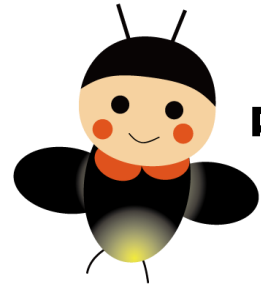


中华数据库
行业协会

2014中华数据库与运维安全大会

官方网址：**www.zhdba.com**

Linux的文件系统对比

目标和范围

普通PC用文件系统，部分可用于服务器。

- btrfs
- ext3
- ext4
- jfs
- reiserfs
- reiser4
- xfs
- ntfs
- zfs

ntfs来凑什么热闹？

ntfs加入对比。
zfs也同样对比。
仅仅是对比而已。

参考标准

特性

- 最大容量
- 日志特性
- 目录结构
- 分配效率
- ACLS支持
- 反碎片
- checksum
- 透明压缩/透明加密
- 在线伸缩
- 设备支持
- 快照

效率

- 容积利用率
- 大文件管理特性
- 小文件管理特性
- 连续读写效率
- 随机读写效率
- 元数据管理效率
- 缓存模型管理效率

特性简表

文件系统	btrfs	ext3	ext4	jfs	reiserfs	reiser4	xfs	ntfs	zfs
最大卷容量	16 EB	32 TB	1 EB (16TB)	32 PB	16 TB	??	16 EB	256 TB	16 EB
最大文件容量	16 EB	2 TB	16 TB	4 PB	8TB	8TB	8 EB	16 TB	16 EB
目录结构	B tree	list/tree	list/Htree	B tree	B+ tree	dancing B* tree	B+ tree	B+ tree	hash table
文件分配	extents	bitmap/table	bitmap/extents	bitmap/extents	bitmap	??	extents	bitmap	??
ACLS	Yes	Yes	Yes	Yes	No	No	Yes	ACLS only	Yes
checksum	Yes	No	journal	No	No	No	No	No	Yes
透明压缩	Yes	No	No	No	No	Plugin	No	Yes	Yes
透明加密	No	No	No	No	No	Plugin	No	Yes	Yes
online defrag	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes
shrink	Yes	Yes	Yes	No	Offline resize	Offline	No	Yes	No
特性	SSD							Stream	

大部分数据来自wikipedia，在此表示感谢。

头好晕，结论是什么？

1.如果你需要用海量数据，跳过ext3/ext4/reiser。（ext4最高的推荐大小是100T，可以视为上限）
2.需要磁盘整理么？跳过ext3/reiser。
3.不想自己编译内核？跳过reiser4。
4.移动媒介到处要用？建议ntfs和ext3。当然，如果媒介size合适，最好的文件系统是vfat。
5.压缩？只有btrfs和reiser4支持。考虑3，推荐btrfs。
6.checksum？只有btrfs和ext4。如果考虑数据checksum，只有btrfs
7.需要shrink？不要考虑jfs/reiserfs/xfs。
8.btrfs支持特性好像很全面阿。问题是它没有磁盘工具，磁盘出错后全靠人品。需要调整的时候也靠人品。而且启用cow后iops非常低。

性能简表

文件系统	btrfs	ext3	ext4	jfs	reiserfs	xfs
全填充速率	1m22.083s	1m34.821s	1m15.495s	1m5.819s	1m34.310s	1m38.953s
全填充利用率	89.45%	90.65%	90.47%	99.59%	99.27%	99.18%
大文件效率	14.676	17.435	10.7255	13.7493	14.319	12.7093
大文件删除	2.693	5.262	2.422	0.037	1.802	0.296
小文件效率	9.949	5.131	2.7866	40.949	13.605	8.978
小文件删除	6.737	10.7227	1.39	16.116	2.756	5.653
循环列文件	0.124	0.089	0.002	0.094	0.19	0.099
大文件read	2046206	1931451	1946598	2003912	1537752	1970242
大文件write	1279625	565960	926461	962617	446841	812466
大文件rndread	2012771	1926287	1934420	1985273	1490199	1976056
大文件rndwrite	1380404	1187010	1294689	1446011	1308210	1384804
小文件read	2375893	2934815	3019732	2708437	2559371	2236197
小文件write	926602	526469	681710	844237	395810	939536
小文件rndread	3324647	3544566	2702282	3737551	4045575	2666753
小文件rndwrite	910277	1525970	1244240	1910756	1790393	1311261

- 时间测量使用time的real，全填充使用dd if=/dev/zero。
- 大文件使用4个iso，共计1.3G，从tmpfs上复制出，三次平均。
- 小文件使用内核源码，49232个，629M，方法同上。
- 循环列文件使用find .，测量时间，方法同上。
- 下面数据采用iozone，相对片面。

还是头晕，结论结论。

1.要保存文件，不要使用btrfs/ext3/ext4。
2.ext3的效率就是个渣。
3.jfs的小文件管理完全无法理解，我只能说，他就是发生了。
4.btrfs上面跑虚拟机慢到死，而且主观体感也很慢。原因是cow导致的iops很差。
5.ext4的小文件管理就是个奇迹。
6.reiserfs现在看的很明白——已经死了。

综上所述

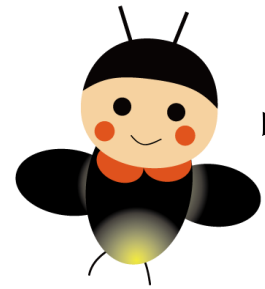
1.大规模文件保存/数据存储，例如媒体支持系统，推荐xfs。但实际上碰到这种情况的概率很小。有多少机器会有单个文件系统大于100T的可能？
2.大量小文件管理，例如源码编译，推荐ext4。
3.如果你需要特性支持，考虑btrfs，但是后果自负。
4.数据库可以适当考虑jfs，随机读写快，但是不活跃，用户少。
5.对于不同需求的目录，分区挂载，并使用不同文件系统。

未来会如何，谁也不知道。

几年前，谁也不知道Hans Reiser会卷入杀人案，锒铛入狱。
豆瓣使用gentoo+reiser，并不是目光问题，只是他们没有预知能力。而且当时ext4并不是一个选项。在ext3/jfs/xfs/reiser中，reiser拥有很好的小文件管理能力——至少比jfs好。
几年后呢？很难说。也许btrfs会修正性能问题，成为linux中的zfs。
但是也有可能，oracle宣布zfs成为专利格式，从此扫地出门。
当然——也有可能oracle直接倒闭了，和sun一样——虽然希望不大。
当然，有不少新的文件格式会出来。
所以，本文也会过期，请在2012来临前谨慎参考。

2014年11月中华架构师大会预告

演讲主题	演讲嘉宾	公司名称	职位/职称
待定	朱超	360	中间件研发负责人
TFS技术架构及运维	张友东	阿里云	TFS研发负责人
待定	黄俊	国药集团	常务副总经理
golang实时消息推送架构实战	毛剑	金山网络	移动游戏技术经理
MyCAT之前世今生	吴治辉	惠普中国	系统架构师
雪球的架构实践	王栋	雪球财经	CTO
待定	刘建平	热璞科技	技术总监



中华数据库
行业协会

中华数据库行业协会

官方网站：www.zhdba.com

官方微信平台：**zhdba2014**

官方微博：中华数据库行业协会**ZHDBA**

技术交流**QQ**群：91596001