编译论坛

先进编译实验室
Advanced Compiler

系统配置优化I

嘉宾：柴晓楠

先进编译实验室
Advanced Compiler

处理器

处理器是操作系统稳定运行的根本，处理器的速度与性能在很大程度上决定了系统整体的性能，因而处理器通常是系统性能分析的首要目标。为了提高性能，优化人员可以通过配置检查来查看处理器的相关信息，以寻找性能改进的空间，明确消耗时间的位置和原因，再通过参数调整来进行调优。本节将分为以下两个部分进行描述：

（1）配置检查；

（2）参数调整；

可以从以下角度检查处理器的当前配置，例如当前可用的处理器数量、是否支持超线程、每个核上运行的线程数、当前处理器的模式以及架构、缓存大小、是否支持硬件虚拟化、处理器的时钟频率、基本输入输出系统BIOS已启用或者禁用的其它处理器相关特性等。

常用more /proc/cpuinfo命令查看处理器信息，输出字段释义如下：

输出结果	结果释义
processor	表示逻辑处理器的唯一标识符
vendor_id	表示处理器类型，若为GenuineIntel则为英特尔处理器
physical id	表示物理处理器的唯一标识符
siblings	表示位于相同物理封装中的逻辑处理器的数量
core id	表示每个内核的唯一标识符
cpu cores	表示位于相同物理封装中的内核数量

如需查看系统物理处理器的个数，可通过如下命令查看：

```
cat /proc/cpuinfo | grep "physical id" | sort | uniq | wc -l
```

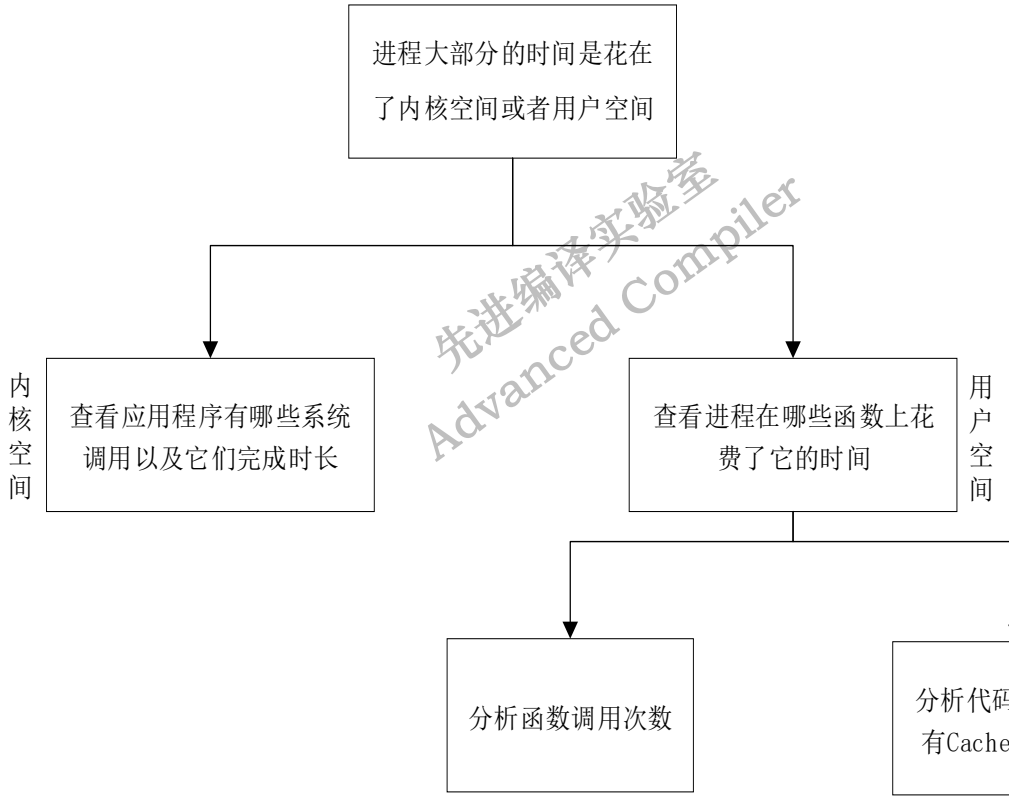
查看每个物理处理器中内核的个数，可通过如下命令查看：

```
cat /proc/cpuinfo | grep "cpu cores"
```

