tuned服务的作用：通过udev来监视硬件设备，然后根据监视所获取的数据来对系统进行动态调优或直接静态调优。

tuned服务使用两类程序：监视程序和调优程序。

1)监视程序：主要负责对系统的硬件设备进行监视，详细信息如下：

序号	监视范围	描述
1	disk	每间隔一定时间获取系统每个磁盘的负载（IO操作的数量）
2	net	每间隔一定时间获取系统每个网卡的网络负载（传输数据包的数量）
3	load	每间隔一定时间获取系统每个CPU的负载（运行时间）

注：默认的间隔时间为10秒，可以通过文件/etc/tuned/tuned-main.conf里的参数update_interval来调整。

2)调优程序：根据监视程序所获得的数据进行动态调优或根据系统指定的profile来对系统进行静态调优。关于profile的详细信息如下：

番号	Tunable	default	network-latency	network-throughput	latency-performance	powersave	throughput-performance	virtual-guest	virtual-host
1	governor※1	ondemand	performance	performance	performance	ondemand	performance	performance	performance
2	energy_perf_bias	normal	performance	performance	performance	powersave	performance	performance	performance
3	force_latency※1	109	1	–	1	–	–	–	–
4	min_perf_pct※1	48	100	100	100	–	100	100	100
5	kernel.sched_autogroup_enabled	0	–	–	–	–	–	–	–
6	kernel.sched_min_granularity_ns※2	3000000	10000000	10000000	10000000	–	10000000	10000000	10000000
7	vm.dirty_ratio	20	10	40	10	–	40	30	40
8	vm.dirty_background_ratio	10	3	10	3	–	10	10	5
9	vm.swappiness	60	10	10	10	–	10	30	10
10	kernel.sched_migration_cost_ns	500000	5000000	–	5000000	–	–	–	5000000
11	vm.laptop_mode	5	–	–	–	5	–	–	–
12	vm.dirty_writeback_centisecs	1500	–	–	–	1500	–	–	–
13	kernel.nmi_watchdog	0	–	–	–	0	–	–	–
14	vm.max_map_count	250 32000 32 128	–	–	–	–	–	–	–
15	kernel.sched_wakeup_granularity_ns※2	65530	–	–	–	–	–	–	–
16	transparent_hugepages	4000000	–	15000000	–	–	15000000	15000000	15000000
17	alpm	always	never	always	never	–	–	–	–
18	readahead	min_power	–	–	–	min_power	–	–	–
19	net.core.busy_read	128	–	4096	–	–	4096	4096	4096
20	net.core.busy_poll	0	50	–	–	–	–	–	–
21	net.ipv4.tcp_fastopen	0	50	–	–	–	–	–	–
22	kernel.numa_balancing※1	0	3	–	–	–	–	–	–
23	net.ipv4.tcp_rmem※3	1	0	–	–	–	–	–	–
24	net.ipv4.tcp_wmem※3	4096 87380 6291450	–	4096 87380 16777216	–	–	–	–	–
25	net.ipv4.udp_mem※3	4096 16384 4194304	–	4096 16384 16777216	–	–	–	–	–
26	net.ipv4.udp_mem※3	767112 1022816 15342	–	3145728 4194304 16777216	–	–	–	–	–

※1具体的值与CPU类型有关，　※2具体的值与CPU个数有关，　※3具体的值Memory大小有关。

注：

1）Native与Host系统下默认的profile为powersave。

2）Guest系统下默认的profile为virtual-guest。

3）系统默认关闭动态调优，开启动态调优可以通过文件/etc/tuned/tuned-main.conf里的参数dynamic_tuning来设置。

4）profile所设置的参数值若不存在，系统会使用默认值。例如系统的CPU参数governor没有ondemand值，如果设置了ondemand，会采用系统默认值powersave。

tuned服务使用方法如下：

序号	命令	描述
1	systemctl start tuned	启动tuned服务
2	tuned-adm list	查看可以使用的profiles
3	tuned-adm active	查看当前正在使用的profile
4	tundd-adm profile -	激活某个指定的profile
5	tuned-adm recommand	让系统推荐一个最适合的profile

