了解IO设备

核心系统数据库组 余锋http://yufeng.info

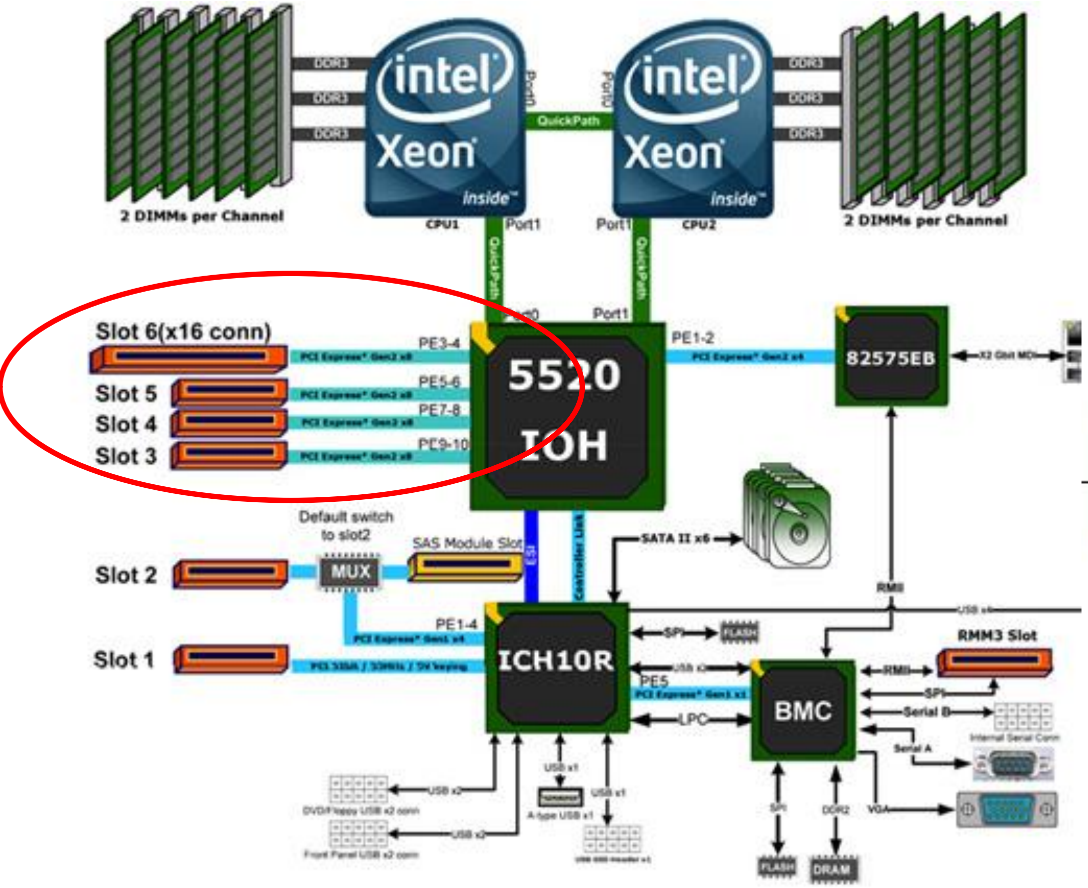
@淘宝褚霸

2012-03-17

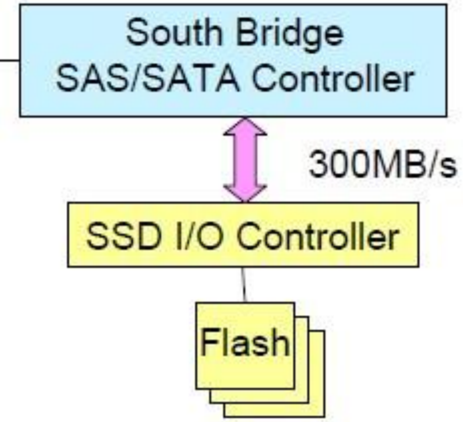
提纲

- 芯片组
- SATA/SAS
- SSD
- PCIe Flash卡
- RAID卡
- NVRAM卡
- 测量工具

IO芯片组



Raid卡和PCIe卡都插在PCIe卡，直接走北桥



