第三章 MPLS技术
PART-3

张喆
zhezhang@njupt.edu.cn
通信与信息工程学院

第三章 MPLS技术

- 3.1 MPLS的标记交换原理及LDP协议
- 3.2 MPLS中的流量工程
- 3.3 MPLS VPN
- 3.4 VPLS
- 3.6 GMPLS
- 3.6 高速路由器的设计

为什么需要高速路由器

- 网络层主要功能: 将数据从一点快速传输至另一点
  - 存储转发-> 路由器

endpoint

Network layer

endpoint

Analogy: postal system

Addressing (IPv4)

Locate, not identify

路由器技术发展

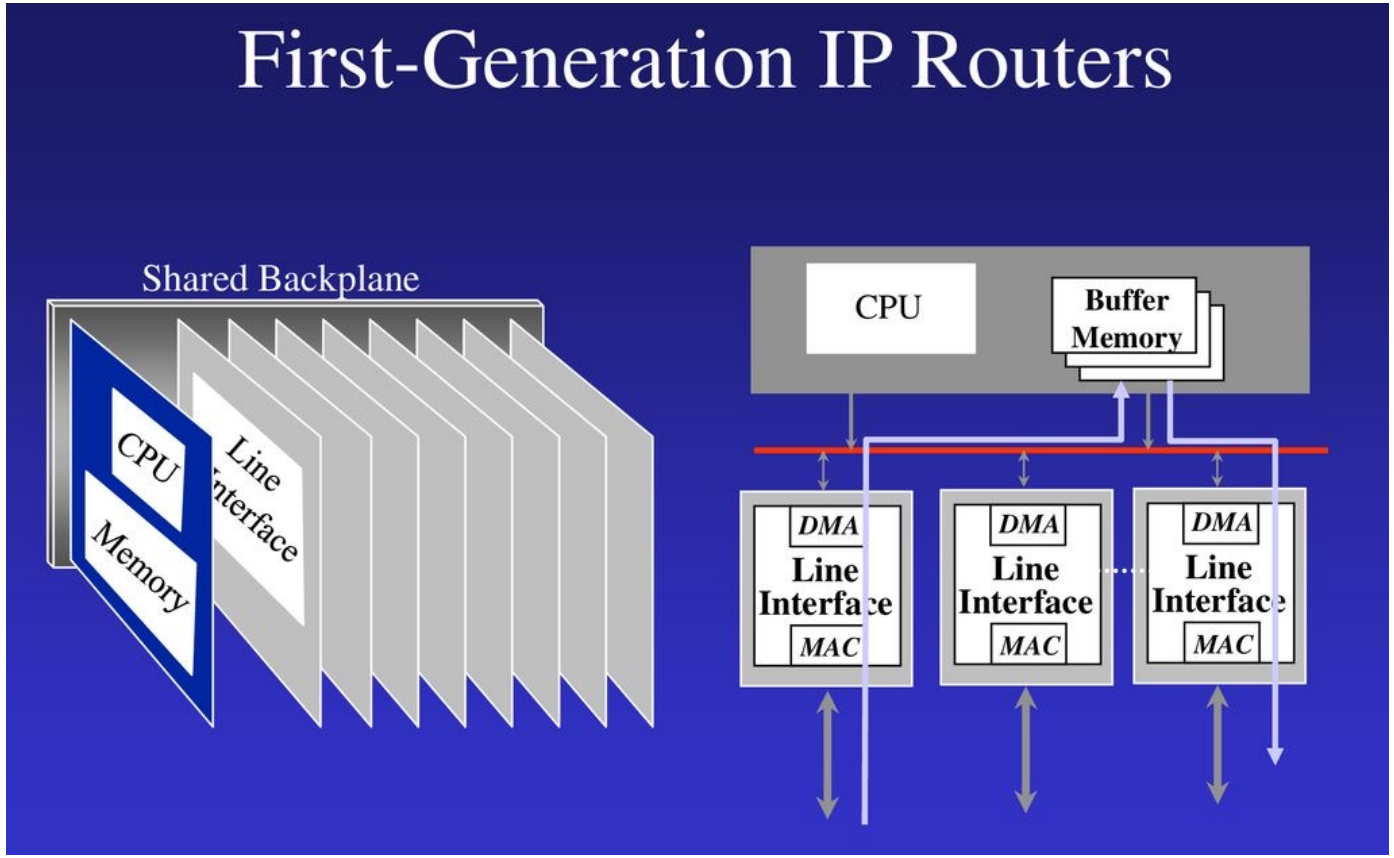
- 第一代：单总线单CPU结构路由器
- 第二代：常用的路由信息采用Cache技术保留在业务接口卡上
- 第三代：采用了并行处理结构
- 第四代：ASIC分布转发、Crossbar结构路由器
  - 全部采用硬件实现，分组直接从输入端经过交叉开关(Crossbar)流向输出端
- 第五代：网络处理器分布转发、Crossbar结构路由器
  - 关键的IP业务流程处理采用网络处理器，而路由查表、分组分类等复杂操作采用硬件协处理器

第一代IP路由器

- Bosack and Lerner
- 于1984年创立了Cisco

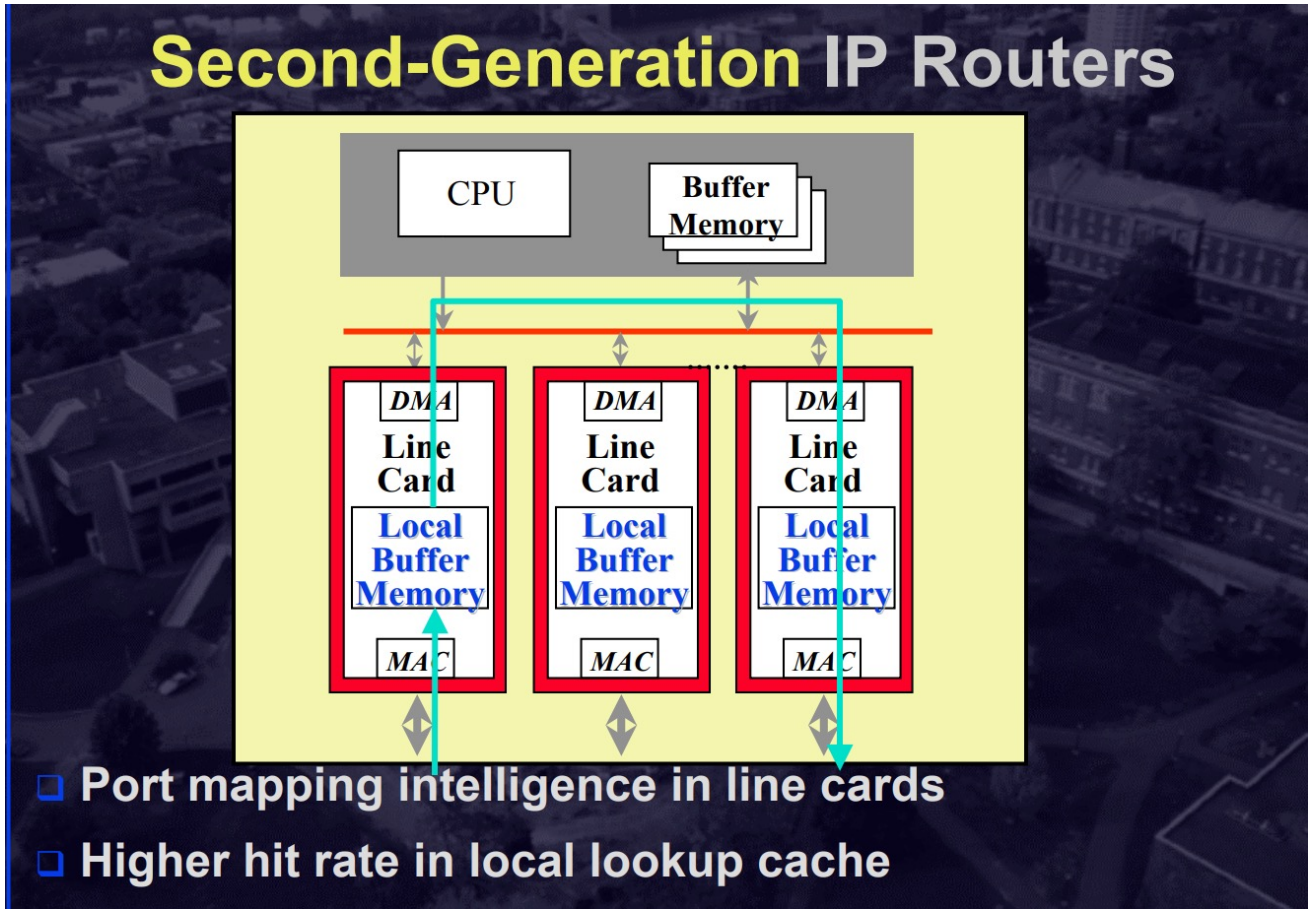
第一代IP路由器

- 设计简单（可用一台计算机插多块网卡来实现）
- 集中转发，总线交换：
  - 单一CPU完成所有路由操作
  - 接口与CPU通过总线连接
- 数据面与控制面高度耦合
- 路由：采用距离矢量算法



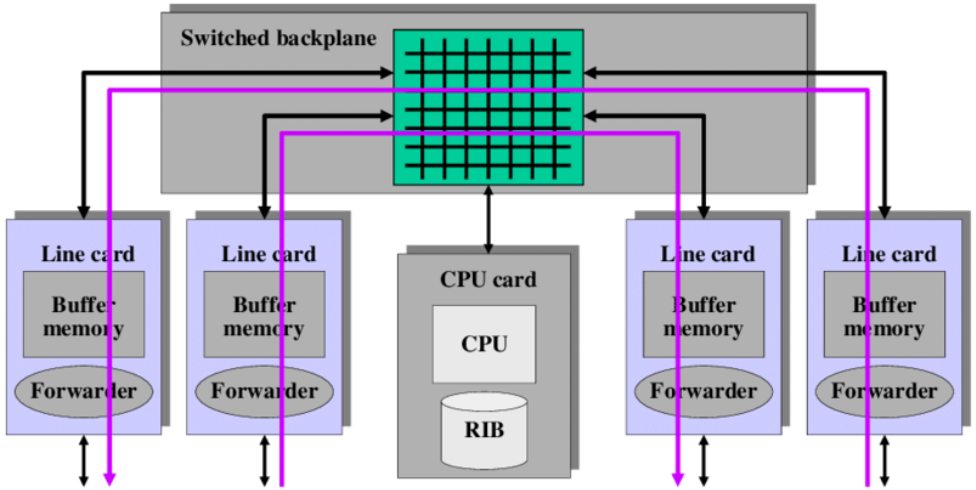
第二代IP路由器

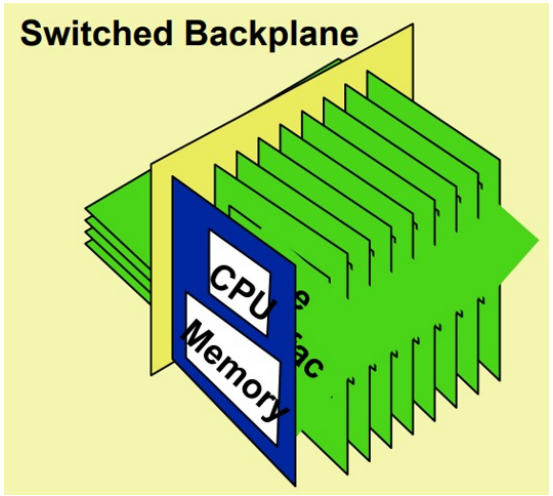
- 网络规模增大->CPU和总线负担
- Solution：
  - 集中+分布转发：
    - 采用cache技术
    - 在接口上缓存常用路由信息
    - 在cache中找不到的报文传输->CPU
  - 接口模块化



