What Every Data Programmer Needs to Know about Disks

OSCON Data – July, 2011 - Portland

Ted Dziuba
@dozba
tjdziuba@gmail.com

Who are you and why are you talking?

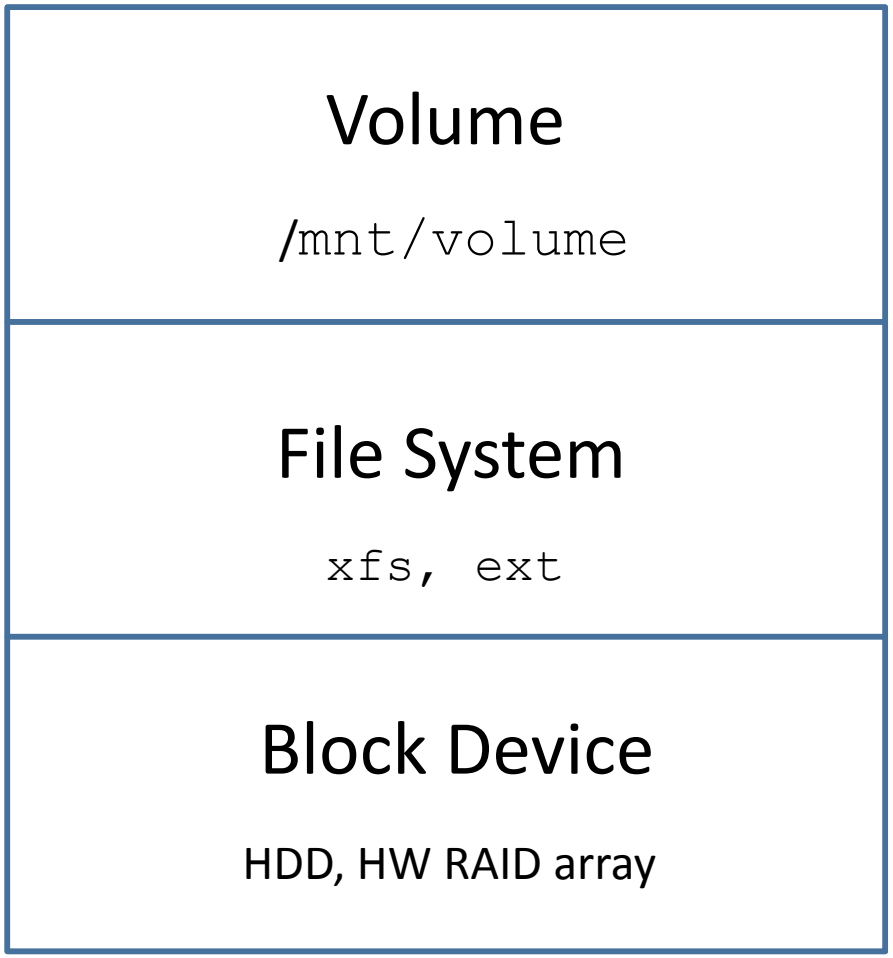
First job: Like college but they pay you to go.

A few years ago: Technical troll for The Register.

Recently: Co-founder of Milo.com, local shopping engine.

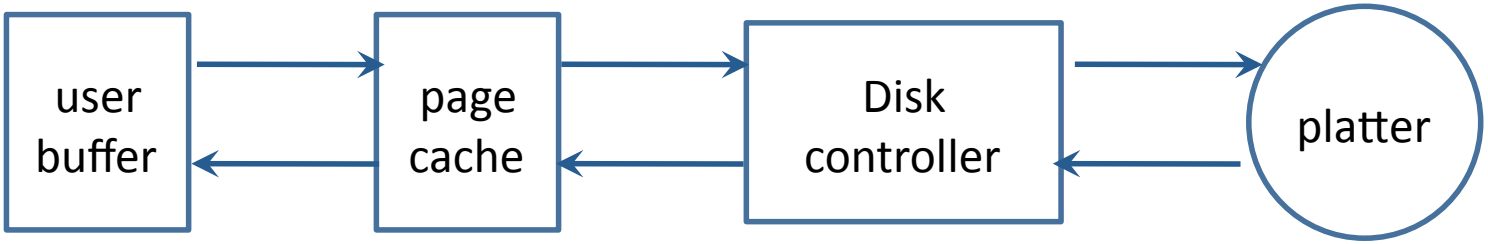
Present: Senior Technical Staff for eBay Local

