Log2fs or how to achieve 150.000 IO/s

Jörn Engel

Lazybastard.org

September 24, 2010

Just a bunch of random hacks

Jörn Engel

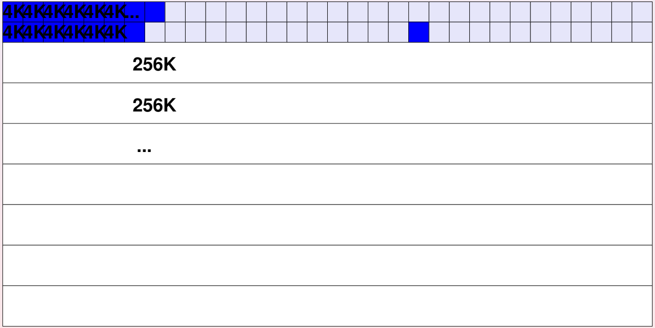
Lazybastard.org

September 24, 2010

Flash basics

- Fast random reads
- Fast somethat-random writes
- Erase before write
- Large eraseblocks

Blocks and Pages



4K4K4K4K4K4K...

4K4K4K4K4K4K4K

256K

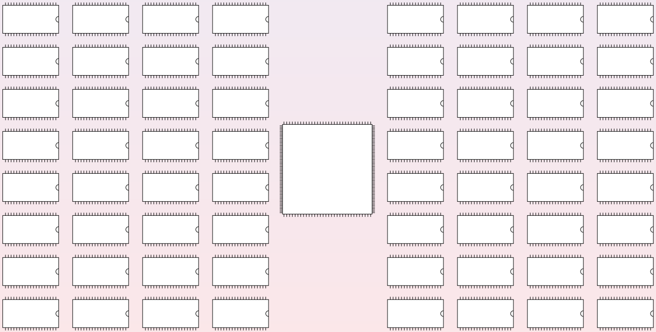
256K

...

Drais card

- PCIe x4 interface
- 1 FPGA
- 64 Flash chips

Drais card



Drais driver

- Simple MTD driver
- 64 queues for requests
- Does error correction
- Adds FIO interface

FIO interface

Adds three new methods to struct mtd_info

- fio_read
- fio_write
- fio_erase

FIO interface

fio_read reads exactly 1 page, then calls fio->fi_end_io

FIO interface

fio_write writes exactly 1 page, then calls fio->fi_end_io

FIO interface

fio_erase erases exactly 1 block, then calls fio->fi_end_io

wait_multiple

wait_multiple waits for N fios to complete

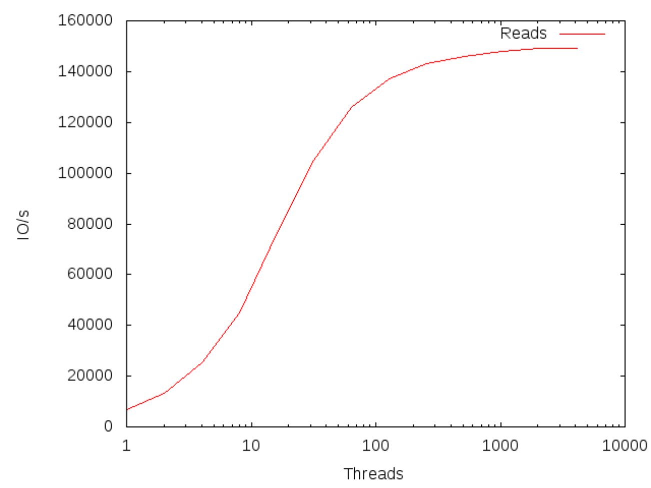
Read Performance

- Single threaded: 6800 IO/s or 27MB/s
- 4096 threads: 149000 IO/s 610MB/s
- Scales 22x