第三代IP路由器

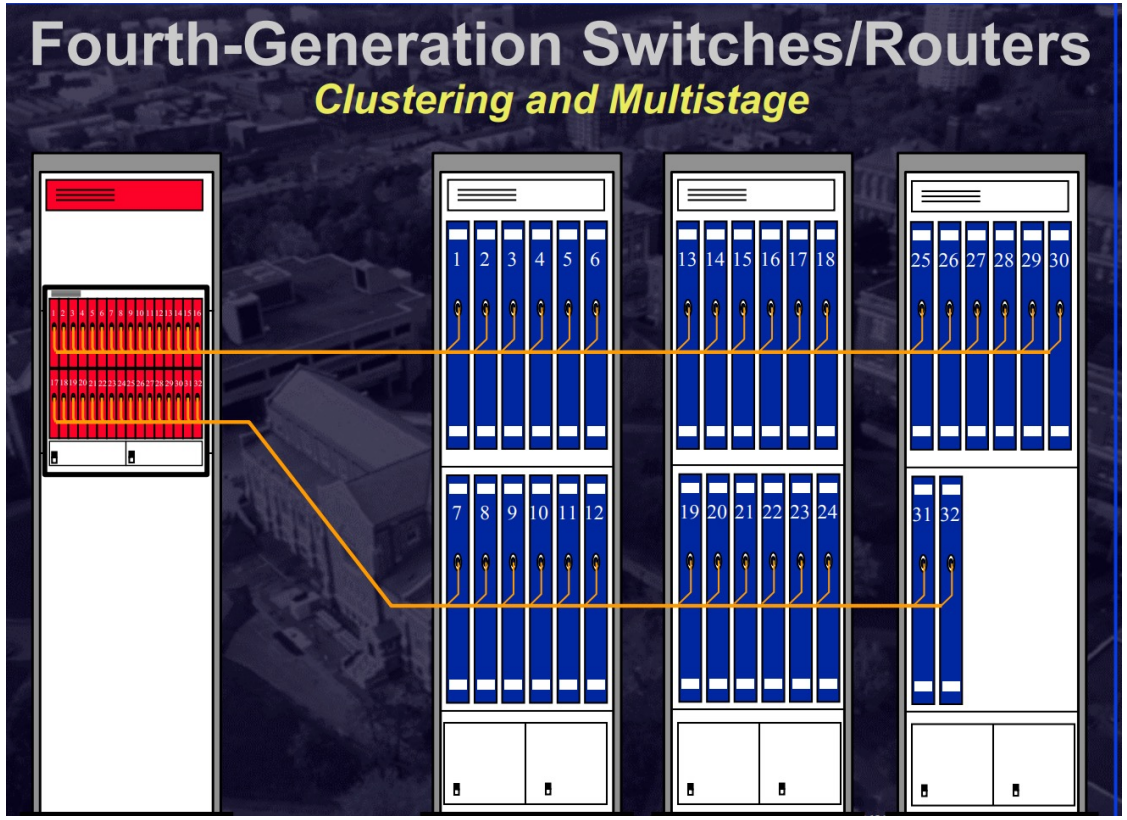
- Web技术的出现->用户访问面激增
  - Cache已无法满足
  - 总线与CPU瓶颈再次出现
  - 接口数量不足的问题也同时出现
- Solution：
  - 分布转发，总线交换
  - 分布式结构：
    - 控制面与数据面诞生（分离）
    - 控制面：负责路由表构建与维护
    - 数据面：负责数据转发





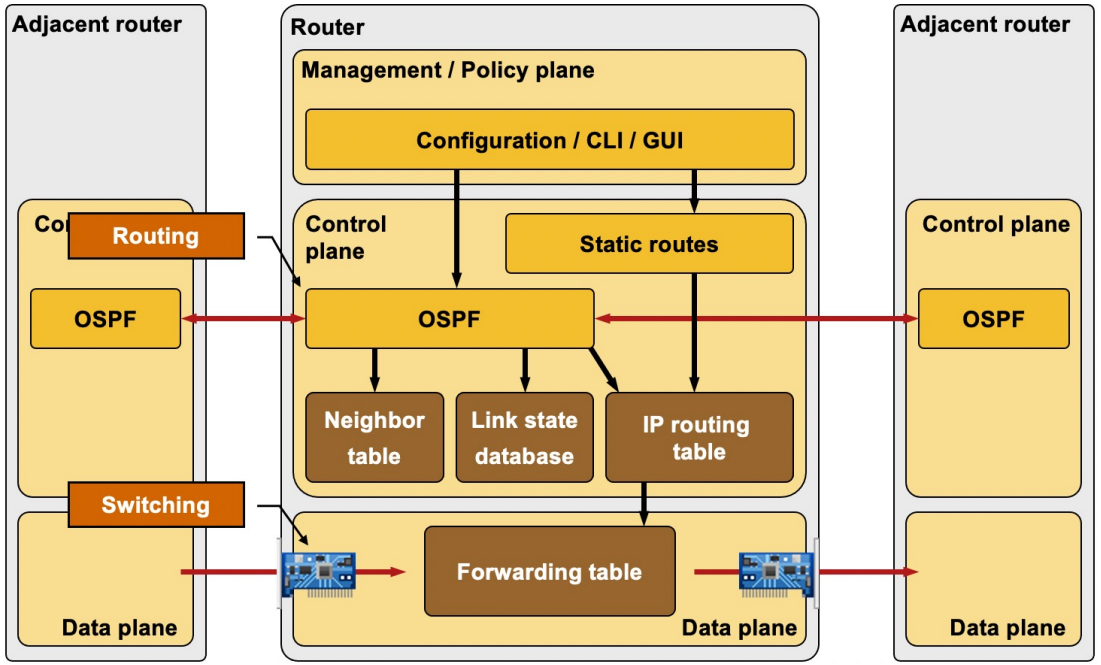
第四代IP路由器

- 九十年代中后期，网络流量呈指数级增长
- 传统CPU处理速度无法满足（2.5G POS端口线速度流量约为6.5 Mbps）
- Solution：
  - ASIC（专用集成电路）分布转发，网络交换转发过程用硬件来实现
  - 交换结构采用crossbar
  - 线速度可以达到千兆



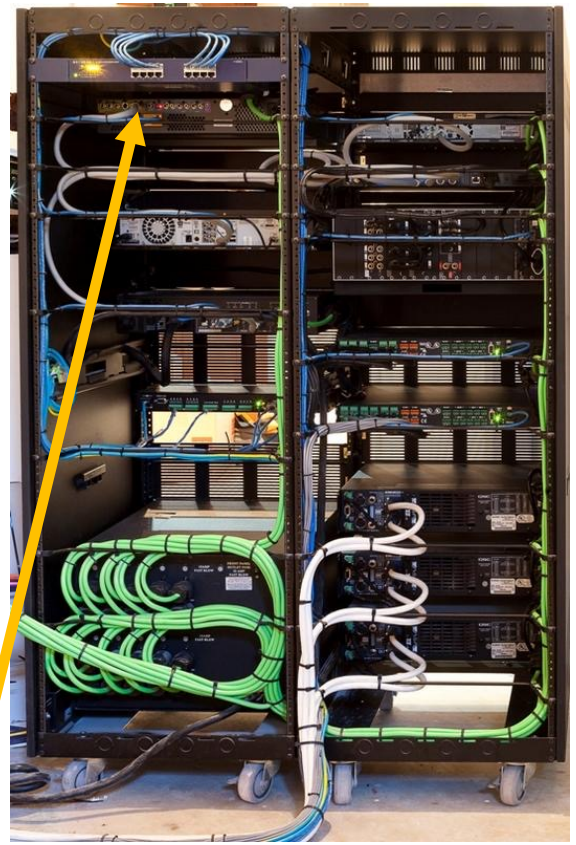
第五代IP路由器

- 新世纪以后，网络管理、用户管理、业务管理、MPLS、VPN、流量工程等新技术加入路由器。
  - 部署复杂性：如配置MPLS，每新增一个PE，都要回去修改每个涉及到的PE
  - 升级复杂：升级新功能极其复杂
- Solution：
  - SDN（software-defined networking）技术
    - 采用可编程器件（网络处理器）
    - 控制面与数据面解耦
      - 控制面可以运行在通用的服务器或云平台
      - 数据面保留在路由器硬件中