South Bridge
SAS/SATA Controller

300MB/s

SSD I/O Controller

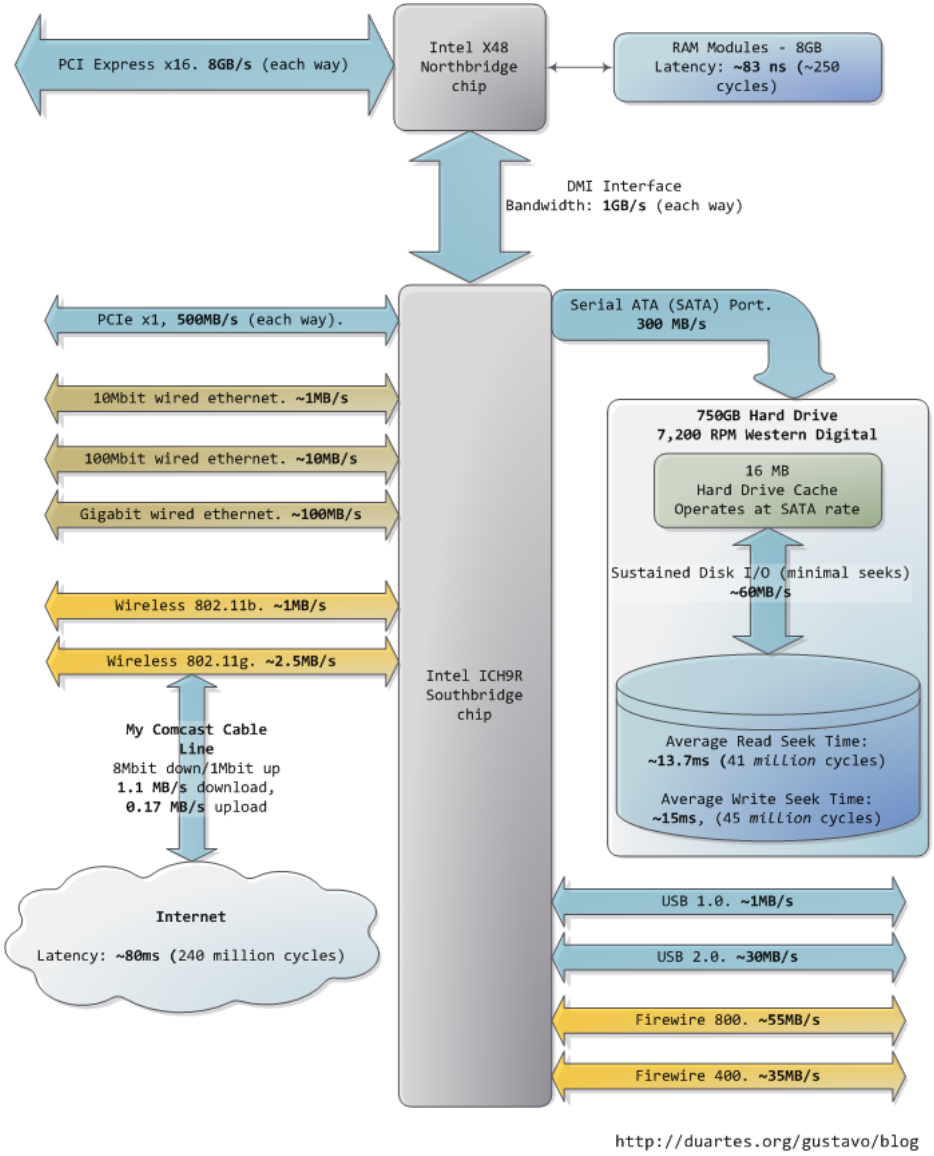
Flash

芯片组型号

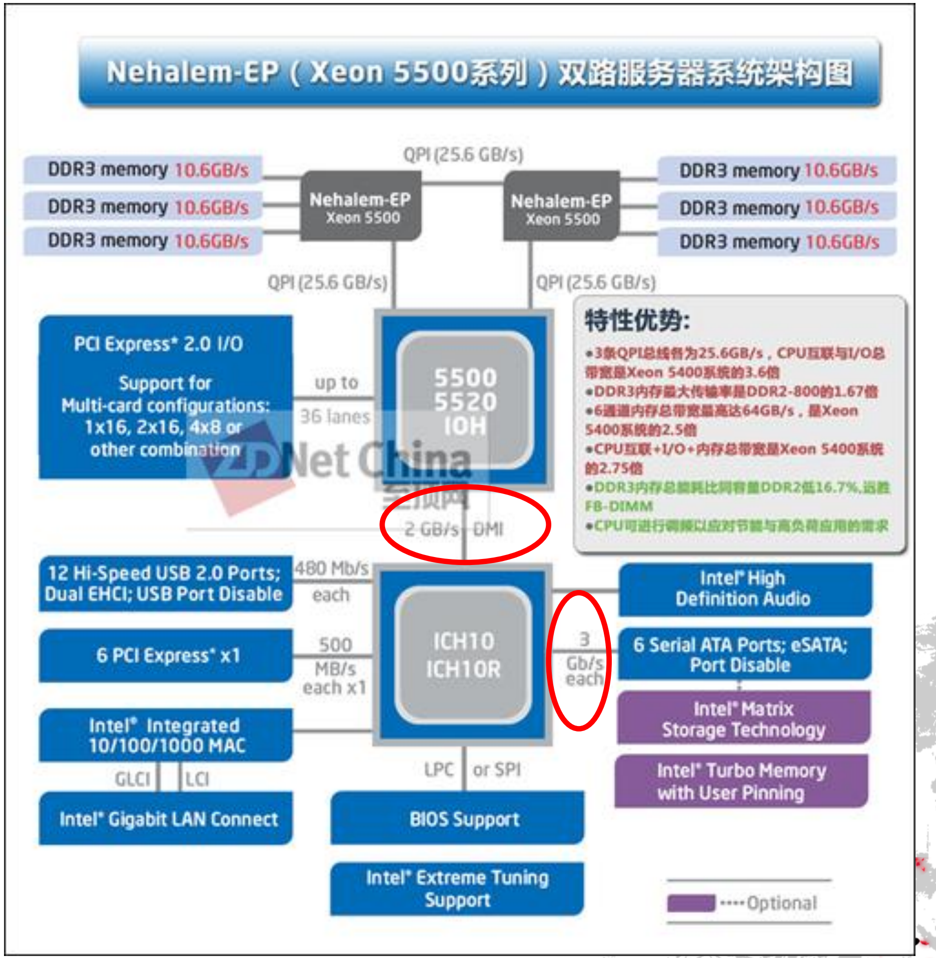
Processors: 2 x Xeon E5645 2.40GHz 5860MHz FSB (HT enabled, 12 cores, 24 threads)

