Chunkfs:
Repair-driven file system design

Val Henson
val@vahconsulting.com
VAH Consulting

Fsck time is growing

Disk hardware improvements, 2006 - 2013

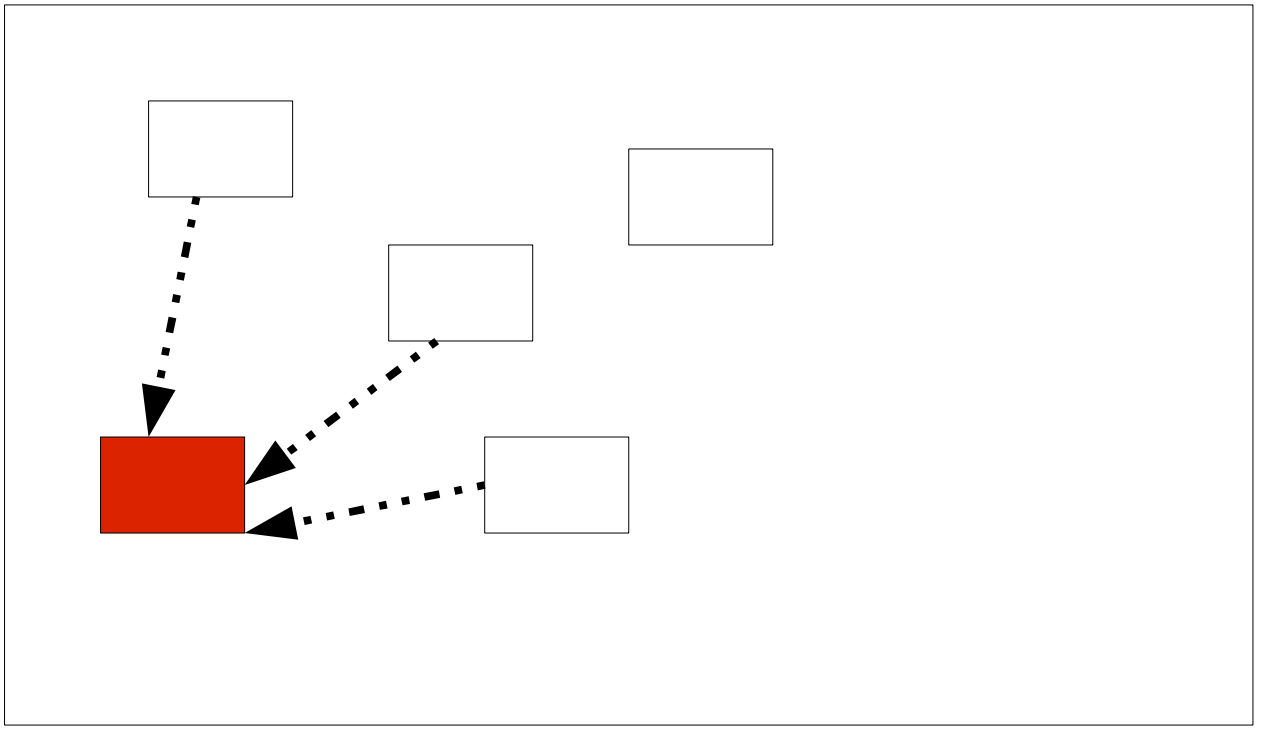
Capacity: 16.0x

Bandwidth: 5.0x

Seek time: 1.2x

=> **10x** increase in fsck time!

Why is fsck O(file system size)?