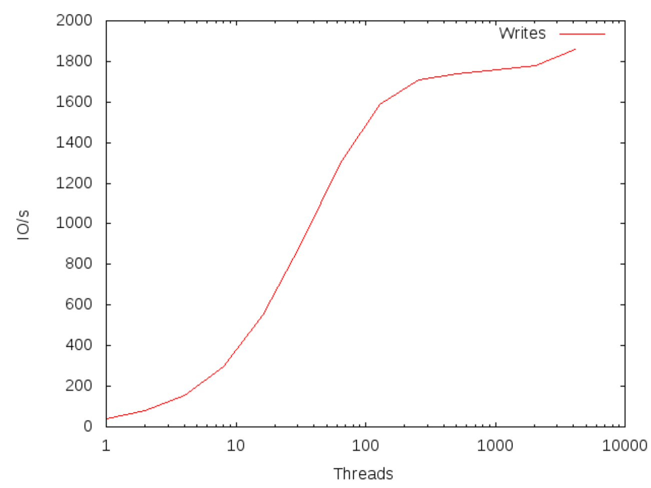
Write Performance

- Single threaded: 40 IO/s or 10MB/s
- 4096 threads: 1859 IO/s or 480MB/s
- Scales 46x

Read performance



Write performance



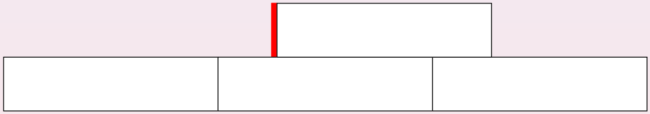
Compression in LogFS

- byte-precise packing
- indirect blocks contain pointers
- block headers contain compressed size

Alignment

- many blocks span a page boundary

Alignment



Alignment

- uncompressed and compressed blocks are mixed

Writes

- write header and compressed data to buffer
- occasionally flush buffer

Reads

- read header plus maximal blocksize to cache
- uncompress

Deletions

- Read header into cache
- Use compressed size for accounting

Cache

- Cache has a granularity of (MMU)PAGE_SIZE

Cache

- Oops!

Deletions

Deletions

Log2

- Don’t mix uncompressed and compressed blocks
- Align uncompressed blocks

Log2

- Move compressed size to indirect blocks
- ...and a number of other fields
- ...and remove (most) direct pointers from inodes

Venti

- Efficient way to store multiple identical copies
- Ideal for large universities
- Horrible for personal computers

VentiLog

- Add a block hashtable
- Check hashtable before writes
- Increment refcount when possible

BtrLog

- Add reference count to block pointers
- copyfile() becomes possible
- clones become possible

Birthday attack

LogFS stores directory entries in a hash table.

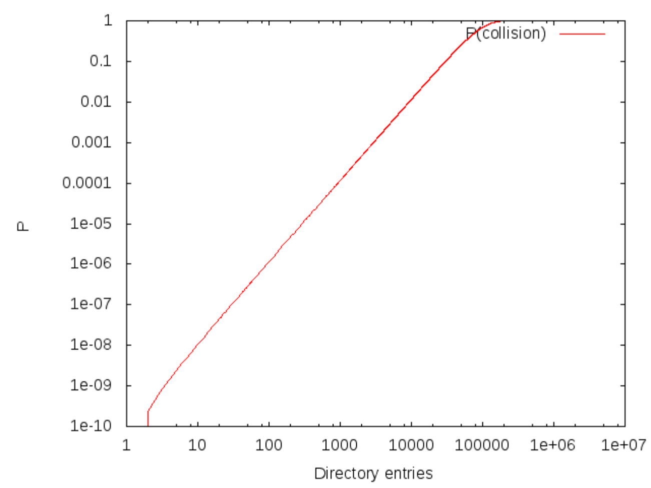
Birthday attack

Given N random numbers between 1 and M ($N \leq M$), what is the probability of having two or more identical numbers?

Birthday attack

$$1 - \frac{M! - N!}{M^N}$$

Birthday attack



Birthday attack

Given N random numbers between 1 and M ($N \leq M$), what is the probability of having O or more identical numbers ($O \leq N \leq O \cdot M$)?