Scaling Btrfs in an Enterprise

Motiejus Jakštys
m.jakstys.lt
@motiejus:jakstys.lt

Chronosphere

This talk is about

- Btrfs cut our storage costs by 74%.
- Productionized Btrfs in Chronosphere and Google Cloud Platform.

Our business & size (mid 2025)

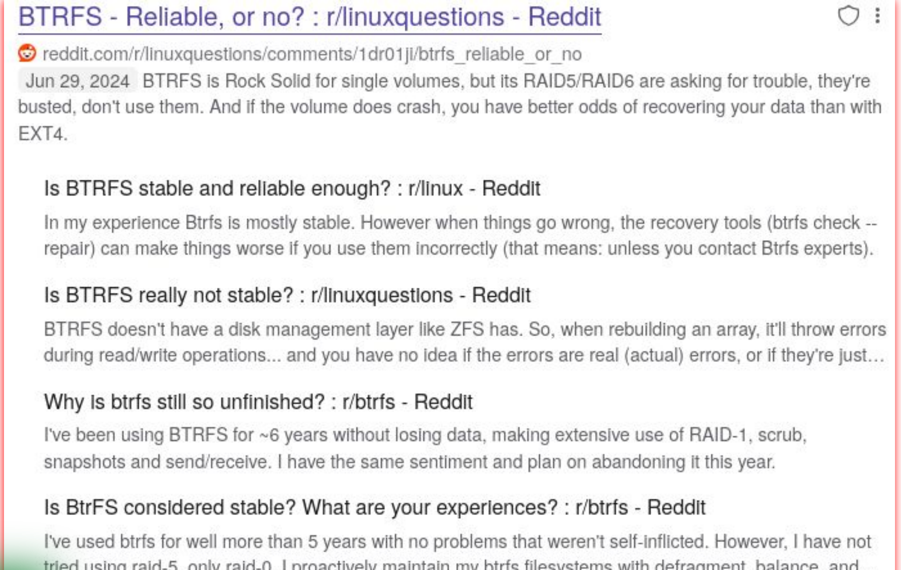
- Observability SaaS, lots of time-series data.
- Customers so big we run thousands of VMs *per customer*.
- Multi-terabyte disks per VM.
  - ⇒ Some PiB in Google's Persistent Disks.
- All on ext4.

Why Btrfs?

1. Compression saves disk space.
2. Compression is fully transparent.
3. Checksums can save on CPU and application complexity.

Disk compression demo

Initial tests showed ~65% disk savings.



[BTRFS - Reliable, or no? : r/linuxquestions - Reddit]

reddit.com/r/linuxquestions/comments/1dr01ji/btrfs_reliable_or_no

Jun 29, 2024 BTRFS is Rock Solid for single volumes, but its RAID5/RAID6 are asking for trouble, they're busted, don't use them. And if the volume does crash, you have better odds of recovering your data than with EXT4.

Is BTRFS stable and reliable enough? : r/linux - Reddit

In my experience Btrfs is mostly stable. However when things go wrong, the recovery tools (btrfs check --repair) can make things worse if you use them incorrectly (that means: unless you contact Btrfs experts).

Is BTRFS really not stable? : r/linuxquestions - Reddit

BTRFS doesn't have a disk management layer like ZFS has. So, when rebuilding an array, it'll throw errors during read/write operations... and you have no idea if the errors are real (actual) errors, or if they're just…

Why is btrfs still so unfinished? : r/btrfs - Reddit

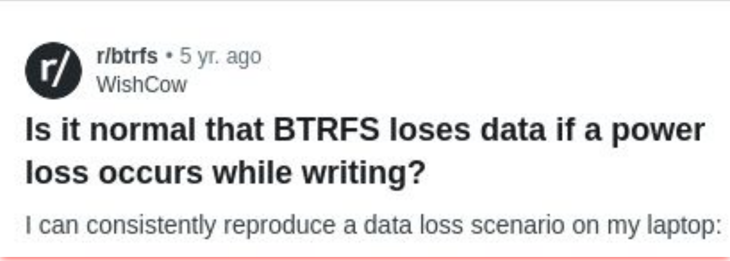
I've been using BTRFS for ~6 years without losing data, making extensive use of RAID-1, scrub, snapshots and send/receive. I have the same sentiment and plan on abandoning it this year.