The Linux Disk Abstraction

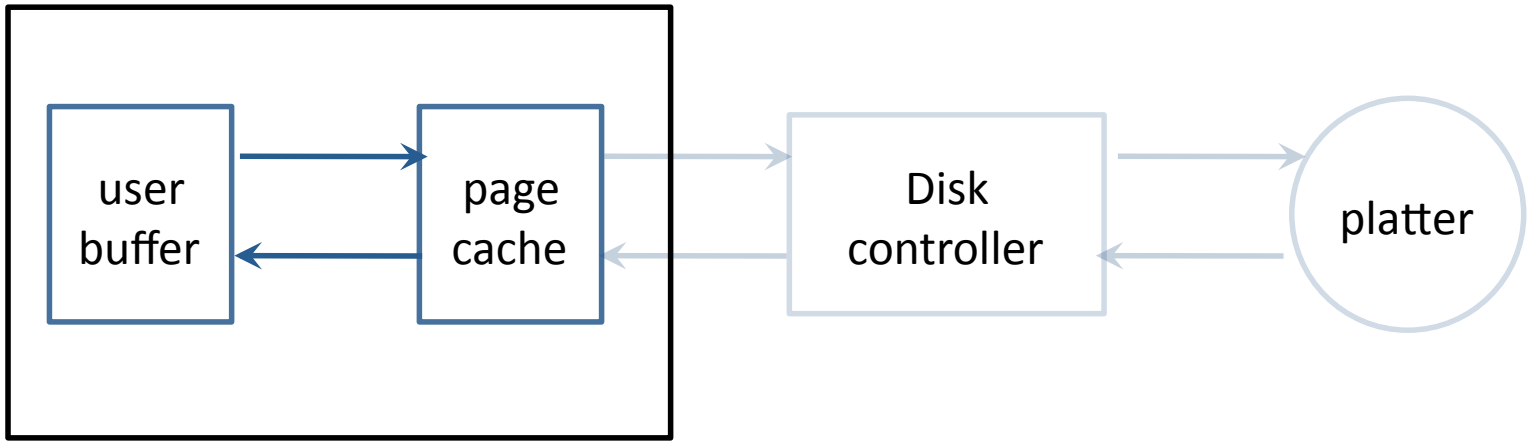


What happens when you read from a file?

```
f = open("/home/ted/not_pirated_movie.avi", "rb")
avi_header = f.read(56)
f.close()
```

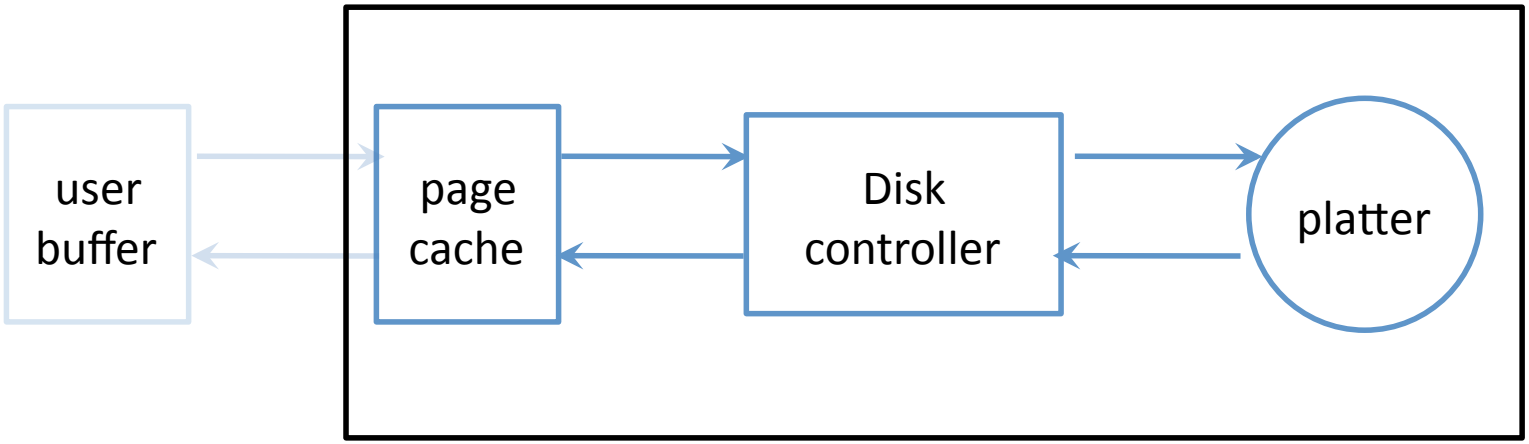


What happens when you read from a file?



- Main memory lookup
- Latency: 100 nanoseconds
- Throughput: 12GB/sec on good hardware

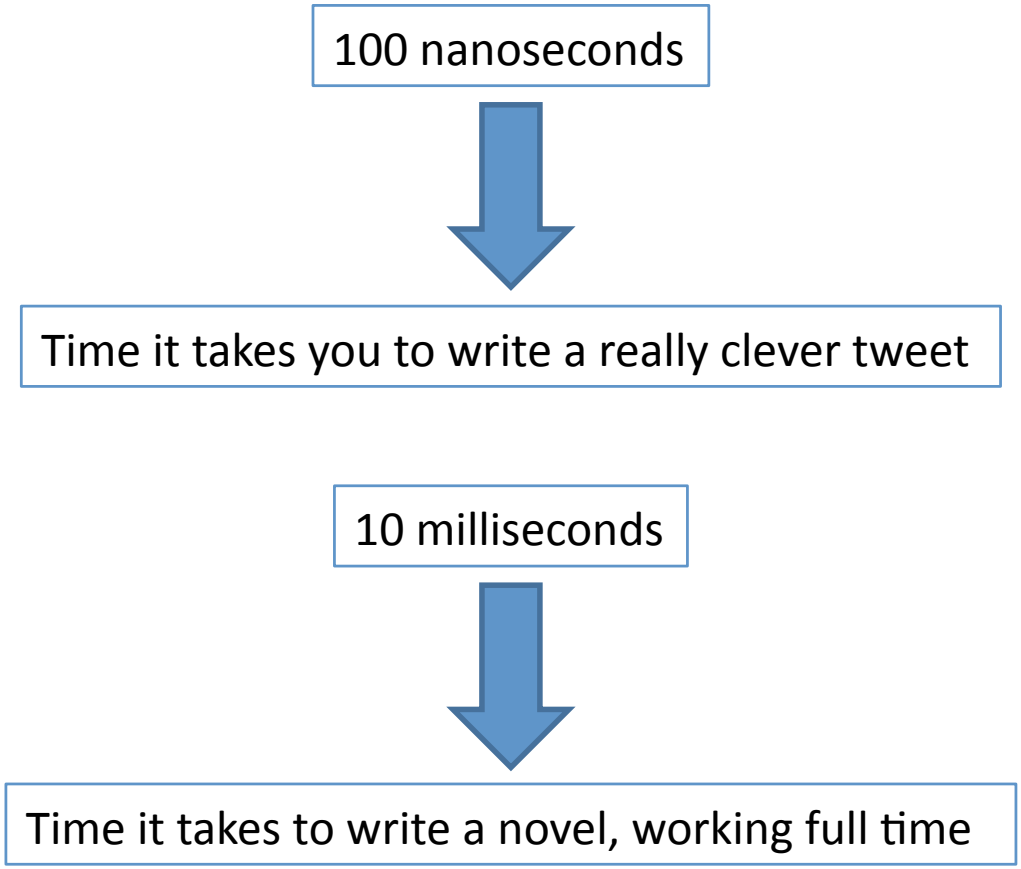
What happens when you read from a file?



