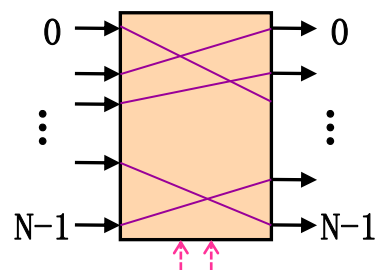
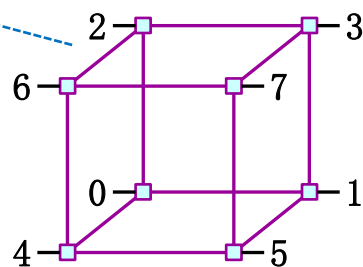
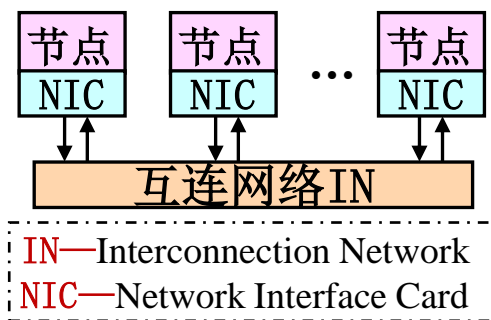


第七章 互连网络

※互连网络 (Interconnection Network)

***定义：** 由开关元件按一定拓扑结构和控制方式构成的网络，用来实现节点间的互连



***抽象：** 所有入端-出端的一组映像 (mapping)，同时只呈现一种

└─端口间连接关系

└─用控制实现

***互连特性：**

节点互连需求—可任意互连 (如 $N!$ 种映像)

←无孤立节点

需求实现策略—IN直接互连，IN多次互连+软件转发
(一次控制) (多次控制)

←效率不同

IN的互连特性—有多种映像 (可 $<N!$ 种)

←软硬取舍结果

※本章主要内容

(1) 互连函数

互连特性的表示方法，基本的互连函数，互连函数的实现

(2) 互连网络结构参数和性能指标

结构参数，性能指标

(3) 互连网络基本组成

组成要素，静态互连网络，动态互连网络

(4) 互连网络控制方式

控制方式，消息传递机制

※总体要求

理解互连网络相关概念，了解互连网络的基本组成

第1节 互连函数

※主要内容：互连特性的表示，基本互连函数，互连函数的实现

1、互连特性的表示方法

*互连函数表示法： $y=f(x)$ ， x 、 y 为入端、出端编码， f 为对 x 的操作函数
 $L \leftarrow b_{n-1} \cdots b_0, n=\log_2 N$

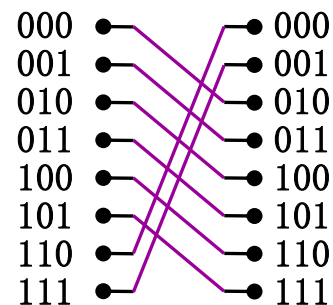
函数类型—排列、置换等

如： $f(b_{n-1}b_{n-2} \cdots b_0) = b_{n-2} \cdots b_0 b_{n-1}$ 、 $f(b_{n-1} \cdots b_0) = b_{n-1} \cdots \overline{b_0}$

*连线图表示法：所有出端-入端的连接关系(即拓扑结构)

如： $f(x) = x+2 \pmod{8}$ 的连线效果

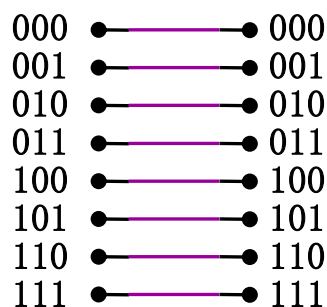
*特点：互连函数便于软件描述，
连线图便于硬件实现



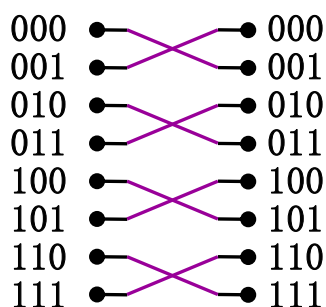
2、基本的互连函数

***恒等函数:** $f_I(b_{n-1} \cdots b_0) = b_{n-1} \cdots b_0$

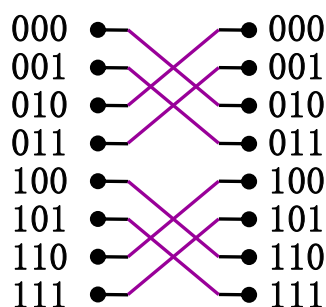
***交换函数:** $f_E(b_{n-1} \cdots b_i \cdots b_0) = b_{n-1} \cdots \bar{b}_i \cdots b_0$, 有n种(某位取反)



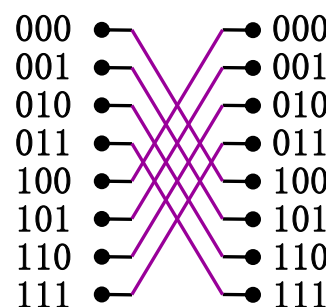
恒等



交换 (Cube₀)



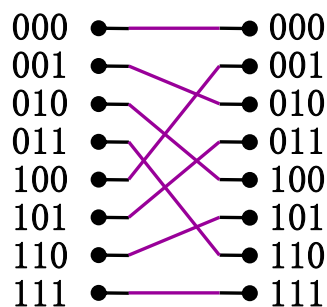
交换 (Cube₁)



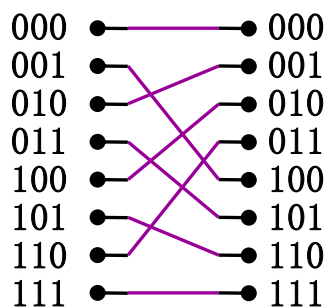
交换 (Cube₂)

