

Dyski SSD a systemy plików

Paweł Wiejacha

Seminarium z Systemów Rozproszonych

27 maja 2010

Plan prezentacji:

- Krótko o SSD
 - czym są SSD, czym się charakteryzują
- Systemy plików a pamięci flash
 - nowe wymagania
 - FLT a MTD i UBI
- Log-structured file systems
 - LogFS
 - NILFS
- Log-structured file systems a zwykłe dyski
- Podsumowanie

Tradycyjne dyski twarde



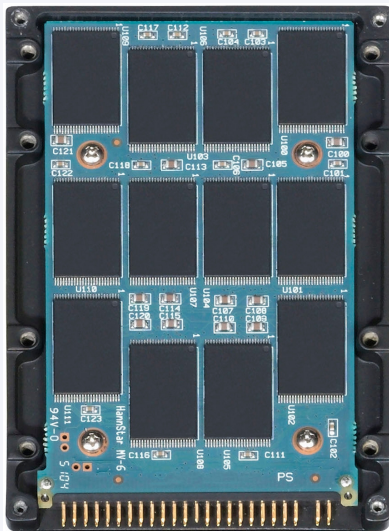
Tradycyjne dyski twarde

- wirujące talerze pokryte nośnikiem magnetycznym
- ruchome głowice elektromagnetyczne
- sensowny sekwencyjny odczyt i zapis
- powolny dostęp losowy:
 - random seek (średnio 7-9 ms)
- wibracje i głośność
- słabo tolerują wstrząsy i przeciążenia

Solid-state drive

- urządzenie pamięci masowej
- zbudowane w oparciu o pamięć typu flash
- brak elementów elektromechanicznych,
 - tylko układy scalone
- dane zapisane są w bramkach przechowujących elektrony

Solid-state drive



Przykładowe SSD obecne na rynku

- **HP StorageWorks IO Accelerator** - 320 GB; transfer 800MB/s; 100 000 IOPS; \$13 000
- **OCZ Z-Drive R2** - 2 TB; zapis: 950MB/s; odczyt: 1.4GB/s; 29 000 IOPS; \$8 778
- **Corsair SSD** - 500GB; odczyt: 240MB/s; zapis 200MB/s; 5900 zł
- **Kingston SSDNow V2 Series** - 64GB; odczyt: 100MB/s; zapis 80MB/s; 609 zł

SSD - zalety

- mniejszy czas dostępu do danych (random access)
- większe prędkości zapisu i odczytu
- mniejszy pobór energii
- odporność na przeciążenia, wstrząsy i wibracje
- szeroki zakres temperatur, w których mogą pracować
- przeważnie są lżejsze

SSD - wady

- ograniczona żywotność
 - MLC do 10 000 zapisów jednej komórki
 - SLC do 100 000 zapisów jednej komórki
- ciągle ceny o rząd wielkości wyższe
- wolniejsze nadpisywanie małych obszarów
- realna awaryjność

SSD - czyszczenie i zapis segmentów

- "pusty" (wymazany) segment flash jest wypełniony jedynkami
- można zmienić pojedyncze bity z 1 na 0
 - czasem granularność jest większa (np. 1KB)
- ale nie można zmienić z 0 na 1
 - by to zrobić trzeba wymazać **cały** segment
- segmenty są dość duże (zazwyczaj 256KB)
- zapis jest szybki jeśli jest dużo czystych segmentów

komenda TRIM

- zazwyczaj dysk twardy nie wie, które sektory są nieużywane
- to system plików obsługuje usuwanie plików
- dla dysku twardego każdy sektor jest używany
- nie można było powiedzieć poprzez interfejs (S)ATA dysкови,
 - że dany sektor jest nie używany
- dlatego rozszerzono interfejs ATA o tę możliwość (TRIM)
 - wsparcie w Linux 2.6.33+, Windows 7

Nowe wymagania dla systemów plików

- ograniczona żywotność komórek
 - równoważenie zużycia (ang. *wear levelling*)
- czyszczenie nieużywanych bloków w tle (np. TRIM)
- duże segmenty - "fragmentacja" wewnętrzna
- problemy z nadpisywaniem segmentów
 - np. utrata zasilania między czyszczeniem a pisaniem
 - również obniżona wydajność tej operacji

Pamięć flash - urządzenie blokowe czy znakowe?