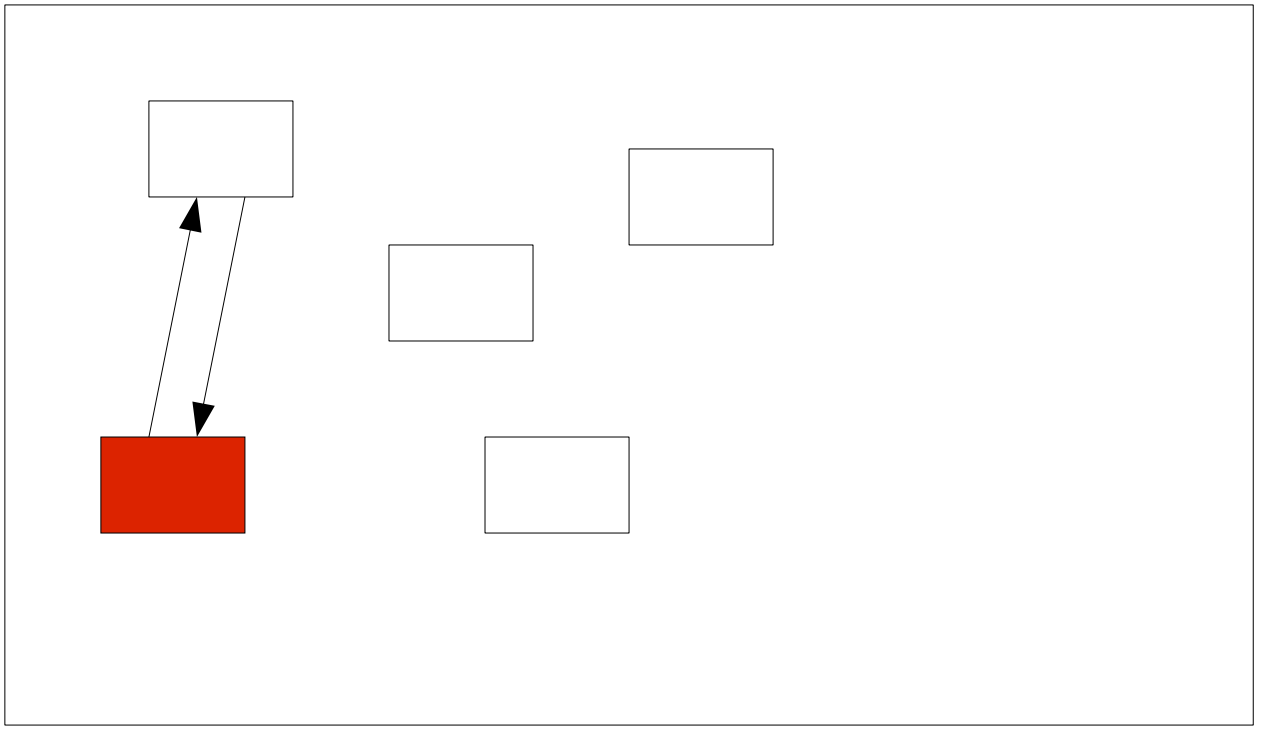
Fsck optimization

- xfs_repair: 75% improvemnt
- e2fsck: 50% improvement
- Erased in 1-2 years

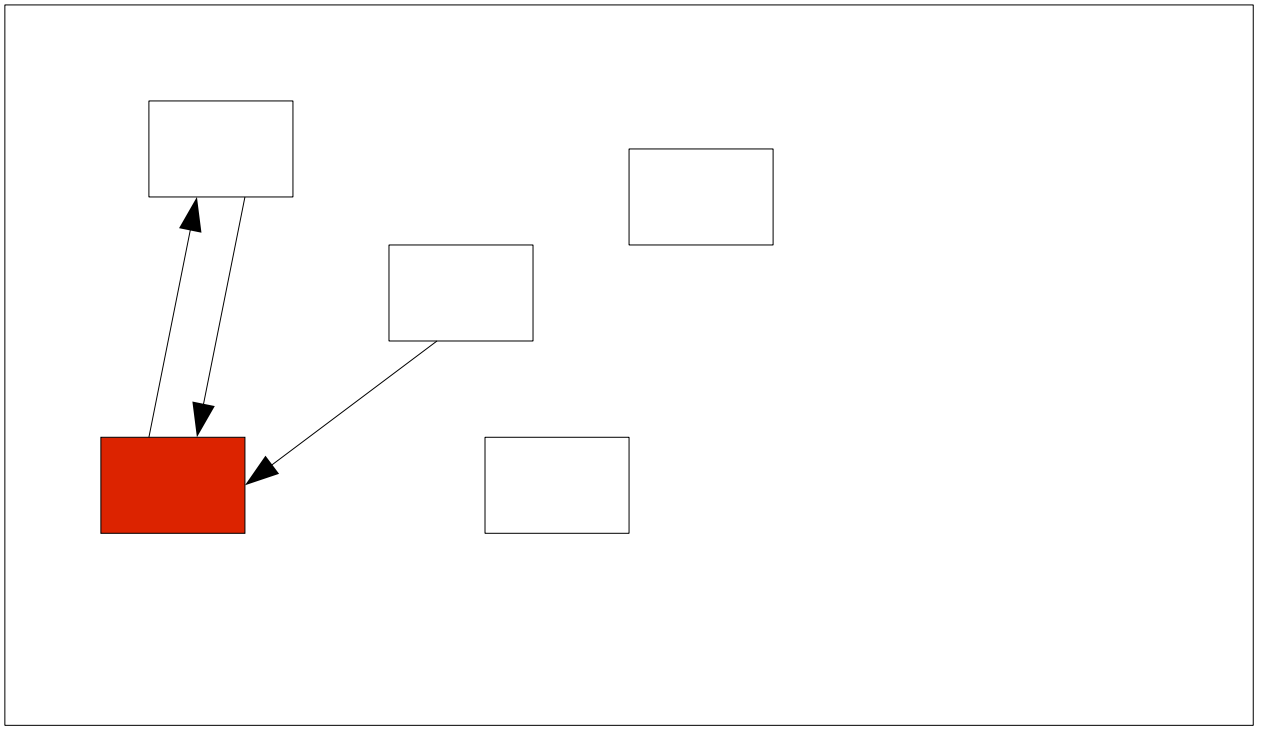
Repair-driven file system design

- On-disk format designed with repair in mind
- Simple data structures
- Optimizations for reading data for repair (e.g., metadata bitmap)
- Fast, incremental file system check
- Checksums, redundancy, scrubbing, etc.
- Metadata isolation

