X86平台RAC安装最佳实践

杜诚文
QQ: 23828728
微信:18047120719

CONTENTS

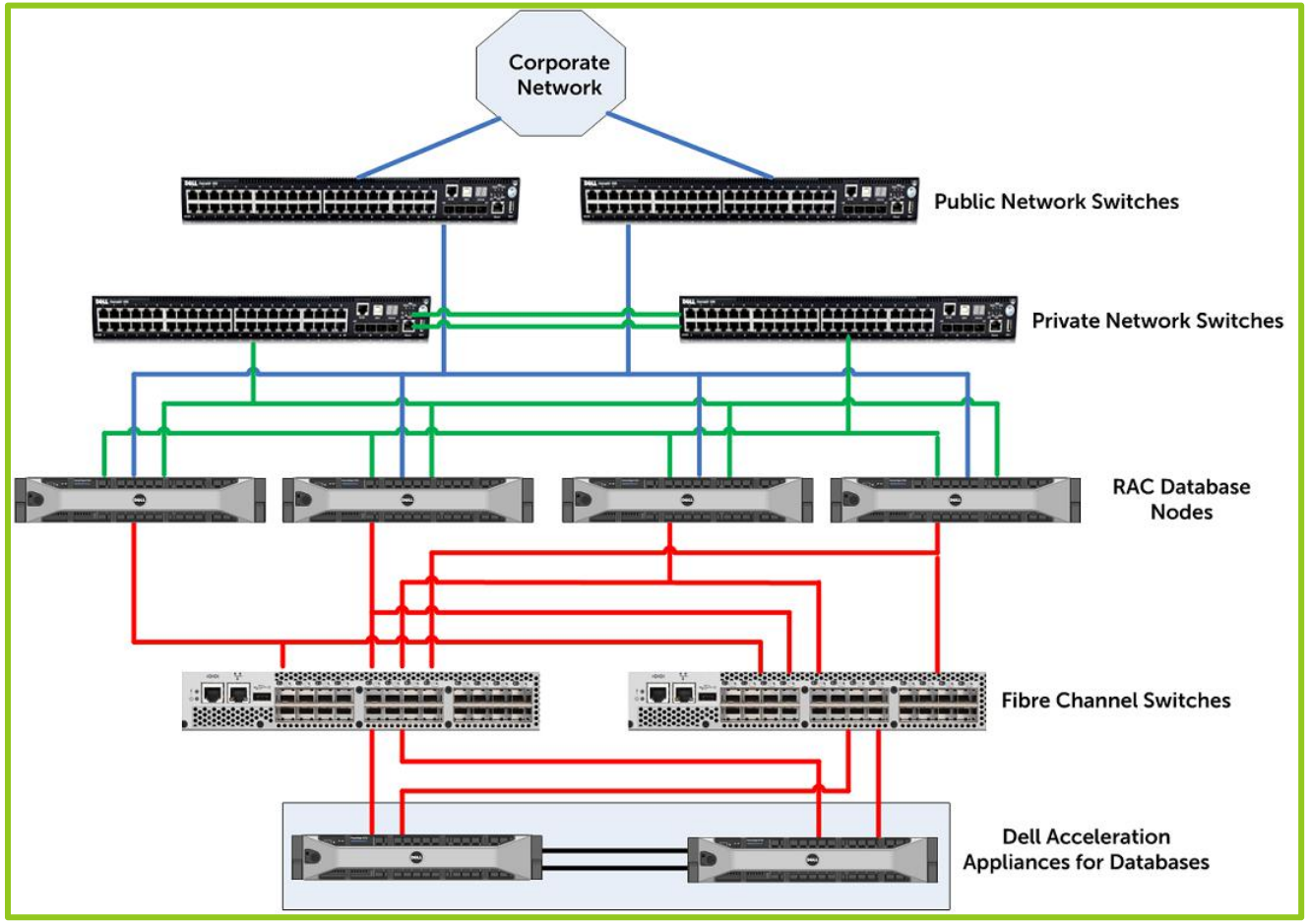
01

硬件设备

硬件架构

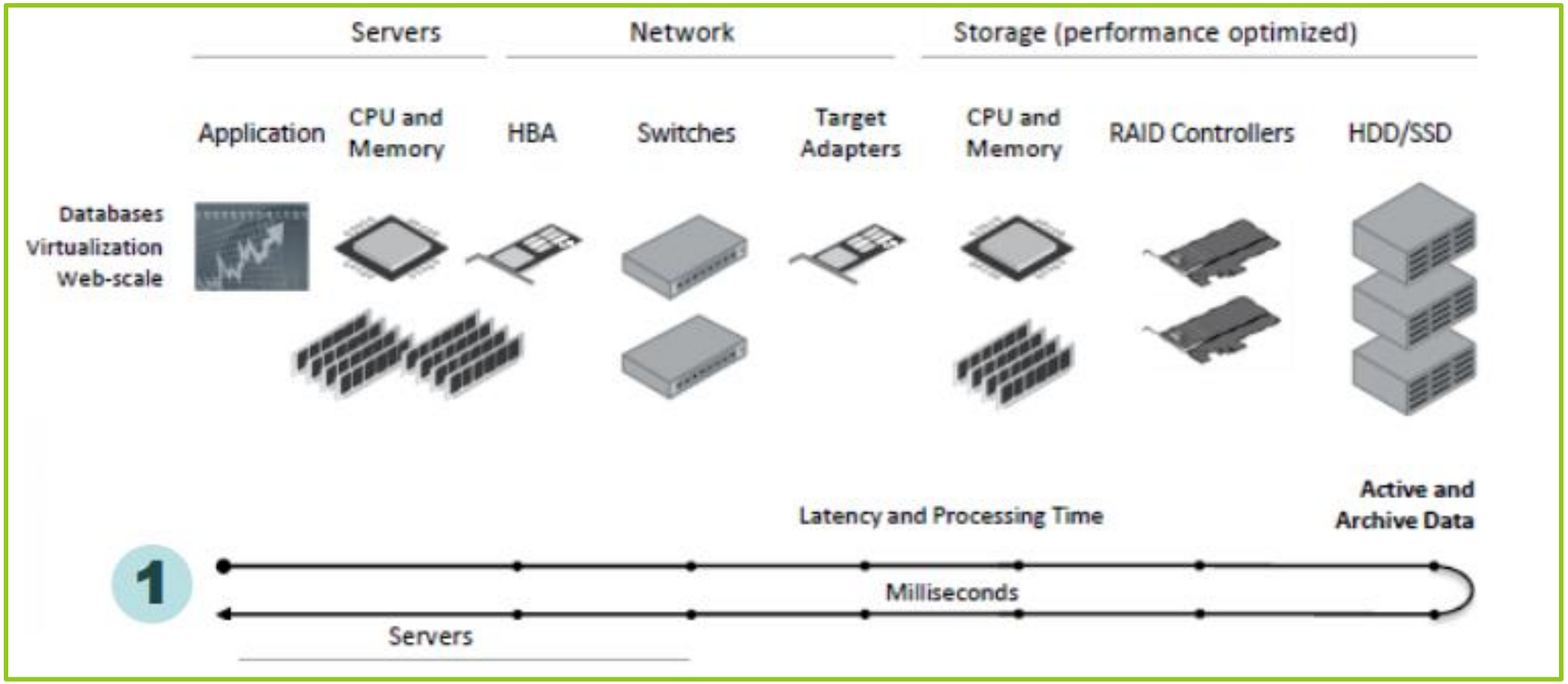
所有设备不存在单点故障

设备	数量
公共网卡	2
心跳网卡	2
HBA卡	2
FC交换机	2
IP交换机	2
服务器	2
存储	2



硬件架构

数据库的性能，受限于速度最慢的设备，传统架构中，最容易出现瓶颈的设备为存储。



硬件架构

选择合适的硬件

1. 服务器选择：2路/4路/8路
2. 存储设备：STAT HDD/FC HDD/SAS SSD/PCIe SSD/NVRAM
3. 数据网络：1/10/25/40/50/100GbE
4. 存储网络：4/8/16/32Gb
5. IB网络：56/100Gb
6. 如何优化服务器

服务器

2U 2路机架服务器

Intel® Xeon® E5-2600 v3/v4 系列处理器

4U 4路机架服务器

Intel® Xeon® E7-4800 v3/v4和E7-8800 v3/v4系列处理器

8U 8路机架服务器

Intel® Xeon® E7-8800 v3/v4系列处理器

华为KunLun开放架构小型机支持
8路、16路、32路Intel CPU

如何选择?