- żadne z nich - Memory Technology Device (MTD)

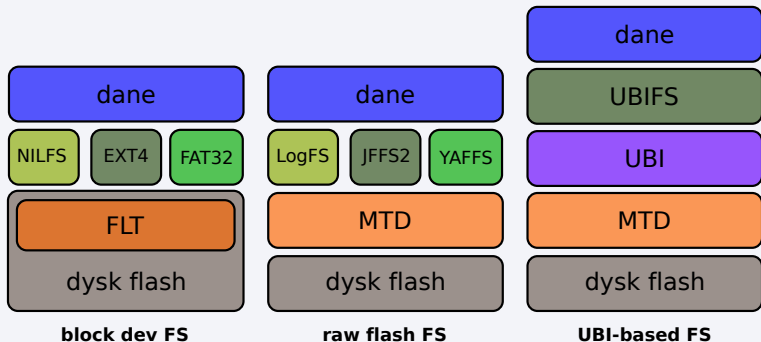
Urządzenie blokowe	MTD
sektory	<i>eraseblocks</i>
małe sektory (512B, 1024B)	większe segmenty 256KB
2 podstawowe operacje: zapisz i odczytaj sektor	3 podstawowe operacje: zapisz, wyczyść, odczytaj segment
ukrywanie zepsutych sektorów	zepsute segmenty trzeba obsłużyć
sektory zużywają się wolno lub równomiernie	segmenty zużywają się po ~10000 zapisach

FTL - Flash Translation Layer

- zazwyczaj firmware
- udający, że dysk flash jest urządzeniem blokowym
- zapewnia:
 - równoważenie zużycia
 - obsługę uszkodzonych sektorów
 - proste odświeżanie i czyszczenie sektorów
- mając pamięć flash z FLT
 - mamy urządzenie blokowe
 - więc można używać dowolnego systemu plików

Systemy plików a pamięci flash

Systemy plików dla pamięci flash w Linuksie



Log-structured file system

- traktuje nośnik jako cykliczny log
- na którego **końcu** zapisuje zmiany w danych i metadanych
- dane zapisywane są **sekwencyjnie**
 - segmenty nie są nadpisywane
- stare wpisy w logu trzeba odświeżać
- umożliwia tworzenie snapshotów systemu plików i wersjonowanie

Log-structured file systems

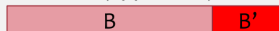
Zapis

Application view:

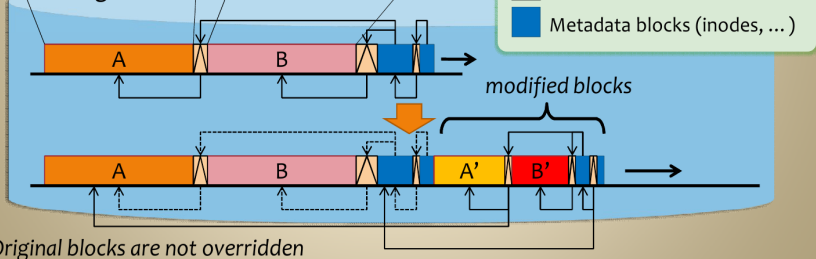
File A (modified)



File B (appended)

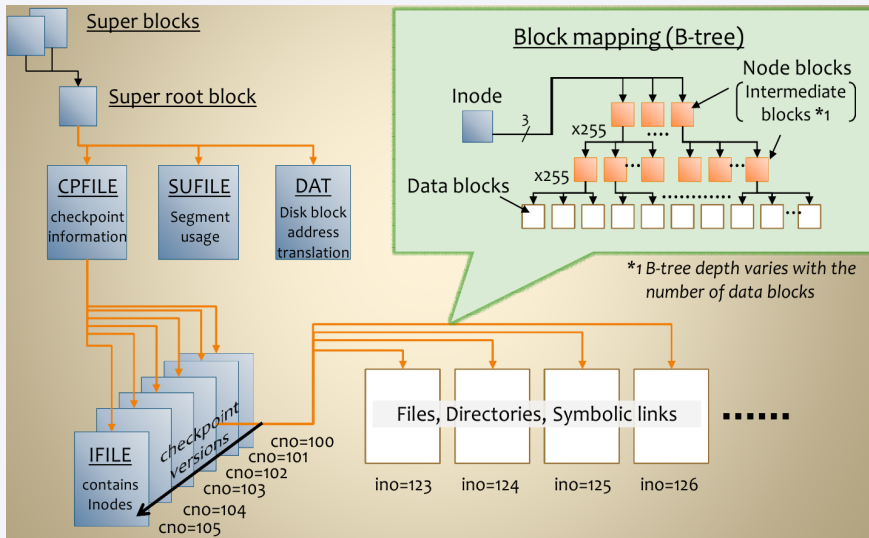


On disk images:



Log-structured file systems

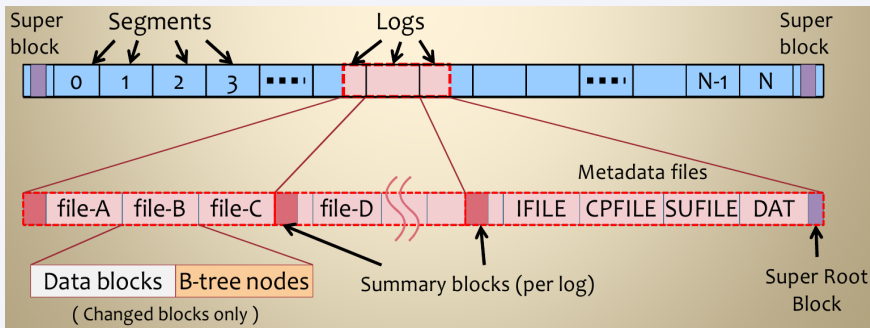
Metadane i struktury



Log-structured file systems

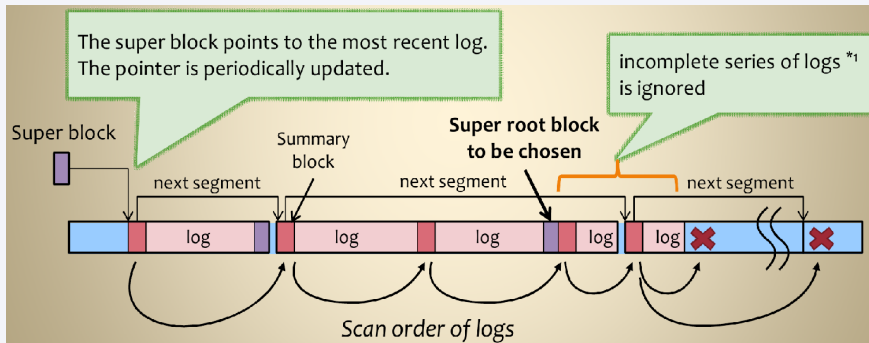
Segmenty i logi

- przestrzeń dyskowa podzielona na segmenty
- segmenty zawierają logi (dety) i checkpointy



