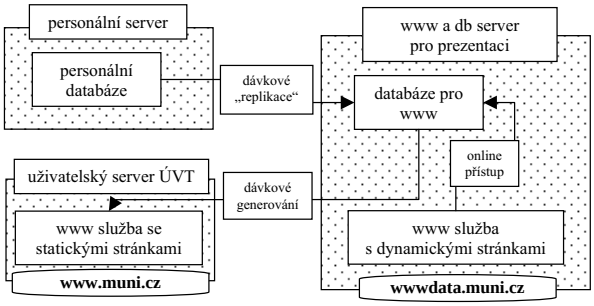
Cluster www-serverů MU

Jaromír Ocelka, ÚVT MU

Internet se dostává do povědomí čím dál tím většího počtu osob; v důsledku toho jsou www servery, které svým výkonem stačily ještě před několika lety, stále častěji přetížené. Lidé si více zvykají hledat informace přímo na webu namísto v tištěných dokumentech, takže případný výpadek informačního serveru má dnes závažnější následky. WWW servery vysokých škol jsou tímto trendem postiženy možná ještě více, neboť každý uchazeč o studium na vysoké škole je obeznámen s Internetem nejpozději na střední škole. Nárůst zájmu a s ním spojené výkonnostní problémy se nevyhnuly ani www-prezentacím Masarykovy univerzity a daly tak impuls ke vzniku univerzitního www-clusteru.

1 Vznik www.muni.cz

Na přelomu let 1996/1997 se MU rozhodla vytvořit zastřešující centrální www server, který by jednotně prezentoval základní profilové informace o univerzitě (viz články v minulých ročnících Zpravodaje). Větší část těchto informací bylo rozhodnuto čerpat z personální databáze univerzity a umožnit tak jejich automatickou aktualizaci – jde například o struktury fakult a seznamy zaměstnanců/studentů. Celkem se tehdy jednalo o cca 13 tisíc stránek (přes 10 tisíc studentských a 2 tisíce zaměstnaneckých osobních stránek, stránky pracovišť atd.), které bylo nutné prezentovat virtuálním návštěvníkům univerzity. Vzhledem k tomuto velkému počtu byla zamítnuta varianta vytvoření (byť automatizovaného) všech stránek ve statické podobě a byla využita technologie dynamických stránek, kdy www-server generuje obsah stránek až na vyžádání. Pouze nejdůležitější část prezentace byla z preventivních důvodů generována do statických stránek prezentovaných jiným www serverem (jednalo se o celkem 300 statických stránek – hlavní stránku univerzity a základní stránky o fakultách a pracovištích). Další otázkou bylo, zda využívat personální databázi přímo. Z bezpečnostních důvodů zvítězila nakonec varianta odděleného databázového serveru, který brání případným útokům na personální databázi.



Obrázek 1: Technické schéma www prezentace

Uživatelský server pracuje pod operačním systémem SunOS, pro www službu (pro statické univerzitní stránky s adresou `www.muni.cz`) je použit software Apache. Pro dynamické stránky a vlastní podpůrnou databázi byl pořízen jediný server Intel Pentium Pro s operačním systémem Windows NT 4.0, www službou IIS 3.0 (veřejnosti přístupným pod hlavičkou `wwwdata.muni.cz`) a databázovým strojem MSSQL 6.5.

2 Rozmach návštěvnosti

Téměř od okamžiku vzniku www prezentace MU jsou udržovány také záznamy o její návštěvnosti; díky tomu bylo možné konfrontovat hardwarové možnosti serverů nejenom s aktuální vytížeností, ale i s předpokládanými navýšeními v budoucnu. Na www službě `wwwdata.muni.cz` bylo koncem roku 1998 zaznamenáno v průměru 90 požadavků[1] na stránky za hodinu. Za rok 2002 se tento ukazatel zvýšil na hodnotu 400 a i podle údajů z prvního čtvrtletí roku 2003 dále trvale roste (aktuální stav je cca 600 požadavků za hodinu). Tyto hodnoty však musíme interpretovat s vědomím, že se může jednat jak o častější požadavky jednoho uživatele, tak o požadavky více uživatelů. U veřejné prezentace totiž není možné obecně rozlišit požadavky různých uživatelů a proto se počet uživatelů nahrazuje počtem různých klientských stanic rozlišených jejich IP adresou. Rostoucí řada hodnot vyjadřující průměrný počet unikátních IP adres za hodinu je od roku 1998 následující: 9, 14, 23, 45, 50 a v letošním roce vykazuje poslední člen řady navýšení na hodnoty vyšší než 80. Podobně je rostoucí průměrný počet různých

[1] Ze statistiky jsou vypuštěny požadavky na obrázky.

IP adres, z čehož vyplývá, že přibývá stále více nových uživatelů. V letech 1998-2001 byly hodnoty (v tis.) této statistiky: 18, 26, 42, 76, 94. Za první čtvrtletí roku 2003 byla již dosažena polovina stavu roku 2002. Vybrané statistiky jsou k dispozici i návštěvníkům prezentace na adrese `http://wwwdata.muni.cz/wwwstat/`.

