If blocked, the rest of packet is stored in the buffer

The same latency as Wormhole

The same buffer requirement as Store and Forward

Hardware router is required.

Virtual Cut Through



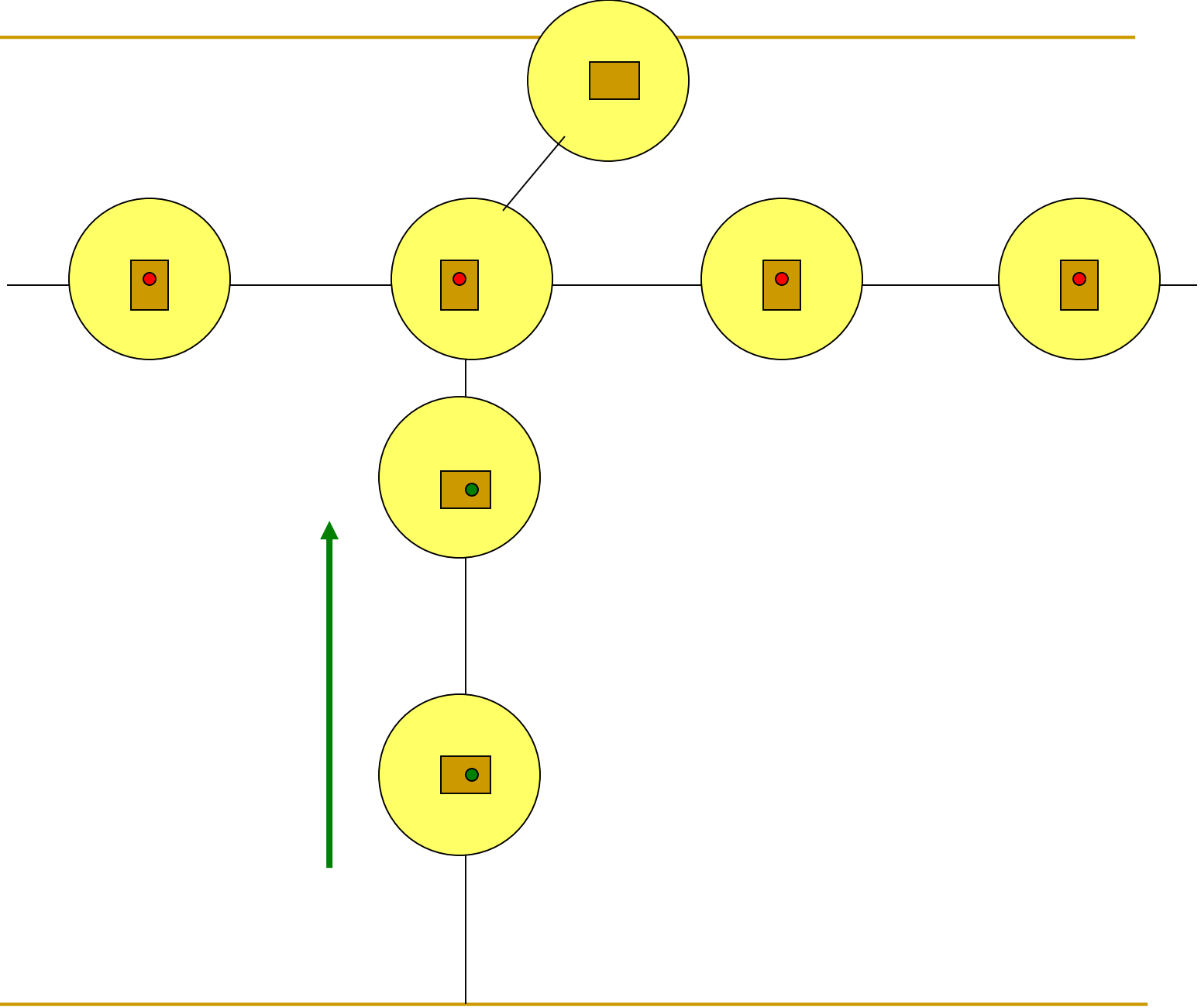
If blocked, the rest of packet is stored in the buffer
The same latency as Wormhole
The same buffer requirement as Store and Forward
Hardware router is required.

LAN,Component networks/SAN and Network on Chip

- LAN(Local Area Network):
  - Store and Forward
- Component network/ SAN(System Area Network):
  - The first generation NORA uses store and forward method.
  - Recent Component networks/SANs:
    - For large packets: Wormhole
    - For multicast: Virtual Cut Through
  - Myrinet, QsNET
- Network on Chip:
  - Wormhole

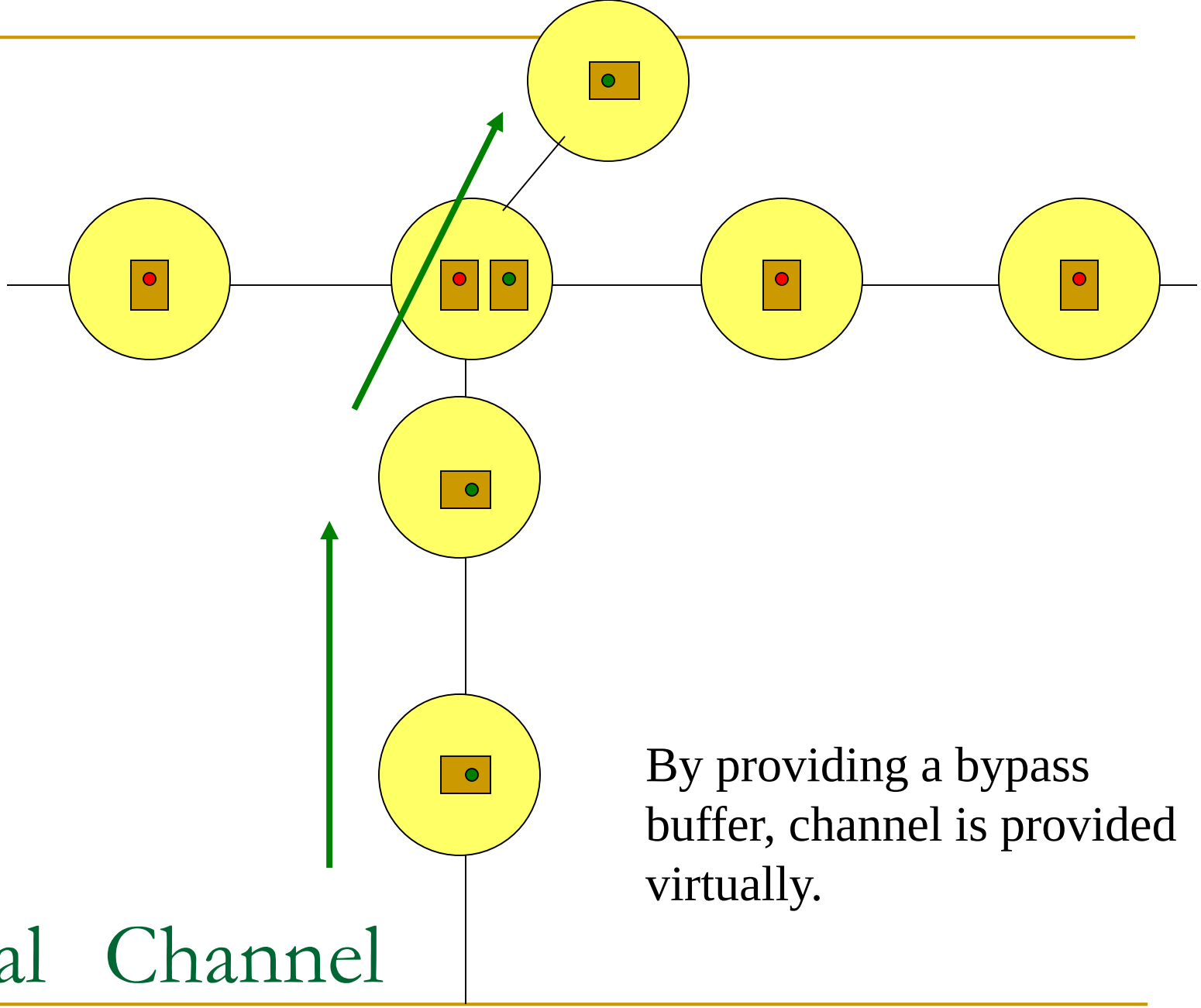
Qiuz

- A packet with 1 flit header and 15 flits body is transferred on a 4-ary 2-cube. Compute the largest number of clocks when it is sent with Store-and-Forward manner, and compared with the case when Wormhole method is used. Ignore the delay caused by congestion.