***混洗函数:** $f_{Shu}(b_{n-1}b_{n-2} \cdots b_0) = b_{n-2} \cdots b_0 b_{n-1}$

变种—逆函数(= $b_0 b_{n-1} \cdots b_1$)、子函数、超函数

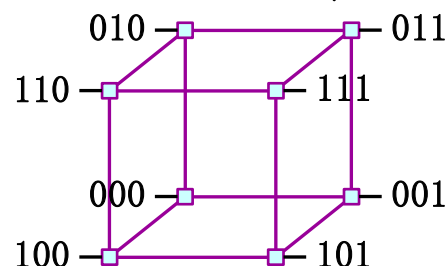


混洗



逆混洗

(b_i 最右) (b_i 最左)

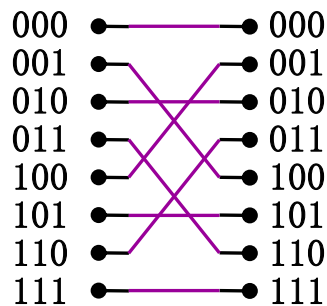


思考: 64张牌中第3张是大王, 2次洗牌后是第几张? 多次洗牌后, 可从第3张取到吗? 若能是几次洗牌?

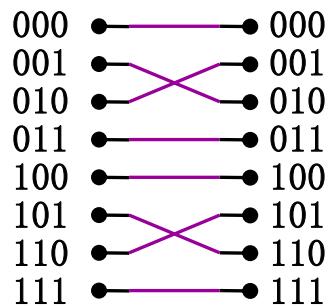
第12张, 可以, 连续6次洗牌

***蝶式函数:** $f_B(b_{n-1}b_{n-2}\cdots b_1b_0) = b_0b_{n-2}\cdots b_1b_{n-1}$

变种—子蝶式等 $f_{B(k)}(b_{n-1}\cdots b_i b_{i-1}b_{i-2}\cdots b_1b_0) = b_{n-1}\cdots b_i b_0 b_{i-2}\cdots b_1 b_{i-1}$



蝶式



子蝶式 $f_{B(2)}$

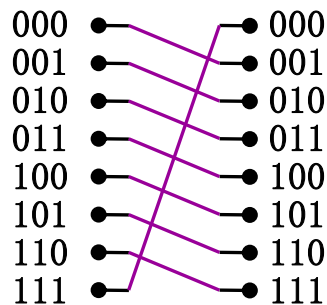
算术运算时应取模
(所有节点互连)

***移数函数:** $f_{Shi}(b_{n-1}\cdots b_0) = b_{n-1}\cdots b_0 + k \pmod{N}$

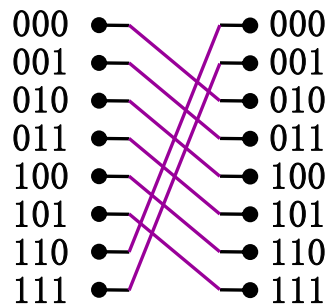
***PM2I函数:** $f_{PM2+i}(b_{n-1}\cdots b_0) = b_{n-1}\cdots b_0 + 2^i \pmod{N}$

$f_{PM2-i}(b_{n-1}\cdots b_0) = b_{n-1}\cdots b_0 - 2^i \pmod{N}$

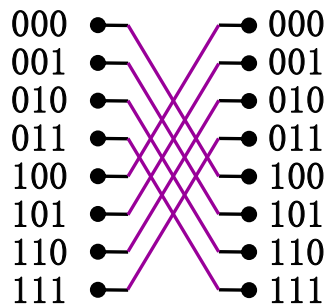
思考①: 闭合螺旋线结构有哪些互连函数?



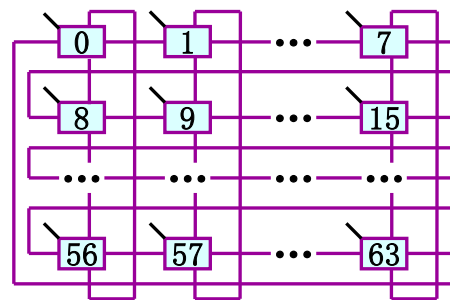
移数 ($k=1$)



$PM2I_{+1}$



$PM2I_{-2}$



思考②: 上述基本互连函数中, 哪些是互逆函数?

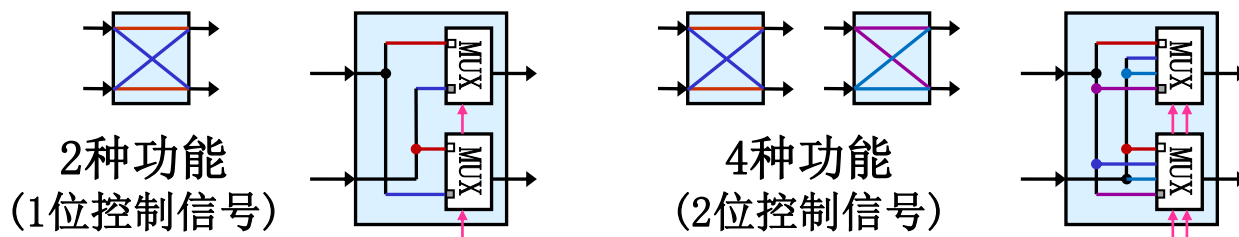
思考①: 移数 ± 1 、 ± 8 , 或 $PM2I_{\pm 0}$ 、 $PM2I_{\pm 3}$; 思考②: 恒等、交换、蝶式

3、互连特性的实现

*互连函数的实现：入端-出端直接连线(连线图表示法)

