Small is Beautiful
or
Workloads Rule!

Erez Zadok

File systems and Storage Lab

Stony Brook University

http://www.fsl.cs.sunysb.edu

FS Complexity Growing

- More file systems being developed
  - Over 60 in Linux
    - From 1-2Kloc to 77Kloc
- FS becoming kitchen sinks
  - ext4: journalling, extents
  - reiser4: plugins
  - btrfs/zfs: storage pool mgmt, encryption, compression, dedup, checksumming, RAID-like, etc.

System Complexity Growing

- More virtualization layers
  - OS, LVM, RAID, networks
- Really hard to analyze complexity
  - OSprof, DARC, MDS/visualization, etc.
- App workloads perturbed
  - looks more “random” in lower layers
  - *“Does Virtualization Make Disk Scheduling Passe?”*

Study: Impact of Workloads

- Linux server
- FS: ext2, ext3, reiser3, xfs
- Vary mount options:
  - journalling, noatime, notail, etc.
- Vary format options:
  - AG size, inode/block size/number, etc.
- Filebench workloads:
  - Web server, OLTP, mail server, file server
- Analyze ops/sec and ops/joule

Study Results

- Default options often suboptimal
  - 50% improvement for same FS
- Change FS, mix mount/format options
  - as much as 9 times improvement
- ext2/3 didn't win for any workload
- reiser3 and xfs best for 2-of-4 workloads
  - B-trees
- LoC:
  - ext2 8k, ext3 24k, reiser3 27k, xfs 77k

Ask the Scientist: FS use?

- Asked colleagues
  - Neutron and X-ray imaging, bio-molecular, structural biology, optical microscopy, macro-molecular imaging, 3D cryo-electron microscopy, astrophysics, and the HDF Group
- run their own small clusters:
  - 10s/100s nodes
  - rent time on larger clusters

Ask the Scientist: Results

- hardlinks, softlinks, EAs/ACLs, open-unlink-close, rename dirs: no
- deep directories: no, often flat
- file names: known names/lengths
- file sizes: known input and output sizes
- reliability, journalling: mostly don't care
  - checkpointing, restart experiment
- Preferred FS: don’t care
- Etc...

You Too Can Develop a FS

- Graduate OS class
- 4 teams of 2-3 first-year MS students
- Develop very simple real FS (VSRFS)
  - fixed/variable no. of files
  - fixed/variable file sizes
  - no directories vs. simple directories
  - partition disk into several large extents
- 3-4 calendar weeks; 1-2K LoC

➔ Dev-time non-linear wrt LoC

STONY BROOK
STATE UNIVERSITY OF NEW YORK

Recommendations

- App/workload specific optimizations
  - *"A Case for Versatile Storage System"*
- "strip" whole systems down to core features needed (slicing?)
- Custom FS, simple and small
  - auto-gen code from high-level language?
- Overhaul POSIX?
  - OS, FS community; LSF workshops
  - hard-to-implement features with little use