不差钱都用4路。

为什么不选8路?

贵、复杂!

服务器

如何选择服务器

① CPU数

② 内存容量

① 本地磁盘数量

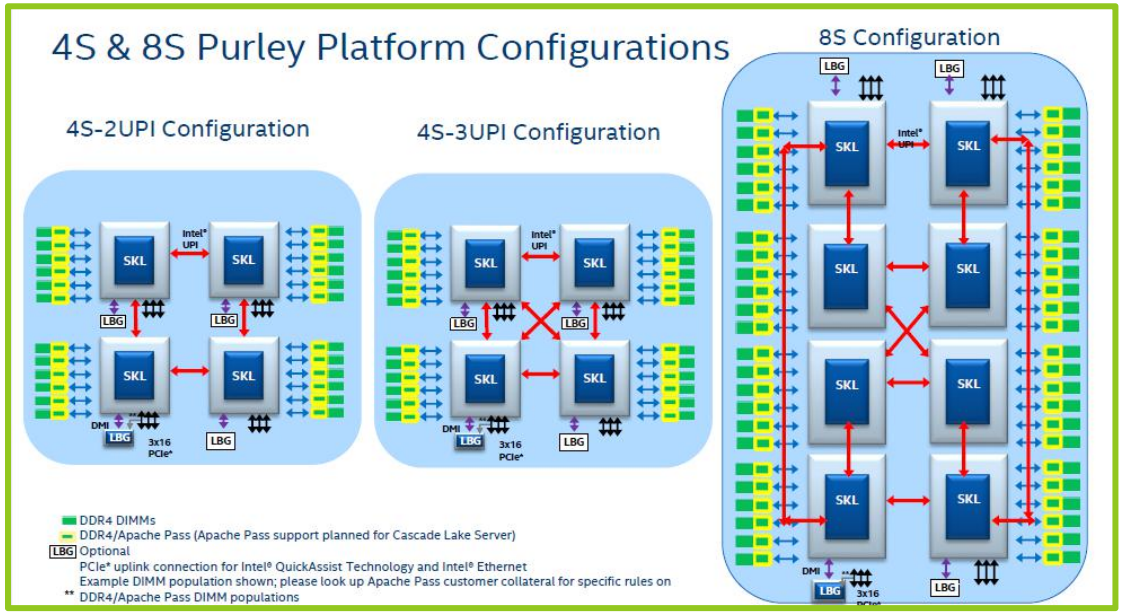
② RAID卡

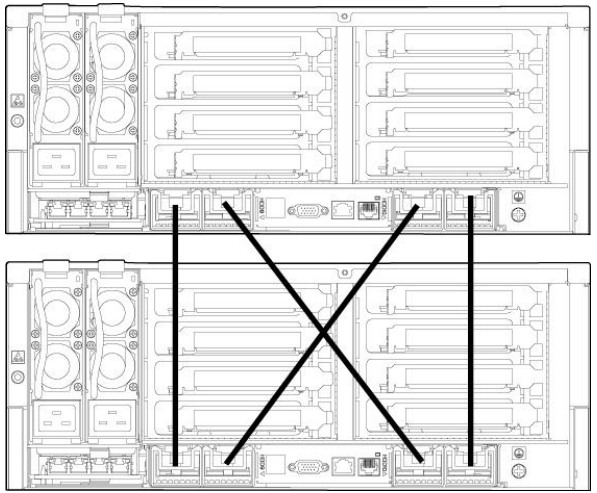
③ PCI-E插槽

8路服务器

① 由2个4路服务器组成

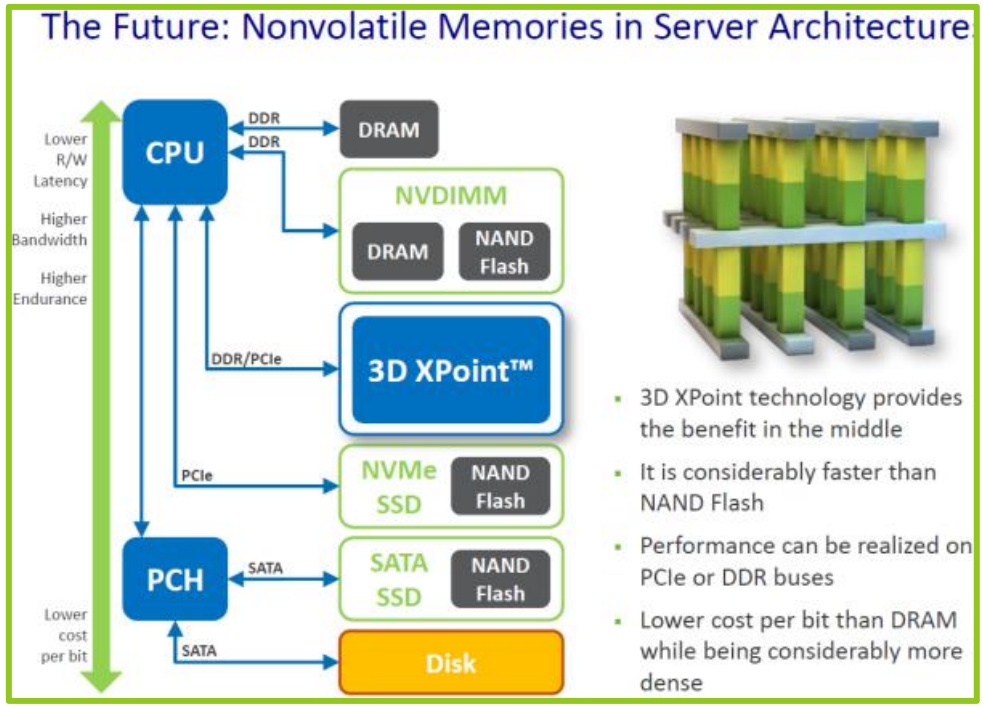
② 通过QPI线缆连接

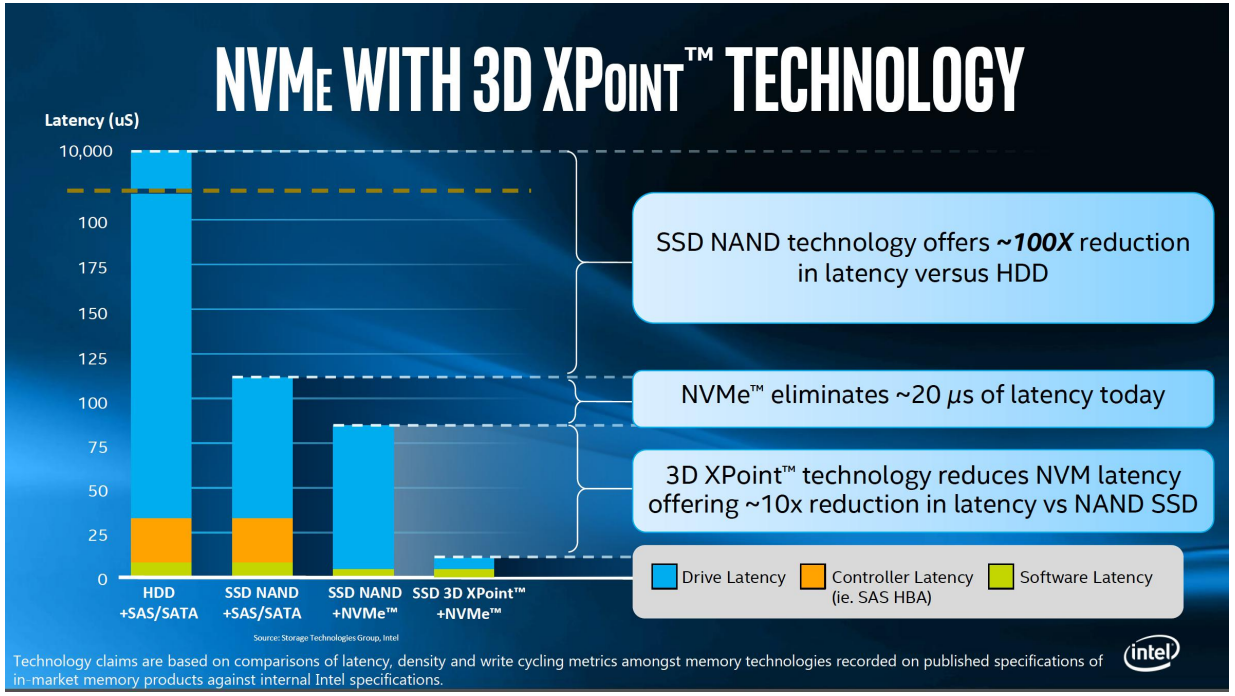




存储设备

使用SSD或者NVMe SSD设备，至少获得数百倍甚至万倍的IOPS提升；





数据网络

① 尽量使用10GE网络设备

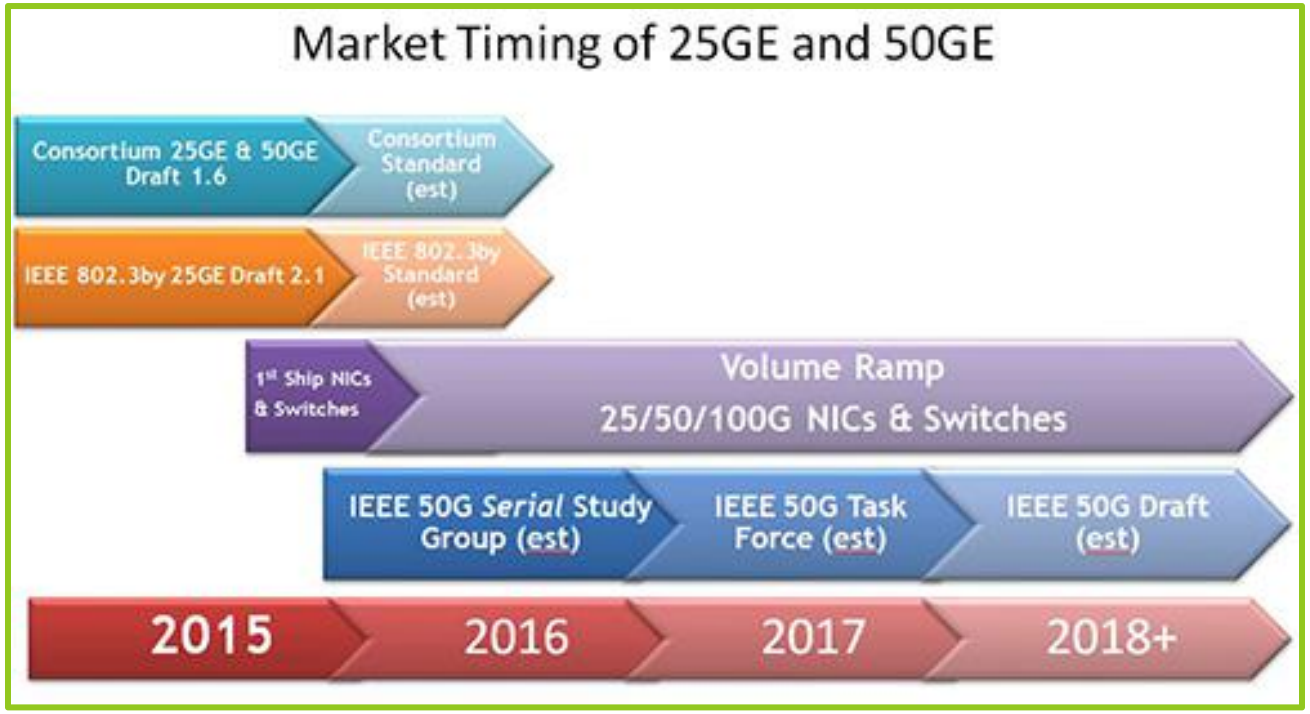
② 最快的设备留给心跳网络

③ 条件允许使用独立心跳交换机

④ 使用操作系统层面的网络绑定

⑤ 心跳网络MTU设置为9000

⑥ 最新的网络设备开始支持RoCE



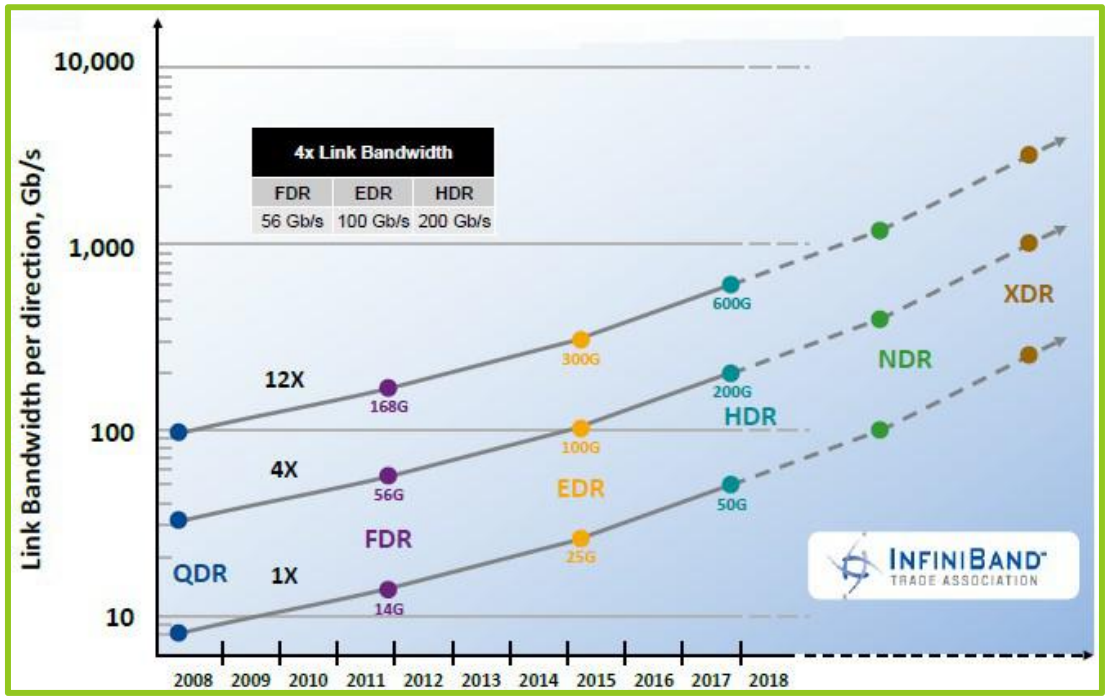
存储网络

- 8Gb HBA卡基本都是库存设备，如果可以尽量使用16Gb设备
- 8Gb HBA卡实际速率为：(8Gb / 8) * (8b / 10b) = 800MB

Name	Line-rate (gigabaud)	Line coding	Nominal throughput per direction; MB/s	Net throughput per direction; MB/s[v 1][v 2]	Availability
4GFC	4.25	8b10b	400	412.9	2004
8GFC	8.5	8b10b	800	825.8	2005
16GFC	14.025	64b66b	1,600	1,652	2011
32GFC "Gen 6"	28.05	256b257b	3,200	3,303	2016[8]
128GFC "Gen 6"	28.05 ×4	256b257b	12,800	13,210	2016[8]

IB网络

- InfiniBand（直译为“无限带宽”技术，缩写为IB）是一个用于高性能计算的计算机网络通信标准，它具有极高的吞吐量和极低的延迟。
- **该技术目前广泛应用于数据库一体机。**
- RDS协议用于心跳网络。
- **RDMA协议用于与存储节点之间进行数据传输。**