*互连特性的实现：用开关元件选择不同函数 ←有多种函数、同时仅1种

例1：2×2开关通过选择恒等/交换等函数实现

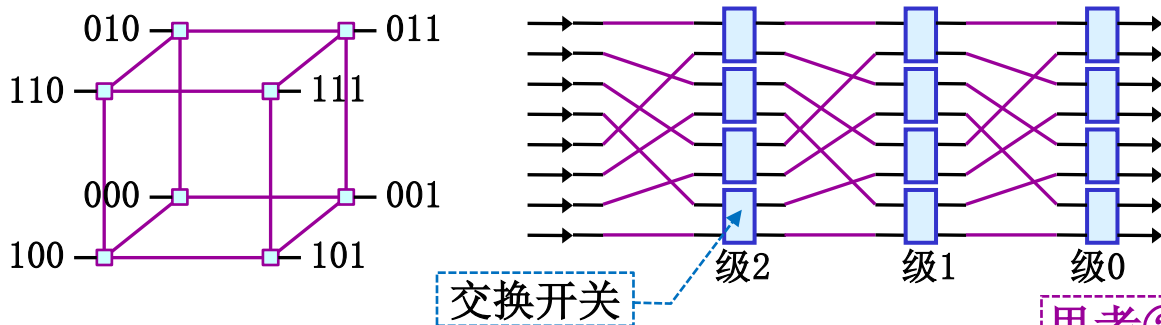


*拓扑结构：指网络内部实现互连函数时构成的几何形状

└→多个互连函数的选择或级联→┐

例2：立方体网络通过选择Cube₀、Cube₁、Cube₂函数实现

例3：多级混洗交换网络通过级联混洗函数、恒等/交换函数实现



思考①：例3可用单级网络实现吗？

思考②：IN的组成要素？

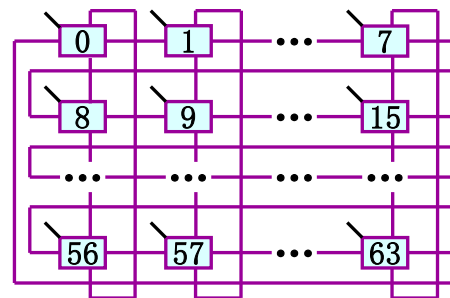
第2节 互连网络的结构参数和性能指标

※主要内容：结构参数，性能指标

1、互连网络的结构参数

网络拓扑的表示—图=节点+边(有向/无向)

└← 悬空边为I/O端口



*网络规模N：节点个数 (本例N=64)

*节点度d：节点所连接边数(节点间)的最大值 (本例d=4)

*节点距离：任意两个节点间的边数的最小值 (本例=1)

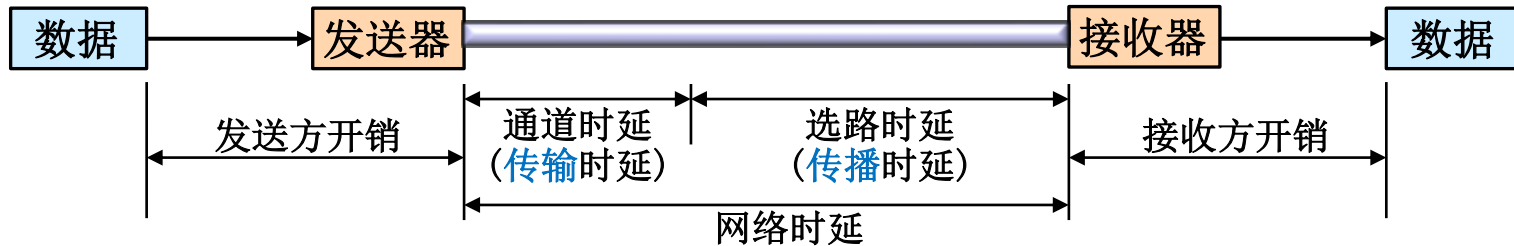
*网络直径D：任意两个节点间的边数的最大值 (本例D=8)

*等分宽度b：IN切成2个子网(N/2)的各种切法中，切口边数的最小值
(本例b=16) ←反映任意互连时的最大流量

*对称性：从任意节点看，拓扑结构是否相同
(本例=对称网络) ←对称网络易实现/编程

2、互连网络的性能指标

通信过程—数据打包+发送+网络传播+接收+数据提取



***网络时延：** = 选路时延 + 通道时延

$\downarrow \sim \text{网络直径}$
 $\downarrow = \text{帧长} / \text{通道带宽}$
即网络带宽
 $\leftarrow \text{硬件性能}$

