



ZFS

Zettabyte File System

Tim Langens & Jef Neefs

Overzicht

- Inleiding
- Capaciteit
- Opslag pools
- Copy On Write
- Snapshots en Klonen
- Variabele blok grootte
- Beperkingen
- Platformen
- Gedistribueerde systemen
- Concurrentie



Inleiding

- “Zettabyte File System”
- Team van Sun
- Jeff Bonwick
- November 2005: deel van OpenSolaris



128 bit



- Nu al data sets in de orde van petabytes (2^{50})
- Levensduur huidige 64-bit systemen:
$$2^{64} - 2^{50} = 2^{14}$$
- *“If 64 bits isn't enough, the next logical step is 128 bits. That's enough to survive Moore's Law until I'm dead, and after that, it's not my problem.” : Jeff Bonwick*

Theoretische Capaciteit



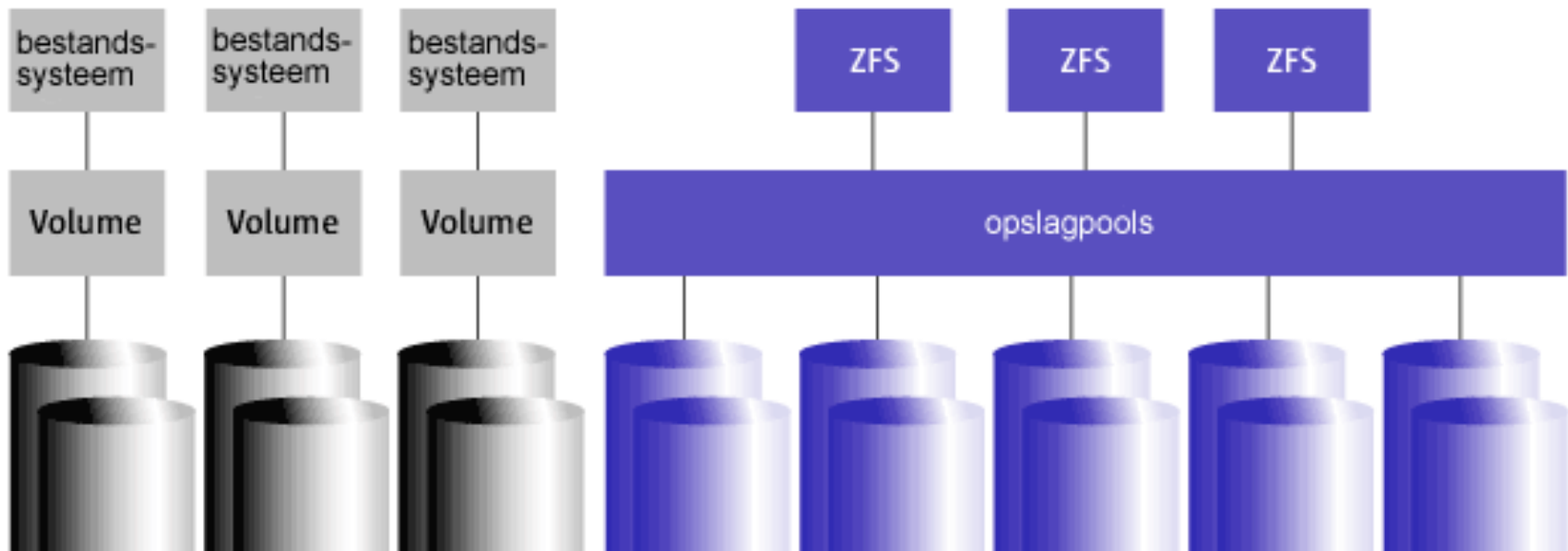
- 128-bit pool heeft 2^{128} blocks nodig
- 2^{128} blocks = 2^{137} bytes = 2^{140} bits
- Maximum grootte van één enkel bestand = 16 exabytes
- Aantal bestanden in bestandssysteem = 2^{48}
- Aantal bestanden in een map = 2^{56}
- Als er elke seconde 1000 bestanden worden gecreëerd duurt het nog 9000 jaar voor deze limieten worden bereikt.