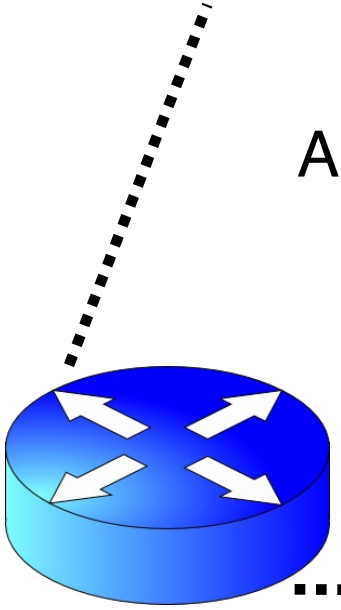
路由器种类

Access routers



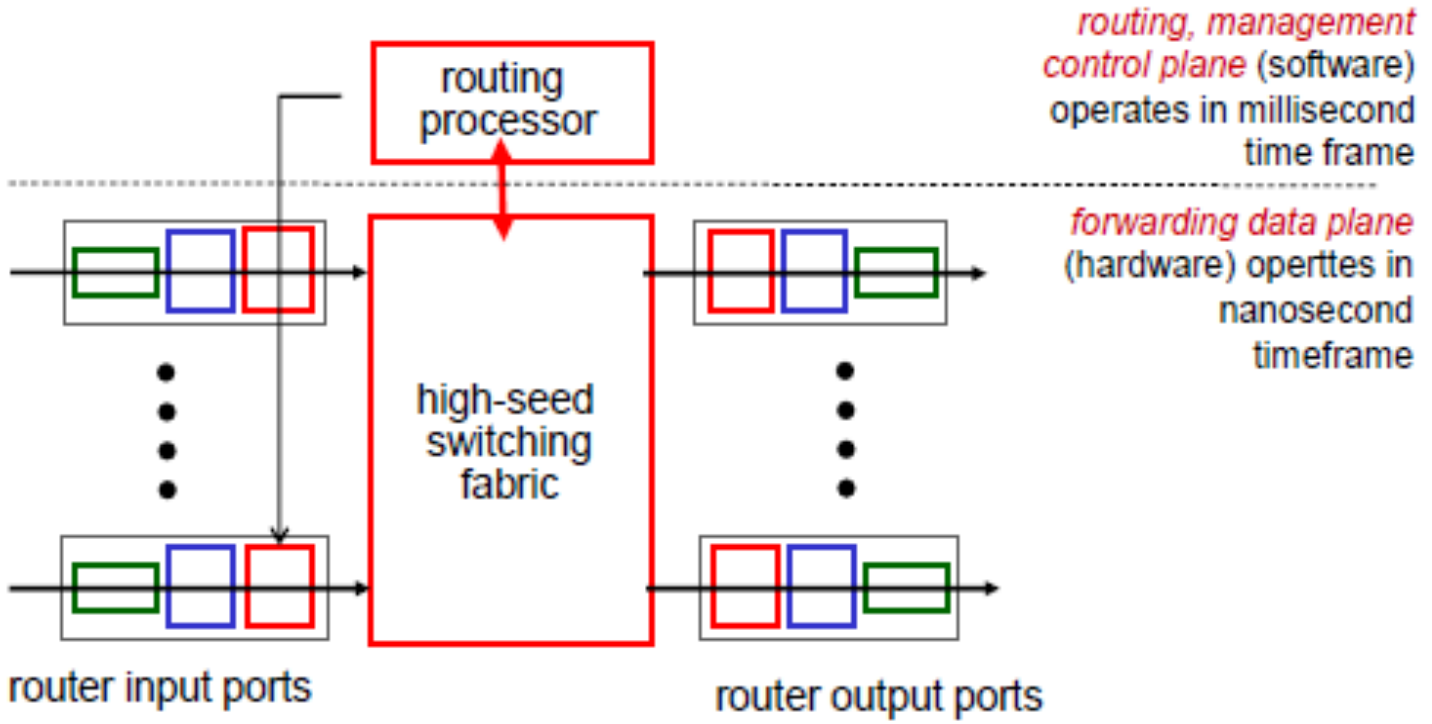
Data center top-of-rack switch

Core router



路由器基本结构

- 路由转发
- 输入输出端口
  - 缓冲队列
- 交换结构
  - Memory（共享内存）
  - Bus（总线）
  - Crossbar（空分交叉开关）



路由表VS转发表：

- 路由表：
  - 目标地址，掩码，下一跳
- 转发表：
  - 目的地址，出接口，下一跳MAC地址

目的网络	下一跳地址	出接口
10.0.0.0/8	192.168.1.1	Eth0/1
192.168.1.0/24	192.168.1.254	Eth0/1
192.168.2.0/24	192.168.1.254	Eth0/1
192.168.3.0/24	192.168.1.254	Eth0/1
0.0.0.0/0	192.168.1.254	Eth0/1

目的地址	出接口	下一跳MAC地址
10.0.0.1	Eth0/1	00:11:22:33:44:55
192.168.1.0/24	Eth0/2	00:11:22:33:44:66
192.168.2.0/24	Eth0/3	00:11:22:33:44:77
0.0.0.0/0	Eth0/0	00:11:22:33:44:88

转发表由路由表生成

最长前缀匹配

目的网络	下一跳地址	掩码长度
192.168.1.0	10.0.0.1	24
192.168.1.0	10.0.0.2	16
192.168.0.0	10.0.0.3	16
0.0.0.0	10.0.0.4	0

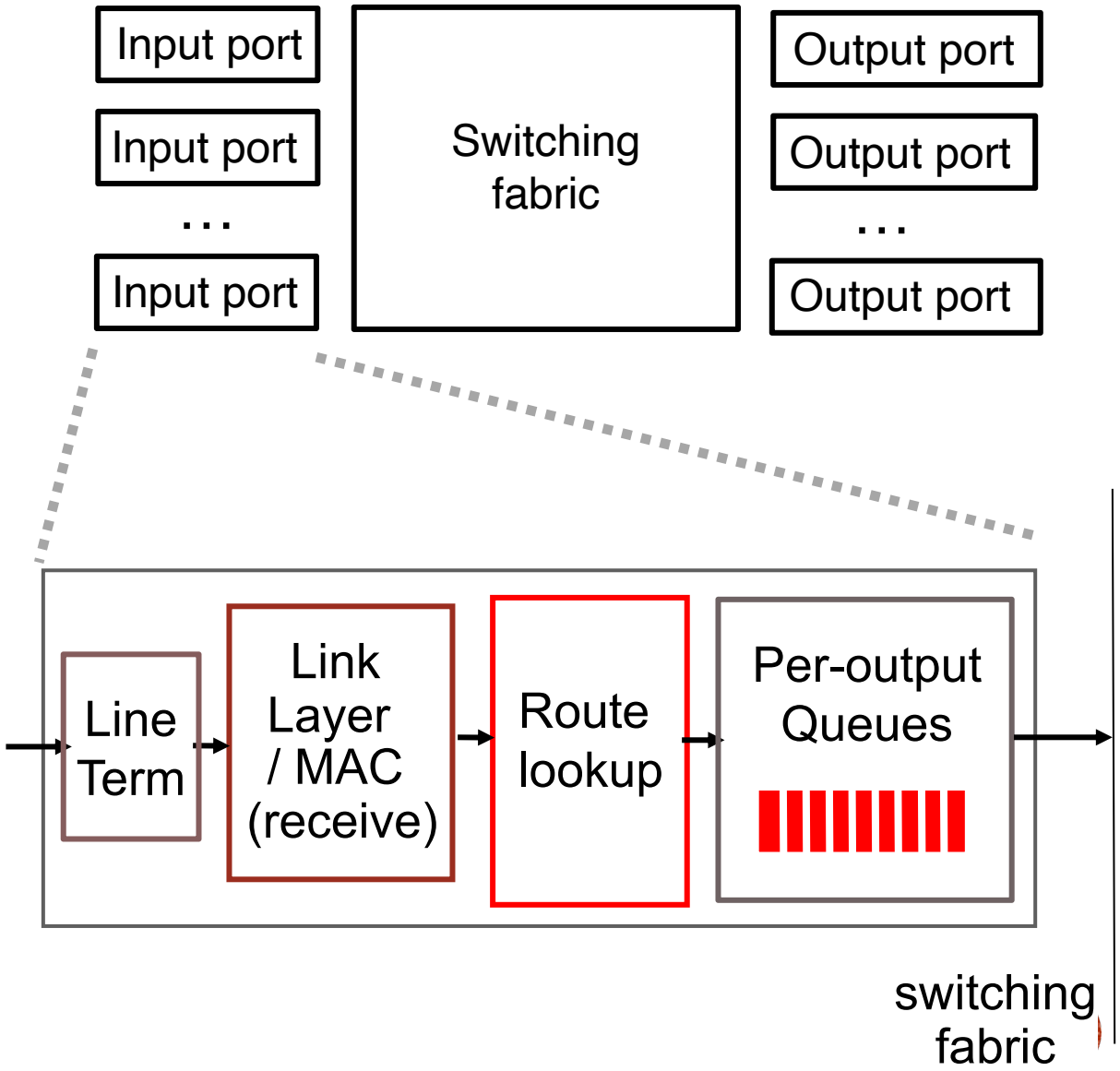
假设有一个数据包的目的地址为192.168.1.10，与第一个条目完全匹配

提高路由器的速度

- 路由器的设计涉及到许多方面：
- 路由器的体系结构
- 路由协议的处理
- 路由表的存储与查找、缓冲区的设计
- 数据包的输出调度等

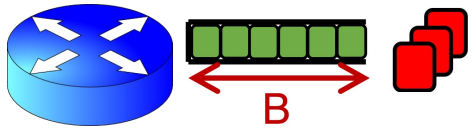
输入

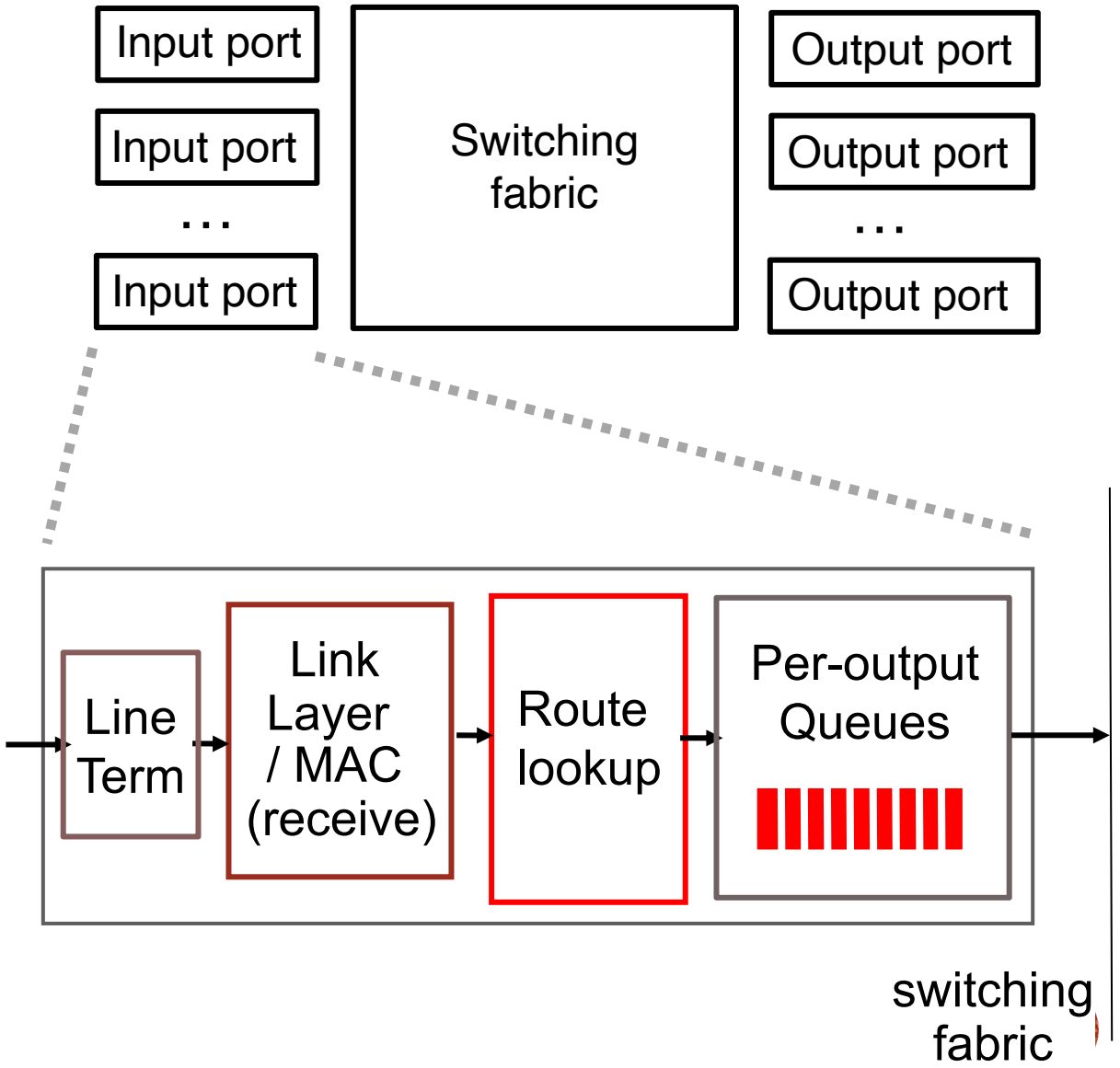
- Line termination: receives physical (analog) signals and turns them into digital signals
- Rate of link connecting to a single port termed line speed or line rate (modern routers: 100+ Gbit/s)
- Link layer: performs medium access control functions (e.g., Ethernet)