忽略网络竞争 \rightarrow

例1：若IN网络带宽为1Gbps，信号在线路上的传播速度为20000km/s，欲传送20kb的数据帧，收/发端距离为1km、1m时的网络时延？

解: $T_{1km} = 1/20000 + 20k/1G = 50 + 20 = 70\mu s$, $T_{1m} = 20.05\mu s$

***通信时延:** = 软件开销(收/发方) + 网络时延 + 竞争时延 ← 整体性能

***端口带宽:** 任意入端-出端的带宽最小值(~路径, ~位置[非对称网络时])

***等分带宽：**IN切成两个子网时，切平面中所有边的带宽之和
(=等分宽度 b *通道带宽)

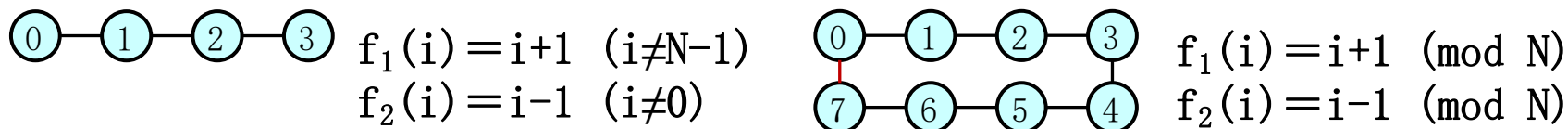
第3节 互连网络的组成

※主要内容：组成要素，静态互连网络，动态互连网络

1、互连网络的组成要素

互连的实现策略一对多种互连函数进行选择及级联

*拓扑结构：指端口间连线的几何形状，用于实现互连函数



*开关元件：指互连函数的改变部件（同时只呈现一种），
用于选择互连函数，或作为互连函数
（如上图连接左/右） （如混洗交换网络）

*控制方式：指各个开关元件的控制时间，用于增强功能
L ← 同时/分时（改变）

※IN的类型：静态互连网络、动态互连网络
（基于节点间的连接通路能否改变）

2、静态互连网络

—又称直接网络

开关元件放在节点中，节点间连接通路固定、运行中不能改变

***拓扑结构：**

网络类型	节点度 d	网络直径 D	链路数 l	等分宽度 b	对称性	网络规模
线性阵列	2	$N-1$	$N-1$	1	非	N
环形	2	$\lceil N/2 \rceil$	N	2	是	N
二叉树	3	$2(h-1)$	$N-1$	1	非	$N, h=\log_2 N$
星型	$N-1$	2	$N-1$	$\lceil N/2 \rceil$	非	N
2D网格	4	$2(r-1)$	$2N-2r$	r	非	$N=r \times r$
2D环网	4	$2\lceil r/2 \rceil$	$2N$	$2r$	是	$N=r \times r$
超立方体	n	n	$nN/2$	$N/2$	是	$N, n=\log_2 N$
全连接	$N-1$	1	$N(N-1)/2$	$(N/2)^2$	是	N

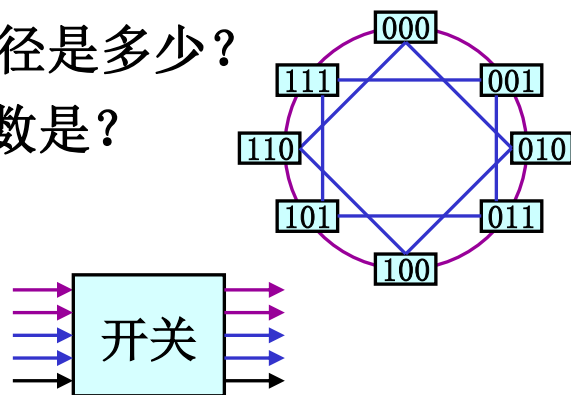
例1：右图带弦环网的互连函数有哪些？网络直径是多少？

离3#节点最远的节点是？开关元件的端口数是？

解： $f(x) = x \pm 2^i, i=0, 1; D = \lceil (8/2)/2 \rceil = 2;$

离3#节点最远($D=2$ 时)的节点=0#、6#、7#;

开关元件的端口数是5(外部-节点内部需1个)。



例2: 右图混洗交换网的互连函数有哪些？开关元件的端口数是？网络直径是多少？离3#节点最远的节点是？

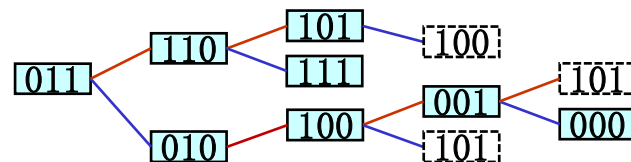
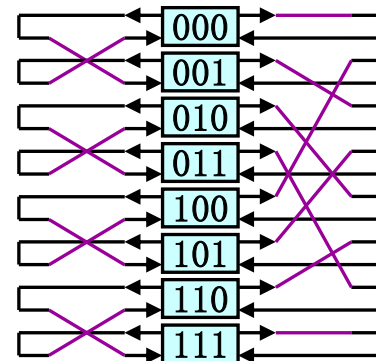
解: $f_{\text{Shu}}(b_2b_1b_0) = b_1b_0b_2$, $f_{\text{Cube0}}(b_2b_1b_0) = b_2b_1\bar{b}_0$;

开关元件的端口数是 $2+1=3$;

0#→7#最远, $D = \text{交换3次} + \text{混洗2次} = 5$;

(000 $\xrightarrow{\text{混洗}}$ 001 $\xrightarrow{\text{交换}}$ 010 $\xrightarrow{\text{混洗}}$ 011 $\xrightarrow{\text{交换}}$ 110 $\xrightarrow{\text{混洗}}$ 111)

离3#节点最远(枚举达到所有节点)的节点是0#。



拓扑结构的特征—维数 \uparrow 导致节点度 \uparrow 、网络直径 \downarrow 、成本 \uparrow

***开关元件:** 端口数 $=d+1$, 复杂度 $=d^2$ \leftarrow 外部-节点内部需1个端口

***互连特性:** 函数个数 $=d$, 函数功能 \sim 通路特性

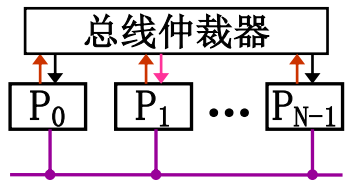
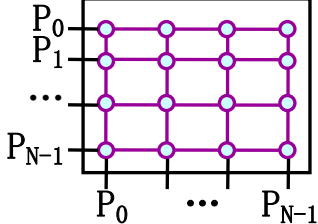
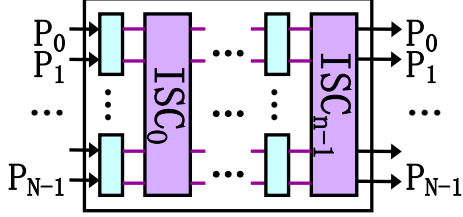
***特点:** 寻径效率 $\sim D$, 网络流量 $\sim b$, 价格 $\sim l$

\leftarrow 可优化(如虫孔寻径)

