PostgreSQL + ZFS

Best Practices and Standard Procedures
"If you are not using ZFS, you are losing data*."
ZFS is not magic, but it is an incredibly impressive piece of software.

Clark's Three Laws

1. When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong.
2. The only way of discovering the limits of the possible is to venture a little way past them into the impossible.
3. Any sufficiently advanced technology is indistinguishable from magic.
Many bits
Lots of bits
Huge bits
It's gunna be great
Very excited
We have the best filesystems
People tell me this is true
Except the fake media, they didn't tell me this
PostgreSQL and ZFS: It's about the bits and storage, stupid.

- Many bits
- Lots of bits
- Huge bits
- It's gunna be great
- Very excited
- We have the best filesystems
- People tell me this is true
- Except the fake media, they didn't tell me this

Too soon?
### PostgreSQL and ZFS

1. Review PostgreSQL from a storage administrator's perspective
2. Learn what it takes to become a PostgreSQL "backup expert"
3. Dive through a naive block-based filesystem
4. Walk through the a high-level abstraction of ZFS
5. See some examples of how to use ZFS with PostgreSQL
   - Tips
   - Tunables
   - Anecdotes

---

Some FS minutiae may have been harmed in the making of this talk. Nit-pick as necessary (preferably after).
PostgreSQL - A Storage Administrator's View

• User-land page cache maintained by PostgreSQL in shared memory
• 8K page size
• Each PostgreSQL table is backed by one or more files in $PGDATA/
• Tables larger than 1GB are automatically shared into individual 1GB files
• `pwrite(2)`'s to tables are:
  • append-only if no free pages in the table are available
  • in-place updated if free pages are available in the free-space map
• `pwrite(2)`'s are page-aligned
• Makes heavy use of a Write Ahead Log (WAL), aka an Intent Log
WAL files are written to sequentially
• append-only IO
• Still 8K page-aligned writes via pwrite(2)
• WAL logs are 16MB each, pre-allocated
• WAL logs are never unlink(2)’ed, only recycled via rename(2)
• Low-latency pwrite(2)’s and fsync(2) for WAL files is required for good write performance
Traditionally, only two SQL commands that you must know:

1. `pg_start_backup('my_backup')`

2. `${some_random_backup_utility} $PGDATA/`

3. `pg_stop_backup()`

Wait for `pg_start_backup()` to return before backing up `$PGDATA/` directory.
PostgreSQL - Backups

Only two three SQL commands that you must know:

1. CHECKPOINT
2. pg_start_backup('my_backup')
3. ${some_random_backup_utility} $PGDATA/
4. pg_stop_backup()

Manual CHECKPOINT if you can't twiddle the args to pg_start_backup().
Only two commands that you must know:

1. `CHECKPOINT`
2. `pg_start_backup('my_backup', true)`
3. `${some_random_backup_utility} $PGDATA/`
4. `pg_stop_backup()`

`pg_start_backup('my_backup', true)`
a.k.a. aggressive checkpointing (vs default perf hit of: 0.5 * `checkpoint_completion_target`)
Achievement unlocked
PostgreSQL Backup Expert
Achievement Pending
ZFS Ninja
TIP: Look for parallels.
Quick ZFS Primer: Features (read: why you must use ZFS)

- Never inconsistent (no fsck(8)'s required, ever)
- Filesystem atomically moves from one consistent state to another consistent state
- All blocks are checksummed
- Compression builtin
- Snapshots are free and unlimited
- Clones are easy
- Changes accumulate in memory, flushed to disk in a transaction
- Redundant metadata (and optionally data)
- Filesystem management independent of physical storage management
- Log-Structured Filesystem
- Copy on Write (COW)
Feature Consequences (read: how your butt gets saved)

• bitrot detected and automatically corrected if possible
  • phantom writes
  • misdirected reads or writes by the drive heads
  • DMA parity errors
  • firmware or driver bugs
  • RAM capacitors aren't refreshed fast enough or with enough power
• Phenomenal sequential and random IO write performance
• Performance increase for sequential reads
• Cost of ownership goes down
• New tricks and tools to solve "data gravity" problems
ELI5: Block Filesystems vs Log Structured Filesystems
Block Filesystems: Top-Down

Userland Application

\[\text{write}(\text{fd}, \text{buffer}, \text{cnt})\]

buffer

Userland
Block Filesystems: Top-Down

Userland Application

```
write(fd, buffer, cnt)
```

Userland

Kernel

VFS Layer

Logical File: PGDATA/global/1
Block Filesystems: Top-Down

Userland Application

```
write(fd, buffer, cnt)
```

buffer

VFS Layer

Logical File: `PGDATA/global/1`

System Buffers
Block Filesystems: Top-Down

Userland Application

\[ \text{write}(\text{fd}, \text{buffer}, \text{cnt}) \]

buffer

Userland

Kernel

VFS Layer

Logical File: `PGDATA/global/1`

System Buffers

Logical File Blocks

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
</table>
Block Filesystems: Top-Down

VFS Layer
Logical File: PGDATA/global/1

System Buffers

Logical File Blocks
0 1 2 3 4

Physical Storage Layer
2: #9971
3: #0016
4: #0317
0: #8884
1: #7014

Pretend this is a spinning disk
Userland Application

write(fd, buffer, cnt)

cnt = 2

8k buffer
Userland Application

```c
write(fd, buffer, cnt)
```

VFS Layer

- **Logical File**: `PGDATA/global/1`
- **System Buffers**
- **Logical File Blocks**

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
</table>

8k buffer

cnt = 2
Block Filesystems: PostgreSQL Edition

Kernel

VFS Layer

Logical File: PGDATA/global/1

System Buffers

Logical File Blocks

0 1 2 3

Physical Storage Layer

2: #9971
3: #0016
0: #8884
1: #7014
Quiz Time

What happens when you twiddle a bool in a row?

```sql
UPDATE foo_table SET enabled = FALSE WHERE id = 123;
```
UPDATE foo_table SET enabled = FALSE WHERE id = 123;

Userland Application

write(fd, buffer, cnt)
ZFS Tip: postgresql.conf: full_page_writes=off

ALTER SYSTEM SET full_page_writes=off;
CHECKPOINT;
-- Restart PostgreSQL

IMPORTANT NOTE: full_page_writes=off interferes with cascading replication
• buffers can be 4K
• disk sectors are 512B - 4K
• ordering of writes is important
• consistency requires complete cooperation and coordination
Physical Storage is decoupled from Filesystems.

