UNIVERSITATEA DIN CRAIOVA

FACULTATEA DE AUTOMATICĂ, CALCULATOARE ȘI ELECTRONICĂ

ȘCOALA DOCTORALĂ „CONSTANTIN BELEA”

DOMENIUL DE DOCTORAT: CALCULATOARE ȘI TEHNOLOGIA INFORMAȚIEI

REZUMATUL TEZEI DE DOCTORAT

OPTIMIZAREA PERFORMANȚEI SOFTWARE ÎN CALCULUL VIRTUAL

CONDUCĂTOR ȘTIINȚIFIC

Prof. Univ. Dr. Ing. Mihai MOCANU

DOCTORAND

Ing. Cătălina MANCAȘ

- 2018 –

CUPRINS

Introducere

Calculul de tip Cloud bazat pe soluții de virtualizare s-a dovedit, de-a lungul timpului, a rezolva problema utilizării ineficiente a resurselor fizice de calcul. Cercetarea demonstrează o reducere a costurilor resurselor necesare (resurse de tip hard, echipamente de climatizare, spațiu etc.), precum și o scădere a consumului de energie, atunci când serverele sunt virtualizate. Virtualizarea poate oferi aplicațiilor critice disponibilitate de nivel înalt și astfel, eficientiza operarea infrastructurii IT și răspunde rapid schimbărilor. De asemenea backup-ul rapid în caz de eroare reprezintă un beneficiu deosebit de important. Dezavantajul major, este, însă, scăderea performanței mașinilor virtuale în comparație cu cele fizice. Dar, cât de mare este raportul cost/performanță? Pentru multe organizații, migrarea spre un mediu virtual este, fără îndoială, prima alegere. Însă, o analiză a performanțelor calculului virtual este necesară.

Lucrarea de faţă reprezintă un demers ştiinţific în direcţia unificării unor metode, modele şi tehnologii într-o metodologie unitară, omogenă, capabilă să permită identificarea soluțiilor optimizate de mapare a unor aplicații complexe, pe o arhitectură hardware de calcul virtualizată, generică, puternic parametrizabilă, de tipul *Cloud Computing* (sistemele de tip Cloud fiind astăzi resurse larg răspândite de calcul eterogen de înaltă performanță), încă din faza de preproiectare a ansamblului hardware-software (co-design). Optimizările vizate au ca principal obiectiv folosirea cât mai eficientă a resurselor disponibile, în scopul obținerii unei performanțe maxime, în condiții restrictive de energie consumată, complexitate etc., fiind tratate ca probleme multicriteriale complexe.

Calculul de tip Cloud a reprezentat obiectul a numeroase cercetări ce, de-a lungul timpului, au demonstrat rezultate precum: reducerea costurilor hardware, reducerea consumului de energie și utilizarea eficientă a mașinilor fizice. Având în vedere utilizarea pe scară tot mai largă a serviciilor de tip Cloud bazate pe virtualizare, determinată în primul rând de considerente economice, continuarea cercetării în domeniul virtualizării în vederea îmbunătățirii performanțelor sistemelor Cloud și a reducerii costurilor pentru utilizatorii de servicii Cloud este o activitate deosebit de utilă. Rezultatele obținute dovedesc că se pot aduce multe îmbunătățiri sistemelor de calcul bazate pe virtualizare și că domeniul calculului de tip Cloud nu este doar de mare actualitate, dar și de viitor.

Astfel, în această lucrare sunt studiate diverse tehnici și tehnologii de virtualizare, folosind soluții de tip Cloud și se propun soluții de optimizare a performanțelor. Au fost evaluate performanțele diverselor tipuri de mașini virtuale făcându-se recomandări în funcție de specificul aplicațiilor utilizatorilor de servicii Cloud. A fost de asemenea realizată a analiză comparativă a Cloud-ului privat vs Cloud-ul public cu scopul recomandării soluției adecvate în funcție de cerințele prioritare ale utilizatorilor.

