57 | 套路篇：Linux 性能工具速查

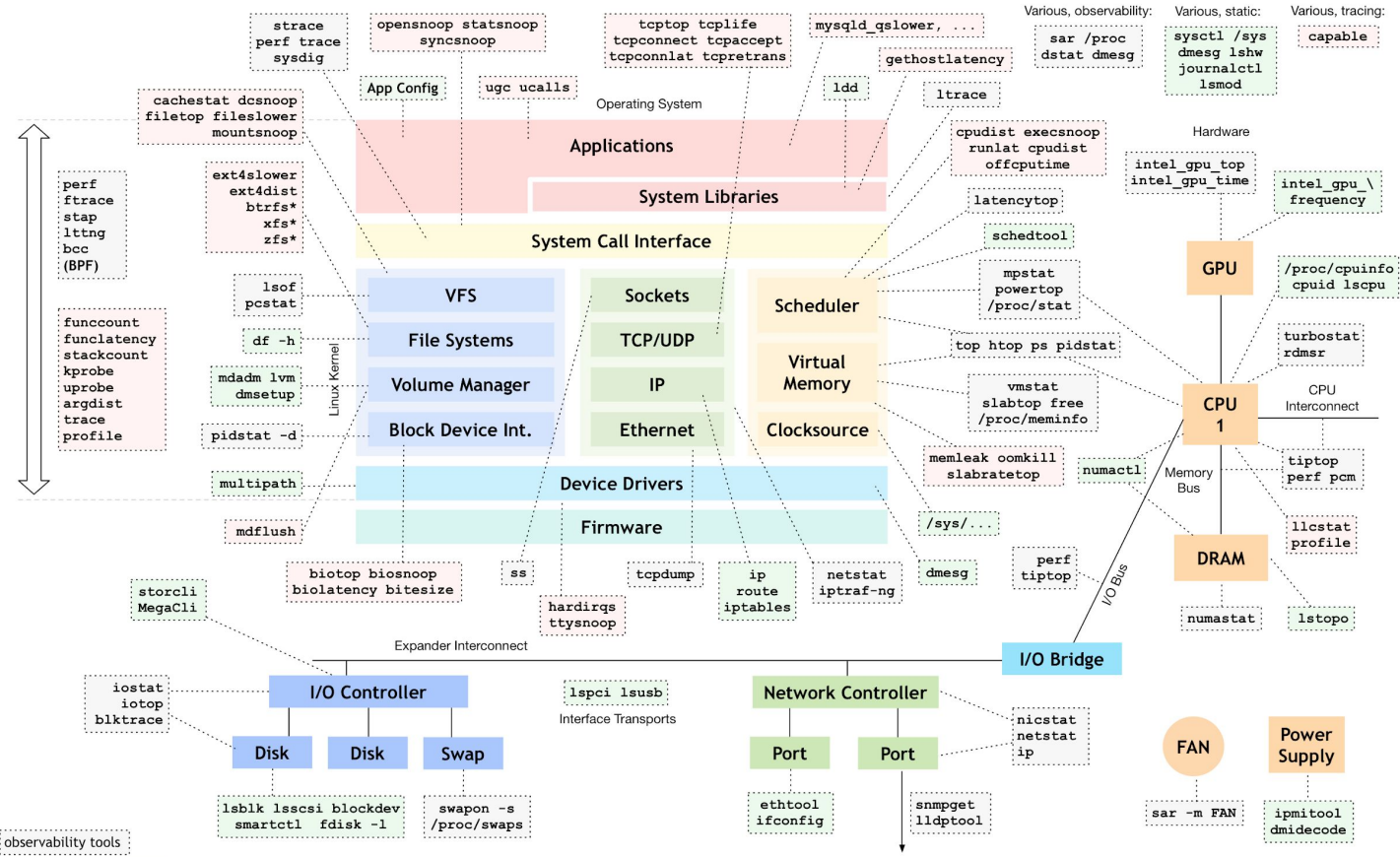
2020-12-24 Jeri

性能優化

- **系統程序**
  - 主要是對CPU、內存、網絡、磁盤I/O 以及內核軟件資源等進行優化。