输入

- Route lookup: high-speed lookup of which output port the packet is destined to
- Goal: must complete this processing at the line rate
- Queueing: packets may wait in per-output-port queues if packets are arriving too fast for the switching fabric to send them to the output port





缓冲策略（排队策略）

- 输入缓存
- 在交换结构的每一条入线上设置缓冲器，通常遵循FIFO；

*问题：队首阻塞现象：“head of line blocking”(*每个输入端的FIFO首先处理的是在队列中最靠前的数据，而这时队列后面的数据对应的出口缓存可能已经空闲，但因为得不到处理而只能等待)
吞吐率低；

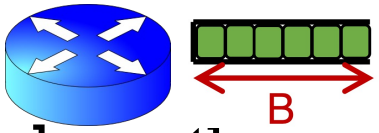
有多种改进方法：

采用FIRO（First in Random out），

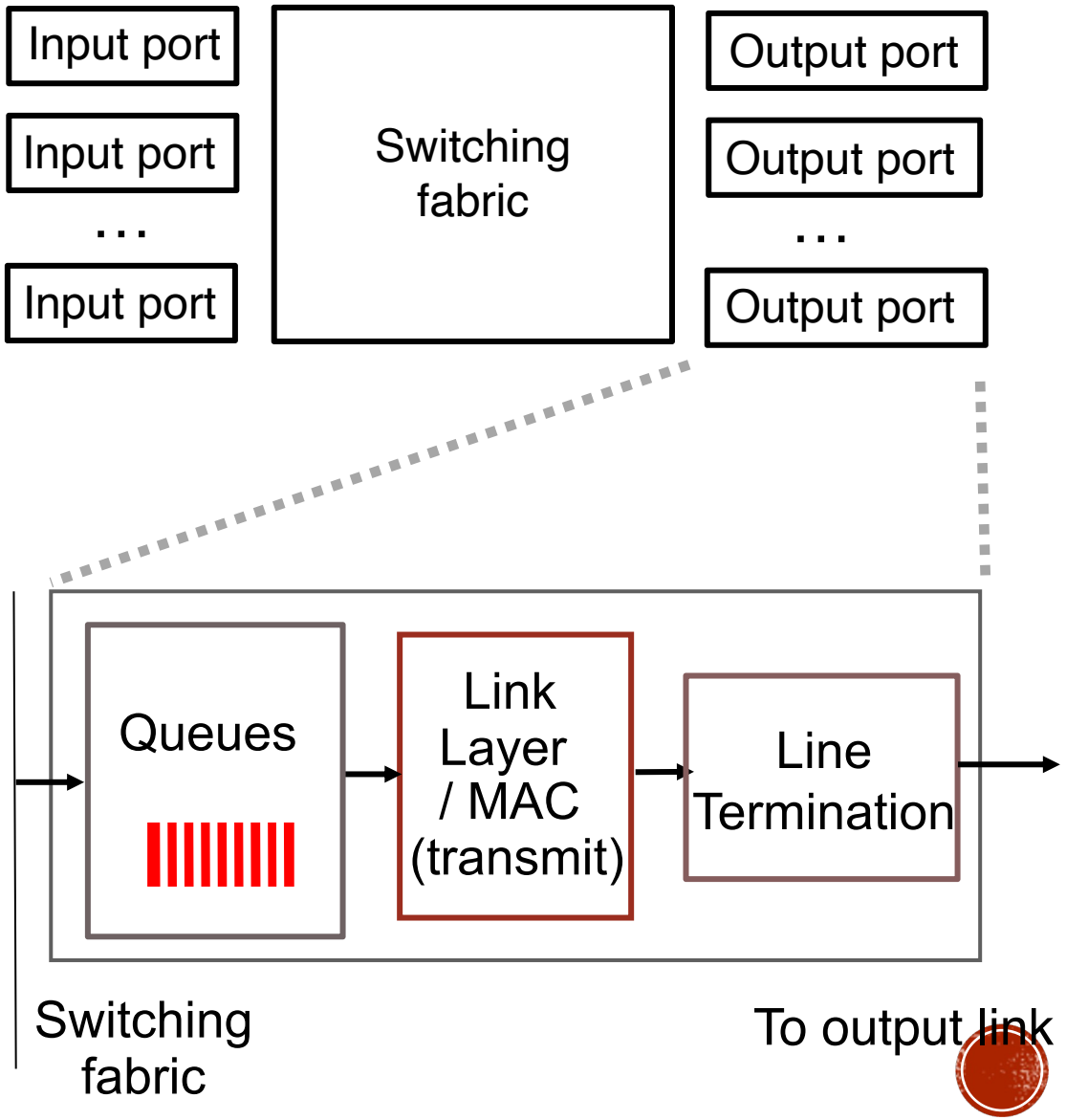
多重输入缓冲等（一条入线设置多个FIFO队列）；

输出

- Components in reverse order of those in the input port

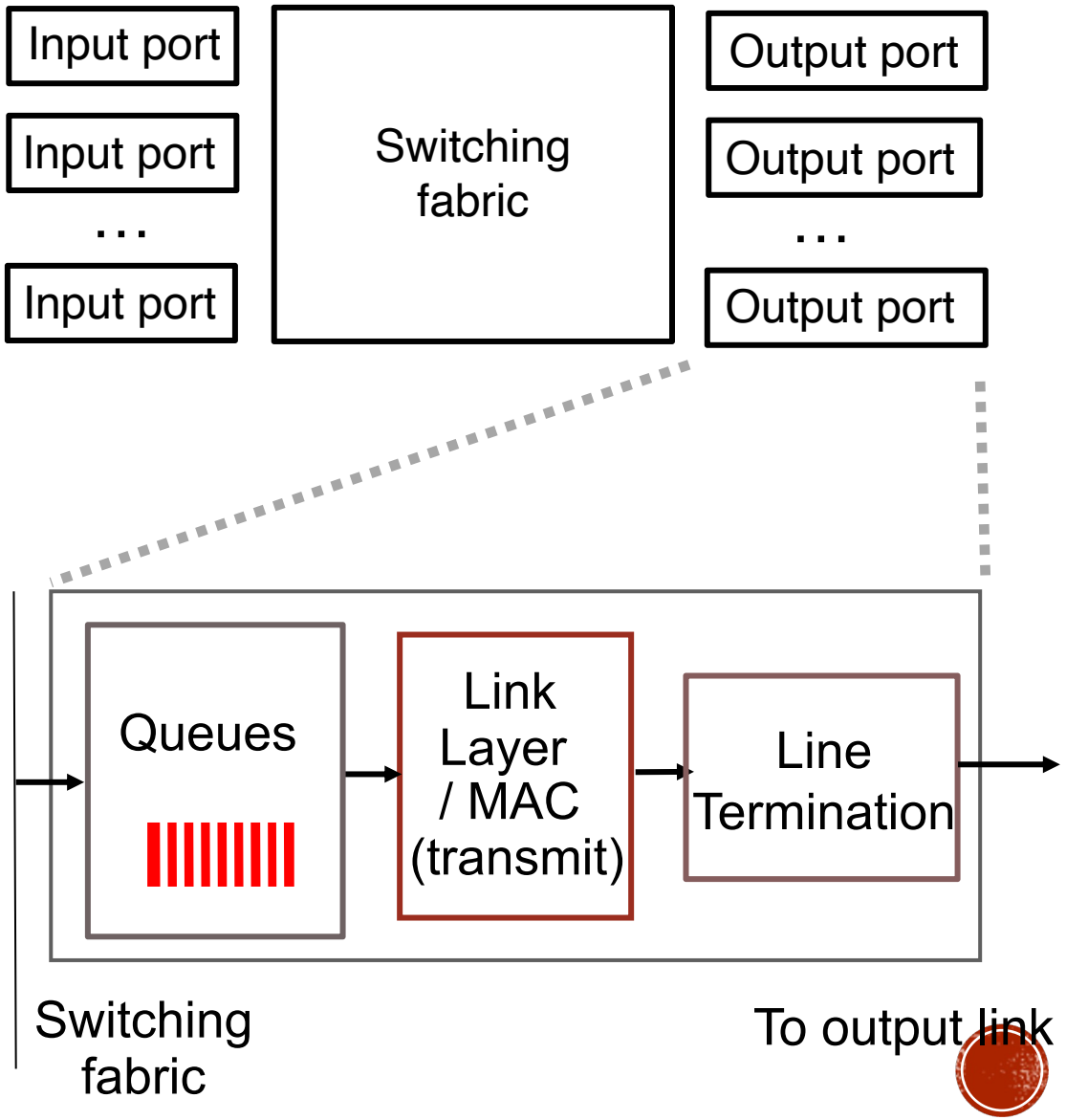


- This is where most routers have the bulk of their packet buffers
  - Recall discussions regarding router buffers from transport
- MGR (multi-gigabit router) uses per-port output buffers, but modern routers have shared memory buffers
  - More efficient use of memory under varying demands



