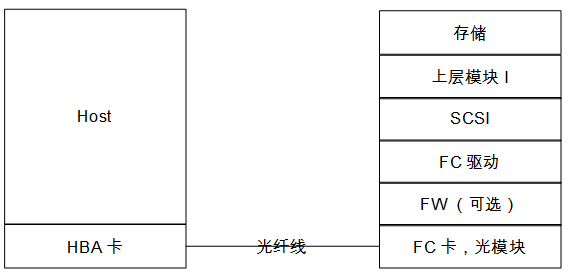
## 1.FC驱动的位置和作用

FC即Fibre Channel，是存储系统中用于前端连接主机的主流协议。FC的常见应用场景如下所示：



## 2.应用场景的软硬件结构示意图如下所示：

存储系统中FC模块（包含软硬件）的主要功能是提供主机和存储之间数据交互的通道：从光纤链路上接收远端设备发来的FC Frame，将Frame的payload中包含的有效数据剥离出来交给SCSI层处理；或者接收SCSI层发来的数据，将其封装到FC Frame中发送给远端设备。



## 3.FC协议的分层

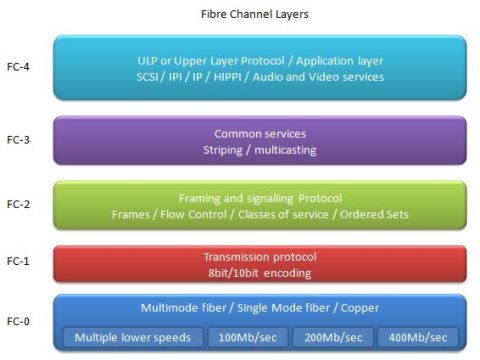
FC-0：连接物理介质的界面、电缆等；定义编码和解码的标准。

FC-1：传输协议层或数据链接层，编码或解码信号。

FC-2：网络层，光纤通道的核心, 定义了帧、流控制、和服务质量等。

FC-3：定义了常用服务，如数据加密和压缩。

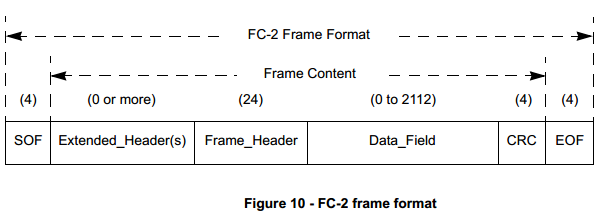
FC-4：协议映射层，定义了光纤通道和上层应用之间的接口，上层应用比如：串行SCSI协 议，HBA卡的驱动提供了FC-4 的接口函数。FC-4 支持多协议，如：FCP-，FC-IP，FC-VI。



## 4.FC Frame和Frame Header结构:

FC frame遵循固定格式，如下图所示。SOF和EOF是Frame的定界符，分别表示Start Of Frame和End Of Frame；光纤中传输的是表示0/1信号的光脉冲，为了确定一个帧的边界就需要定义帧的定界符，假设0101010作为SOF, 00110011作为EOF，则在接收端就可以就可以根据SOF和EOF将连续的光脉冲切分成一个个FC帧进行解析。

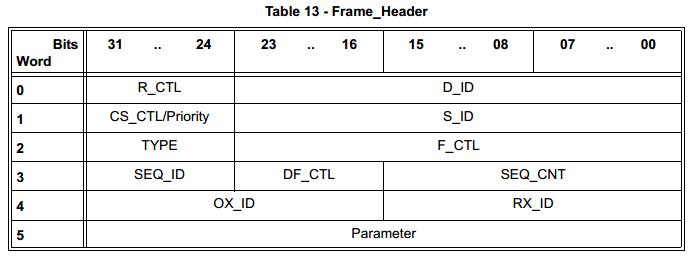
Extended\_Header可选，Frame\_Header详细格式见后一张图；Data\_field是有效载荷数据，也叫payload; CRC是帧校验码，用于检测在链路传输过程中是否发生错误。



