Xen – virtualizace na Západočeské univerzitě

Michal Švamberg

20. října 2007

Virtualizace se dříve v rutinním provozu příliš nepoužívala, zvláště z důvodů potřeby šáhnout ke speciálnímu hardwaru nebo komerčnímu softwaru. Na poli svobodného softwaru existovaly virtualizační nástroje, které převážně na platformě i386 dokázaly vytvořit virtuální stroj. Jejich velkou nevýhodou byla pomalost nebo velmi úzké zaměření. Do vzniku projektu Xen neexistovalo řešení virtualizace založené na svobodném softwaru, které by bylo možné nasadit do běžného provozu na stranu serverů.

Právě možnost provozu virtuálních strojů v produkčním prostředí způsobila velkou oblibu virtualizace v posledním období. Nejčastěji zmiňované projekty v tomto světle jsou Xen a Linux VServer. O tom proč jsme pro provoz virtuálních strojů vybrali právě Xen, jak jsme jej nastavili a s jakými zkušenostmi jej provozujeme se dozvíte v tomto článku.

1 Virtualizace

Jak již bylo naznačeno, v příspěvku se budeme nadále věnovat pouze softwarovému řešení virtualizace. Zde existují tři základní přístupy:

Virtualizace emulací kompletního hardware se vytváří virtuální stroj, na němž běží hostovaný systém. Díky tomuto přístupu nevyžaduje hostovaný systém žádné úpravy; lze takto dokonce provozovat systémy určené pro jinou architekturu než je fyzická.

Nevýhodou tohoto pojetí virtualizace je vysoká režie vzniklá vlivem softwarového překladu instrukcí nebo I/O operací. Existuje několik optimalizací (dynamická rekompilace, přímé vykonání instrukce), avšak režie je i tak vysoká.

Typickými zástupci jsou BOSCH, QEMU nebo DOSEMU.

Virtualizace na úrovni OS kde je virtuálnímu stroji předložena část fyzického stroje. Jejich vzájemná izolace je zajištěna vytvořením bariér mezi těmito systémy. Všechny virtuální stroje běží nad společným jádrem, nelze tedy mít virtuální stroje různých operačních systémů. Typickým zástupcem tohoto způsobu virtualizace je VServer.

paravirtualizace Třetí skupina se pokusila spojit to lepší z obou předchozích přístupů. Virtuální stroj se nesnaží hardware simulovat, namísto toho poskytuje hostovanému systému speciální rozhraní, které k němu umožní přistupovat.

Pro tento případ je nutné hostované systémy přizpůsobit tak, aby dokázaly využít virtualizační rozhraní. Nutnost takové úpravy je určitou komplikací tohoto řešení, nicméně univerzalita a výkon, kterého lze tímto přístupem dosáhnout jednoznačně předčí oba výše zmíněné.

Paravirtualizace není jen doménou Xenu, tuhle myšlenku používá také User Mod Linux (UML). Ovšem UML bylo konstruováno za účelem ladění a vývoje linuxového jádra, proto neoplývá výrazným výkonem.

Dále existuje skupina *a*plikační virtualizace do níž patří známé Wine. Známý VMWare, ale nejen on, spadá do skupiny *n*ativní virtualizace, což je plná virtualizace s částečnou

emulací hardware. Aby byla zachována únosná režie je třeba, aby architektura hostitele byla totožná s virtuálním strojem. V případě této virtualizace není potřeba provádět modifikace hostovaného systému.

2 Xen

Umožňuje bezpečný provoz několika virtuálních strojů – každý se svým vlastním operačním systémem – na jediném fyzickém stroji. Výkon virtuálních strojů spravovaných Xenem se blíží jejich výkonu při nativním chodu.

Xen byl původně zamýšlen jako podkladní vrstva pro klastrová řešení. A to jak například pro vývoj aplikací, kdy na jednom železe lze připravovat a testovat klastrové úlohy, tak hlavně pro možnosti správy jednotlivých uzlů, aby bylo možné vzdáleně uzly uspávat či je přesouvat mezi fyzickými stroji. Projekt ale našel daleko širší uplatnění.