查看系统所有逻辑处理器个数，可通过如下命令查看：

```
cat /proc/cpuinfo | grep "processor" | wc -l
```


可以参考右图调查进程处理器使用情况的分析检查过程，通过查出某特定进程或应用程序所在处理器瓶颈，进而查明其消耗时间的位置和原因。



使用nice或renice命令设置nice值可以调整进程优先级，Linux支持的nice值范围为-20到19，nice值越大表示进程优先级越低，而nice值越小表示优先级越高，负nice值仅能由超级管理员设置。nice命令可以指定nice值并启动程序，renice命令用来调整已经在运行的进程。

renice -n 9 12345

查看进程号为12345的进程的调度策略：

chrt -p 12345

使用chrt命令还可以修改调度策略，例如超级管理员账户下可调整为先到先服务，使用参数-f表示，优先级为10：

chrt -p -f 10 1234

进程绑定是把一个或多个进程绑定到某个或多个处理器上，这样可以增加进程的处理器缓存，提高它的内存I/O性能。

在Linux上可以通过taskset命令实现进程绑定的，此方法可以使用处理器掩码或者范围设置处理器与进程的关联性。例如：

taskset -pc 7-10 10790

pid 10790's current affinity list： 0-15

pid 10790's new affinity list： 7-10

设置限定进程号为10790的进程只能运行在处理器7到处理器10上。

中断是一种来自硬件或者软件的信号之一，表明这里有项工作现在就要做。当一个中断信号达到操作系统内核的时候，内核必须从当前执行的进程切换到一个新的进程，以处理这个中断。

如需确定内核在哪个处理器上执行特定的中断处理程序，可以查看/proc/irq/number文件夹下的smp_affinity文件，该文件中的数据表示处理器位掩码，以十六进制数表示。当一个中断被允许在某处理器上处理，则将相应的比特位设置为1，否则设为0。

假如想设置中断50仅在处理器0、2、7上处理，则计算方法为：2^0 +2^2 +2^7 =133=0x85，将从shell计算得到的结果直接设置为affinity-mask命令如下：

printf '%0x' $[2**0+2**2+2**7]>/proc/irq/50/smp_affinity

关于多核技术，处理器0是很关键的，如果0号处理器使用过度，则别的处理器性能也会下降。这是因为处理器0具有调整功能，所以不能任由操作系统对其进行负载均衡，可以手动地为其分配处理器核，从而不会过多地占用0号处理器，或是让关键进程和一堆别的进程挤在一起。

传统的多核运算是使用对称多处理模式，多个处理器共享一个集中的存储器和I/O总线。于是就会出现一致存储器访问的问题，一致性通常意味着性能的损失。而在非一致内存访问模式下，处理器被划分成多个节点，每个节点有自己的本地存储器空间。如需只让进程访问当前运行节点，可使用如下命令：

```
numactl --membind 1 --cpunodebind 1 --localalloc myapplication
```

内存

内存是连接处理器和其它设备的通道，起到缓冲和数据的交换作用。内存包括物理内存和虚拟内存swap，内存资源的充足与否直接影响到系统性能，因此内存的管理和优化是系统性能优化的重要组成部分。本节仍从以下两个部分讨论：