Plecând de la teoria rețelelor de cozi de așteptare Jackson, lucrarea propune un model matematic menit să asigure satisfacerea unui nivel impus de calitate a serviciilor (QoS). Au fost de asemenea propuse o politică de alocare a mașinilor virtuale la mașinile fizice, cu scopul reducerii consumului de energie electrică și o politică de alocare a sarcinilor de lucru la mașinile virtuale disponibile, cu scopul reducerii timpului de răspuns al sistemului. Experimentele prezentate se bazează pe o analiză riguroasă a literaturii și au drept scop, validarea modelelor matematice propuse și a politicilor de planificare elaborate având ca obiectiv reducerea timpului de răspuns și a costului prin scăderea consumului de energie electrică.

Un alt aspect tratat în prezenta lucrare îl reprezintă evaluarea utilizării resurselor de calcul într-un centru de calcul ce oferă servicii de tip Cloud și a impactului scalabilității asupra timpului de răspuns al sistemului. Concluziile desprinse în urma experimentelor realizate vin în sprijinul utilizatorilor în alegerea soluțiilor care să satisfacă un anumit nivel de QoS impus dar cu costuri minime.

Performanța sistemului Cloud are un impact semnificativ asupra performanței viitoarei infrastructuri de calcul. Evaluarea amănunțită a performanței serviciilor Cloud este crucială și benefică atât pentru

furnizorii de servicii, cât și pentru consumatori, formând astfel un domeniu activ de cercetare. Unele tehnologii cheie pentru Cloud computing, cum ar fi virtualizarea și arhitectura orientată pe servicii (SOA), aduc provocări evaluării performanței serviciilor. Până în prezent, comunitatea de cercetare a investit eforturi substanțiale pentru a aborda aceste provocări, iar rezultatul a constat într-un progres remarcabil. Printre lucrările de analiză a performanțelor Cloud, abordările de evaluare dezvoltate prin perspectiva modelării sistemelor joacă un rol important. Cu toate acestea, lucrările recente au vizat diferite secțiuni ale literaturii, lipsind astfel o imagine de ansamblu care să prezinte cel mai recent stadiu al acestui domeniu. Scopul acestei lucrări este de a contribui la formarea imaginii de ansamblu a stadiului actual al evaluării performanței în sisteme virtualizate prin rezultatele obținute în urma unor experimente de evaluare a performanței. Această lucrare analizează, de asemenea, problemele deschise și provocările abordărilor de evaluare realizate și identifică posibile oportunități de cercetare viitoare în acest important domeniu.

Virtualizarea a fost revoluţionată de apariţia Calculul de tip Cloud, în primul rând, prin furnizarea de servicii la cerere. Dacă un client are nevoie de putere de calcul pentru o perioadă de timp limitată, acesta o poate cumpăra pentru acea perioadă în Cloud, fără să fie nevoie să achiziționeze resursele fizice necesare pentru furnizarea puterii de calcul solicitate. Prin achizitionarea de servicii Cloud, clientul are acces la o gamă largă de resurse partajate, fără să aibă cunoștințe despre localizarea fizică a acestora. Resursele furnizate în Cloud sunt scalabile, imediat. Astfel, clientul poate să aleagă, pe loc, dacă dorește să crească sau să scadă resursele de calcul. Această opțiune introduce proprietatea de modularitate în Cloud. În plus, resursele Cloud-ului sunt controlabile, măsurabile și optimizabile. Optimizarea utilizării mașinilor fizice este facilitată de virtualizare, care a câștigat popularitate odată cu apariția procesorului cu nuclee multiple și suport pentru virtualizare.

Calculul de tip Cloud a reprezentat obiectul a numeroase cercetări ce, de-a lungul timpului, au demonstrat rezultate precum: reducerea costurilor hardware, reducerea consumului de energie și utilizarea eficientă a mașinilor fizice. Literatura de specialitate atestă că adoptarea de soluții de calcul de tip Cloud se produce, în primul rând din motive economice (Wang & Varela, 2011) (Indrani, Sudhakar, & Lizy, 2012), (Kundu, Rangaswami, Dutta, & Zhao, 2010), (Zhang, și alții, 2013). Calculul de tip Cloud permite clienților să reducă costurile cu resursele de calcul, precum și consumul de energie electrică deoarece clienții plătesc doar pentru resursele pe care le folosesc. O altă concluzie care se desprinde este aceea că multe mașini fizice sunt utilizate ineficient pentru că sunt insuficient încărcate. Utilizarea calculului de tip Cloud în asociere cu virtualizarea reprezintă o soluție adusă problemei de subutilizare a serverelor și, de asemenea, centrele de date de astăzi se îndreaptă rapid către utilizarea virtualizării serverului ca mod preferat de partajare a resurselor hardware ale unui server între multiple instanțe virtuale care găzduiesc diferite aplicații (Abirami & Shalini, 2012), (Chang, Tsai, & Chen, 2013). Cu toate acestea, performanțele virtualizării în calculul de tip Cloud, nu pot egala performanțele native.