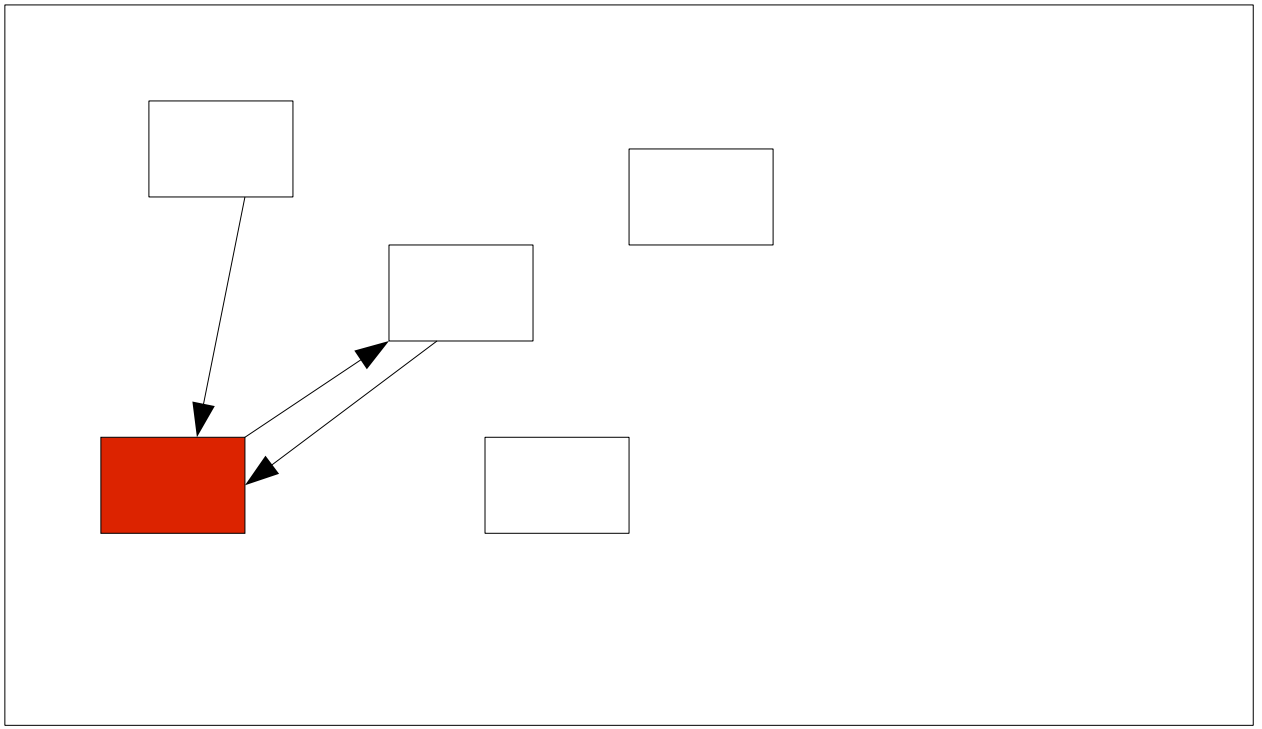
Need fast check AND repair



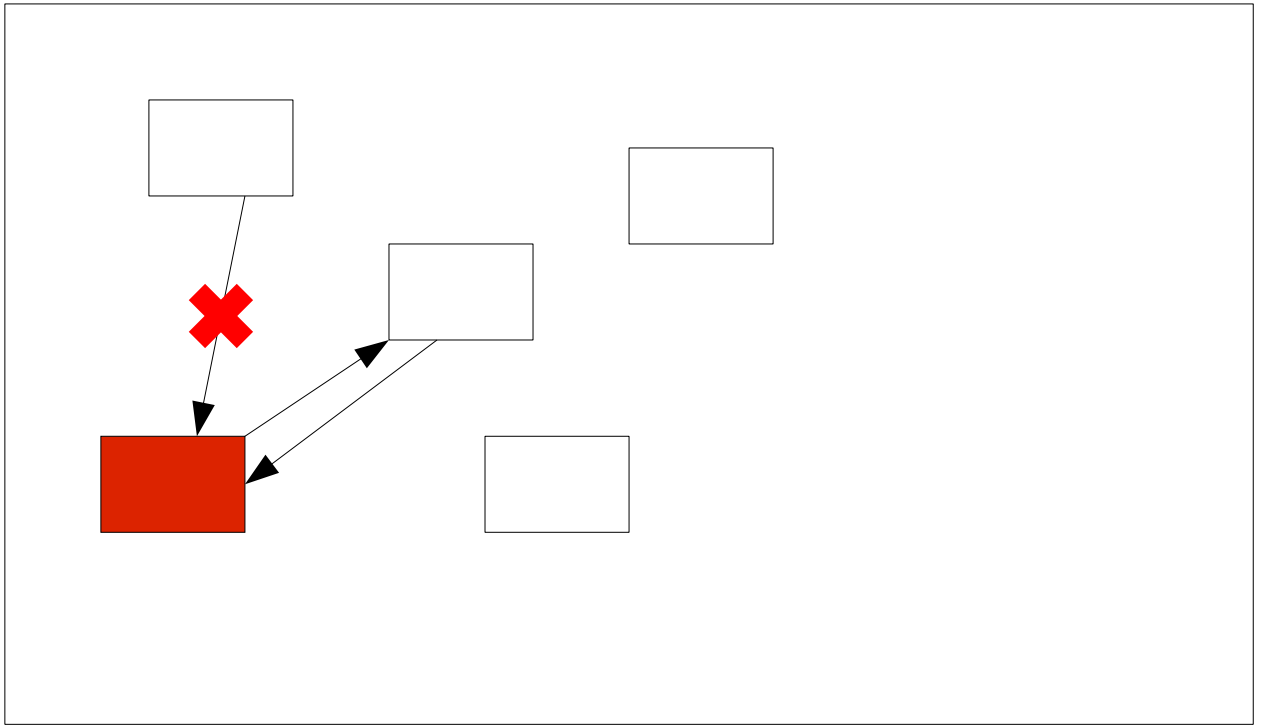
Double allocation bug



Double allocation bug

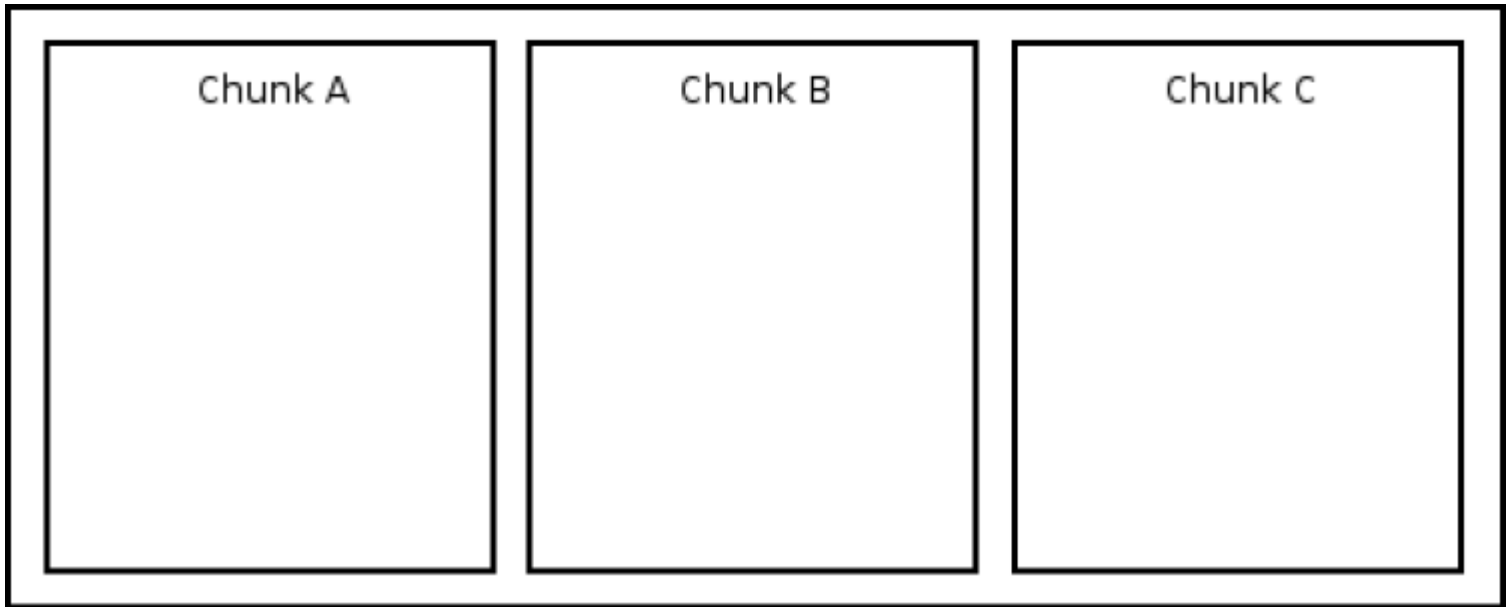