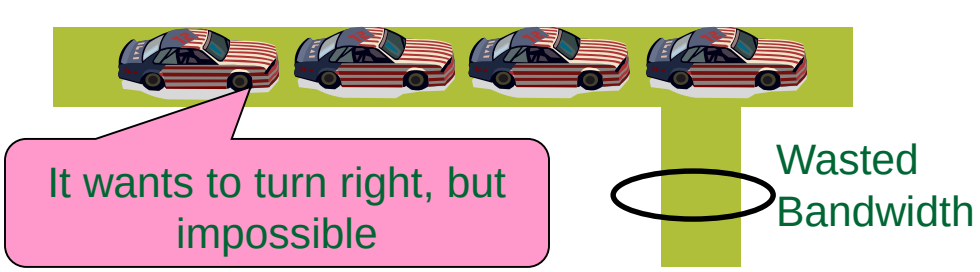
A problem of Wormhole

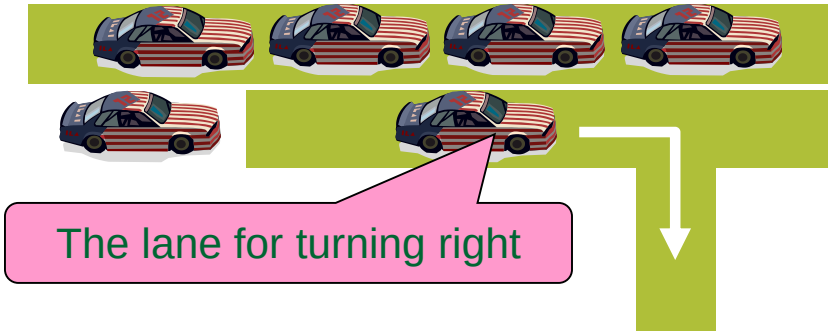
Virtual Channel



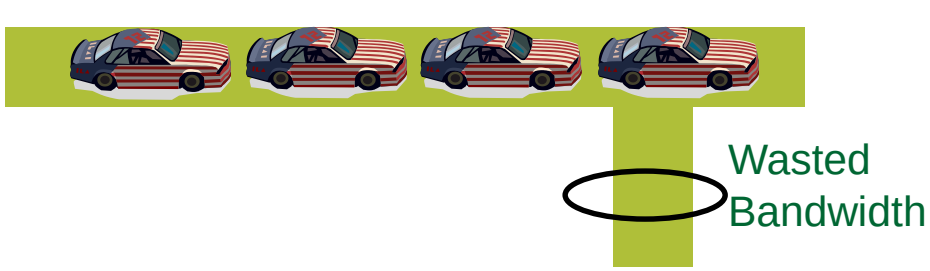
By providing a bypass buffer, channel is provided virtually.

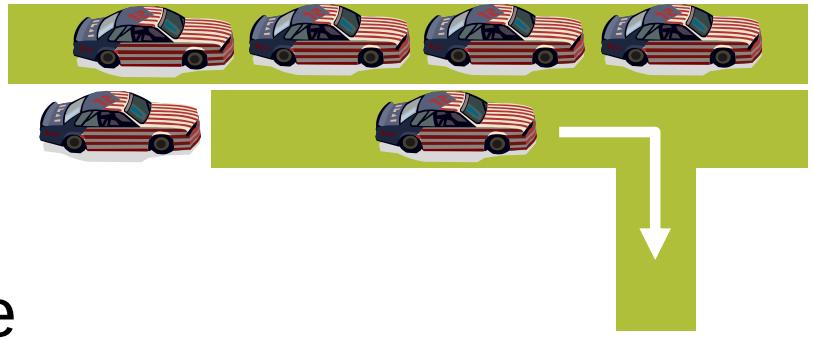
Implementation of Virtual Channel





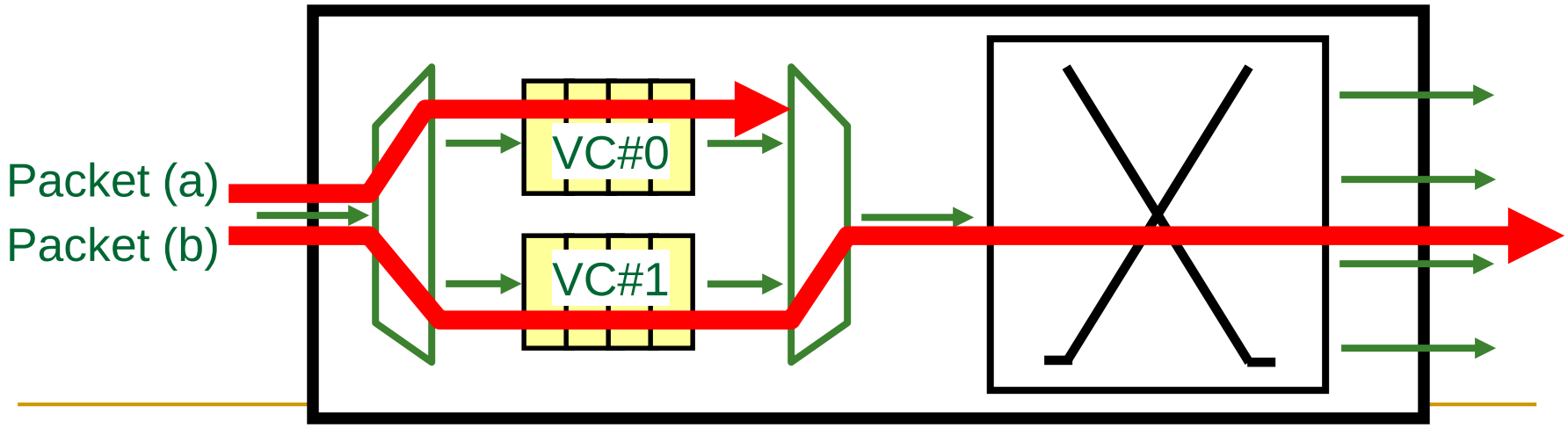
Implementation of Virtual Channel



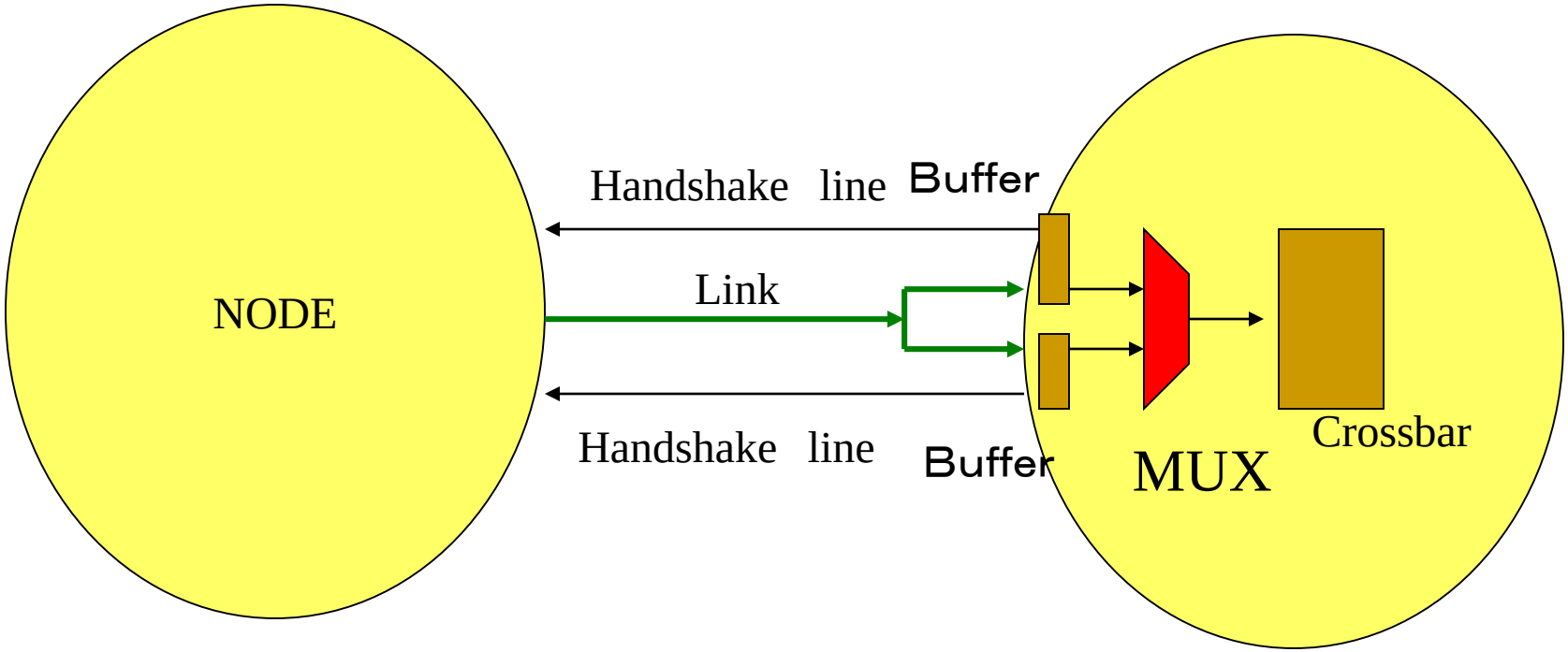


- VC → Providing another lane
  - But the physical wires are not increased.

[Dally,TPDS'92]