如何优化服务器配置(BIOS)

BIOS中有CPU和内存参数可以设置为节能模式，节能在降低能耗的同时也意味着性能的下降。

① 选择Performance Per Watt Optimized(DAPC)模式，发挥CPU最大性能，数据库服务器不需要节能和休眠；

② 关闭C1E和C States等选项，提升CPU效率；

③ Memory Frequency选择Maximum Performance；

④ 内存设置菜单中，启用Node Interleaving，避免NUMA问题；

02

操作系统

操作系统认证

数据库只建议安装在，已经通过ORACLE官方认证的操作系统版本上面。

操作系统认证

通过-ignoreSysPrereqs方式可以跳过操作系统验证，但是不建议这么做。(文档 ID 820135.1)

认证结果

显示 Oracle Database 11.2.0.4.0 认证。 查看认证简讯 分组方式 产品类别 显示 主要认证

查看 共享链接

已认证	发行版/版本号
▽ 操作系统 (13 项)	
Fujitsu BS2000	5 个版本 (V9.0, V8.0, V7.0, V11.0, V10.0)
Fujitsu BS2000/OSD (SQ series)	4 个版本 (V9.0, V8.0, V11.0, V10.0)
HP OpenVMS Itanium	1 个版本 (8.4)
HP-UX Itanium	1 个版本 (11.31)
HP-UX PA-RISC (64-bit)	1 个版本 (11.31)
IBM AIX on POWER Systems (64-bit)	4 个版本 (7.2, 7.1, 6.1, 5.3)
IBM: Linux on System z	6 个版本 (SLES 11, SLES 10, Red Hat Enterprise Linux 7, Red Hat Enterprise Linux 6, Red Hat Enterprise Linux 5, Red Hat Enterprise Linux 4)
Linux x86	11 个版本 (SLES 11, SLES 10, Red Hat Enterprise Linux 6, Red Hat Enterprise Linux 5, Red Hat Enterprise Linux 4, Oracle Linux 6, Oracle Linux 5, Oracle Linux 4, Asianux 4, Asianux 3) 和 1 个其他结果
Linux x86-64	16 个版本 (SLES 12, SLES 11, SLES 10, Red Hat Enterprise Linux 7, Red Hat Enterprise Linux 6, Red Hat Enterprise Linux 5, Red Hat Enterprise Linux 4, Oracle Linux 7, Oracle Linux 6, Oracle Linux 5) 和 6 个其他结果
Microsoft Windows (32-bit)	8 个版本 (XP, Vista, 8.1, 8, 7, 2008, 2003 R2, 2003)
Microsoft Windows x64 (64-bit)	11 个版本 (XP, Vista, 8.1, 8, 7, 2012 R2, 2012, 2008 R2, 2008, 2003 R2) 和 1 个其他结果
Oracle Solaris on SPARC (64-bit)	2 个版本 (11, 10)
Oracle Solaris on x86-64 (64-bit)	2 个版本 (11, 10)
▷ 桌面应用程序, 浏览器和客户端 (1 项)	

另请参阅: Oracle Real Application Clusters 11.2.0.4.0, Advanced Compression 11.2.0.4.0, Advanced Networking Option 11.2.0.4.0, Advanced Security 11.2.0.4.0 和 9 个其他结果

安装配置建议

使用推荐的操作系统版本、安装补丁程序

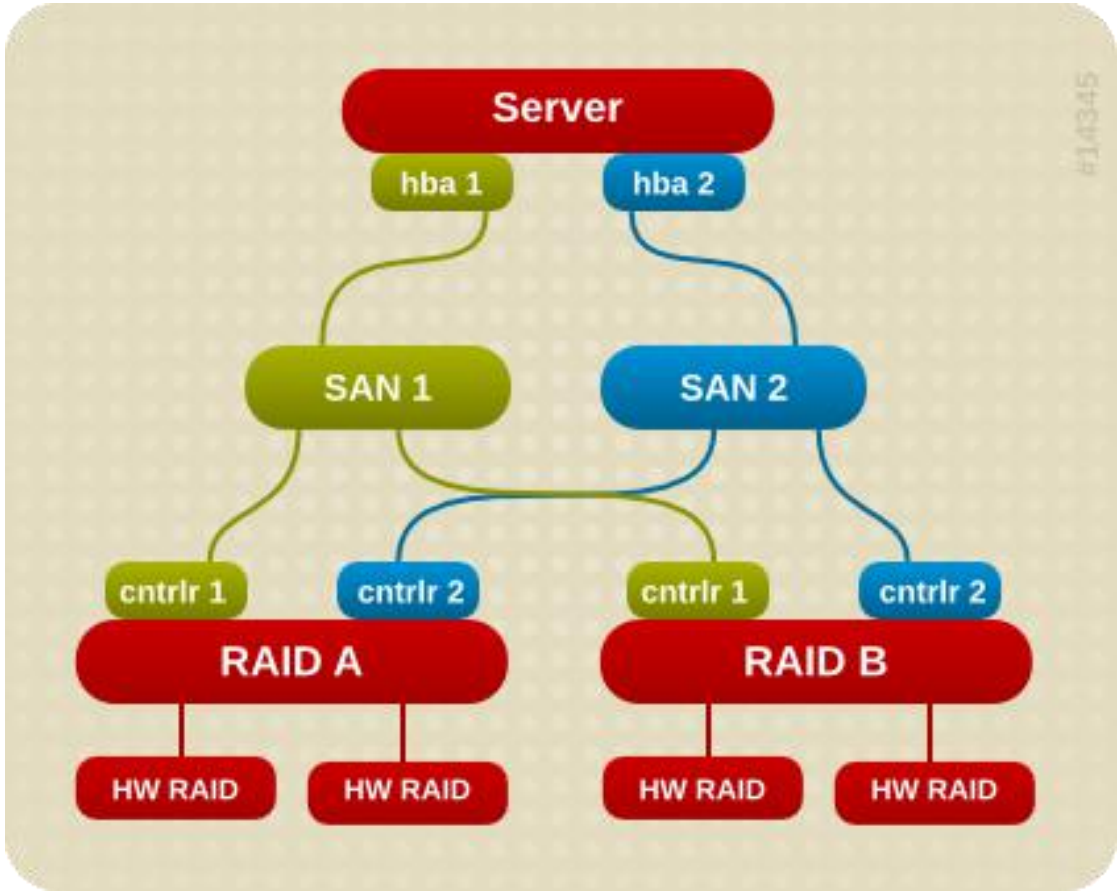
配置系统参数、网络参数、IO参数、内存参数、NUMA、网卡MTU

官方提供的最佳实践文档：

- Oracle Database (RDBMS) on Unix AIX,HP-UX,Linux,Mac OS X,Solaris,Tru64 Unix Operating Systems Installation and Configuration Requirements Quick Reference (8.0.5 to 11.2) (文档 ID 169706.1)
- Oracle Database (RDBMS) on Unix AIX,HP-UX,Linux,Solaris and MS Windows Operating Systems Installation and Configuration Requirements Quick Reference (12.1/12.2) (文档 ID 1587357.1)
- Information Center: Install and Configure Database Server/Client Installations (文档 ID 1351051.2)

存储多路径

设备映射器多路径（DM-Multipath）可让您将服务器节点和存储阵列间的多个 I/O 路径配置为一个单一设备。这些 I/O 路径是可包含独立电缆、交换机以及控制器的物理 SAN 连接。多路径集合了 I/O 路径，并生成由这些整合路径组成的新设备。



存储多路径

每个多路径设备都有一个全球识别符（WWID），它是一个全球唯一的无法更改的号码。默认情况下会将多路径设备的名称设定为它的 WWID。

```
[root@rpsdb1 ~]# multipath -ll asmmgmtdisk01
asmmgmtdisk01 (360002ac0000000000000001990001a7f3) dm-11 3PARdata,VV
size=100G features='1 queue_if_no_path' hwhandler='1 alua' wp=rw
`-+- policy='round-robin 0' prio=50 status=active
  |- 1:0:0:40 sdx                65:112 active ready running
.....
  `- 3:0:1:40 sdas                66:192 active ready running

[root@rpsdb1 ~]# /usr/lib/udev/scsi_id --whitelisted --replace-whitespace --device=/dev/sdx
```