Chipset: Intel 5500 IOH-24D B3 (Tylersburg), 82801JIR A0 (ICH10R)

接口速率


PCI Express x16. 8GB/s (each way)
Intel X48 Northbridge chip
RAM Modules - 8GB Latency: ~83 ns (~250 cycles)
DMI Interface Bandwidth: 1GB/s (each way)
PCIe x1, 500MB/s (each way).
Serial ATA (SATA) Port. 300 MB/s
10Mbit wired ethernet. ~1MB/s
100Mbit wired ethernet. ~10MB/s
Gigabit wired ethernet. ~100MB/s
750GB Hard Drive 7,200 RPM Western Digital
16 MB Hard Drive Cache Operates at SATA rate
Sustained Disk I/O (minimal seeks) ~60MB/s
Wireless 802.11b. ~1MB/s
Wireless 802.11g. ~2.5MB/s
Intel ICH9R Southbridge chip
My Comcast Cable Line
8Mbit down/1Mbit up
1.1 MB/s download,
0.17 MB/s upload
Average Read Seek Time: ~13.7ms (41 million cycles)
Average Write Seek Time: ~15ms, (45 million cycles)
Internet
Latency: ~80ms (240 million cycles)
USB 1.0. ~1MB/s
USB 2.0. ~30MB/s
Firewire 800. ~55MB/s
Firewire 400. ~35MB/s

http://duartes.org/gustavo/blog


Nehalem-EP（Xeon 5500系列）双路服务器系统架构图
QPI (25.6 GB/s)
DDR3 memory 10.6GB/s
DDR3 memory 10.6GB/s
DDR3 memory 10.6GB/s
Nehalem-EP Xeon 5500
Nehalem-EP Xeon 5500
DDR3 memory 10.6GB/s
DDR3 memory 10.6GB/s
DDR3 memory 10.6GB/s
QPI (25.6 GB/s)
QPI (25.6 GB/s)
PCI Express* 2.0 I/O
Support for Multi-card configurations: 1x16, 2x16, 4x8 or other combination
up to 36 lanes
5500 5520 IOH
2 GB/s DMI
12 Hi-Speed USB 2.0 Ports; Dual EHCI; USB Port Disable
480 Mb/s each
6 PCI Express* x1
500 MB/s each x1
ICH10 ICH10R
3 Gb/s each
Intel® High Definition Audio
6 Serial ATA Ports; eSATA; Port Disable
Intel® Matrix Storage Technology
Intel® Turbo Memory with User Pinning
Intel® Integrated 10/100/1000 MAC
GLCI
LCI
LPC or SPI
Intel® Gigabit LAN Connect
BIOS Support
Intel® Extreme Tuning Support
Optional
ZDNet China

特性优势:

- 3条QPI总线各为25.6GB/s，CPU互联与I/O总带宽是Xeon 5400系统的3.6倍
- DDR3内存最大传输率是DDR2-800的1.67倍
- 6通道内存总带宽最高达64GB/s，是Xeon 5400系统的2.5倍
- CPU互联+I/O+内存总带宽是Xeon 5400系统的2.75倍
- DDR3内存总能耗比同容量DDR2低16.7%,远胜FB-DIMM
- CPU可进行调频以应对节能与高负荷应用的需求

PCIe每个X接口速率:

v1.x: 250 MB/s

v2.x: 500 MB/s

SATA/SAS机械磁盘

SATA II
7200 RPM IOPS: ~90

SAS
15K RPM IOPS: ~180

Disk: sda (scsi0): 100GB JBOD == 1 x HITACHI-HUSSL4010ASS600

SSD

寿命:200T

SATA II
Intel X25-M IOPS: ~8600

思考

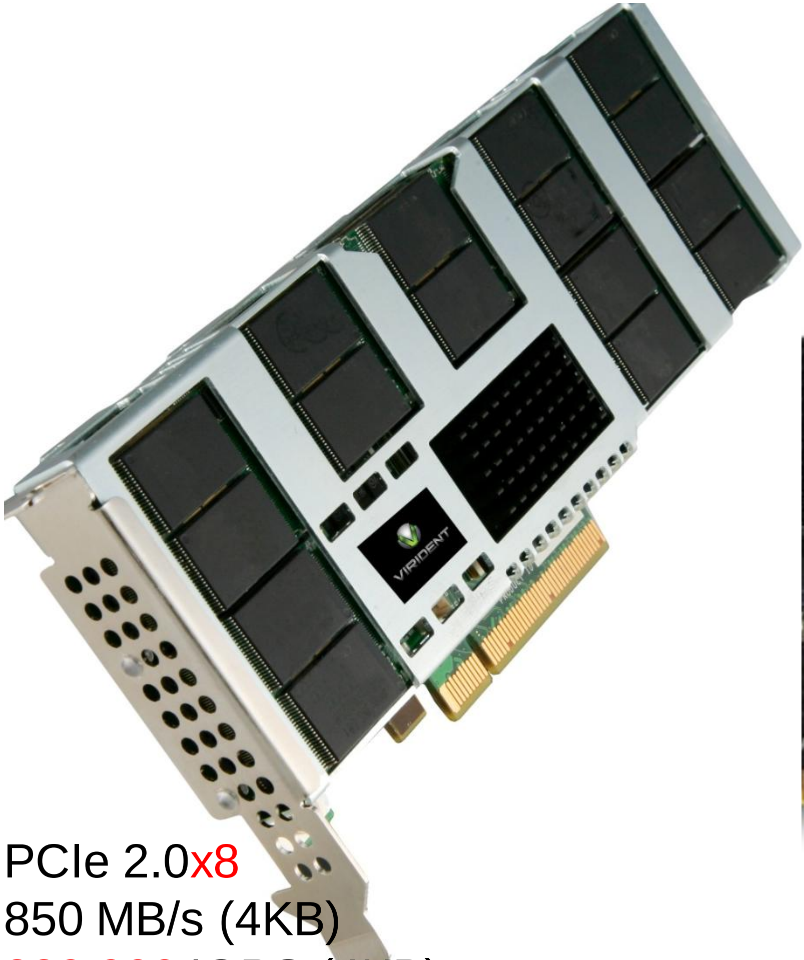
为什么

要有RAID或者HBA卡

接SATA磁盘阵列？

解决什么问题？

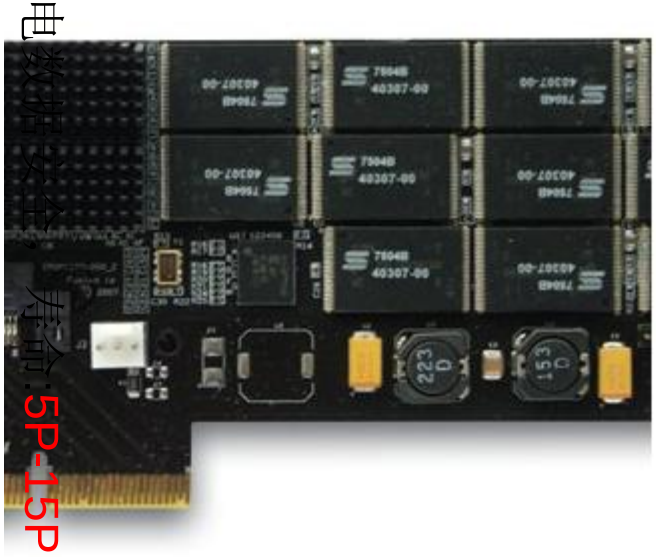
PCIe Flash卡



PCIe 2.0x8
850 MB/s (4KB)
220,000 IOPS (4KB)