3 Analýza architektury systému

Pro posuzování potřeby posílení hardware serveru se statistiky o návštěvnosti dávaly do souvislosti se statistikami o vytížení systému. Zde je nutné poznamenat, že server `www.muni.cz` „trpí“ zvýšením počtu návštěvníků mnohem méně, neboť jeho jedinou činností v reakci na www požadavek je vrátit obsah souboru.

Jelikož z ekonomických důvodů není obvykle možné server dimenzovat tak, aby vyhověl všem hypotetickým (špičkovým) situacím, nezbývá než odhadnout další osud křivky návštěvnosti a podle ní celý systém rozumně nastavit. Správce www serveru by měl přitom vědět, že je možné www server zahltit, aniž by se jednalo nutně o záměrný útok. Typickým příkladem je rozšíření informace o zajímavé www adrese v některém on-line sdělovacím prostředku. Za zmínku stojí dva obecně známé případy z ČR. První se udál na jedné komerční TV, kdy její ředitel sdělil divákům adresu stránky a teprve poté chtěl ukázat obsah stránky na obrazovce. Již ji neukázal, neboť server byl totálně přetížen vysokým počtem návštěvníků. V nedávné době vznikl obdobný problém při zveřejňování seznamů spolupracovníků STB.

Protože www služby by ze své povahy měly umět zpracovávat požadavky uživatelů paralelně, a v případě dynamických stránek je tato nutnost přenesena i na databázový stroj, je vhodné, aby webový systém podpořil paralelní zpracování více procesory. Mimo jiné i z tohoto důvodu byl na počátku roku 2000 pořízen nový server se dvěma procesory Intel Pentium III 500 MHz, 256 MB paměti a jedním SCSI diskem o kapacitě 8 GB. Pro provoz serveru byl zvolen operační systém Windows 2000 Server s www službou IIS 5.0 a databází MSSQL 7.0. Aby nedošlo k zahlcení serveru zvyšujícími se počty uživatelů, byly statistiky návštěvnosti průběžně vyhodnocovány a na základě toho docházelo k postupnému rozšiřování hardwaru. Paměť byla rozšířena na celkovou kapacitu 512 MB a postupně byly přidávány další SCSI disky až do počtu 4, aby bylo možné zrcadlením disků rozložit zátěž databáze a dosáhnout co nejmenšího počtu neplánovaných výpadků (což se v roce 2002 ukázalo jako velmi užitečné při havárii jednoho z disků).

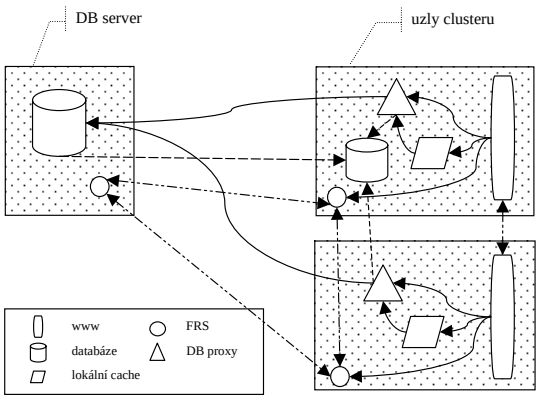
Během let 2000 až 2002 se průměrný počet požadavků na www-server více než zdvojnásobil; bylo proto nutné vymyslet takovou strategii, která by dlouhodobě umožnila výkonnostně držet krok se stále rostoucím tempem uživatelů. Jednou z možností bylo průběžně pořizovat nové výkonnější hardwarové komponenty či servery. Tím však okamžitě vyvstane řada otázek, počínaje „Kde na to vzít?“ a konče „Kam s tím?“. Navíc jediný server je rizikový z hlediska možné hardwarové poruchy. Vhodným řešením se ukázalo propojování více nezávislých www-serverů do www-clusteru.

Idea rozšířit www prezentaci MU z jednoho počítače na clusterové řešení se zrodila na jaře roku 2002. Cílem bylo dosáhnout jak vyšší odolnosti proti výpadku hardwaru, tak zvýšit výpočetní schopnosti a možnosti škálovatelnosti celého systému v souvislosti s neustálým přílivem uživatelů.

WWW prezentace MU se skládá ze dvou částí – www služby a databáze. Vzhledem k tomu, že není nutné v samotné www službě veřejné prezentace udržovat stavové informace, postačuje pro rozložení zátěže mechanismus *load balancing*, tj. rovnoměrné rozdělování www požadavků mezi více serverů s identicky konfigurovanou www službou. Pro intranet www prezentace, který si po dobu přihlášení uživatele udržuje v paměti stavové informace, bylo navíc využito možnosti definovat výjimku, aby všechny požadavky uživatele s protokolem https byly vyřizovány vždy jedním stejným uzlem www-clusteru. Problémem zůstává cluster databází, kde load balancing nepostačuje z důvodů například on-line modifikací z intranetu. Řešením by bylo použití databáze s podporou clusteringu. Avšak zatímco pro realizaci mechanismu load balancing postačuje instalace OS Windows 2000 Advanced Server, pro clusterování databází je nutné navíc

