Kontejnery + Linux

Pavel Šnajdr

InstallFest 2014

Kontejnery + Linux

- O mně, background
- Proč virtualizujeme, historie
- Typy virtualizace
- Hypervisor vs. Kontejner
- Další typické vlastnosti kontejnerů
- OpenVZ
- LXC
- Kontejnery ve vpsFree.cz a v Relbitu
- Kompromisní řešení?

O mně

- Od 2009 vpsFree.cz
- Od 2010 Relbit

Virtualizace

Proč virtualizujeme?

- Konsolidace HW, šetření nákladů, outsourcing
- Abstrakce od specifik HW
- Oddělení nesouvisejících služeb

Hlavní požadavky?

- Stabilita, hustota, nízká režie, bezpečnost, snadná správa

Virtualizace

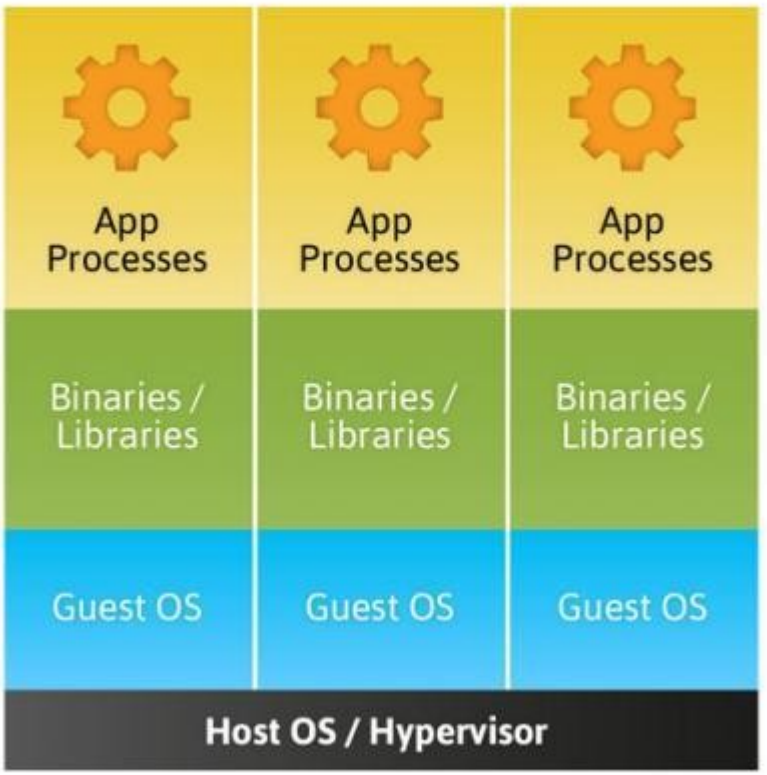
- 1998 vmware
- 2005 Xen 3
- 2008 Red Hat přebírá KVM

- Cloud
  - “Infrastructure as a service” v plném proudu
  - “Platform as a service” na vzestupu
  - “Software as a service” tu je od začátku dynamického webu

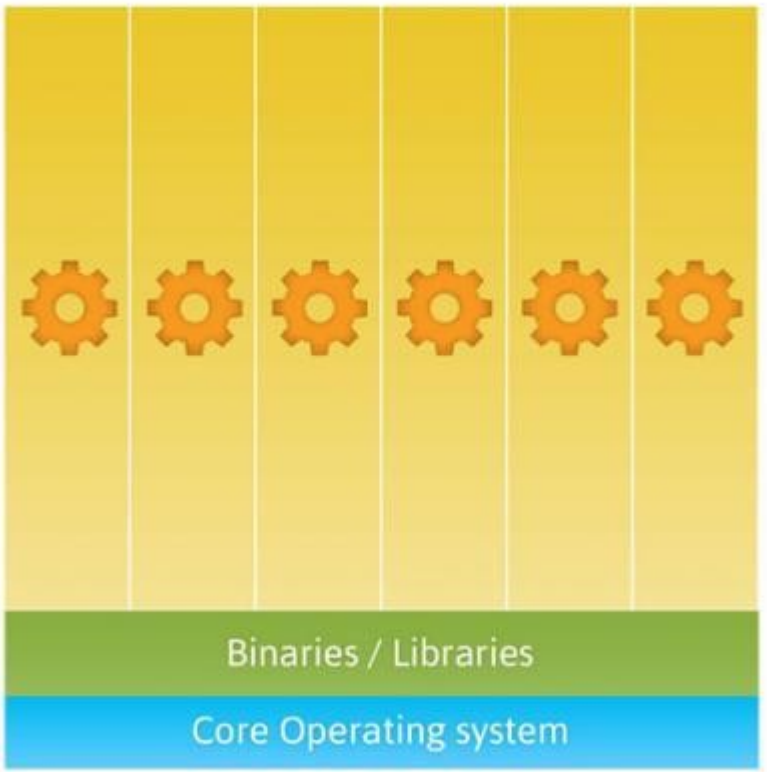
Typy virtualizace

- Hypervisor-based
  - Enterprise: vmware, Hyper-V, KVM, OVM, Xen
  - Mimo enterprise: KVM, Xen, VirtualBox
- Kernel-based
  - 1998 FreeBSD 4.0 – Jails
  - 2001 linux-vserver
  - 2001 SWSoft Virtuozzo
  - 2005 OpenVZ
  - 2006 – 2014 – LXC
- Co je hypervisorová virtualizace a co kontejner?

