Optimalizace správy a zabezpečení virtuálních strojů

Ing. Michal Švamberg, Ing. Jan Krčmář

Západočeská univerzita v Plzni
Centrum informatizace a výpočetní techniky
email: svamberg@civ.zcu.cz, honza801@civ.zcu.cz

14. července 2017

Závěrečná zpráva projektu Fondu rozvoje CESNET, z.s.p.o. pro rok 2017 vedená pod číslem 571R1/2015. Tento projekt je zařazen do oblasti **I.**, tématického okruhu **A**. Řešitel projektu Ing. Michal Švamberg.

Struktura dokumentu

Struktura tohoto dokumentu je v souladu s podklady pro závěrečené oponentní řízení Fondu rozvoje CESNET, z.s.p.o. rozdělena následujícím způsobem:

- Způsob řešení
- Dosažené cíle
- Zdůvodnění změn v projektu
- Konkrétní výstupy
- Přínosy projektu
- Tisková zpráva

Způsob řešení

Projekt byl řešen v souladu se žádostí. Jako první krok jsme provedli vyhledání a srovnání existujících řešení, abychom vybrali vhodného kandidáta pro správu virtualizačního clusteru. Na základě kritérii

- náročnosti implementace,
- nákladů na provoz,
- přehlednosti uživatelského i administrátorského rozhraní,
- využití standardních nástrojů,
- snadnost rozšiřitelnosti,
- aktuálního vývoje

jsme vybrali projekt WebVirtCloud[1]. V porovnání s ostatními projekty jako je OpenNebula[2], OpenStack[3], oVirt[4] se nám jevil jako nejvhodnější.

Podle plánu byly nakoupeny tři servery vhodné pro virtualizaci v celkové hodnotě 380.420,- Kč bez DPH.[5] Servery byly nakoupeny z výběrového řízení[6] dle platné legislativy.

Architektura diskového prostoru pro virtualizaci byla upravena do podoby (od nejvyšší vrsty):

[1] https://github.com/honza801/webvirtcloud

[2] Komplexní, určený pro větší nasazení.

[3] Příliš komplexní pro naše potřeby.

[4] Je silně postavený na Red Hat distribuci a jeho odvozenin.

[5] Zvolili jsme 2 různé konfigurace, původně se počítalo s třemi stroji se 128GB RAM, ale klesající ceny umožnily koupit jeden stroj s 384GB RAM. Oproti žádosti jsme také pořídili vícejádrové procesory, z plánovaných 6 jader na 10 a 12 jaderné. Ostatní parametry se oproti žádosti nezměnily.

[6] https://zakazky.zcu.cz/contract_display_1009.html

1. souborový systém virtuálního stroje, např. `xfs`
2. virtuální zařízení[7] uvnitř stroje, např. `/dev/vda`
3. soubor `.qcow2` exportovaný jako virtuální zařízení do hosta, např. `/fc-sas/libvirt/storage/ourea22.qcow2`
4. sdílený souborový systém GFS2[8], bude probráno detailněji níže
5. CLVM[9] pro snažší správu diskového prostoru, vznikne např. `/dev/fc-sas/gfs2`
6. multipath[10] pro zvýšení dostupnosti[11], např. `/dev/dm-1`
7. disková FibreChannel infrastruktura např. `/dev/sda`, `/dev/sdb`

Celé řešení nám takto dává cluster pro virtualizaci. Nejdůležitější je správa souborového systému a Cluster LVM, která musí být přes všechny uzly synchronní. O to se starají nástroje z Red Hat cluster suite[12], které byly portovány do mnoha dalších distribucí včetně Debianu.

