Container Images Considered Harmful

I'm sorry, I couldn't resist.

Aleksa Sarai
Senior Software Engineer — SUSE
asarai@suse.de

SUSE

OPEN CONTAINER INITIATIVE

Who is that maniac? Get him off the stage!

- Wait! Let me explain!
- Effectively every container ecosystem uses `tar` archives for layers.
  - And to be frank, they suck … a lot.
- The full rant is about 3 hours long and is *definitely* R-rated.
  - I'll try to condense it into 5 minutes.
  - If you want more grizzly details, buy me a beer.
- There are lots of other issues with how opaque they are, but we don't have time.

In my defence, I made this late last night.

Why does `tar` suck?

- Not standardised. At all.
  - It has three different (effectively incompatible) extension formats.
  - They have partially overlapping feature-sets, and different support levels.
- Order of `tar` entries is implementation-defined and might not be reproducible.
  - GNU `tar` generates different archives to Go's `archive/tar` with trailing bits.
  - Go 1.10's `archive/tar` creates slightly different archives to Go 1.9's.
- You can even have duplicate entries for the same file in a single archive!

Why does `tar` extraction suck?

- Each entry is stored one-after-another.
  - Finding the header and contents of a particular file requires a linear scan.
- **Cannot** be extracted in parallel, so it's very slow.

Why does `tar` layering suck?

- Any metadata change causes the whole file to copied to the new layer's archive.
- Layer-level deduplication is effectively useless.
  - Any small change breaks it for the whole layer.
  - Not to mention that different distributions ship bit-similar software.
- Files changed in multiple layers are extracted multiple times.

Can we do any better?

- Of course!
- Backup tools have solved this problem for at least a *decade*.
  - With a more clever indexed format to allow parallel extraction.
  - Using content-defined chunking with rolling hashes for **intra-layer** deduplication.
  - Content-addressable store means we can get **intra-image** deduplication.
- And a new format means we can make it sanely extensible!