tuned可调节参数的含义：

序号	参数	取值范围	设置方法	描述
1	governor	performance 、 powersave	echo [performance|powersave] > /sys/devices/system/cpu/cpuN/cpufreq/scaling_governor	CPU频率调速器
2	energy_perf_bias	performance 、normal、 powersave	x86_energy_perf_policy [performance | normal | powersave]	CPU在权衡performance和 energy efficiencyman的模式，详细信息可参考"man x86_energy_perf_policy"。
3	force_latency	>0	echo N >/dev/cpu_dma_latency	CPU从idle状态切换C0状态的最大唤醒时间（ms）
4	min_perf_pct	0 - 100	echo N > /sys/devices/system/cpu/intel_pstate/min_perf_pct	设置P-state的最小百分比（相对于cpufreq）
5	kernel.sched_autogroup_enabled	0、1	使用命令sysctl -p 参数 = ***	是否启动进程自动分组调度
6	kernel.sched_min_granularity_ns	>0		进程被调度前最少运行时间（ns）
7	vm.dirty_ratio	>0		设置脏页数据占系统内存的比例，阻塞式启动pdflush内核线程
8	vm.dirty_background_ratio	>0		设置脏页数据占系统内存的比例，非阻塞式启动pdflush内核线程
9	vm.swappiness	0 - 100		积极使用swap空间的比例，例如0表示尽量不使用swap设备，100表示尽可能使用swap设备。
10	kernel.sched_migration_cost_ns	>0		利用该值（ns）来判断一个进程是否是"cache hot"。如果是的话，就尽可能不对这个进程进行迁移。
11	vm.laptop_mode	>=0		将所有磁盘I/O操作、脏缓存写到磁盘的时间间隔（s）
12	vm.dirty_writeback_centisecs	>0		设置脏页数据在内存中的最大驻留时间，超过此值，pdflush内核线程将会将这些脏数据写入磁盘
13	kernel.nmi_watchdog	0、1		是否启动watchdog，watchdog用于检测系统是否hang。
14	kernel.sem	-		设置信号量的相关参数值
15	vm.max_map_count	>0		限制一个进程可以拥有的VMA(虚拟内存区域)的数量。
16	kernel.sched_wakeup_granularity_ns	>0		表示进程被唤醒后至少应该运行的时间(ns)
17	transparent_hugepages	always、 madvise、 never	echo [always | madvise | never] > /sys/kernel/mm/transparent_hugepage/enabled	是否启用透明大页面
18	alpm	min_power、medium_power、max_performance	echo [min_power|medium_power| max_performance] >/sys/class/scsi_host/host1/link_power_management_policy	针对磁盘（SATA控制器）在IO空闲状态的模式
19	readahead	>0	echo-个值到文件文件/sys/block/sd*/queue/read_ahead_kb	预取数据加载到内存(Kb),该参数可通过文件/sys/block/sd*/queue/read_ahead_kb来设置
20	net.core.busy_read	>=0	sysctl -p parameter = N	设置自旋（spin）等待从设备队列读取socket数据的时间（us）
21	net.core.busy_poll	>=0		设置自旋（spin）等待从设备队列的socket poll与select的时间（us）
22	net.ipv4.tcp_fastopen	0、1、2、3		是否开启快速打开TCP。详细参考：https://lwn.net/Articles/508865/
23	kernel.numa_balancing	0、1		是否启动自动numa balancing。启动之后系统会自动移动任务或数据更加接近内存，缩短访问时间。
24	net.ipv4.tcp_rmem	-		为TCP socket预留用于接收缓冲的内存大小（字节）
25	net.ipv4.tcp_wmem	-		为TCP socket预留用于发送缓冲的内存大小（字节）
26	net.ipv4.udp_mem	-		为UDP socket预留用于发送缓冲的内存大小（字节）

