Intel® Lustre* Layout Enhancement
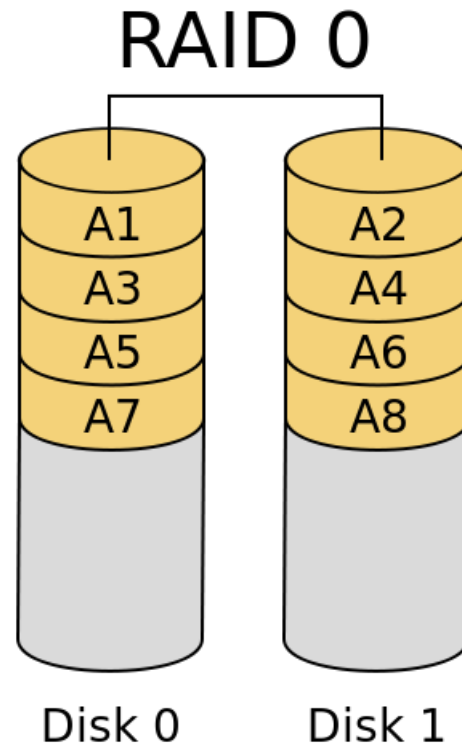
John L. Hammond

April 8, 2013

Outline

1. What is *layout* and why enhance it?

2. Composite Layouts

3. Layouts for Widely Striped Files

What is Layout

Layout is file striping metadata

– Stripe size

– Stripe count

– Stripe pattern

– Pool

– List of object identifiers

 (and OST indices)

RAID 0

Disk 0   Disk 1

What is Layout (2)
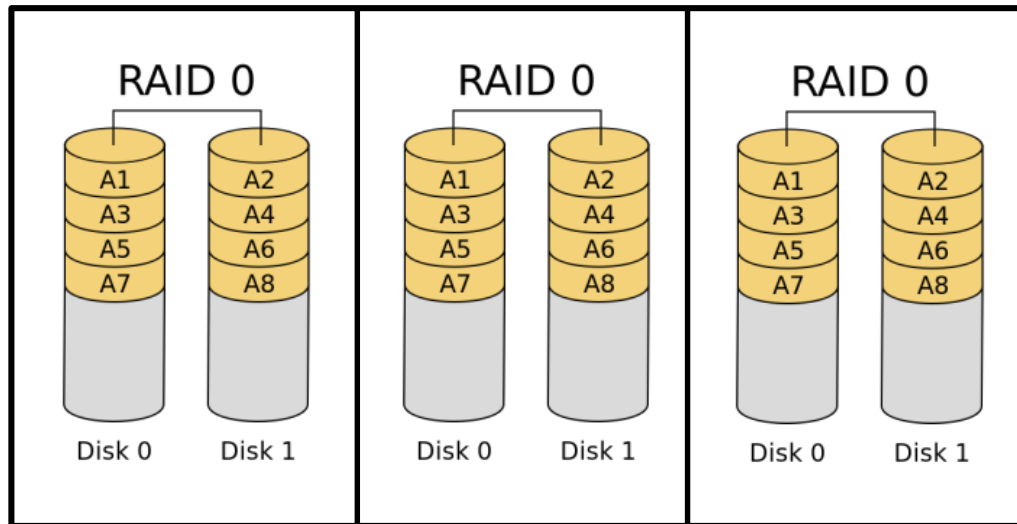
struct lov_mds_md_v1

    u32   lmm_magic

    u32   lmm_pattern   /* RAID-0 */

    u32   lmm_stripe_size

    …

    /* followed by array of object identifiers */

Stored as an extended attribute ("trusted.lov") of the file on the MDT inode.

(intel)

Why layout enhancement (1)

File Level Replication

- bundle v1 layouts to replicate file data

Why layout enhancement (2)

Extent Based Layouts

- Different layouts in different regions

0   1MB
∞

First 1MB on SSD backed OST, rest on SATA

Why layout enhancement (3)

Extended attribute (EA) and RPC limits:

- ~64KB upper limit on EA

- 2000 stripes < 48 KB layout (12 4K pages)

- 64K stripes > 1.5MB > LNET_MTU

Composite Layouts (1)