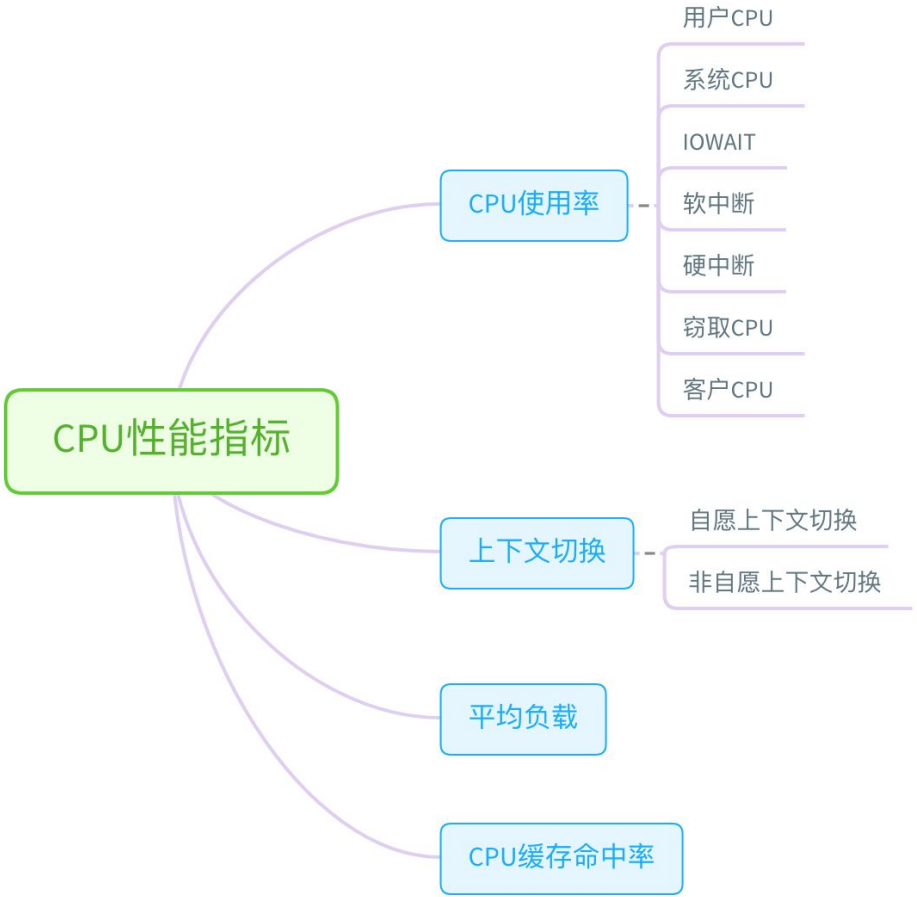
- **應用程序**
  - 主要是簡化代碼、降低 CPU 使用、減少網絡請求和磁盤 I/O，並藉助緩存、異步處理、多進程和多線程等，提高應用程序的吞吐能力。