（1）配置检查；

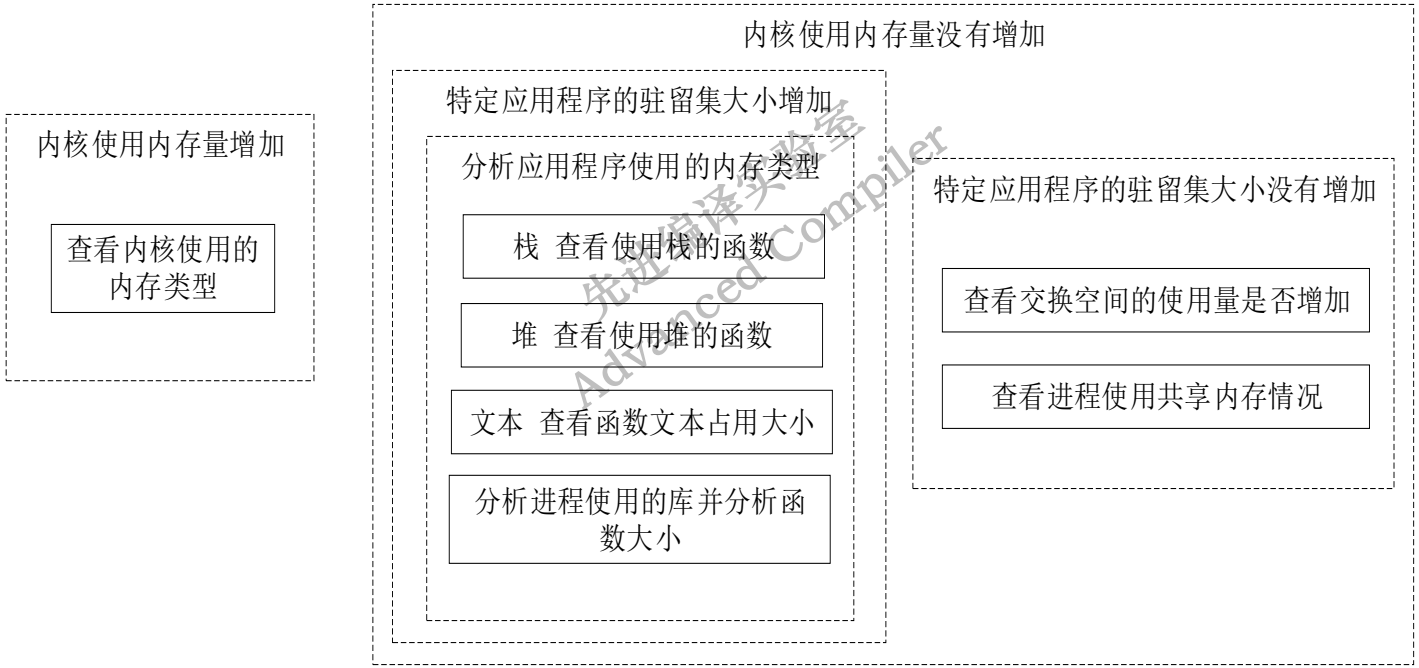
（2）参数调整；

查找内存可以优化的空间首先需要进行内存配置检查，可以从以下几个方面展开，如内存空间大小、硬件允许的最大内存数量、是否支持非统一内存访问、主存的访存速度、内存总线数量、是否使用了大页面、是否有其它内存可调参数等。

在Linux下查看内存信息可以使用/proc/meminfo文件查看操作系统内存的使用状态，使用命令cat /proc/meminfo即可。

查看内存硬件信息可以使用命令sudo dmidecode -t memory，查看现有的内存硬件信息可以使用命令dmidecode | grep -A16 "Memory Device"。

右图所示显示了在配置检查时分析内存数据的策略。当系统的内存使用量快速增加时，需要分析清楚是何原因造成了系统内存需求的增长。若特定进程的驻留集大小在增加，可以追踪是哪个进程该为内存使用量的增加负责。



分页是将程序分配到磁盘的过程，页面空间是操作系统在磁盘分区上创建的文件，用于存储当前未使用的用户程序。在Linux中，页面大小通常为4KB或8KB，页面大小是通过/usr/src/kernels/3.10.0-1160.el7.x86_64/arch/x86/include/asm/elf.h目录下内核头文件中的变量EXEC_P AGESIZE定义的。

在Linux中有许多设置大页面的方法，通常用于创建巨页面。

#echo 50 > /proc/sys/vm/nr_hugepages（//设置页面大小）

#grep Huge /proc/meminfo（//查看已经设置的页面大小）

当物理内存完全被使用或系统需要额外内存时，此时需要使用虚拟内存swap设备。当系统上没有空闲内存可用的时候，操作系统开始将内存中最少被使用的数据分页调度到磁盘的swap区域。

改进活跃和非活跃内存的处理。当内核想释放内存中的一个分页时，它需要在两种选择之间作出权衡：一种是从进程的内存中换出一个分页，另一种则是它能从分页Cache中丢弃一个分页。为了做出这个决定，内核将执行下面的计算：

swap_tendency = mapped_ratio/2 + distress + vm_swapiness

如果swap_tendency低于100，内核将从分页中回收一个分页；如果大于等于100，一个进程内存空间的一部分将有资格获得交换。

当虚拟机运行完全相同的操作系统或工作量的时候，一些内存分页将大概率有相同的内容。使用内核同页合并（Kernel Samepage Merging,KSM）功能可使总共的内存使用减少，因为它可以将那些完全相同的分页合并到一个内存分页中。