- Needs to actuate a physical device
- Latency: 10 milliseconds
- Throughput: 768 MB/sec on SATA 3
- *(Faster if you have a lot of money)*

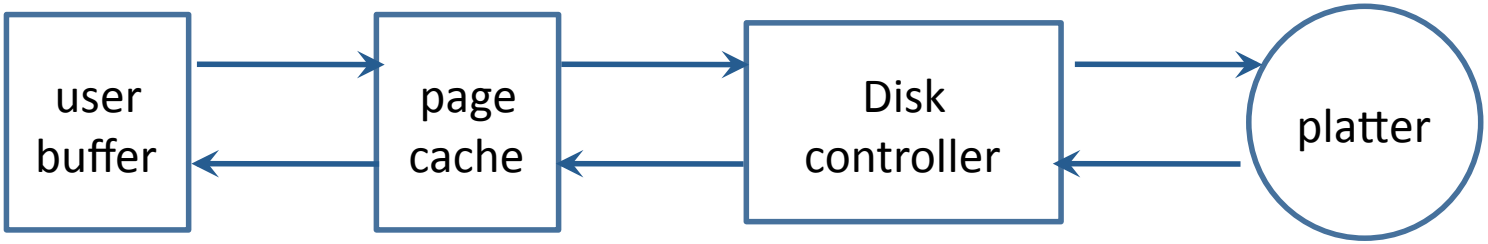
Sidebar: The Horror of a 10ms Seek Latency

A disk read is 100,000 times slower than a memory read.



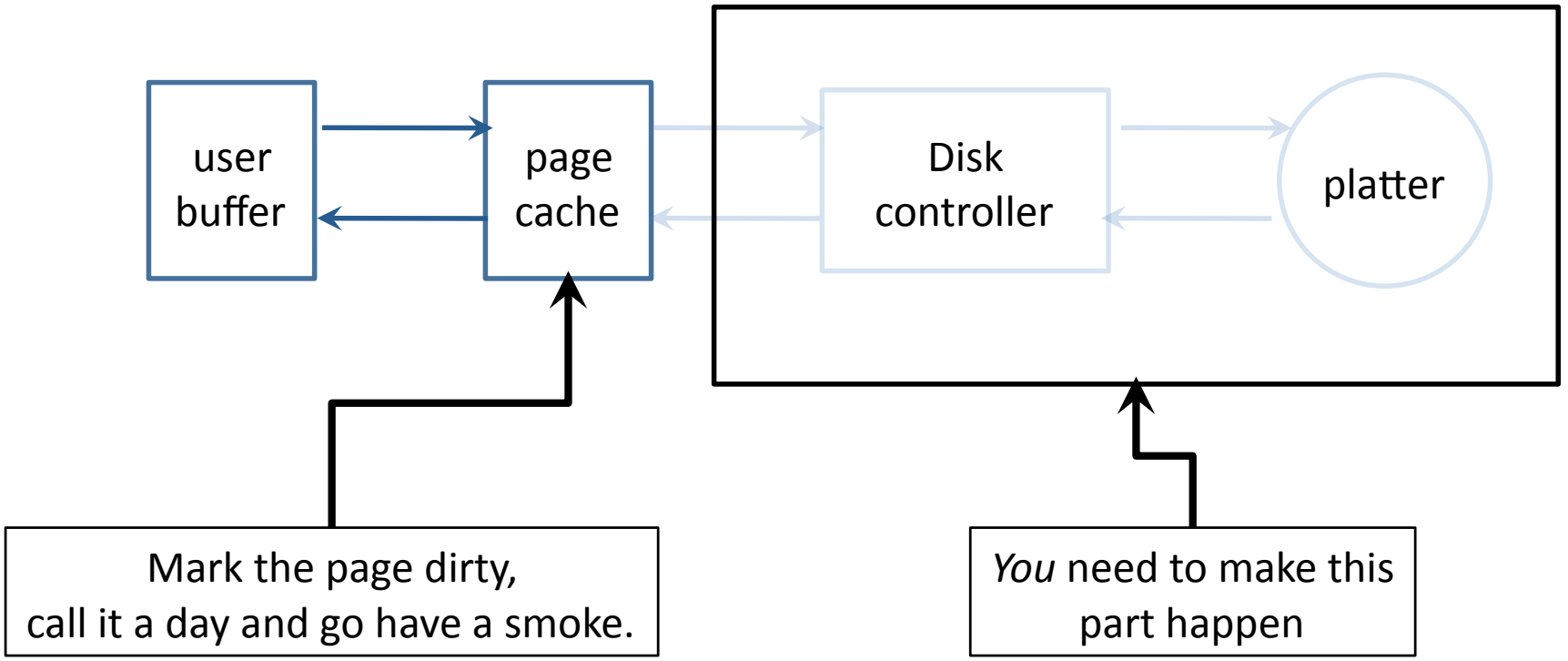
What happens when you write to a file?

```
f = open("/home/ted/nosql_database.csv", "wb")
f.write(key)
f.write(",")
f.write(value)
f.close()
```



