ZFS z pohľadu systémového administrátora

Martin Matuška
mm@FreeBSD.org

VX Solutions s. r. o.

Projekt HOW-KNOW
YNET, IAESTE
8.12.2010

O tejto prednáške

Cieľom tejto prednášky je uviesť Vás do sveta ZFS a priniesť odpovede na nasledovné otázky:

- Čo je ZFS?
- Čím sa ZFS odlišuje od iných súborových systémov?
- Ktoré operačné systémy (distribúcie) podporujú ZFS?
- Aké je právne hľadisko používania, vývoja a distribúcie ZFS?
- Ako bude pokračovať vývoj ZFS v budúcnosti?

Úvod

- Čo je ZFS?
- Základné koncepty
- Rozšírené koncepty
- História
- Základná štruktúra
- Limity
- Verzie

Čo je ZFS?

ZFS je súborový systém prezývaný "Zettabyte filesystem"

Wikipedia: "Súborový systém je spôsob ukladania a organizovania počítačových súborov a údajov, ktoré obsahujú, tak, aby k nim bol umožnený jednoduchý prístup"

Základné koncepty ZFS

ZFS bolo navrhnuté s nasledovnou funkcionalitou:

- úložiská (pool - zabudovaný volume manager)
- transakčná sémantika (copy on write)
- kontrolné súčty a autokorektúra (scrub, resilver)
- rozšíriteľnosť (škálovateľnosť)
- instantné snímky a klony
- kompresia dátových setov (lzjb, gzip)
- zjednodušená delegovateľná administrácia

Rozšírené koncepty ZFS

Vybraná funkcionalita pridaná počas ďalšieho vývoja:

- RAID s dvojitou (v3) a trojitou (v17) paritou
- sekundárny cache (L2 cache, v10)
- používateľské a skupinové kvóty (v15)
- deduplikácia dát (v21)
- kryptovanie datasetov (v30)

História ZFS

- 2005/10: OpenSolaris - ZFS uvedené v revízii 789
- 2005/12: Solaris Express - prvé verejné vydanie
- 2006/06: Solaris 10 update 6 - pool verzia 3
- 2007/04: FreeBSD - pool verzia 6
- 2008/11: FreeBSD - pool verzia 13
- 2009/10: Solaris 10 update 8 - pool verzia 15
- 2010/07: FreeBSD - pool verzia 15
- 2010/08: OpenSolaris - projekt ukončený revíziou 13149 (v28)
- 2010/09: Solaris 10 update 9 - pool verzia 22 (bez deduplikácie)
- 2010/11: Solaris 11 Express - pool verzia 31 (ZFS crypto)

Základná štruktúra ZFS

ZFS je reprezentované nasledovnými dvoma základnými objektami:

- pool (úložisko)
- dataset (dátová sada)

ZFS pool

ZFS pool je objekt pozostávajúci z virtuálnych zariadení (vdev).
Tieto zariadenia môžu byť:

- disk (napr. pevný disk alebo jeho partícia)
- súbor (určené na testovanie)
- mirror (spojenie dvoch alebo viacerých vdev zariadení)
- raidz, raidz2, raidz3 (RAID s jednoduchou, dvojitou a trojitou paritou)
- spare (pseudo-vdev určené ako záloha pre RAIDZ)
- log (samostatné vdev pre ZIL, nemôže byť raidz)
- cache (L2 cache, nemôže byť mirror alebo raidz)

ZFS dataset

Každý ZFS pool (úložisko) obsahuje ZFS datasety (dátové sady).
ZFS dataset je všeobecným pomenovaním pre:

- filesystem (súborový systém, vrstva POSIX)
- volume (virtuálne diskové zariadenie)
- snapshot (snímok súborového systému resp. volume, iba na čítanie)
- clone (klon - súborový systém s počiatočným stavom snapshotu)

Limity ZFS

Aké hranice má ZFS?

- ZFS je 128-bitový súborový systém
- Maximálna veľkosť poolu je: 256 quadriliónov zettabajtov (= 256 * 10^{36} bajtov)
- Maximálna veľkosť datasetu, súbora alebo jeho atribútu: 16 exabajtov
- Maximálny počet poolov/filesystémov/snapshotov: 2^{64}