Linux Performance Tools



（圖片來自 brendangregg.com）

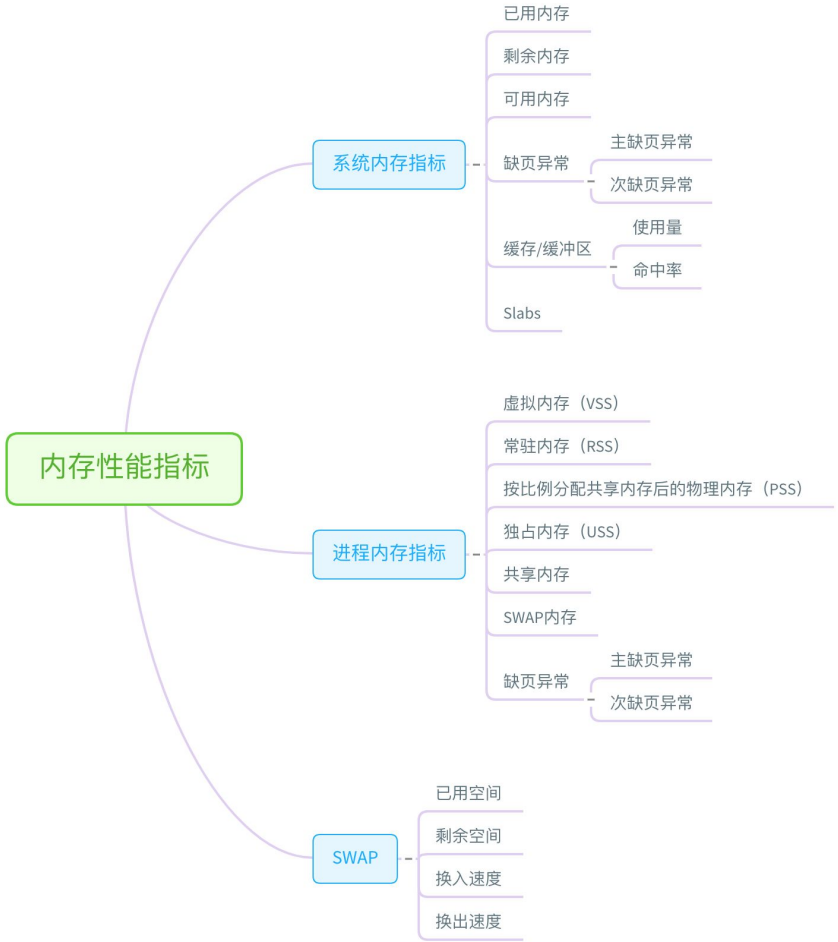
CPU 性能工具


CPU性能指标
CPU使用率
用户CPU
系统CPU
IOWAIT
软中断
硬中断
窃取CPU
客户CPU
上下文切换
自愿上下文切换
非自愿上下文切换
平均负载
CPU缓存命中率

CPU 性能工具速查表

CPU性能工具		
性能指标	**性能工具**	**说明**
平均负载	uptime top /proc/loadavg	uptime最简单；top提供了更全的指标；/proc/loadavg常用于监控系统
系统CPU使用率	vmstat mpstat top sar /proc/stat	top、vmstat、mpstat 只可以动态查看，而 sar 还可以记录历史数据；/proc/stat 是其他性能工具的数据来源，也常用于监控
进程CPU使用率	top ps pidstat htop atop	top和ps可以按CPU使用率给进程排序，而pidstat只显示实际用了CPU的进程；htop和atop以不同颜色显示更直观
系统上下文切换	vmstat	除了上下文切换次数，还提供运行状态和不可中断状态进程的数量
进程上下文切换	pidstat	注意加上 -w 选项
软中断	top mpstat /proc/softirqs	top提供软中断CPU使用率，而/proc/softirqs和mpstat提供了各种软中断在每个CPU上的运行次数
硬中断	vmstat /proc/interrupts	vmstat提供总的中断次数，而/proc/interrupts提供各种中断在每个CPU上运行的累积次数
网络	dstat sar tcpdump	dstat和sar提供总的网络接收和发送情况，而tcpdump则是动态抓取正在进行的网络通讯
I/O	dstat sar	dstat和sar都提供了I/O的整体情况
CPU缓存	perf	使用 perf stat 子命令
CPU数	lscpu /proc/cpuinfo	lscpu更直观
事件剖析	perf、火焰图 execsnoop	perf和火焰图用来分析热点函数以及调用栈，execsnoop用来监测短时进程
动态追踪	ftrace bcc、systemtap	ftrace用于跟踪内核函数调用栈，而bcc和systemtap则用于跟踪内核或应用程序的执行过程（注意bcc要求内核版本>=4.1）