PCIe 2.0x4
ioDrive IOPS: with Flash 140,000
Read IOPS, 135,000 Write IOPS

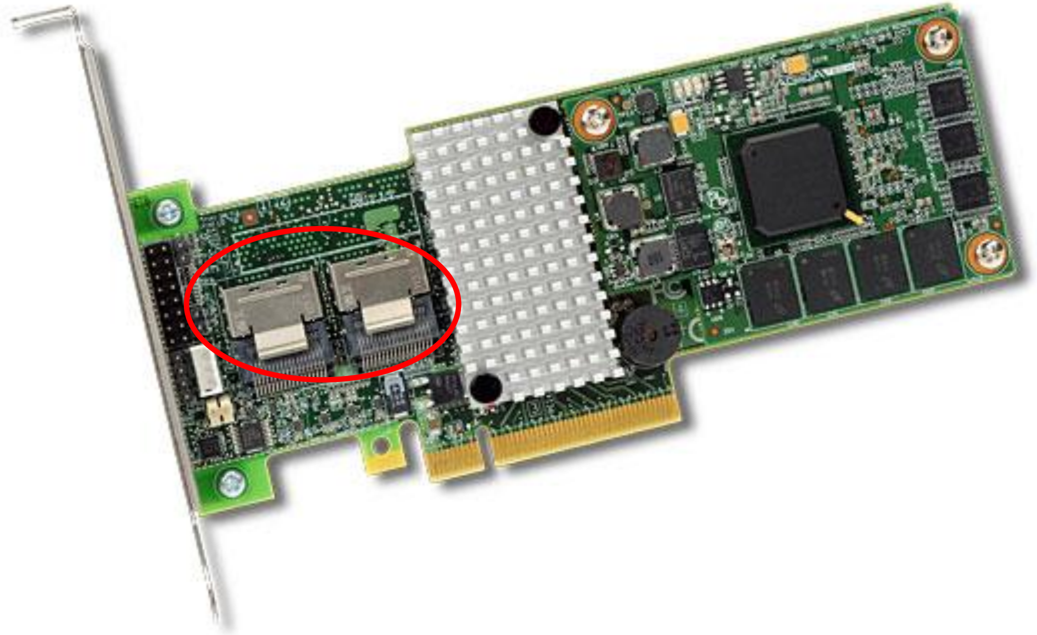
掉电数据安全，寿命:5P-15P



Disk-Control: iodrive0: Fusion-io ioDIMM3 320GB

Raid卡

- PCIe 2.0x8
- Support Up to 128 SATA Devices
- Dual Core ROC
- 1 GB cache



Disk-Control: megaraid_sas0: LSI Logic / Symbios Logic MegaRAID SAS 1078

Raid卡(续)-Cache

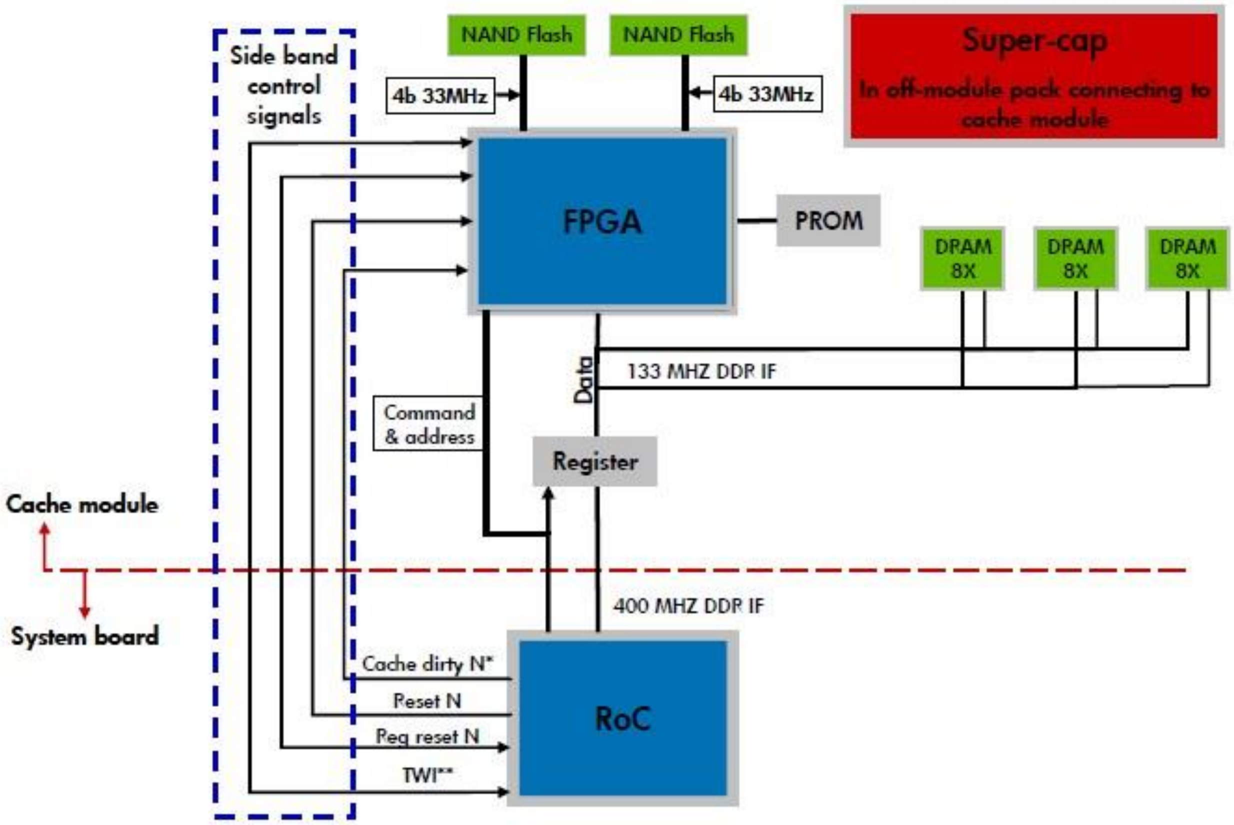
- **虚拟卷**
- **预读缓存**
  - NORA (No read ahead)
  - RA (Read ahead)
  - ADRA (Adaptive read ahead)
- **写缓存**
  - WT (Write through),
  - WB (Write back)
- **Disk Cache**
  - 关闭，考虑到数据安全

Raid卡(续)-BBWC

- **Nickel Metal Hydride (NiMH)**
  - 100 full discharge cycles.
  - 48-hour battery life .
  - Typical capacity for the HP Smart Array battery pack reduces by 5 to 10 percent over a 3-year period.
  - Battery recharge takes between 30 minutes and 2 hours
- **模块化设计可替换**

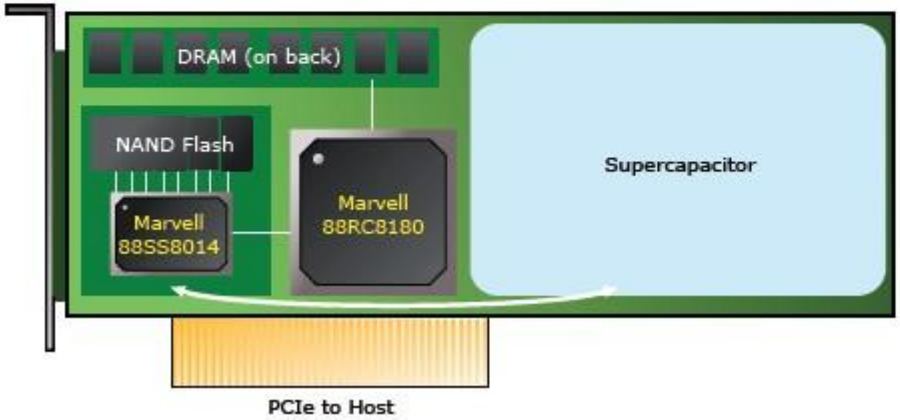
Raid卡 (续) - FBWC

Figure 1. FBWC block diagram



NVRAM卡

寿命:1M hours



DDR backup to persistent flash on powerfailure
Automatic restore from Flash to DDR when power is restored