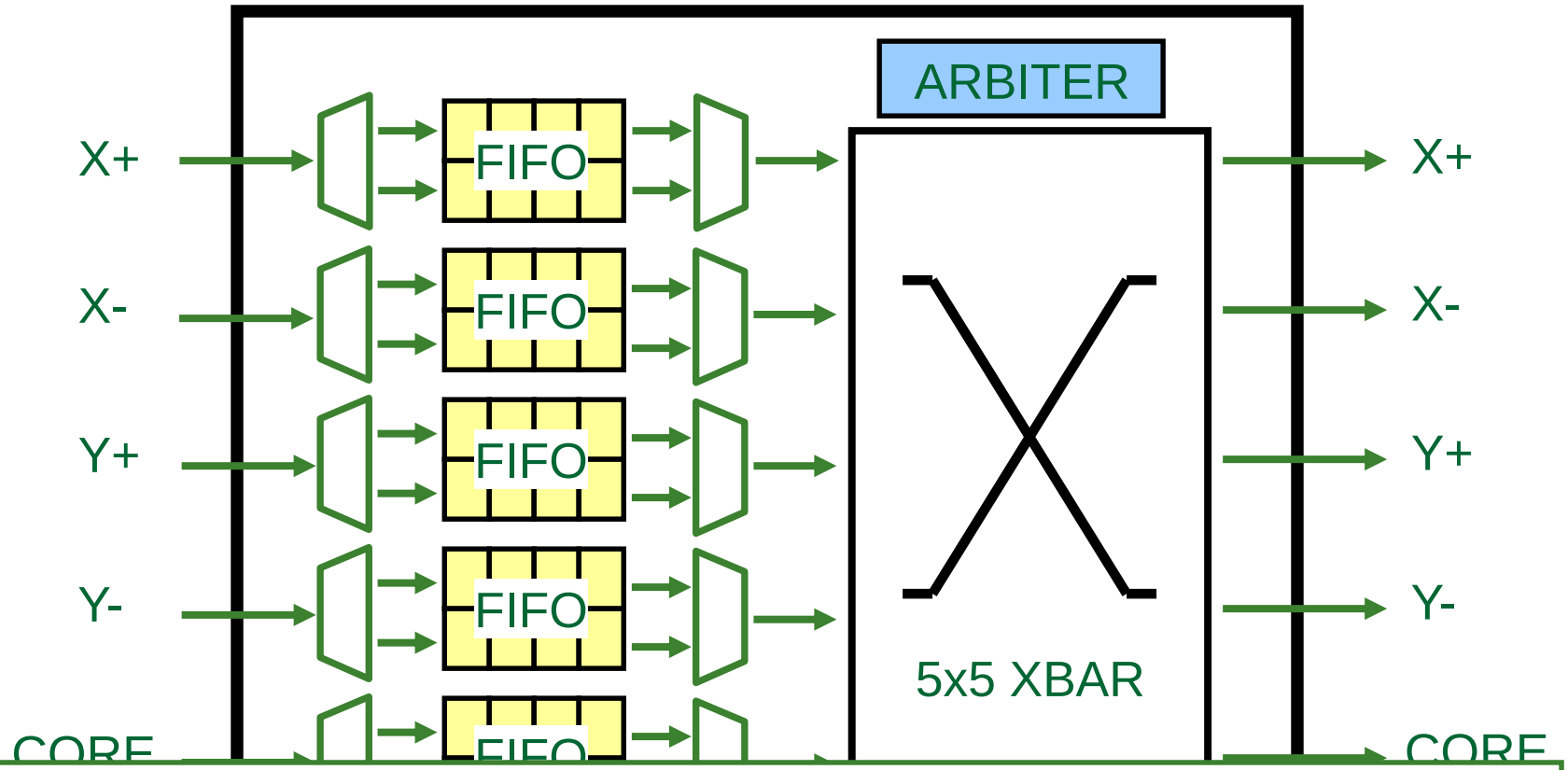
Handshake of Virtual Channel



An example of a modern router

- WH router with two virtual channels

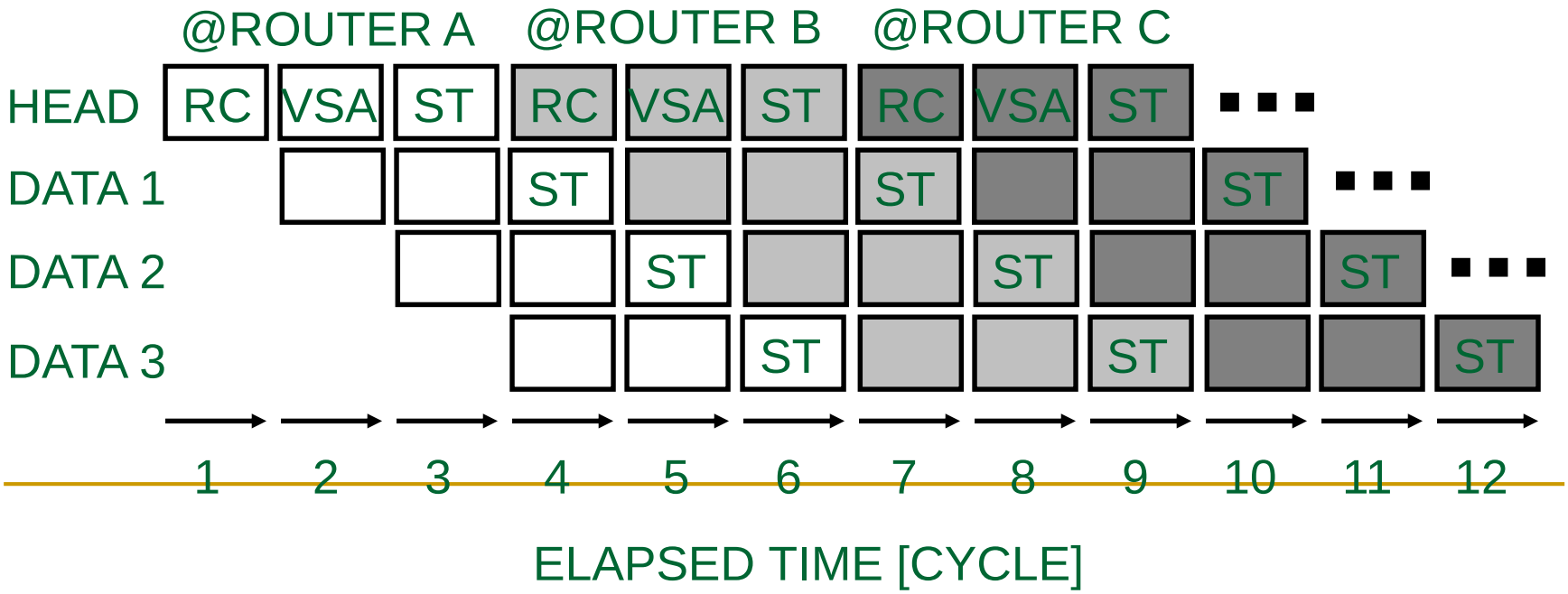
(Introduced later in this lecture)



If the bitwidth is 64bits, it uses 30～40 [kgates] FIFO occupies 60%

Pipelined operation

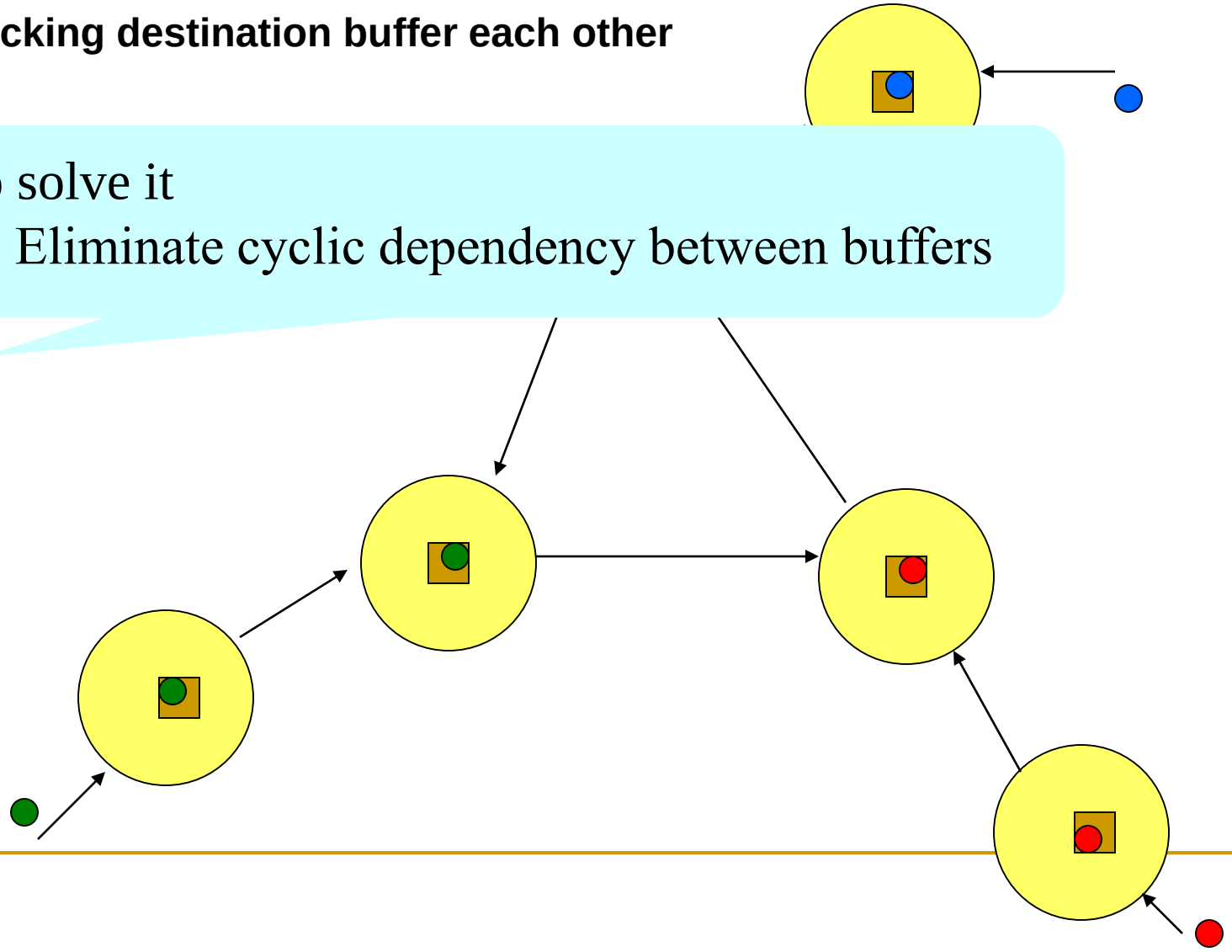
- It takes three clocks to pass through the switch
  - RC (Routing Computation)
  - VSA (Virtual Channel / Switch Allocation)
  - ST (Switch Traversal)