Double allocation bug

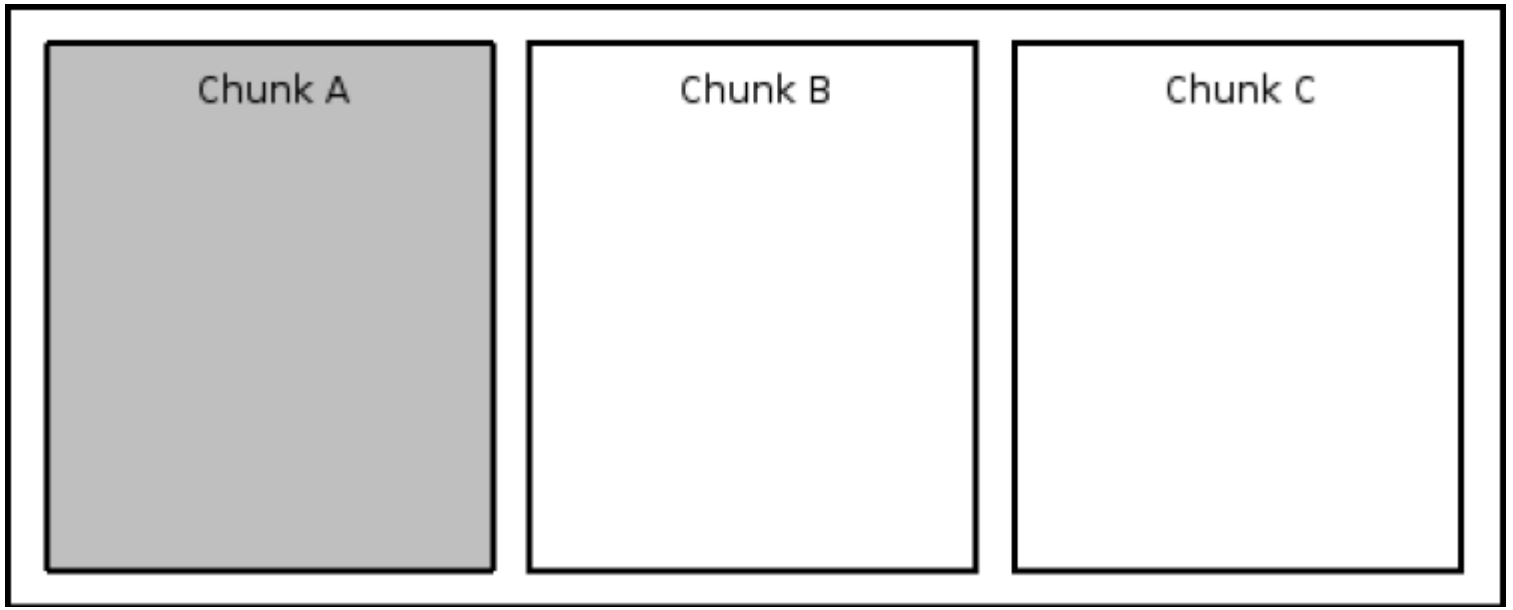


Chunkfs

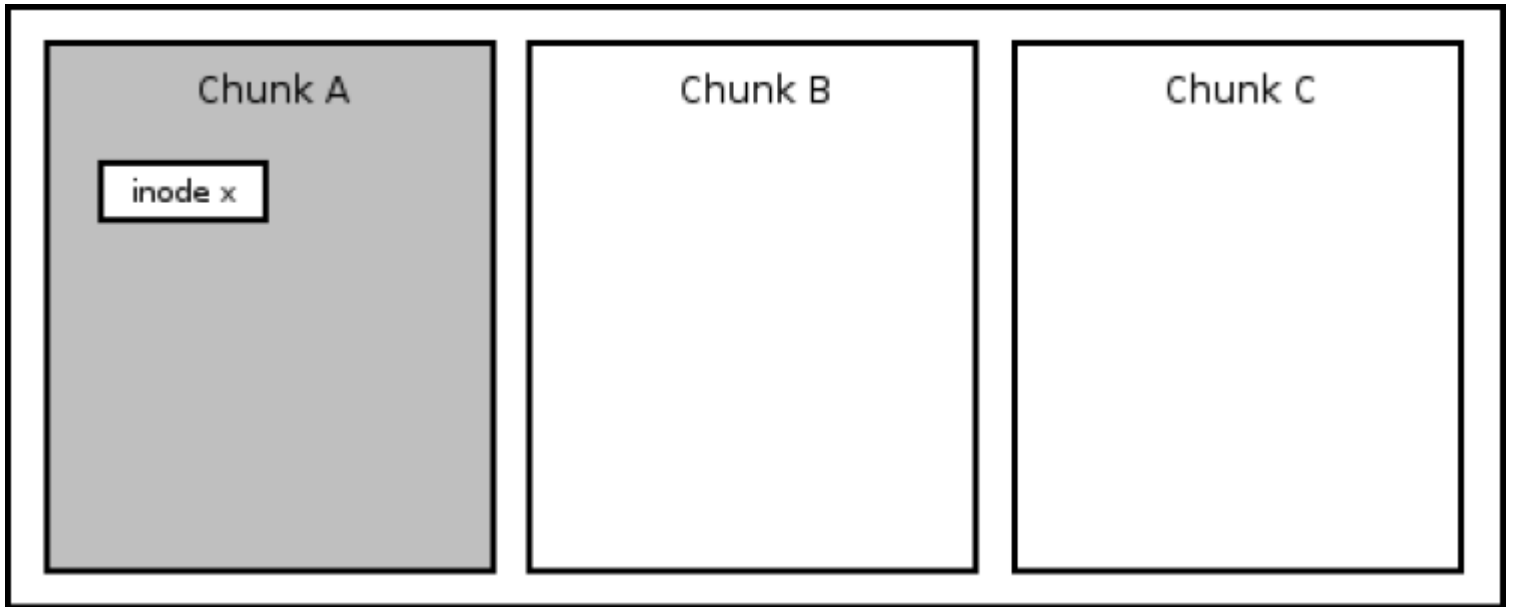
Divide fs into metadata isolation groups - chunks



