Ext2FS vs Ext3FS vs ReiserFS

Un anàl·lisi comparatiu del sistemes de fitxers més comuns sota Linux

Albert Astals Cid

Una mica d'història

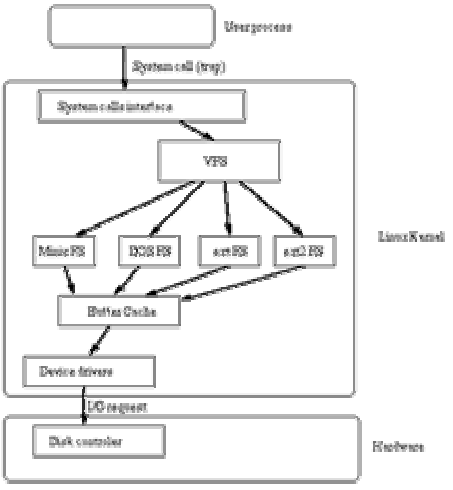
Inicialment, Linux usava el sistema de fitxers MinixFS, el sistema de fitxers que usava el sistema operatiu Minix. Però aquest sistema de fitxers tenia bastants limitacions. Per fer que fos més fàcil afegir suport per nous sistemes de fitxers, Linux va afegir al seu kernel una capa VFS (Virtual File System). Sobre aquesta capa, l'abril del 92, es va escriure l'Extended FileSystem, que corregia alguna de les limitacions de MinixFS, però encara en mantenia unes quantes. Per intentar eliminar aquestes limitacions, el gener del 93, es presentaven les versions Alpha de dos sistemes de fitxers, Ext2FS i XiaFS, el primer extensió del Extended File System i el segon extensió de MinixFs.

	Minix FS	Ext FS	Ext2 FS	Xia FS
Max FS size	64 MB	2 GB	4 TB	2 GB
Max file size	64 MB	2 GB	2 GB	64 MB
Max file name	16/30 c	255 c	255 c	248 c
3 times support	No	No	Yes	Yes
Extensible	No	No	Yes	No
Var. block size	No	No	Yes	No
Maintained	Yes	No	Yes	?

VFS

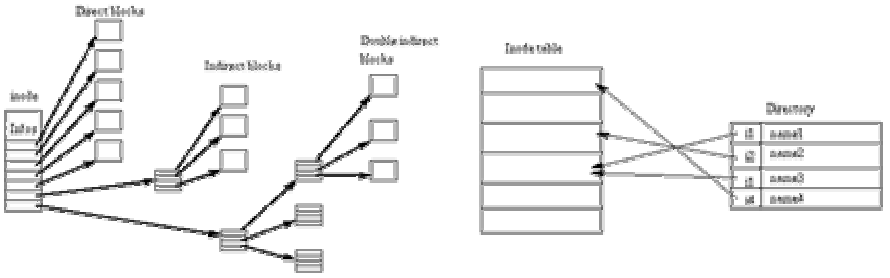
El VFS (Sistema de Fitxers Virtual) és una capa del kernel que processa les crides que han de treballar amb fitxers i crida les rutines necessàries del sistema de fitxers corresponent. EL VFS defineix un conjunt d'operacions que tot sistema de fitxers ha d'implementar. Aquestes operacions són les de muntar un sistema de fitxers, aquelles que es poden fer sobre tot inode(crear, esborrar, etc.) i aquelles que es poden fer sobre els fitxers oberts (llegir, escriure, etc.).

VFS



Ext2FS

Ext2FS segueix la estructura clàssica d'un sistema de fitxers Unix.