GFS2 bylo vybráno z důvodů využít existující FibreChannel diskovou infrastrukturu. V obecném pojetí stačí jakýkoliv souborový systém, například NFS. Důležité je, aby byl k dispozici prostor, v němž lze uložit .qcow2 obrazy virtuálních strojů. Ukládání virtuálních disků do souboru oproti vyhrazeným oddílům[13] přineslo výhody:

- úspora místa – virtuální stroj nemá žádný vyhrazený oddíl, tudíž nealokuje volné místo
- snadnější manipulace – klonování, vytváření snapshotů či migrace je otázkou vteřin
- snadnější rozšíření prostoru – stačí navýšit limit a zvětšit souborový systém
- jednodušší vytváření snapshotů – podporováno přímo knihovnou `libvirt`
- snadnější migrace virtuálních strojů – opět přímá podpora v `libvirt`

Nejzásadnější změna byla tedy v přechodu z vyhrazených LVM oddílů na .qcow2 formát umístěný na sdíleném souborovém systému GFS2. Vyzkoušeli jsme také souborový systém OCFS2[14]. OCFS2 jsme ale po úvodních pokusech opustili, občas se stalo, že pádem jednoho uzlu souborového clusteru byly sestřeleny všechny uzly (chyba v jaderném modulu)

[7] Pro linuxové systémy nepoužíváme rozdělování disků, pro operační systém Windows je situace opačná

[8] https://en.wikipedia.org/wiki/GFS2

[9] https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Logical_Volume_Manager_Administration/LVM_Cluster_Overview.html

[10] https://en.wikipedia.org/wiki/Linux_DM_Multipath

[11] Z více cest k jednomu zařízení vytvoří jedno zařízení

[12] https://en.wikipedia.org/wiki/RedHat_cluster_suite

[13] V původním řešení byly vyhrazeny LVM oddíly pro každý virtuální stroj, minimálně kořen a swap.

[14] A to na základě zkušeností z provozu Oracle databází.

a servery musely být resetovány. Je třeba říci, že GFS2 není také bez chyby, ale problémy nám přišly snadněji řešitelné a nemívaly tak fatální dopad.

Dále jsme přešli od technologie virtualizace Xen na KVM. Od nasazení KVM si slibujeme lepší podporu pro virtualizaci operačního systému Windows, zvláště pak při aktualizacích virtualizační platformy. Druhou výhodu spatřujeme ve větší komunitě uživatelů (a potažmo vývojářů), navíc podpora KVM je přímou součástí linuxového jádra, což u Xenu donedávna nebylo.

Jak bylo řečeno již dříve, rozhodli jsme se pro správu systémem WebVirtCloud[15]. Jedná se o menší projekt využívající rozhraní `libvirt`[16]. Je psaný v jazyce Python nad frameworkem Django[17]. WebVirtCloud jsme rozšířili a doplnili, změny byly autorem do projektu přijaty[18], takže nyní jsou již jeho součástí. Ke dni 16.6.2017 je Jan Krčmář (spoluřešitel) druhým nejčastějším přispěvatelem[19] do WebVirtCloudu hned po samotném autorovi.

Uživatelé si pochvalují jednoduché a snadné ovládání WebVirtCloudu. Pro nejčastější zásahy je k dispozici administrátorské rozhraní. Pouze pokud přijdou speciální požadavky (nejčastěji týkající se kombinace diskových prostorů) je nutné použít přímo nástroje pro libvirt nebo operaci provést z příkazové řádky hostitele.