各参数的详细解析：

1、

参数	默认值	取值范围	描述
governor	ondemand	performance 、 ondemand 、powersave	CPU频率调速器

说明：

1）该参数用于设置CPU的频率模式，有3个值可供设置。

2）当参数值为performance时，CPU的频率将会一直维持在最高主频（/proc/cpuinfo中显示的主频），电力消耗增加，对某些应用程序来说性能会有提升。

3）当参数值为ondemand时，CPU的频率将会根据CPU的利用率来变化，电力消耗较少，对应用程序性能影响也较小。
注：机器RX300S7没有ondemand模式可供选择。

4）当参数值为powersave时，CPU的频率在CPU空闲时将处于最低值（由参数min_perf_pct控制），电力消耗最少，对某些应用程序性能有影响。

2、

参数	默认值	取值范围	描述
energy_perf_bias	normal	performance 、normal、 powersave	CPU权衡performance和energy efficiency的模式

说明：

1）该参数表示CPU在性能和电力节约之间做出选择，有3个值可供设置。

2）当参数值为performance时，CPU不会了节省电力牺牲一点性能，这种情况下，电力消耗增加，程序性能会有提升。

3）当参数值为normal时，CPU会在电力消耗和性能之间做一个折中，这也是默认的模式。

4）当参数值为powersave时，CPU将会最大化的节省电力，这种情况下，程序性能会受到影响。

5）可动态调节每个CPU的模式，使用命令“x86_energy_perf_policy”即可调整。

3、

参数	默认值	取值范围	单位	描述
force_latency	109	[0,109]	ms	CPU从idle状态切换C0状态的最大唤醒时间

说明：

1）该参数表示CPU从idle状态切换到C0状态（运行状态）的最大唤醒时间。

2）调大该参数的值，使得CPU能够进入深度睡眠，减少CPU对电力的消耗。但是，由于CPU的唤醒时间增加，可能对某些应用程序的性能有影响。

3）调小该参数的值，能够减少CPU从idle状态切换到C0状态的延迟，对某些应用程序来说（如网络程序）能够能够减少延时。但是由于CPU不能进入深度睡眠，电力消耗增加。

4、

参数	默认值	取值范围	描述
min_perf_pct	48	[1,100]	设置CPU最低的频率

说明：

1）该参数可用于设置CPU的最低频率，参数值为CPU频率的百分比。

2）调大该参数值，将提高CPU在空闲时的最低频率，不利于节电。

3）调小改参数值，将降低CPU在空闲时的最低频率，利于节电。

4）该参数可以通过文件 /sys/devices/system/cpu/intel_pstate/min_perf_pct 来设置。

5、

参数	默认值	取值范围	描述
kernel.sched_autogroup_enabled	0	0、1	是否启动进程自动分组调度特性

说明：

1）该特性依据进程的类型，将不同的进程放到不同的组内，进程调度单位是组。这样启动该特性后低响应的进程（比如编译内核）就不会影响高响应的进程（交互性强的进程）。

2）该特性主要用于Desktop环境。

3）该特性不适用于Server环境下，因为该特性可能会导致一些daemon的子进程不停的进行移植，影响性能。

6、

参数	默认值	取值范围	单位	描述
kernel.sched_min_granularity_ns	3,000,000	[1,2^32-1]	ns	进程被调度前最少运行的时间

说明：

1）该参数表示多久内核会检查调度另外一个进程，也就是表示被调度前进程最少运行时间。

2）调大该参数值，会使得进程被频繁的切换，对于交互系统，可以保证交互得到更快的响应。

3）调小该参数值，会减少进程被频繁的切换，即减少了上下文切换，CPU利用率将提高，对某些应用程序来说这将提高性能。

7、

参数	默认值	取值范围	描述
vm.dirty_ratio	20	[0,100]	当脏页数据占系统内存达到一定比例时，阻塞式将脏页回写到磁盘

说明：

1）应用程序在向page cache写数据的过程中,系统会首先检查脏页占内存的百分比是否达到了dirty_ratio阈值,如果达到，应用程序则阻塞等待直到将脏页写回磁盘。