Ve své podstatě je Xen sada patchů pro jádra Linuxu[1] či OpenBSD a několik skriptů, kterými lze virtualizaci ovládat. Základním kamenem je hypervisor. Hypervisor je malá část jádra, která má za úkol řídit nedělitelné operace. Je to zvláště plánování procesorového času, řízení přerušení, mapování stránek či bootovací sekvence stroje a start *Domain-0*, tzv. privilegovaného stroje.

Pro řízení Xenu je třeba mít odpovídající prostředí, nejlépe s plnohodnotným operačním systémem a ovladači k hardwaru. Tuto službu nám zajišťuje *Domain-0*, který se startuje po zavedení hypervisora. Ovladače se v pojetí Xenu dělí na dvě skupiny: *front-end* a *back-end*. Ovladače typu *back-end* spravují přístup k fyzickému zařízení, ke každému back-end ovladači tedy přísluší nějaký běžný ovladač. *Front-end* ovladač je pak jakási trubka, která se z jiných strojů připojuje k back-end ovladači spravující normální ovladač. Nejběžnější způsob je, že back-end ovladače jsou v Domain-0 a front-end ovladače ve virtuálních strojích. Vzhledem k tomu, že události v systému nám spravuje hypervisor, je možné, aby zařízení bylo "exportováno" do samostatného virtuálního stroje, který bude přes back-end ovladač zpřístupňovat toto zařízení dalším virtuálním strojům.

Aby to bylo ještě zajímavější, lze mezi sebou kombinovat různé druhy systémů. Například na Domain-0 s jádrem v2.6 lze provozovat linuxové stroje s jádry v2.4 i v2.6 a navíc si jako bonus pustit OpenBSD. S nástupem Xenu 3.0 se do toho přimíchává 64bitová architektura a také možnost spouštět nemodifikované systémy. Tím se otevírá další možnost využití Xenu – vývojář softwaru již nebude potřebovat pro testování své aplikace více strojů, ale bude mu stačit jeden s mnoha virtuálními stroji.

Xen je založen na principu paravirtualizace. S nízkou režií je rozumně univerzální. Vzhledem k nutnosti zásahu do zdrojových kódů jej lze použít jen s operačními systémy, které jsou otevřené. Ve verzi Xen 3.0 přibyla podpora hardwarové virtualizace, s jejíž kombinací je možné provozovat též neupravené hostované systémy. Technologie IntelVT obsahuje podporu virtualizace v samotném hardware. Zjednodušeně lze říci, že tyto hardwarové čipy maskují existenci hypervisora jako správce prostředků, čímž si hostovaný operační systém myslí, že má celý systém pod kontrolou.

[1] Od verze 2.6.23 je Xen součástí zdrojových kódů jádra.

Zde se jedná o tzv. *k*ruhy ochrany (rings), které jsou číslovány od 0 ke 3. Operační systémy jsou zvyklé běžet v kruhu 0 a rozhodovat o prostředcích počítače. Ale při virtualizaci už je v kruhu 0 zaveden hypervisor a ten spouští hosty v kruhu 1. Zamaskovat tento problém pro nemodifikované operační systémy lze pouze při spolupráci softwarových a hardwarových prostředků. V kruhu 3 jsou spouštěny aplikace, kruh 2 nebývá využíván.

2.1 Přínosy Xenu

Zásadním přínosem nasazení Xenu jsou peníze. Použitím jednoho stroje pro více hostovaných systémů se šetří hlavně za hardware, elektřinu na napájení, místě v serverovně, ale také na chlazení nebo záložních zdrojích elektrického proudu. V případě, kdy na jednom stroji je hostováno více aplikací, roste důležitost takového stroje a jeho výpadek může být kritičtější. Proto lze doporučit pro virtualizaci zakoupit lepší hardware, nejlépe s redundantním zdrojem a vyšším počtem disků tak, aby bylo možné zapnout HW nebo SW RAID. Tím sice rostou pořizovací náklady, ale zároveň každý hostovaný systému dostává kvalitní základ.