Există mai multe tipuri de Cloud-uri care pot fi utilizate pentru a îndeplini nevoi diferite. Dintre acestea, Cloud-ul public - administrat de un furnizor de servicii accesibile clienților prin intermediul Internetului, și Cloud-ul privat – în interiorul rețelei unei companii, sunt cele mai populare și servesc drept medii pentru experimentele din această lucrare.

Calculul de tip Cloud este o tehnologie recentă, dar din ce în ce mai populară deoarece din ce în ce mai multe companii încep să o utilizeze. Din acest motiv, în această direcție este investită o activitate de cercetare intensă. De asemenea, relația dintre calculul de tip Cloud și virtualizare reprezintă un subiect curent de cercetare orientat pe îmbunătățirea performanțelor de virtualizare în calculul de tip Cloud (Chang, Tsai, & Chen, 2013), (Diogo & Ferraz, 2012), (Indrani, Sudhakar, & Lizy, 2012). Obiectivul studiilor în domeniu, așa cum reiese din cercetările curente, este acela de a obține performanțe în mediul virtualizat cât mai apropiate de performanțele native.

Virtualizarea serverului se ocupă de mascarea resurselor fizice ale serverului, inclusiv numărul și identitatea de servere fizice individuale, procesoare și sisteme de operare. Administratorul serverului utilizează un software specific pentru a împărți un server fizic în mai multe medii virtuale izolate. Aceste medii virtuale sunt, de asemenea, cunoscute sub denumirile de oaspeți, instanțe, containere sau emulări. Virtualizarea este o metodă prin care pot fi rulate mai multe sisteme de operare pe o singură mașină fizică. Există trei abordări populare de virtualizare a serverului: virtualizarea completă, virtualizarea sistemului de operare și virtualizarea hardware-ului.

Datorită popularității Cloud Computing, au apărut pe piață numeroase soluții de virtualizare a serverului, printre care lideri de piață sunt VMware, KVM, Microsoft, Xen sau Oracle. Acestea manifestă rezultate diferite privind performanța, astfel că, un subiect curent de interes este identificarea hipervizorului recomandat în funcție de specificul aplicațiilor utilizatorului (Graniszewski & Arciszewski, 2016). Acest subiect a servit drept punct de pornire pentru unele analize de performanță din această lucrare. Astfel, lucrarea prezintă un studiu comparativ al performanțelor hipervizorilor VMware vSphere, Microsoft Hyper-V și KVM , în ceea ce privește gradul de utilizare a procesorului, memoriei și dispozitivelor I/O.

Utilizarea virtualizării introduce o degradare a performanțelor native, afectând gradul de utilizare a procesorului, a rețelei, a memoriei și a spațiului de stocare sau dispozitivelor I/O (Johnsen, Schlatte R., & Tapia, 2012). Obiectul de studiu al acestei lucrări îl reprezintă evaluarea performanțelor virtualizării serverului în calculul de tip Cloud, iar unul dintre obiectivele lucrării este acela de a recomanda tipul de mașină virtuală care afectează cel mai puțin performanțele native, în funcție de specificul aplicației utilizatorului. Pentru aceasta, au fost utilizate mai multe instrumente de măsurare a performanței, cu diverse metrici. Printre instrumentele alese pentru măsurarea performanțelor se numără: UNIXBench (Yamato, 2017), IOzone, (Tarasov, Bhanage, Zadok, & Seltzer, 2011) sau RAMSpeed[1].