Verzie ZFS

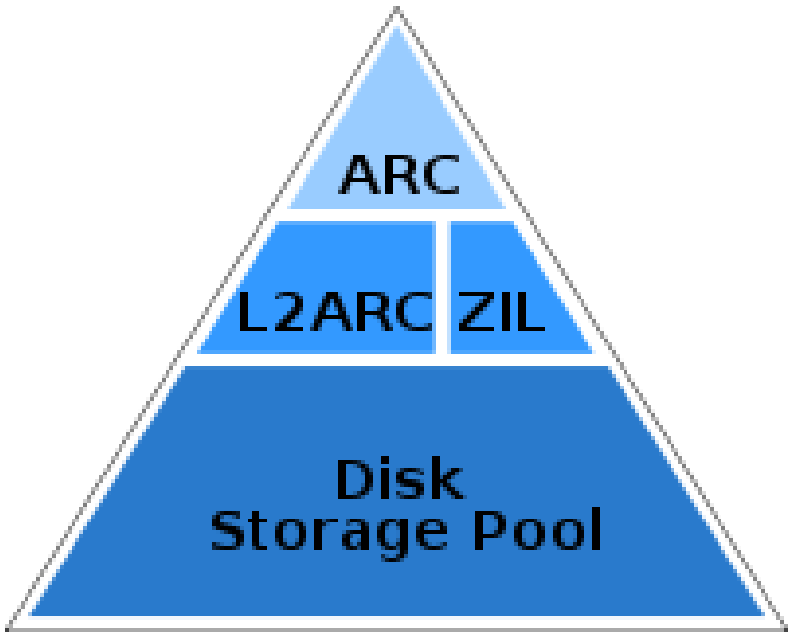
- Pool a filesystem majú v ZFS číslo verzie
- nekompatibilné zmeny v štruktúre vedú k zvýšeniu čísla verzie
- ZFS je spätne kompatibilné
- nekompatibilita s novšími verziami
- ZFS neumožňuje downgrade
- najnovšia verzia ZFS pool: 31
- najnovšia verzia ZFS filesystem: 5

Vybrané vlastnosti ZFS

- Vyrovnávacia pamäť (ARC)
- Integrovaný volume manager
- ZFS snapshot a clone
- ZFS send a receive
- Zálohovanie pomocou send a receive
- Delegovaná administrácia

Vyrovnávacia pamäť (ARC)

ZFS pracuje s modifikovanou verziou vyrovnávacej pamäte Adaptive Replacement Cache (ARC), ktorá zrýchľuje prístup k často čítaným dátam.

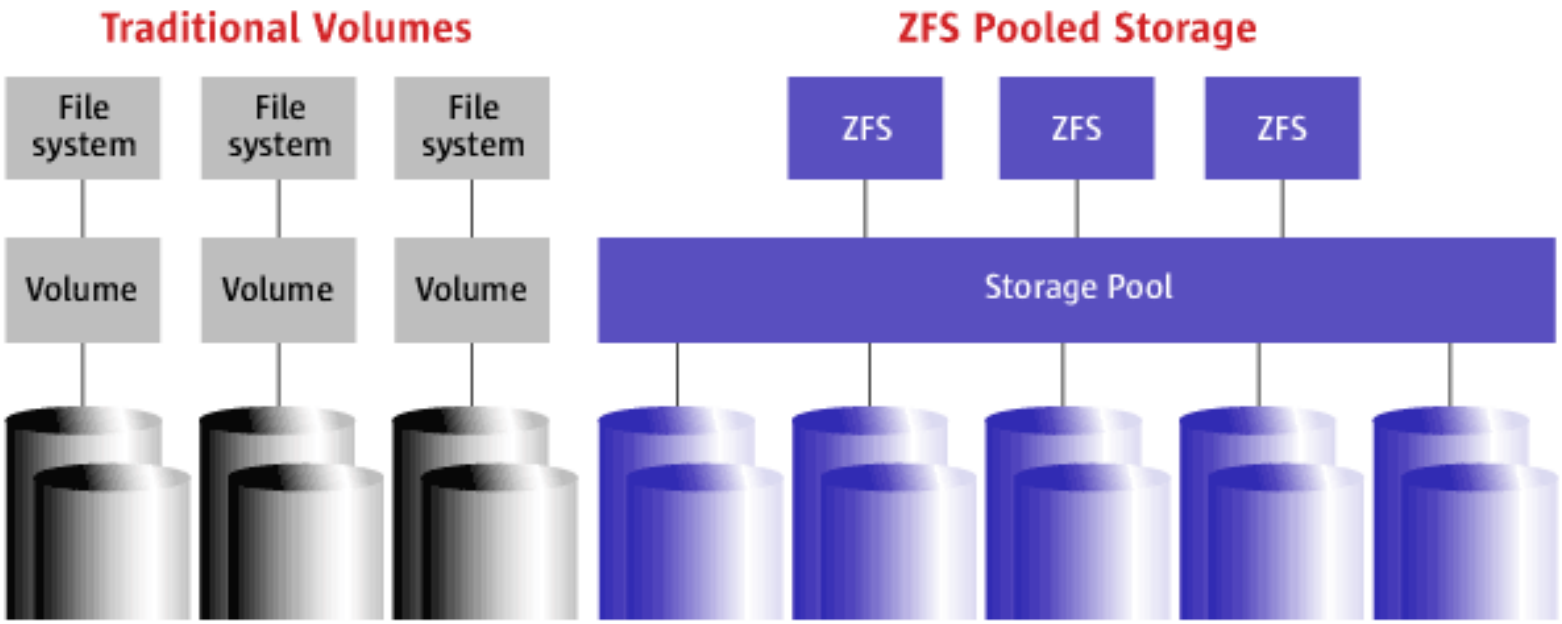


Integrovaný volume manager

Integrovaný volume manager umožňuje lepšiu organizáciu dát:

- možnosť rozdelenia diskov do mirror-ov a kombinácií RAID
- synchronizujú sa iba dáta, voľné miesto nie
- detekcia chýb a autokorekcia pomocou kontrolných súčtov
- možnosť plynulého prechodu na väčšie disky

Integrovaný volume manager



ZFS snapshot a clone

Snapshoty (snímky) umožňujú ochrániť dáta pred zmazaním a rôzne s nimi nakladať bez obmedzenia prevádzky.

- snapshot je snímka systému k určitému času
- dáta a súbory v snapshote sú ľahko dostupné
- súborový systém možno navrátiť do stavu jeho snapshotu
- zo snapshotu sa dá vytvoriť zapisovateľný súborový systém - klon
- klon sa dá okrem iného výborne použiť na testovacie účely so živými dátami
- klon možno "povýšiť" aby nahradil pôvodný súborový systém

ZFS send a receive

Pomocou send a receive sa posielajú snapshoty medzi systémami (počítačmi, servermi).

- príkaz "zfs send" vytvára zo snapshotu dátový tok
- tento sa dá uložiť do súboru alebo poslať na iný systém
- príkaz "zfs receive" premení tieto dáta opäť na snapshot
- možnosť posielať aj iba zmenené dáta medzi snapshotami

Zálohovanie pomocu send a receive

ZFS send a receive tvoria veľmi efektívny spôsob zálohovania dát až po celé systémy:

- spolu s dátami sa zálohuje aj celá história ("zálohy záloh")
- ideál: prenášať iba zmenené dáta
- záloha je na druhej strane rovnako ľahko čitateľná
- pohodlne kombinovateľné s klasickou formou zálohovania
- nevýhody: zálohovanie je asynchrónne, model všetko-alebo-nič, obmedzené na datasety
- výhody: operácie na pozadí, (takmer) úplne bez downtime, rýchle pri veľmi veľkom počte súborov

Delegovaná administrácia

ZFS umožnuje oprávniť používateľov vykonávať rôzne operácie.

- príkazy "zfs allow" a "zfs unallow"
- systém prístupových práv na spravovanie ZFS
- spolupráca s virtualizáciou (Solaris: zones, FreeBSD: jails)

Operačné systémy

- Distribúcie založené na Solaris-e/OpenSolaris-e
- Ostatné operačné systémy a distribúcie

OS založené na Solaris-e/OpenSolaris-e

- OpenSolaris (projekt ukončený)
- Oracle Solaris 10 / Solaris Express 11
- Nexenta Core
- OpenIndiana
- SchilliX
- Belenix

OpenSolaris

- Zdroj kódu ZFS pre všetkých ostatných
- ZFS zavedené 31.10.2005 v revízii 789
- Posledné vydanie: OpenSolaris 0906 (jún 2009)
- Posledné vývojárske vydanie: build 134 (marec 2010)
- Posledná zmena v zdrojových kódoch: 18.08.2010 (revízia 13147)
- Budúcnosť: projekt ukončený spoločnosťou Oracle
- Voľné pokračovanie: projekt Illumos

Oracle Solaris

- Komerčný operačný systém (potrebné zakúpenie licencie)
- ZFS prvý krát dostupné v Solaris 10 update 6 (jún 2006)
- Posledné vydanie: update 9 (2010/09) s ZFS v22 (bez dedup.)
- Dokumentácia: Oracle Solaris ZFS Administration Guide
- Oracle pokračuje v neverejnom vývoji ZFS

Nexenta Core

- OpenSolaris s debiánovskou správou balíkov
- Posledné vydanie: 3.0.1 (2010/09) so ZFS v26
- Kompatibilné s OpenSolaris-om
- Relatívne stabilné, ale veľmi slabá dokumentácia
- Budúcnosť: spolupráca s Illumos-om