存储多路径

为什么需要关注这么细节的内容？ 因为通过对设备的监控，可以获取到设备全链路的状况。

```
[root@rpsdb1 ~]# lspci | grep -i hba
c1:00.0 Fibre Channel: QLogic Corp. ISP2532-based 8Gb Fibre Channel to PCI Express HBA (rev 02)
……
c4:00.1 Fibre Channel: QLogic Corp. ISP2532-based 8Gb Fibre Channel to PCI Express HBA (rev 02)

[root@rpsdb1 ~]# ls -la /dev/disk/by-path/* | grep -E '(sdx|sdy|sdar|sdas)' | awk '{print $9}'
/dev/disk/by-path/pci-0000:c1:00.0-fc-0x20640002ac01a7f3-lun-40
……
/dev/disk/by-path/pci-0000:c4:00.0-fc-0x21630002ac01a7f3-lun-40
```

IO调度策略调整

I/O调度器的工作是管理块设备的请求队列。它决定队列中的请求排列顺序及在什么时候派发请求到块设备。这样做有利于减少磁盘寻址时间，从而提高全局吞吐量。目前主流Linux发行版本使用三种I/O调度器：DeadLine、CFQ、NOOP。

从原理上看，DeadLine是一种以提高机械硬盘吞吐量为思考出发点的调度算法， NOOP 对于闪存设备和嵌入式系统是最好的选择。对于固态硬盘来说使用NOOP是最好的，DeadLine次之，而CFQ效率最低。实际表现如何，还需要进行测试验证。

```
# cat
/sys/block/sdc/queue/scheduler
noop [deadline] cfq

# vi /etc/default/grub
GRUB_CMDLINE_LINUX=“.....
elevator=deadline“
# grub2-mkconfig -o
/boot/efi/EFI/redhat/grub.cfg
```

03

数据库配置 (11G)

数据库版本选择

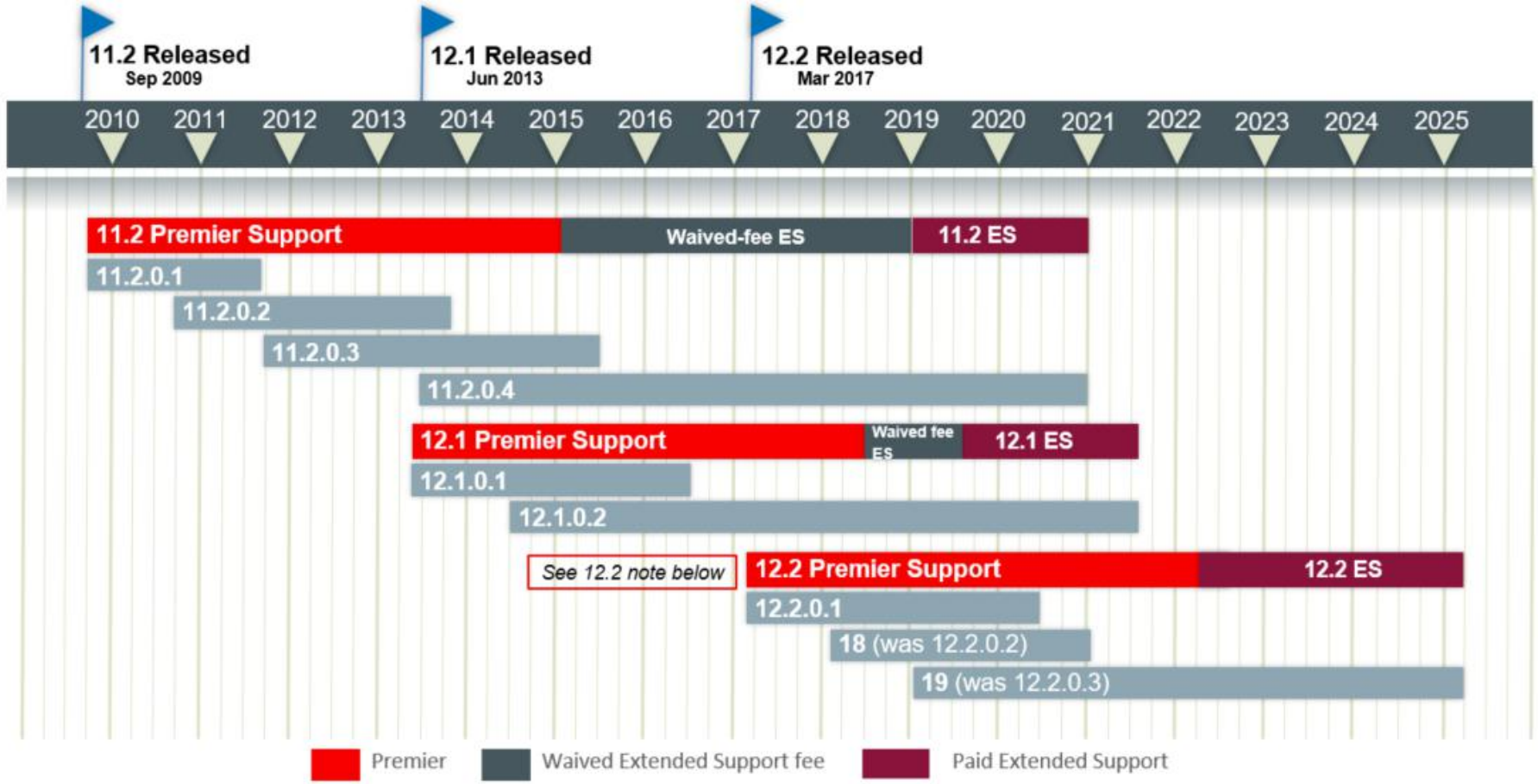
数据库版本选择的依据

- 应用需要用到功能和特性？分析函数，分区表。
- 是否需要使用数据库的特性来实现某些功能？例如：ADG，闪回数据库。
- 数据库软件生命周期，什么时候开始停止补丁支持？例如：11.2.0.4 是 11.2 的最终补丁集。
- CPU以及操作系统的支持情况？例如：11.2 将是 HP-UX PA-RISC 上发布的最后一个数据库版本。
- 数据库升级过程中应用是否需要改造？例如：WMSYS.WM_CONCAT的变化。
- 引导客户而不是让客户被应用引导。

Release Schedule of Current Database Releases (文档 ID 742060.1)

数据库版本选择

如果可以的话，生产系统的版本尽量中庸



集群调整

安装最新的PSU补丁

Master Note for Database Proactive Patch Program (文档 ID 756671.1)

降低vip资源对网络的依赖 (11.2.0.3之前问题较多)

```
crsctl modify res ora.racdb1.vip -attr "STOP_DEPENDENCIES=hard(intermediate:ora.net1.network)"
crsctl modify res ora.racdb2.vip -attr "STOP_DEPENDENCIES=hard(intermediate:ora.net1.network)"
crsctl modify res ora.scan1.vip -attr "STOP_DEPENDENCIES=hard(intermediate:ora.net1.network)"
crsctl modify resource ora.net1.network -attr "CHECK_INTERVAL=30"
```

集群调整调整

修改资源为总是启动

grid 用户： crsctl modify resource "ora.LISTENER.lsnr" -attr "AUTO_START=always"

oracle 用户： crsctl modify resource "ora.racdb.db" -attr "AUTO_START=always"

ASM参数调整

- 调整内存相关参数；

  alter system set memory_max_target=4g scope=spfile;
  alter system set memory_target=4g scope=spfile;

- ASM是否需要存放低版本的数据库，DG兼容性参数；
- 规划的数据库容量，考虑设置合理的DG AU_SIZE；
- 显示指定asm_diskstring，优化磁盘设备发现时间；

