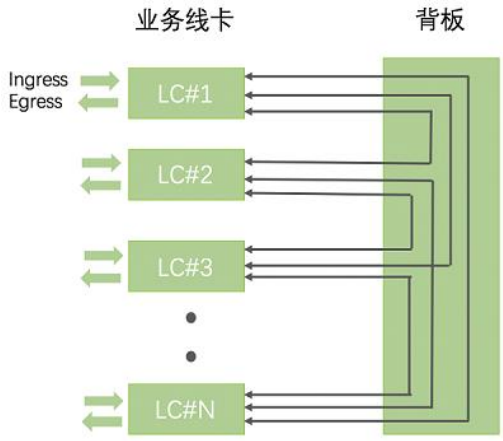
每日一学：人人都学交换机之框式交换机架构及技术原理

框式核心交换机先后出现了多种硬件架构，而现在最为常用的有三种：Full-Mesh 交换架构、Crossbar 矩阵交换架构和基于 Cell 的 CLOS 交换架构。今天主要通过对这三种硬件架构、报文转发流程等原理的分析，全面剖析三种架构的优劣势。

Full-Mesh 架构说明



▲图 1：Full-Mesh 架构图

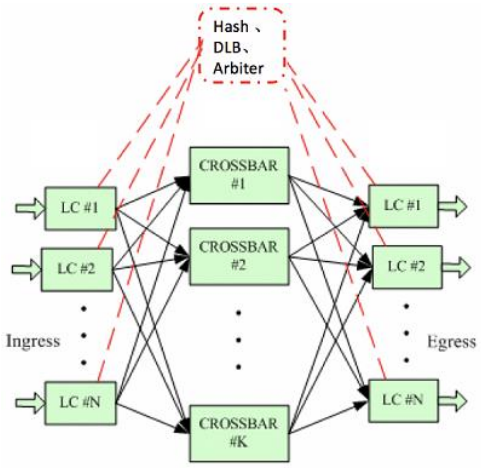
如图 1 所示，所有业务线卡通过背板走线连接到其它线卡，因为 Full-Mesh 不需要外部的交换芯片，而是任意两个节点间都有直接连接，故得名全连接。

由于各线卡需要 Full-Mesh 互联，一个节点数为 N 的 Full-Mesh，连接总数为[N×(N-1)]÷2，所以随着节点数量增加连接总数也急剧上升，因而可扩展性较差，仅适用于槽位数量较少的核心设备。

报文转发流程

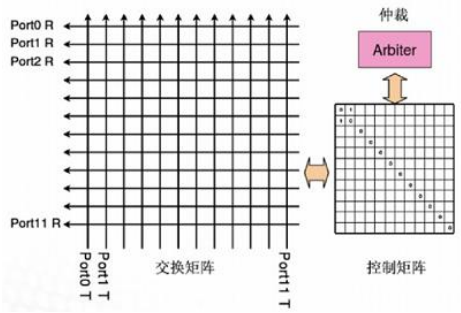
1、报文从线卡进入，跨卡报文送到与目的线卡连接的背板通路；
2、报文到达目的线卡。

Crossbar 架构说明



▲图 2:Crossbar 架构图

如图 2 所示，业务线卡通过背板走线连接到 Crossbar 芯片上，Crossbar 芯片集成在主控引擎上。



▲图 3:Crossbar 芯片架构

Crossbar 芯片架构如图 3 所示，每一条输入链路和输出链路都有一个 CrossPoint，在 CrossPoint 处有一个半导体开关连接输入线路和输出线路，当来自某个端口的输入线路需要交换到另一个端口的输出点时，在 CPU 或交换矩阵的控制下，将交叉点的开关连接，数据就被发到另一个接口。

简单地说，Crossbar 架构是一种两级架构，它是一个开关矩阵，每一个 CrossPoint 都是一个开关，交换机通过控制开关来完成输入到特定输出的转发。如果交换具有 N 个输入和 N 个输出，那么该 Crossbar Switch 就是一个带有 $N*(N-1)\approx N^2$ 个 CrossPoint 点的矩阵，可见，随着端口数量的增加，交叉点开关的数量呈几何级数增长。对于 Crossbar 芯片的电路集成水平、矩阵控制开关的制造难度、制造成本都会呈几何级数增长。所以，采用一块 Crossbar 交换背板的交换机，所能连接的端口数量也是有限的。