OpenIndiana, Belenix, SchilliX

- distribúcie založené na OpenSolaris-e
- OpenIndiana: "voľné pokračovanie" OpenSolaris-u (Illumos)
  Posledné vydanie: dev build 147 (2010/09)
- BeleniX: Indická LiveCD-distribúcia
  Posledné vydanie: 0.8 beta 1
- SchilliX: nemecká distribúcia (teraz už zakladá na Illumos-e)
  Spravujú: Jörg Schilling a Fabian Otto (Fraunhofer-Institut für Offene Kommunikationssysteme)
  Posledné vydanie: 0.7.2 (2010/09)

Ostatné systémy

ZFS pochádza z OpenSolaris-u, ostatní ho musia importovať

- FreeBSD
- MacOS X
- Linux (cez FUSE alebo kernelový modul)
- Debian (GNU/kFreeBSD) - "iba" distribúcia

FreeBSD

- Podpora ZFS od apríla 2007 (verzia 6)
- Posledné vydanie: verzia 14 v 8.1-RELEASE
- Plány: verzia 15 v 8.2, verzia 28 v 8.3 a 9.0
- Dokumentácia: wiki, stránky s manuálmi
- Podpora: e-mailové konferencie a diskusné fóra
- Budúcnosť: spolupráca s projektom Illumos

MacOS X

- MacOS X ZFS bol ukončený v októbri 2009
- Dustin Sallings: mac-zfs na googlecode a github-e, dostupný inštalátor
- Beta (nevhodné na produkčné nasadenie)

Linux

- Projekt ZFS-fuse
  verzia 0.6.9 so ZFS poolom v23
- Kernelový modul ZFS (Brian Behlendorf)
  verzia 0.5.2 - pool v28, bez vrstvy POSIX (ZPL)
- ZFS Posix Layer (ZPL) od KQ Infotech
  stavia na skorších vydaniach od Briana Behlendorfa, beta kód
- KQ Infotech (hlavný vývojár Anand Mitra) na zdokonaľovaní kódu pracuje

Právne hľadisko

- Licencia CDDL
- Patentové nároky (spoločnosť Netapp)

Licencia CDDL

Zdrojové kódy ZFS majú licenciu s názvom
Common Development and Distribution License (CDDL)

- zakladá na Mozilla Public License (MPL)
- nie je kompatibilná s GPL
- s binárnou distribúciou musí byť zverejnený aj zdrojový kód, ale iba z "dotknutého software" = originál + zmeny
- ak je ZFS časťou "väčšieho diela", toto dielo nesmie porušiť CDDL
- zmeny musia byť licencované CDDL, prispievateľ ("Contributor") musí zverejniť svoje meno
- licencia zaniká, ak sú podané patentové žaloby voči autorovi alebo prispievateľom

Patenty

Medzi spoločnosťami Netapp a Sun Microsystems bol patentový súdny spor.
Netapp žaloval porušenie okrem iného troch dôležitých amerických patentov, číslo:

- 5,819,292 (copy on write) - takmer kompletne zamietnuté (prípad uzavretý)
- 7,174,352 (snapshots) - takmer kompletne zamietnuté (prípad neuzavretý)
- 6,857,001 (writable snapshots) - začalo konanie

Spor sa skončil mimosúdnou dohodou v septembri 2010, obidve strany vzali späť svoje žaloby. Podrobnosti dohody nie sú známe.

Budúcnosť ZFS

- Vývoj ZFS v spoločnosti Oracle
- Projekt Illumos
- Budúcnosť v ostatných systémoch

Vývoj ZFS v spoločnosti Oracle

Zo spoločnosti Oracle unikla na verejnosť interná správa, ktorá obsahovala okrem iného nasledovné informácie:

- Oracle bude pokračovať vo vývoji ZFS - neverejne
- zdrojové kódy ZFS si zachovajú CDDL licenciu
- zdrojové kódy pod licenciou CDDL budú zverejnené s novými vydaniami Solaris
- vývojové verzie budú dostupné iba vybraným komerčným partnerom cez OTN (Oracle Technology Network)

Projekt Illumos

- po zavretí OpenSolaris-u ho opustilo veľa vývojárov
- niekoľkí z nich založili projekt Illumos
- projekt sponzorovaný a podporovaný spoločnosťou Nexenta
- cieľ: poskytnúť voľne zdrojové kódy (a nahradiť uzavreté časti otvorenými)
- distribúcie založené na Illumos-e: Nexenta, Belenix, Schillix

Akým kódom chcú nahradiť uzavreté časti OpenSolaris-u?
Kódom z FreeBSD! (sed, tr, em, msk)

Budúcnosť v ostatných systémoch

- FreeBSD: verzia 28 (malý tím maintainerov)
- Linux: verzia 28 (Brian Behlendorf, KQ Infotech)
- nové verzie - až keď Oracle zverejní zdrojové kódy

Ďakujem za pozornosť.

http://blog.vx.sk
http://www.vx.sk