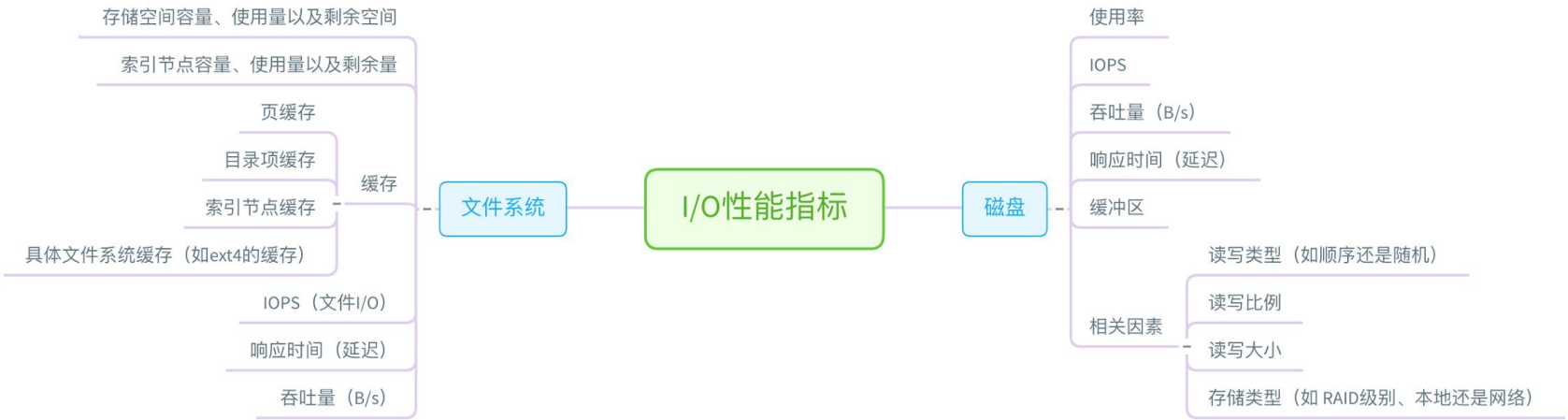
内存性能工具


内存性能指标
系统内存指标
已用内存
剩余内存
可用内存
缺页异常
主缺页异常
次缺页异常
缓存/缓冲区
使用量
命中率
Slabs
进程内存指标
虚拟内存（VSS）
常驻内存（RSS）
按比例分配共享内存后的物理内存（PSS）
独占内存（USS）
共享内存
SWAP内存
缺页异常
主缺页异常
次缺页异常
SWAP
已用空间
剩余空间
换入速度
换出速度

内存性能工具速查表

内存性能工具		
性能指标	**性能工具**	**说明**
系统已用、可用、剩余内存	free、vmstat、sar /proc/meminfo	free最为简单，而vmstat、sar更为全面；/proc/meminfo是其他工具的数据来源，也常用于监控系统中
进程虚拟内存、常驻内存、共享内存	ps、top、pidstat /proc/pid/stat /proc/pid/status	ps和top最简单，而pidstat则需要加上-r选项；/proc/pid/stat和/proc/pid/status是其他工具的数据来源，也常用于监控系统中
进程内存分布	pmap /proc/pid/maps	/proc/pid/maps是pmap的数据来源
进程Swap换出内存	top、/proc/pid/status	/proc/pid/status是top的数据来源
进程缺页异常	ps、top、pidstat	注意给pidstat加上-r选项
系统换页情况	sar	注意加上-B选项
缓存/缓冲区用量	free、vmstat 、sar cachestat	vmstat最常用，而cachestat需要安装bcc
缓存/缓冲区命中率	cachetop	需要安装bcc
SWAP已用空间和剩余空间	free、sar	free最为简单，而sar还可以记录历史
Swap换入换出	vmstat、sar	vmstat最为简单，而sar还可以记录历史
内存泄漏检测	memleak、valgrind	memleak需要安装bcc，valgrind还可以在旧版本（如3.x）内核中使用
指定文件的缓存大小	pcstat	需要从源码下载安装

