路由器基本原理和结构体系

一、路由器在IP网络中的位置

IP是一种网络间的互连协议。整个IP网络，由许多子网络构成，各子网络又由许多主机组成。子网之间可以使用不同的链路层协议，如Ethernet或PPP等，同一子网必须使用相同的链路协议。在网络层，主机用IP地址寻址，IP地址实行全网统一管理。IP地址通过子网掩码而分成两部分：Net ID和Host ID。同一子网内部使用相同的Net ID，而Host ID各不相同。子网内部的主机通信，由链路协议直接进行；子网之间的主机通信，要通过路由器来完成。路由器是多个子网的成员，在它的内部有一张表示Net ID与下一跳端口对应关系的路由表。通信起点主机发出IP包被路由器接收后，路由器查路由表，确定下一跳输出端口，发给下一台路由器，这台路由器又转发给另外一台路由器，用这样一跳接着一跳的方式，直到通信终点另一台主机收到这个IP包。 IP协议的网络层是无连接的，路由器中没有表示连接状态的信息。路由器在网络层也没有重发机制和拥塞控制。IP协议重发机制和拥塞控制由传输层TCP来处理，按端到端的方式运行。传输层拥塞控制通过TCP慢启动实现。

IP协议把网络划分为物理层（L1）、链路层（L2）、网络层（L3）、传输层（L4）及应用层（L7）五个层次。处理物理层的设备有Hub集线器，处理链路层的设备有L2以太交换机，路由器是在网络层转发数据的设备。L3以太交换机是IP网络路由器的特例，通常只有以太线路接口，工作在纯以太网络环境中。

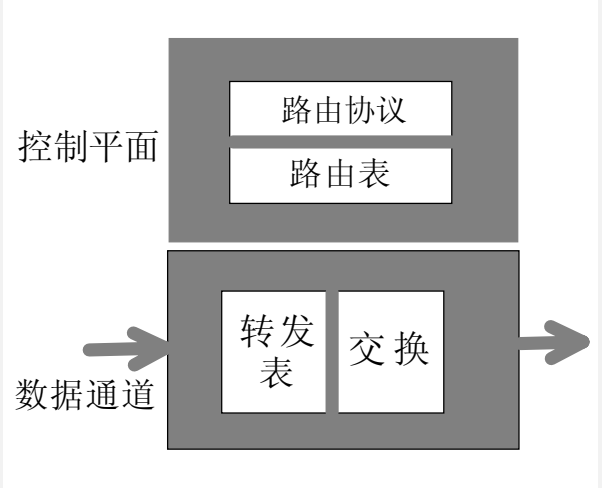
二、路由器工作原理

路由表是工作在IP协议网络层实现子网之间转发数据的设备。路由器内部，如图，可以划分为控制平面和数据通道。在控制平面上，路由协议可以有不同的类型，如OSPF、BGP等。路由器通过路由协议交换网络的拓扑结构信息，依照拓扑结构动态生成路由表。在数据通道上，转发引擎从输入线路接收IP包后，分析与修改包头，使用转发表查找输出端口，把数据交换到输出线路上。转发表是根据路由表生成的，其表项和路由表项有直接对应关系，但转发表的格式和路由表的格式不同，它更适合实现快速查找。

转发的主要流程包括线路输入、包头分析、数据存储、包头修改和线路输出。

IP包从不同的线路上到达路由器的接口卡，线路输入处理部分对它进行信号恢复、解码和CRC校验，然后放进输入FIFO。输入FIFO的数据要送入数据存储器，数据存储器可以是CPU控制主内存或逻辑控制的专用内存。新输入数据放在系统输入队列尾部，CPU或逻辑从输入队列取出报文进行分析，需要分析的内容主要是L3包头中的目的IP地址，有些情况也L3包头的其他部分，甚至包括L2和L4包头。包头分析首先滤掉IP头校验和有错的报文，然后确定是协议报文还是转发报文。协议报文送协议软件处理，转发报文要查转发表确定输出端口，查流分类表确定输出队列。 每个端口可以有若干个输出队列，他们对应于不同的优先级别。 输出队列调度模块根据特定的规则，把选中的报文交给输出FIFO。报文在进入输出FIFO之前，要修改包头。修改包头包括IP TTL值减一，更新IP头校验和，替换L2的地址等。线路输出处理部分从输出FIFO中取出数据，更新链路层CRC数值，然后编码，经信号调制发送到输出线路上。这就是IP包转发的基本流程，如果支持更多的IP业务，如ACL，NAT等，在上述流程中还要增加额外的过滤和处理。