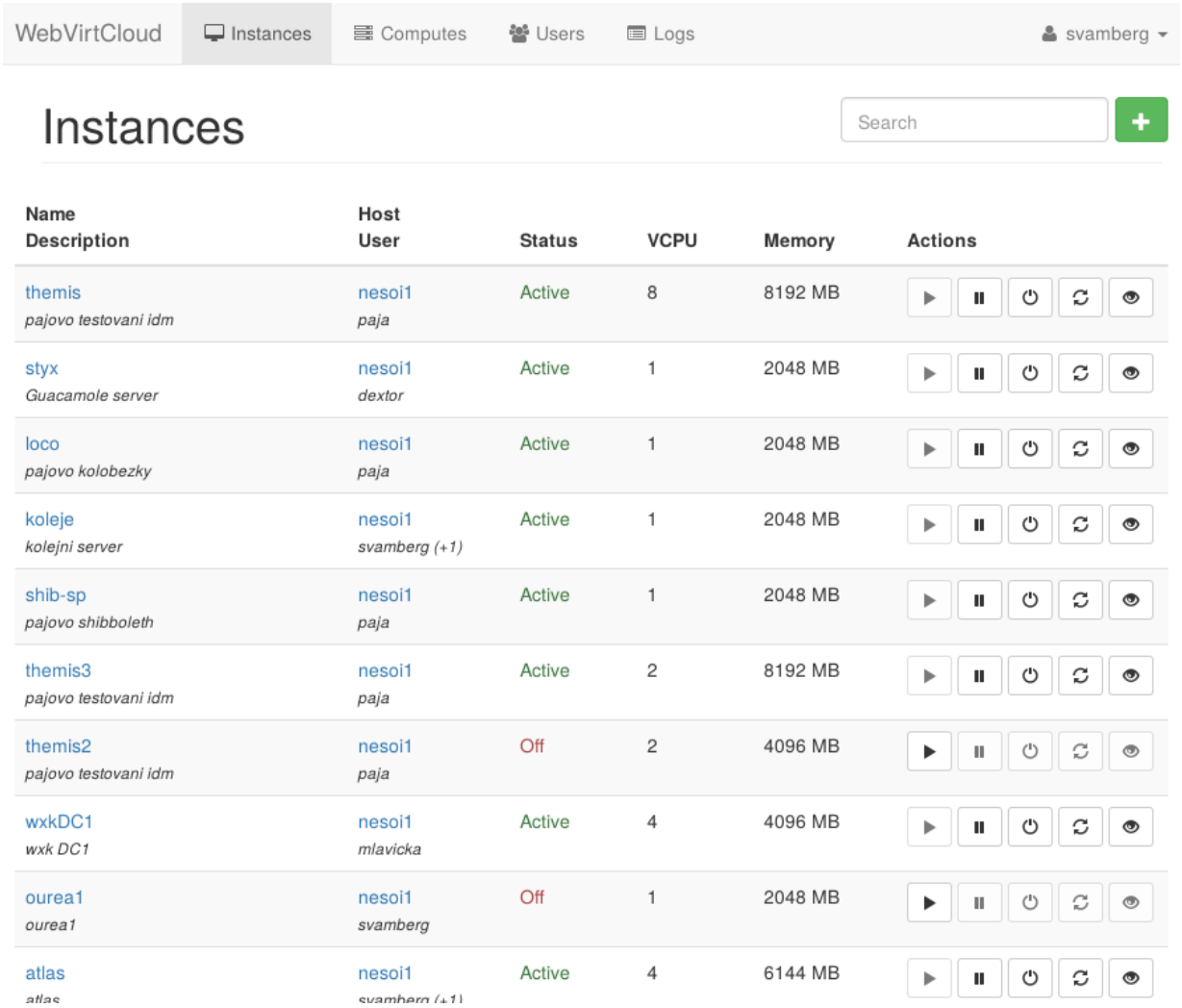
Virtuální stroje jsme taktéž chtěli nabídnout běžným uživatelům. Dnes existuje mnoho služeb, kde si takový server lze zřídit, často ze začátku zdarma. Ale my jsme chtěli uživatelům nabídnout server s návazností na existující univerzitní infrastrukturu (jednotné přihlašování, předpřipravený souborový systém AFS, atd.). Zároveň jsme takové stroje chtěli lépe zabezpečit než jsou standardní distribuce linuxu. Další bezpečnostní službou uživatelům měla být aktualizace operačního systému a jeho vybavení v době, kdy vyjdou záplaty chyb.

Shodně s žádostí projektu byla tato (a mnohá další) nastavení zajištěna konfiguračním managementem CFEngine. Konfigurační management je předpřipraven v šablonách (templates) ze kterých vznikají virtuální stroje. Podstatná část projektu je tak zcela oddělená a nebrání přenést znalosti do jiných organizací. Bezpečnostní nastavení pak záleží na přípravě obrazu. My jsme zvolili minimalistický obraz jen s potřebným minimem nástrojů a přednastaveným konfiguračním managementem, který zajistí úpravu operačního systému tak, aby odpovídal aktuálním politikám konfiguračního serveru.