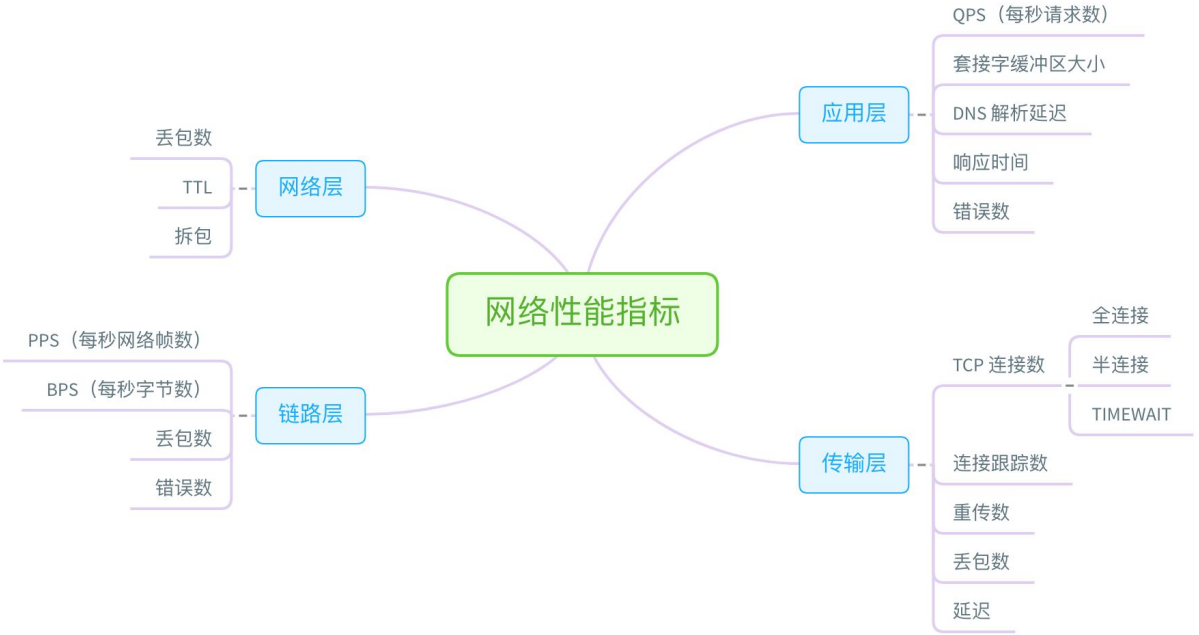
磁盘I/O 性能工具


I/O性能指标
文件系统
存储空间容量、使用量以及剩余空间
索引节点容量、使用量以及剩余量
缓存
页缓存
目录项缓存
索引节点缓存
具体文件系统缓存（如ext4的缓存）
IOPS（文件I/O）
响应时间（延迟）
吞吐量（B/s）
磁盘
使用率
IOPS
吞吐量（B/s）
响应时间（延迟）
缓冲区
相关因素
读写类型（如顺序还是随机）
读写比例
读写大小
存储类型（如 RAID级别、本地还是网络）

磁盘I/O 性能工具速查表

文件系统和磁盘I/O性能工具		
性能指标	**性能工具**	**说明**
文件系统空间容量、使用量以及剩余空间	df	详细文档可以执行 info coreutils 'df invocation' 命令查询
索引节点容量、使用量以及剩余量	df	注意加上 -i 选项
页缓存和可回收Slab缓存	/proc/meminfo sar、vmstat	注意sar需要加上-r选项，而/proc/meminfo是其他工具的数据来源，也常用于监控
缓冲区	/proc/meminfo sar、vmstat	注意sar需要加上-r选项，而/proc/meminfo是其他工具的数据来源，也常用于监控
目录项、索引节点以及文件系统的缓存	/proc/slabinfo slabtop	slabtop更直观，而/proc/slabinfo常用于监控
磁盘 I/O 使用率、IOPS、吞吐量、响应时间、I/O平均大小以及等待队列长度	iostat、sar、dstat /proc/diskstats	iostat最为常用，注意使用 iostat -d -x 或 sar -d 选项；/proc/diskstats则是其他工具数据来源，也常用于监控
进程I/O大小以及I/O延迟	pidstat、iotop	注意使用 pidstat -d 选项
块设备 I/O 事件跟踪	blktrace	需要跟blkparse配合使用，比如 blktrace -d /dev/sda -o- | blkparse -i-
进程 I/O 系统调用跟踪	strace、perf trace	strace只可以跟踪单个进程，而perf trace还可以跟踪所有进程的系统调用
进程块设备I/O大小跟踪	biosnoop、biotop	需要安装bcc
动态追踪	ftrace bcc、systemtap	ftrace用于跟踪内核函数调用栈，而bcc和systemtap则用于跟踪内核或应用程序的执行过程（注意bcc要求内核版本>=4.1）