2）调大该参数值，应用程序到达dirty_ratio的次数会减少，使得应用程序在调用write函数时等待page cache的回写时间减少，因此会缩短程序的运行时间。
由于内核保证了dirty page的量不会超过内存总量的50%，所以dirty_ratio大于50时程序运行时间基本和dirty_ratio=50时程序运行时间相同。

3）调小该参数值，应用程序到达dirty_ratio的次数会增多，导致回写次数增多,使得应用程序调用write函数时等待page cache的回写时间变长,因此会增长程序的运行时间。

8、

参数	默认值	取值范围	描述
vm.dirty_background_ratio	10	[0,100]	当脏页数据占系统内存达到一定比例时，启动内核线程pdflush，非阻塞式将脏页回写到磁盘

说明：

1）应用程序在向page cache写数据的过程中，系统会检查脏页占内存的百分比是否达到dirty_background_ratio阈值,如果达到，系统会启动pdflush内核线程回写脏页
直到脏页占内存的比例小于dirty_background_ratio或回写了指定脏页数，而应用程序继续写数据。

2）调大该参数值，在这种情况下（应用程序每次写之间存在一定间隔）会使得被回写的脏页量减少，从而导致脏页维持在dirty_ratio的时间段变长，因此应用程序将会
因为等待时间变长而使整个运行时间变长。

3）调小参数dirty_background_ratio值，在这种情况下（应用程序每次写之间存在一定间隔）会使得被回写的脏页量增多，从而导致脏页维持在dirty_ratio的时间段变短,
因此应用程序将会因为等待时间变短而使整个运行时间变短。

9、

参数	默认值	取值范围	描述
vm.swappiness	60	[0,100]	调整用户态地址空间的页的回收策略

说明：

1）当系统内存紧张时，系统有可能会回收两种类型的页框来获得内存：一种是用于存放进程用户空间页；另一种是供I/O使用的page cache。

2）用户调整swappiness参数值，可以影响用户态地址空间的页的回收。如下所示：

swappiness	userspace	pagecache
值小(0)	不回收	回收
值大(100)	回收	回收

3）调大该参数的值，两种页都会被回收，回收的page cache就会减少，IO用程序性能可能会得到提升。

4）调小该参数的值，不回收用户态地址空间的页，用户空间应用程序性能可能会得到提升。

10、

参数	默认值	取值范围	单位	描述
kernel.sched_migration_cost_ns	500,000	[1,2^32-1]	ns	用于判断一个进程是否处于“cache hot”状态

说明：

1）该参数用于判断一个进程是否处于“hot”状态。在系统需要对进程进行移植的时候，如果该进程距离上一次运行的时间间隔小于该参数值，则判定该进程处于“hot”状态，那么系统将尽量不移植该进程，否则系统将会对该进程进行移植。

2）调大该参数的值，能够减少进程进行移植的操作，特别是对于某些进程在CPUs或nodes来回切换的这种情况下，调大参数能够提高应用程序的性能。

3）调小改参数的值，能够增加进行进行移植的操作，在CPU空闲时间比较多的情况下，调小该参数可以CPU资源更加合理的被运用。

11、

参数	默认值	取值范围	单位	描述
vm.laptop_mode	5	[0,INT_MAX]	s	用于设置系统是否启用laptop_mode模式

说明：

1）laptop_mode模式是一种特殊的页回写策略，该策略主要意图是将硬盘转动的机器化行为最小化，尽量使硬盘处于低能耗的状态下，节省电力。

2）laptop_mode模式周期性的启动pdflush线程将许多的I/O操作组织在一起，一次完成，这样可以减少磁盘启动的次数。

3）laptop_mode模式需要与参数vm.dirty_writeback_centisecs和dirty_expire_centisecs配合使用来达到节省电力的目的。

12、

参数	默认值	取值范围	单位	描述
vm.dirty_writeback_centisecs	1500	[0, INT_MAX]	1/100 s	系统触发pdflush内核线程的周期