Orice sistem de calcul de tip Cloud este alcătuit din două componente numite Front-End și Back-End. Front-End-ul reprezintă componenta accesibilă clientului de servicii Cloud. Aceasta cuprinde aplicația folosită pentru a accesa Cloud-ul printr-o interfață cum ar fi browser-ul web și prin rețeaua accesibilă clientului. Back-End-ul reprezintă componenta din spate, ascunsă clientului și cuprinde diverse servere, dispozitive de stocare, echipamente de rețea etc. Teoretic, un Cloud încorporează toate tipurile de servere dedicate să găzduiască, practic, orice tip de aplicații. Un server central administrează sistemul, monitorizează sistemul, traficul și cererile clientului, pentru a asigura orice funcționalitate solicitată. În plus, există seturi de reguli și middleware care ajută la conectarea în rețea a resurselor de calcul care necesită comunicare între ele. Lucrarea propune un model matematic pentru Back-End-ul unui sistem Cloud, implementat și testat prin simulare discretă bazată pe evenimente.

Existența mașinilor virtuale se reflectă asupra performanțelor sistemului Cloud, deoarece rularea mașinilor virtuale crește numărul schimbărilor de context, consumând considerabil puterea procesorului pentru translatarea adreselor. Monitorizarea gradului de încărcare a procesorului vizează îmbunătățirea performanței și reprezintă o acțiune utilă în investigațiile de performanță deoarece gradul de încărcare a procesorului prezintă implicații critice asupra altor metrici de performanță a sistemului, cum ar fi consumul de energie electrică.

Din punct de vedere al managementului virtualizării în Cloud, hipervizorii serverelor virtualizate utilizează diverse politici de planificare pentru a gestiona funcționarea eficientă a instanțelor virtuale. Performanțele alicației ce rulează într-un mediu virtual sunt direct influențate de parametrii politicii de planificare utilizate. Astfel că, variind parametrii politicii de planificare poate fi influențată performanța aplicației. De exemplu, autorii (Chang, Tsai, & Chen, 2013) au demonstrat că performanțele I/O pot fi îmbunătățite prin modul de planificare a sarcinilor mașinii virtuale. Astfel, una dintre contribuțiile acestei lucrări o reprezintă propunerea unei politici de planificare, structurată pe două niveluri, astfel încât să

[1] RAMSpeed, a cache and memory benchmarking tool (http://alasir.com/software/ramspeed/)

încorporeze atât aspectul de alocare a mașinilor virtuale la cele fizice, dar și cel de alocare a mașinilor virtuale la sarcinile de lucru. Politica propusă a fost evaluată prin simulare discretă, bazată pe evenimente, iar rezultatele au fost comparate cu cele ale unei politici populare de planificare – cea bazată pe algoritmul Round-robin.

Într-un scenariu din viața reală, resursele de calcul din Cloud sunt împărțite între clienții care plătesc pentru servicii în funcție de utilizarea resurselor. În general, prin virtualizare, detaliile resurselor de calcul devin transparente pentru clienți. Observate din perspectiva utilizatorului, serviciile sunt identice în ceea ce privește funcționalitatea și interfața. Însă modul de alocare a resurselor pentru a îndeplini sarcinile de lucru solicitate de diverși utilizatori ai sistemului, dar și scalabilitatea sistemului Cloud reprezintă factori cruciali pentru optimizarea timpului de răspuns al sistemului (Vinothina, Sridaran, & Ganapathi, 2012), (Falatah & Batarfi, 2014), (Goswami & Saha, 2013). Dacă alocarea resurselor nu este gestionată eficient, gradul de deservire al sistemului este afectat, implicit nivelul QoS, un factor important în evaluarea perfomanțelor sistemelor virtualizate. Lucrarea studiază modul în care timpul de răspuns este influențat de alocarea resurselor prin scalarea sistemului.

Pentru evaluarea performanțelor diferitelor metode de virtualizare încă din faza de proiectare, este nevoie de un model de ansamblu (hardware-software) care să permită estimarea acestor performanțe, fără a apela la codul aplicației și la deciziile concrete de alegere a arhitecturii hardware, considerat doar ca un concept (arhitectură Cloud generică, parametrizată). Modelul trebuie să reprezinte ansamblul în mod uniform, astfel încât să poată fi descris prin asocierea componentelor concurente ale aplicației cu nodurile elementare sau agregate de execuție de la nivelul componentelor Cloud-ului și totodată să poată evidenția în ce măsură maparea exploatează anumite facilități cum ar fi paralelismul hardware existent, sau restricții precum cele jurisdicționale.

