Predikce zátěže v cloudu pomocí analýzy časových řad

Řešitel: Ing. Tomáš Vondra, FEL ČVUT

***Cloud computing*, a obzvláště privátní cloud, patří mezi rychle se rozvíjející odvětví výpočetní techniky. Cloud typu IaaS (*Infrastructure as a Service*) je nadstavbou nad virtualizací umožňující dynamické využití serverové infrastruktury.**

Výhodou oproti statické virtualizaci je úspora v případě, že počet serverů lze škálovat dle aktuální zátěže. Autoškálovače, kterých pro privátní nasazení není mnoho (kromě našeho ScaleGuru známe z lokálně nasaditelných jen Scalr), fungují reaktivně, tj. dokážou přidat kapacitu po překročení prahu nějaké monitorované veličiny. Pokud by autoškálovač obsahoval predikční mechanismus, který by dokázal zvýšení zátěže odhadnout dříve, než k překročení prahu dojde, bylo by možno zlepšit záruku kvality škálované služby.

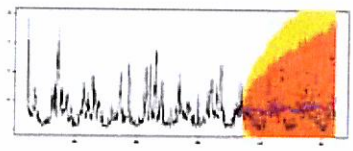
Problém, jak zodpovědět otázky typu „Kolik bude v cloudu volných slotů na virtuální stroje během noci?" nebo „Kolik procesorového času bude aplikace potřebovat příští hodinu s pravděpodobností 95 %?" řeší odvětví statistiky zvané analýza časových řad. Nabízí řadu algoritmů, které dokážou předikovat vývoj časové řady v budoucnosti, včetně konfidenčních intervalů. V zásadě jde o metody průměrovací, autoregresní nebo vyhledávání vzorů. Z oblasti strojového učení pak pocházejí markovovské modely a neuronové sítě.

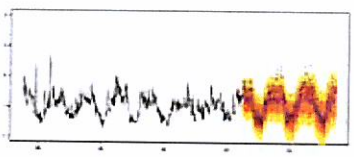
Cílem projektu bylo aplikovat metody predikce časových řad na průběhy zátěže z webových serverů a v případě jejich dostupnosti i z privátních cloudů. Data se podařilo získat pouze z malého webhostingu, přesto jsme otestovali dvě metody (Holt-Winters a ARIMA) na šesti dosti rozdílných časových řadách. Výsledky ukazují, že obzvláště metoda ARIMA je vhodná, ale jen pokud je server pod stálou zátěží. Pro nízké zátěže bude nutno modelovat příchody uživatelů pomocí teorie hromadné obsluhy.

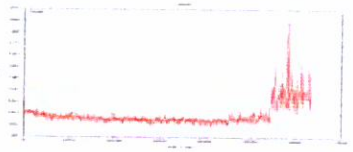
Byla také provedena rešerše podobných článků – z přibližného počtu 250 výsledků původního hledání do ní bylo zahrnuto 20 relevantních článků, žádný však neřešil stejný problém.

Od sponzora, firmy HP (Hewlett-Packard), jsme se bohužel nedozvěděli, kam by projekt měl dále směřovat. Byla vyžadována další simulace pro potvrzení výsledků. Obrátili jsme se proto k simulátoru CloudSim, který však bohužel neodpovídal požadavkům, takže bylo nutné ho výrazně upravit. Dosažené výsledky byly v rozporu s reálným zátěžovým testem. V tomto směru, tedy v simulaci autoškálovače, jsme se rozhodli pokračovat, a to metodami teorie hromadné obsluhy. Jak ukázala rešerše, jiné vhodné simulátory pro cloud neexistují.

Vyvíjený simulátor počítá z profilu zátěže a parametrů systému jeho zatížení. Je-li nad 100 %, zjistí stav při přidání dalších prostředků a aktualizuje stav systému pro další krok profilu. Jeho základem jsou volně dostupné knihovny pro řešení frontových sítí (předmět další rešerše).







Z dalších příbuzných projektů stojí za zmínku dokončený pokus s replikací databáze, který ukázal limity vhodnosti použití hybridního cloudu v porovnání se vzdáleným přístupem k datům a navrhl vzorce pro hodnocení dalších aplikací. Pracujeme také na škálování Hadoopu v cloudovém prostředí a na provisioningu aplikací řízeném výkonnostním modelem.

Výsledky predikce

Holt-Winters se fituje rychle, ale může být nepřesný: 20 % při šestihodinové předpovědi na třech vzorcích, jeden vzorek pod 50 % při třídenním horizontu, dva zcela selhaly.

ARIMA má mnoho parametrů, autotuning by bylo možno vylepšit: 20 % při šestihodinové předpovědi na třech až čtyřech vzorcích, pět vzorků pod 50 % při třídenním horizontu, jeden zcela selhal.

Výsledky simulace

Chyby vyšly extrémně vysoké: 7 %, 38 %, 79 %, 168 %, 217 %, 347 %, a 235 % – osm kroků škálování.