Frame Header中的字段含义：

D\_ID: Destination\_ID，24bit，表明一个帧发往哪个端口，FC帧通过交换机时交换机根据D\_ID转发该FC帧。

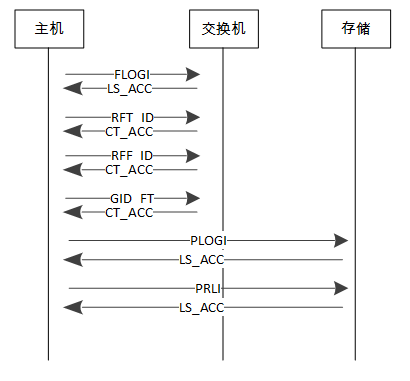
S\_ID:Source\_ID，24bit，表明产生该FC帧的端口的PORT\_ID；如果接收端口需要对该FC帧做出响应，则分别以该帧的D\_ID、S\_ID作为响应帧的S\_ID、D\_ID。



## 5.与生俱来的身份证号：WWPN,WWNN

与以太网卡的MAC地址类似，每个FC设备都有自己的全球唯一的编号WWPN和WWNN，WWPN表示World Wide Port Name, 每个FC端口有自己的一个WWPN；WWNN表示World Wide Node Name，每个节点有自己的一个WWNN，Node的范围由各个厂商自己界定。例如对于FC HBA卡厂商，通常一张HBA卡作为一个node，HBA卡上的多个端口分别有自己的WWPN，但是共享一个WWNN。而对于存储厂商而言，通常一个存储控制器作为一个node，也有厂商将多个控制器，包括通过PCIE扩展的含有多个控制器的一整套存储设备作为一个Node。

WWPN和WWNN的格式相同，都是64bit的二机制数据，通常表示为10:00:00:90:fa:7b:e2:d8或者0x10000090fa7be2d8的形式。WWPN和WWNN通常写死在硬件当中或者根据硬件信息计算得出，一般不会改变。一些针对存储的常见的维护操作，如更换光模块等不会影响端口的WWPN/WWNN。



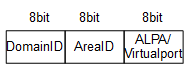
|  |
| --- |
| Wwpn规则： |

## 6.两台设备通过交换机的交互过程：

### 1）第一张学籍卡：FLOGI（Fabric Login）

以太网设备在网络中进行通信，除自身的MAC地址外，还需要一个IP地址；与之类似，FC设备进行通信时，也需要一个用于通信的地址:NportID；NportID由交换机进行分配或者直连的两台设备进行协商，当设备通过光纤线接入交换机时，会向交换机发送Flogi请求，Flogi的Frame Header中S\_ID为0x000000，D\_ID为0xFFFFFE；交换机收到FLOGI请求后，给设备返回LS\_ACC，LS\_ACC的Frame Header中D\_ID就是交换机分配给设备端口的NportID。

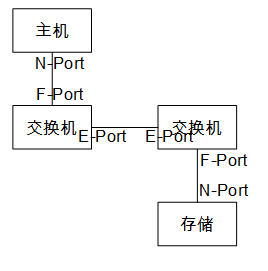
交换机分配NportID时，通常遵循以下格式：



其中最高8bit DomainID表示交换机的编号；中间8bit AreaID表示交换机端口号；最低8bit在设备以loop拓扑模式连接时为环路地址，端口有虚拟端口时为虚拟端口编号，其他情况下为0。如果一台设备接入到domain=3的交换机的4号端口，则通常交换机为设备端口分配的NportID为0x030400；按照此规则分配的NportID可用于链路故障诊断。

设备端口通过FLOGI登录到交换机并获取到NportID以后，还会通过RXX\_ID向交换机注册一些XX信息，告知交换机该ID的端口携带的XX属性是什么；还能通过GID\_XX向交换机发送特定的属性，请求交换机返回连接在交换机上的带有这些属性的端口的NportID。

备注：FC网络中，连接到交换机的普通设备端口称为N-port，连接普通设备的交换机端口称为F-Port；交换机级联端口称为E-Port。



### 2）初次相识：PLOGI（Port Login）

在上文部分中，我们看到主机完成了向交换机的注册，并且通过GID\_FT知道了交换机上连接有哪些端口，获取到了这些端口的NportID。但此时并不知道这些端口的WWPN/WWNN是什么；类似于两位同学只知道彼此的学号，但彼此不认识。