报文转发流程

• 无缓存 Crossbar

每个交叉点没有缓存，业务调度采用集中调度的方式，对输入输出进行统一调度，报文转发流程如下：

1、报文从线卡进入，线卡先向 Arbiter 请求发送；
2、Arbiter 根据输出端口队列拥塞情况，决定是否允许线卡发送报文到输出端口；
3、报文通过 Crossbar 转发到目的线卡输出端口。

由于是集中调度，所以仲裁器的调度算法复杂度很高，扩展性较差，系统容量大时仲裁器容易形成瓶颈，难以做到精确调度。

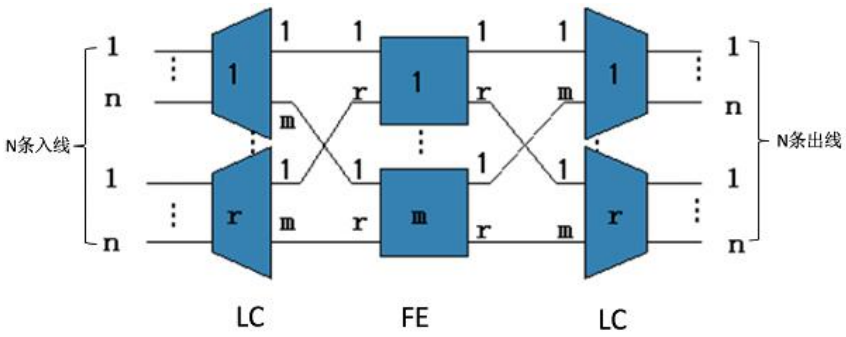
- **缓存式 Crossbar**

最早的缓存式 Crossbar 只有交叉节点带缓存，而输入端是无缓存的，被称为”bus matrix”,后来，CICQ 的概念被引入，即在输入端用大的 Input Buffer，在中间节点用小的 CrossPoint Buffer。

这种结构采用分布式调度的方式进行业务调度，即输入和输出端都有各自的调度器，报文转发流程如下：
1、报文从线卡进入，输入端口通过特定的调度算法（如 RR 算法）独立地选择有效的 VOQ；
2、将 VOQ 队列头部分组发送到相应的交叉点缓存；
3、输出端口通过特定的算法在非空的交叉点缓存中选择进行服务。

由于输入和输出的调度策略相互独立，所以很难确保交换系统在每个时隙整体上达到理想匹配状态，并且调度算法复杂度和交换系统规模有关，限制了其扩展性。

CLOS 架构说明



▲图 4:CLOS 架构图

如图 4 所示，每块业务线卡和所有交换网板相连，交换芯片集成在交换网板上，实现了交换网板和主控引擎硬件分离。CLOS 架构是一种多级架构，每个入口级开关和每个中间级开关之间只有一个连接，并且，每个中间级开关正好连接到每个出口级开关，这种架构的优点是可以通过多个小型 Crossbar 开关来实现大量输入和输出端口之间的连接，CrossPoint 数量级别低于 Crossbar 架构的 N 的 2 次方，降低了芯片实现难度。

报文转发流程

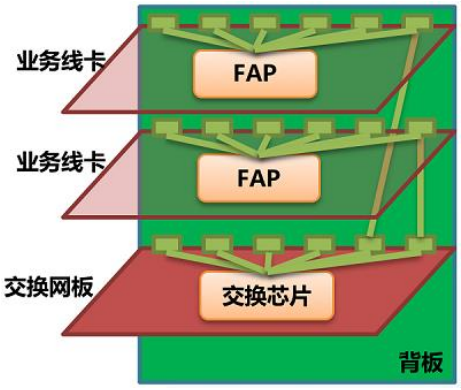
• 基于 Cell 的动态负载

1、入方向线卡将数据包切分为 N 个 cell，其中：N=下一跳可用线路数量；
2、交换网板采用动态路由方式，即根据下一级各链路的实际可用交换能力，动态选路和负载均衡，通过多条路径将分片发送到出方向线卡；
3、出方向线卡重组报文。