路由协议根据网络拓扑结构动态生成路由表。IP协议把整个网络划分为管理区域，这些管理区域称为自治域，自治域区号实行全网统一管理。这样，路由协议就有域内协议和域间协议之分。域内路由协议，如OSPF、IS-IS，在路由器间交换管理域内代表网络拓扑结构的链路状态，根据链路状态推导出路由表。 域内路由协议相邻节点之间，采用多播或广播方式通信。域间路由协议，如BGP，根据距离向量和过滤策略生成全网路由表。 域间路由协议相邻节点交换数据，不能使用多播方式，只能采用指定的点到点连接。域间路由协议不能使用缺省路由，BGP路由表必须表达IP网络全部子网的信息，所以路由表项较多。尽管使用IP

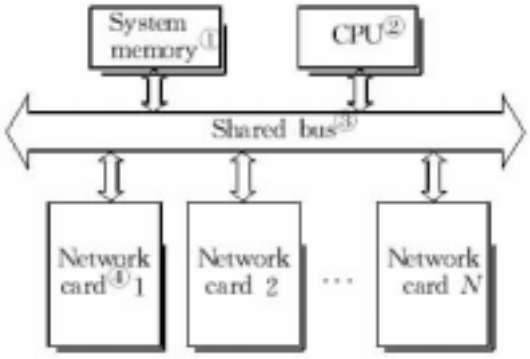


地址子网聚合技术后，路由表项得到有效减少，到2001年，BGP路由表项已经接近100K。不断增大的路由表项，是IP网络必须面对的现实之一。小型企业网络，位于网络边沿，采用人工配置的静态路由或简单协议RIP即可。使用缺省路由后，路由表项的大小只受企业内部子网划分的影响。

三、路由器结构体系

路由器内部可以划分为控制平面和数据通道。路由器的控制平面，运行在通用CPU系统中，多年来一直没有多少变化。在高可用性设计中，可以采用双主控进行主从式备份，来保证控制平面的可靠性。路由器的数据通道，为适应不同的线路速度，不同的系统容量，采用了不同的实现技术。 路由器的结构体系正是根据数据通道转发引擎的实现机理来区分。简单而言，可以分为软件转发路由器和硬件转发路由器。软件转发路由器使用CPU软件技术实现数据转发，根据使用CPU的数目，进一步区分为单CPU的集中式和多CPU的分布式。硬件转发路由器使用网络处理器硬件技术实现数据转发，根据使用网络处理器的数目及网络处理器在设备中的位置，进一步细分为单网络处理器的集中式、多网络处理器的负荷分担并行式和中心交换分布式。