What happens when you write to a file?

```
f = open("/home/ted/nosql_database.csv", "wb")
f.write(key)
f.write(",")
f.write(value)
f.close()
```



Aside: Stick your finger in the Linux Page Cache

Pre-Linux 2.6 used “pdflush”, now per-Backing Device Info (BDI) flush threads

Dirty pages: `grep –i “dirty” /proc/meminfo`

`/proc/sys/vm` Love:

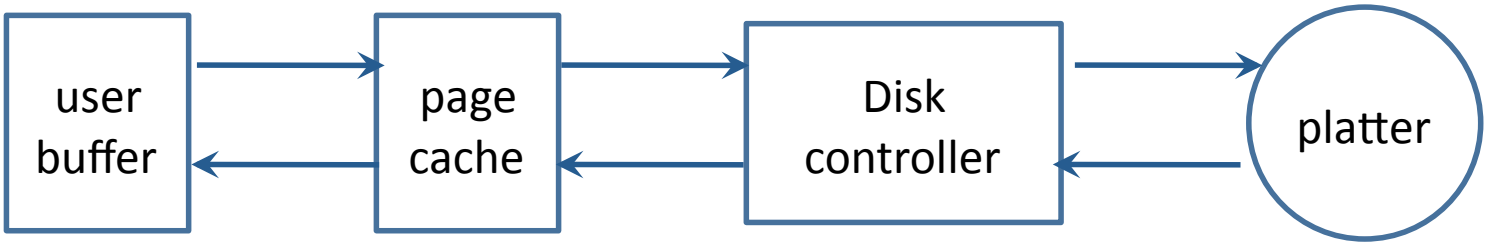
- `dirty_expire_centisecs` : flush old dirty pages
- `dirty_ratio`: flush after some percent of memory is used
- `dirty_writeback_centisecs`: how often to wake up and start flushing

Clear your page cache : `echo 1 > /proc/sys/vm/drop_caches`

Crusty sysadmin’s hail-Mary pass: `sync; sync; sync`

Fsync: force a flush to disk

```
f = open("/home/ted/nosql_database.csv", "wb")
f.write(key)
f.write(",")
f.write(value)
os.fsync(f.fileno())
f.close()
```



Also note, fsync() has a cousin, fdatasync() that does not sync metadata.

Aside: point and laugh at MongoDB

Mongo's "fsync" command:

```
> db.runCommand({fsync:1,async:true});
```

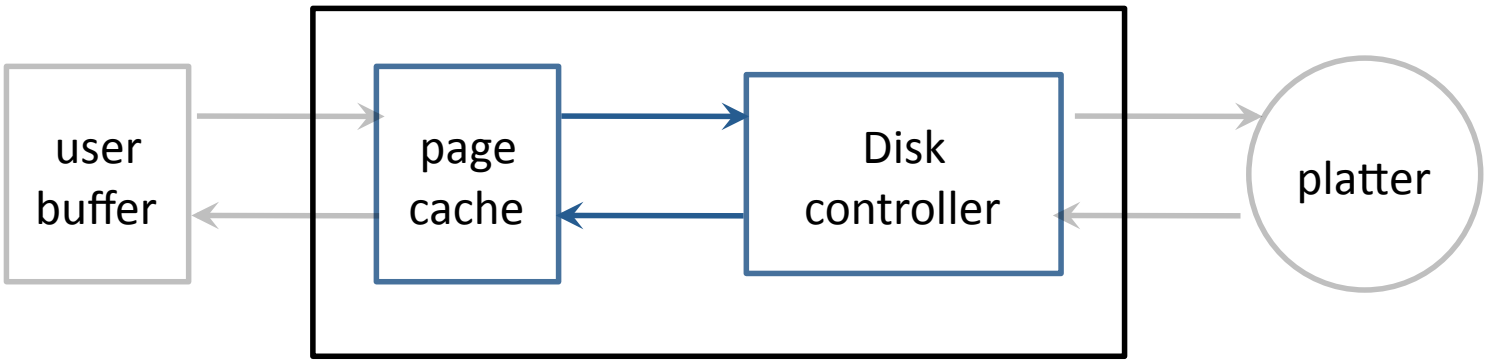
wat.

Also supports "journaling", like a WAL in the SQL world, however…

- It only fsyncs() the journal every 100ms…"for performance".
- It's not enabled by default.