KSM将使用两个服务：ksm服务实际地扫描内存和合并分页，ksmtuned服务控制ksm是否扫描内存及如何积极扫描内存。使用ksmtuned服务手段调整ksm通常更为有用，配置ksmtuned服务可以使用/etc/ksmtuned.conf文件。

基础的资源控制包括设置主存限制和虚拟内存限制，可以用ulimit命令实现。Linux中，控制组cgroup的内存子系统可提供多种附加控制，如下：

Memory.memsw.limit_in_bytes：允许的最大内存和交换空间，单位是字节；

Memory.limit_in_bytes：允许的最大用户内存，包括文件缓存，单位是字节；

Memory.swappiness：类似之前描述的vm.swappiness，差别是可以设置于cgroup；

Memory.oom_contro1：设置为0，允许内存溢出终结者运用于这个cgroup，或者设置为1，禁用。

AdvancedCompiler
Tel: 13839830713

编译论坛

先进编译实验室
Advanced Compiler

系统配置优化II

嘉宾：柴晓楠

先进编译实验室
Advanced Compiler

文件系统

文件系统是操作系统与磁盘设备之间交互的一个桥梁，对磁盘的任何写操作都要经过文件系统然后才到磁盘。文件系统除了在磁盘上存储和管理数据，还负责保证数据的完整性。本节从以下两个部分讨论：

（1）配置检查；

（2）参数调整；

文件系统性能调优时，需要检查其静态配置情况：如挂载的文件系统数量、文件系统记录大小、是否启用访问时间戳、是否配置文件系统缓存、使用的文件系统版本是什么、文件系统是否有补丁等。

可以使用提到的vmstat、sar工具进行查阅、配置。

Linux系统可以使用SystemTap动态跟踪文件系统事件，其它一些用于调查文件系统性能以及刻画使用情况的工具和监控框架。

文件系统也是有缓存的，为了让文件系统有最大的性能，首要的事情就是分配足够大的内存，这个非常关键。

Linux上的ext2、ext3、ext4文件系统调整工具为tune2fs，而在挂载时指定多种选项有两种方法：一种是手动的mount命令，另一种是启动时的/boot/grub/menu.lst和/etc/fstab。

参数调整分为两部分：

（1）文件描述符调优；

（2）内核参数调优；

文件描述符可以说是服务器程序的宝贵资源，大部分系统调用都是和文件描述符打交道的。而系统分配给应用程序的文件描述符有限，所以关闭那些不再使用的文件描述符以释放其所占的资源有利于提高性能。

通过ulimit -n命令可以查看用户级的文件描述符数；

通过ulimit -SHn max-file-number可以将用户级文件描述符限制设定为max-file-number；不过这种设置是临时的，只在当前的会话中有效。例： ulimit -SHn 512

要想永久修改用户级文件描述符数限制，可以在/etc/security/limits.conf文件中加入如下两项：

hard nofile max-file-number

soft nofile max-file-number

几乎所有的操作系统内核模块，包括内核核心模块和驱动程序都在/proc/sys文件系统下提供了某些配置文件，以供用户调整模块的属性和行为。通常一个配置文件对应一个内核参数，文件名就是参数的名字，文件的内容是参数的值。可以通过sysctl -a查看所有这些内核参数。

其中最重要的两个参数如下：

/proc/sys/fs/file-max，系统级文件描述符限制。直接修改这个参数与上述最大文件描述符中讨论的修改方法有相同的效果；

/proc/sys/fs/epoll/max_user_watches中的值是一个用户能够往epoll内核事件表中注册的事件的总量，epoll是Linux中I/O多路复用的一种机制，该值是指该用户打开的所有epoll实例总共监听的事件数目。