Ext2FS (II)

A més a més de les possibilitats bàsiques d'un sistema de fitxers Unix, Ext2FS afegeix d'altres capacitats al sistema de fitxers com:

- Els atributs dels fitxers poden modificar el comportament del kernel sobre els mateixos fitxers.
- En muntar el sistema de fitxers es pot escollir si volem que segueixi semàntica System V (més complexa) o BSD (més simple) en quant als grups dels fitxers en la creació.
- Es pot escollir que l'escriptura de les dades es faci en el moment que l'usuari les fa (gran pèrdua de rendiment)
- Implementa enllaços simbòlics ràpids, que no usen cap bloc de dades per fer la redirecció sinó el mateix inode.
- Hi ha fitxers immutables (només es poden llegir) i fitxers de només afegir (útils per a logs)

Ext2FS (III)

La estructura física del sistema de fitxers és:

Cada bloc conté una còpia redundant de les informacions de control crucials per al sistema i també conté una part del sistema de fitxers. La estructura de cada bloc és:

El fet d'usar aquesta estructura és important en quant a estabilitat, ja que en cas de corrupció del superbloc, es fàcil recuperar-lo ja que cada bloc en conté una còpia. Aquesta estructuració també ajuda al rendiment: el fet de que les taules de inodes i els blocs de dades estiguin propers en el disc pot fer que es redueixi el temps perdut esperant a que el capçal busqui les dades al disc

Journal File Systems

Perquè?
Els sistemes de fitxers amb journal han aparegut per a solucionar el problema introduït pel fet de que les escriptures a disc mai no son directes si no que es fan a través d'un buffer. Això pot provocar inconsistència en els sistemes de fitxers quan l'ordinador es penja. Aquest fet, juntament amb el temps necessari per recuperar el sistema de fitxers a un estat consistent i que provoca que els servidors no pugui estar funcionant, ha fet aparèixer els sistemes de fitxers amb journal que intenten corregir aquests problemes.

Com?
El que es fa és utilitzar el concepte d'atomicitat de transaccions usat en les bases de dades. Una operació, o es duu a terme o es cancel·la, però no es fa només la meitat. Això juntament amb un log que guarda totes les operació i arguments permet en cas de trobar una inconsistència en el sistema de fitxers, tornar ràpidament a un estat consistent.
Hi ha dos tipus de journaling bàsic, el que guarda només les dades de control i el que a part també guarda les dades en sí.

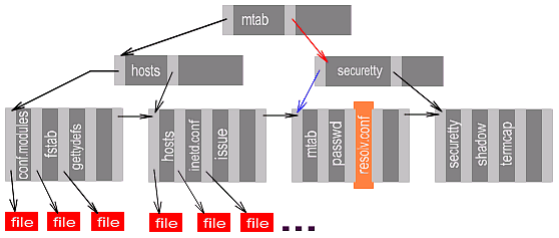
Ext3FS

El sistema de fitxers Ext3FS no és més que el sistema de fitxers Ext2FS amb journaling. Això s'ha pogut fer degut a que Ext2FS va ser programat pensant en possibles ampliacions.

Hi ha tres modes de funcionament:
- **writeback**: fa journalling de metadades. És el mètode més ràpid. Hi pot haver corrupció de dades en cas de que es pengi el sistema.
- **ordered**: El mètode per defecte, només fa journaling de metadades però usa un mètode d'escriure a disc que protegeix més les dades, de fet, només hi pot haver corrupció de dades al sobreescriure fitxers, però no al afegir-hi dades.
- **journal**: Journalling complet de dades i metadades. Hauria de ser el més lent ja que les dades s'escriuen dues vegades, una al journal i un altra al fitxer real. En canvi alguns tests mostren que el seu rendiment és millor que el dels dos altres en entorns que tinguin molta feina contínuament. També necessita molt espai a disc per guardar el journal.

ReiserFS

El sistema de fitxer ReiserFS utilitza arbres B+ per guardar les dades.



Els arbres B+ han estat usats des de fa molt temps en les bases de dades. ReiserFS es va desenvolupar amb el propòsit de demostrar que era posible fer un sistema de fitxers eficient usant arbres B+ i com a base per a un sistema encara més potent (Reiser4 del qual s'espera una beta per aquest Juny), un sistema amb característiques similars a les de les bases de dades i els sistemes d'hipertext.

ReiserFS només fa journaling només de les metadades.

ReiserFS (II)

Què és un arbre B+?
Un arbre B+ es un arbre arbre binari de cerca. Els arbres B+ son completament equilibrats. Tenen dos tipus de nodes, els nodes interns i les fulles. Tots els apuntadors a les dades es troben a els nodes fulla. Els valors estan ordenats.