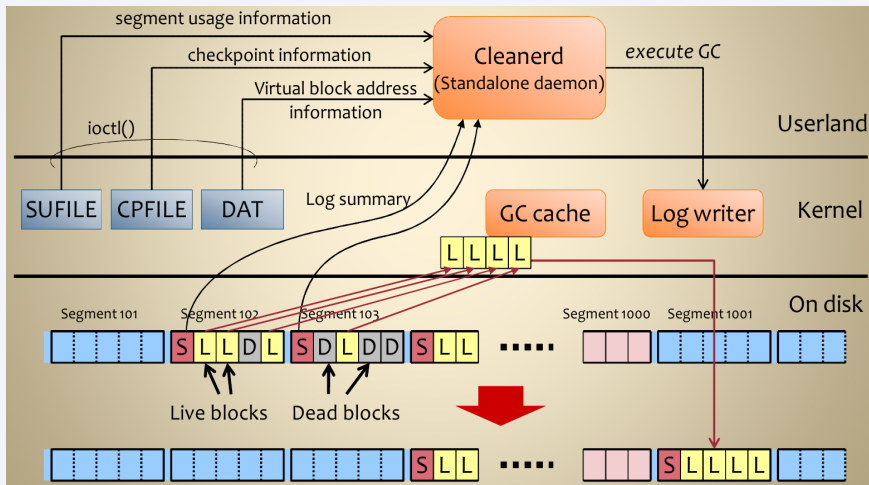
Log-structured file systems

Montowanie i przywracanie po awarii



Log-structured file systems

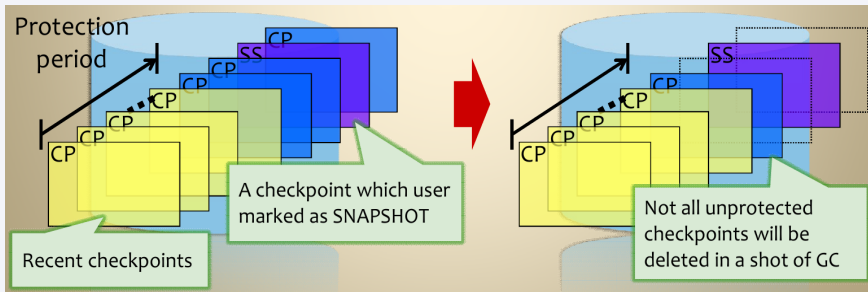
Odśmianie



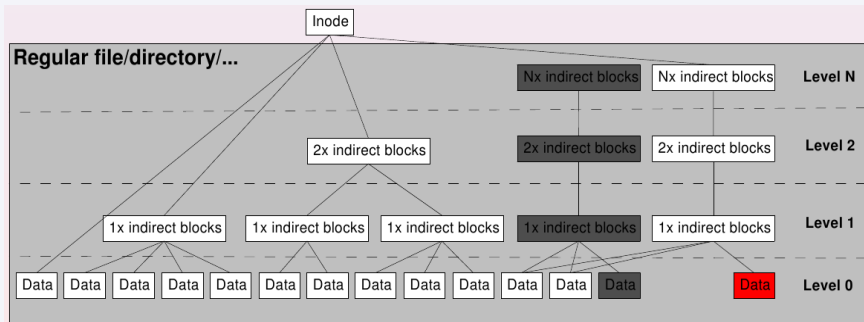
Log-structured file systems

Migawki

- użytkownik może oznaczyć punkt kontrolny jako migawkę
- GC usuwa niepotrzebne checkpointy, ale zostawia migawki



Wandering tree użyte do delt B-drzewa



Cechy systemów plików o budowie loga

- szybki zapis danych
- migawki
- spójny stan systemu plików po awarii
- cofnięcie przypadkowego **rm -rf ***
- przystosowane do pamięci flash
- wspiera inkrementalne kopie zapasowe

LogFS a NILFS

- oba mają budowę loga
- LogFS - raw flash FS, NILFS - blockdev FS
- LogFS jest prostszy i przeznaczony do mniejszych dysków
- LogFS nie wspiera migawek
- LogFS - montowanie w $O(\log(n))$

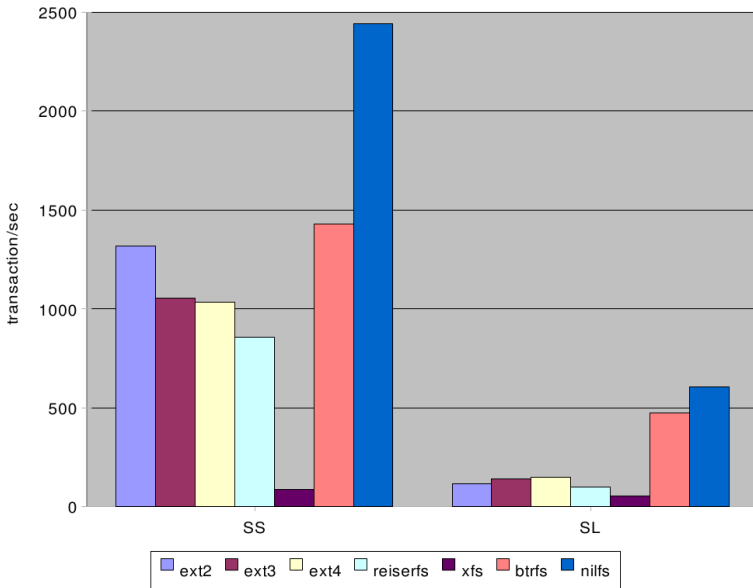
Testy NILFS na SSD - zestawy danych

Workload	File size	# of file (work-set)	# of transaction	Total app read/write
SS	9-15K	10,000	100,000	630M/755M*
SL	9-15K	100,000	100,000	600M/1.8G
LS	0.1-3M	1,000	10,000	9.7G/12G
LL	0.1-3M	4,250	10,000	9G/17G