PCIe 1.1x4
4K Block Writes: 165,000 IOPS
4K Block Reads: 185,000 IOPS

Disk-Control: mvloki0: Marvell Device 8180

DDR3 Non-Volatile DIMM

```
static unsigned long ram_start=0xa40000000UL;
static unsigned long ram_size= 0x80000000UL;
```

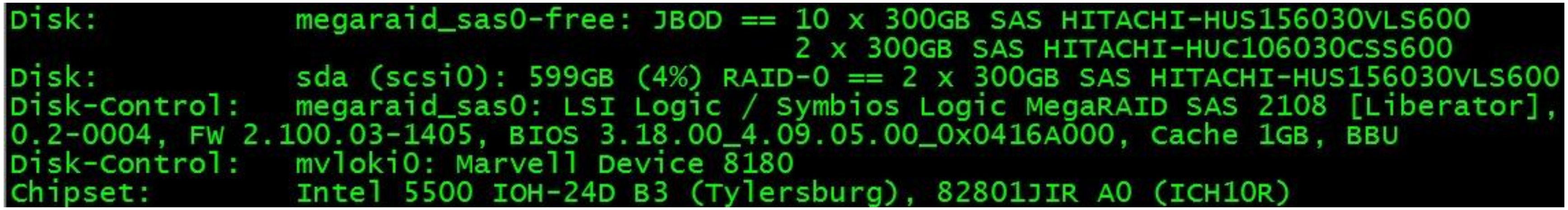
思考

PCIe卡的寿命和安全如何保证？

掉电数据安全吗？

hwconfig

hwconfig



```
Disk:           megaraid_sas0-free: JBOD == 10 x 300GB SAS HITACHI-HUS156030VLS600
                                            2 x 300GB SAS HITACHI-HUC106030CSS600
Disk:           sda (scsi0): 599GB (4%) RAID-0 == 2 x 300GB SAS HITACHI-HUS156030VLS600
Disk-Control:   megaraid_sas0: LSI Logic / Symbios Logic MegaRAID SAS 2108 [Liberator],
0.2-0004, FW 2.100.03-1405, BIOS 3.18.00_4.09.05.00_0x0416A000, Cache 1GB, BBU
Disk-Control:   mvloki0: Marvell Device 8180
Chipset:        Intel 5500 IOH-24D B3 (Tylersburg), 82801JIR A0 (ICH10R)
```

hwconfig –x sample.cfg

firmware="E516"
handle="69"
interface="SAS"
serial="JXYGHLAN"
size="299999690752"
status="free"
volume="megaraid_sas0-free"
volume_handle="74"
wwn="0x5000cca018c378f1”
model="HITACHI-HUS156030VLS600"