Prezentarea sintetică a capitolelor tezei de doctorat

Teza de doctorat este structurată în șapte capitole. În capitolul introductiv este prezentată motivația din spatele lucrării și stadiul domeniului, obiectivele și structura și organizarea tezei. Plecând de la stadiul actual al cercetării în domeniu și de la nevoile pieței calculului de tip Cloud, au fost identificate direcțiile de cercetare considerate utile la creșterea nivelului cunoașterii în domeniu. Prin obiectivele propuse, se dorește sprijinirea utilizatorilor de servicii Cloud, ușurându-le procesul de alegere a tipului de serviciu ce se pliază cel mai bine pe specificul aplicațiilor lor, iar pe de altă parte propunerea de soluții menite să contribuie la eficientizarea utilizării resurselor și la scăderea costurilor. Apariția pe piață a diverselor tipuri de servicii Cloud, Cloud public, Cloud privat sau Cloud hibrid, reprezintă motivația de a veni în sprijinul numărului tot mai mare de utilizatori în alegerea tipului de serviciu ce se potrivește cel mai bine cerințelor acestora.

Virtualizarea a reprezentat un subiect de interes constant în activitatea de cercetare dezvoltată în ultimii ani (Mancaș C. , 2015), (Mancaș C. , 2014), (Mancaș, Mocanu, & Mancaș, 2013). Rezultatele activității de cercetare au demonstrat că performanțele mediului virtualizat, respectiv ale sistemului Cloud depind semnificativ de modul de planificare a sarcinilor și de volumul de lucru al sistemului. Acest argument, împreună cu dorința de a găsi soluții pentru reducerea timpului de răspuns al sistemului și a consumului de energie electrică, au reprezentat una dintre motivațiile pentru continuarea cerecetărilor în această direcție. De asemenea, necesitatea impusă pe piața serviciilor Cloud de satisfacere a unui anumit nivel de calitate a serviciilor oferite, reprezintă motivația din spatele activității de cercetare în vederea modelării sistemelor Cloud bazate pe virtualizare cu scopul de a putea oferi nivelul solicitat de calitate a serviciilor Cloud, la un preț de cost minim.

Această lucrare își propune o evaluare a performanțelor calculului virtual, cu precădere într-un mediu de tip Cloud, cu scopul de a propune noi soluții de optimizare a calculului virtual în vederea creșterii performanțelor și reducerii costurilor. Lucrarea evaluează performanțele soluțiilor existente pe piață și formulează recomandări utile utilizatorilor de servicii Cloud în vederea alegerii soluțiilor optime din punct de vedere al specificului aplicațiilor lor dar propune și soluții noi ce au în vedere reducerea timpului de răspuns al sistemului sau reducerea consumului de energie electrică. Experimentele cuprinse în evaluările de performanță sunt realizate atât în medii reale de testare, cât și prin simulare.

Obiectivul principal al lucrării: *Optimizarea performanțelor în calculul virtual* este descompus în subobiective.

1. Formularea de recomandări privind utilizarea hipervizorilor KVM, VMware și Microsoft Hyper-V, în funcție de cerințele și specificul aplicațiilor utilizatorului. În acest scop au fost evaluate performanțelor mașinilor virtuale KVM, VMware și Hyper-V în diverse scenarii de test.

2. Formularea de recomandări cu privire la tipul de serviciu Cloud care se recomandă a fi adoptat în funcție de specificul cerințelor impuse de utilizatori. Pentru îndeplinirea acestui obiectiv s-a realizat o analiză comparativă între un Cloud-ul public și unul privat în ceea ce privește performanțele resurselor de calcul: procesor, memorie și sistem I/O.

3. Reducerea costurilor prin reducerea consumului de energie electrică.
   În acest sens, a fost elaborată și implementată o politică de planificare a procesului de mapare al mașinilor virtuale la mașinile fizice.