Els nodes interns només tenen com a objectiu dirigir la cerca i per tant no contenen l'apuntador a les dades, només apuntadors als arbres que hi ha per sota.

Els nodes fulla contenen tant el apuntador al següent valor com l'apuntador a les dades.

Els arbres B+ són molt eficients per a fer cerques, com a exemple, una cerca en 23.667.000 elements costa accedir a 4 nodes, mentre que en un arbre binari normal costa accedir-ne a 26.

Comparant amb el mètode de taules i llistes (usat a ext2fs i ext3fs) té el problema de que les insercions i, sobretot, les supressions més costoses degut a que s'ha de mantenir l'ordre de l'arbre.

Rendiment Bonnie++

PC de proves:
Bi-PentiumIII 1 Ghz
256 MB de RAM
Disc dur SCSI FUJITSU UW2 18 GB
Red Hat 7.2
http://www.linux-france.org/article/sys/ext3fs/Benchmarks/bonnie++.txt

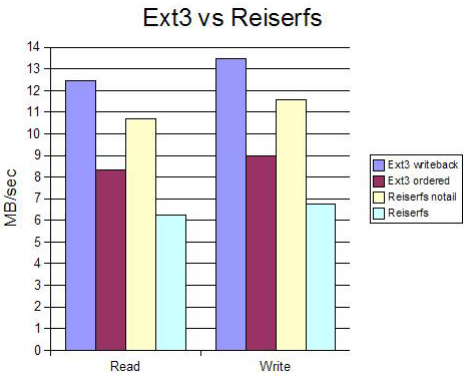
	---Sequential Output (nosync)---						---Sequential Input--				--Rnd Seek-	
	-Per Char-		--Block---		-Rewrite--		-Per Char-		--Block---		--04k (03)-	
Filesystem	K/sec	%CPU	K/sec	%CPU	K/sec	%CPU	K/sec	%CPU	K/sec	%CPU	/sec	%CPU
Ext2	14099	99	52190	26	10368	7	12766	96	45891	15	258.8	1
Ext3	12925	97	46427	40	20047	16	12890	98	44359	17	240.8	1
Reiserfs	12829	98	48603	50	10274	8	12609	96	46880	21	257.8	0

	------Sequential Create------						--------Random Create--------					
	-Create--		--Read---		-Delete--		-Create--		--Read---		-Delete--	
Filesystem	/sec	%CP	/sec	%CP	/sec	%CP	/sec	%CP	/sec	%CP	/sec	%CP
Ext2	4621	54	980	4	52984	80	5022	59	130	1	11269	90
Ext3	2526	51	934	5	21157	80	3407	68	125	2	366	4
Reiserfs	954	42	165	1	1240	13	988	46	118	1	188	3

Rendiment PostMark

Pc de proves:
- Intel Pentium III 1133MHz
- 768MB 133MHz ECC SDRAM
- 3 x 18GB 15K RPM Ultra160 SCSI RAID5

http://www.gurulabs.com/ext3-reiserfs-3.html



Conclusió

Veient que el journaling ens pot ajudar a salvar les nostres dades i que a més en el cas d'un apagat incorrecte farà que la comprovació del disc sigui molt més ràpida, eliminaria a Ext2FS com a sistema de fitxers a recomanar

Entre Ext3FS i ReiserFS és més complicat, en quant a velocitat, com hem vist en aquestes dues proves de rendiment, el rendiment pot variar molt segons l'aplicació que fem servir per fer la prova. Per exemple en Bonnie++ veiem com ReiserFS dona una mica més de rendiment en els tests de lectura i escriptura. En canvi si anem a PostMark mostra un escenari diferent on guanya Ext3FS

En quant a estabilitat Ext3FS té la avantatge de ser fill directe de Ext2FS, un sistema molt provat, i per tant amb molta estabilitat; ReiserFS no és inestable, però al ser més nou és possible que tingui algun error de més. Per tant, tot depèn de l'ús que se li vulgui donar al PC. Com a exemple, la majoria de distribucions de Linux usen Ext3FS com a sistema de fitxers recomanats, encara que alguna, com Gentoo, comença a recomanar ReiserFS