If you remember one thing from this section, this is it.
VDEVs On the Bottom

VDEV: raidz

IO Scheduler

disk1  disk2  disk3  disk4

VDEV: mirror

IO Scheduler

disk5  disk6

zpool: rpool or tank
## Filesystems On Top

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>VFS Mountpoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>tank/ROOT</td>
<td>/</td>
</tr>
<tr>
<td>tank/db</td>
<td>/db</td>
</tr>
<tr>
<td>tank/ROOT/usr</td>
<td>/usr</td>
</tr>
<tr>
<td>tank/local</td>
<td>none</td>
</tr>
<tr>
<td>tank/local/etc</td>
<td>/usr/local/etc</td>
</tr>
</tbody>
</table>

(canmount=off)
Offensively Over Simplified Architecture Diagram

- **ZPL - ZFS POSIX Layer**
- **Filesystem**
- **zvol**

**DSL - Dataset and Snapshot Layer**

**VDEV: raidz**
- IO Scheduler
- disk1, disk2, disk3, disk4

**VDEV: mirror**
- IO Scheduler
- disk5, disk6

**zpool**: rpool or tank
ZFS is magic until you know how it fits together

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>VFS Mountpoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>tank/ROOT</td>
<td>/</td>
</tr>
<tr>
<td>tank/db</td>
<td>/db</td>
</tr>
<tr>
<td>tank/ROOT/usr</td>
<td>/usr</td>
</tr>
<tr>
<td>tank/local</td>
<td>none</td>
</tr>
<tr>
<td>tank/local/etc</td>
<td>/usr/local/etc</td>
</tr>
</tbody>
</table>

**ZPL - ZFS POSIX Layer**

**Filesystem**

**DSL - Dataset and Snapshot Layer**

**Dataset**

**VDEV: raidz**

IO Scheduler

disk1 disk2 disk3 disk4

**VDEV: mirror**

IO Scheduler

disk5 disk6

**zpool: rpool or tank**
Log-Structured Filesystems: Top-Down
Log-Structured Filesystems: Top-Down

Disk Block with `foo_table` Tuple
Disk Block with foo_table Tuple
ZFS: User Data + File dnode

\[ t_1 \]
ZFS: Object Set
ZFS: Meta-Object Set Layer
ZFS: Uberblock
At what point did the filesystem become inconsistent?
At what point could the filesystem become inconsistent?

Neglected to highlight ZFS is Copy-On-Write (read: knowingly committed perjury in front of a live audience)
How? I lied while explaining the situation. Alternate

ZFS is Copy-On-Write
What what's not been deleted and on disk is immutable.

(read: I nearly committed perjury in front of a live audience by knowingly withholding vital information)
ZFS is Copy-On-Write

Disk Block with foo_table Tuple
At what point did the filesystem become inconsistent?
At what point did the filesystem become inconsistent?
At what point did the filesystem become inconsistent?
At what point could the filesystem become inconsistent?

NEVER
TIL about ZFS: Transactions and Disk Pages

• Transaction groups are flushed to disk every N seconds (defaults to 5s)
• A transaction group (txg) in ZFS is called a "checkpoint"
• User Data can be modified as it is written to disk
• All data is checksummed
• Compression should be enabled by default
**ZFS Tip: ALWAYS enable compression**

```bash
$ zfs get compression
NAME      PROPERTY     VALUE     SOURCE
rpool     compression  off       default
rpool/root compression  off       default
$ sudo zfs set compression=lz4 rpool
$ zfs get compression
NAME      PROPERTY     VALUE     SOURCE
rpool     compression  lz4       local
rpool/root compression lz4       inherited from rpool
```

- Across ~7PB of PostgreSQL and mixed workloads and applications: compression ratio of ~2.8:1 was the average.
- Have seen >100:1 compression on some databases (*cough* this data probably didn't belong in a database *cough*)
- Have seen as low as 1.01:1
ZFS Tip: ALWAYS enable compression

I have yet to see compression slow down benchmarking results or real world workloads. My experience is with:

- spinning rust (7.2K RPM, 10K RPM, and 15KRPM)
- fibre channel connected SANs
- SSDs
- NVME
ZFS Tip: ALWAYS enable compression

$ zfs get compressratio

<table>
<thead>
<tr>
<th>NAME</th>
<th>PROPERTY</th>
<th>VALUE</th>
<th>SOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpool</td>
<td>compressratio</td>
<td>1.64x</td>
<td>-</td>
</tr>
<tr>
<td>rpool/db</td>
<td>compressratio</td>
<td>2.58x</td>
<td>-</td>
</tr>
<tr>
<td>rpool/db/pgdb1-10</td>
<td>compressratio</td>
<td>2.61x</td>
<td>-</td>
</tr>
<tr>
<td>rpool/root</td>
<td>compressratio</td>
<td>1.62x</td>
<td>-</td>
</tr>
</tbody>
</table>