\leftarrow 网络控制中解释

3、动态互连网络

开关元件放在网络中，节点间连接通路可动态地改变 ←节点中无开关

种类	总线网络	交叉开关网络	多级互连网络
组成示例			
拓扑结构	总线	全连接	多个级间拓扑的级联
开关元件	无(节点识别地址)	交叉开关, $O(N^2)$	多个 $k \times k$ 开关, $O(k^2)$
互连特性	0个函数(N 次=1个)	常为 $N!$ 个(可为 N^N)	$\leq N^{N/2}$ 个
特点	带宽窄、成本低	无阻塞、成本高	可扩展性好、成本较低

***多级网络的互连函数：** (3种因素叠加，级数常= $\log_k N$)

级间拓扑结构—可不同，混洗、蝶式、立方体等 ←**决定**互连函数

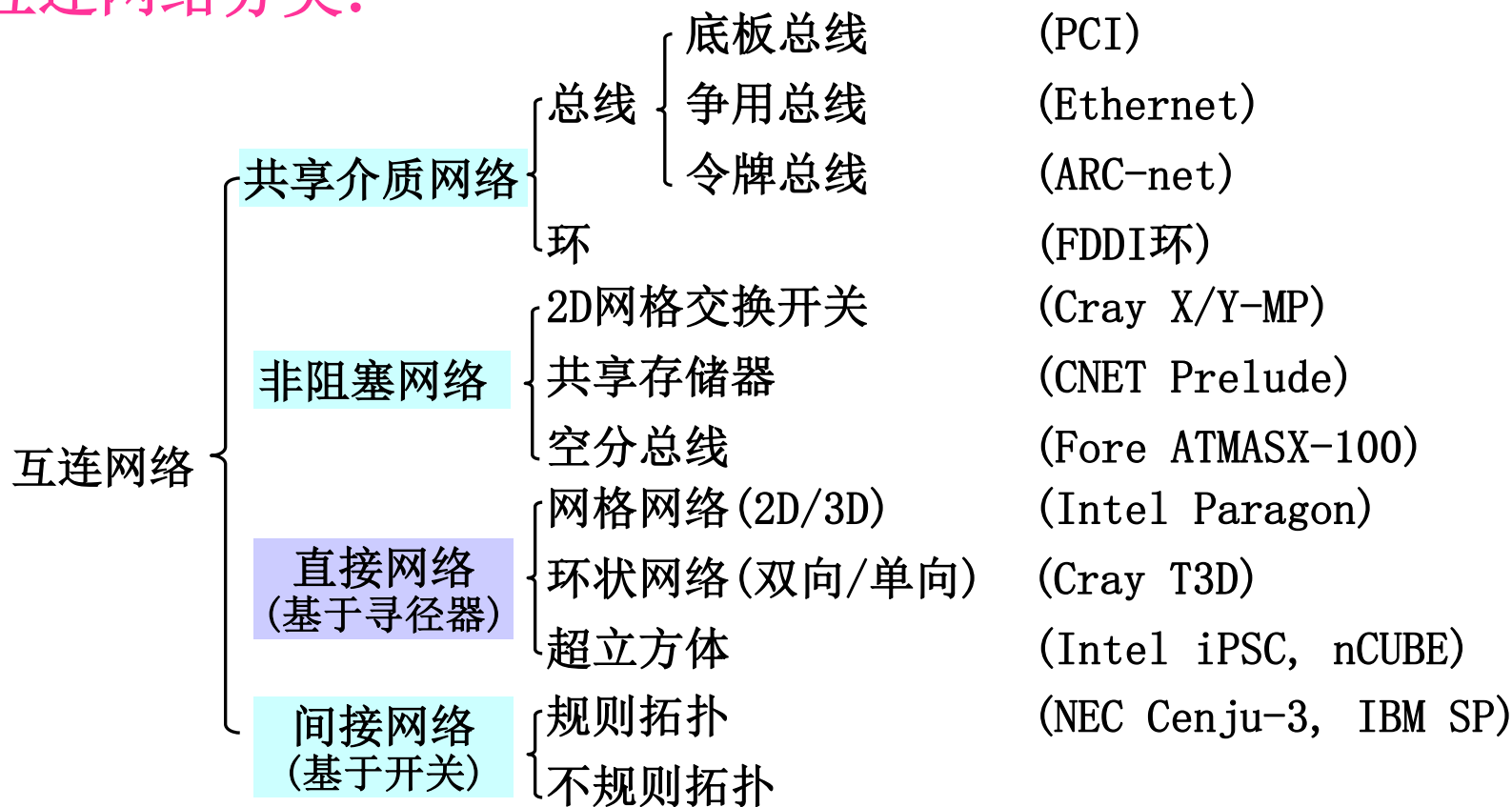
开关类型— $k \times k$ ，互连函数常为置换 ←**增加**互连功能

开关控制方式—级控制、单元控制、部分级控制 ←**决定**所增加函数

(指取值类型) ($k!^{\log_k N}$ 种) ($N^{N/2}$ 种) ($k!^{\log_k N} \sim N^{N/2}$ 种)

非取值时间(网络控制方式)