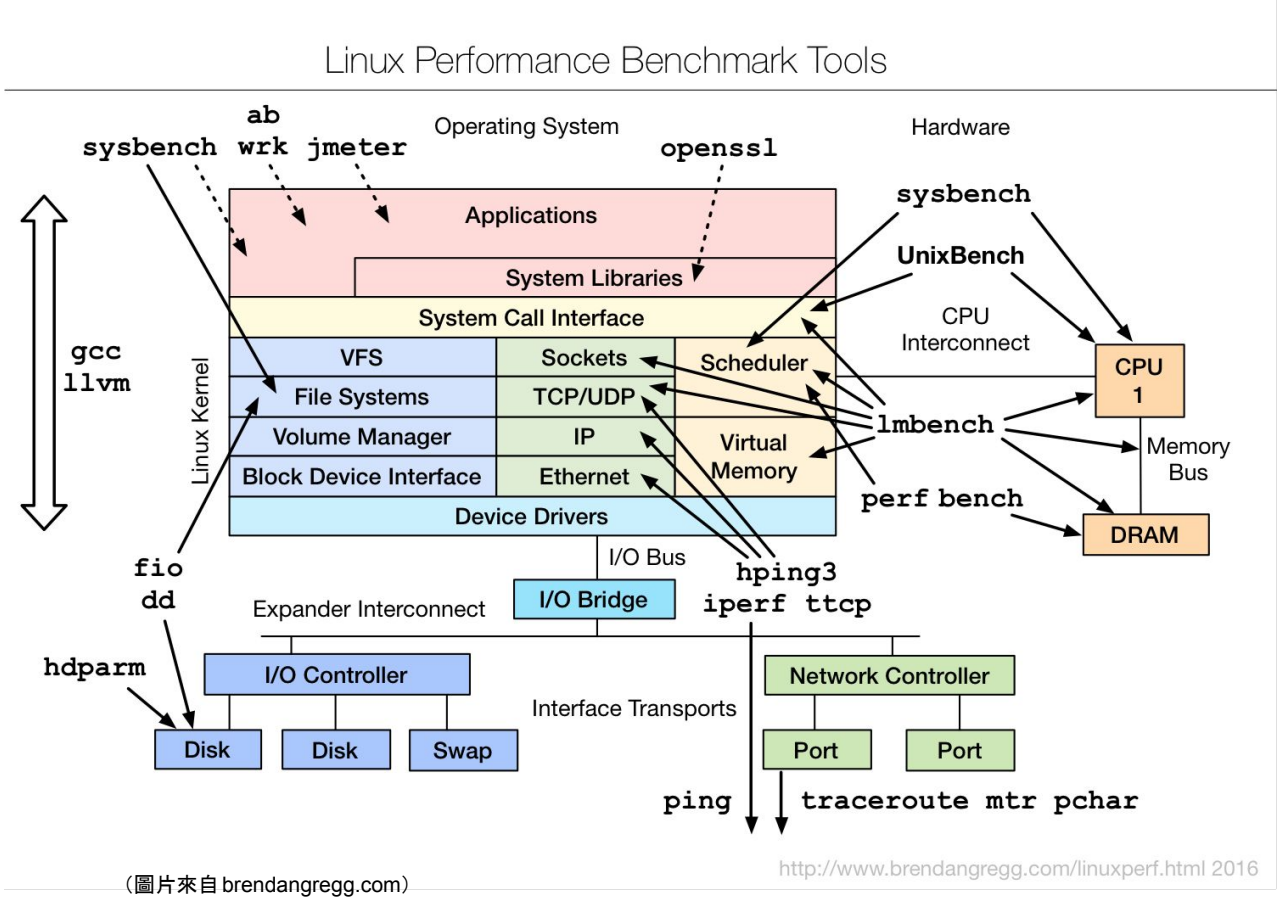
網絡性能工具


网络性能指标
应用层
QPS（每秒请求数）
套接字缓冲区大小
DNS 解析延迟
响应时间
错误数
网络层
丢包数
TTL
拆包
链路层
PPS（每秒网络帧数）
BPS（每秒字节数）
丢包数
错误数
传输层
TCP 连接数
全连接
半连接
TIMEWAIT
连接跟踪数
重传数
丢包数
延迟

網絡性能工具速查表

网络性能工具		
性能指标	**性能工具**	**说明**
吞吐量（BPS）	sar、nethogs、iftop /proc/net/dev	分别可以查看网络接口、进程以及IP地址的网络吞吐量；/proc/net/dev常用于监控
吞吐量（PPS）	sar、/proc/net/dev	注意使用sar –n DEV选项
网络连接数	netstat、ss	ss速度更快
网络错误数	netstat、sar	注意使用netstat –s或者sar –n EDEV/EIP选项
网络延迟	ping、hping3	ping基于ICMP，而hping3则基于TCP协议
连接跟踪数	conntrack /proc/sys/net/netfilter/nf_conntrack_count /proc/sys/net/netfilter/nf_conntrack_max	conntrack可用来查看所有连接跟踪的相信信息，nf_conntrack_count只是连接跟踪的数量，而nf_conntrack_max 则限制了总的连接跟踪数量
路由	mtr、traceroute、route	route用于查询路由表，而mtr和traceroute则用来排查和定位网络链路中的路由问题
DNS	dig、nslookup	用于排查DNS解析的问题
防火墙和NAT	iptables	用于排查防火墙及NAT的问题
网卡选项	ethtool	用于查看和配置网络接口的功能选项
网络抓包	tcpdump、Wireshark	通常在服务器中使用tcpdump抓包后再复制出来用Wireshark的图形界面分析
动态追踪	ftrace bcc、systemtap	ftrace用于跟踪内核函数调用栈，而bcc和systemtap则用于跟踪内核或应用程序的执行过程（注意bcc要求内核版本>=4.1）

基準測試工具



（圖片來自 brendangregg.com）

小結

- 當分析性能問題時，大的來說，主要有這麼兩個步驟：
  - 第一步，從性能瓶頸出發，根據系統和應用程序的運行原理，確認待分析的性能指標。
  - 第二步，根據這些圖表，選出最合適的性能工具，然後了解並使用工具，從而更快觀測到需要的性能數據。