Is BtrFS considered stable? What are your experiences? : r/btrfs - Reddit

I've used btrfs for well more than 5 years with no problems that weren't self-inflicted. However, I have not tried using raid-5, only raid-0. I proactively maintain my btrfs filesystems with defragment, balance, and…

s BTRFS stable and reliable enough? : r/linuxquestions - Reddit

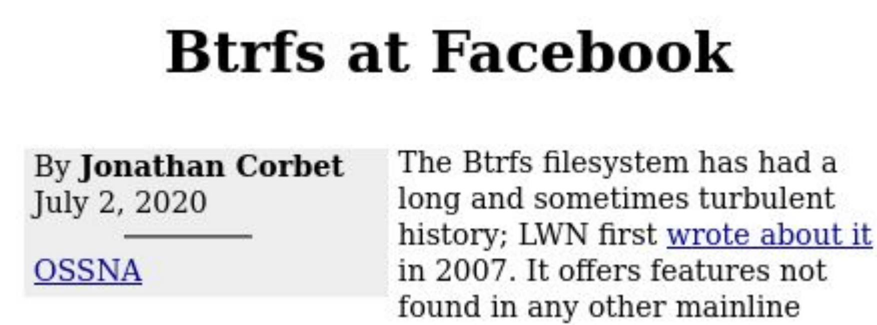
o far, ext4 has been 100% reliable for me. I don't want to do any kind of BTRFS-related system aintenance. If BTRFS is not at least as stable and reliable as ext4, I'll stick to LVM-ext4. Even if the…



r/btrfs • 5 yr. ago
WishCow

Is it normal that BTRFS loses data if a power loss occurs while writing?

I can consistently reproduce a data loss scenario on my laptop:

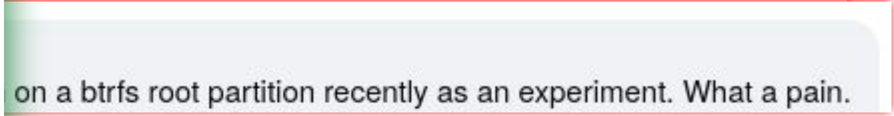


Btrfs at Facebook

By **Jonathan Corbet**
July 2, 2020

[OSSNA]

The Btrfs filesystem has had a long and sometimes turbulent history; LWN first [wrote about it] in 2007. It offers features not found in any other mainline