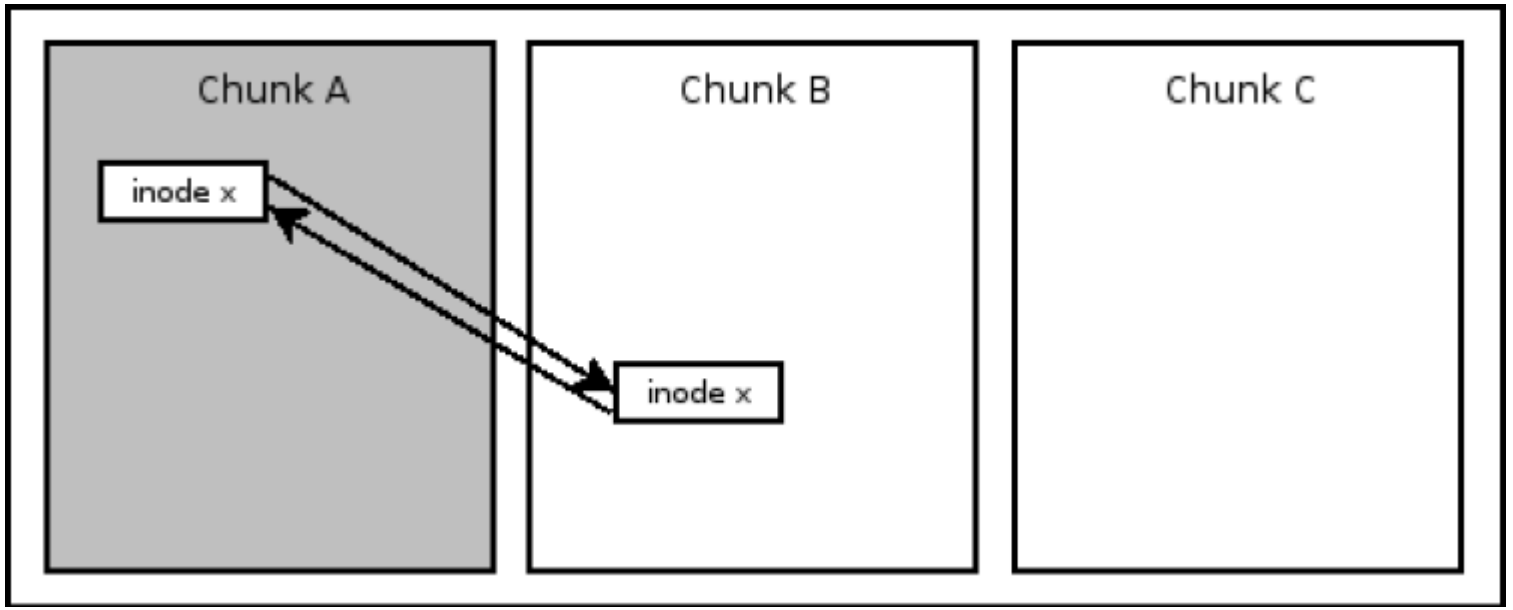
How to glue it back together?