Projekt předpokládal maximálně automatizovat vytváření a rušení serverů včetně návazností na další systémy. Registraci virtuálních strojů jsme vyřešili tak, že máme předgenerované doménové jména a nastaveny k nim statické adresy včetně MAC adres pro DHCP.

Další systém, který jsme chtěli napojit na virtuální stroje byl monitoring. Po prozkoumání možností jsme zjistili, že nejsnažší způsob by byl přístup přes REST API do monitoringu a stroj se sám nastavil. Bohužel aktuálně provozovaný monitoring[20] nám toto

[15] `https://github.com/honza801/webvirtcloud`
[16] `https://libvirt.org/`
[17] `https://www.djangoproject.com/`
[18] `https://github.com/retspen/webvirtcloud/commits?author=honza801`
[19] `https://github.com/retspen/webvirtcloud/graphs/contributors`
[20] Nagios a správa konfigurace přes webové rozhraní NConf.



Obrázek 1: Přehledová stránka všech spravovaných strojů

neumožňuje. Navíc připravujeme zcela nově postavený monitoring na Icinga 2[21], která má REST API vestavěné. Prozatím jsme vyřešili dohled tak, že CFEngine zná konfiguraci Nagiosu a hlídá, zda odpovídá provozovaným službám.[22] Toto řešení není dokonalé, ale do doby než bude připraveno nové monitorovací prostředí, je dostačující.

Jak již bylo naznačeno dříve, instalaci jsme se rozhodli řešit formou předpřipravených obrazů, které se po startu přizpůsobují (roztáhnou souborový systém, nastaví hostname, inicializují konfigurační management, ...). Dříve zvažovaný systém FAI[23] jsme opustili a přešli k nástroji DebianInstaller s preseed[24] nastavením. Uživatelé pro instalaci používají