在上图中主机以存储端口的NportID为D\_ID，以自身的NportID为S\_ID发送PLOGI,PLOGI中携带有主机的WWPN/WWNN信息；存储端收到Plogi后，以主机端口的NportID为D\_ID，以自身的NportID为S\_ID发送LS\_ACC, LS\_ACC中携带有存储的WWPN/WWNN信息。

通过PLOGI/LS\_ACC交互后，主机和存储端口就知道了对方的WWPN/WWNN信息，确定了NportID和WWPN的关联关系，后续的交互只需要利用NportID交互即可，如同网络中两台设备的交互可以通过IP完成，而不需要MAC。

### 3）深入了解达成共识：PRLI（Process Login）

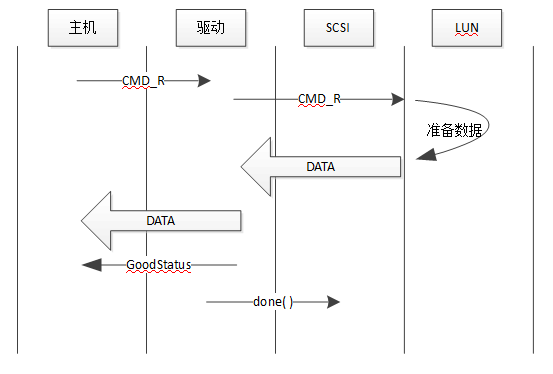
通过上文的PLOGI两个端口获取了对方的WWPN/WWNN信息，但是FC协议作为传输类协议，既可以承载SCSI数据，也可以承载其他协议的数据。如果主机FC模块上层承载了IP协议，而存储FC模块上层承载了SCSI协议，则两台设备无法进行通信。类似于两个不同专业的同学无法就专业问题展开探讨。

PRLI用于交互FC协议承载的协议等信息，类似交换所学的专业信息，如果两个设备该信息不一致，则回复LS\_RJT。

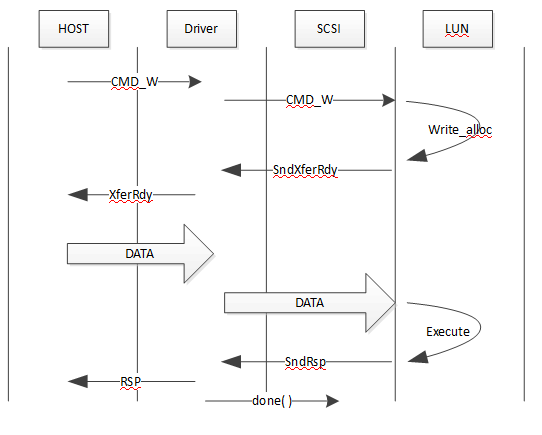
### 4）例行业务往来：IO交互

上文中的连接建立过程完成后，就可以进行IO交互了；类似于两个同学在学校注册了、相互认识了、而且所学专业相同；就可以关于专业问题进行深入交流了。

主机与存储之间，以及存储内部模块间IO的交互流程如下图所示：



读IO流程

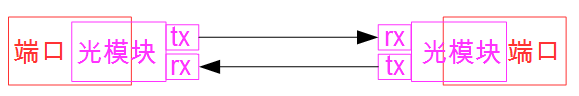


写IO流程

## 7.硬件形态：

发送端口将电信号发送给光模块，光模块将电信号转换为光信号发出；接收端口的光模块接收链路上的光信号，将其转换为电信号传递给接收端口。光模块的接收发送单元相互独立，光纤线的两根光通路也相互独立。因为只有接收端能够统计到误码，当链路质量差发生大量丢包时，可能在一端无法统计到，此时需要两端排查确认是否光纤链路导致的问题，确定链路后对光纤线/光模块进行更换。