数据库参数调整

① 避免自动内存管理，它就像红绿灯，系统正常的时候，它表现的很良好，系统异常的时候，它会让系统更混乱；

② 建议关闭DRM，虽然它设计的出发点是好的，但生产系统中经常会造成不可控的场面；

③ 如果没有经过严格的测试，新特性建议关闭比较稳妥；

④ 优化器有大量的潜在bug，如果语句执行计划或者结果集有问题，可以考虑调整优化器相关参数；

⑤ 从MOS上看看还有哪些常见的问题：

- 11.2.0.4 Patch Set - Availability and Known Issues (文档 ID 1562139.1)
- 12.1.0.1 Base Release - Availability and Known Issues (文档 ID 1565082.1)
- 12.2.0.1 Base Release - Availability and Known Issues (文档 ID 2239820.1)

数据库调整

表空间

调整表空间容量(SYS/SYSAUX/UNDO/TEMP)

重做日志

调整日志文件大小，调整日志组数量

口令生命周期

password_life_time

审计序列缓存

AUDSES$

AWR保留周期

建议大于31天

数据库资源管理及定时任务

DBA_AUTOTASK_CLIENT

DBA_SCHEDULER_WINDOWS

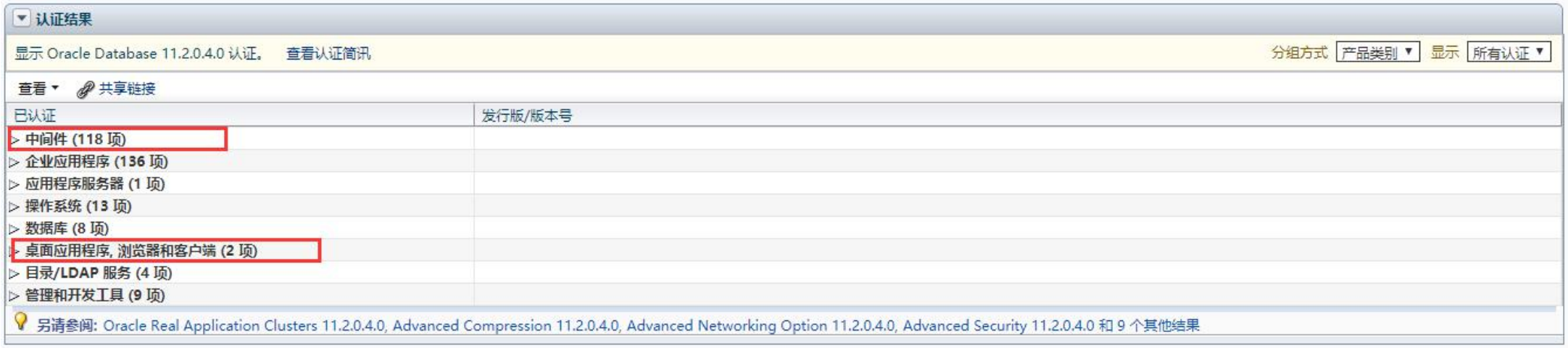
RMAN备份参数

CONFIGURE CONTROLFILE AUTOBACKUP ON;

CONFIGURE SNAPSHOT CONTROLFILE NAME TO '+ASMDG';

客户端兼容性

- Client / Server Interoperability Support Matrix for Different Oracle Versions (文档 ID 207303.1)
- 其他相关兼容性，通过MOS认证页面进行查询



TFA 收集器

TFA 收集器 - 加强版诊断日志收集工具 (文档 ID 2179484.1)

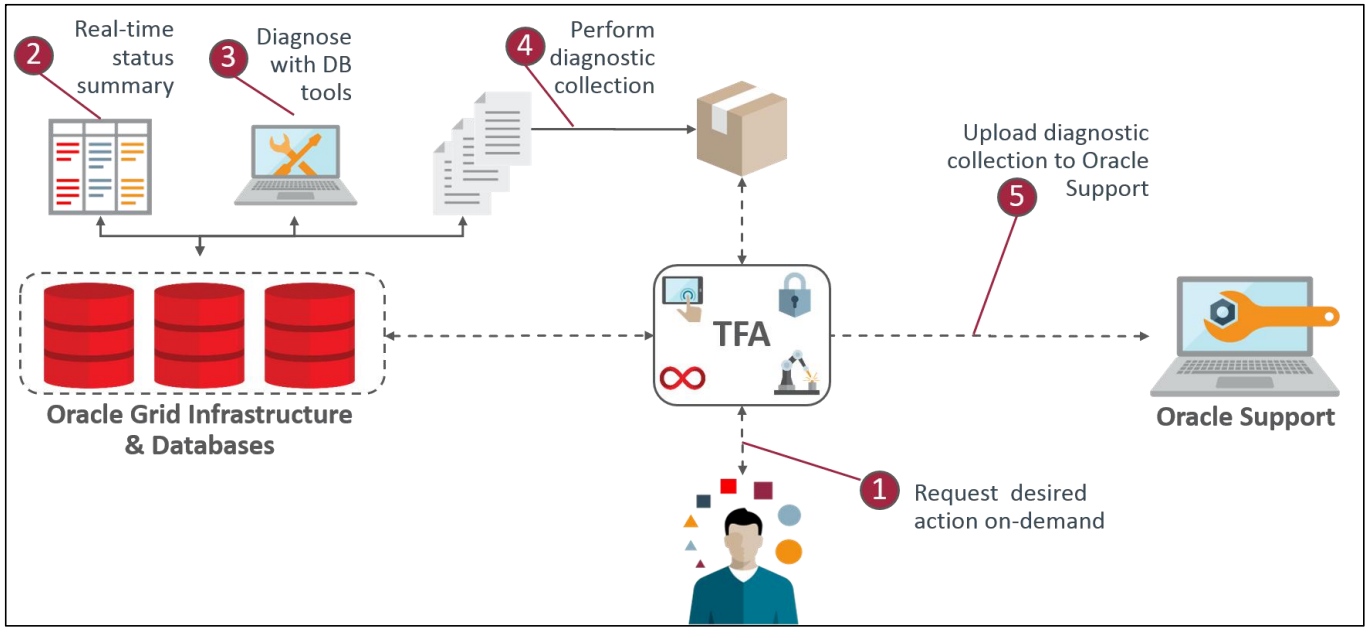
① 为所有诊断需要提供单一接口

② 收集整个集群的日志并放在一个地方

③ 收集问题发生时段的所有诊断日志

④ 通过快速诊断问题来节省资源

⑤ 所有需要的数据库支持工具一站式提供



04

高可用测试

高可用测试原则

- 高可用测试一定要伴随压力测试，无负载情况下，系统表现基本都是稳定的，有负载的情况下，系统表现是五花八门的；
- 使用Swingbench可以很好的辅助我们进行简单压力测试；
- 高可用测试不仅要观察系统能否恢复正常，还要想办法降低故障给系统带来的抖动；

① 多网卡绑定(主/备模式)

② 存储多路径绑定

③ 应用使用VIP连接数据库

④ 使用service控制应用的连接属性

硬件高可用测试

主机故障模拟测试	有计划停止单机模拟测试
	无计划停止单机模拟测试
	所有机器停机模拟测试
网络故障模拟测试	公网单网络故障模拟测试
	公网多网络故障模拟测试
	私网单网络故障模拟测试
	私网多网络故障模拟测试
	私网交换机故障模拟测试
存储故障模拟测试	单个存储链路故障模拟测试
	ASM单个磁盘丢失及恢复测试
	丢失OCR/Voting故障模拟测试
	丢失单个OCR/Voting故障模拟测试