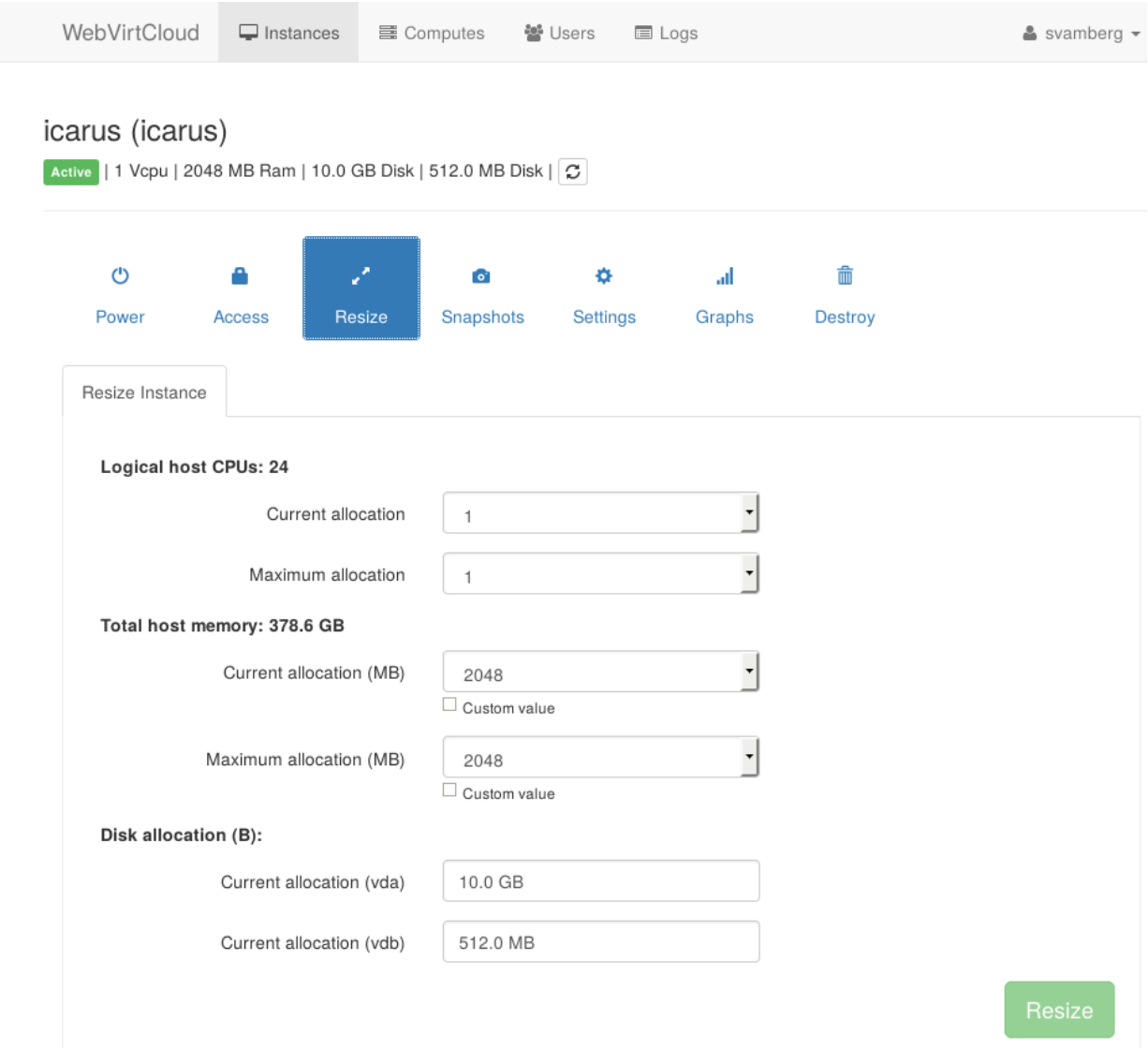
[21] https://www.icinga.com/

[22] Pokud se na stroj nainstaluje databáze a v Nagiovi není sonda na její kontrolu, tak upozorní na tuto nesrovnalost správce. Ten ve webovém rozhraní NConf sondy danému hostu přidá.

[23] http://fai-project.org/

[24] https://wiki.debian.org/DebianInstaller/Preseed

klonování obrazů, které je rychlé a v cloudovém prostředí běžné. Chtěli jsme ale instalaci virtálních a fyzických strojů co nejvíce sblížit, tedy instalovat jen potřebné minimum. Přípravu instalace fyzických strojů tak nyní můžeme otestovat snadno na virtálním stroji, což nám přináší zjednodušení a urychlení práce.



Obrázek 2: Možnost změnit parametry virtuálního stroje

Dosažené cíle

Byla vytvořena zcela nová virtualizační platforma, která zjednodušuje práci administrátorům a uživatelům nabízí přehledné a jednoduché rozhraní. Uživatelé si mohou vytvořit

vlastní virtuální stroje a to do limitů stanovených správci. Otázka bezpečnosti byla zahrnuta jako součást předpřipravených obrazů, zároveň konfigurační management běží také na všech serverech virtualizační platformy. Konfigurační management na strojích (hostitelích i hostech) hlídá aktuálnost nainstalovaných balíků, nastavení firewalu, hesel a mnohé další volby, které ovlivňují bezpečnost. Další možnosti přidáváme podle toho, jak přicházejí nové požadavky.

Novým virtualizačním systémem se podařilo odstranit největší bolesti toho starého. Šlo zejména o rozdělování disků a přiřazování diskového prostoru, chybějící uživatelské rozhraní, komplikovanou instalaci a inicializaci virtuálního stroje a nízká úroveň bezpečnosti. To byly hlavní důvody, proč jsme nemohli virtuální servery nabídnout všem uživatelům univerzity.

