Linux之TCPIP内核参数优化

本文以Ubuntu 12.04 LTS Desktop (x64)默认配置为例（机器的内存为4GB），推荐先阅读《TCP连接的状态与关闭方式，及其对Server与Client的影响》、《Windows系统下的TCP参数优化》，以了解TCP优化的相关知识。

/proc/sys/net目录

所有的TCP/IP参数都位于/proc/sys/net目录下（请注意，对/proc/sys/net目录下内容的修改都是临时的，任何修改在系统重启后都会丢失），例如下面这些重要的参数：

参数（路径+文件）	描述	默认值	优
/proc/sys/net/core/rmem_default	默认的TCP数据接收窗口大小（字节）。	229376	25
/proc/sys/net/core/rmem_max	最大的TCP数据接收窗口（字节）。	131071	51
/proc/sys/net/core/wmem_default	默认的TCP数据发送窗口大小（字节）。	229376	25
/proc/sys/net/core/wmem_max	最大的TCP数据发送窗口（字节）。	131071	51
/proc/sys/net/core/netdev_max_backlog	在每个网络接口接收数据包的速率比内核处理这些包的速率快时，允许送到队列的数据包的最大数目。	1000	2
/proc/sys/net/core/somaxconn	定义了系统中每一个端口最大的监听队列的长度，这是个全局的参数。	128	2
/proc/sys/net/core/optmem_max	表示每个套接字所允许的最大缓冲区的大小。	20480	81
	确定TCP栈应该如何反映内存使用，每个值的单位		

/proc/sys/net/ipv4/tcp_mem	都是内存页（通常是4KB）。第一个值是内存使用的下限；第二个值是内存压力模式开始对缓冲区使用应用压力的上限；第三个值是内存使用的上限。在这个层次上可以将报文丢弃，从而减少对内存的使用。对于较大的BDP可以增大这些值（注意，其单位是内存页而不是字节）。	94011 125351 188022	131 262 524
/proc/sys/net/ipv4/tcp_rmem	为自动调优定义socket使用的内存。第一个值是为socket接收缓冲区分配的最少字节数；第二个值是默认值（该值会被rmem_default覆盖），缓冲区在系统负载不重的情况下可以增长到这个值；第三个值是接收缓冲区空间的最大字节数（该值会被rmem_max覆盖）。	4096 87380 4011232	87 256 408
/proc/sys/net/ipv4/tcp_wmem	为自动调优定义socket使用的内存。第一个值是为socket发送缓冲区分配的最少字节数；第二个值是默认值（该值会被wmem_default覆盖），缓冲区在系统负载不重的情况下可以增长到这个值；第三个值是发送缓冲区空间的最大字节数（该值会被wmem_max覆盖）。	4096 16384 4011232	87 256 408
/proc/sys/net/ipv4/tcp_keepalive_time	TCP发送keepalive探测消息的间隔时间（秒），用于确认TCP连接是否有效。	7200	18
	探测消息未获得响应时，		

/proc/sys/net/ipv4/tcp_keepalive_intvl	重发该消息的间隔时间（秒）。	75	
/proc/sys/net/ipv4/tcp_keepalive_probes	在认定TCP连接失效之前，最多发送多少个keepalive探测消息。	9	
/proc/sys/net/ipv4/tcp_sack	启用有选择的应答（1表示启用），通过有选择地应答乱序接收到的报文来提高性能，让发送者只发送丢失的报文段，（对于广域网通信来说）这个选项应该启用，但是会增加对CPU的占用。	1	
/proc/sys/net/ipv4/tcp_fack	启用转发应答，可以进行有选择应答（SACK）从而减少拥塞情况的发生，这个选项也应该启用。	1	
/proc/sys/net/ipv4/tcp_timestamps	TCP时间戳（会在TCP包头增加12个字节），以一种比重发超时更精确的方法（参考RFC 1323）来启用对RTT 的计算，为实现更好的性能应该启用这个选项。	1	
/proc/sys/net/ipv4/tcp_window_scaling	启用RFC 1323定义的window scaling，要支持超过64KB的TCP窗口，必须启用该值（1表示启用），TCP窗口最大至1GB，TCP连接双方都启用时才生效。	1	
/proc/sys/net/ipv4/tcp_syncookies	表示是否打开TCP同步标签（syncookie），内核必须打开了CONFIG_SYN_COOKIES项进行编译，同步标签可以防止一个套接字在有过多试图连接到达时引起过	1	

	载。		
/proc/sys/net/ipv4/tcp_tw_reuse	表示是否允许将处于TIME-WAIT状态的socket（TIME-WAIT的端口）用于新的TCP连接 。	0	
/proc/sys/net/ipv4/tcp_tw_recycle	能够更快地回收TIME-WAIT套接字。	0	
/proc/sys/net/ipv4/tcp_fin_timeout	对于本端断开的socket连接，TCP保持在FIN-WAIT-2状态的时间（秒）。对方可能会断开连接或一直不结束连接或不可预料的进程死亡。	60	
/proc/sys/net/ipv4/ip_local_port_range	表示TCP/UDP协议允许使用的本地端口号	32768 61000	10 65
/proc/sys/net/ipv4/tcp_max_syn_backlog	对于还未获得对方确认的连接请求，可保存在队列中的最大数目。如果服务器经常出现过载，可以尝试增加这个数字。	2048	20
/proc/sys/net/ipv4/tcp_low_latency	允许TCP/IP栈适应在高吞吐量情况下低延时的情况，这个选项应该禁用。	0	
/proc/sys/net/ipv4/tcp_westwood	启用发送者端的拥塞控制算法，它可以维护对吞吐量的评估，并试图对带宽的整体利用情况进行优化，对于WAN 通信来说应该启用这个选项。	0	
/proc/sys/net/ipv4/tcp_bic	为快速长距离网络启用Binary Increase Congestion，这样可以更好地利用以GB速度进行操作的链接，对于WAN通信应该启用这个选项。	1	

/etc/sysctl.conf文件

/etc/sysctl.conf是一个允许你改变正在运行中的Linux系统的接口。它包含一些TCP/IP堆栈和虚拟内存系统的高级选项，可用来控制Linux网络配置，由于/proc/sys/net目录内容的临时性，建议把TCPIP参数的修改添加到/etc/sysctl.conf文件, 然后保存文件，使用命令“/sbin/sysctl －p”使之立即生效。具体修改方案参照上文：

net.core.rmem_default = 256960

net.core.rmem_max = 513920

net.core.wmem_default = 256960

net.core.wmem_max = 513920

net.core.netdev_max_backlog = 2000

net.core.somaxconn = 2048

net.core.optmem_max = 81920

net.ipv4.tcp_mem = 131072 262144 524288

net.ipv4.tcp_rmem = 8760 256960 4088000

net.ipv4.tcp_wmem = 8760 256960 4088000

net.ipv4.tcp_keepalive_time = 1800

net.ipv4.tcp_keepalive_intvl = 30

net.ipv4.tcp_keepalive_probes = 3

net.ipv4.tcp_sack = 1

net.ipv4.tcp_fack = 1

net.ipv4.tcp_timestamps = 1

net.ipv4.tcp_window_scaling = 1

net.ipv4.tcp_syncookies = 1

net.ipv4.tcp_tw_reuse = 1

net.ipv4.tcp_tw_recycle = 1

net.ipv4.tcp_fin_timeout = 30

net.ipv4.ip_local_port_range = 1024 65000

net.ipv4.tcp_max_syn_backlog = 2048