- Use `lz4` by default everywhere.
- Use `gzip-9` only for archive servers
- Never mix-and-match compression where you can't suffer the consequences of lowest-common-denominator performance
- Anxious to see `ZStandard` support (I'm looking at you Allan Jude)
ZFS Perk: Data Locality

• Data written at the same time is stored near each other because it's frequently part of the same record
• Data can now pre-fault into kernel cache (ZFS ARC) by virtue of the temporal adjacency of the related `pwrite(2)` calls
• Write locality + `compression=lz4` + `pg_repack` == PostgreSQL Dream Team
ZFS Perk: Data Locality

• Data written at the same time is stored near each other because it's frequently part of the same record
• Data can now pre-fault into kernel cache (ZFS ARC) by virtue of the temporal adjacency of the related `pwrite(2)` calls
• Write locality + `compression=lz4` + `pg_repack` == PostgreSQL Dream Team

If you don't know what `pg_repack` is, figure out how to move into a database environment that supports `pg_repack` and use it regularly.

Ask after if you are curious, but here's a teaser:

What do you do if the `dedup` hash tables don't fit in RAM?

- This is not just my recommendation, it's also from the community and author of the feature.
- These are not the droids you are looking for
- Do not pass Go
- Do not collect $200
- Go straight to system unavailability jail
- The feature works, but you run the risk of bricking your ZFS server.
Bitrot is a Studied Phenomena

A Large-Scale Study of Flash Memory Failures in the Field

Justin Meza  Carnegie Mellon University  meza@cmu.edu
Qiang Wu  Facebook, Inc.  qwu@fb.com
Sanjeev Kumar  Facebook, Inc.  skumar@fb.com
Ourur Mutlu  Carnegie Mellon University  onur@cmu.edu

ABSTRACT

Servers use flash memory-based solid state drives (SSDs) as a high-performance alternative to hard disk drives to store persistent data. Unfortunately, recent increases in flash density have also brought about decreases in chip-level reliability. In a data center environment, flash-based SSD failures can lead to downtime and, in the worst case, data loss. As a result, it is important to understand flash memory reliability characteristics over flash lifetime in a realistic production data center environment running modern applications and system software.

This paper presents the first large-scale study of flash-based SSD reliability in the field. We analyze data collected across a majority of flash-based solid state drives at Facebook data centers over nearly four years and many millions of operational hours in order to understand failure properties and trends of flash-based SSDs. Our study considers a variety of SSD characteristics, including the access of data written to SSD read from flash chips; how data is mapped within the SSD address space; the amount of data copied, erased, and discarded by the flash controller; and flash-based temperature and host power.

Based on our field analysis of how flash memory errors manifest when running modern workloads on modern SSDs, this paper is the first to make several major observations: (1) SSD failure rates do not increase monotonically with flash chip wear; instead they go through several distinct periods corresponding to how failures emerge and are subsequently detected, (2) the effects of read disturbance errors are not prevalent in the field, (3) space logical data layout across an SSD’s physical address space (e.g., non-contiguous data), as measured by the amount of metadata required to track logical address translations stored in an SSD internal buffer, can greatly affect SSD failure rates, (1) higher temperatures lead to higher failure rates, but techniques that throttle SSD operation appear to greatly reduce the negative reliability impact of higher temperatures, and (5) data written by the operating system in flash-based SSDs does not always survive.
3.1 Bit Error Rate

The bit error rate (BER) of an SSD is the rate at which errors occur relative to the amount of information that is transmitted from/to the SSD. BER can be used to gauge the reliability of data transmission across a medium. We measure the uncorrectable bit error rate (UBER) from the SSD as:

$$UBER = \frac{Uncorrectable\ errors}{Bits\ accessed}$$

For flash-based SSDs, UBER is an important reliability metric that is related to the SSD lifetime. SSDs with high UBERs are expected to have more failed cells and encounter more (severe) errors that may potentially go undetected and corrupt
Bitrot is a Studied Phenomena

data than SSDs with low UBERs. Recent work by Grupp et al. examined the BER of raw MLC flash chips (without performing error correction in the flash controller) in a controlled environment [20]. They found the raw BER to vary from $1 \times 10^{-1}$ for the least reliable flash chips down to $1 \times 10^{-8}$ for the most reliable, with most chips having a BER in the $1 \times 10^{-6}$ to $1 \times 10^{-8}$ range. Their study did not analyze the effects of the use of chips in SSDs under real workloads and system software.

Table 1 shows the UBER of the platforms that we examine. We find that for flash-based SSDs used in servers, the UBER ranges from $2.6 \times 10^{-9}$ to $5.1 \times 10^{-11}$. While we expect that the UBER of the SSDs that we measure (which correct small errors, perform wear leveling, and are not at the end of their rated life but still being used in a production environment) should be less than the raw chip BER in Grupp et al.'s study (which did not correct any errors in the flash controller, exercised flash chips until the end of their rated life, and accessed flash chips in an adversarial manner), we find that in some cases the BERs were within around one order of magnitude of each other. For example, the UBER of Platform B in our study, $2.6 \times 10^{-9}$, comes close to the lower end of the raw chip BER range reported in prior work, $1 \times 10^{-8}$. 
Figure 2 (bottom) shows the average yearly uncorrectable error rate among SSDs within the different platforms – the sum of errors that occurred on all servers within a platform over 12 months ending in November 2014 divided by the number of servers in the platform. The yearly rates of uncorrectable errors on the SSDs we examined range from 15,128 for Platform D to 978,806 for Platform B. The older Platforms A and B have a higher error rate than the younger Platforms C through F, suggesting that the incidence of uncorrectable errors increases as SSDs are utilized more. We will examine this relationship further in Section 4.