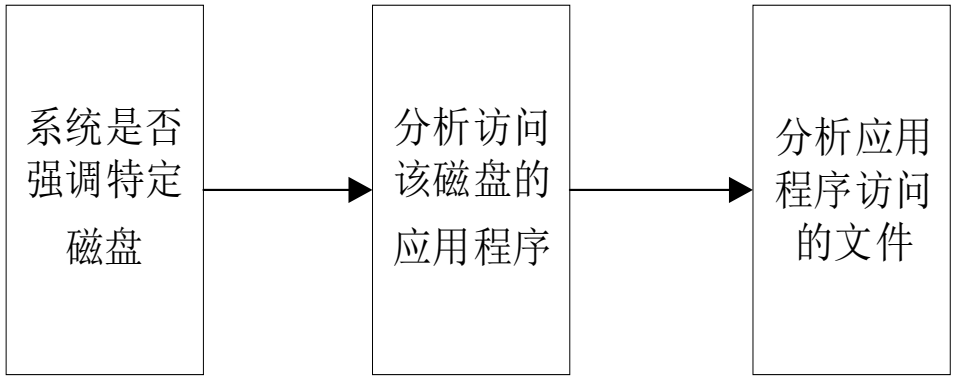
磁盘是速度较慢的存储子系统，通常会成为影响整个系统性能的瓶颈。若在高负载下磁盘成为了瓶颈，处理器会持续空闲以等待磁盘I/O结束。此时，及时发现并消除这些瓶颈很有可能会提升数倍的性能。本节仍分为以下两个部分：

（1）配置检查；

（2）参数调整；

查看系统硬盘信息和使用情况可以在超级管理员权限下使用命令fdisk或者命令df，也可以直接使用命令cat /proc/partitions进行查看。也可以利用iostat、sar命令评估磁盘性能。此外也可以通过top命令确定系统是否受I/O限制。

当确定是磁盘I/O问题时，需要分析出是哪个应用程序引起了I/O磁盘问题，下图给出了确定磁盘I/O使用原因的步骤。



linux中的ionice工具可以限制一个特定进程的磁盘子系统使用率，ionice将磁盘I/O调度分为三类，实时、尽力、空闲。

一个ionice的例子，将进程号为1623的进程放入了空闲I/O调度级别：

ionice –c 3 –p 1623

上例适用于对读写被长时间允许的备份任务，以尽量避免与生产负载产生冲突。

通常有3个调度器可供选择：

（1）完全公平调度器（Completely Fair Scheduler,CFQ），它将由进程提交的同步请求放到多个进程队列中，然后为每个队列分配时间片以访问磁盘，通常是Linux系统的默认调度器；

（2）noop是Linux内核里最简单的I/O调度器，基于先进先出队列算法；

（3）deadline又称截止时间调度器，它尝试保证请求的开始服务时间。适当的更改调度器有助于系统发挥最佳性能。

操作系统选择I/O调度器策略的可调参数为/sys/block/sda/queue/scheduler

Linux操作系统的服务器中同样会遇到网络问题，随着计算节点规模的扩大，网络在性能方面扮演着越来越重要的角色，因网络而生的糟糕性能常常备受指责。本节从配置检查和参数调整两方面展开讨论，分析并对网络性能进行优化。

网络性能调优之前需要检查的配置，如网络可使用接口数量、当前接口使用情况、接口的速度、是否使用链路聚合、设置驱动的参数、域名系统是否设置、是否有已知的可能性能问题等。