※互连网络分类：



※互连网络应用：主干网为静态网络，子网为动态网络

直径小 → 性能好

← 灵活性好

思考：设计阵列机IN时，宜采用静态/动态网络？影响网络功能的因素？

思考：静态网络（直径小），常用并行算法类型

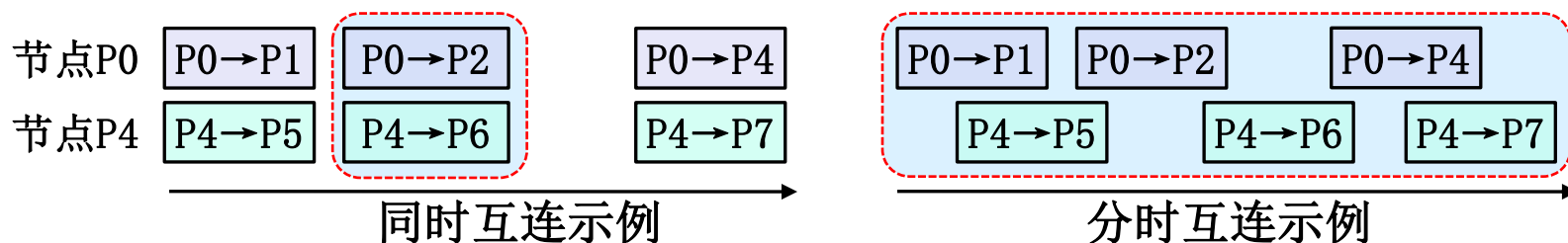
第4节 互连网络的控制

※主要内容：控制方式，消息传递机制

1、互连网络的控制方式

指各开关元件控制时间的类型

互连时间需求— SIMD的互连函数需同时实现， ←指令内部互连
MIMD的互连函数可分时实现 ←线程之间交互



互连功能的表示—

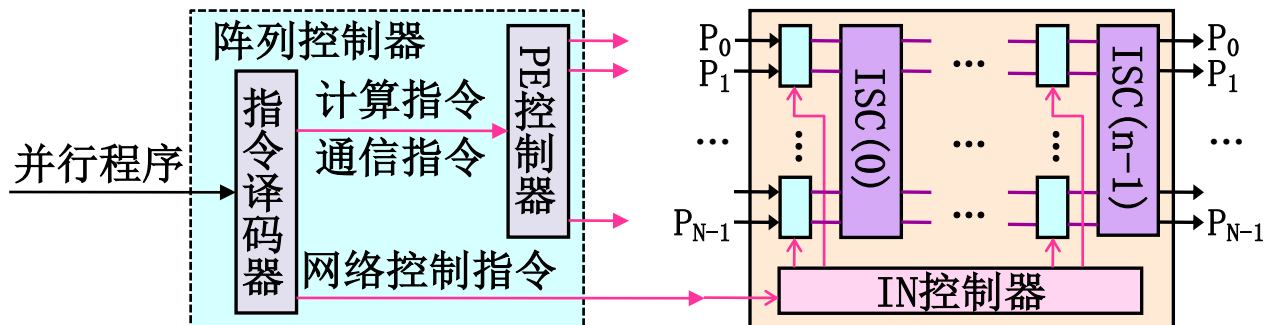
思考：MIMD的拓扑结构最低要求？

同时互连时：IN外部用控制信号表示 ←涉及IN中所有开关

分时互连时：数据包中用源-目地址表示 ←涉及IN中部分开关
└→ IN内部形成路由、控制相应开关

思考：节点间可任意互连即可，不要求同时互连；
对拓扑结构无要求，如总线。

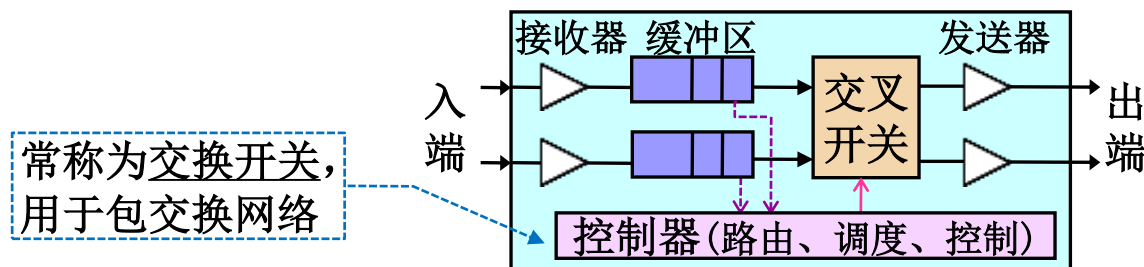
***集中式控制：**同时控制所有开关的状态，直到再次控制为止
 组织—IN接收控制信号，控制器产生所有开关的控制信号



应用—外部部件产生IN控制信号，然后可进行通信

***分布式控制：**各端口独立控制数据包经过路径上开关的状态(分时)，直到数据包通过为止

组织—开关设置缓冲区，控制器只产生所需出端的控制信号



思考：控制器为何要有路由功能？调度可解决什么问题？

***应用：**集中式控制适于SIMD，分布式控制适于MIMD

思考：源-目有多种路径，路由功能选择最佳路径；调度处理多个数据包要求同一出端的冲突

集中式控制的开关元件没有缓冲器，分布式必须有。

2、消息传递机制 —— 分布式控制

***通信过程：** 寻径+消息传递

← 确定经过开关+传递数据包

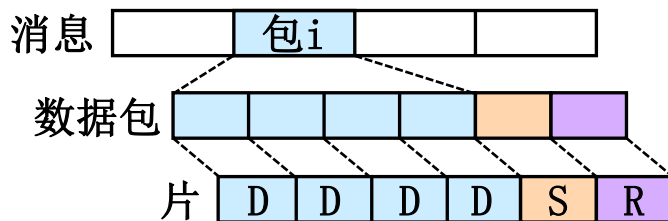
***传递方式：** 线路交换、包交换

← 面向连接、面向无连接的传递

思想——先寻径、再传输，边寻径、边传输

← 如电话网、铁路网

传输粒度——消息 (大小可变)，数据包 (大小固定 [如1500B])



数据包——大小固定、独立传送

R—导径信息 (目的地址)

S—数据包序号 (消息中)

D—数据片 (大小固定)