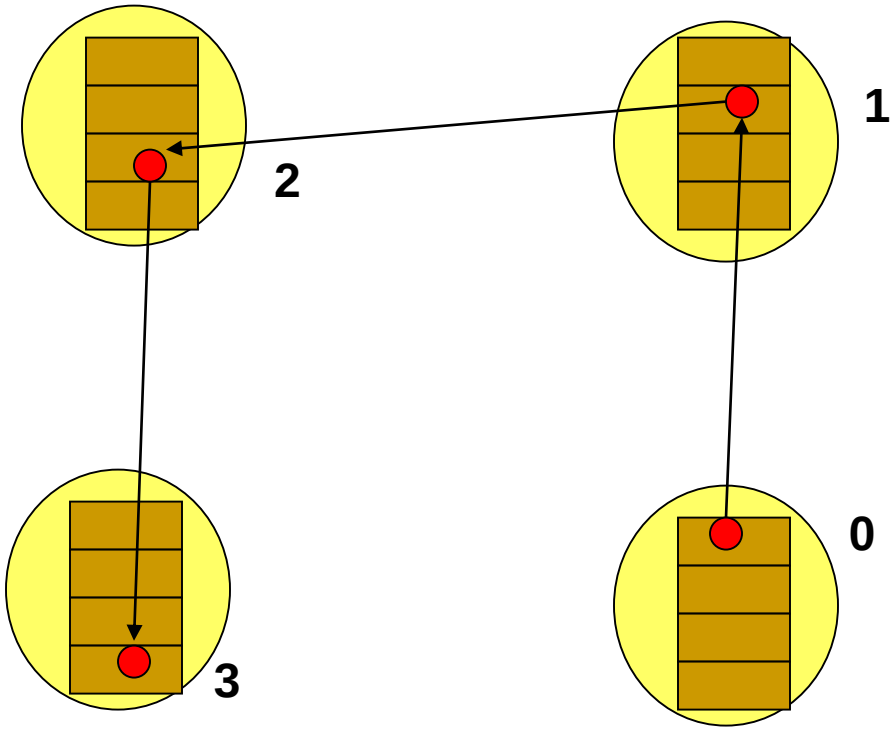
Deadlock avoidance

Blocking destination buffer each other

To solve it
→ Eliminate cyclic dependency between buffers



Structured buffer pool

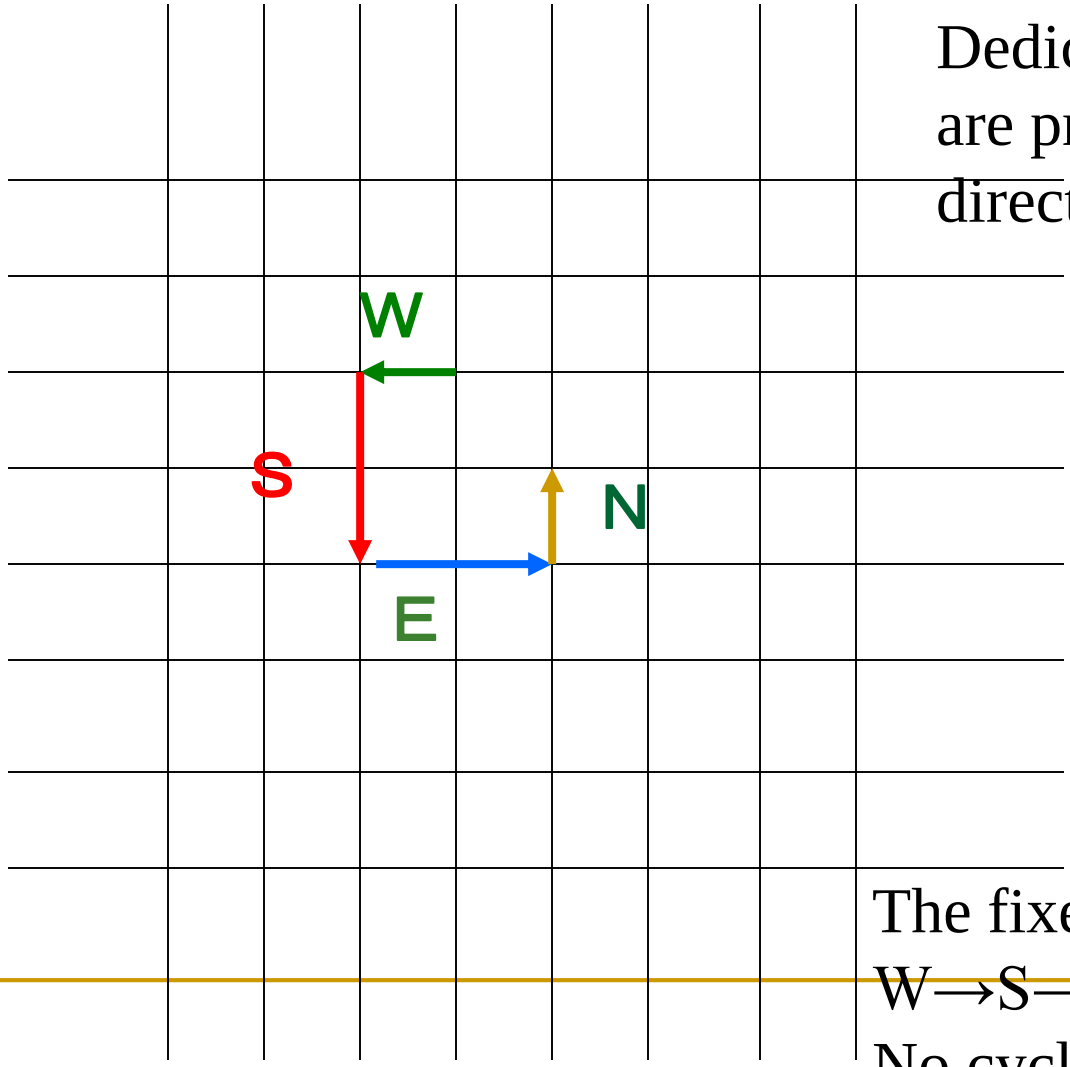


Packet is sent to Buf#＋1

No cyclic dependency between buffers

Structured channel for Wormhole

Dimension order (e-cube) routing : DOR

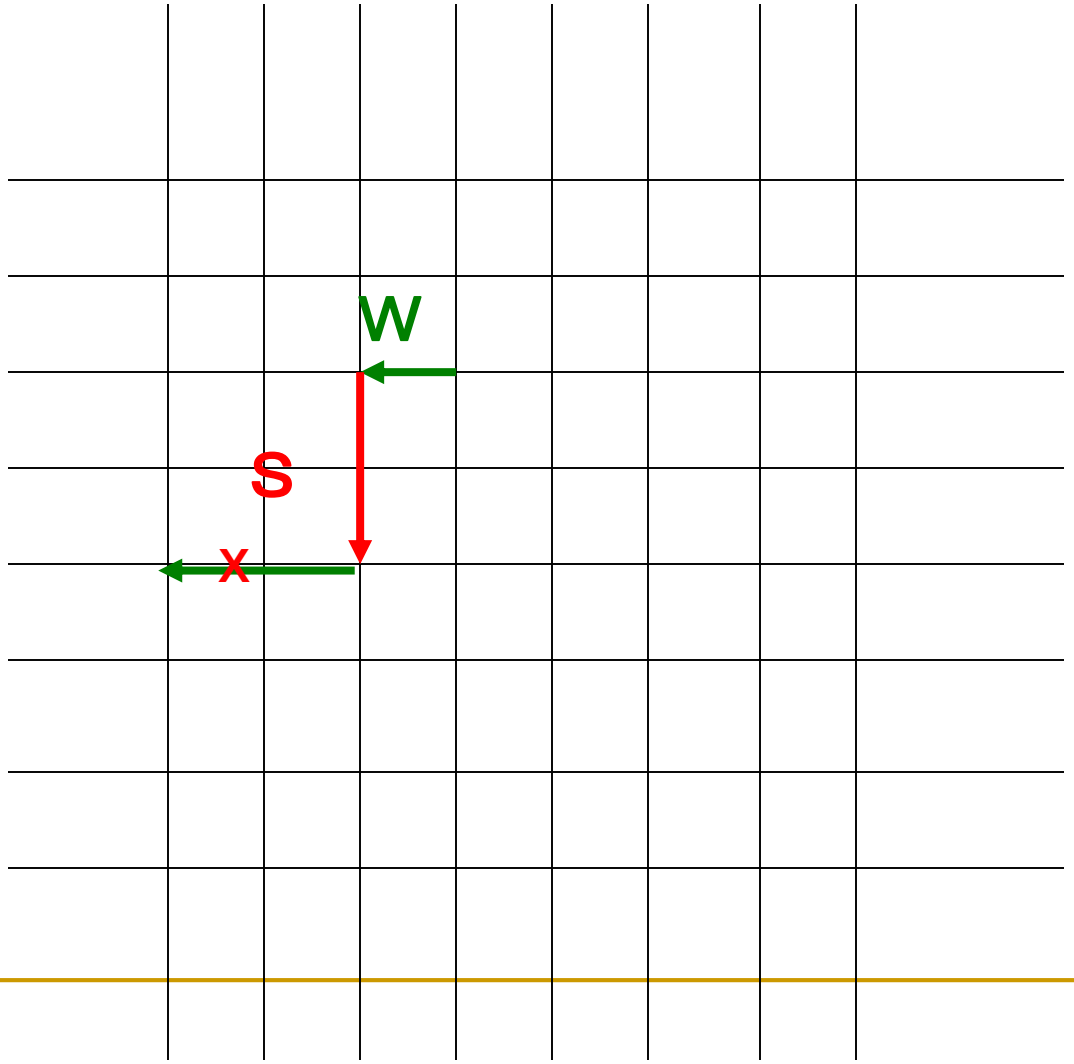


Dedicated buffers are provided for each direction.