输出

- Two important policy decisions
- Scheduling: which among the waiting packets gets to be transmitted out the link?
  - Ex: First-In-First-Out (FIFO)
- Buffer management: which among the packets arriving from the fabric get space in the packet buffer?
  - When and which packet to drop
  - Ex: Tail drop: later packets dropped first



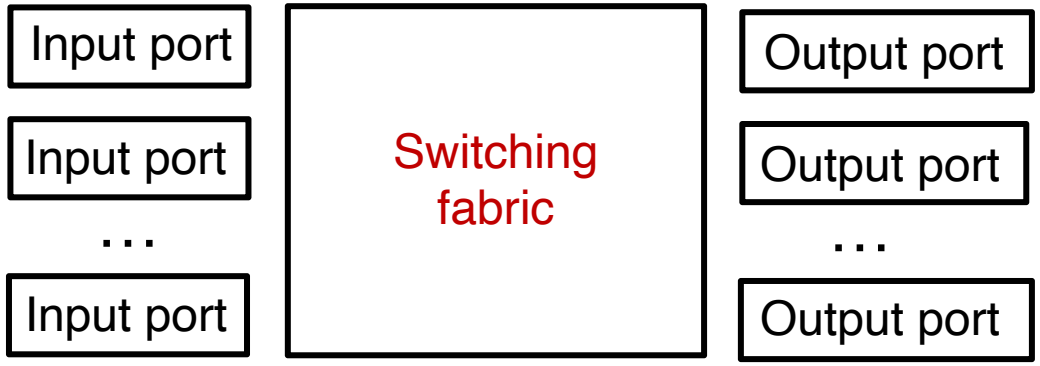
交换结构分类

- 单级交换结构：
  - 共享内存（memory）
  - 总线（bus）
  - crossbar
- 多级交换结构：
  - Banyan
  - Clos
  - Butterfly

多级交换网络的内部阻塞

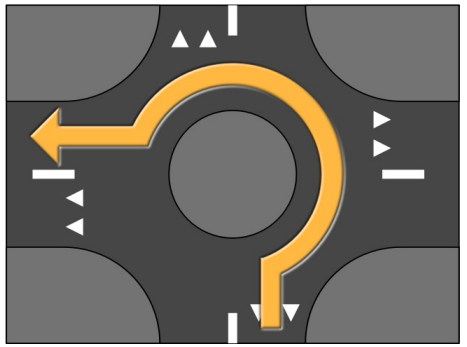
- 若出、入线空闲，但因交换网络级间链路被占用而无非接通的现象，称为多级交换网络的内部阻塞。
- 严格非阻塞网络：
  - 不管网络处于何种状态，任何时刻都可以在交换网络中建立一个连接，只要这个连接的起、终点是空闲的，而不会影响网络中已建立的连接。
- 可重排无阻塞交换网络:对原有连接重新选路后无阻塞
- 广义无阻塞交换网络:遵循规律选路可无阻塞

非阻塞交换结构

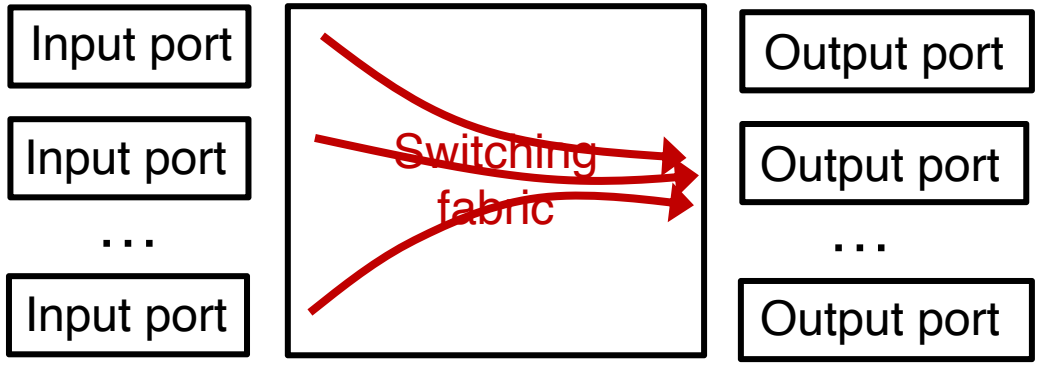


- 高速交换结构被设计为非阻塞的（nonblocking）：
  - 如果一个输出端口是“空闲的"，一个输入端口总是可以向它传输，而不会被交换结构本身所阻挡。
- Crossbars也被设计为非阻塞的

- Shared memory can be designed to be nonblocking if memory is optimized to be fast enough

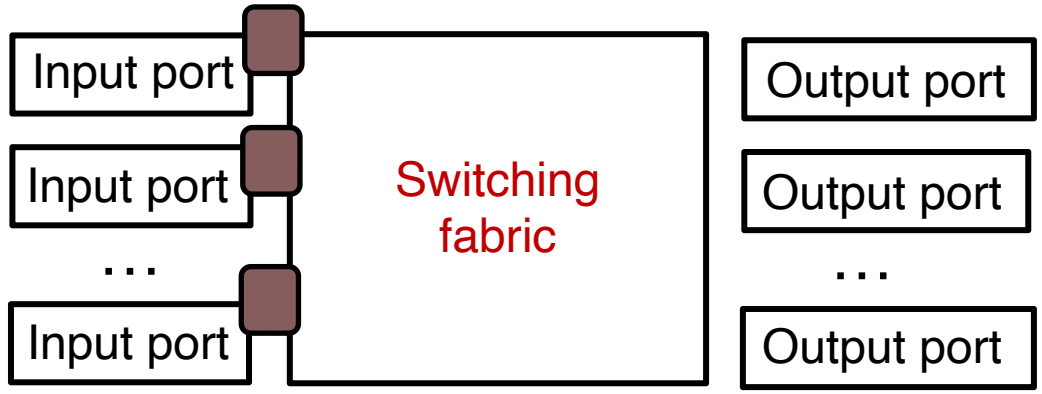


非阻塞交换结构

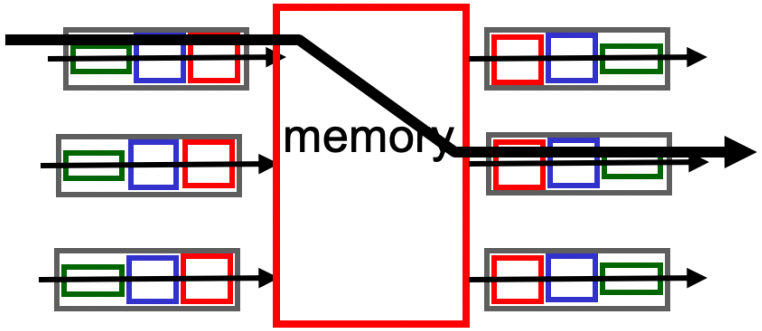


- 在非阻塞结构中，由于交换结构的原因，不会形成队列。
- 有了非阻塞结构，就不会因为输入端口或交换结构的低效率而出现队列。
- 队列只会由于对输出端口的争夺而产生
  - **Fundamental, unavoidable, given the route**
- 通常队列只会在输出端口产生
  - 但如果对输出端口的争夺很激烈，也可以向输入端 "反压"。
  - i.e.:由于缓冲区已满，无法将数据包移至输出端口，所以需要在输入端口加buffer

典型的单级交换结构

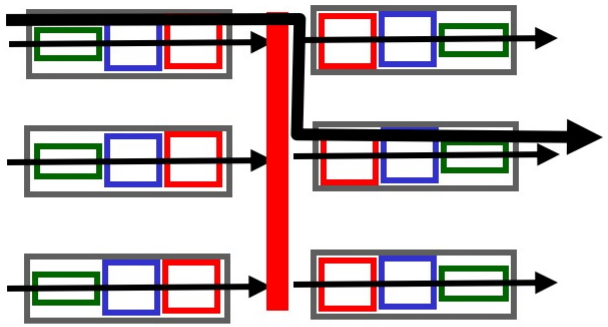


Fabric goal: Ferry as many packets as possible from input to output ports as quickly as possible.



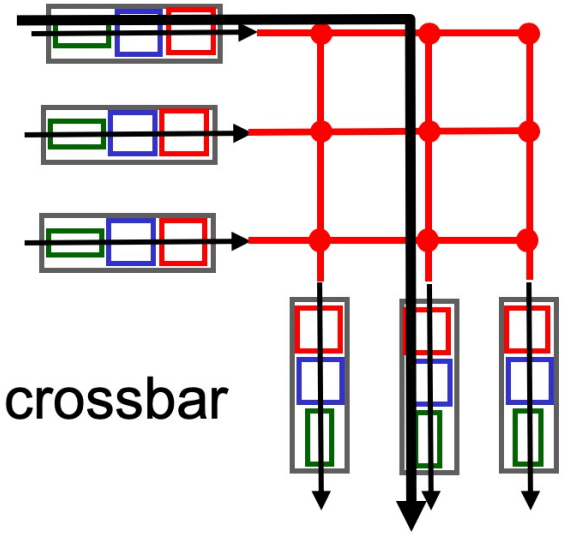
memory

Input port writes packets into shared memory.
Output port reads the packet when output link ready to transmit.



bus

Single shared channel to move data from input to output port.
Easy to build buses; technology is quite mature.