Dalším přínosem je jednodušší správa. Snadno vytvoříte nový stroj, nemusíte napřed shánět železo, umístit jej někde a zapojovat k elektrickým rozvodům, ale ani hledat volný port na přepínači. Velmi snadno lze zjistit stav hostovaných systémů a vytvořit nový stroj včetně instalace.[2] Samozřejmě zrušení nebo přeinstalace virtuálního stroje je opět snazší a výrazně šetří čas obsluhy.

Přestože přínosy virtualizace jsou obecně stejné pro všechny virtualizační nástroje, má Xen jednu velkou výhodu – *migraci*. Pokud jsou pro Xen splněny i některé podmínky z pohledu IT infrastruktury, pak vám umožní migrování (přesunutí) virtuálních strojů. Tím lze zvýšit dostupnost hostovaných strojů, protože před plánovanou odstávkou lze virtuální stroje přemigrovat na jiný Xen server.

Nasazení Xenu má i jedno negativum, je jím riziko HW poruchy a tím odpadnutí všech hostovaných strojů. Tento problém je nevyzpytatelný a každý s ním má osobní zkušenost, ale lze jej alespoň omezit redundancí nejproblémovějších částí a poctivým zahořením jednotlivých komponent. Pokud je k dispozici migrace, pak v případě nečekaného pádu, lze virtuální stroj okamžitě nastartovat na jiném serveru a není třeba čekat na opravení původního stroje. To vše lze samozřejmě udělat vzdáleně.

2.2 Omezení

Z pohledu hostovaného operačního systému by se měl dle konceptu virtualizace chovat virtuální stroj na němž systém běží jako reálný hardware. V praxi však přece objevujeme určité rozdíly, které jsou v tomto smyslu více či méně jistým omezením. Pro virtuální stroje spravované Xenem se jedná zejména o následující případy:

Nelze virtualizovat pevné disky jako celky, pouze jednotlivé oddíly, které jsou při konfiguraci virtuálního stroje vybrány pro export. Jinými slovy, pro virtuální stroj neexistuje

[2] Automatická instalace virtuálního stroje metodou FAI včetně přípravy konfiguračních souborů pro FAI i Xen zvládáme na ZČU do 10 minut.

zařízení `/dev/hda`, přestože zařízení `/dev/hda1` existuje. S tím je třeba počítat při instalaci systému na virtuální stroj (pokud se instalátor bude pokoušet rozdělit disk na oddíly dojde patrně k selhání).

2.3 Xen na ZČU

O nasazení nějakého virtualizačního nástroje se začalo uvažovat v polovině roku 2003, kdy se nahromadilo několik požadavků na samostatný testovací stroj umístěný na serverovně. V té době chybělo skoro vše: místo v racku, záložní zdroje a hlavně volné porty na přepínačích. Naštěstí tou samou dobou byl již projekt Xen stabilizován v podobě řady 2.x. V porovnání s jinými virtualizačními nástroji měl vždy navrch, a to jak s cenou, tak rychlostí či dokumentací. Následovalo testování možností Xenu, ale i kompatibility s ostatním vybavením IT prostředí ZČU, zvláště AFS[3], FAI[4] a podpora VLAN (802.1q).

Na základě zkušeností byl zakoupen stroj s potřebným HW vybavením a rozhodnuto, že datové disky virtuálních strojů budou umístěny na Fibre Channel[5]. První rok byly na takto vzniklém stroji provozovány virtuální stroje, kterým by případný pád nehrozil, převážně pro testování. Ovšem jak čas plynul, software určený k testování přecházel v ostrý provoz a s ním narůstala i důležitost zachování provozu Xenu. Po naplnění kapacity se pořídil druhý stroj a život šel dál.