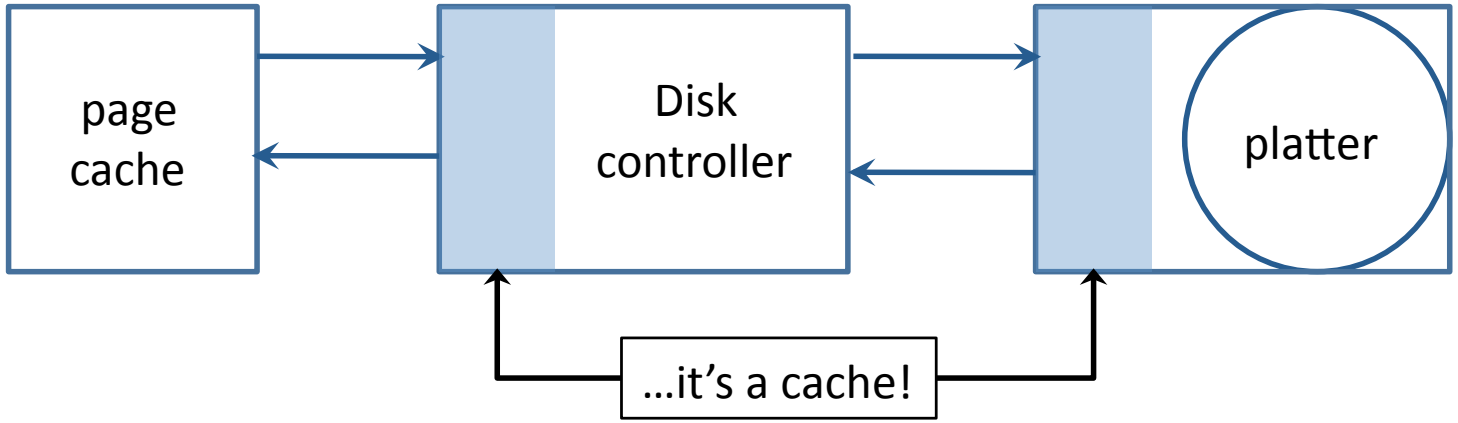
Fsync: bitter lies

```
f = open("/home/ted/nosql_database.csv", "wb")
f.write(key)
f.write(",")
f.write(value)
os.fsync(f.fileno())
f.close()
```



Drives will lie to you.

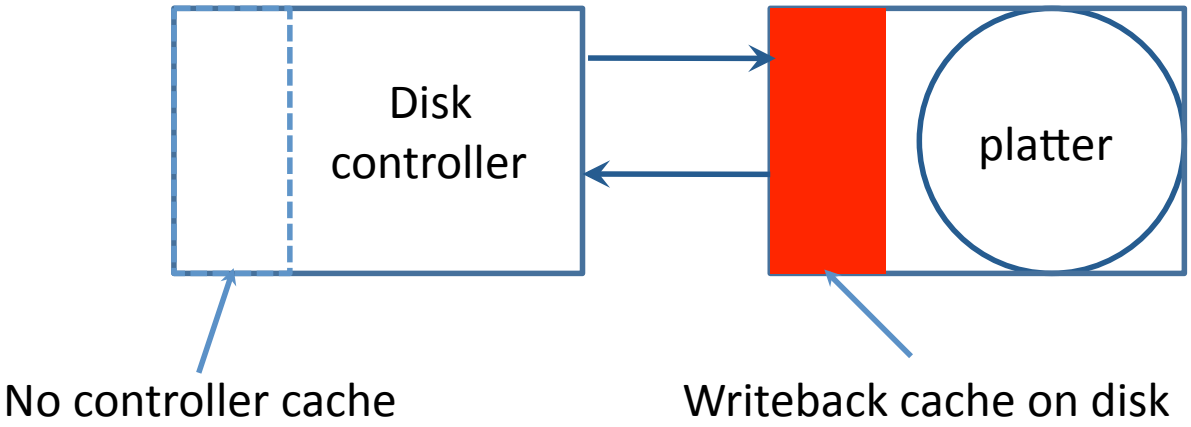
Fsync: bitter lies



- Two types of caches: write*through* and write*back*
- Write*back* is the demon

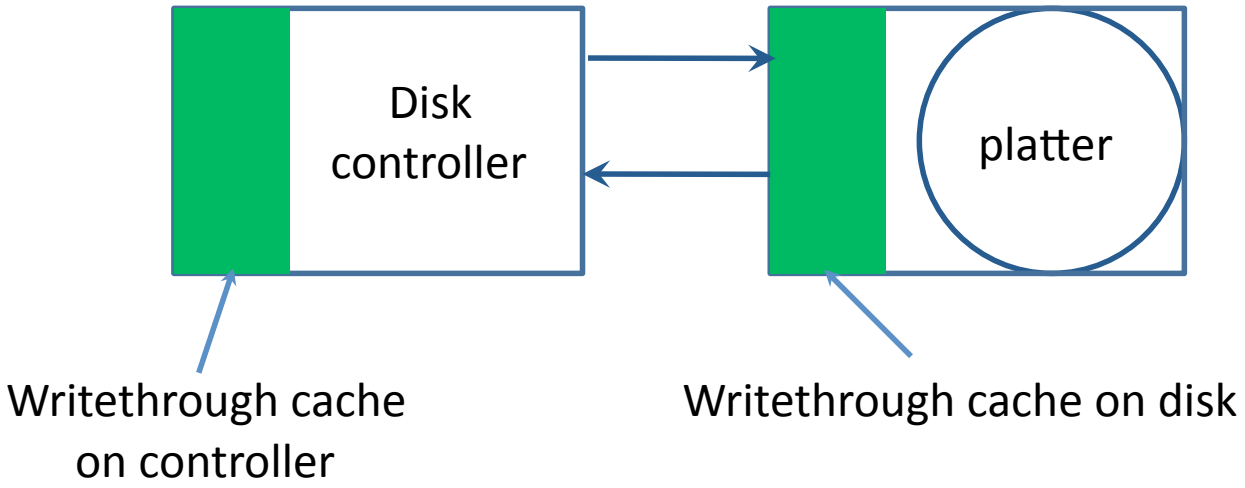
(Just dropped in) to see what condition your caches are in