两个端口使用一根光纤线连接时，必须满足每个端口的TX连接对端端口的RX才能正常连接。







## 8.FC端口的速率协商

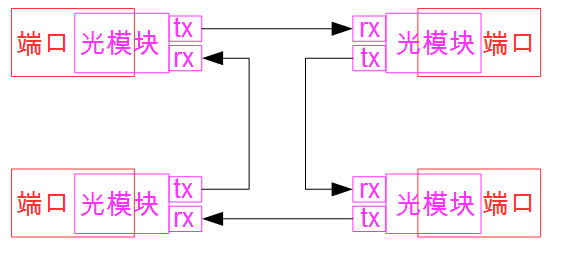
当前主流FC端口支持速率有2G/4G/8G/16G,如一个最高支持8G速率的端口，其设置速率通常为2G/4G/8G/Auto，其中Auto表示速率自适应；

一个最高支持8G速率的端口和一个最高支持16G速率的端口连接，如果他们的速率都设置为Auto，则只能协商为8G；如果一侧设置了速率为4G，则这两个端口只能协商为4G。FC端口不支持与光口的ETH/ISCSI端口对接。

## 9．拓扑：LOOP, P2P/Fabric

FC协议定义了三种拓扑模式，LOOP/P2P（点对点，直连）/Fabric，每种拓扑模式在建立连接时有自己的状态机，其中P2P/Fabric两种拓扑模式差异较小，几乎可以作为一种。

在没有交换机的情况下可以利用LOOP模式将多台设备连接在一起，如下图所示。通过loop连接的所有端口都是设备端口，则称为private loop，如果loop中存在交换机端口，则称为public loop（回忆FLOGI部分，NportID的最低8bit）。



Loop模式的缺点在于：

1. 链路利用率较低；
2. 容易造成误码扩散。

## 10.流量控制机制BB\_CREDIT

FC协议传输数据过程中采用基于Buffer的流控机制，具体实现方式为AB两个FC端口通过光纤进行物理对接，在FLOGI过程中端口A会告知端口B自己的buffer数量--buffer\_count\_A，同时B端口会记录该值保存到X中。此后B端口每发送一帧数据，X减1；当A端口收到B端口的FC帧后会发送R\_RDY信号，B端口每收到一个R\_RDY信号；X加1。若A端口由于各种异常无法处理B端口发送的FC帧，导致无法回复R\_RDY信号，B端口的X减到0时就不能再向外发送数据。

todo补充：BB\_CREDIT导致端口流量反压。

Todo：Exchange/Sequence/Frame概念介绍、ABTS、Port Fencing

Todo:RSCN

## 11.FC相关操作

### 1）存储设备FC端口配置操作

FC协议通过复杂的机制提供了简单的用户接口，其中FC端口标识WWPN/WWNN由硬件固定，Nport ID由交换机自动分配或者两个端口自行协商，不需要人工介入。

端口建立连接由软硬件自动完成：针对新接入FC网络中的设备A，FC协议提供了查询机制，新接入的设备A可以查询网络中有哪些设备可以与之通信；针对已接入网络中的设备B，网络中其他设备加入或者离开网络时，有主动通知机制（RSCN）将加入或者离开网络的设备信息通知给B；这些都不需要人工参与。

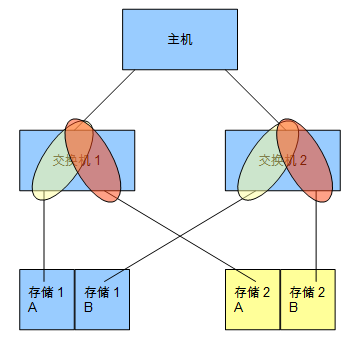
IO的处理由底层物理线路触发，或者上层的SCSI模块调用，也不需要人工介入。

FC协议可配置的参数主要有两个：拓扑模式和端口速率，通常端口拓扑模式有LOOP/FABRIC/AUTO三种选项，端口根据实际支持的最大速率(假设为8G)有2G/4G/8G/AUTO几种选项；上文中的AUTO表示自适应，两个端口通过协商确定拓扑类型或者速率。除需要保证端口在非AUTO配置下，两个端口参数一致外，FC很少需要其他配置操作；基本可以通过光纤线即连即用。