4. Reducerea timpului de execuție total al aplicațiilor ce rulează în cadrul infrastructurii Cloud.

5. Reducerea timpului de execuție a sarcinilor de lucru paralele, cu scopul de a garanta nivelul QoS necesar. Astfel, a fost propusă o politică de planificare a procesului de mapare a saricinilor la mașinile virtuale, în vederea reducerii timpului de execuție a sarcinilor paralele și creșterii performanței globale a sistemului de calcul Cloud.

6. Îmbunătățirea performanțelor unui sistem Cloud prin analizarea impactului utilizării resurselor de calcul și a scalabilității infrastructurii Cloud asupra timpului de răspuns total al sistemului.

Al doilea capitol al tezei *Tehnologii și soluții curente de virtualizare* prezintă noțiuni generale despre virtualizare, pornind de la primele forme de virtualizare și oferind o vedere retrospectivă asupra conceptului de virtualizare. Subcapitolele capitolului 2 cuprind noțiuni teoretice despre: *Virtualizarea procesorului,* cu precădere *Virtualizarea procesorului x86* și limitări ale acestuia, *Virtualizarea memoriei, Virtualizarea dispozitivelor I/O, Tehnici de virtualizare*, *Implicațiile tipurilor de virtualizare asupra performanței*, Tehnologii de virtualizarea dintre care de interes este *Virtualizarea serverului* și *Soluții curente de virtualizare a serverului.*

Capitolul prezintă terminologia virtualizării, fiind explicate noțiuni precum *mașină fizică, mașină virtuală, monitor al mașinii virtuale (VMM), virtualizarea serverului* sau *resurse de calcul*. Coborând un nivel de detaliu mai jos, tot în Capitolul 2 este abordată **Error! Reference source not found.** și detaliat modul în care este realizată virtualizarea procesorului, prezentând caracteristicile dar și limitările virtualizării procesorului x86 de către VMware. VMware marchează istoria virtualizării cu lansarea primei soluții de virtualizare a procesorului, însă limitările familiei de procesoare x86, au dat naștere mai multor tehnici de virtualizare: *virtulizarea completă, paravirtulizarea* sau *virtualizarea asistată de sistemul de operare (OS)* și *virtualizarea asistată de hardware*. Prin virtualizarea procesorului este introdus un nivel de abstractizare cheie în domeniul virtualizării, care poartă denumirea de *hipervizor.* Pentru a rula multiple

mașini virtuale pe un singur sistem fizic, *virtualizarea procesorului*, nu este suficientă. În plus, este necesară introducerea altor niveluri de virtualizare: *virtualizarea memoriei* și *virtualizarea dispozitivelor I/O*. Componenta cheie a virtualizării memoriei este unitatea de administrare a memoriei (Memory Management Unit), responsabilă pentru maparea memoriei fizice virtuale la memoria fizică, pe când virtualizarea dispozitivelor I/O devine posibilă prin intermediul hipervizorului.

Performanțele virtualizării procesorului, memoriei și dispozitivelor I/O constituie obiectul de studiu al acestei lucrări. În experimentele de evaluare a performanței, virtualizarea resurselor de calcul a fost realizată folosind soluțiile de virtualizare: *VMware vSphere*, *Microsoft Hyper-V* și *KVM*. Aceste soluții au fost alese în urma unui studiu al pieței și al literaturii ce a condus la motivația de a realiza experimente noi de evaluare a performanței. Rezultatele studiului și situația curentă a celor trei soluții de virtualizare, sunt descrise tot în Capitolul 2, iar rezultatele acestui studiu reprezintă un punct de pornire necesar în evaluările de performanță prezentate în Capitolul **Error! Reference source not found.** al tezei.