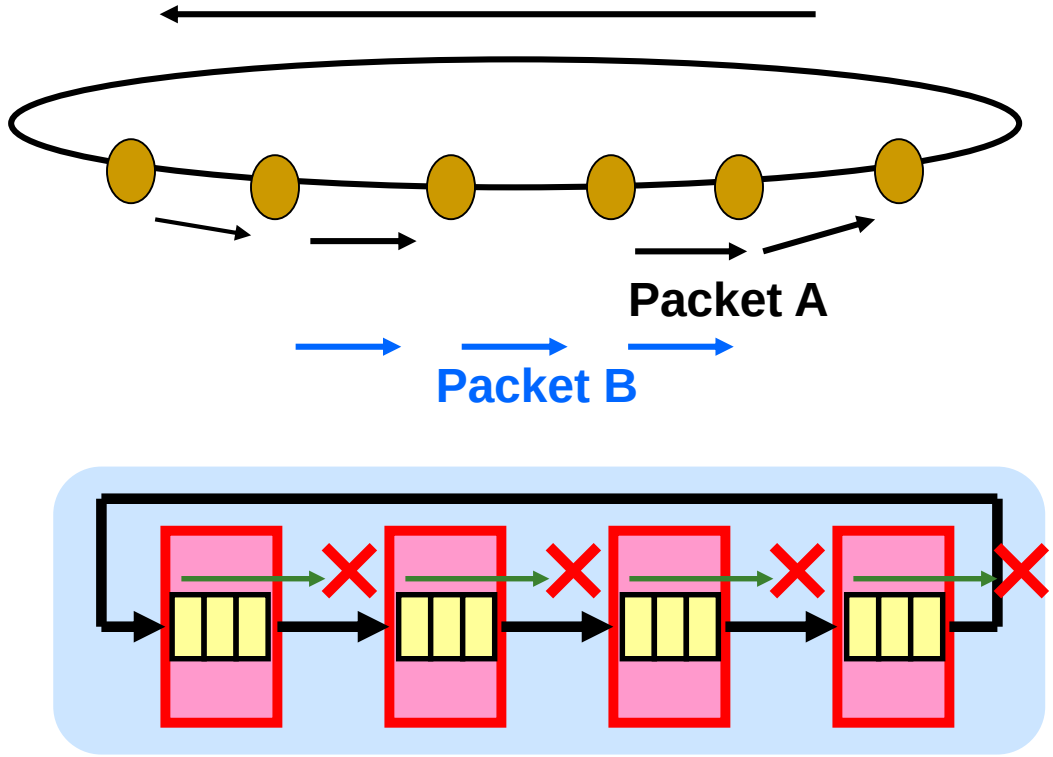
The fixed order:
W→S→E→N
No cyclic dependency

Dimension order routing

Once the direction is changed, it cannot be used again.

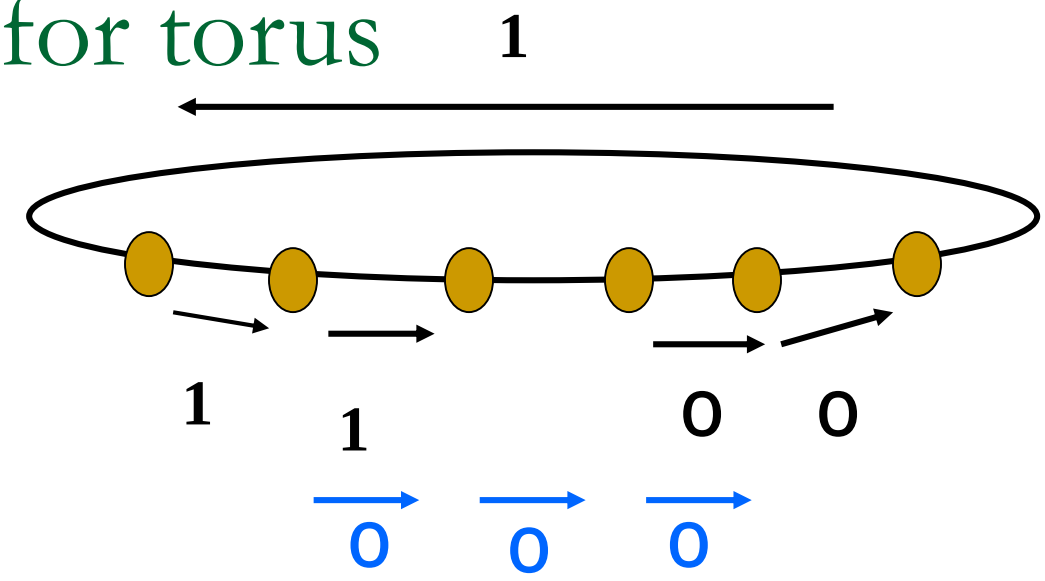


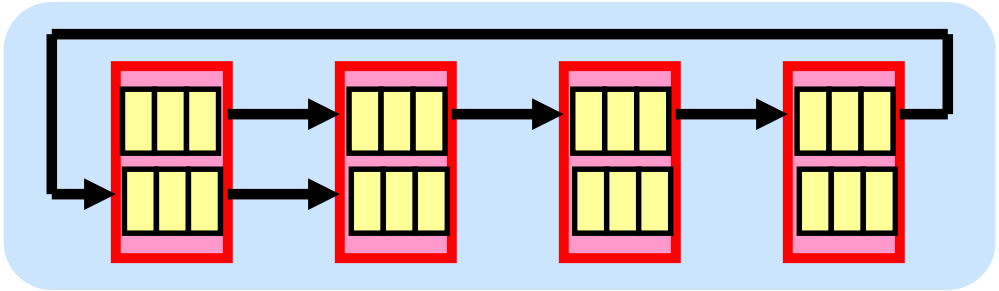
DOR for torus



Single direction packet transfer also makes a cyclic dependency

DOR for torus





Virtual channel number is changed when the round trip link is used.

Glossary 1

- Flit:パケットの基本（最小）転送単位、必ずしもリンクのビット幅と等しい必要はないが、これ以上細かいデータ単位で転送を制御することはできないもの
- Wormhole routing:いも虫が穴を開けながら進んで行く様子から出た単語。日本語でもこのまま読む。
- Virtual Cut through:仮想的にパケットが突き抜けたように見えることから出た単語。日本語でもこのまま読む
- Virtual Channel:仮想チャネル。バッファとハンドシェークラインを独立に用意することで、仮想的に複数の転送チャネルを実現する。リンクの利用率を上げ、デッドロックを防ぐ。
- Deadlock:すくみ、デッドロック、パケットが用いるバッファがCyclic dependency（互いに循環的にバッファを要求すること）を生じることにより、先に進めなくなる現象
- Structural buffer pool：構造化バッファ法、デッドロックを防ぐための古典的な手法

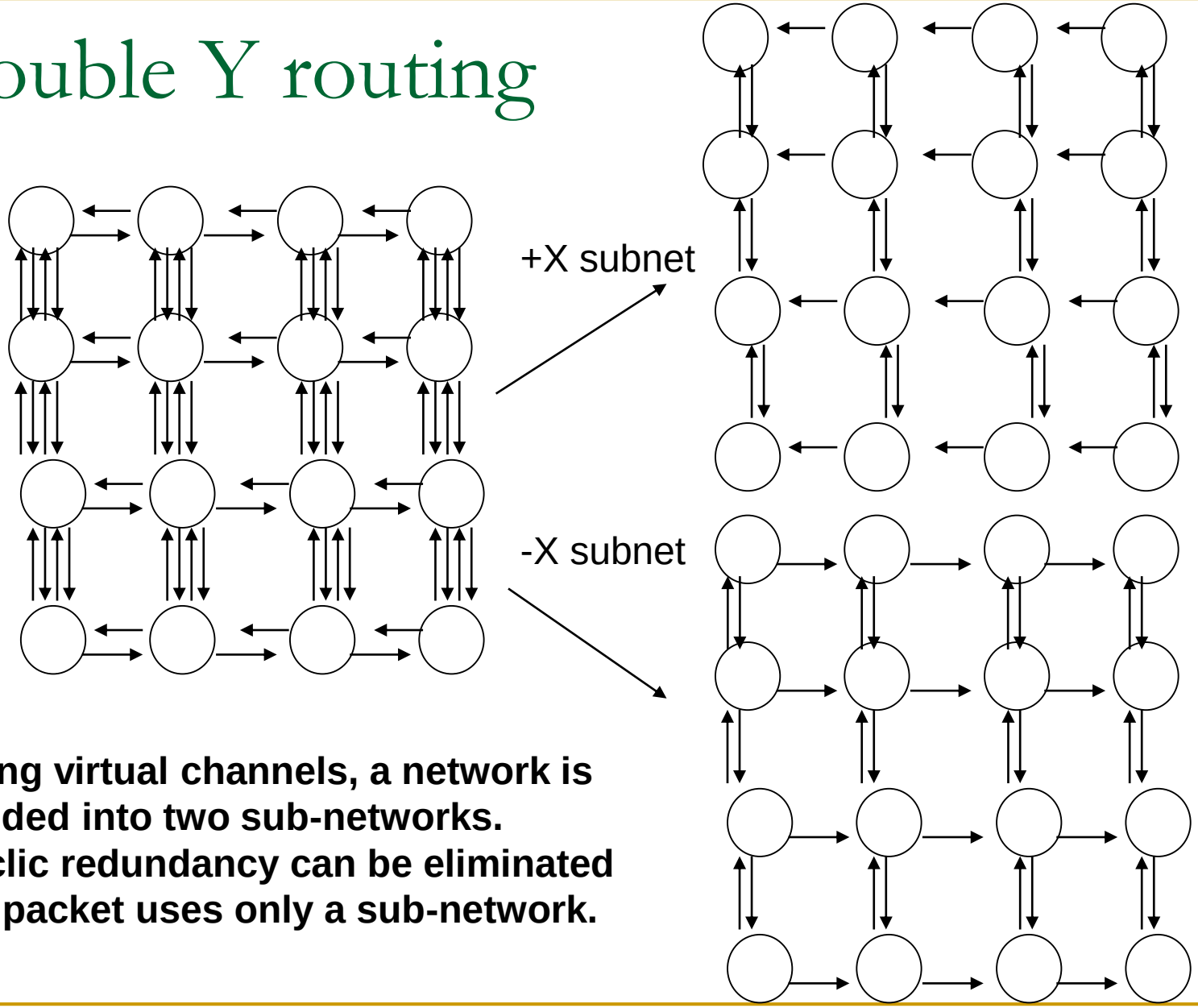
Adaptive routing

- A fixed path is used, and not changed dynamically.
  - Fixed/Deterministic routing
- A path is dynamically changed in order to bypass the congested point (hot spot).
  - Adaptive routing
  - However, deadlock should be avoided.