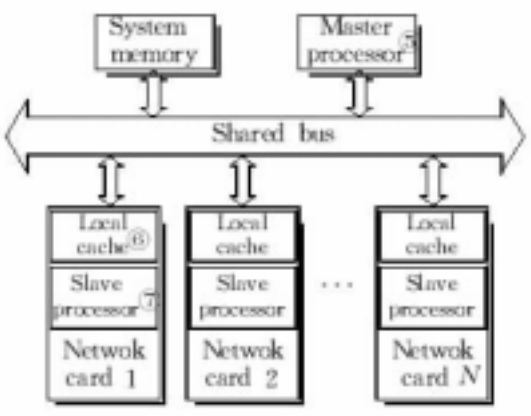
（下面的结构图可以删去）



1、软件转发集中式

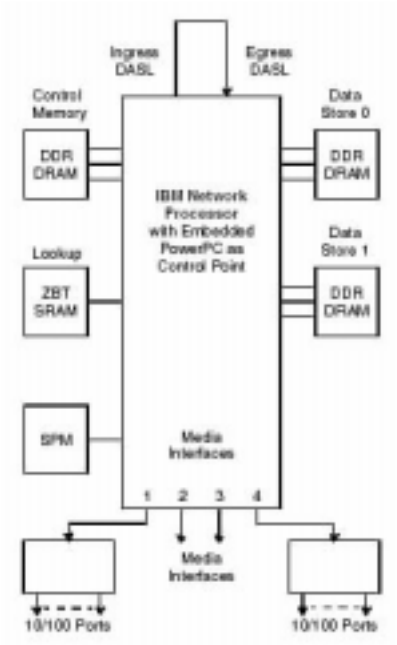
软件转发集中式路由器本质上是一台单CPU专用计算机系统，数据转发和路由协议由CPU分时处理，线路接口是CPU外设。系统中，所有线路接口、所有软件共享唯一的CPU资源，所以整机性能不高。在交换带宽方面，系统总线或接口总线决定了系统中各线路接口的总有效带宽。总线带宽一般小于1Gbps。在转发性能方面，

CPU性能、转发流程和查表算法决定了系统的总处理能力。查表算法对系统性能影响很大，需要使用快速算法提升转发性能。常用快速查表算法包括Hash Table、Patricia Tree和256 Way Multiway Tree等。使用快速算法后，系统的转发速度也不会超过400Kpps。机壳结构方面，软件转发集中式路由器可以封装成主板上出线路接口的集成式路由器，或者使用插卡式线路接口的模块式路由器。



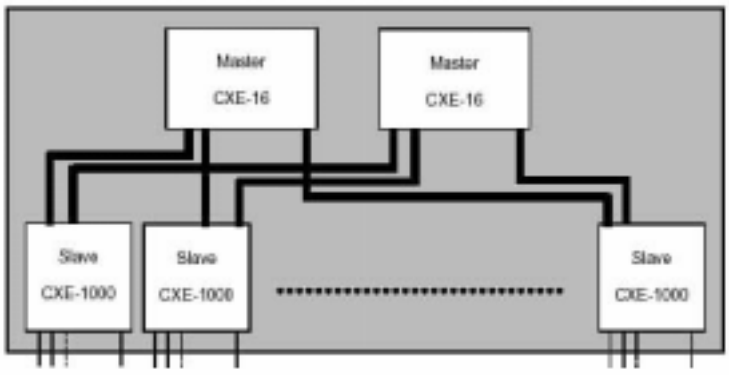
2、软件转发分布式

软件转发分布式路由器采用多CPU设计，CPU之间通过背板总线而实现内存共享。这种系统的总线背板上，可以插接多个CPU处理板，其中，有专门处理路由协议的主控板，专门处理报文转发的接口板。接口板上有局部总线，局部总线上可挂接多个子插卡，子插卡作为线路卡。 这种体系的路由器是为提高转发性能而设计的，有多个接口板并行处理报文转发，转发能力大幅提升。在交换带宽方面，跨接口板转发的数据要经过多个总线桥，所以如何提高背板总线效率，是系统设计的关键。这类路由器适合于跨接口板通信较少而业务处理复杂的场合。在硬件结构上，这类设备可以支持处理板热插拔、主控板双备份等高可用性特性。