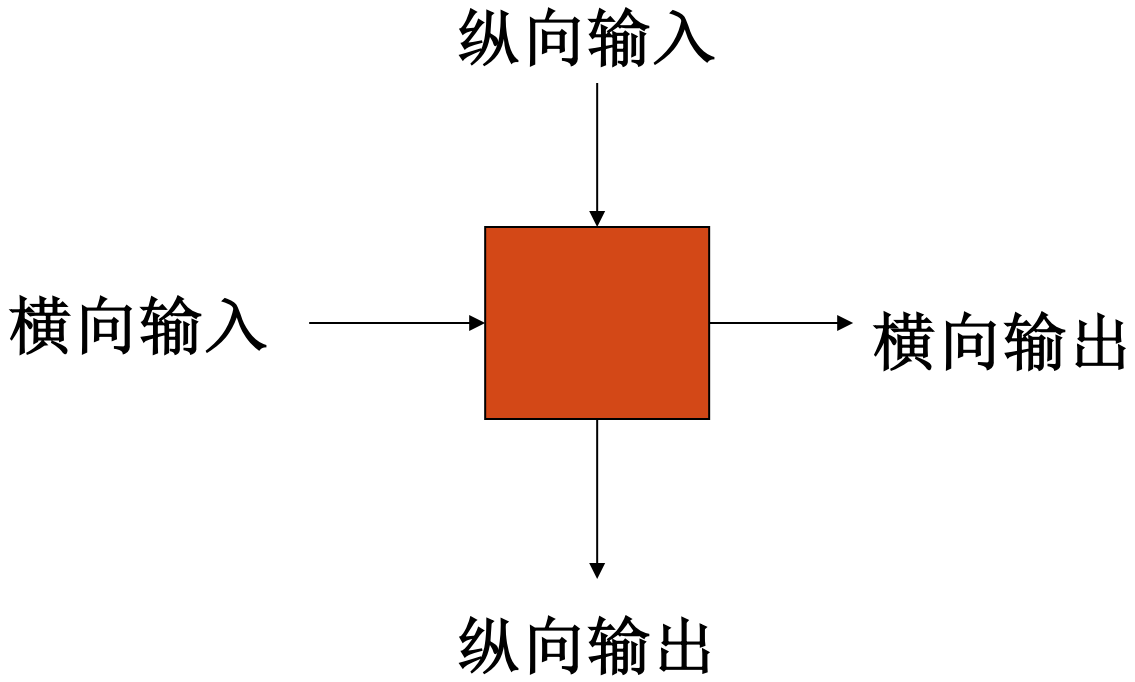


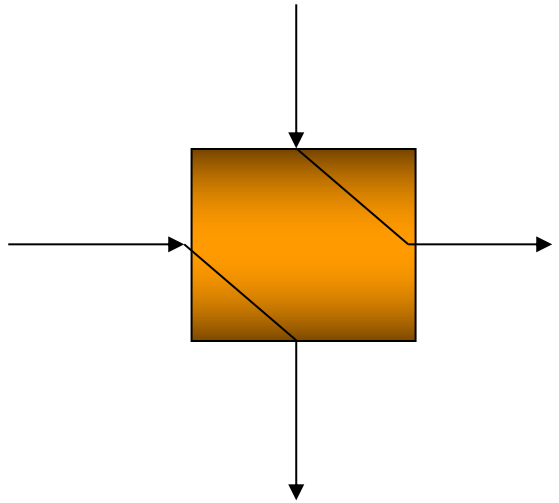
Each input port has a physical data path to every output port.
Switch at the cross-over points turns on to connect pairs of ports.

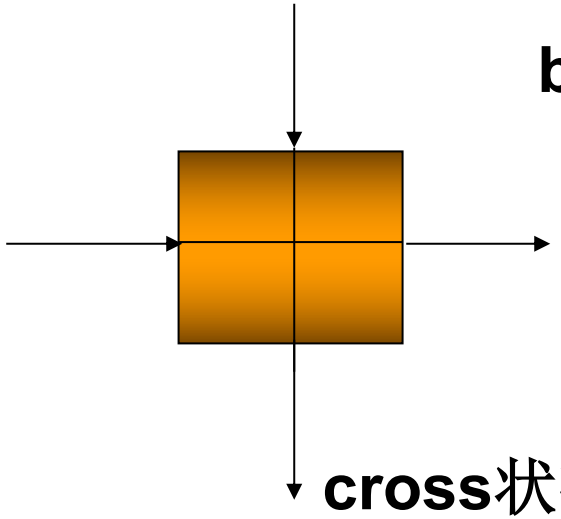
基于CROSSBAR的交换结构

- Crossbar（空分交叉开关）交换结构是一种基于矩阵交叉开关的交换结构，其中交换单元的输入和输出端口数量相等，可以通过交叉开关将输入端口和输出端口之间建立直接的连接，实现任意输入端口和输出端口之间的数据包转发。
- Crossbar是一种单级交换结构

开关阵列交叉点的实现

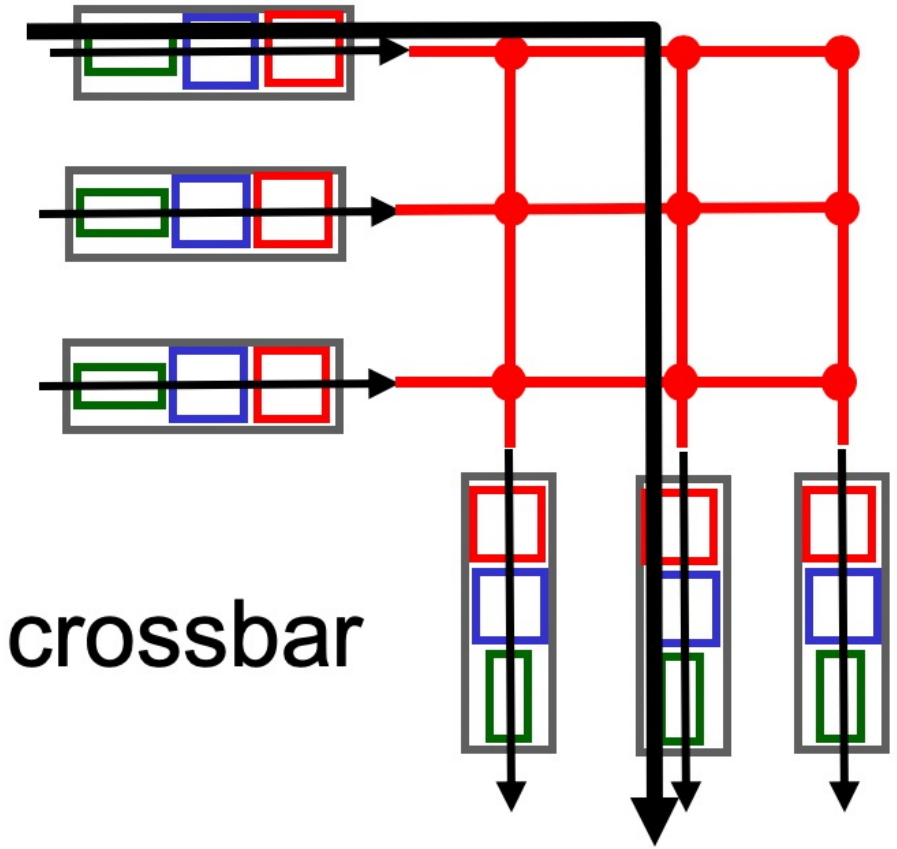






CROSSBAR特点

- 优点：
  - 具有良好的灵活性
  - 可以快速建立任意in-out直接连接
- 缺点：
  - 交叉点随着输入输出数量的增加呈指数增长（N^2 connections for N ports）



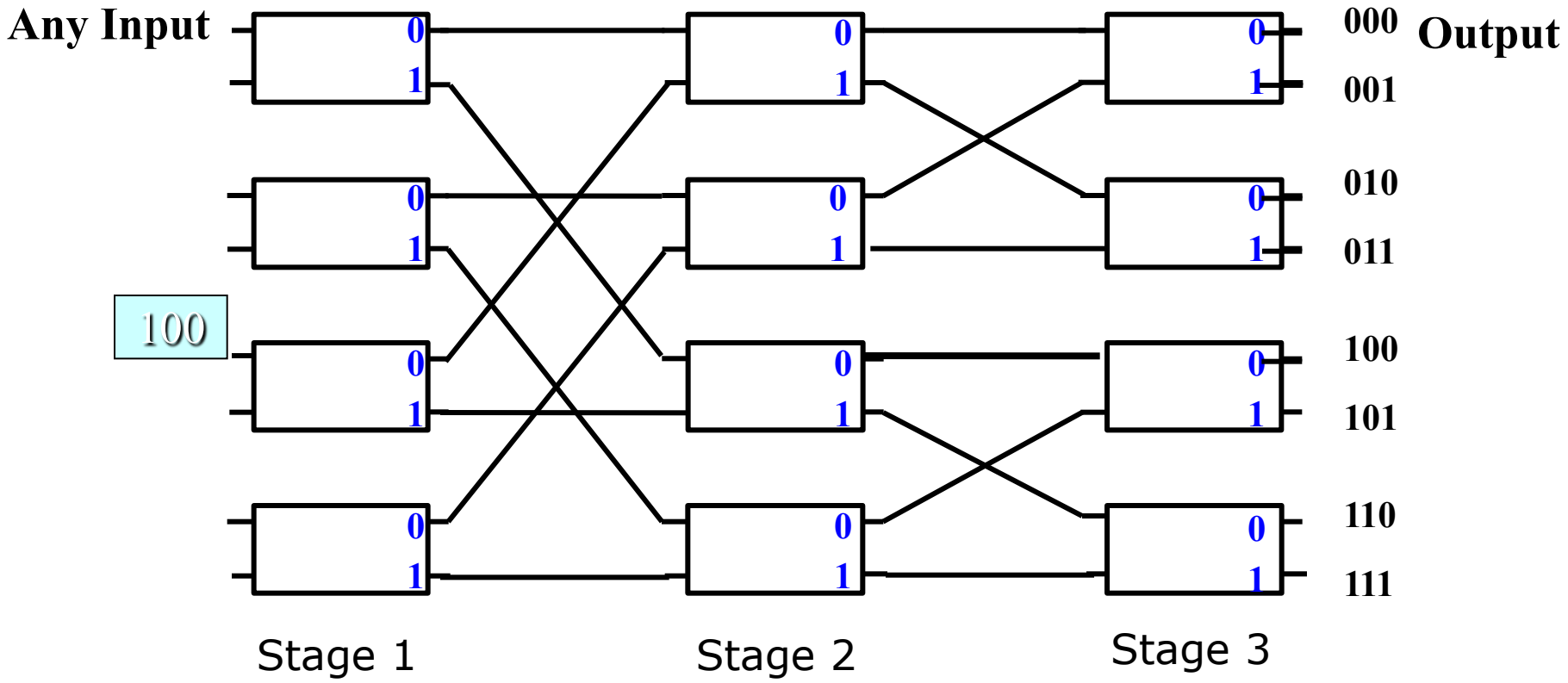
CROSSBAR可扩展性问题

- 虽然crossbar比共享内存扩展性好，但是仍属于单级交换结构，其扩展性仍不够完善，不足以支撑大容量路由器的发展。

- Solution?