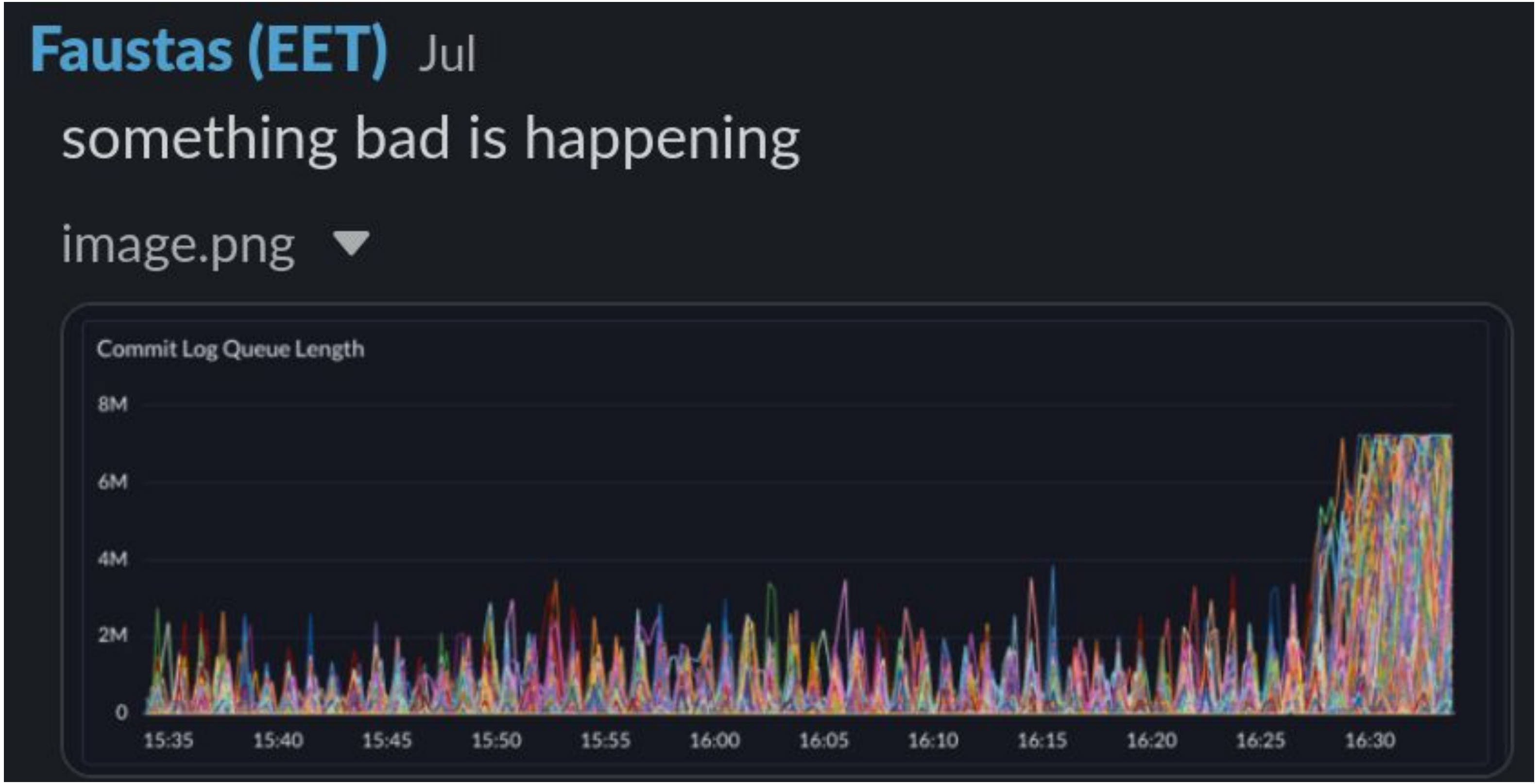
on a btrfs root partition recently as an experiment. What a pain.

Hurdles

1. Confidence & reputation.
   a. Savings significant enough to worth trying.
2. ENOSPC.
   a. bg_reclaim_treshold + dynamic_reclaim.
3. Unsupported by Google — why?
   a. Didn’t work? Or nobody just did it?
4. How will IO patterns affect performance?
   a. Will get to that 😈.
5. Unknown unknowns?