Platform B has a much higher average yearly rate of uncorrectable errors (978,806) compared to the other platforms (the second highest amount, for Platform A, is 106,678). We find that this is due to a small number of SSDs having a much
TIL: Bitrot is here

• TL;DR: 4.2% -> 34% of SSDs have one UBER per year
TIL: Bitrot Roulette

\[
(1-(1-\text{uberRate})^{\text{numDisks}}) = \text{Probability of UBER/server/year}
\]

\[
(1-(1-0.042)^{20}) = 58\%
\]

\[
(1-(1-0.34)^{20}) = 99.975\%
\]

Highest quality SSD drives on the market

Lowest quality commercially viable SSD drives on the market
Causes of bitrot are Internal and External

External Factors for UBER on SSDs:

- Temperature
- Bus Power Consumption
- Data Written by the System Software
- Workload changes due to SSD failure
...except maybe they can.
Take Care of your bits

$ zpool status tank | head -n 3
  pool: tank
  state: ONLINE
  scan: scrub repaired 4.50K in 53h44m with 0 errors on Tue May 26 21:36:26 2015

Answer their cry for help.
Similar studies and research exist for:

- Fibre Channel
- SAS
- SATA
- Tape
- SANs
- Cloud Object Stores
"...I told you all of that, so I can tell you this..."
ZFS Terminology: VDEV

**VDEV** | vē-dēv
noun
a virtual device

- Physical drive redundancy is handled at the VDEV level
- Zero or more physical disks arranged like a RAID set:
  - mirror
  - stripe
  - raidz
  - raidz2
  - raidz3
Loose a VDEV, loose the zpool.

**ZFS Terminology: zpool**

*zpool*  |  zē-pōōl
---|---
noun
an abstraction of physical storage made up of a set of VDEVs
ZFS Terminology: ZPL

ZPL | zē-pē-el
noun
ZFS POSIX Layer

• Layer that handles the impedance mismatch between POSIX filesystem semantics and the ZFS "object database."
ZFS Terminology: **ZIL**

**ZIL** | zil
---|---
noun

ZFS Intent Log

- The ZFS analog of PostgreSQL's WAL
- If you use a ZIL:
  - Use disks that have low-latency writes
  - Mirror your ZIL
- If you loose your ZIL, whatever data had not made it to the main data disks will be lost. The PostgreSQL equivalent of: `rm -rf pg_xlog/`
ZFS Terminology: ARC

ARC | ärk
noun
Adaptive Replacement Cache

• ZFS's page cache
• ARC will grow or shrink to match use up all of the available memory

TIP: Limit ARC's max size to a percentage of physical memory minus the shared_buffer cache for PostgreSQL minus the kernel's memory overhead.
ZFS Terminology: Datasets

dataset | dædəˌset
noun
A filesystem or volume ("zvol")

- A ZFS filesystem dataset uses the underlying zpool
- A dataset belongs to one and only one zpool
- Misc tunables, including compression and quotas are set on the dataset level
<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZAP</td>
<td>ZFS Attribute Processor</td>
</tr>
<tr>
<td>DMU</td>
<td>Data Management Unit</td>
</tr>
<tr>
<td>DSL</td>
<td>Dataset and Snapshot Layer</td>
</tr>
<tr>
<td>SPA</td>
<td>Storage Pool Allocator</td>
</tr>
<tr>
<td>ZVOL</td>
<td>ZFS Volume</td>
</tr>
<tr>
<td>ZIO</td>
<td>ZFS I/O</td>
</tr>
<tr>
<td>RAIDZ</td>
<td>RAID with variable-size stripes</td>
</tr>
<tr>
<td>L2ARC</td>
<td>Level 2 Adaptive Replacement Cache</td>
</tr>
<tr>
<td>record</td>
<td>unit of user data, think RAID stripe size</td>
</tr>
</tbody>
</table>
$ sudo zfs list
NAME        USED  AVAIL  REFER  MOUNTPOINT
rpool       818M  56.8G  96K    none
rpool/root  817M  56.8G  817M   /

$ ls -lA -d /db
ls: cannot access '/db': No such file or directory

$ sudo zfs create rpool/db -o mountpoint=/db
$ sudo zfs list
NAME        USED  AVAIL  REFER  MOUNTPOINT
rpool       818M  56.8G  96K    none
rpool/db    96K  56.8G  96K    /db
rpool/root  817M  56.8G  817M   /

$ ls -lA /db
total 9
drwxr-xr-x  2 root root  2 Mar  2 18:06 ./
drwxr-xr-x 22 root root 24 Mar  2 18:06 ../
Storage Management

• Running out of disk space is bad, m'kay?
• Block file systems reserve ~8% of the disk space above 100%
• At ~92% capacity the performance of block allocators change from "performance optimized" to "space optimized" (read: performance "drops").
ZFS doesn't have an artificial pool of free space: you have to manage that yourself.

- Running out of disk space is bad, m'kay?
- Block file systems reserve ~8% of the disk space above 100%
- At ~92% capacity the performance of block allocators change from "performance optimized" to "space optimized" (read: performance "drops").
The pool should never consume more than 80% of the available space
$ sudo zfs set quota=48G rpool/db
$ sudo zfs get quota rpool/db
NAME     PROPERTY  VALUE  SOURCE
rpool/db  quota     48G    local
$ sudo zfs list
NAME    USED  AVAIL  REFER  MOUNTPOINT
rpool   818M  56.8G  96K    none
rpool/db 96K   48.0G  96K    /db
rpool/root 817M  56.8G  817M   /
Dataset Tuning Tips