### 2）FC交换机的基本操作

FC交换机通过划zone的方式确定两个端口是否连通。

如下图所示，在交换机1上连接主机的交换机端口分别与连接存储1A的交换机端口以及连接存储2A的交换机端口划到同一个zone中，主机端口可以分别与这两个存储端口连通；两个存储端口则相互隔离。



新组网时如果使用交换机，需要对连接在交换机上的设备进行划zone操作；在使用交换机之前需要先对交换机端口进行规划，规划好哪些端口连接主机那些端口连接存储；针对需要通信的端口进行划zone，建议两个端口一个zone，不要采用大zone的形式。

在交换机上配置zone的步骤如下所示：

**1.创建zone**

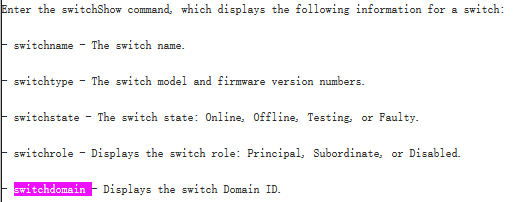
zonecreate "zone\_name","domain\_id,port;domain\_id,port;....."

示例：zonecreate "zone\_1","1,0;1,1;1,2;1,3"

还可以通过zone\_add命令添加zone成员：

zoneadd "zone\_1","1,4"

domain\_id是交换机的编号，多台交换机级联组网时，每台交换机的domain\_id各不相同，可通过switchshow命令执行结果中的switchdomain查看，如下图所示：



**2.创建zoneconfig**

zoneconfig是一个或多个zone的组合，初始创建zoneconfig时，至少要包含一个zone成员，否则无法创建。

语法：cfgcreate "zoneconfig\_name","zone\_name1;zone\_name2...."

示例：cfgcreate "zoneconfig\_1","zone\_1;zone\_2"

后续可通过cfgadd命令为zoneconfig添加其他zone成员。

语法：cfgadd "zoneconfig\_name","zone\_name1;zone\_name2..."

范例：cfgadd "zoneconfig\_1","zone3;zone4;zone5"

**3.执行cfgsave命令保存配置**

**4.激活zonecfg**

语法： cfgenable "zoneconfig\_name"

示例：cfgenable "zoneconfig\_1"

上述是交换机的最基本操作；目前FC交换机市场基本由Brocade垄断，Brocade系列交换机操作基本类似，通过互联网可以方便的查询到需要的操作命令。

## 12.FC组网最佳实践

ISL（Inter-switch links）可以增加SAN的扩展性和灵活性，同时也有可能带来了阻塞.合理规划和配置ISL对SAN非常重要。IBM best practice里面提到如下应该注意的点：

1. ISL占用比率：一般情况下7:1，高性能SAN建议3:1配置；（3个host使用1个ISL），如图12- 1所示，SAN交换机上连接8个host端口，则级联链路条数为8/3=3(向上取整)条。
2. 避免不必要的ISL traffic；对于规模较小的网络，建议图中A路B路各保留一台交换机，无需级联；使用级联时采用图中的垂直级联，尤其尽量避免A路和B路交换机之间的水平级联。
3. 增加ISL链路的冗余，确保交换机打开trunking license来获得较好的性能；
4. 避免主机和存储之间超过2级switch；

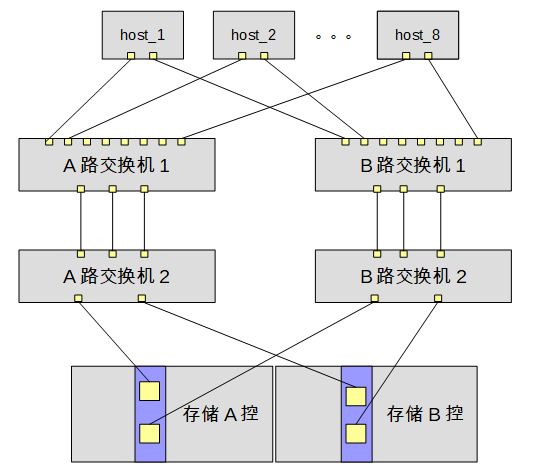


图12- 1 FC组网最佳实践

AB两路之间水平级联交换机实际上形成了一个fabric，稳定性弱于两个独立的网络。