Všechna bezpečnostní opatření popsaná výše závisí na tom, zda bude aktivní konfigurační management na daném virtuálním stroji. U strojů, které jsou ve správě, tuto vlastnost ohlídá monitorovací systém. Problém nastává u „rychloobrátkových“[25] virtuálních strojů nebo těch, kde práva administrátora má také uživatel, který stroj vytvořil. Administrátor může totiž konfigurační management vypnout, čímž také vypne všechna bezpečnostní opatření. Tomu jsme se rozhodli zabránit tak, že na serveru konfiguračního managementu se kontroluje časová značka, kdy naposledy klient konfiguračního managementu kontaktoval server. Pokud se stane tato značka neaktuální, je vlastníkovi poslána výstraha emailem a virtuální stroj je vypnut. Uživatel si jej může z webového rozhraní opět nastartovat a provést nápravu, pokud tak opět neučiní, server se zase vypne.

Bohužel se nám nepodařilo najít žádného studenta, který by byl ochotný na projektu hlouběji spolupracovat. Není mnoho studentů, kteří se chtějí věnovat správě linuxových operačních systémů. Alespoň pro testování jsme získali Marka Zimmermanna, který nám již v rané fázi přípravy přinesl zpětnou vazbu.

Zdůvodnění změn v projektu

Oproti původnímu plánu jsme provedli několik drobných změn v projektu. Místo původní konfigurace serverů jsme nakoupili jeden v lepší konfiguraci. Důvodem byla klesající cena serverů. Pro stanovení qvóra clusteru jsme potřebovali minimálně 3 stroje. Ke konci projektu jsme k přidali další 2 stroje, které do hospodaření projektu nezapočítáváme. Jsou to stroje, které navazují na projekt a rozšiřují jej.

V implementační části projektu jsme se rozhodli nenapojovat systém těsněji na monitoring než je v současnosti. Zjistili jsme, že vhodné řešení by vyžadovalo významnější zásahy do monitorovacího systému. Proto jsme s kolegy začali diskutovat o možné změně monitoringu[26] tak, aby to vyhovalo i dynamickému vzniku či zániku serverů v cloudovém prostředí. Současné řešení nás upozorňuje na změny, ale ty se musí zanášet do systému ručně, což není úplně dle představ, ale do nového monitoringu s tím vydržíme.

[25] Pokud potřebuji stroj jen na krátkodobý pokus. Stroj může být používán v jednotkách hodin nebo několika dní.

[26] Aktuálně se testují možnosti přechodu z Nagios 3.5.1 na Icinga 2.

V projektu se nám nepodařilo úplně dobře čerpat z cestovného a konferenčného. Důvodem byl fakt, že původně plánované výdaje na konferenci OpenAlt se rozhodl zaplatit organizátor konference sám.

V žádosti jsme uvedli, že bude třeba oslovit uživatele a toto řešení jim nabídnout. Zatím řešení virtualizace nabízíme, pokud přijde požadavek na zřízení serveru. Hromadněji chceme uživatele oslovit až po aktualizaci operačního systému, pravděpodobně se začátkem nového akademické roku 2017/18.

Konkrétní výstupy

K 10.7.2017 jsme na celkem 5 hostovských serverech provozovali 104 virtuálních strojů. Požadovaný diskový prostor virtuálních strojů je 3,5TB, ale reálně zabráno 1,7TB. To znamená, že oproti původnímu Xen řešení šetříme 52% diskového prostoru.

Během projektu jsme uspořádali několik seminářů, na kterých jsme představili naše řešení, ukázali jeho výhody a předvedli používání.

- Seminář pro IT techniky CIV, jaro 2016, Plzeň (neveřejné)
  Sloužilo administrátorům pro seznámení s novým prostředím virtualizace a jejich možnostmi. Cílem byla diskuse o použitých technologiích, navázání na infrastrukturu a hlavně získání zpětné vazby po tom, co jej začali používat.

- OpenAlt, 5.11.2016, Brno
  Přednáška o řešení projektu[27] s krátkou ukázkou používání. K dispozici je také video záznam[28].