A Typical Workstation



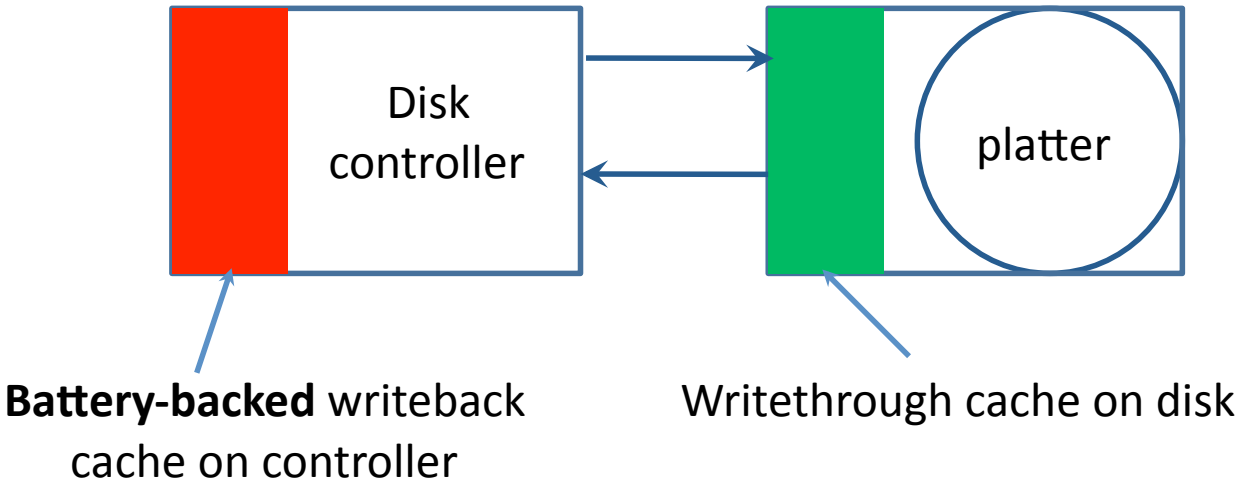
(Just dropped in) to see what condition your caches are in

A Good Server



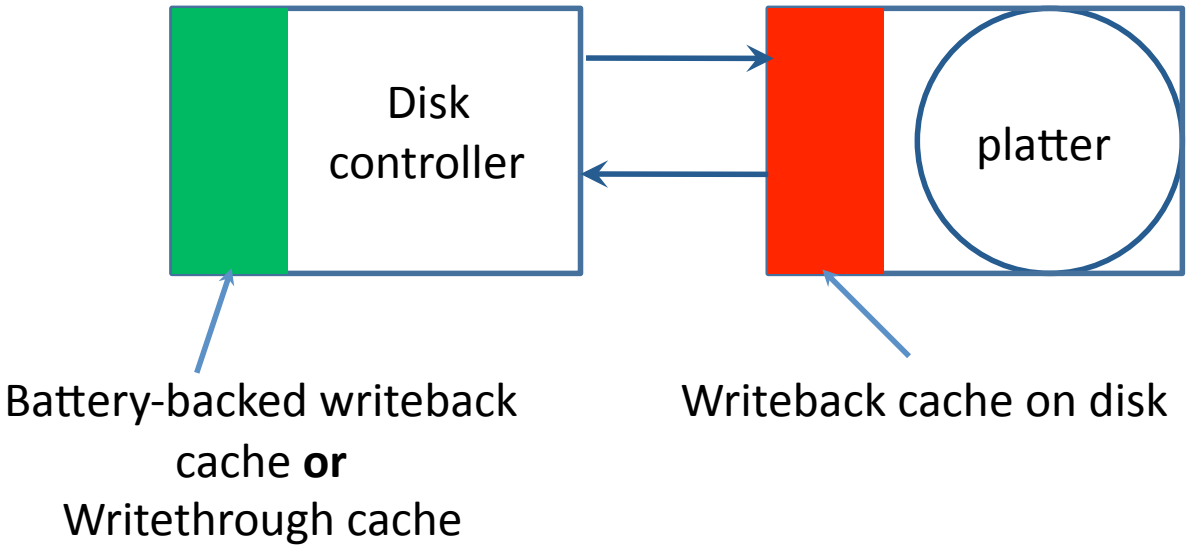
(Just dropped in) to see what condition your caches are in

