Hybrid XFS: Supercharging HDDs with SSDs

Skanda Shamasunder
Production Engineer (Storage), Facebook
Who am I?
Outline

1. The IO wall
2. The Opportunity
3. The Risks
4. The Analysis
5. Rolling it out
6. Lessons
The IO wall
IO workloads are only getting hotter

Thanks, ML

- AI/ML analytics
- Ephemeral photos
- Live Video
- Offloading Cold Data
Capacity-bound

IO-bound

Time

IO Wall

IOPs demand

Bytes demand
The Opportunity
41 N  <---- Un-categorized
8571 R  <---- Data reads
14 RM  <---- Metadata reads
4773 WM  <---- Metadata writes
6901 WS  <---- (Synchronous) Data writes
Options to reduce metadata IO

- Write less data
- Smaller metadata
- Fewer flushes for writes
- Create a new filesystem
- Put the metadata elsewhere
XFS Realtime mode - a hidden gem
Traditional XFS

- Metadata
- Intent Log (Journal)
- Data Blocks

Data Sub-volume
- Metadata
- Intent Log (Journal)
- Data Blocks

XFS w/ Realtime Subvolume

Realtime Sub-volume
- Realtime Data Blocks

image by Richard Wareing, Facebook
But... we only have one SSD
Partition the heck out of it
And gave it a name - Hybrid XFS
Keep your eye on the Random Writes
<10 coalesced IOs/sec; more than 50% reduction

Nearly eliminated
DENIED
The Risks
1. SSD Failures
2. Endurance
3. Hardware changes?
4. Operational headache
Risks.....
Many many risks

• Changes in workload?
• Rollout?
• and...?
Let the Data guide you
The Analysis
1. SSD Failures

1%  1

5%  10

1000
1. SSD Failures

\[ \text{1000} \times 1\% = 10 \text{ SSDs or 100 HDDs} \]

\[ \text{10000} \times 5\% = 500 \text{ HDDs} \]

\[ \text{\cancel{1500 HDDs}} \]

\[ \text{\cancel{600 HDDs}} \]
1. SSD Failures - en masse!
1. SSD Failures

- Rescue mode
- Metadata
- Data
2. Endurance

*DWPD = Drive Writes Per Day
3. Will there always be SSDs?
4. Operational headache

- Patched XFS statfs call
- Collect stats for both metadata and data
- Make systemd wait for BOTH devices
How do you roll out a destructive change to tens of thousands of hosts?
PLEASE BE CAREFUL!

PREVENT WILDFIRES

NEW JERSEY FOREST FIRE SERVICE
Look out for...

- Impact to:
  - Durability
  - Capacity
  - Performance
- Fallout from failures
- Automation flying blind
Success!
“Hard problems can have simple solutions”
“Gut feelings can be wrong”
“Data wins arguments”
“Better safe than sorry”
Thank You!