IO子系统架构图

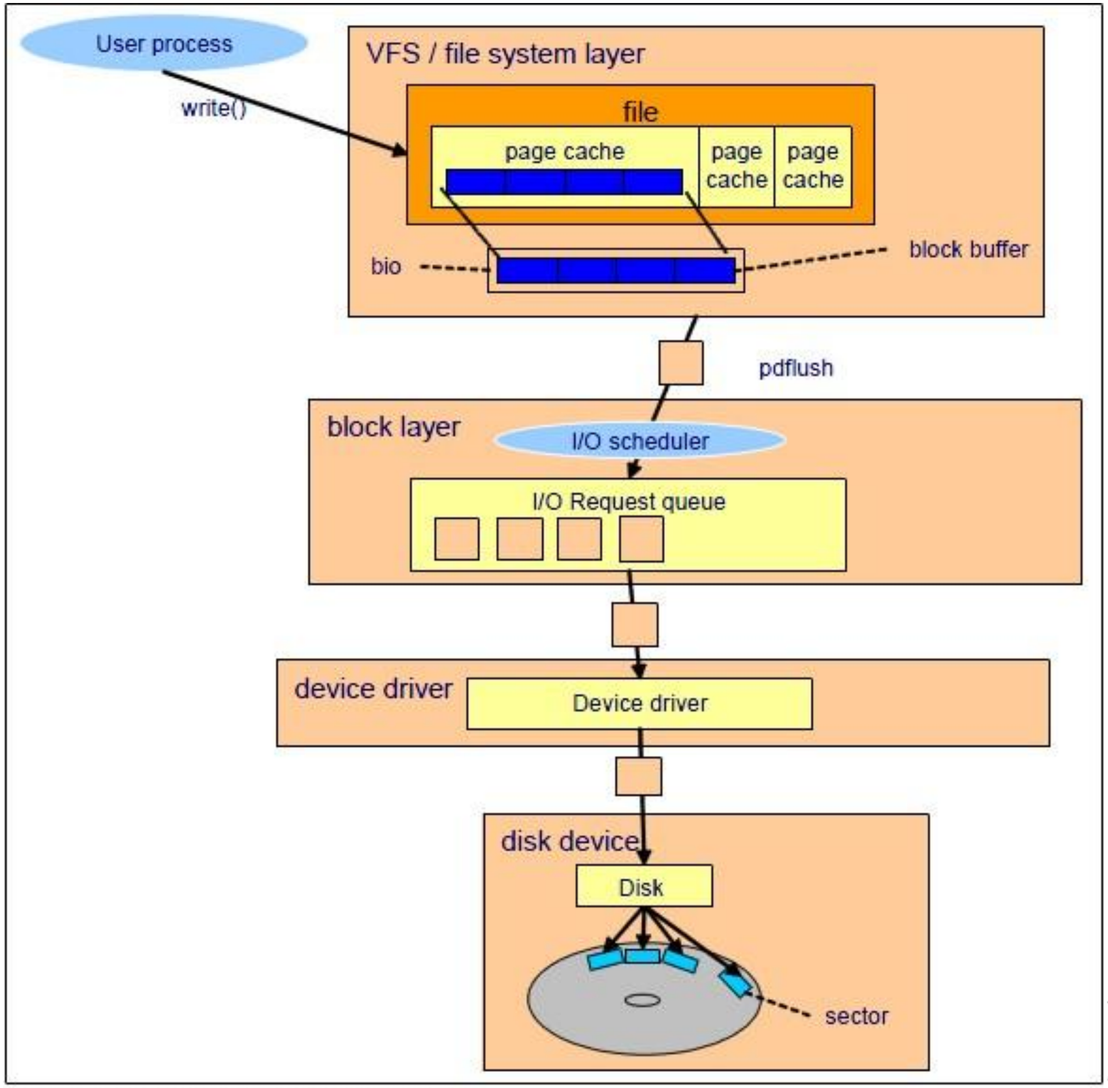


Figure 1-18 I/O subsystem architecture

lsblk

```
# lsblk -i
NAME    MAJ:MIN RM   SIZE RO MOUNTPOINT
sda       8:0    0 557.8G  0
sda1      8:1    0   500M  0 /boot
sda2      8:2    0 146.5G  0 /
sda3      8:3    0     2G  0 [SWAP]
sda4      8:4    0     1K  0
sda5      8:5    0 408.8G  0 /disk0
nvdisk0 252:0    0     8G  0 /u05
```

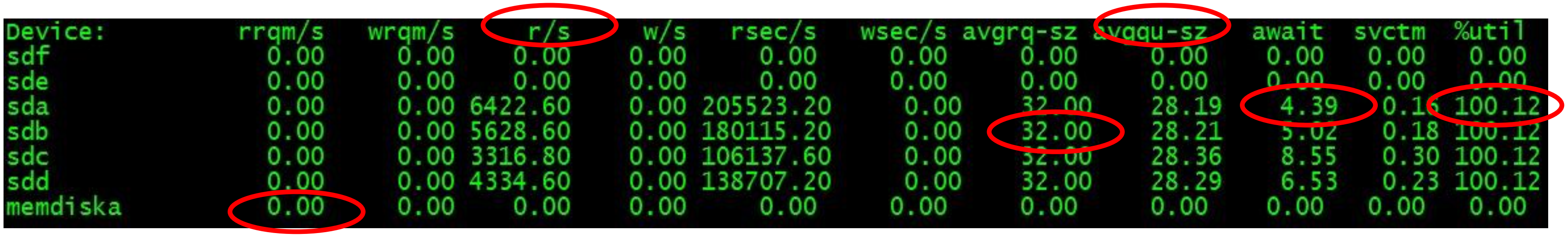
fio设备写饱和脚本

$ fio a_b_c_d_test

```
[global]
bs=4K
ioengine=libaio
rw=randrw
rwmixwrite=100
time_based
runtime=3600
direct=1
group_reporting
randrepeat=0
norandommap
invalidate=1
iodepth=8
iodepth_batch=4
iodepth_low=4
iodepth_batch_complete=8
numjobs=1

[test_sda]
filename=/dev/sda
[test_sdb]
filename=/dev/sdb
[test_sdc]
filename=/dev/sdc
[test_sdd]
filename=/dev/sdd
```

