Gelato @ UNSW

Performance and Scalability on Itanium

www.gelato.unsw.edu.au

Which Filesystem????

Peter Chubb

Gelato Project

National ICT Australia

The University of New South Wales

October 2005

Linux Filesystems

- Dos-like filesystems: VFAT family, isofs, UDF
- ‘inode-based’ filesystems: Ext2
- Journalling filesystems: Ext3, XFS, JFS
- Weirdos: ReiserFS versions 3 and 4, freevxfs
- Special purposes: JFFS, ROMFS,
- Plus others for cross-machine portability: amigaFS, befs, hfs, qnx4, Minix, SysV, UFS

Linux has many different filesystems available. Which one do you use? The DOS-like filesystems are mostly for data transfer (USB stick, compact flash, floppy, DVD or CD); but of the others ... which?

Data Integrity

(and performance)

You need to bear in mind that the essential use for a filesystem is to keep your data safe. If it doesn't cope with the kinds of stresses (power failures, disk failures, controller failures, etc.,) that you environment may impose, then it's almost useless.

If you have no backups you will be sorry.

RAID is **not** a substitute for backups.

Keith Bostic used to tell a story sometimes, when he was on the helpline for OpenBSD. A woman called him, and wanted to know how to get the data back after she'd accidentally reformatted the disk. He asked her if she'd read the manual.

Woman: Yes
Keith: What does it say on page two?
Woman: 'If you have no backups you will be sorry'.
Keith: Do you have a backup?
Woman: No
Keith: Are you sorry?
Woman: Yes
Keith: There you are then....Works as advertised

I should also point out the obvious... RAID protects only against

some classes of problems. It wouldn't have helped that woman at all.

Trusted, Tried and True...

- (minix)
- ext2
- ext3 (ext2 plus journal)
- XFS
- (JFS)

Noone much uses the Minux filesystem any more, because of its low performance and short file name limitation.

Many distributions install on ext2 or ext3 by default.

XFS has been used on 64-bit machines for many years.

JFS doesn't work well on IA64 (yet).

ext{2,3}

- ext2 very stable.
- ext3 has journal, otherwise same as ext2.
- Good for small systems.
- e2fsck can fix most problems.

The really nice thing about ext2 is that its performance is good (at least for small numbers of spindles) and that fsck can fix almost all errors. It has a reputation for being a very robust and full-featured filesystem.

ext{2,3} Limits

Filesystem Block Size	1k	2k	4k
Max file size	16GB	256GB	2048GB
Max filesystem size	2047GB	8192GB	16384GB

Possible Problems with $>$ 2TB on 32-bit platforms

With 4k fileystem blocks, you can have a single 2TB file, and a filesystem up to 16TB. However, on 32-bit systems I have heard unconfirmed reports of filesystem corruption when a filesystem bigger than 2TB is configured.

ext{2,3} tradeoffs

+ Very mature in Linux
+ Good small-scale performance
+ Mostly benign failure modes
– Doesn't scale well to large filesystems
– Poor performance under large-scale SMP
– Poorer than optimal performance with RAID
– Needs occasional fsck even with journal

XFS

- From SGI — long history of use in IRIX
- 64-bit clean
- Journalled FS
- Scales well
- Extent-based
  - Can preallocate contiguous region on disk
  - Possibility of a *real-time* partition

XFS is newer to Linux, but is a mature filesystem introduced first in Irix 6. It was designed to replace SGI's EFS (extent-based filesystem), removing EFS's limitations while maintaining the good performance and scalability.

XFS tradeoffs

+ Scales to 9EB (9.10^{18} bytes)
+ Performs very well over RAID
– Large, complex code base
– Disk/Power Failures can leave blocks of nulls in recently changed files.

I wouldn't use XFS on my laptop or PDA — the extra complexity leads to added battery drain. But for almost every other prupose it is my filesystem of choice.

ReiserFS v3

- New kid on the block.
- Novel storage mechanisms: *Not block based*
  - Good for lots of small files?
- ReiserFS version 4 around the corner...

ReiserFS uses variants of the B-tree for all file and metadata storage. This makes it more prone to corruption on error, but allows packing data much more tightly. ReiserFS's stated goals are to solve the 'lots of small files' problem.

Version four is apparently ready for inclusion in Linux now. I haven't tried it yet.