Surprises came up

Reclaim causing massive IO on large deletes



Faustas (EET) Jul

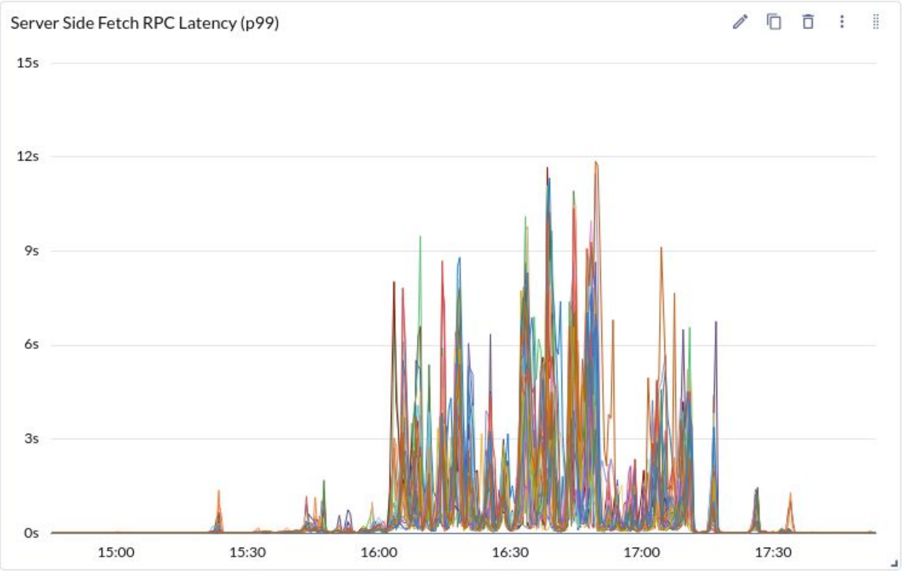
something bad is happening

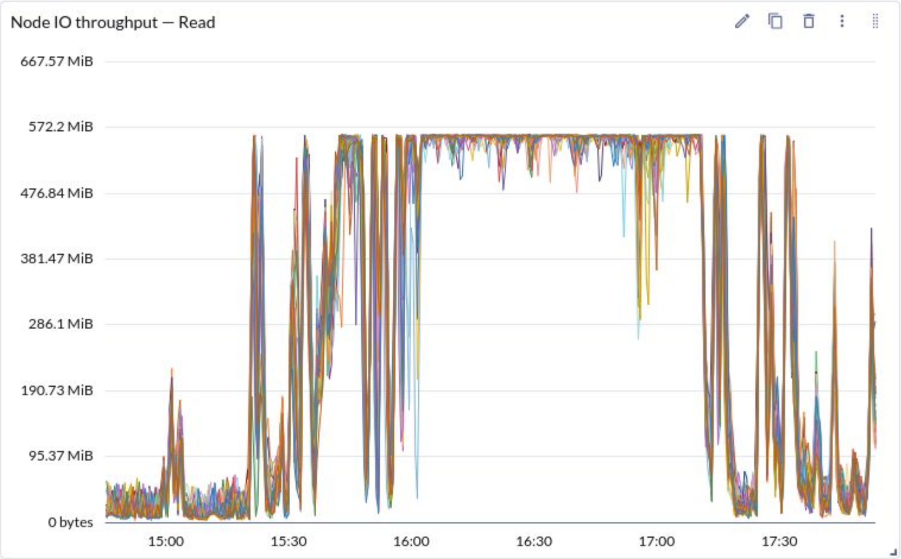
image.png ▼

Commit Log Queue Length

Surprises came up

Read Ahead!


Server Side Fetch RPC Latency (p99)
15s
12s
9s
6s
3s
0s
15:00
15:30
16:00
16:30
17:00
17:30


Node IO throughput — Read
667.57 MiB
572.2 MiB
476.84 MiB
381.47 MiB
286.1 MiB
190.73 MiB
95.37 MiB
0 bytes
15:00
15:30
16:00
16:30
17:00
17:30

Read Ahead!

```
/sys/block/<...>/queue/read_ahead_kb
```

```
128
```

```
/sys/fs/btrfs/<...>/bdi/read_ahead_kb
```

```
4096
```

Can you tell the difference?

Timeline

- 2024-06: Wrote a doc and redirected everyone there.
  - Every question ever asked was put to the doc.
- 2024-08: kick off infra work.
- 2025-01: first internal cluster.
- 2025-06: first customer cluster.
- 2025-08: mass migration.
- 2025-10: done.
- 2026-01: yours truly @ FOSDEM.

Takeaways

- All metrics data @ Chronosphere is on Btrfs.
    - Our savings: 65% predicted, 74% actual.
- Btrfs is available upstream in GCP:
    - (CSI) driver and Container Optimized OS (COS).
    - Planning to go full-upstream in the next 1-2 months.
- You know one more company trusting Btrfs.

Special thanks

- Boris Burkov, Josef Bacik from Meta.
  - Reclaim (defragment).
  - Spreading the word!
- All Btrfs maintainers.
- GCP Storage team for being Btrfs champions.

Motiejus Jakštys, Chronosphere
m.jakstys.lt

@motiejus:jakstys.lt

Bonus slide: unknown unknowns

- Filesystem deadlock on significant memory pressure.
  - Fixed in mainline and backported to all stable kernels.
- Disk snapshots grew by 6-10x.
  - Replaced with file-based snapshots.