Add a new layout type: struct lov_comp_md_v1
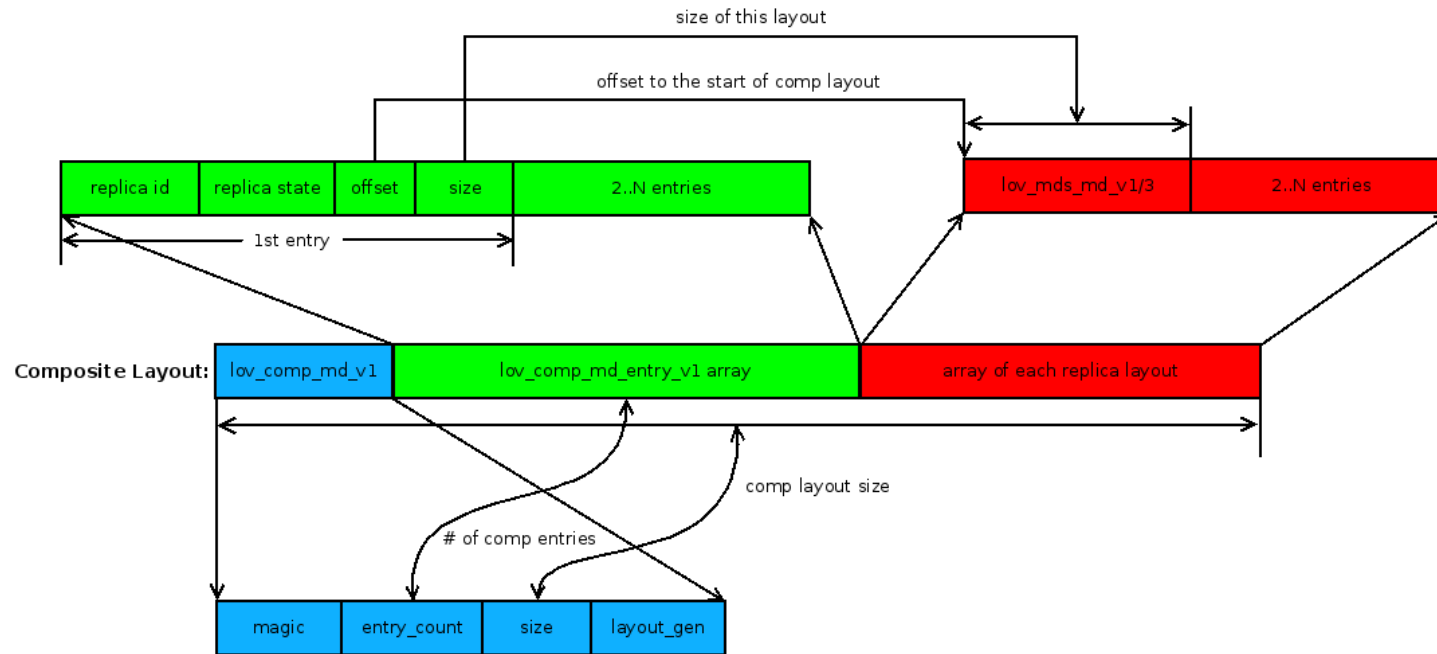Bundle simple (v1/3) layouts together

Diagram of composite layout for replication

By Jinshan Xiong

Composite Layouts (2)

Add a new layout operations

- Add, move, remove entries (replicas) from a composite layout

- Iterate, select entries with various properties

- Operations to support replication
  - Get and set bits encoding replication state

Compact Layouts for Widely Striped Files

- Allocate a contiguous range of FID sequences s_0, s_1, s_2, …. such that s_i belongs to OST i.

- struct lov_wide_md_v1
  - stripe count and size as before
  - sequence (begin, end, stride)
  - single OID shared by all objects
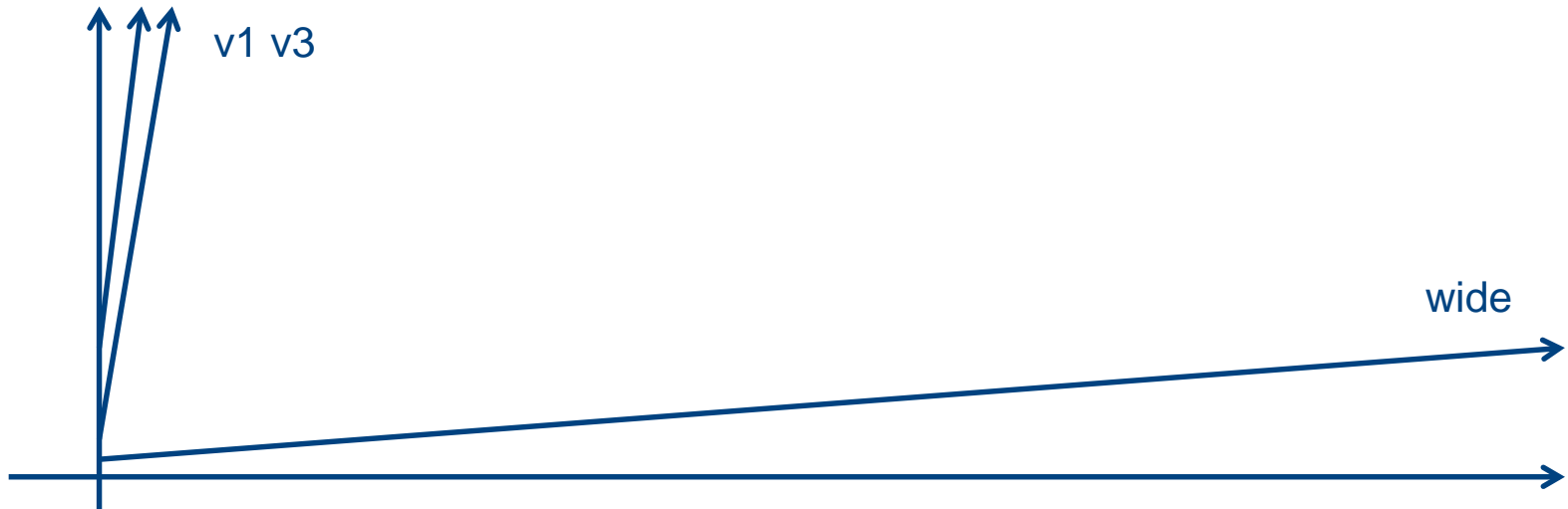 - bitmap of OST indexes

Compact Layouts for Widely Striped Files

type	size in bytes
struct lov_mds_md_v1	32 + 24 * stripe_count
struct lov_mds_md_v3	48 + 24 * stripe_count
struct lov_wide_md_v1	~72 + stripe_count / 8

v1 v3

wide

Questions?

Thanks!