***寻径方式：** (线路交换仅1种，包交换有3种)

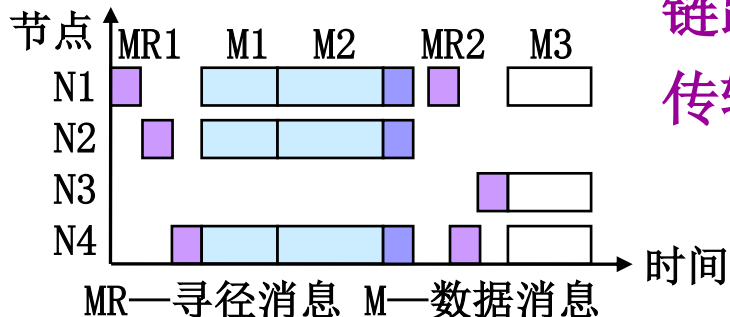
线路交换——先建立物理链路、再传送信息

链路建立&释放： 均用消息实现

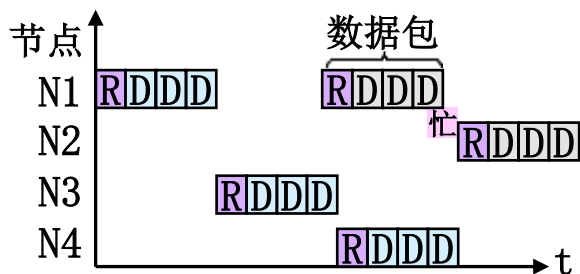
传输时延： $T_{\text{寻径}} = L_{\text{寻径}} * (D+1) / B$,

$T_{\text{传输}} = L_{\text{消息}} / B$

其中，D—中间节点数，B—通路带宽



存储转发—数据**包**到达时，先**存储**(整个包)，再**寻径**和**转发**(**包**)

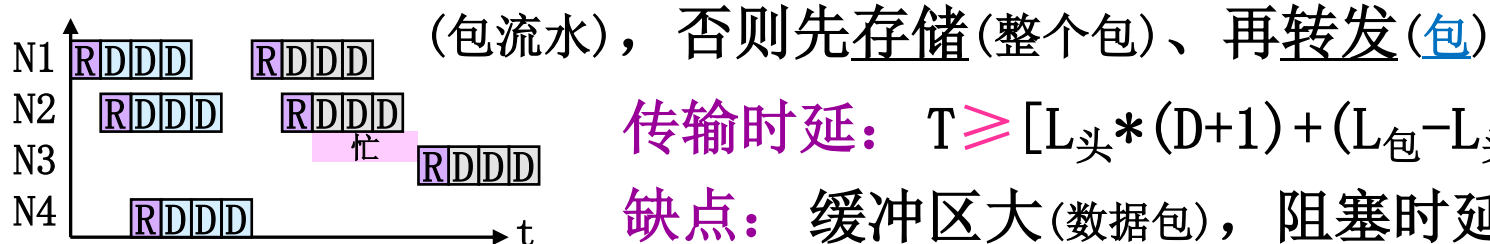


传输时延: $T = [L_{\text{包}} * (D+1)] / B$

链路释放: 包**通过**时自动(**包交换方式通用**)

缺点: 缓冲区大(数据包)，传输时延大(\sim 距离)

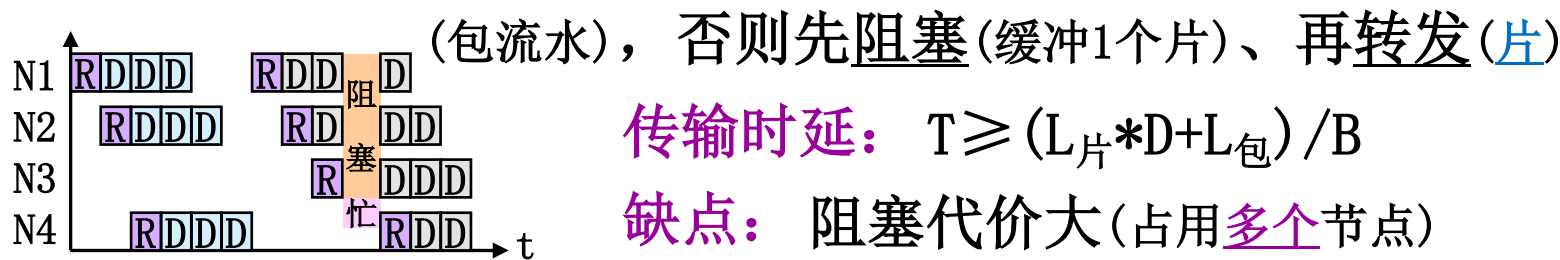
虚拟直通—数据**包头**(片)到达时，立即**寻径**；链路闲时**片**立即**传送**



传输时延: $T \geq [L_{\text{头}} * (D+1) + (L_{\text{包}} - L_{\text{头}})] / B$

缺点: 缓冲区大(数据包)，阻塞时延大(=存储转发)

虫蚀寻径—数据**包头**(片)到达时，立即**寻径**；链路闲时**片**立即**传送**



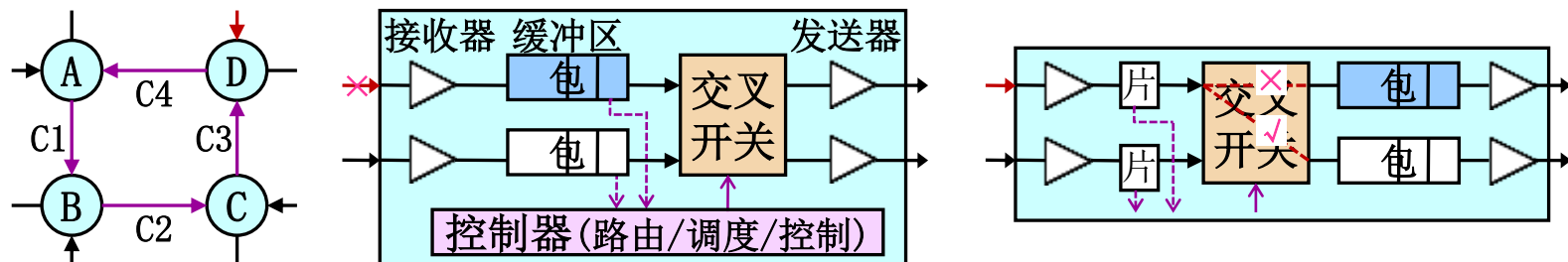
传输时延: $T \geq (L_{\text{片}} * D + L_{\text{包}}) / B$