* Mostly write-only

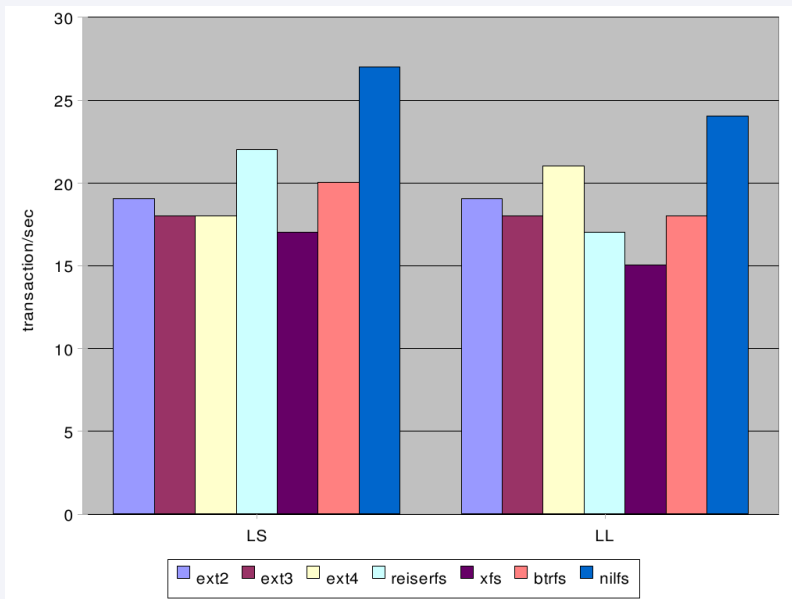
Log-structured file systems

NILFS - wydajność na SSD (2)



Log-structured file systems

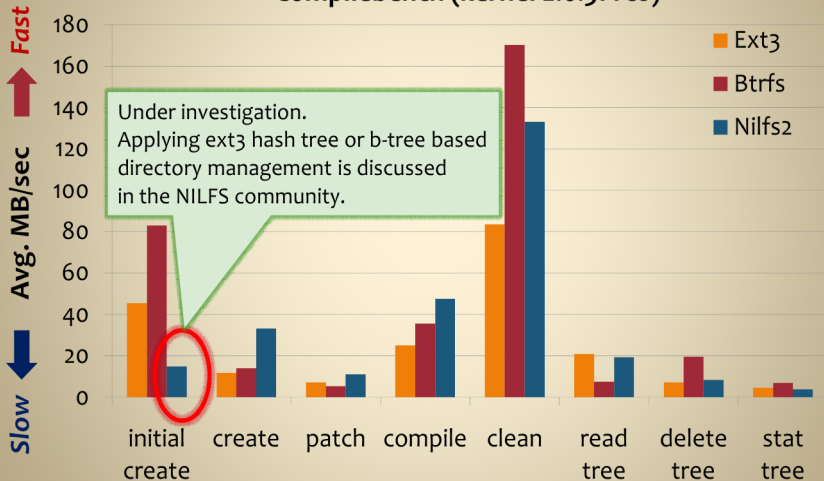
NILFS - wydajność na SSD (3)



Log-structured file systems

NILFS - wydajność na HDD

Compilebench (kernel 2.6.31-rc8)



Hardware specs:

Processor: Pentium Dual-Core CPU E5200 @ 2.49GHz, Chipset: Intel 4 Series Chipset + ICH10R, Memory: 2989MB, Disk: ST3500620AS

- **LogFS** - <http://logfs.org/logfs/>
- **NILFS** - <http://www.nilfs.org/papers/jls2009-nilfs.pdf>
- **MTD, UBI, UBIFS** - <http://www.linux-mtd.infradead.org/>
- **SSD benchmarks** -
http://www.benchmark.pl/testy_i_recenzje/a-2695/strona/8397.html
- **YAFFS** - <http://www.yaffs.net/>
- **JFFS2** - <http://linux-mtd.infradead.org/dwmw2/jffs2.pdf>

Pytania?