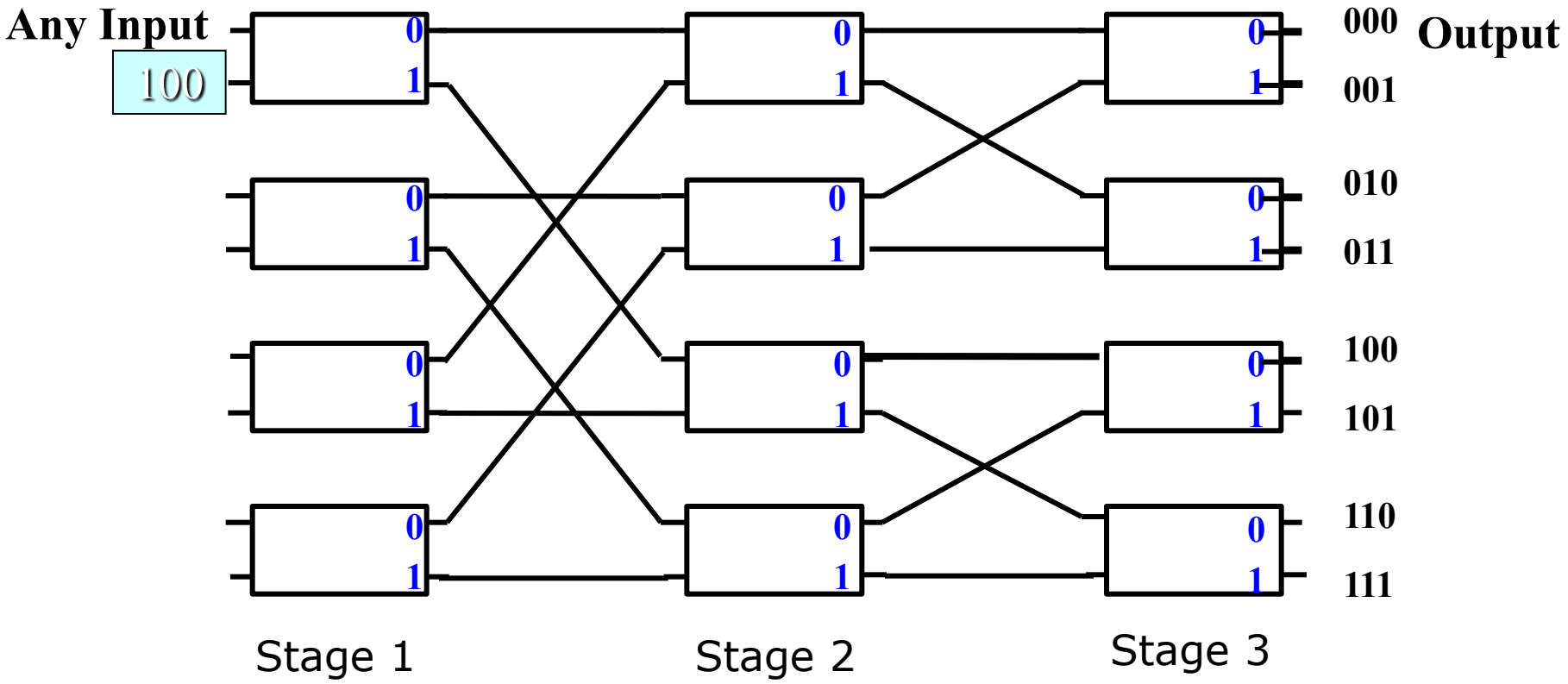
多级交换结构

BANYAN网络结构



- 2 X 2 crossbar基本单元：输入1为下输出端口，输入0为上输出端口。
- $N \times N$网络有$k=log_2N$ 级，每级$N/2$个单元
- 构成规律:蝶形连接
- 惟一路特性:任一入出线间有且只有一条路径
- 自选路由:给定出线地址，不用外加控制命令，就可选到出线。可以使用对应于出端号的二进制码的选路标签来自动选路。

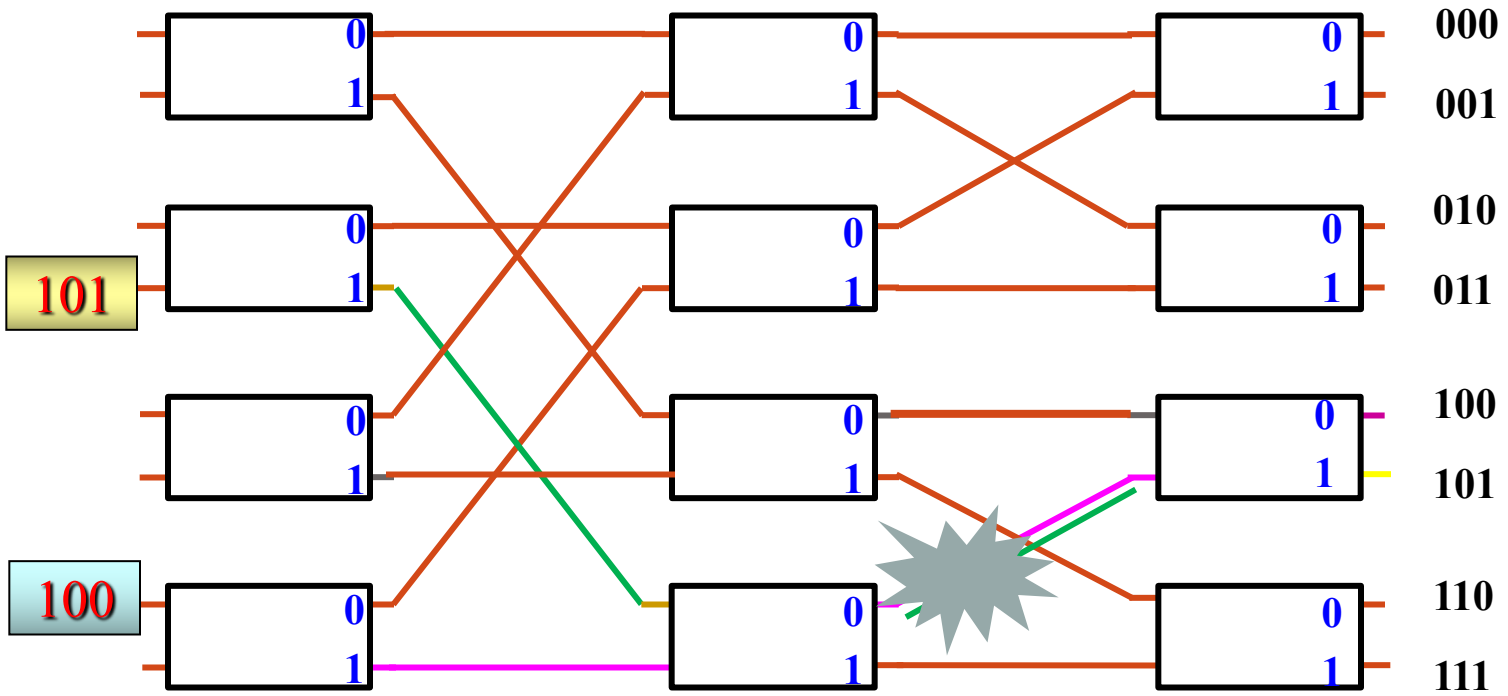
BANYAN网络结构



- 将一个路由标签任意更换位置（起始位置），该数据包仍然会从100那个端口输出

BANYAN网络结构的内部阻塞性质

- 内部阻塞：在交换网络的内部链路中出现阻塞（竞争）
  - Banyan网络不仅有内部阻塞，而且这种内部阻塞随着阵列级数的增加而增加。



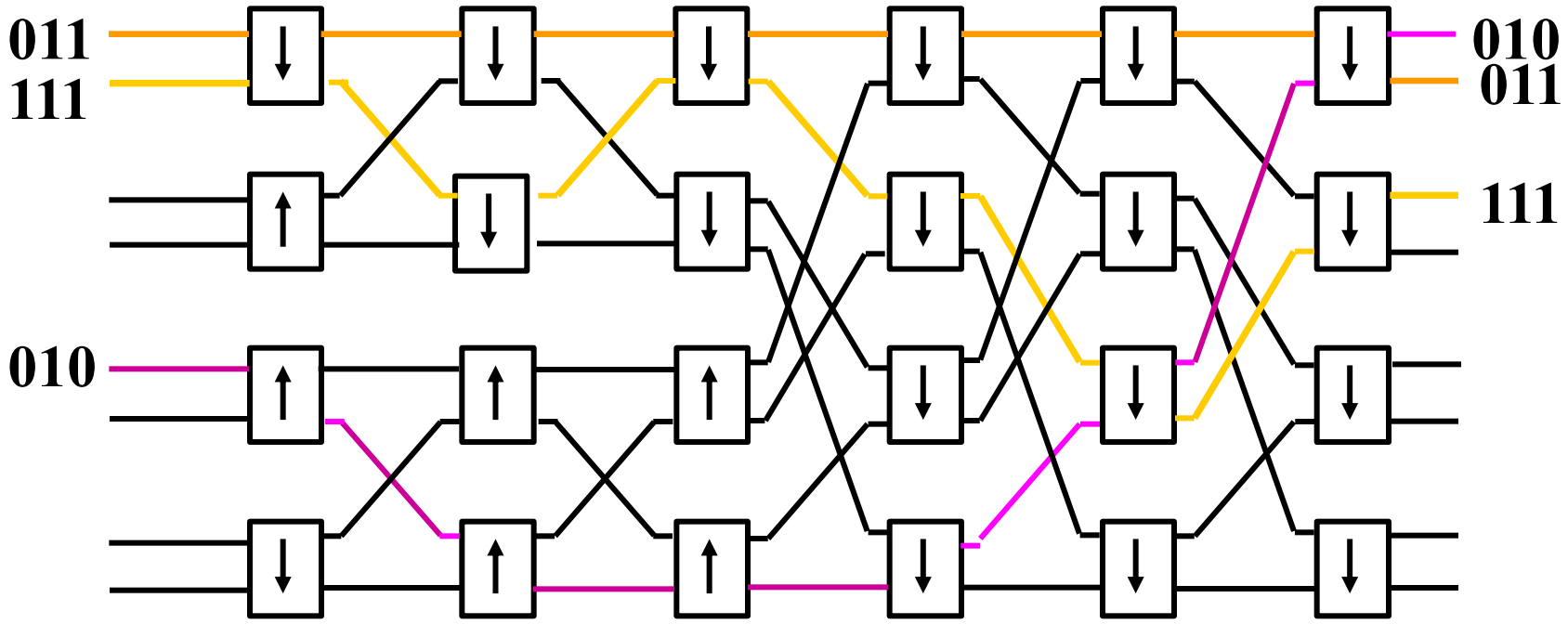
解决内部阻塞的方法

1）内部阻塞是在**2X2**交换单元的两条入线要向同一个出线上发送信元时产生的，最坏情况下概率为**50%**，若减少入线上的信息量，就可减少阻塞的概率，故可通过适当限制入线上的信息量或加大缓冲存储器来减少内部阻塞。

2）可以通过增加多级交换网络的级数来消除内部阻塞。已有证明，若要完全消除**N X N**的**banyan**网络的内部阻塞，至少需要**2**$\log_2$**N-1**级。