Praktische capaciteit



- Kwantum mechanische limiet gaat wet van Moore moeten doorbreken.
- “1kg materie in 1L ruimte kan 10^{51} operaties per seconde doen met maximum 10^{31} bits aan informatie.” (Seth Lloyd)
- Volledig beschreven 128-bit opslagpool 136 miljard kg aan materie nodig
- $E(\text{data creëren}) > E(\text{oceanen koken})$

Opslagpools



Copy on write



- Als 2 processen een bestand gebruiken krijgen ze pointers naar dit bestand
- Als dan 1 van de 2 processen het bestand wil aanpassen wordt dit naar een kopie van het origineel geschreven zodat het 2^e proces niet wordt beïnvloed
- In ZFS zijn ALLE bewerkingen copy-on-write. Resultierend in een praktisch foutloos bestandssysteem.
- Als een kopie beschadigd is zal ZFS een andere kopie gebruiken om de kopie te herstellen

Snapshots en klonen



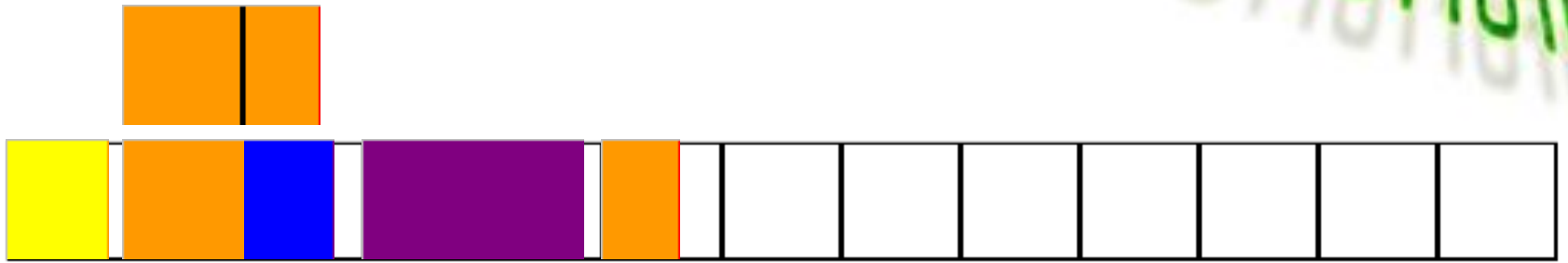
- Een snapshot is een moment-opname van een bestandssysteem dat enkel te lezen is.
- Een kloon is een snapshot die aangepast kan worden
- Een snapshot kan een volledige back-up genereren
- Een incrementele back-up kan gegenereerd worden door verschillende snapshots te gebruiken

Variabele blokgröote

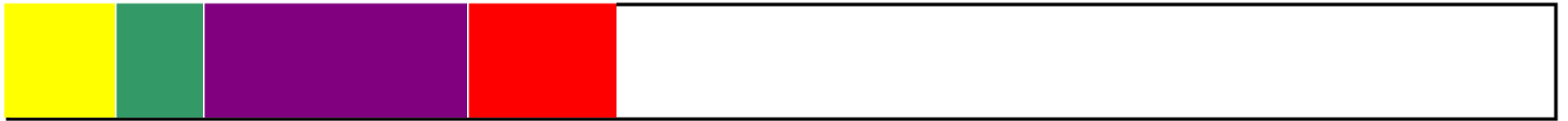


- Huidige bestandssystemen verliezen veel schijfruimte aan overhead als gevolg van een vaste blokgröote
- ZFS gebruikt een variabele blokgröote tot 128KB

Vaste blokgröte



Variabele blokgröte



Beperkingen



- Huidige bootsystemen ondersteunen ZFS niet
- Geen transparante encryptie zoals NTFS
- Checksums bij copy-on-write trekt de processorbelasting omhoog.
- Men mag enkel EFI gelabelde harde schijven gebruiken

Platformen



- Heden:
 - Solaris en OpenSolaris
- Toekomst:
 - Apple toont interesse
 - De BSD wereld ziet ZFS ook wel zitten
 - Linux is moeilijker omwille van licenties

Gedistribueerde Systemen



- Zpool
- Adaptive endiannes
 - De block pointer kan in big middle of little endian opgeslagen worden
 - Dit zorgt ervoor dat ZFS leesbaar en schrijfbaar is voor eender welke endiannes

Concurrentie



- Veritas File System of VxFS
- Speciaal voor gedistribueerde systemen
- Defragmentatie is miniem tot onbestaande
- Heeft niet de praktisch onbereikbare datalimiet van ZFS

Vragen?