Adaptive routing techniques

- Using sub-networks
  - double-Y routing
  - Planner Adaptive routing
- Using virtual channels
  - Dimension reversal routing
  - ∗channel (Duato's Protocol)
- Probability based methods
  - Chaos routing
- Restrict the direction of paths
  - Turn model

double Y routing

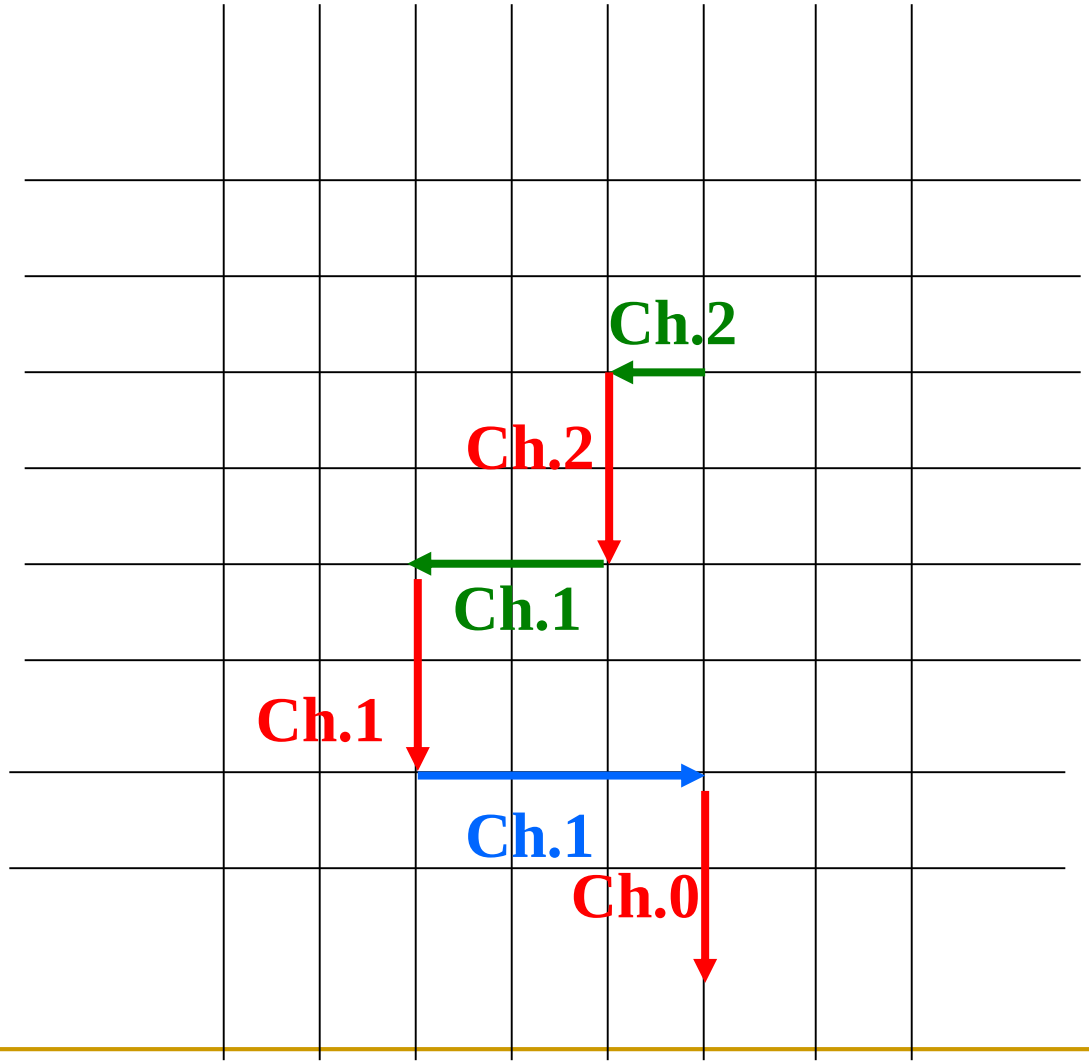


Using virtual channels, a network is divided into two sub-networks. Cyclic redundancy can be eliminated if a packet uses only a sub-network.

Dimension reversal routing

- Providing N virtual channels, and start from channel N.
- E-cube routing is basically used.
- When the packet is routed to the direction which is forbidden in e-cube routing, then decrement the virtual channel number.
- On channel 0, DOR is strictly used.

Dimension reversal routing

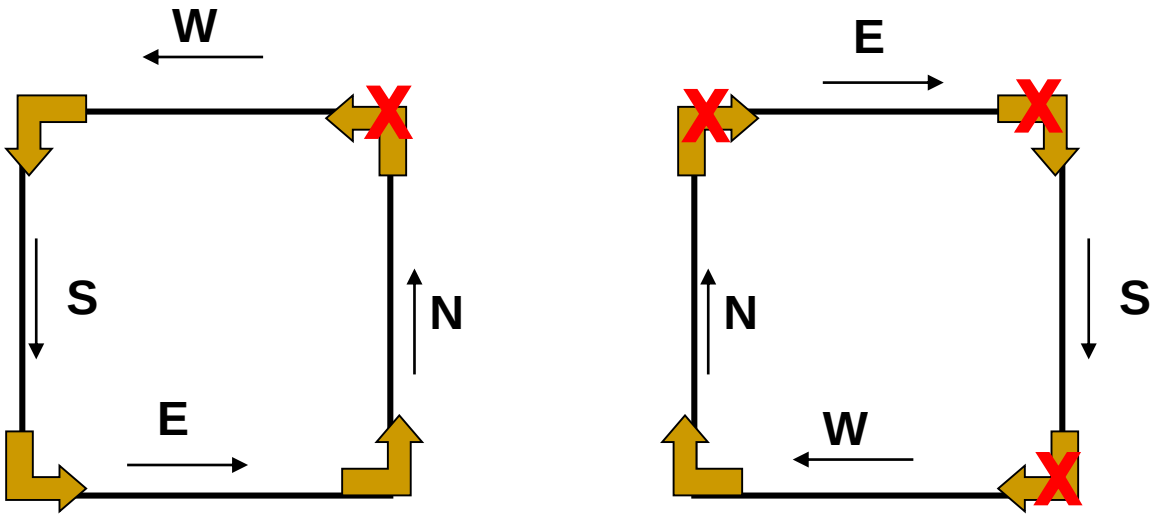


When a packet goes to irregular direction, the virtual channel number is decremented.