3、硬件转发集中式

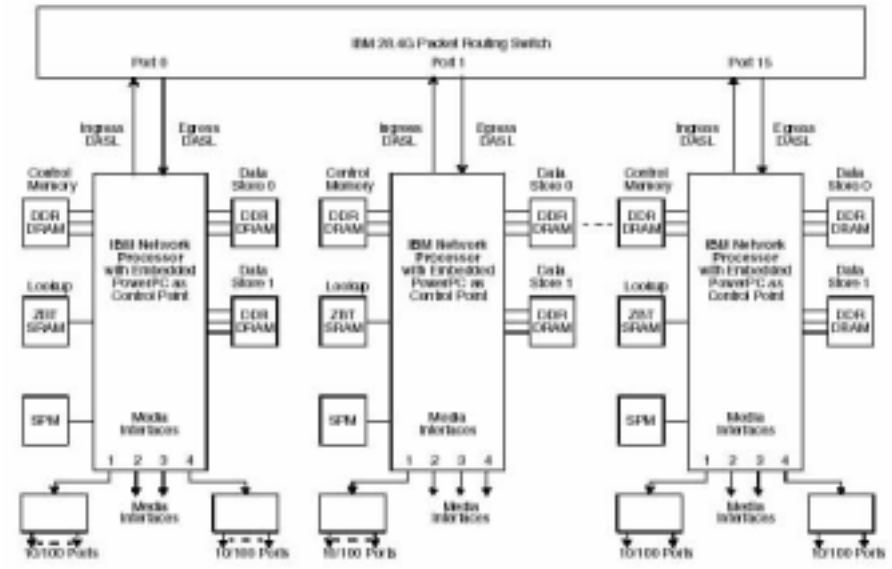
硬件转发集中式路由器在软件转发集中式路由器的基础上增加了网络处理器。这样，数据转发用专门的网络处理器来完成，而CPU用于处理路由协议和系统管理。网络处理器可以使用微码编程的微引擎，固化逻辑ASIC芯片以及可重配置的FPGA来实现。经过精心设计的网络处理器能够保证所有线路接口达到线速。 在交换带宽方面，网络处理器内部，使用了独立报文存储系统，交换带宽主要由内存读写速度决定，可以通过提高内存时钟或增加内存位宽增加交换带宽。目前，处理2.5Gbps线路微码逻辑网络处理器基本成熟，也可以见到几十Gbps交换带宽的固化逻辑网络处理器。在转发处理方面，使用硬件查表技术提升系统处理能力。硬件查表技术主要有两种：基于逻辑算法设计的SRAM技术和基于半导体工艺设计TCAM技术。SRAM技术可以达到5 Mpps的性能，TCAM技术可以达到100 Mpps的性能。微码编程的网络处理器可以现场升级，修改或增加新业务很方便。硬逻辑网络处理器能获得最高的性能，更低的价格。



4、硬件转发并行式

硬件转发并行式路由器是使用负荷分担方式提高提高系统容量的一种办法。一般使用2或4个网络处理器作为主芯片，主芯片外接副芯片，是一种主从式结构。主芯片不带线路接口，副芯片带线路接口。各主芯片以负荷分担的方式增大系统处理能力。副芯片上有流量分配机制和流量会聚机制，线路接口的报文转发流量按特定算法，均匀分配给中心网络处理器，网络处理器后由副芯片会聚交换到线路上。合理设计副芯片的带宽，可以保证交换无阻，流量分配机制的实现是系统的关键点。这种系统的总容量为网络处理器的2或4倍。

5、硬件转发分布式



硬件转发分布式路由器是使用中心交换方式提高提高系统容量的另一种办法。这种系统的特征是，核心为大容量交换网络，交换网络外挂网络处理器，网络处理器上挂线路接口。整个转发系统是星形结构，中心是交换网络。交换网络使用信元

交换，边沿网络处理器负责IP转发。这类路由器的系统总容量由交换网络容量决定，最高线路接口速率由网络处理器性能决定。交换网络只实现定长信元交换，容量可以作到很大，可见这是一种大容量设计方案。 网络处理器的数目在16个左右。

四、路由器性能分析

我们知道：依据路由表的操作方法的不同，在路由器内部划分了控制平面部分和数据通道部分，这两部分对路由器的使用有着不同的影响。

在数据通道上，IP报文处理能力主要受交换部件的有效带宽和转发部分的处理速度影响。交换带宽，用技术指标bps（比特/秒）来衡量，一般在纯大包的条件下测定。转发速度，用技术指标pps（包/秒）来衡量，一般在纯小包的条件下测定。软件转发单CPU路由器，交换带宽和转发能力为系统共享，线路卡个数不同，各线路卡会有不同的性能表现。软件转发多CPU路由器，转发能力有明显优势，交换带宽没有明显提高。硬件转发路由器，一般要针对线速进行设计，系统交换带宽大于各线路接口的总和，转发处理能力在纯小包的情形下也能胜任，区别的只是容量和价格。线速路由器，保证线路接口在各种情况下能够达到满速，这时候，是线路而不是设备是网络的瓶颈。另外，IP业务，如ACL，NAT、IPSec等，是否造成性能下降，也是设计和使用路由器要考虑的问题。

在控制平面上，小型网络问题不大，只要支持选用的路由协议就可以了。对于大型网络，特别是有独立自治域号的运营商网络，在路由协议类型满足要求的情况下，应该考察路由表项大小是否满足，域间路由协议相邻节点的连接数目是否满足，路由表项更新速度如何，路由更新时对数据通道上的处理有无影响，等等。