În Capitolul **Error! Reference source not found.** ***Error! Reference source not found.****: modele și arhitecturi de referință,* focusul se mută de la general la specific, de la conceptul de virtualizare la componenta sa: calculul de tip Cloud. Și aici, se dorește formularea de definiții și revizuirea unei terminologii. Descrierea cea mai recunoscută a calculului de tip Cloud o reprezintă standardizarea conceptului ISO/IEC 17788:2014 (ISO-International Standard for Organizations, 2014). Primele eforturi în standardizarea calculului de tip Cloud au fost, însă, realizate de NIST (Institutul Național de Standarde și Tehnologii al Guvernului S.U.A., (Mell & Grance, 2011). Acest model descris de NIST întrunește cinci caracteristici esențiale, trei modele de servicii și patru modele de dezvoltare. Cele cinci caracteristici sunt descrise în subcapitolul ***Error! Reference source not found.***. Modelele de servicii formulate de NIST sunt descrise în subcapitolul ***Error! Reference source not found.***. Acestea se regăsesc în standardul ISO/IEC 17788:2014 sub denumirea de tipuri de funcționalități. Totuși, standardul alege să încadreze cele trei modele propuse de NIST într-o altă clasificare pe care o numește categorii de servicii Cloud; o categorie poate cuprinde funcționalități din toate cele trei tipuri menționate mai sus, deci standardul delimitează clar categoriile de servicii Cloud, de tipurile de funcționalități: Infrastructure as a Service (IaaS), Network as a Service (NaaS), Platform as a Service (PaaS), Software as a Service (SaaS), etc. În plus, sunt previzionate noi categorii de servicii ce urmează a fi standardizate în 2019, aflate momentan în stadiul de categorii informale. Aceste categorii informale pot servi drept direcții viitoare de cercetare.

Subcapitolul ***Error! Reference source not found.*** tratează modelele standardizate de Cloud și anume: Cloud-ul public, Cloud-ul privat, Cloud-ul hibrid și Cloud-ul comunitate. Dintre acestea, analiza de performanță evaluează primele două modele: Cloud-ul public și Cloud-ul privat. Nu, în ultimul rând, în Capitolul 3, sunt enumerate aspectele transversale ale calculului de tip Cloud standardizate, dintre care performanța este aspectul principal tratat în această lucrare. Pentru a realiza o analiză de performanță riguroasă, este necesar să se definească parametrii ce urmează a fi măsurați și evaluați.

O definiție completă a calculului de tip Cloud este reflectată de arhitectura de referință a calculul de tip Cloud (en. Cloud Computing Reference Architecture - CCRA), descrisă în subcapitolul ***Error! Reference source not found.***. Acest model arhitectural întrunește descrierea rolurilor și a sub-rolurilor, activităților, aspectelor transversale și componentelor funcționale ale calculului de tip Cloud, precum și a relațiilor dintre acestea. CCRA reprezintă o unealtă de referință în comunitatea internațională Cloud Computing, punând accent pe ceea ce furnizează serviciile Cloud. Din punct de vedere arhitectural, sistemele Cloud pot fi privite și descrise din mai multe unghiuri. În această lucrere arhitectura Cloud de referință este descrisă din punctul de vedere al utilizatorului și din punct de vedere funcțional.

CCRA a condus la dezvoltarea de modele arhitecturale ale celor două infrastructuri Cloud: public și privat, utlizate în experimentele de evaluare a performanțelor. Cele două modele arhitecturale sunt prezentate în subcapitolul ***Error! Reference source not found.****ă*.

Pornind de la modelul arhitectural al Cloudului privat, a fost proiectată modelarea matematică a acestei infrastructuri, în vederea realizării mai multor experimente de evaluare a performanțelor. Modelarea matematică se bazează pe Teoria cozilor de așteptare și pe Teoria rețelelor deschise Jackson. Fundamentul matematic necesar generării unui model matematic se regăsește și modelele propuse sunt descrise în capitolul 4. Modelarea matematică prezintă o variantă de bază, dar și o extensie virtualizată, ambele validate prin simulare bazată pe eveniment.

Capitolul **Error! Reference source not found.** ***Error! Reference source not found.***, descrie mediul în care s-au desfășurat experimentele de evaluare a performanțelor, realizate cu scopul de a identifica scenariile reale în care un sistem Cloud manifestă rezultate de performanță ridicată. Cel mai concludent indicator de performanță îl reprezintă timpul de răspuns al sistemului. În vederea reducerii timpului de răspuns total, dar și al execuției sarcinilor paralele, au fost efectuate mai multe experimente. De asemenea, un alt obiectiv vizat îl reprezintă reducerea consumului de energie. Pentru a îndeplini aceste obiective, au fost propuse două politici de planificare. Prima politică propusă se referă la planificarea procesului de mapare a mașinilor virtuale la mașinile fizice și are ca obiectiv reducerea consumului de energie al sistemului. Această politică este descrisă în secțiunea *Politica de planificare de nivel înalt* și evaluată în secțiunea *Evaluarea performanțelor politicii de alocare a mașinilor virtuale la mașinile fizice*. **Error! Reference source not found.** Cea de-a doua politică vizează reducerea timpului de răspuns total al sistemului, dar și a timpului de execuție a sarcinilor paralele și propune o politică de alocare al sarcinilor la mașinilor virtuale. Aceasta este descrisă în secțiunea ***Error! Reference source not found.*** și evaluată în secțiunea ***Error! Reference source not found.***. Rezultatele și concluziile care se desprind din evaluările celor două politici propuse pot fi consultate în secțiunea ***Error! Reference source not found.***.