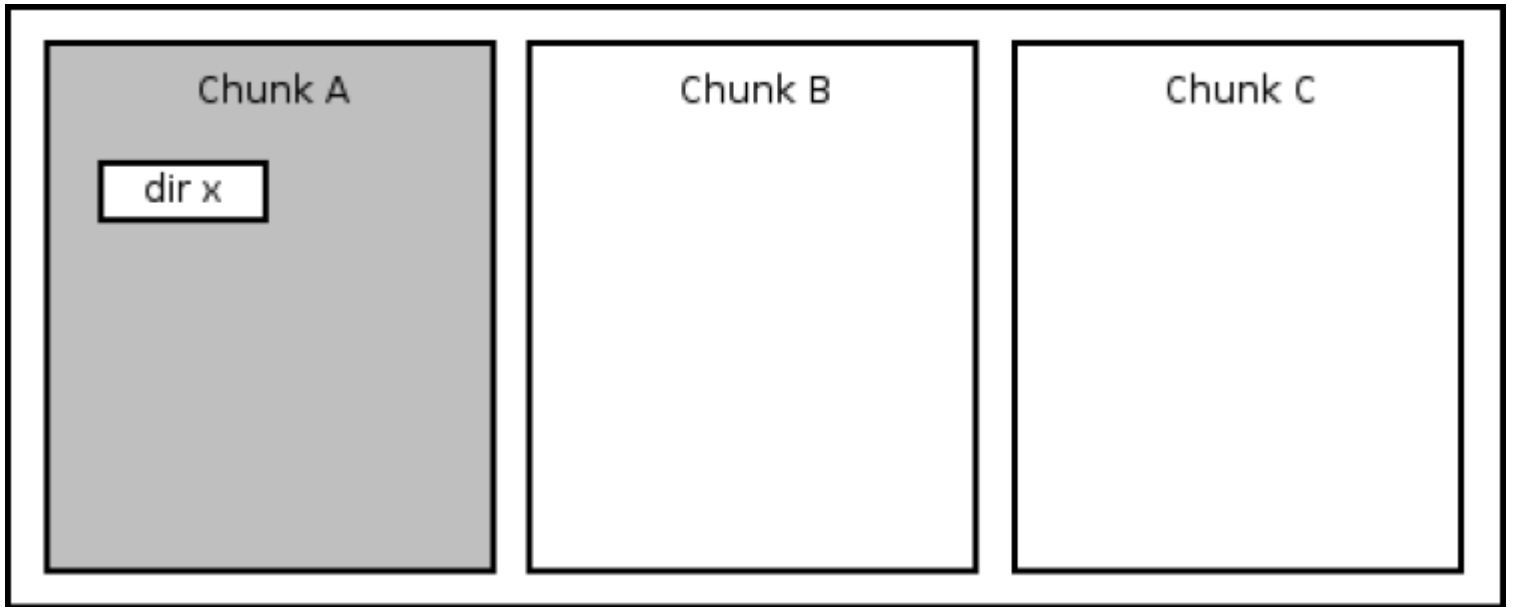
Problem: File data outgrows chunk



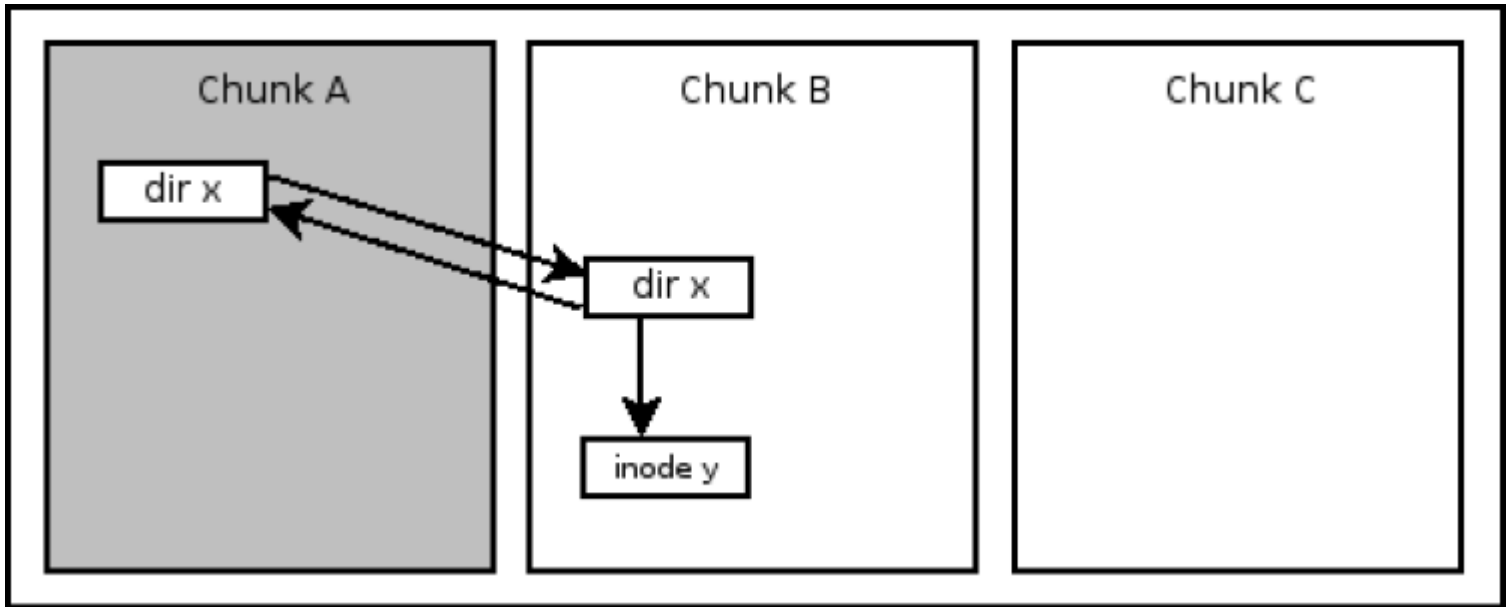
Solution: Continuation inodes



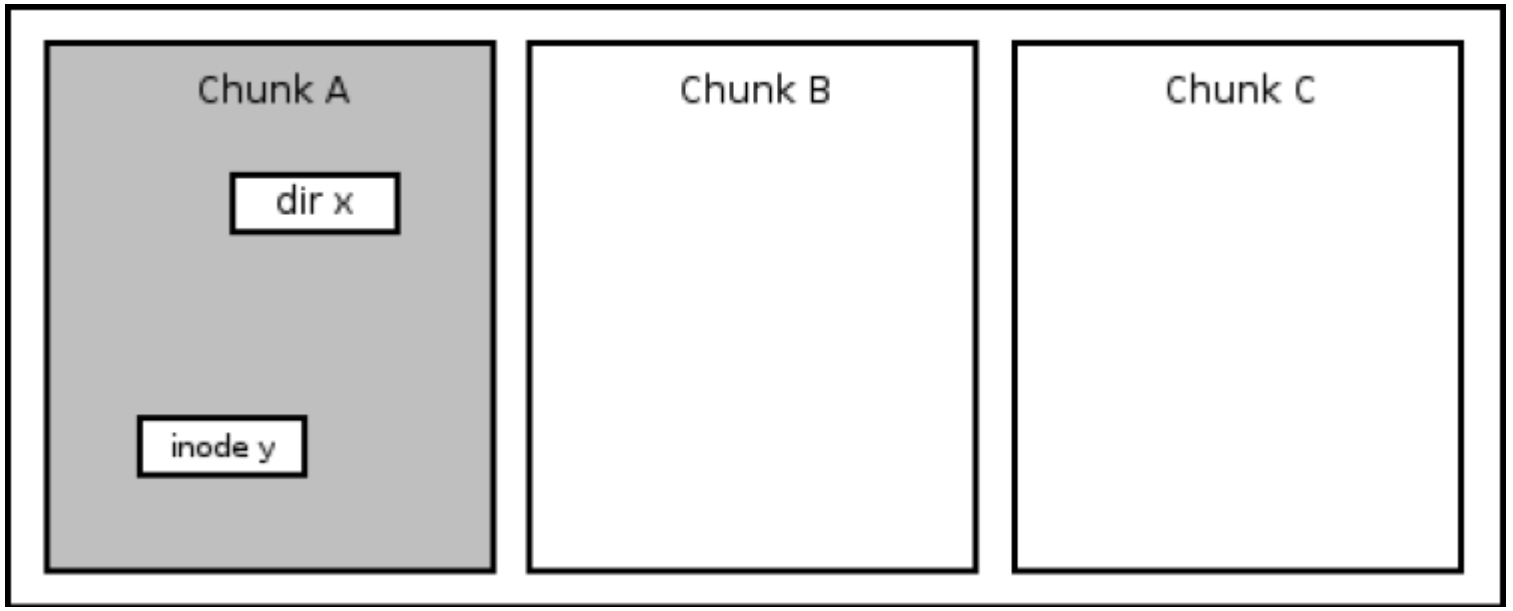