iostat



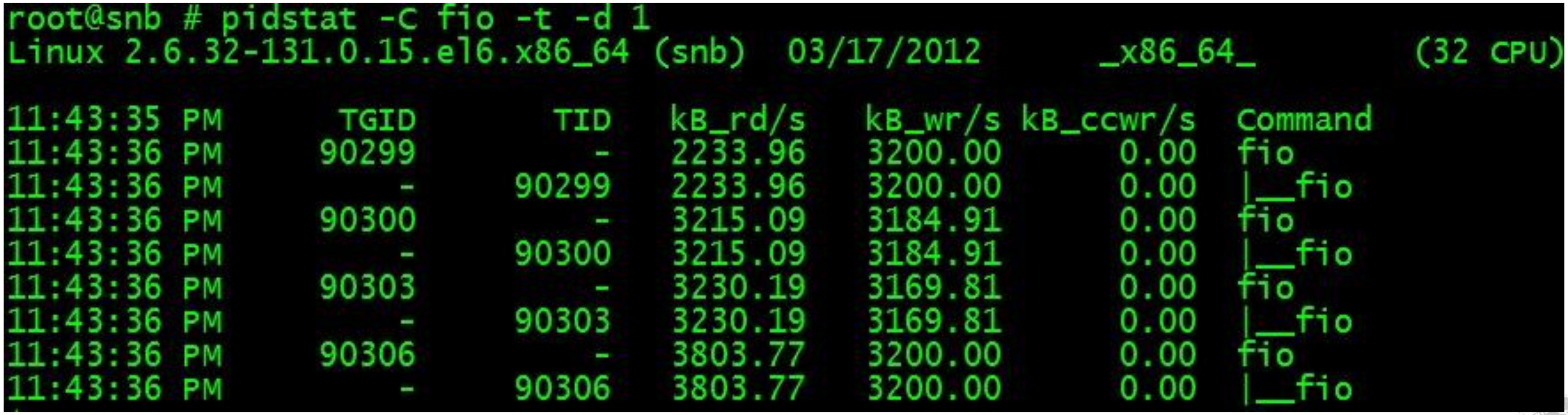
```
Device:         rrqm/s   wrqm/s     r/s     w/s    rsec/s   wsec/s avgrq-sz avgqu-sz   await  svctm  %util
sdf               0.00     0.00    0.00    0.00      0.00     0.00     0.00     0.00    0.00   0.00   0.00
sde               0.00     0.00    0.00    0.00      0.00     0.00     0.00     0.00    0.00   0.00   0.00
sda               0.00     0.00 6422.60    0.00 205523.20     0.00    32.00    28.19    4.39   0.16 100.12
sdb               0.00     0.00 5628.60    0.00 180115.20     0.00    32.00    28.21    5.02   0.18 100.12
sdc               0.00     0.00 3316.80    0.00 106137.60     0.00    32.00    28.36    8.55   0.30 100.12
sdd               0.00     0.00 4334.60    0.00 138707.20     0.00    32.00    28.29    6.53   0.23 100.12
memdiska          0.00     0.00    0.00    0.00      0.00     0.00     0.00     0.00    0.00   0.00   0.00
```

思考

IO depth对设备性能

有什么样的影响？

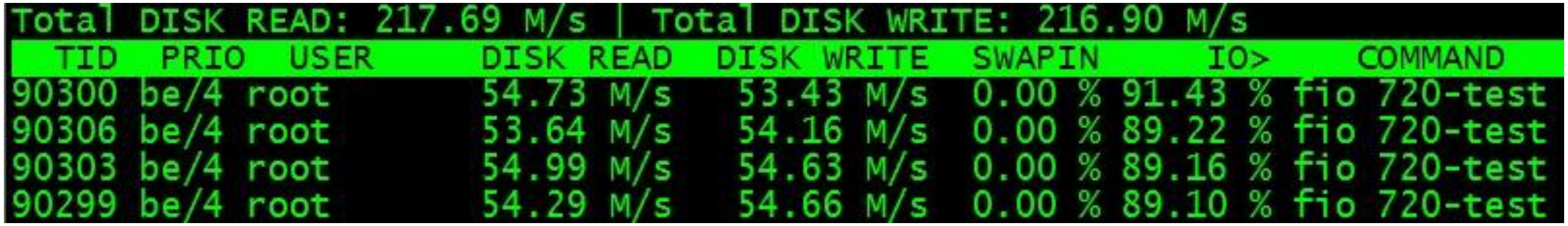
pidstat



```
root@snb # pidstat -C fio -t -d 1
Linux 2.6.32-131.0.15.el6.x86_64 (snb)  03/17/2012      _x86_64_        (32 CPU)

11:43:35 PM      TGID       TID   kB_rd/s   kB_wr/s kB_ccwr/s  Command
11:43:36 PM     90299         -   2233.96   3200.00      0.00  fio
11:43:36 PM         -     90299   2233.96   3200.00      0.00  |__fio
11:43:36 PM     90300         -   3215.09   3184.91      0.00  fio
11:43:36 PM         -     90300   3215.09   3184.91      0.00  |__fio
11:43:36 PM     90303         -   3230.19   3169.81      0.00  fio
11:43:36 PM         -     90303   3230.19   3169.81      0.00  |__fio
11:43:36 PM     90306         -   3803.77   3200.00      0.00  fio
11:43:36 PM         -     90306   3803.77   3200.00      0.00  |__fio
```

iotop



```
Total DISK READ: 217.69 M/s | Total DISK WRITE: 216.90 M/s
  TID  PRIO  USER     DISK READ  DISK WRITE  SWAPIN      IO>    COMMAND
90300 be/4 root      54.73 M/s   53.43 M/s  0.00 % 91.43 % fio 720-test
90306 be/4 root      53.64 M/s   54.16 M/s  0.00 % 89.22 % fio 720-test
90303 be/4 root      54.99 M/s   54.63 M/s  0.00 % 89.16 % fio 720-test
90299 be/4 root      54.29 M/s   54.66 M/s  0.00 % 89.10 % fio 720-test
```

参考材料

- **Fio测试工具使用：**

  http://blog.yufeng.info/archives/tag/fio

- **hwconfig查看硬件信息:**

  http://blog.yufeng.info/archives/2086

- **Linux下方便的块设备查看工具lsblk**

  http://blog.yufeng.info/archives/1882

- **Linux TASK_IO_ACCOUNTING功能以及如何使用:**

  http://blog.yufeng.info/archives/2138

提问时间

谢谢大家！