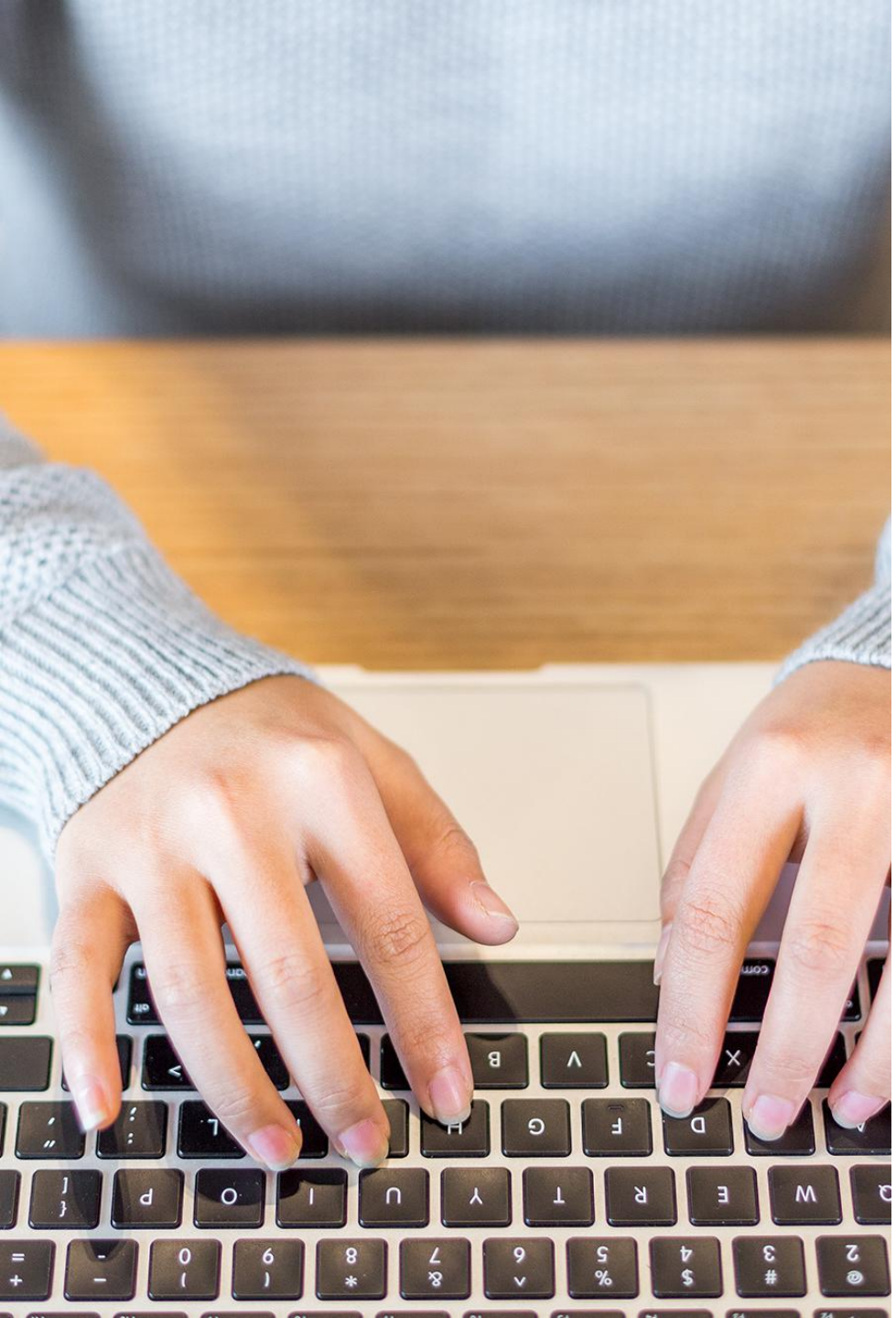
集群高可用测试/数据库高可用测试

CRS进程故障模拟测试	ohasd进程crash模拟测试
	CRSD进程crash模拟测试
	EVMD进程crash模拟测试
	OCSSD进程crash模拟测试
	oraagent进程crash模拟测试
	orarootagent进程crash模拟测试
	cssdagent进程crash模拟测试
	cssdmonitor进程crash模拟测试

LISTENER故障模拟测试	LISTNER故障模拟测试
	SCAN LISTNER故障模拟测试
数据库实例故障模拟测试	无计划实例故障模拟测试
	有计划实例故障模拟测试

05

性能测试



存储IO性能测试

物理磁盘IO性能测试工具

- FIO
- Vdbench
- ORION
- DD

数据库IO性能测试工具

- DBMS_RESOURCE_MANAGER.CALIBRATE_IO
- Swingbench

IO性能监控指标

- 数据库IO相关等待
- 磁盘性能 (iops/mbps/平均响应时间/队列)
- 多路径负载
- HBA卡性能

存储IO性能测试

--测试随机写IOPS，运行以下命令：

fio --name=/dev/sdb --ioengine=libaio --iodepth=64 --rw=randwrite --bs=8K --direct=1 --size=10240m --numjobs=64 --runtime=60 --group_reporting

--测试随机读IOPS，运行以下命令：

fio --filename=/dev/sdb --ioengine=libaio --iodepth=64 --rw=randread --bs=8K --direct=1 --size=10240m --numjobs=64 --runtime=60 --group_reporting

--测试顺序写吞吐量，运行以下命令：

fio --name=/dev/sdb --ioengine=libaio --iodepth=64 --rw=write --bs=1M --direct=1 --size=10240m --numjobs=64 --runtime=60 --group_reporting

--测试顺序读吞吐量，运行以下命令：

fio --name=/dev/sdb --ioengine=libaio --iodepth=64 --rw=read --bs=1M --direct=1 --size=10240m --numjobs=64 --runtime=60 --group_reporting

监控磁盘设备IO情况

iostat -dx sdt sdu sdan sdao dm-6 2 10000

网络性能测试

网络性能测试工具

① netperf

② iperf

③ ping

网络监控

① 可用性

② 网络带宽

③ 网络利用率

④ 响应时间

⑤ 丢包重组

网络性能测试

Netperf

--服务器端# netserver -4 -L 192.168.30.76 -p 8000

--客户端# netperf -H 192.168.30.76 -p 8000 -n 20 -l 60

Iperf

--服务器端# iperf3 -s 192.168.40.58 -p 5001 -i 2

--客户端# iperf3 -c 192.168.40.58 -P 4 -t 30 -i 2 -p 5001 -f

Swingbench压力测试

Swingbench是一款免费的数据库压力测试工具，支持11g，12c。

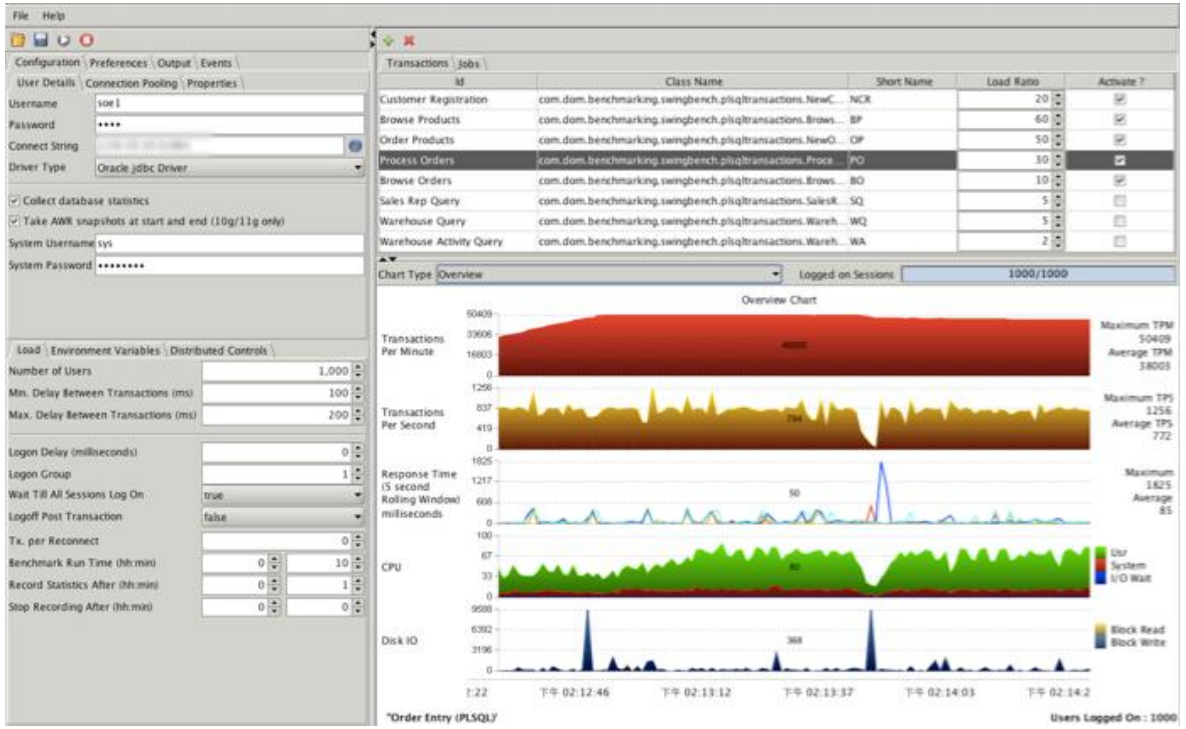
http://www.dominicgiles.com/swingbench.html