Problem: Directory outgrows chunk



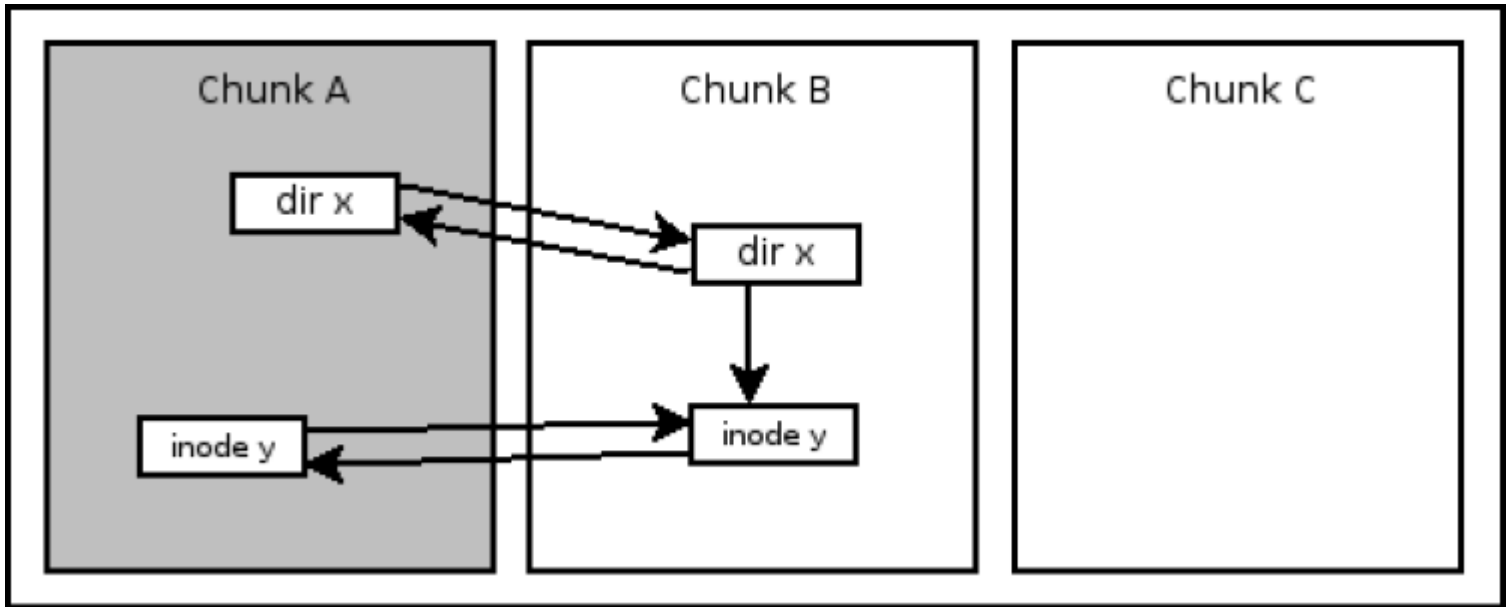
Solution: Continuation inodes



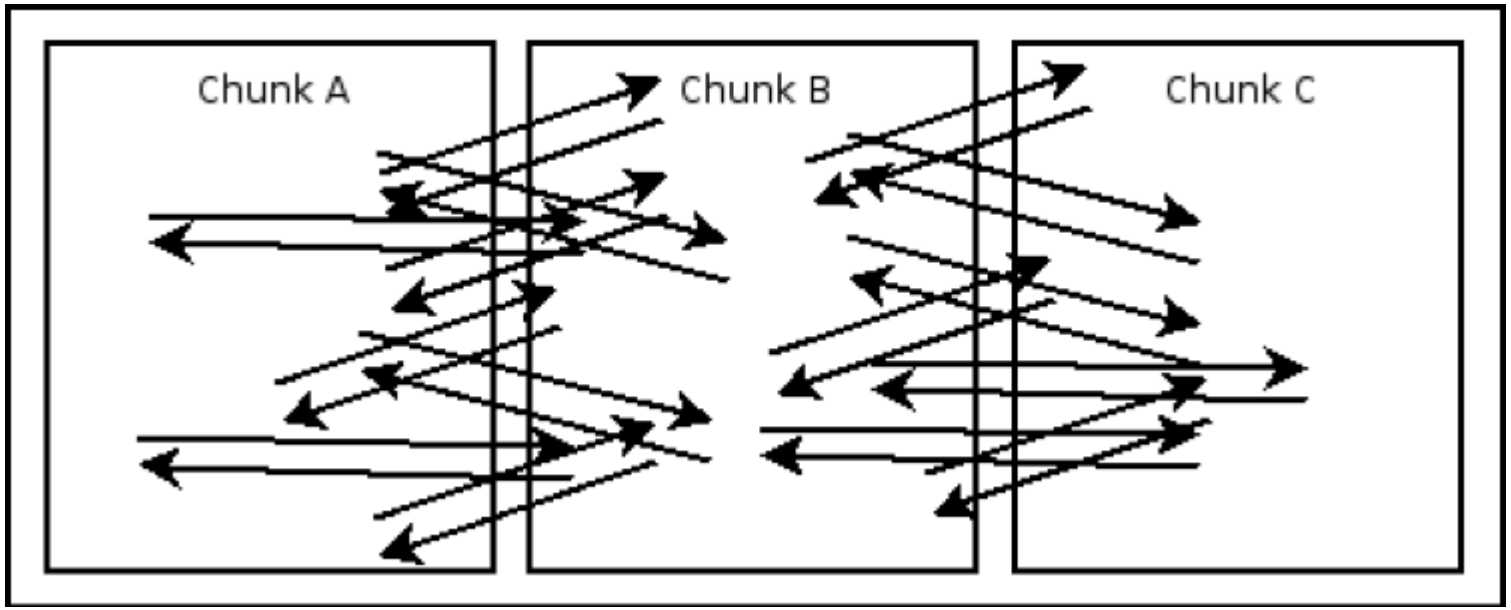
Problem: Hard link to inode in full chunk