Ch. 0 must use e-cube routing

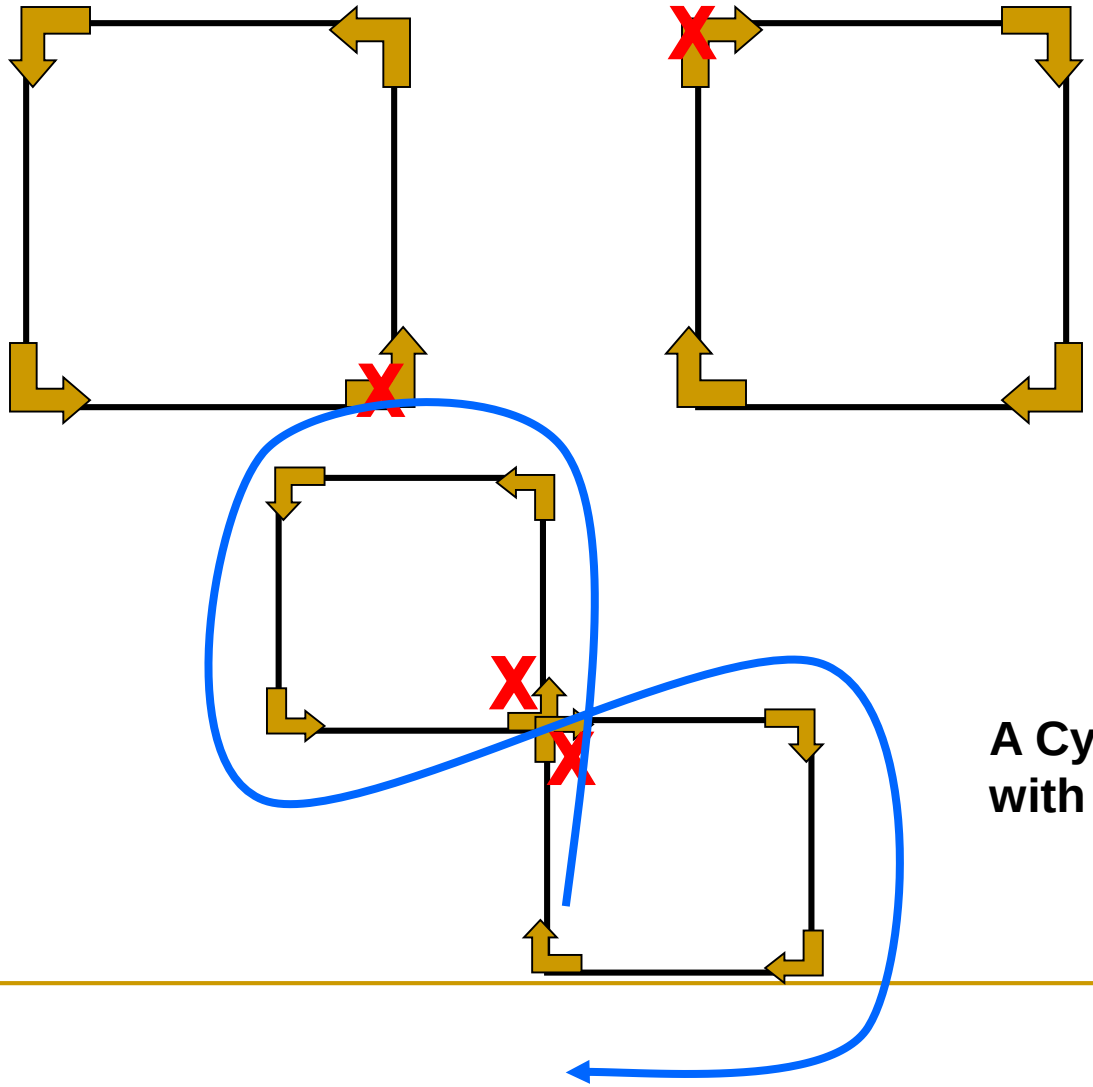
Turn model: Motivation

e-cube routing which allows only W→S→E→N



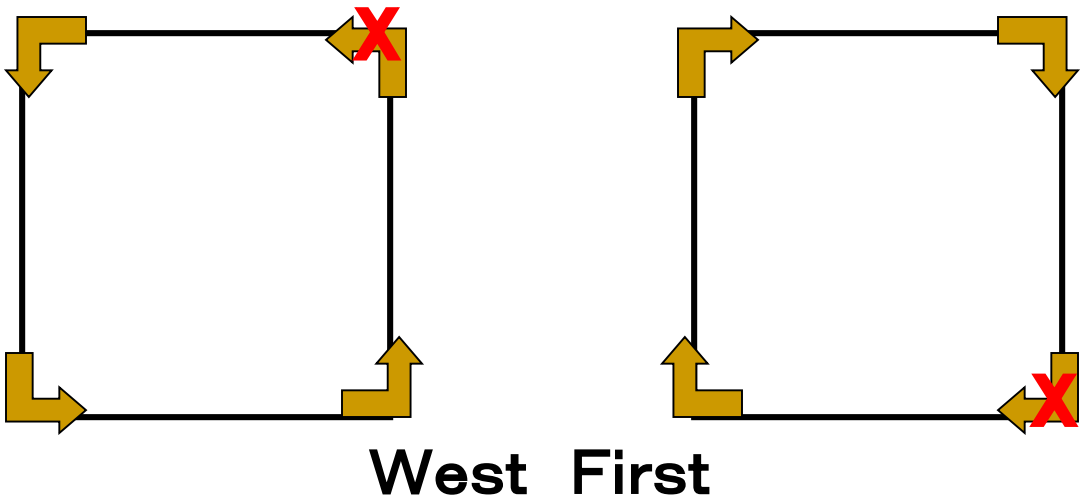
e-cube routing forbids too many turns.
Cycles can be broken with less forbidden turns.

Forbidden turns must be set considering complex combinations

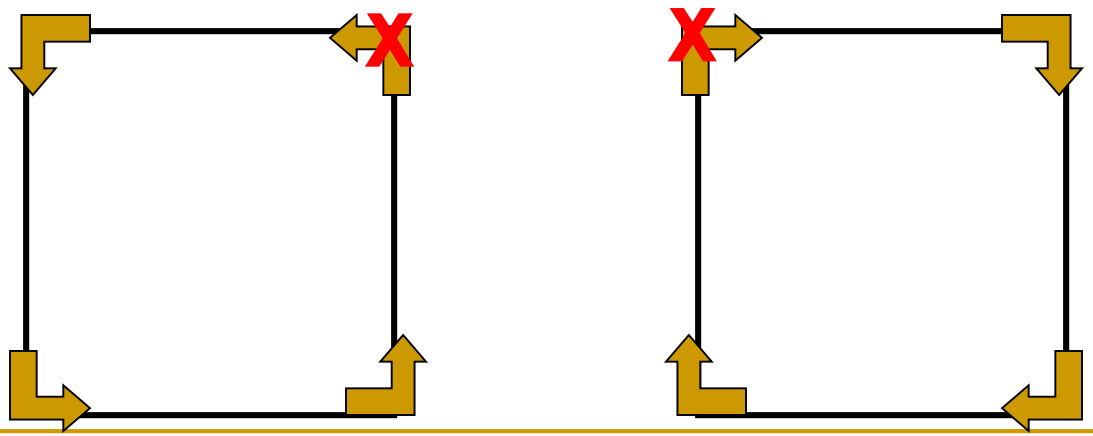


A Cycle is formed with a combination.