HVM


App Processes
App Processes
App Processes
Binaries / Libraries
Binaries / Libraries
Binaries / Libraries
Guest OS
Guest OS
Guest OS
Host OS / Hypervisor

Kontejnery



Hypervisor vs. kontejner

- N + 1 běžících jader
- Emulace hardware, paravirtuální HW
- Izolace hypervisorem + HW
- Komplikovanější netransparentní správa HW zdrojů
- Každá další generace HW dává další a další berličky hypervisorům
- Závislost na HW podpoře

- Jedno sdílené jádro
- Standardní rozhraní operačního systému
- Izolace v jádře, namespaces
- Globální plánování jedním jádrem
- Funguje všude, kde funguje dané jádro

Výhody kontejnerů oproti hypervisorům

- Efektivní správa HW prostředků
- Mnohem jednodušší overcommit
- Neměřitelný overhead
- Vyšší hustota CT na jednom HW
- Rychlý start
- Sdílená cache (dcache)
- Aplikační kontejnery

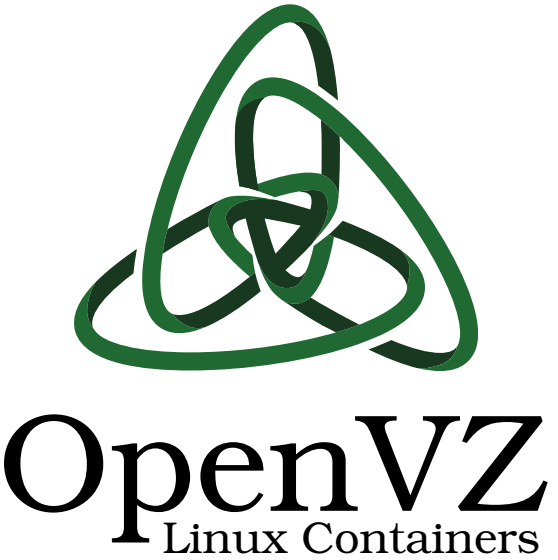
Nevýhody kontejnerů oproti hypervisorům

- Omezení na jedno jádro
- Ne všechna funkcionalita jádra je podporovaná v kontejneru
  - Správa HW, disků, oddílů
  - Omezená podpora FS (OpenVZ => FUSE)
  - Komplikovanější iptables moduly
  - Bezpečnostní moduly

Kontejnery

- Procesy viditelné na hostiteli
- Filesystem je subtree hostitele
- Spouštění příkazů z hostitele
- Možnost sdílení spustitelných objektů

- Většina Internetu může běžet v kontejnerech



- Patch pro jádro z RHEL 6 (2.6.32)
  - Userspace např. i Fedora 20 (glibc)
- Namespaces (PID, network, FS/simfs, user)
- UBC – User Beancounters
- VSwap
- vzquota
- ploop
- iptables, NFS client + server, tun, FUSE, ppp, IPSec, ...
- checkpointing / live migration
- vzctl (+ vzlist, atd.)
- venet, veth

LXC

- 2.6.24 cgroups od Google (>> UBC v OpenVZ)
- Namespaces (PID, network, FS, user) kompletní okolo 3.9
- CRIU
- V userspace je zmatek
  - vzctl
  - lxc utility
  - systemd-nspawn
  - docker.io
  - libvirt
  - OpenShift
  - Android

vpsFree.cz

- ~590 členů
- ~750 CT (VPS)
- 144 Xeon E5 jader
- 2.25 TB RAM
- 40 TB RAID10
- 5 TB SSD

- Since 2009
- OpenVZ
- Debian -> 2010
- SL6 2010+

začátky Relbitu

- PHP PaaS v utajení
- Debian 6, SL6
- KVM kdysi dávno v roce 2010
- GlusterFS k replikaci VM HD images
- Problémy se škálováním
- Nízká hustota per HW
- Skoro bez agregace
- Nestabilní, uptime zřídkakdy přes 2 týdny

Relbit Evia

- PHP PaaS jako operační systém
- SL6, OpenVZ, ZFS, NFS
- Cílová skupina: clustery ~ 10 - 1000 jader CPU
- Stabilní
- Open Source, http://eviaproject.org

Nejlepší z obou světů

- OpenVZ + KVM
  - 2.6.32 RHEL6 kernel
  - libovolné distro, doporučuje se RHEL (CentOS/SL)
  - ZFS on Linux git HEAD – snapshot, clone, send/recv, ARC, SSD
  - kontejnery kde to jen jde, KVM na zbytek
  - bridged network, MAC filter
  - Proxmox VE pro líné

Otázky?

- Kontakt:
  - pavel.snajdr@vpsfree.cz