An Even Better Server



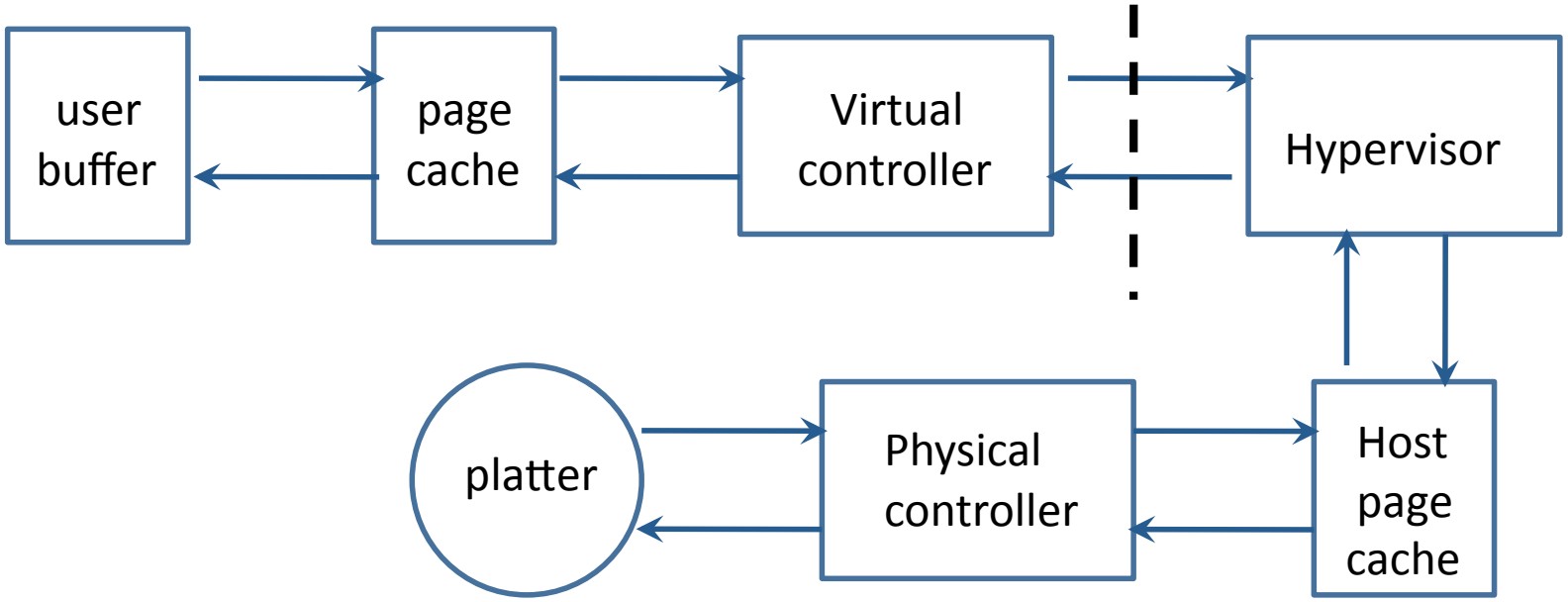
(Just dropped in) to see what condition your caches are in