Nyní máme téměř čtyřletou zkušenost s provozem Xenu a směle jej mohu prohlásit za velmi stabilní. Nikdy jsme neřešili SW problém pádu hypervisora a naštěstí ani HW problémy. Nyní máme celkem 4 stroje, z nichž jeden je určen pro provoz ostrých virtuálních systémů, druhý pro testovací virtuální stroje. Třetí stroj byl koupen z grantu fondu Rozvoje CESNET[6] na ověření migrace virtuálních strojů. Konfigurace tří serverů je téměř stejná:

- 2xCPU Xeon na 3,2GHz s podporou Hyper Threadingu
- 4GB RAM
- 2xGbit ethernet (zatím používán jen jeden)
- 2x80GB SATA disk (partitiony v SW RAIDU - mdadm)
- Fibre Channel adapter QLA2300
- redundantní zdroj

Čtvrtý stroj byl také zakoupen v rámci grantu Fondu Rozvoje CESNET[7], který navazoval na předchozí projekt, tentokrát byl směrován k testování 64bitové architektuře a možnosti provozu nemodifikovaných virtuálních strojů při užití HW podpory virtualizace. Stroj je obdobné konfigurace, je ale 64bitový a má podporu hardwarové virtualizace.

[3] Andrew File System – `http://www.openafs.org/`

[4] Fully Automated Installation – `http://www.informatik.uni-koeln.de/fai/`

[5] Fibre Channel – `http://hsi.web.cern.ch/HSI/fcs/`

[6] Grant je veden pod číslem 154R1/2005, elektronické zdroje použité při řešení jsou k dispozici na `http://support.zcu.cz/index.php/CIV:Granty/Overeni_migrace_Xen_virtualnich_stroju`

[7] Cislo grantu je 192R2/2006, grant v současnosti probíhá a jeho součástí je také tento příspěvek.

2.4 Problémy při nasazování

Při zavádění Xenu jsme se potýkali s několika zásadnějšími problémy. Jedna z významných potíží byla kompilace jádra do `.deb` balíků nástrojem `make-kpkg`. V tomto případě šlo zvláště o nastavení správných parametrů a přejmenování defaultních názvů patchovaných jader. Zprovozněním kompilace přes `make-kpkg` byla také úspěšně zkompilován modul klienta pro OpenAFS do balíčku.

Kompilace jádra a rozšiřujících modulů do debianích balíčků byla důležitá prerekvizita pro nasazení FAI. Ten se ale potýkal při rozdělování disků s problémem neexistence disku (partitiony ale existují). Tento problém byl překonán vlastním nástrojem pro rozdělování disků; nebylo potřeba zasahovat do kódu FAI.

Virtuální stroje měly být umísťovány na různých síťových segmentech dle požadavků provozované aplikace. V hypervisorovi je síťové rozhraní nastaveno tak, že obsahuje několik interních bridgů (pro každý segment jeden) a ke každému bridgi je přivedena daná tagovaná podsíť dle 802.1q. Xen při vytváření nového hosta sám podle konfigurace napojí síťové rozhraní na daný bridge.

2.5 Migrace

Jak jste si všimli, byla z popisu vynechány informace o nástrojích a provozu Xenu. Důvodem je nošení dříví do lesa, tato část je totiž velmi dobře zdokumentována na stránkách projektu `http://xen.sf.net/` a k dispozici je také LiveCD na kterém si lze Xen vyzkoušet. Mimo jiné existuje mnoho různých howto, které vás do problematiky instalací a provozu uvedou.

Proto rovnou odskočíme k tématu migrace, kterou úspěšně využíváme na ZČU přes rok. Xen dokáže hibernovat hosta do souboru (virtuální stroj je zastaven a obsah paměti uložen) a později jej znovu probudit. Pokud bychom jej chtěli probudit na jiném stroji, musíme splnit dvě podmínky:

- zachování disku jako blokového zařízení (SAN, NAS, iSCSI, ...),
- zachování síťového segmentu.