说明：

1）系统会周期性地触发pdflush线程，将系统中标记为脏时间过长（由参数dirty_expire_centisecs来判定）的脏页回写到磁盘。

2）调大该参数的值，内核线程pdflush被触发周期变长，系统因回写脏页而占用的io资源变少，在这种情况下（系统中存在多个文件被不断更新产生脏页），其它不通过内存cache使用IO资源的程序可用IO资源量变多，性能提高。

3）调小该参数的值，pdflush被触发周期变短，系统因回写脏页而占用的io资源变多，在这种情况下（系统中存在多个文件被不断更新产生脏页），其它不通过内存cache使用IO资源的程序可用IO资源量变少，性能降低。

4）参数为0时，pdflush不会被周期性触发，不会因回写而占用io资源，在这种情况下（系统中存在多个文件被不断更新产生脏页），其它不通过内存cache使用IO资源的程序可用IO资源最多，性能最优。

13、

参数	默认值	取值范围	描述

kernel.nmi_watchdog	0	0、1	用于设置系统是否启动nmi_watchdog特性

说明：

1）nmi_watchdog(Non Maskable Interrupt Watchdog) 通过周期性的向系统发送不可屏蔽的中断来检测系统是否hang。

2）启动该特性可以使内核有效的检测到CPU是否被锁住，并及时作出一些措施使得CPU恢复正常运行状态。

3）关闭该特性可以减少CPU对watchdog发送的不可屏蔽的中断进行处理，从而可以提供应用程序的性能。

14、

参数	默认值	取值范围	描述
kernel.sem	250 32000 32 128	-	系统关于信号量的一些限制

说明：

1）该参数由4个部分组成：

SEMMSL:控制每个信号集可以包括最多的信号数量
SEMMNS:控制系统最多可以拥有的信号数量
SEMOPM:控制系统调用semop一次最多可以操作的信号数量
SEMMNI:控制系统最多可以同游的信号集数量

2）在应用程序工作过程中需要大量信号量的时候（如oracle数据库），需要调大该参数，否则应用程序可能因内核限制的信号量数导致应用程序无法正常工作。

15、

参数	默认值	取值范围	描述
vm.max_map_count	65530	[1, INT_MAX]	一个进程最多可以拥有的VMA(虚拟内存区域)的数量

说明：

1）该参数用来限制一个进程最多可以拥有vma的数量。

2）像malloc、mmap、mprotect以及加载共享库这样的操作都会影响vma的数量。

3）增加该参数可以避免某些程序因大量增加vma数量到达限制而产生错误。

16、

参数	默认值	取值范围	单位	描述
kernel.sched_wakeup_granularity_ns	4,000,000	-	ns	被wake-up的进程进行抢占的粗粒度

说明：

1）它用来判断被wake-up的进程是否抢占当前正在运行的进程，该参数越小，抢占发生的概率越高，该参数越大，抢占发生的概率越小。

2）调大该参数值，可以减少进程抢占的发生概率，也就减少了进程上下文的切换所带来的资源消耗，因此可以提高某些应用程序的throughput值。

3）调小改参数值，可以增大进程抢占的发生概率，对于某些交互性强的应用程序来说，可以减少Latency。

17、

参数		默认值	取值范围	描述
	transparent_hugepages	always	always、 madvise、 never	系统是否启动透明大页面

说明：

1）使用大页面可以减少应用程序TLB miss的发生，因此可以提高应用程序的性能。

2）透明大页面不需要应用程序做任何修改或设置，系统会自动为应用程序使用透明大页面。

注：透明大页面只能适用于匿名映射的内存区域。

3）当参数设为[always]时，系统会尽可能的为应用程序使用大页面。

4）当参数设为[madvise]时,系统只会为应用程序的内存区域标有MAD_HUGEPAGE的内存使用大页面。

5）当参数设为[never]时，系统不会为任何应用程序使用大页面。

18、

参数	默认值	取值范围	描述
alpm	min_power	min_power、medium_power、max_performance	I/O为idle状态下磁盘的省电模式

说明：

1）alpm（aggressive link power management）是一个power-saving技术，在没有I/O操作时，系统通过一些设置来降低disk的电力消耗从而达到省电。