- Disable `atime`
- Enable `compression`
- Tune the `recordsize`
- Consider tweaking the `primarycache`
ZFS Dataset Tuning

```
# zfs get atime,compression,primarycache,recordsize rpool/db
NAME      PROPERTY      VALUE         SOURCE
rpool/db  atime         on            inherited from rpool
rpool/db  compression   lz4           inherited from rpool
rpool/db  primarycache  all           default
rpool/db  recordsize    128K          default
# zfs set atime=off rpool/db
# zfs set compression=lz4 rpool/db
# zfs set recordsize=16K rpool/db
# zfs set primarycache=metadata rpool/db
# zfs get atime,compression,primarycache,recordsize rpool/db
NAME      PROPERTY      VALUE         SOURCE
rpool/db  atime         off           local
rpool/db  compression   lz4           local
rpool/db  primarycache  metadata      local
rpool/db  recordsize    16K           local
```
**Discuss: recordsize=16k**

- Pre-fault next page: useful for sequential scans
- With `compression=lz4`, reasonable to expect ~3-4x pages worth of data in a single ZFS record

Anecdotes and Recommendations:

- Performed better in most workloads vs ZFS's prefetch
- Disabling prefetch isn't necessary, tends to still be a net win
- Monitor arc cache usage
Discuss: `primarycache=metadata`

- metadata instructs ZFS's ARC to only cache metadata (e.g. `dnode` entries), not page data itself
- Default: cache all data

Two different recommendations based on benchmark workloads:
- Enable `primarycache=all` where working set exceeds RAM
- Enable `primarycache=metadata` where working set fits in RAM
Discuss: \texttt{primarycache=metadata}

- metadata instructs ZFS's ARC to only cache metadata (e.g. \texttt{dnode} entries), not page data itself
- Default: cache all data
- Double-caching happens

Two different recommendations based on benchmark workloads:
- Enable \texttt{primarycache=all} where working set exceeds RAM
- Enable \texttt{primarycache=metadata} where working set fits in RAM

Reasonable Default anecdote: Cap max ARC size ~15%-25% physical RAM + ~50% RAM \texttt{shared_buffers}
Performance Wins

```bash
# dtrace -s vfs-io-postgres.d
Latencies (ns)
postgres Write
```

Latency distribution:

<table>
<thead>
<tr>
<th>Value</th>
<th>Distribution</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>2048</td>
<td>@</td>
<td>1325</td>
</tr>
<tr>
<td>4096</td>
<td>@@@</td>
<td>267</td>
</tr>
<tr>
<td>8192</td>
<td>@@@</td>
<td>72</td>
</tr>
<tr>
<td>16384</td>
<td>@</td>
<td>0</td>
</tr>
<tr>
<td>32768</td>
<td>@</td>
<td>0</td>
</tr>
<tr>
<td>65536</td>
<td>@</td>
<td>0</td>
</tr>
<tr>
<td>131072</td>
<td>@@@@</td>
<td>19</td>
</tr>
<tr>
<td>262144</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
### Performance Wins

```
# zpool iostat tank 1

<table>
<thead>
<tr>
<th>pool</th>
<th>alloc</th>
<th>free</th>
<th>read</th>
<th>operations</th>
<th>write</th>
<th>read</th>
<th>bandwidth</th>
<th>write</th>
</tr>
</thead>
<tbody>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>210K</td>
<td>1022</td>
<td>330M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>1</td>
<td>207K</td>
<td>4.99K</td>
<td>326M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>32</td>
<td>30.5K</td>
<td>79.9K</td>
<td>46.9M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>22</td>
<td>9.62K</td>
<td>202K</td>
<td>15.9M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>15</td>
<td>10.2K</td>
<td>169K</td>
<td>16.5M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>36</td>
<td>10.5K</td>
<td>198K</td>
<td>14.9M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>6</td>
<td>10.8K</td>
<td>39.4K</td>
<td>17.4M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>12</td>
<td>189K</td>
<td>209K</td>
<td>298M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>1</td>
<td>210K</td>
<td>7.96K</td>
<td>340M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>10</td>
<td>218K</td>
<td>23.0K</td>
<td>355M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>2</td>
<td>224K</td>
<td>4.49K</td>
<td>359M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>6</td>
<td>228K</td>
<td>12.5K</td>
<td>367M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>7</td>
<td>140K</td>
<td>53.4K</td>
<td>225M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>9</td>
<td>26.9K</td>
<td>40.9K</td>
<td>44.0M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>9.43K</td>
<td>0</td>
<td>13.9M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>9.69K</td>
<td>0</td>
<td>16.3M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>1</td>
<td>74.0K</td>
<td>3.49K</td>
<td>120M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>6</td>
<td>226K</td>
<td>17.0K</td>
<td>366M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>225K</td>
<td>0</td>
<td>385M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>176K</td>
<td>0</td>
<td>515M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>84.7K</td>
<td>0</td>
<td>382M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>39.6K</td>
<td>0</td>
<td>163M</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
## Performance Wins

```
# zpool iostat tank 1

<table>
<thead>
<tr>
<th>pool</th>
<th>alloc</th>
<th>capacity</th>
<th>operations</th>
<th>bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>free</td>
<td>read</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>write</td>
<td>read</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>210K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>1</td>
<td>207K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>32</td>
<td>30.5K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>22</td>
<td>9.62K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>15</td>
<td>10.2K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>36</td>
<td>10.5K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>6</td>
<td>10.8K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>12</td>
<td>189K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>1</td>
<td>210K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>16</td>
<td>218K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>2</td>
<td>224K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>6</td>
<td>228K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>7</td>
<td>140K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>9</td>
<td>26.9K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>9.43K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>9.69K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>1</td>
<td>74.0K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>6</td>
<td>226K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>225K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>176K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>84.7K</td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>39.6K</td>
</tr>
</tbody>
</table>
# zpool iostat tank 1