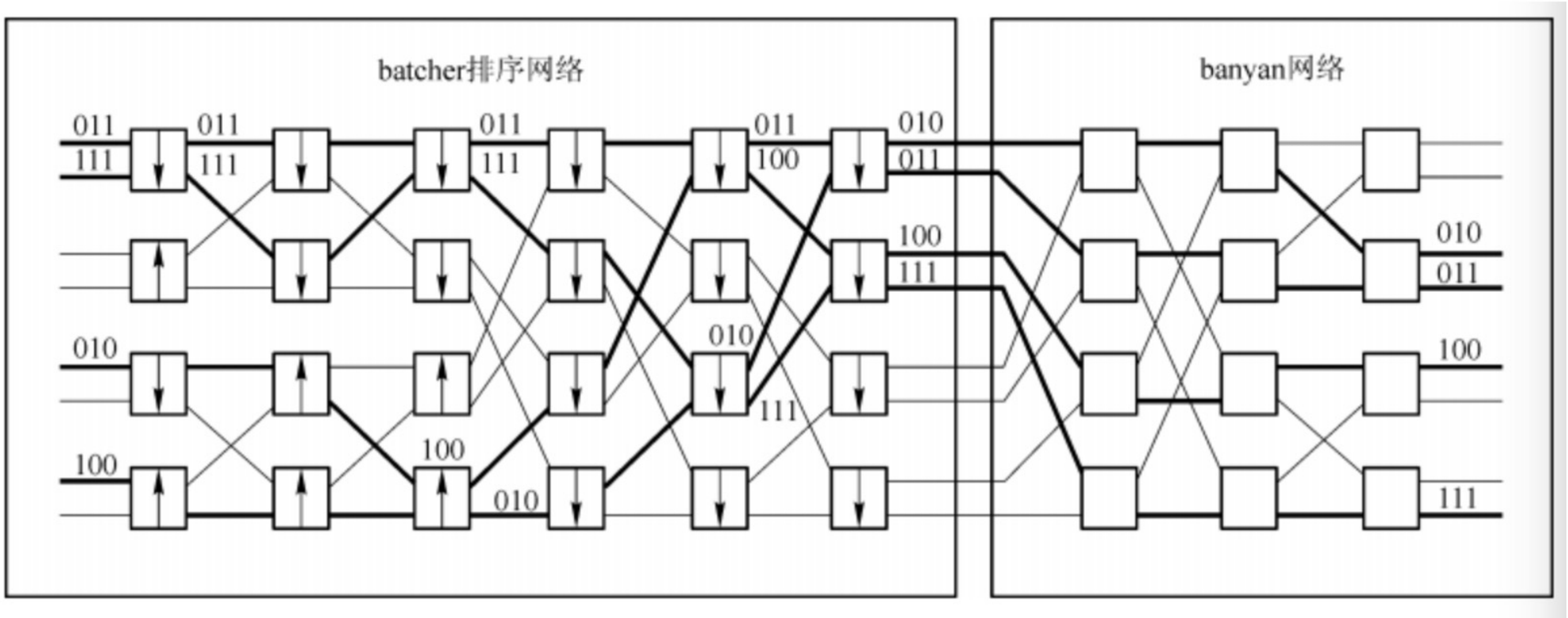
3）通过在**Banyan**网络前面添加一个排序网络使其成为一个无阻塞网络。

BATCHER 排序（SORTING）网络



- **Batcher**排序规则：在每级，如果两个标记到达同一个交换单元，则较大号的按箭头指示的输出端输出。 如果仅一个标记到达，则按箭头指示反向的输出端输出。
- 选择**3**个标记 **(001, 100** 和 **101)**。将其在任意输入端，任意次序输入；排序电路将其以升序安排在右侧的上部 **3** 个输出端。

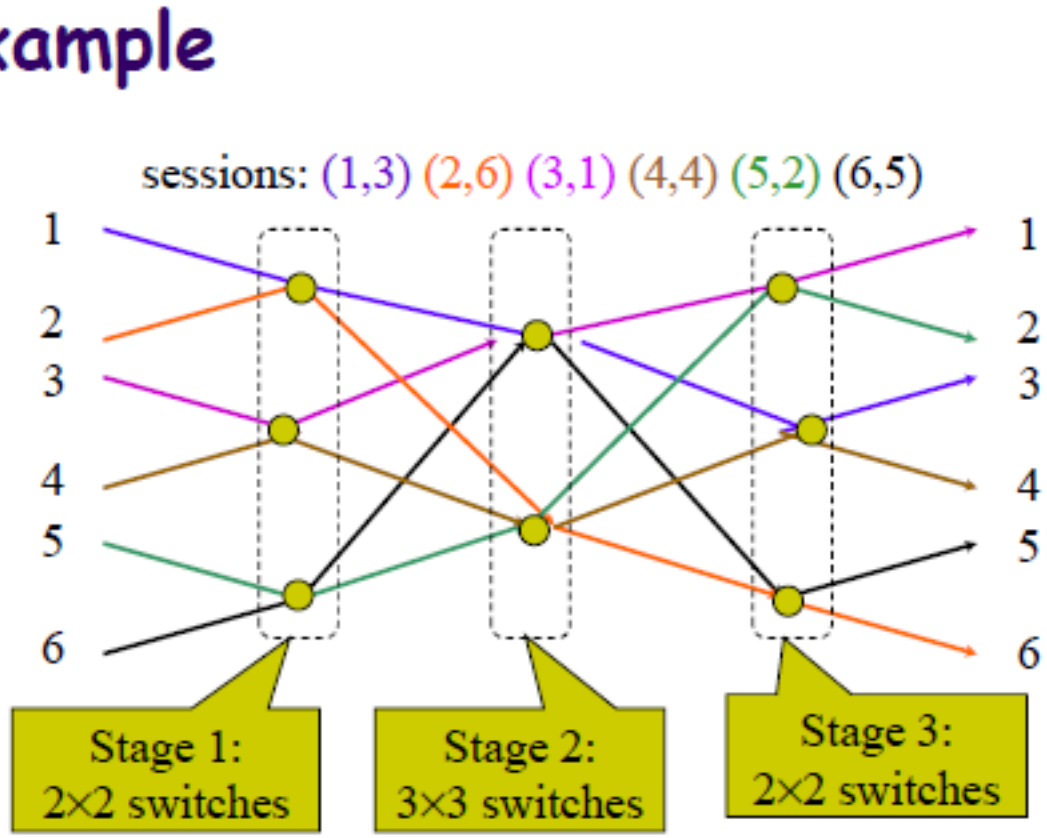
BATCHER-BANYAN网络



BATCHER-BANYAN网络

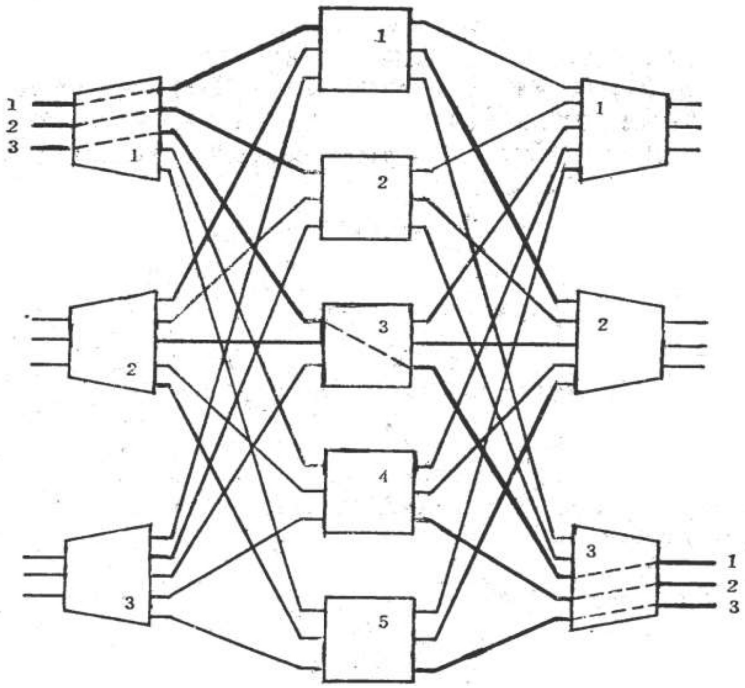
CLOS网络结构

- 多级交换，典型为三级交换架构
- 在每一级的每个单元都与下一级的设备全连接
- 到指定目的地，在第1级交换单元存在多条路由，而后续交换单元都只存在唯一的一条路由
- 严格意义上的无阻塞
- 支持递归，可无限扩展



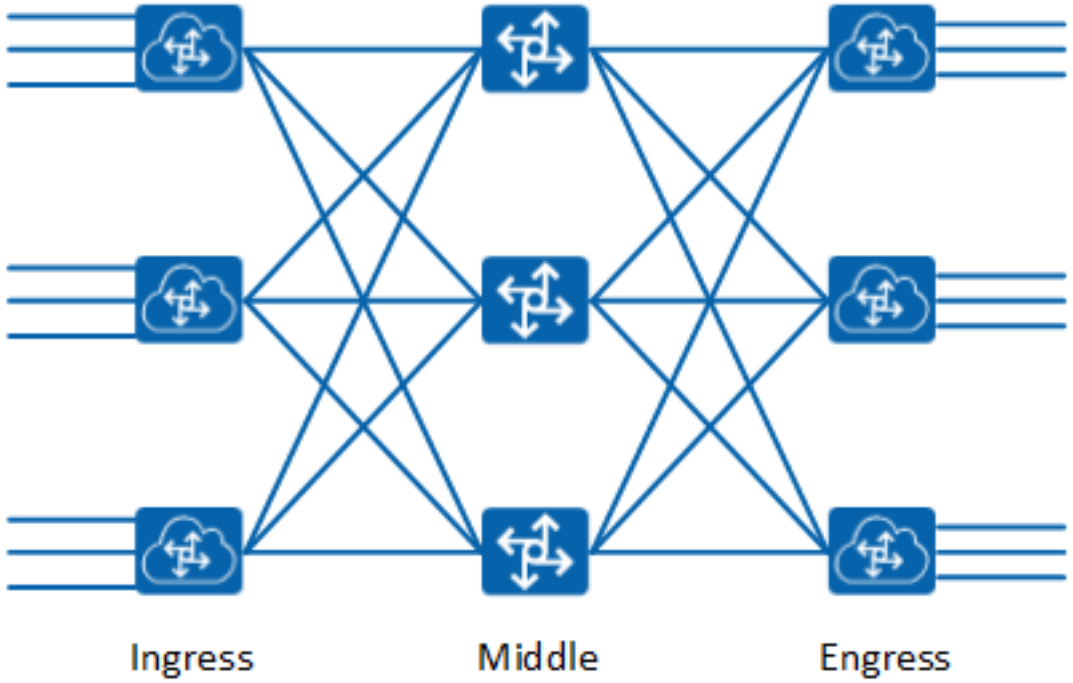
CLSO特点

- 3级对称CLOS网络严格无阻塞条件：
  - m≥2n-1;
- 可重排无阻塞条件：
  - m≥n
  - 其中m是第2级单元数，n是第1、3级单元数



CLOS进一步扩展

- 路由器之间同样可以采用clos结构进行互连



Thank You