- EurOpen, 16.5.2017, Myslovice
  Prezentace stavu virtualizace[29] a jeho řešení. Součástí je příspěvěk do sborníku konference[30]. V něm lze najít porovnání nejčastějších cloudových rozhraních i konkrétní technické řešení tohoto projektu.

Vzhledem k velmi silné pozitivní vazbě od správců IT, kteří virtuální prostředí začali aktivně používat, jsme se rozhodli informaci o možnosti zřídit si vlastní virtuální stroj uvolňovat trochu obezřetněji. A to minimálně do doby než virtualizační infrastrukturu posílíme[31]. Na druhou stranu, pokud je příležitost, tak systém nabídneme. Máme tak již virtuální servery pro katedry, projekty i studenty.

[27] https://openalt.cz/2016/data/Svamberg%20-%20webvirtcloud.pdf
[28] https://www.superlectures.com/openalt2016/jednoduchy-cloud-vhodny-i-pro-univerzitu
[29] http://europen.cz/Proceedings/50/Svamberg_europen_webvirtcloud.pdf
[30] http://europen.cz/Anot/50/eo-1-17.pdf
[31] Aktuálně je uzavřené výběrové řízení na dodávku serverů mezi kterými je také posila virtualizační platformy. Dodání serveru očekáváme v průběhu léta 2017.

Z hlediska správy operačního systému se nám podařilo sjednotit management serverů včetně zohlednění bezpečnostních hrozeb a aktualizací. Nyní můžeme upravit hromadně všechny servery do 10 minut[32].

Vznikla uživatelská dokumentace s návodem[33], jak si vytvořit virtuální stroj. Vytvořena dokumentace s poznatky o provozu clusteru[34] pro virtualizaci v Debianu.

Cluster se aktuálně sestává z 5 nodů[35] s celkovou pamětí 2T a počtem procesorů 192. Na této infrastruktuře aktuálně provozujeme 104 virtuálních strojů. Navýšení bylo potřebné z toho důvodu, že připravujeme převod původních virtuálních strojů z technologie Xen na KVM a také aktualizovat infrastrukturu z Debian Jessie na Stretch[36].

Přínosy projektu

Zjednodušení přípravy a samotné správy virtuálních strojů je hlavním přínosem pro administrátory operačních systémů. Pro uživatele pak snadný způsob vytvoření, ovládání, přístupu i zrušení stroje. O oblíbenosti tohoto rozhraní hovoří čísla. V systému je k 11.7.2017 aktivních 45 uživatelů, z toho 40 jsou běžní uživatelé, tj. nemají aktivovanou roli administrátora.

Přechodem na virtualizaci KVM očekáváme lepší podporu pro nelinuxové operační systémy, zvláště pak pro MS Windows. Zde byly potíže při aktualizaci hypervizorů.

Velká úspora je pak na diskových polích, kde je prostor velmi drahý a je ho omezené množství. Nyní je úspora místa 52% oproti původnímu řešení.

Tisková zpráva

Projekt na Západočeské univerzitě zvýšil dostupnost a snadnost vytvoření vlastního virtálního stroje. Umožnil lepší využití hardware a významně šetří diskový prostor oproti původnímu řešení. Byla vylepšena automatizace životního cyklu virtuálního stroje a napojení na návazné informační systémy. Pro správu virtuálních strojů byl nasazen projekt WebVirtCloud. Získané znalosti je možno využít i v dalších organizacích.

[32] Synchronizace master serveru s GIT repozitářem se provádí v 5 minutových intervalech, klienti kontrolují konfiguraci ze serveru každých 5 minut.

[33] http://support.zcu.cz/index.php/Cloudov%C3%A9_slu%C5%BEby

[34] http://support.zcu.cz/index.php/Public:Honza801/debian_cluster

[35] Oproti původně plánovaným třem jsme museli již během projektu počet navýšit.

[36] Debian Stretch byl jako stabilní distribuce vydán 17.6.2017.