Právě tomuto se říká v pojetí Xenu migrace. Tu rozlišujeme na *off-line* a *on-line*. Rozdíl mezi nimi spočívá v rychlosti migrace a dostupnosti systému během migrace. Off-line migrace probíhá prakticky ve výše popsaných třech fázích: uspání, přesun obrazu paměti na nového hypervisora, probuzení. Během těchto fázích není hostovaný systém dostupný, ale migrace je provedena bez zbytečného zdržování. Naproti tomu on-line migrace umožňuje jen minimální výpadek (méně než 1 vteřina[8]), ale provedení migrace je časově i procesorově náročnější – musí se navíc hlídat zápisové operace do paměti.

2.6 Nemodifikované operační systémy

Mezi novinky v řadě 3.x patří možnost provozu nemodifikovaných operačních systémů. Má to ale háček, je potřeba mít HW podporu virtualizace. Podporou se rozumí, že musíte mít

[8] Záleží na zatížení stroje, zvláště počtu prováděných operací v paměti.

správný čipset, dále procesor a zapnutou podporu v BIOSu. V Xenu jsou podporovány technologie od Intelu[9] i AMD[10].

Na ZČU se objevila potřeba občas otestovat novou aplikaci v MS Windows. Kupovat a instalovat nový stroj pro tyto občasné potřeby není z ekonomického hlediska ideální.

Podle různých návodů a vlastních zkušeností se nám podařilo na stroji s podporou IntelVT[11]. Virtuálnímu stroji jsou potřebná rozhraní k fyzickému stroji emulována, jde zejména o BIOS, grafickou kartu, síťovou kartu a disky. Xen využívá pro tyto účely kód z projektu QEMU, což je patrné při startu. Tato emulace I/O rozhraní může být přímo navázána na jednotlivé komponenty díky řízení přístupu, který zajišťuje hardwarová podpora virtualizace. To nám zajistí dostatečné oddělení virtuálních strojů a zároveň neztrácíme rychlost daných operací. Zároveň není potřeba provozovat modifikovaný systém, který by si rozuměl s hypervisorem v Xenu.

V říjnu 2007 jsme měli k dispozici jeden stroj s hardwarovou podporou virtualizace. Zároveň je tento stroj plně 64bitový, což nám otevřelo cestu ubírat se směrem ke zkoušení nemodifikovaných operačních systémů, ale poskytnout správcům služeb testovací platformu pro 64bitovou architekturu.

3 Závěr

Tento příspěvek si nekladl za cíl seznámit uživatele s detaily virtualizace, proto se vyhnul i některým důležitým pojmům nebo je jen okrajově zmínil. Z článku by měl být patrný způsob nasazení a vývoj virtualizace na Západočeské univerzitě, možnosti jeho využití i nástin plánů do budoucnosti. Dnes je k dispozici více virtualizačních technologií, které poskytují takový výkon, aby mohly být nasazeny v provozu. Velmi zdárně se vyvíjí projekt Linux VServer[12] nebo KVM[13], který je od verze 2.6.20 přímo v jádře. Záleží pouze na vašich potřebách, protože každá technologie má něco pozitivního i negativního a bez kompromisů se neobejdete.

Literatura

[1] Miroslav Suchý: *Úvod do virtualizace pomocí XENu*
`http://www.root.cz/clanky/uvod-do-virtualizace-pomoci-xenu/`

[2] Ian Pratt a kol.: *Live Migration of Virtual Machines*
`http://www.cl.cam.ac.uk/research/srg/netos/papers/2005-migration-nsdi-pre.pdf`

[3] Kolektiv autorů: *Xen 3.0 User Manual*
`http://www.cl.cam.ac.uk/research/srg/netos/xen/readmes/user.pdf`

[9] Nazývá se kódovým označením *Vanderpool*, procesor lze poznat podle příznaku `vmx`.
[10] AMD ji označuje názvem *Pacifica*, procesory obsahují příznak `vms`.
[11] IntelVT je starší název podpory virtualizace, která nakonec dostala název Vanderpool.
[12] Domovská stránka projektu VServer: `http://linux-vserver.org`
[13] Informace o projektu KVM jsou na `http://kvm.qumranet.com/kvmwiki`