2）alpm技术只适用于采用高级主机控制接口（Advanced Host Controller Interface）的SATA控制器。

3）alpm的三种模式：

a）min_power:这种模式最省电，适用于I/O操作长时间处于idle状态。

b）medium_power:这种模式较省电，适用于一会有连续繁重的I/O操作，一会长时间的处于idle I/O状态。

c）max_performance:禁用alpm技术，那么即使磁盘没有I/O操作，也不会进入省电模式。

4）设置alpm为min_power或max_performance模式，将会自动使“Hot Plug”特性失效。

19、

参数	默认值	取值范围	单位	描述
readahea	128	[1, LONG_MAX]	KB	设置预读取数据到内存的大小(KB)

说明：

1）当系统需要读取某个文件时，无论实际需要多少，默认一次会读取128KB的数据。

2）当顺序读大文件时，提高该参数值，一次可以多读取点数据，这样可以有效的减少读seek的次数，从而提高性能。

3）该参数可以通过命令"blockdev --setra /dev/sd*"或通过文件/sys/block/sd*/queue/read_ahead_kb来设置。

20、

参数	默认值	取值范围	单位	描述
net.core.busy_read	0	[0,2^32-1]	ms	设置自旋（spin）等待从设备队列读取socket数据的时间

说明：

1）Busy polling特性会使socket底层代码poll网络设备的接收队列，这样可以减少了网络中断和进程上下文切换，增加CPU利用率，
但是由于CPU不会进行sleep从而增加电力消耗。

2）该参数用于设置接收网络数据时poll的最长近似时间。

3）调大该参数的值，能够降低应用程序的Latency，增加电力消耗。

21、

参数	默认值	取值范围	单位	描述
net.core.busy_poll	0	[0,2^32-1]	ms	设置自旋（spin）等待从设备队列的socket poll与select的时间

说明：

1）Busy polling特性会使socket底层代码poll网络设备的接收队列，这样可以减少了网络中断和进程上下文切换，增加CPU利用率，
但是由于CPU不会进行sleep从而增加电力消耗。

2）该参数表示系统调用select()与poll()所监视的socket文件（需要打开SO_BUSY_POLL选项）没有发生任何事件时，监视这个socket文件的最长近似时间。

3）调大该参数的值，能够降低应用程序的Latency，调大该参数值，将要增加电力消耗。

22、

参数	默认值	取值范围	描述
net.ipv4.tcp_fastopen	0	0、1、2、3	是否开启快速打开TCP连接

说明：

1）TCP Fast Open（TFO）会利用TCP三次握手的SYN报文来传输应用数据，这样客户端与服务器端的交互过程中就减少一个RTT（(Round-Trip Time)的开销。

2）TCP三次握手是页面延迟时间的重要组成部分，因此启用TFO可以减少客户端加载页面的时间。

3）各个参数值的含义：
0：禁止TFO特性
1：客户端启用TFO特性
2：服务器端启用TFO特性
3：客户端和服务器端都启用TFO特性

23、

参数	默认值	取值范围	描述
kernel.numa_balancing	1	0、1	是否开启Automatic NUMA Balancing特性

说明：

1）CPU访问同一Node上的Memory要比访问其它Node上的Memory速度快，因此运行应用程序的CPU和要访问的Memory一直处于同一Node上的话，应用程序性能将变好。

2)开启 Automatic NUMA Balancing这个特性，系统会在应用程序运行时自动做些操作使其运行的CPU和访问的Memory处于同一Node，从而提高应用程序性能。

3）Automatic NUMA Balancing可以使用如下方法来达到其特性：
a)Migrate-on-Fault (MoF) - moves memory to where the program using it runs
b)task_numa_placement - moves running programs closer to their memory

4）在NUMA架构的系统上支持Automatic NUMA Balancing特性需要满如下两个条件：
a)使用命令# numactl --hardware 能够看到多个nodes。
b)使用命令#cat /sys/kernel/debug/sched_features 能够看到NUMA标记。