<table>
<thead>
<tr>
<th>pool</th>
<th>alloc</th>
<th>free</th>
<th>read</th>
<th>capacity</th>
<th>operations</th>
<th>read</th>
<th>bandwidth</th>
<th>write</th>
</tr>
</thead>
<tbody>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>9.94T</td>
<td>210K</td>
<td>1022</td>
<td>330M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>1</td>
<td>9.94T</td>
<td>207K</td>
<td>4.99K</td>
<td>326M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>32</td>
<td>9.94T</td>
<td>30.5K</td>
<td>79.9K</td>
<td>46.9M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>32</td>
<td>9.94T</td>
<td>632K</td>
<td>203K</td>
<td>15.9M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>36</td>
<td>9.94T</td>
<td>10.5K</td>
<td>198K</td>
<td>14.9M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>12</td>
<td>9.94T</td>
<td>189K</td>
<td>209K</td>
<td>298M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>1</td>
<td>9.94T</td>
<td>210K</td>
<td>7.96K</td>
<td>340M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>16</td>
<td>9.94T</td>
<td>218K</td>
<td>23.0K</td>
<td>355M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>2</td>
<td>9.94T</td>
<td>224K</td>
<td>4.49K</td>
<td>359M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>6</td>
<td>9.94T</td>
<td>228K</td>
<td>12.5K</td>
<td>367M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>7</td>
<td>9.94T</td>
<td>140K</td>
<td>53.4K</td>
<td>225M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>9</td>
<td>9.94T</td>
<td>26.9K</td>
<td>40.9K</td>
<td>44.0M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>9.94T</td>
<td>9.43K</td>
<td>0</td>
<td>13.9M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>9.94T</td>
<td>9.69K</td>
<td>0</td>
<td>16.3M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>1</td>
<td>9.94T</td>
<td>74.0K</td>
<td>3.49K</td>
<td>120M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>6</td>
<td>9.94T</td>
<td>226K</td>
<td>17.0K</td>
<td>366M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>9.94T</td>
<td>225K</td>
<td>0</td>
<td>385M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>9.94T</td>
<td>176K</td>
<td>0</td>
<td>515M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>9.94T</td>
<td>84.7K</td>
<td>0</td>
<td>382M</td>
<td></td>
</tr>
<tr>
<td>tank</td>
<td>958G</td>
<td>9.94T</td>
<td>0</td>
<td>9.94T</td>
<td>39.6K</td>
<td>0</td>
<td>163M</td>
<td></td>
</tr>
</tbody>
</table>

P.S. This was observed on 10K RPM spinning rust.
ZFS Always has your back

- ZFS will checksum every read from disk
- A failed checksum will result in a fault and automatic data reconstruction
- Scrubs do background check of every record
- Schedule periodic scrubs
  - Frequently for new and old devices
  - Infrequently for devices in service between 6mo and 2.5yr

PSA: The "Compressed ARC" feature was added to catch checksum errors in RAM

Checksum errors are an early indicator of failing disks
Schedule Periodic Scrubs

```bash
# zpool status
pool: rpool
state: ONLINE
scan: none requested
config:

<table>
<thead>
<tr>
<th>NAME</th>
<th>STATE</th>
<th>READ</th>
<th>WRITE</th>
<th>CKSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpool</td>
<td>ONLINE</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sda1</td>
<td>ONLINE</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

errors: No known data errors

# zpool scrub rpool
# zpool status
pool: rpool
state: ONLINE
scan: scrub in progress since Fri Mar 3 20:41:44 2017
753M scanned out of 819M at 151M/s, 0h0m to go
0 repaired, 91.97% done
config:

<table>
<thead>
<tr>
<th>NAME</th>
<th>STATE</th>
<th>READ</th>
<th>WRITE</th>
<th>CKSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpool</td>
<td>ONLINE</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sda1</td>
<td>ONLINE</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

errors: No known data errors

# zpool status
pool: rpool
state: ONLINE
scan: scrub repaired 0 in 0h0m with 0 errors on Fri Mar 3 20:41:49 2017

Non-zero on any of these values is bad™
One dataset per database

- Create one ZFS dataset per database instance
- General rules of thumb:
  - Use the same dataset for $PGDATA/ and pg_xlogs/
  - Set a reasonable quota
  - Optional: reserve space to guarantee minimal available space

Checksum errors are an early indicator of failing disks
One dataset per database

```bash
# zfs list
NAME        USED  AVAIL  REFER  MOUNTPOINT
rpool       819M  56.8G  96K none
rpool/db    160K  48.0G  96K /db
rpool/root   818M  56.8G  818M /

# zfs create rpool/db/pgdb1
# chown postgres:postgres /db/pgdb1
# zfs list
NAME        USED  AVAIL  REFER  MOUNTPOINT
rpool       819M  56.8G  96K none
rpool/db    256K  48.0G  96K /db
rpool/db/pgdb1 96K  48.0G  96K /db/pgdb1
rpool/root   818M  56.8G  818M /

# zfs set reservation=1G rpool/db/pgdb1
# zfs list
NAME        USED  AVAIL  REFER  MOUNTPOINT
rpool       1.80G  55.8G  96K none
rpool/db    1.00G  47.0G  96K /db
rpool/db/pgdb1 96K  48.0G  12.0M /db/pgdb1
rpool/root   818M  55.8G  818M /
```
Encode using UTF8, sort using C

Only enable locale when you know you need it
• ~2x perf bump by avoiding calls to `iconv(3)` to figure out sort order
• **DO NOT** use PostgreSQL checksums or compression


```bash
# zfs list -t snapshot
no datasets available
# pwd
/db/pgdb1
# find . | wc -l
895
# head -1 postmaster.pid
25114
# zfs snapshot rpool/db/pgdb1@pre-rm
# zfs list -t snapshot
NAME                    USED  AVAIL  REFER  MOUNTPOINT
rpool/db/pgdb1@pre-rm      0      -  12.0M  -
# psql -U postgres
psql (9.6.2)
Type "help" for help.