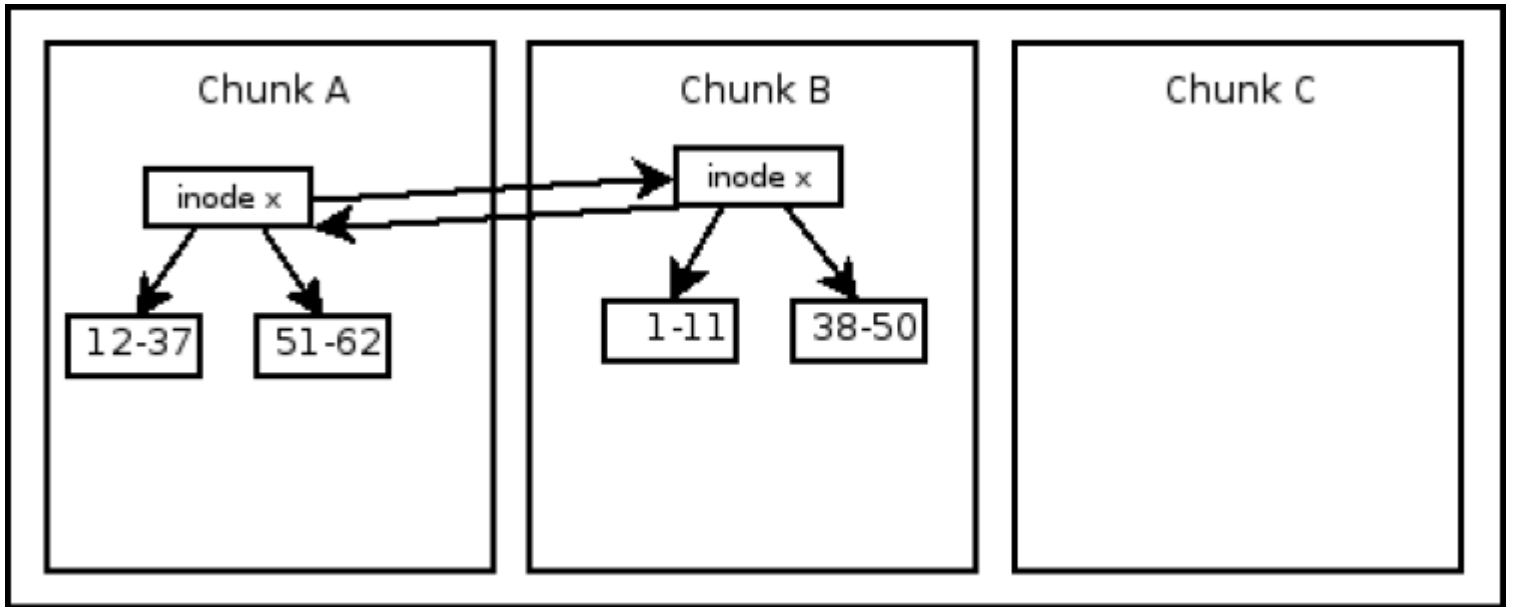
Solution: Continuation inodes



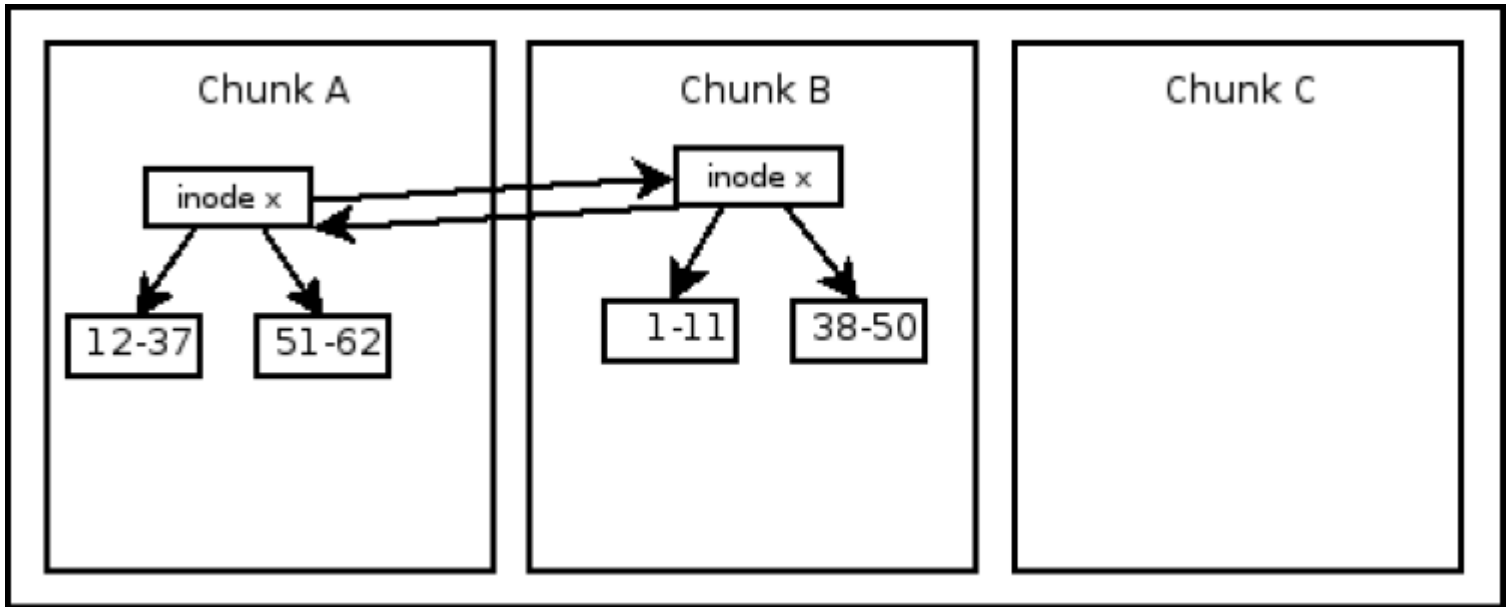
Problem: Too many continuation inodes



Solution: Smart allocation & sparse files



Problem: Quickly finding file offset



Solution: Embedded lookup structure in continuation inode

```
struct chunkfs_inode {
    ...
    struct continuation_data;
}

struct continuation_data {
    ...
    struct tree_node;
}
```

Implementing chunkfs

- Options:
  - Hack into existing file systems
  - Layer on top of existing file systems
  - Design into fancy new file system - btrfs:

    http://oss.oracle.com/projects/btrfs/

Demo

Thanks!

- Funding: Intel, EMC^2, VAH Consulting
- Design: Arjan van de Ven, Zach Brown, Theodore Y. T'so
- Code: Amit Gud, Karuna Sagar
- Making file systems seem cool: Jeff Bonwick

Chunkfs code, docs, etc.

http://valhenson.org/chunkfs

Q &A

val@vahconsulting.com