Deadlock free set of forbidden turns

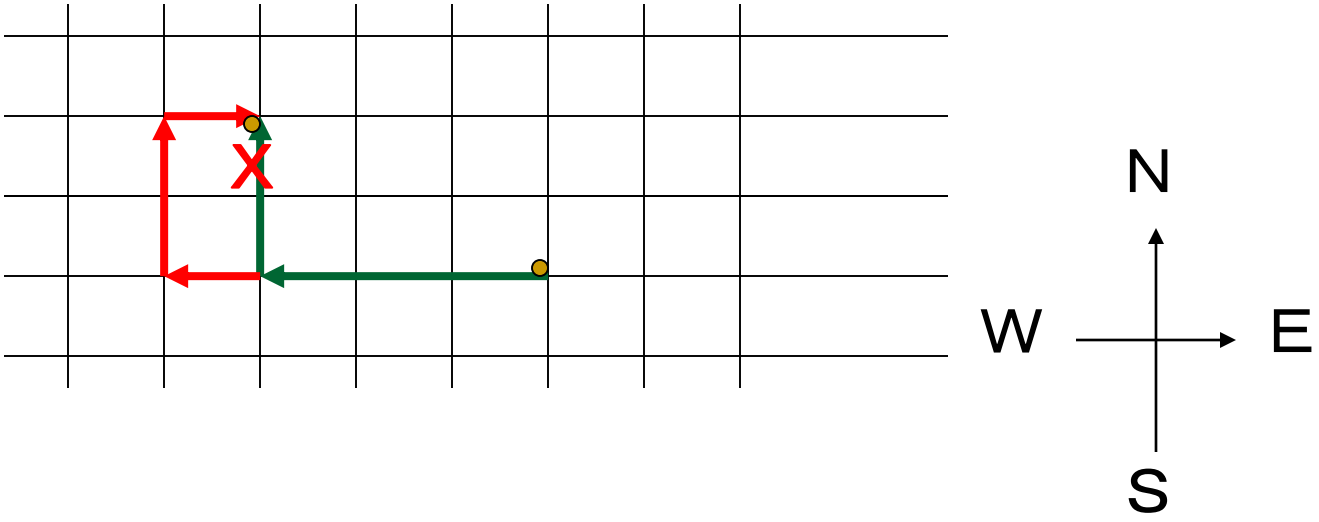


West First



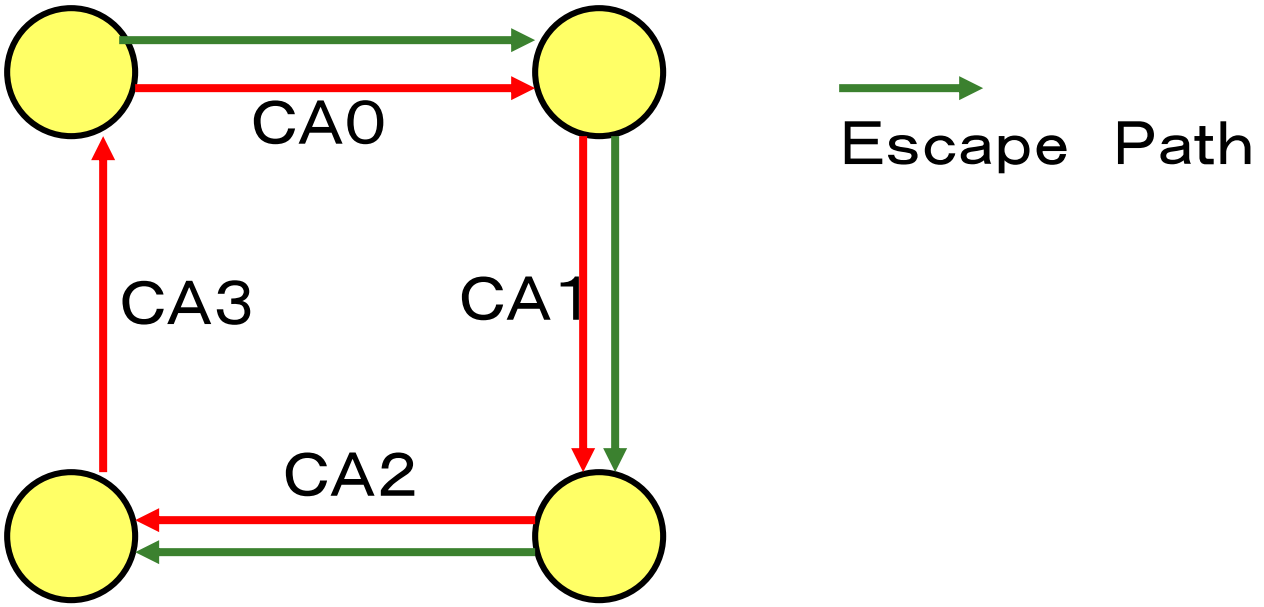
North Last

Congestion avoidance with West First

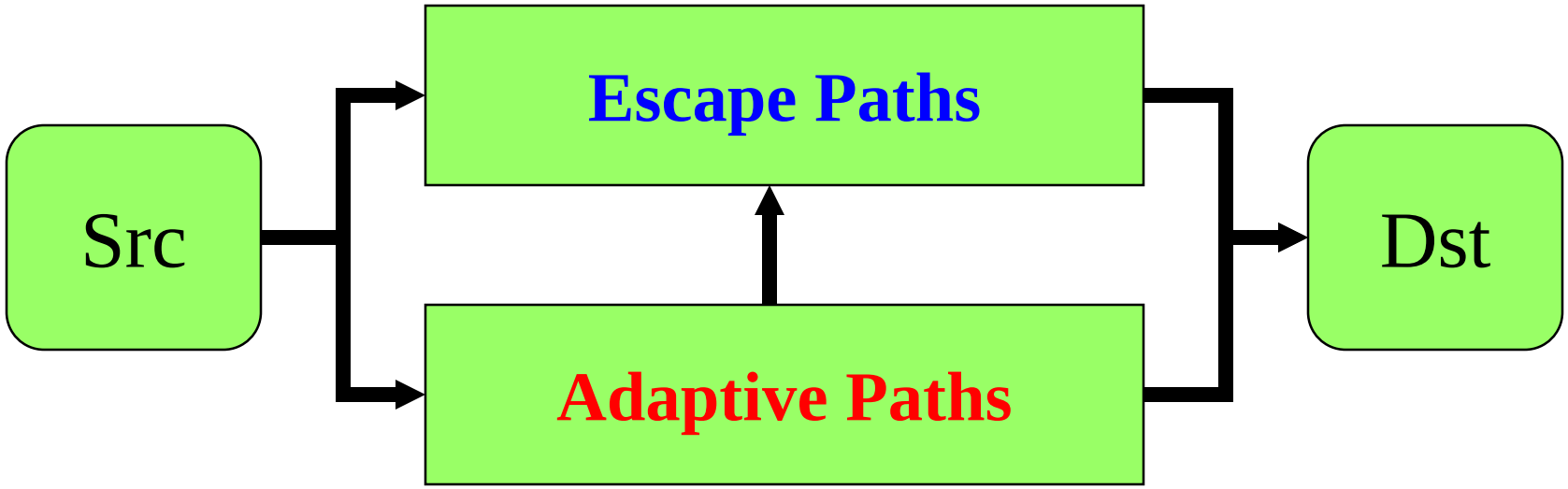


Once a packet goes to West and turns, it cannot go again.

Duato's Protocol (*-channel)



Minimal routing [F.Silla,1997]



- Overall the network, a path without cycle is provided. (Escape path)
- A packet can be moved from Adaptive path to Escape path at any node.
- Once a packet uses Escape path, it cannot go back to Adaptive path

Adaptive routing for Irregular networks: Up*/Down* routing

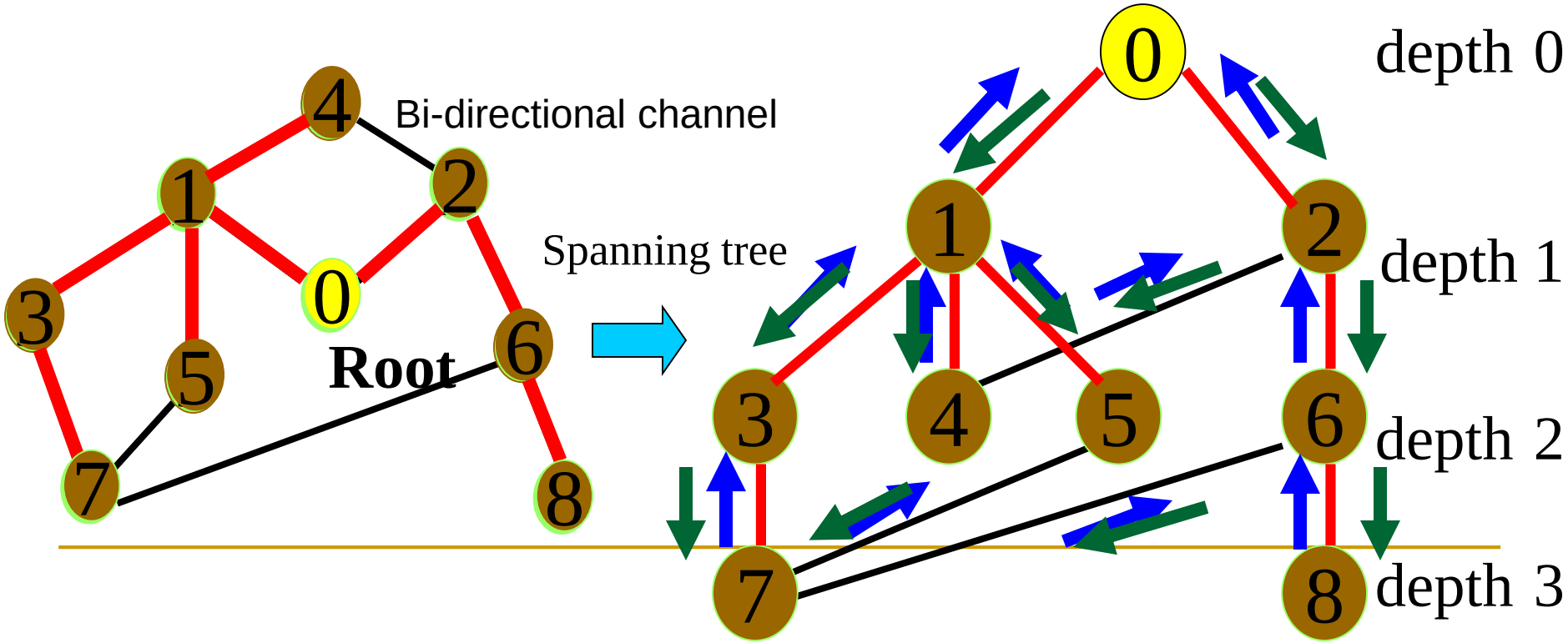
- Typical partially adaptive routing
- Eliminates a channel cyclic-dependency in order to avoid deadlock.
- Algorithm:
  1. Build a spanning-tree.
     - BFS(Breadth First Search)
     - DFS(Depth First Search)
  2. Build an up/down directed-graph.
  3. Set a restriction to avoid the deadlock.

Building an up*/down* directed graph

1. Select a root node.
2. Add the rest nodes to the tree.
3. Allocate the direction (up or down) for each channel.

a. Up direction
destination node is closer to the root node.

b. Down direction



Up*/Down* routing algorithm

- After using up channel(if any), use down channel(if any).
- Non-minimal partially adaptive routing

A cyclic dependency between up and down channels is broken.

 deadlock-free

Drawbacks of up*/down* routing

- Many forbidden turns are concentrated on certain leaf nodes.
- Congestion around root node.
- Improvement proposals
  - Using DFS tree
  - Introducing another dimension

Researches on adaptive routing for regular networks

- Duato's Protocol or Turn model is mostly used.
- For irregular networks, up*/down* routing is also popular.
- Deadlock detection and drop protocol vs. deadlock free routing.

Summary: adaptive routing

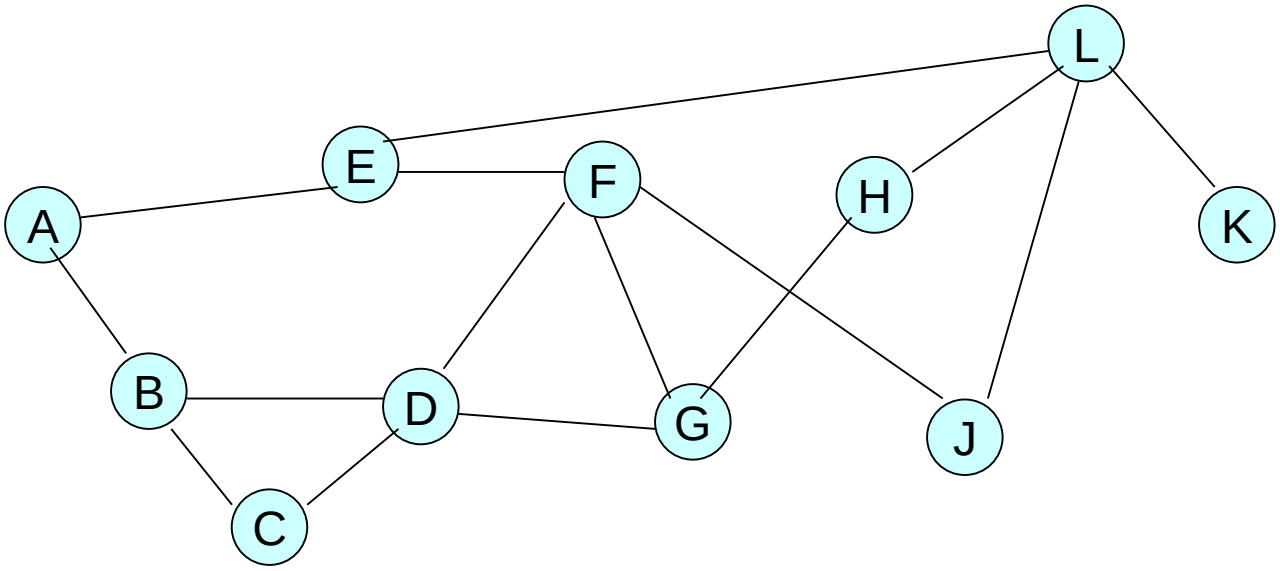
- Drawbacks:
  - FIFO assumption is not guaranteed.
  - Difficult to debug, if trouble occurs.

However, the benefits will overcome the drawbacks.

Recent high performance networks use adaptive routing.

Exercise

- For an irregular network shown below:
  - Pick up a node as a root node and draw a spanning tree.
  - Add up/down direction to each link.
  - Show the longest path between a source and destination node.



Glossary2

- Adaptive routing:適応型ルーティング。ネットワークの混雑状況に応じて動的に経路を変えるルーティング。変えることができない方法をDeterministic routing（固定ルーティング）と呼ぶ。
  - 経路を勝手に変えるとデッドロックしてしまうので、様々な方法が提案されている。Double Y　Routing、Dimension Reversal Routing,Turn model, Duato's protocolは、全てこの方法の名前。
  - Minimal routingつまり最短経路を必ず選ぶ方法と、non-minimal routing最短経路でなくても迂回可能な方法がある。
- SAN(System Area Network): PCクラスタなどで用いられるネットワーク、代表選手はMyrinet、QsNet。ちなみにサーバー屋さんは、SANを（Storage Area Network)のことだと思っているので注意。
- Irregular　Network:不規則なネットワーク、多くのSANでは規則的ではなく、不規則なネットワークを許容する。これはPCクラスタなどでは、場合に応じて、ノードが欠けたりするため。