postgres=# \q
# rm -rf *
# ls -l | wc -l
0
# psql -U postgres
psql: FATAL:  could not open relation mapping file "global/pg_fileno\nde.map":
No such file or directory
```
$ psql
psql: FATAL: could not open relation mapping file "global/pg_fileno.node.map": No such file or directory

# cat postgres.log

LOG: database system was shut down at 2017-03-03 21:08:05 UTC
LOG: MultiXact member wraparound protections are now enabled
LOG: database system is ready to accept connections
LOG: autovacuum launcher started
FATAL: could not open relation mapping file "global/pg_fileno.node.map": No such file or directory
LOG: could not open temporary statistics file "pg_stat_tmp/global.tmp": No such file or directory
LOG: could not open temporary statistics file "pg_stat_tmp/global.tmp": No such file or directory
... 
LOG: could not open temporary statistics file "pg_stat_tmp/global.tmp": No such file or directory
LOG: could not open file "postmaster.pid": No such file or directory
LOG: performing immediate shutdown because data directory lock file is invalid
LOG: received immediate shutdown request
LOG: could not open temporary statistics file "pg_stat/global.tmp": No such file or directory
WARNING: terminating connection because of crash of another server process
DETAIL: The postmaster has commanded this server process to roll back the current transaction and exit, because another server process exited abnormally and possibly corrupted shared memory.
HINT: In a moment you should be able to reconnect to the database and repeat your command.
LOG: database system is shut down

# ll
total 1
drwx------ 2 postgres postgres 2 Mar 3 21:40 ./
drwxr-xr-x 3 root root 3 Mar 3 21:03 ../
Restores: As Important as Backups

# zfs list -t snapshot
NAME                    USED  AVAIL  REFER  MOUNTPOINT
rpool/db/pgdb1@pre-rm  12.0M      -  12.0M  -

# zfs rollback rpool/db/pgdb1@pre-rm
# su postgres -c '/usr/lib/postgresql/9.6/bin/postgres -D /db/pgdb1'
LOG:  database system was interrupted; last known up at 2017-03-03 21:50:57 UTC
LOG:  database system was not properly shut down; automatic recovery in progress
LOG:  redo starts at 0/14EE7B8
LOG:  invalid record length at 0/1504150: wanted 24, got 0
LOG:  redo done at 0/1504128
LOG:  last completed transaction was at log time 2017-03-03 21:51:15.340442+00
LOG:  MultiXact member wraparound protections are now enabled
LOG:  database system is ready to accept connections
LOG:  autovacuum launcher started

Works all the time, every time, even with kill -9
(possible dataloss from ungraceful shutdown and IPC cleanup not withstanding)
Clone: Test and Upgrade with Impunity

# zfs clone rpool/db/pgdb1@pre-rm rpool/db/pgdb1-upgrade-test
# zfs list -r rpool/db

<table>
<thead>
<tr>
<th>NAME</th>
<th>USED</th>
<th>AVAIL</th>
<th>REFER</th>
<th>MOUNTPOINT</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpool/db</td>
<td>1.00G</td>
<td>47.0G</td>
<td>96K</td>
<td>/db</td>
</tr>
<tr>
<td>rpool/db/pgdb1</td>
<td>15.6M</td>
<td>48.0G</td>
<td>15.1M</td>
<td>/db/pgdb1</td>
</tr>
<tr>
<td>rpool/db/pgdb1-upgrade-test</td>
<td>8K</td>
<td>47.0G</td>
<td>15.2M</td>
<td>/db/pgdb1-upgrade-test</td>
</tr>
</tbody>
</table>

# echo "Test pg_upgrade"
# zfs destroy rpool/db/pgdb1-clone
# zfs clone rpool/db/pgdb1@pre-rm rpool/db/pgdb1-10
# echo "Run pg_upgrade for real"
# zfs promote rpool/db/pgdb1-10
# zfs destroy rpool/db/pgdb1

Works all the time, every time, even with kill -9
(possible dataloss from ungraceful shutdown and IPC cleanup not withstanding)
Be explicit: codify the tight coupling between PostgreSQL versions and $PGDATA/.

**Tip: Naming Conventions**

- Use a short prefix not on the root filesystem (e.g. /db)
- Encode the PostgreSQL major version into the dataset name
- Give each PostgreSQL cluster its own dataset (e.g. pgdb01)
- Optional but recommended:
  - one database per cluster
  - one app per database
  - encode environment into DB name
  - encode environment into DB username

<table>
<thead>
<tr>
<th>Suboptimal</th>
<th>Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpool/db/pgdb1</td>
<td>rpool/db/prod-db01-pg94</td>
</tr>
<tr>
<td>rpool/db/myapp-shard1</td>
<td>rpool/db/prod-myapp-shard001-pg95</td>
</tr>
<tr>
<td>rpool/db/dbN</td>
<td>rpool/db/prod-dbN-pg10</td>
</tr>
</tbody>
</table>
Defy Gravity

- Take and send snapshots to remote servers
- `zfs send` emits a snapshot to stdout: treat as a file or stream
- `zfs receive` reads a snapshot from stdin
- TIP: If available:
  - Use the `-s` argument to `zfs receive`
  - Use `zfs get receive_resume_token` on the receiving end to get the required token to resume an interrupted send: `zfs send -t <token>`

Unlimited flexibility. Compress, encrypt, checksum, and offsite to your heart's content.
Defy Gravity