ReiserFS3 tradeoffs

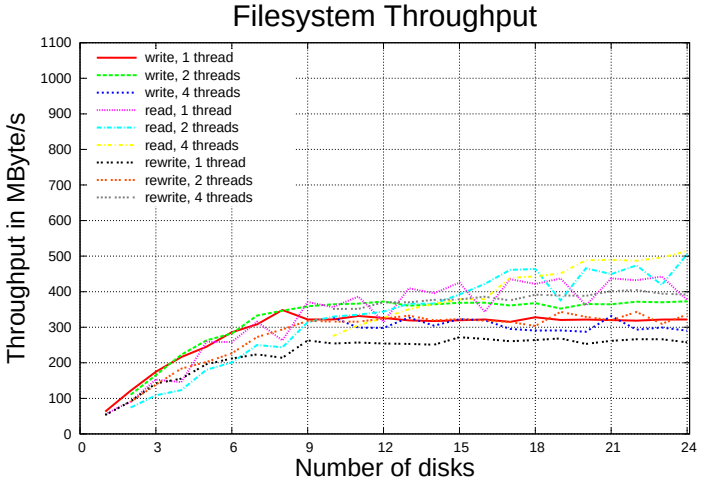
+ Reasonably fast
+ Less internal fragmentation than others
– poor scaling over RAID or to many processors.

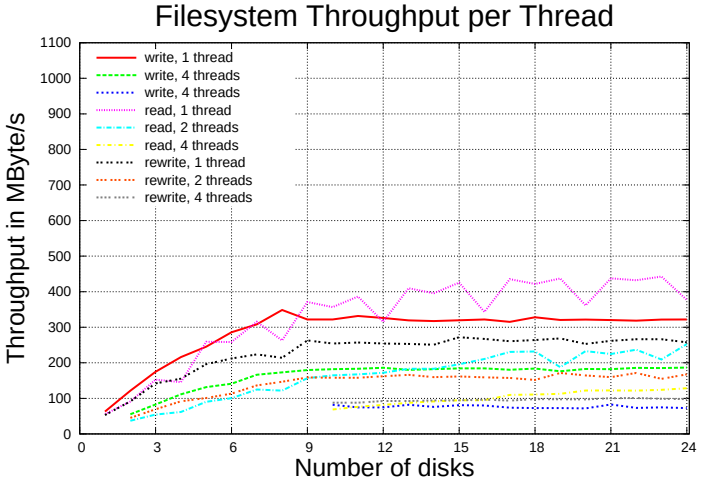
On testing ReiserFS version 3 we found it scaled neither to large numbers of processors nor large numbers of spindles in RAID.

Others

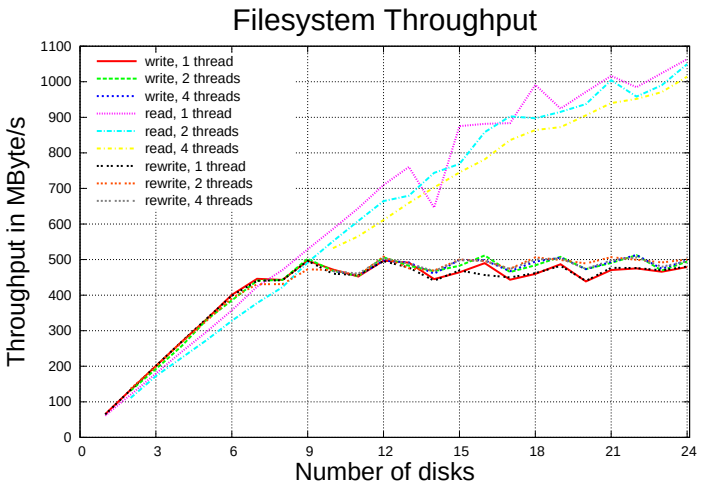
JFS	fails to work with *pagesize* $\neq$ 4 k
NTFS	Only really good for dual-boot machines at present...
VFAT	For file exchange. Non-unix semantics.
JFFS	For NAND flash ram — not usually found on servers or desktops.
tmpfs	Ram-based, for `/tmp` or `/dev/shm`
hugeTLBfs	Ram-based, mapped using huge pages. Primarily for Oracle :-)