pořídit MS SQL 2000 v Enterprise verzi a doplnit diskové úložiště speciálně sdílené mezi servery. Realizace clusteru databází by tedy způsobila výrazný finanční nárůst celé investice jak v položce softwarové, tak i hardwarové. Proto bylo nakonec zvoleno řešení, kdy je v clusteru začleněna pouze www služba, přičemž snížení zátěže databáze je řešeno pomocí lokálních vyrovnávacích pamětí (lokální cache) nejvíce používaných dat, a zajištění případného hardwarového výpadku databázového serveru je realizováno čistě softwarově pomocí záložní databáze a automatické detekce výpadku.

Ze stávajícího serveru byla tedy přemístěna www služba na dva nové servery, které dohromady tvoří dva uzly www-clusteru. Protože na uzlech není uložena žádná informace primárně a výpadek jednoho z nich není kritický, bylo rozhodnuto, že hardwarem těchto serverů budou v podstatě obyčejné pracovní stanice: Intel Pentium 4 1,8 GHz, 1xIDE disk 20 GB, 384 MB RAM. Navíc jsou tyto servery osazeny dvojicí síťových karet, z nichž jedna je použita pro load balancing www služby a druhá je určena pro administrativní přístup a systémovou komunikaci služby load balancing. Operační systém na obou uzlech je Windows 2000 Advanced Server, na datovém serveru byl ponechán Windows 2000 Server. Programové vybavení se skládá z hotových produktů firmy Microsoft - tj. MSSQL 7.0, IIS 5.0, network load balancing, file replication service (dále FRS) a z vlastních modulů vytvořených pro účely www prezentace – na systémové úrovni se jedná o *lokální cache a databázovou proxy*. FRS bylo zvoleno pro účely automatické replikace ASP skriptů mezi uzly www-clusteru, přičemž jako „primární“ úložiště je dál používáno původní úložiště z dob jediného serveru. Ke konfiguraci www služby nedochází příliš často, a proto byla pro replikaci konfigurace zvolena neautomatická ruční varianta. Aby se zabránilo výpadku celého systému v případě výpadku databáze, byla na jeden uzel clusteru nainstalována záložní databáze, která se periodicky aktualizuje z primární databáze. Pro zajištění automatického přepnutí v případě výpadků vznikla *databázová proxy*, která si udržuje aktuální přehled o funkčních databázích. Jelikož jsou veškeré požadavky směrovány na TCP port této proxy, jsou dále s její pomocí přesměrovány do primární resp. záložní databáze. Pro odlehčení databázové zátěže byl vyvinut modul *lokální cache*, který při svém startu načte z databáze základní a většinou neměnné číselníky (pracovišť, adres, ...) a z nich případně vypočítá složitější struktury (např. hierarchii pracovišť). Tyto údaje periodicky z databáze aktualizuje.



Obrázek 2: Schéma cluster řešení

Celkové řešení (viz obrázek č.2) bylo již důkladně prověřeno – jednak testováním, jednak extrémními situacemi nastalými v reálném provozu (při havárii jednoho z disků primární databáze a jeho výměně běžel systém po automatickém přepojení po nutnou dobu na záložní databázi). Nezanedbatelným přínosem popsaného řešení je instalace záplat a oprav, které je možné provádět za plného provozu postupně – nejprve vypojením prvního uzlu a instalací oprav na něm, následně pak provedením téže akce na druhém uzlu.

4 Závěr

Jednotlivá parciální technologická řešení popsaná výše mají ještě některé drobné nedokonalosti. Například intranet www prezentace není možné používat při výpadku primární databáze z důvodu možných on-line zápisů. Aktualizace databáze není z pohledu celého systému plně transakční – v případě neúspěchu vygenerování stránek pro www.muni.cz zůstává stará verze

se včerejšími daty, ač dynamické stránky zveřejňují data nová. V současné době jsou však tyto nedokonalosti již zanedbatelné.

Současné řešení není konečné a o jeho dalším rozvoji budou převážně nepřímo rozhodovat návštěvníci www prezentace univerzity. Nedávno proběhlo také zdvojení síťového switche tak, aby v případě jeho výpadku mohly být požadavky od uživatelů na wwwdata.muni.cz vyřízeny alespoň jedním uzlem clusteru za použití, když ne jinak, alespoň záložní databáze. Díky již navržené robustní síťové topologii univerzity by mohl být každý switch napojen na páteřní linky jinou síťovou cestou. Paralelně s těmito úvahami se připravuje návrh pořízení nového databázového serveru, přičemž stávající by „osvobodil“ jeden z uzlů www clusteru od záložní databáze. Kromě záložní databáze by mohl tento „vysloužilý“ server v případě náhlého zvětšení návštěvnosti www stránek provozovat na dočasnou dobu třetí uzel www-clusteru. A konečně, na základě dosavadních pozitivních zkušeností zvažujeme variantu přesunout na www-cluster také službu www.muni.cz. □