初始化

使用sys用户初始化，其他用户会有出现授权失败的问题，需要手工处理

12C需要连接PDB进行初始化，如果使用CDB，创建用户为C##SOE

```
./charbench -c /home/oracle/rpsdb.xml -uc 1000 -a -v users,tpm,tps,cpu,disk
```



SPA/DBReplay性能测试

Oracle 11g为数据库变更提供的性能测试新特性RAT(Real Application Testing)包含两个组件：

SQL Performance Analyzer

Database Replay

适用场景

SPA 用来评估SQL性能及执行计划的影响

DB Replay 评估变化后整体负载

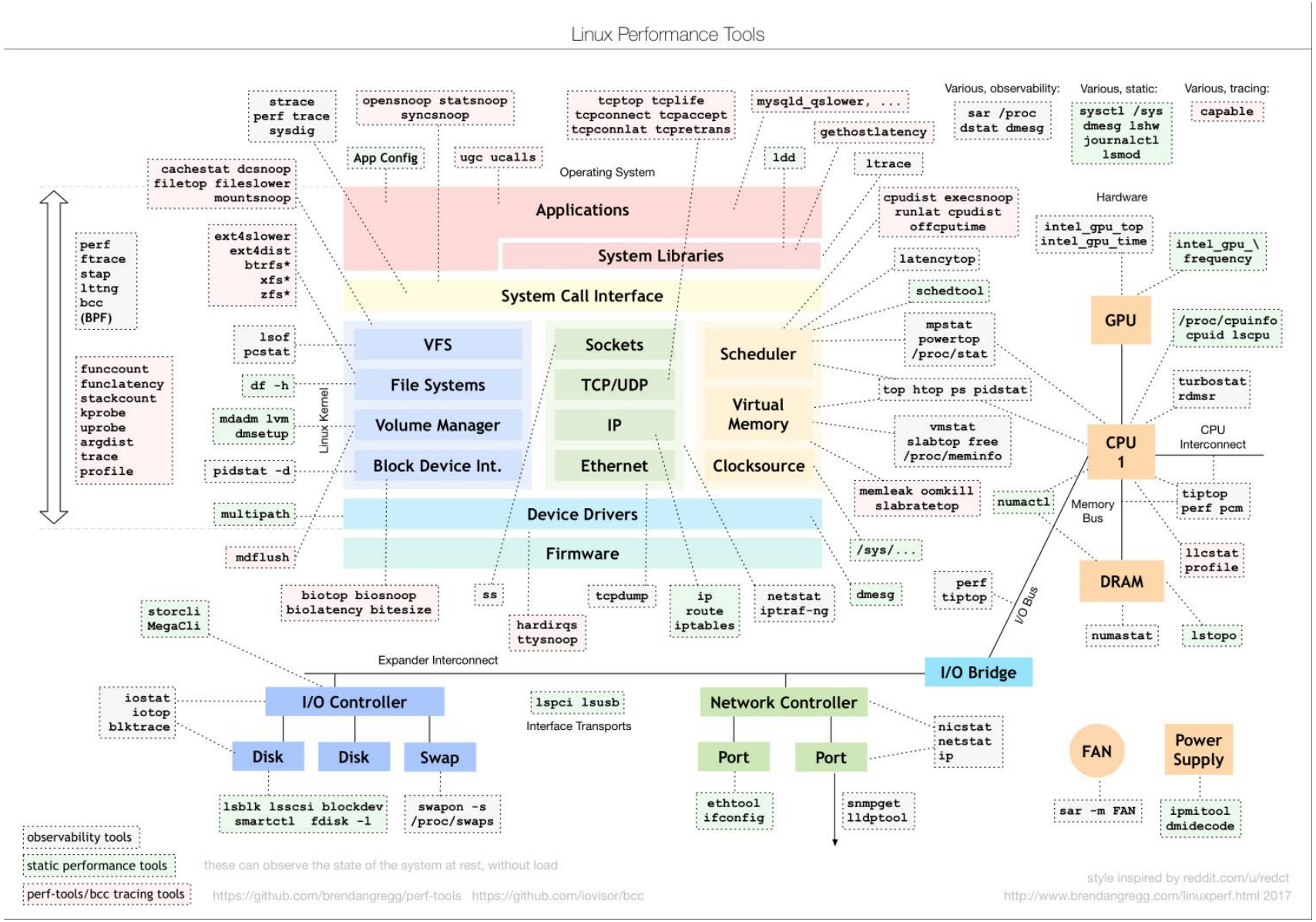
参考文档

Master Note for Real Application Testing Option (文档 ID 1464274.1)

SQL Performance Analyzer Summary (文档 ID 1577290.1)

Using Workload Capture and Replay (文档 ID 445116.1)

操作系统性能监控


Linux Performance Tools
Various, observability:
Various, static:
Various, tracing:
sar /proc dstat dmesg
sysctl /sys dmesg lshw journalctl lsmod
capable
strace perf trace sysdig
opensnoop statsnoop syncsnoop
tcptop tcplife tcpconnect tcpaccept tcpconnlat tcpretrans
mysqld_qslower, ...
gethostlatency
App Config
ugc ucalls
ldd
ltrace
Operating System
cachestat dcsnoop filetop fileslower mountsnoop
Applications
System Libraries
System Call Interface
cpudist execsnoop runlat cpudist offcputime
Hardware
intel_gpu_top intel_gpu_time
intel_gpu_\ frequency
perf ftrace stap lttng bcc (BPF)
ext4slower ext4dist btrfs* xfs* zfs*
latencytop
schedtool
GPU
/proc/cpuinfo cpuid lscpu
lsof pcstat
VFS
Sockets
Scheduler
mpstat powertop /proc/stat
turbostat rdmsr
funccount funclatency stackcount kprobe uprobe argdist trace profile
df -h
File Systems
TCP/UDP
top htop ps pidstat
Linux Kernel
mdadm lvm dmsetup
Volume Manager
IP
Virtual Memory
vmstat slabtop free /proc/meminfo
CPU 1
CPU Interconnect
pidstat -d
Block Device Int.
Ethernet
Clocksource
memleak oomkill slabratetop
numactl
tiptop perf pcm
multipath
Device Drivers
Memory Bus
mdflush
Firmware
/sys/...
llcstat profile
DRAM
perf tiptop
I/O Bus
biotop biosnoop biolatency bitesize
ss
tcpdump
ip route iptables
netstat iptraf-ng
dmesg
hardirqs ttysnoop
storcli MegaCli
numastat
lstopo
Expander Interconnect
I/O Bridge
iostat iotop blktrace
I/O Controller
lspci lsusb
Interface Transports
Network Controller
nicstat netstat ip
Disk
Disk
Swap
Port
Port
FAN
Power Supply
lsblk lsscsi blockdev smartctl fdisk -l
swapon -s /proc/swaps
ethtool ifconfig
snmpget lldptool
sar -m FAN
ipmitool dmidecode
observability tools
static performance tools
perf-tools/bcc tracing tools
these can observe the state of the system at rest, without load
https://github.com/brendangregg/perf-tools https://github.com/iovisor/bcc
style inspired by reddit.com/u/redct
http://www.brendangregg.com/linuxperf.html 2017

谢谢！