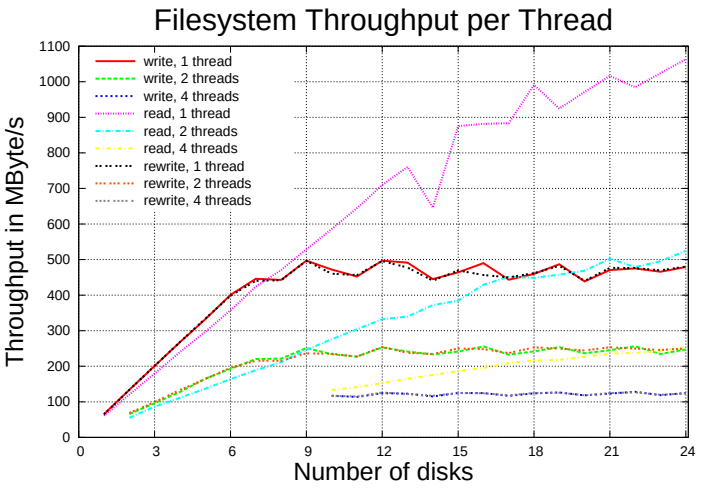
ext3 performance





XFS performance

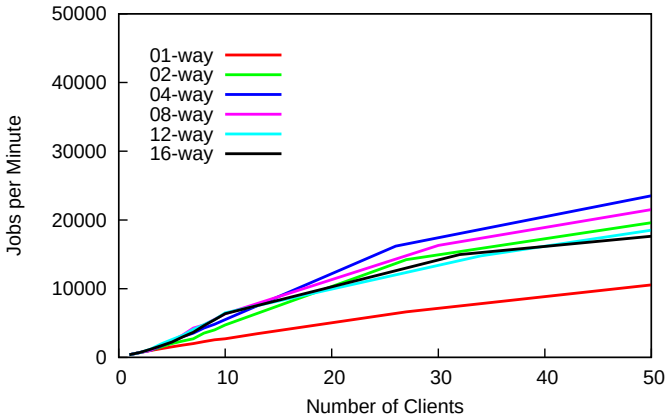




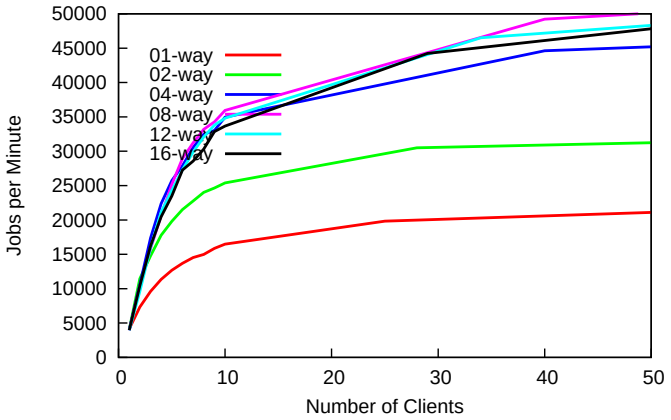
These graphs show IOZONE results XFS and Ext3, using the deadline scheduler and a 64k linux page size. The setup uses three 3ware controllers and up to 12 disks on each channel, in a RAID0 configuration.

OSDL AIM7 results

ext3



XFS



Performance Compared

File system	Write	Rewrite	Read
XFS	490	492	1089
ext2 (1M, deadline)	495	290	431
ext3 (4M, as)	330	264	386
ReiserFS v3	498	441	444
JFS	35	46	23

I’ve chosen the best figure from each graph.

File system	Write	Rewrite	Read
ext3 (256k, deadline, wb)	346	295	577
ext3 (1M, deadline, wb)	360	296	424
ext2 (1M, as)	485	289	431
ext2 (4M, as)	498	280	121
ext2 (4M, deadline)	498	276	123

In conclusion

- Horses for Courses
- Test for your workload
- Test on your disk array
- XFS seems a good bet for high performance systems...
- On single spindle, few processors not much difference

Thanks To

Andreas Hirstius of CERN – Disk Benchmarking

Darren Williams from Gelato@UNSW – REAIM7 benchmarking

This work was supported by HP, UNSW, the ARC, and National ICT Australia.

Quick update on other work

- Superpages — Mostly working!!
- IPbench/NFS testing — started
- NUMA Visualisation — stalled
- New Page Table Interface
  - Early version sent to Linux-MM list
  - Guarded Page Table works for some radices