```bash
# zfs send -v -L -p -e rpool/db/pgdb1@pre-rm > /dev/null
send from @ to rpool/db/pgdb1-10@pre-rm estimated size is 36.8M
total estimated size is 36.8M
TIME        SENT   SNAPSHOT
# zfs send -v -L -p -e
  rpool/db/pgdb1-10@pre-rm | |
zfs receive -v |
  rpool/db/pgdb1-10-receive
send from @ to rpool/db/pgdb1-10@pre-rm estimated size is 36.8M
total estimated size is 36.8M
TIME        SENT   SNAPSHOT
received 33.8MB stream in 1 seconds (33.8MB/sec)
# zfs list -t snapshot
NAME                               USED  AVAIL  REFER
MOUNTPOINT
rpool/db/pgdb1-10@pre-rm             8K      -  15.2M  -
rpool/db/pgdb1-10-receive@pre-rm      0      -  15.2M  -
```
Defy Gravity: Incrementally

- Use a predictable snapshot naming scheme
- Send snapshots incrementally
- Clean up old snapshots
- Use a monotonic snapshot number (a.k.a. "vector clock")

Remember to remove old snapshots.
Distributed systems bingo!
# echo "Change PostgreSQL's data"
# zfs snapshot rpool/db/pgdb1-10@example-incremental-001
# zfs send -v -L -p -e \
#   -i rpool/db/pgdb1-10@pre-rm \ 
#   rpool/db/pgdb1-10@example-incremental-001 \ 
#   > /dev/null
send from @pre-rm to rpool/db/pgdb1-10@example-incremental-001
estimated size is 2K
total estimated size is 2K
#
# zfs send -v -L -p -e \
#   -i rpool/db/pgdb1-10@pre-rm \ 
#   rpool/db/pgdb1-10@example-incremental-001 |
# zfs receive -v \
#   rpool/db/pgdb1-10-receive
send from @pre-rm to rpool/db/pgdb1-10@example-incremental-001
estimated size is 2K
total estimated size is 2K
receiving incremental stream of rpool/db/pgdb1-10@example-incremental-001 into rpool/db/pgdb1-10-receive@example-incremental-001
received 312B stream in 1 seconds (312B/sec)
# echo "Change more PostgreSQL's data: VACUUM FULL FREEZE"
# zfs snapshot rpool/db/pgdb1-10@example-incremental-002
# zfs send -v -L -p -e \n   -i rpool/db/pgdb1-10@example-incremental-001 \n   rpool/db/pgdb1-10@example-incremental-002 \n > /dev/null
send from @example-incremental-001 to rpool/db/pgdb1-10@example-incremental-002 estimated size is 7.60M
total estimated size is 7.60M
TIME        SENT   SNAPSHOT
# zfs send -v -L -p -e \n   -i rpool/db/pgdb1-10@example-incremental-001 \n   rpool/db/pgdb1-10@example-incremental-002 | | \n   zfs receive -v \n   rpool/db/pgdb1-10-receive
send from @example-incremental-001 to rpool/db/pgdb1-10@example-incremental-002 estimated size is 7.60M
total estimated size is 7.60M
receiving incremental stream of rpool/db/pgdb1-10@example-incremental-002 into rpool/db/pgdb1-10-receive@example-incremental-002
TIME        SENT   SNAPSHOT
received 7.52MB stream in 1 seconds (7.52MB/sec)
Defy Gravity: Cleanup

```
# zfs list -t snapshot -o name,used,refer
NAME                              USED  REFER
rpool/db/pgdb1-10@example-incremental-001     8K  15.2M
rpool/db/pgdb1-10@example-incremental-002     848K 15.1M
rpool/db/pgdb1-10-receive@pre-rm          8K  15.2M
rpool/db/pgdb1-10-receive@example-incremental-001     8K  15.2M
rpool/db/pgdb1-10-receive@example-incremental-002      0   15.1M
# zfs destroy rpool/db/pgdb1-10-receive@pre-rm
# zfs destroy rpool/db/pgdb1-10@example-incremental-001
# zfs destroy rpool/db/pgdb1-10-receive@example-incremental-001
# zfs list -t snapshot -o name,used,refer
NAME                              USED  REFER
rpool/db/pgdb1-10@example-incremental-002     848K 15.1M
rpool/db/pgdb1-10-receive@example-incremental-002      0   15.1M
```
**Controversial: logbias=throughput**

- Measure tps/qps
- Time duration of an outage (OS restart plus WAL replay, e.g. 10-20min)
- Measure cost of back pressure from the DB to the rest of the application
- Use a `txg` timeout of 1 second

Position: since ZFS will never be inconsistent and therefore PostgreSQL will never lose integrity, 1s of actual data loss is a worthwhile tradeoff for a ~10x performance boost in write-heavy applications.

Rationale: loss aversion costs organizations more than potentially losing 1s of data. Back pressure is a constant cost the rest of the application needs to absorb due to continual `fsync(2)`'ing of WAL data. Architectural cost and premature engineering costs need to be factored in. Penny-wise, pound foolish.
Controversial: logbias=throughput

```bash
# cat /sys/module/zfs/parameters/zfs_txg_timeout
5
# echo 1 > /sys/module/zfs/parameters/zfs_txg_timeout
# echo 'options zfs zfs_txg_timeout=1' >> /etc/modprobe.d/zfs.conf
# psql -c 'ALTER SYSTEM SET synchronous_commit=off'
ALTER SYSTEM
# zfs set logbias=throughput rpool/db
```
Thank you to people who helped make this happen:

• Percona:
  • Peter Boros
  • Robert Barabas
• Groupon:
  • Sergio Murilo
  • Chris Schneider
  • Filippos Kalamidas
  • Jose Finotto
QUESTIONS?

Email  sean@chittenden.org
       seanc@joyent.com

Twitter: @SeanChittenden

We're hiring! Ask me what we're up to!