动态负载关键点在于能负载分担地均衡利用所有可达路径，由此实现了无阻塞交换。

CLOS 架构交换机的分类

• 非正交背板设计

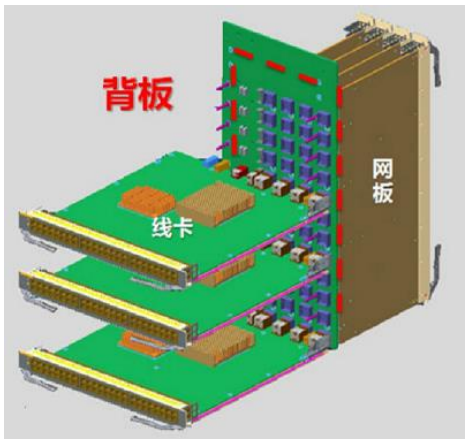


▲图 5:非正交背板

如图 5 所示，业务线卡与交换网板互相平行，板卡之间通过背板走线连接。

背板走线会带来信号干扰，背板设计也限制了带宽的升级，同时，背板上 PCB 的走线要求很高，从背板开孔就成了奢望，这直接导致纯前后的直通风道设计瓶颈一直无法突破。

• 正交背板设计

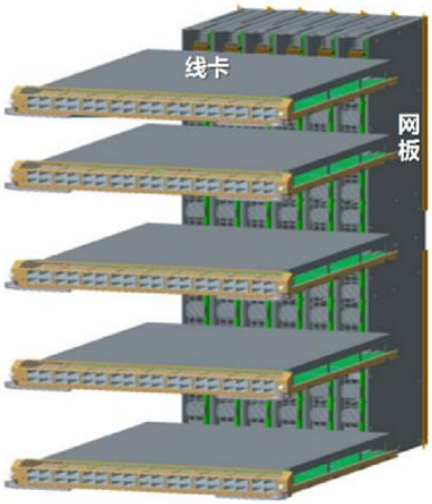


▲图 6:正交背板

如图 6 所示，交换机线卡与交换网板分别与背板对接。

同非正交背板设计一样，背板带宽限制了带宽的升级，同时也增加了散热的难度。

- **正交零背板设计**



▲图 7:正交零背板

如图 7 所示，业务线卡与交换网板互相垂直，背板走线为零，甚至无中板。

正交设计能减少背板走线带来的高速信号衰减，提高了硬件的可靠性，无背板设计能够解除背板对容量提升的限制，当需要更大带宽的时候，只需要更换相应板卡即可，大大缩短业务升级周期，并且因为没有了背板的限制，交换机直通风道散热问题迎刃而解，匹配数据中心机房空气流的走向，形成了贯穿前后板卡的高速、通畅的气流。

总结

下表将对以上三种架构做出总结：

对于高端机架式交换机，以 Crossbar 交换架构和 CLOS 交换架构为主。

	Full-Mesh	Crossbar		CLOS		
分类	-	无缓存	有缓存	非正交背板	正交背板	正交零背板
硬件架构	● 无交换网板 ● 线卡之间通过背板走线相连	● 单平面交换 ● 集中调度 ● 交叉点无缓存	● 单平面交换 ● 分布式调度 ● 交叉点有缓存	● 多平面交换 ● 线卡和交换网板平行 ● 背板长走线	● 多平面交换 ● 线卡和交换网板正交 ● 背板短走线	● 多平面交换 ● 线卡和交换网板正交 ● 无背板无走线
性能特点	● 受限于背板带宽和连接总数，扩展性差 ● 背板带宽是瓶颈	● 随端口数增加 CrossPoint 数量呈几何增长 ● 系统容量大时仲裁器易形成瓶颈	● 随端口数增加 CrossPoint 数量呈几何增长 ● 调度算法复杂度限制扩展	● 背板限制带宽扩展且无法实现直通散热 ● 走线带来信号衰减 ● 基于 cell 的动态负载实现无阻塞	● 背板限制带宽扩展且无法实现直通散热 ● 基于 cell 的动态负载实现无阻塞	● 带宽扩展更换相应网板即可 ● 无背板设计实现交换机直通散热 ● 基于 cell 的动态负载实现无阻塞
适用设备	● 低密度槽位	● 高密度槽位 ● 可面向未来 1-3 年扩展		● 高密度槽位 ● 可面向未来 1-3 年扩展		● 高密度槽位 ● 可面向未来 10 年扩展