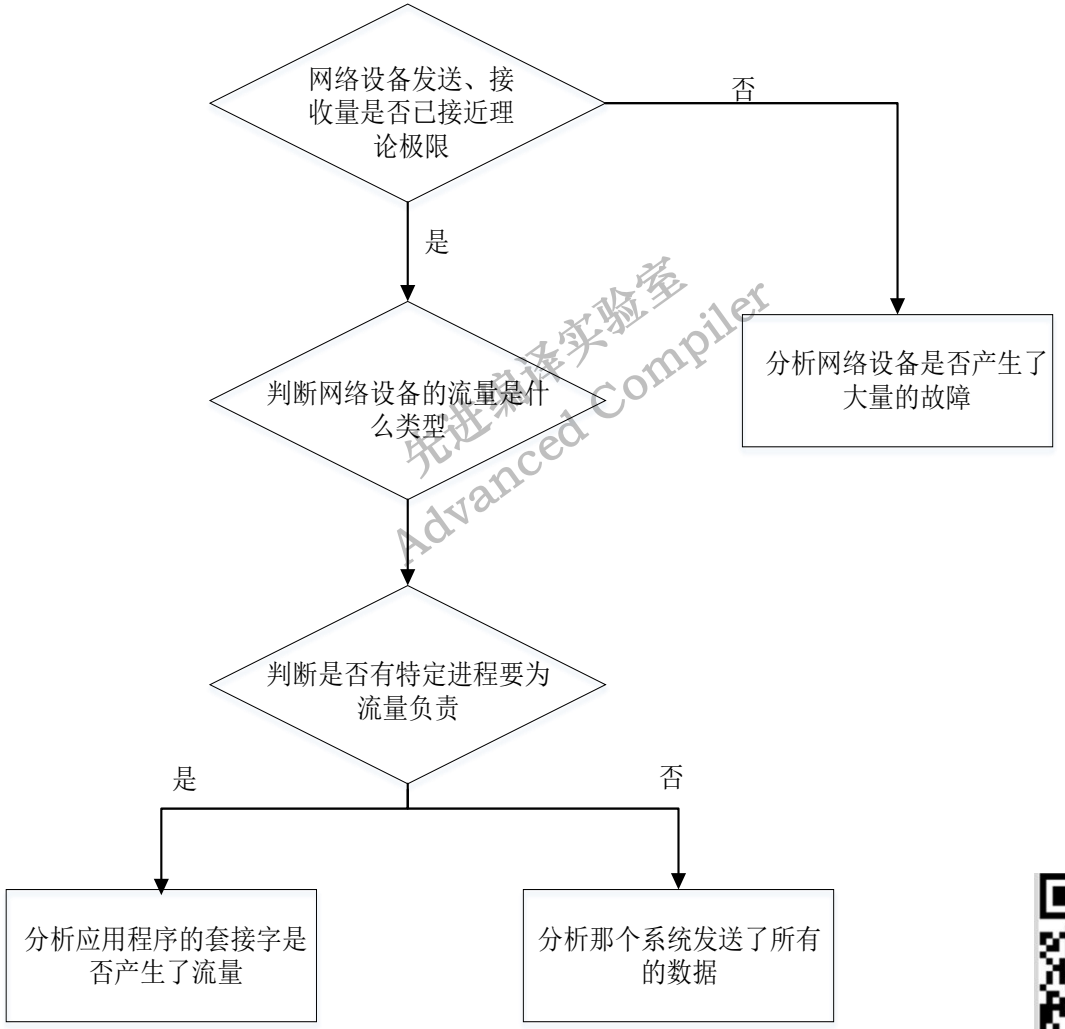
在Linux中，可以借助ifconfig 查看网络接口信息，也可以使用netstat、sar命令查看系统网络信息。例：

netstat -rn；//查看网关地址

netstat –r//检测系统的路由表信息

netstat -i //查看网络接口状态信息，输出结果中，RX-ERR/TX-ERR、RX-DRP/TX-DRP和RX-OVR/TX-OVR的值都应该为0

当知道网络发生了问题时，Linux提供了一组工具来确定哪些应用程序涉及其中。右图展示了调查分析网络性能问题的步骤。



如果认为网络子系统存在瓶颈，就要对网络性能采取优化措施，因为网络问题可能会对其它子系统带来影响。比如，当数据包太小的时候，CPU使用率会受到明显的影响；如果有过多数量的TCP连接，内存使用会增加。为了提高网络性能可以对网络参数进行调整，本节介绍以下两种调整方法：

（1）服务器相关参数调优；

（2）TCP参数调优；

可调参数可用sysctl命令查看和设置，并写到/etc/sysctl.conf。内核中网络模块的相关参数都位于/proc/sys/net目录下，其中和TCP/IP协议相关的参数主要位于以下三个子目录中：core、ipv4和ipv6。可以调整其中的/proc/sys/net/core/somaxconn、/proc/sys/net/ipv4/tcp_wmem等相关参数。

/proc/sys/net/ipv4/tcp_rmem，它包含3个值，分别指定一个socket读缓冲区的最小值、默认值和最大值。

/proc/sys/net/ipv4/tcp_wmem，它包含3个值，分别指定一个socket的TCP写缓冲区的最小值、默认值和最大值。

在net目录下有很多参数，包括IP、Ethernet、路由和网络接口参数，一般常调试以下几个参数：

（1）套接字和TCP缓冲；
（2）TCP积压队列；
（3）设备积压队列；
（4）TCP阻塞控制；
（5）TCP选项；
（6）网络接口；

AdvancedCompiler
Tel: 13839830713