The Demon Setup



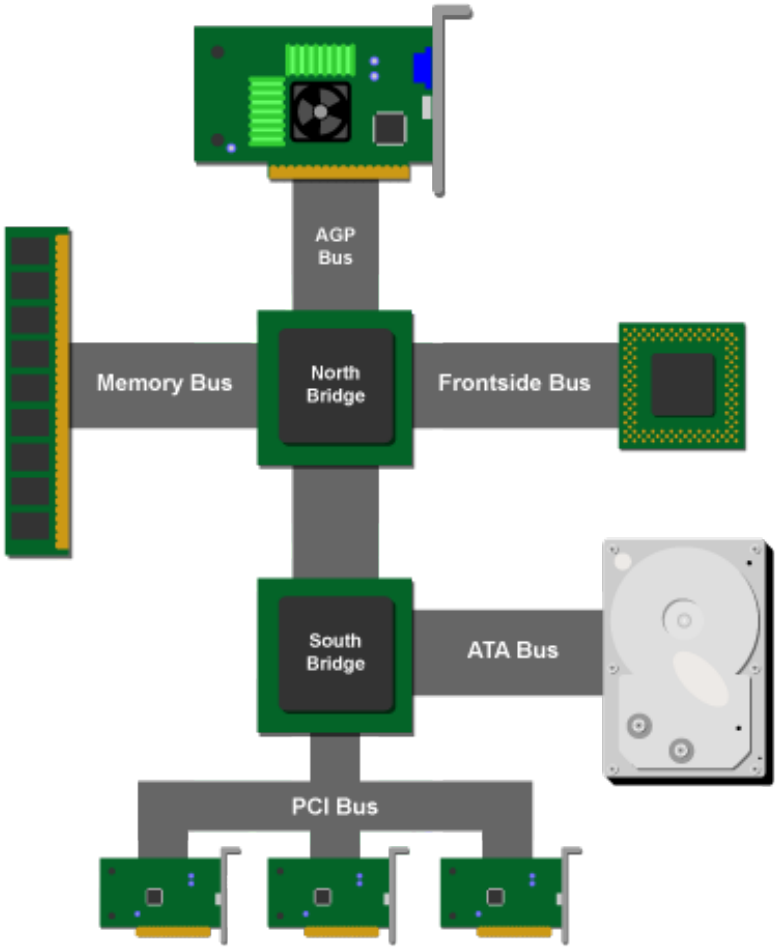
Disks in a virtual environment

The Trail of Tears to the Platter



Disks in a virtual environment

Why EC2 I/O is Slow and Unpredictable

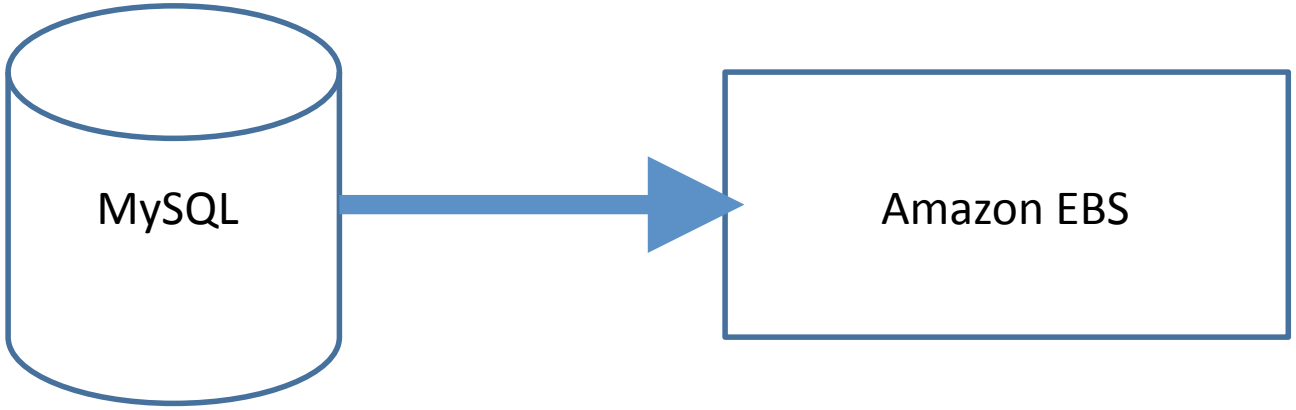


Shared Hardware
- Physical Disk
- Ethernet Controllers
- Southbridge

- How are the caches configured?
- How big are the caches?
- How many controllers?
- How many disks?
- RAID?

Image Credit: Ars Technica

Aside: Amazon EBS



Please stop doing this.

What's Killing That Box?

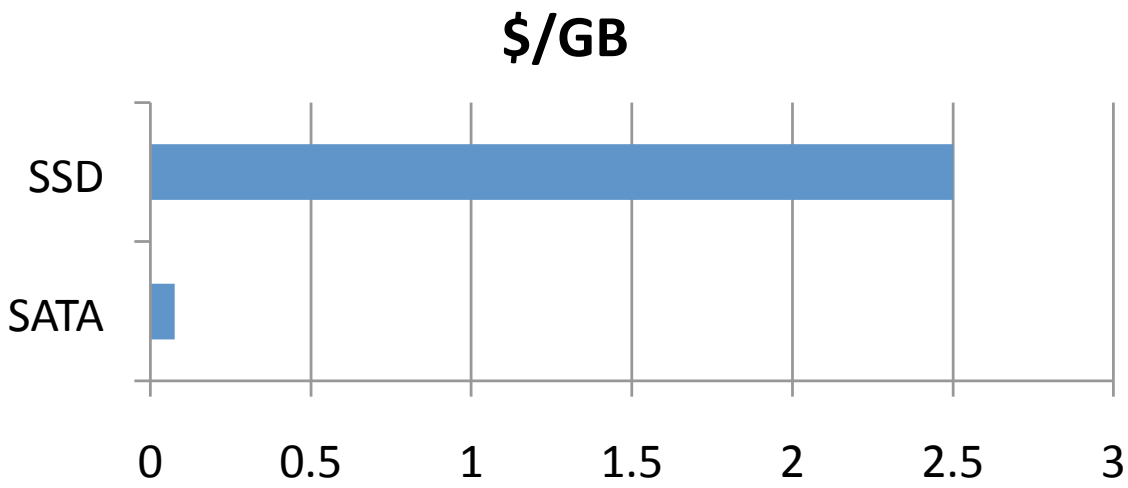
```
ted@u235:~$ iostat -x
Linux 2.6.32-24-generic (u235)  07/25/2011      _x86_64_        (8 CPU)

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           0.15    0.14    0.05    0.00    0.00   99.66

Device:         rrqm/s   wrqm/s     r/s     w/s   rsec/s   wsec/s avgrq-sz   %util
sda               0.00     3.27    0.01    2.38     0.58    45.23    19.21    0.24
```

Cool Hardware Tricks

Beginner Hardware Trick: SSD Drives



- $2.50/GB vs 7.5c/GB
- Negligible seek time vs 10ms seek time
- Not a lot of space

Cool Hardware Tricks

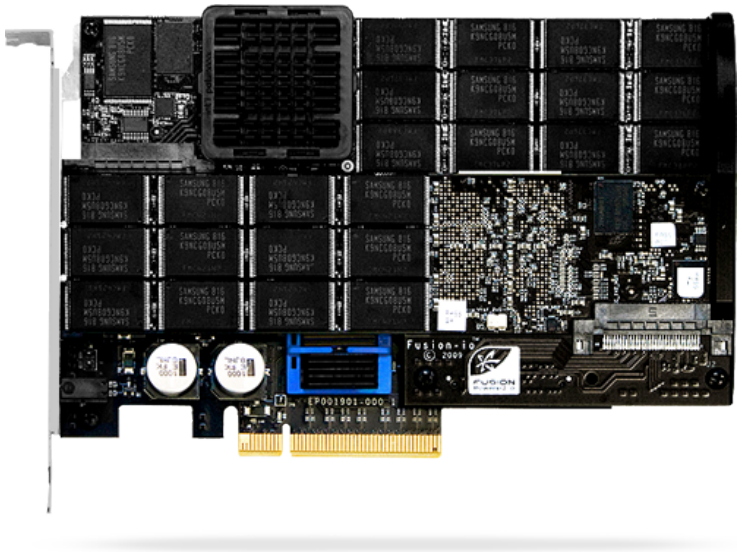
Intermediate Hardware Trick: RAID Controllers

- Standard RAID Controller
- SSD as writeback cache
- Battery-backed
- Adaptec “MaxIQ”
- $1,200

Image Credit: Tom’s Hardware

Cool Hardware Tricks

Advanced Hardware Trick: FusionIO



- SSD Storage on the Northbridge (PCIe)
- 6.0 GB/sec throughput. Giga***bytes***.
- 30 microsecond latency (30k ns)
- Roughly $20/GB
- Top-line card > $100,000 for around 5TB

Questions & Heckling

Thank You

http://teddziuba.com/

@dozba