缺点: 阻塞代价大(占用**多个**节点)

思考: 如何选择? 基于性能、性/价

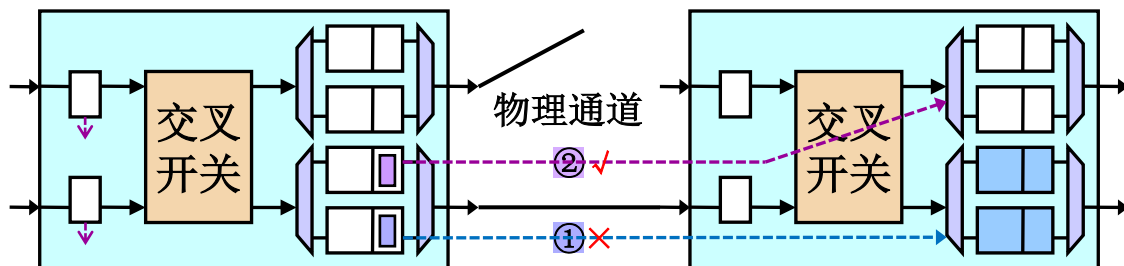
*寻径的死锁避免:

死锁的产生—节点缓冲区满、选路需求构成闭环时(并发传递所致)



输出缓冲—将输入缓冲区改为输出缓冲区, 避免排头阻塞现象

虚拟通道—设置多个缓冲区, 多个逻辑链接共享物理通道

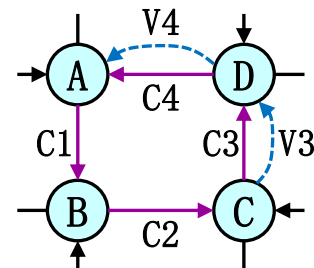


端口的缓冲区个数: \leq 输出端口个数

←性能-成本的权衡

虚拟通道的个数: 部分路径上增设

└←打破闭环即可、降低成本



思考①: 可避免排头阻塞现象(同一入端第一个包被阻, 第二个包[使用不同出端]也被阻)。

思考②: \leq 输出端口个数。

*寻径中的包冲突处理:

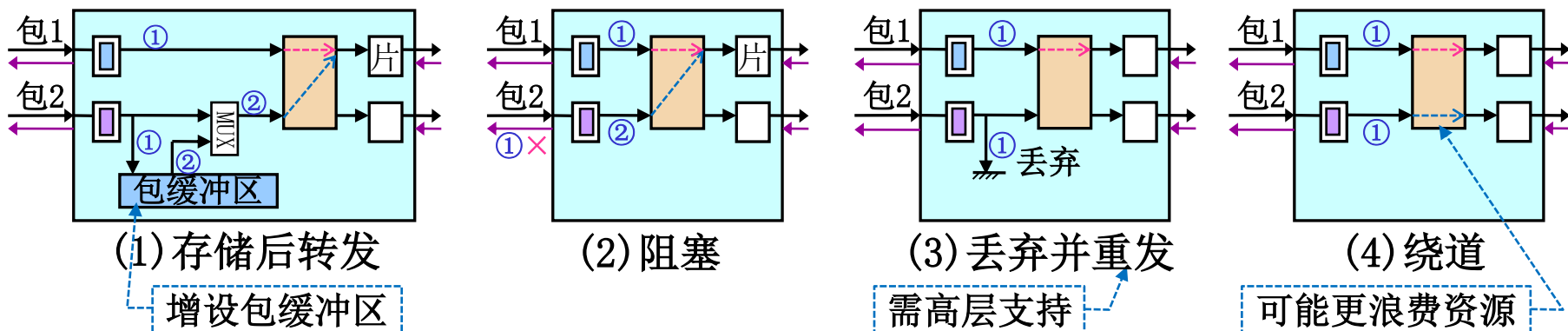
可传送条件—①源缓冲区已存储 ②通道已分配 ③目节点可接收

冲突的产生—2个包同时请求同一缓冲区/输出通道

└←解包时条件①已满足

处理的任务—①缓冲区/通道的分配 ②被拒绝包的处理

处理方案—有4种 (假设通道分配给包1)



应用选择—短时冲突时采用前2种, 网络拥塞时采用(4)

└←如存储转发、虚拟直通采用(1)

└← 虫蚀寻径采用(2)

第七章小结&思考

- (1) IN中，节点互连的需求、实现策略？ IN的功能？
 - (2) 一个互连函数如何实现？ 不同互连函数的选择如何实现？ 拓扑结构与互连函数的关系？
 - (3) IN的组成要素？ 静态IN、动态IN中开关元件的位置、作用？ 不同控制方式适应的体系结构？
 - (4) IN的集中式控制、分布式控制如何实现、开关控制何时失效？
 - (5) 分布式控制IN中，消息的传递方式类型、传输粒度？
-
- (1) 可任意互连($\geq N!$ 种映像)； IN直接连接(1次控制)或IN连接+软件转发(多次控制)；
有多种映像(可 $< N!$ 种)
 - (2) 所有入端-出端的连线； 使用开关元件； 所有互连函数实现时的几何形状
 - (3) 拓扑结构+开关元件+控制方式；
静态IN：节点中、选择互连函数，动态IN：网络中、用作互连函数；
集中式控制：适于SIMD，分布式控制：适于MIMD
 - (4) 集中式：IN接收外部控制信号，所有开关同时控制，下次控制时；
分布式：各端口独立解析包中源-目的地址、选择出端、控制开关状态，包通过时
 - (5) 线路交换、包交换，消息、数据包