Linux 对多队列网卡的支持

Herbert Xu
Red Hat Inc.

背景

以太网速和 CPU 频率的演变：

10Mb/s	几十兆赫
100Mb/s	几百兆赫
1Gb/s	几千兆赫
10Gb/s	同上：多核＋ SMP

一张 10Gb/s 网卡必须由多个核来运用。

多核与多队列

多核与 SMP 相似，必须用锁。

锁导致了 CPU 资源的浪费。

多核浪费不起 CPU 资源。

多队列能解决浪费问题：

- 每个核可有独立的队列和中断。
- 传送： CPU 负责包的分配。
- 接收：网卡负责包的分配。

多队列传送的支持

原先：

一个网络设备有一个软件队列。

一个软件队列对应一个硬件队列。

07 年 7 月（ Intel ）：

一个软件队列对应多个硬件队列。

硬件队列的瓶颈解决了。

软件队列（ qdisc ）成为瓶颈。

多队列传送的支持

08 年 7 月（David S. Miller）：

默认：每个设备有多个软件队列。

一个软件队列对应一个硬件队列。

非默认：每个设备有一个软件队列。

一个软件队列对应多个硬件队列。

基本解决了默认软件队列的瓶颈问题。

多队列接收的支持

分配由硬件决定。

通常对包头用散列函数。

软件支持可由驱动程序直接实现。

但 NAPI 的实现比较难。

07 年 10 月：多队列 NAPI 接收的支持。

仍待解决的问题

全局分配仍需优化：

传送队列和接收队列的分配应该同步。

队列分配和 CPU 调度也应同步。

对非默认的软件队列的支持：

暂时仅只持单一软件队列。

应把多队列更加细化（如 sfq 等）。

虚拟化和其他应用的支持。

问题