Datorită expansiunii adoptării de soluții de vitualizare s-a dovedit necesară o analiză a soluțiilor de virtualizare din categoria liderilor de pe piață, în vederea formulării de recomandări cu privire la performanțele soluțiilor în funcție de specificul aplicației utilizatorului. Astfel, au fost evaluate performanțele hipervizorilor KVM, VMware și Hyper-V, din punct de vedere al comportamentului procesorului, al operațiilor asupra memoriei și al operațiilor asupra sistemului de fișiere. Experimentul a fost realizat într-un Cloud privat, utilizând diferite aplicații open-source pentru măsurarea unei game variate de metrici cum ar fi: rata de Input/Output (I/O), viteză la scriere/citire, latența, creare/ștergere fișiere, viteză de scriere/citire memorie, viteză CPU (Unitatea Aritmetico-Logică și Unitatea în virgulă mobilă) etc. Indicatorii de performanță, uneltele utilizate, precum și rezultatele experimentului sunt descrise în secțiunea ***Error! Reference source not found.***. Rezultatele obținute sunt descrise în secțiunea ***Error! Reference source not found.***, ***Error! Reference source not found.*** și ***Error! Reference source not found.***. Concluziile și recomandările formulate în urma rezultatelor obținute, precum și direcțiile de cercetare ce se desprind în urma studiului de performanță se regăsesc în capitolul **Error! Reference source not found.**, și urmăresc îndeplinirea obiectivelor formulate în capitolul 1.

Un alt subiect de interes al lucrării a fost identificarea modelului de implementare a Cloud-ului (secțiunea ***Error! Reference source not found.***) recomandat a fi adoptat în funcție de cerințele utilizatorului. În acest sens, experimentul realizat a comparat performanțele obținute într-un Cloud privat cu cele obținute într-un Cloud public (secțiunea *Evaluarea performanțelor în Cloud-ul public vs. Cloud-ul privat*). Performanța a fost evaluată din punctul de vedere al comportamentului procesorului, memoriei și al sistemului I/O, iar furnizorul de Cloud public a fost Rackspace. Cele două infrastructurii sunt descrise în secțiunea introductivă a capitolului **Error! Reference source not found.**, în paragrafele ***Error! Reference source not found.Error! Reference source not found.*** și ***Error! Reference source not found.***.

Timpul de răspuns al sistemului este, de asemenea, analizat din punctul de vedere al utilizării resurselor, dar și al scalabilității centrelor de date. Modul în care a fost realizată această analiză de performanță, dar și rezultatele acesteia sunt descrise în secțiunea ***Error! Reference source not found.***.

Capitolul final, **Error! Reference source not found.** încorporează rezultatele și concluziile tuturor experimentelor de evaluare descrise în capitolul 6. Acestea vizează obiectivele identificate în capitolul introductiv, dar servesc și drept punct de pornire pentru direcții de cercetare viitoare (capitolul **Error! Reference source not found.**).

Contribuțiile și rezultatele tezei de doctorat

O primă contribuție se referă la formularea de recomandări cu privire la adoptarea unei soluții de virtualizare. Trei dintre cele mai populare soluții curente de virtualizare au fost supuse unor evaluări ale mai multor criterii de performanță, ale căror rezultate au menirea de a veni în întâmpinarea nevoilor utilizatorilor.

Apariția pe piață a diverselor tipuri de servicii Cloud, Cloud public, Cloud privat sau Cloud hibrid, a reprezentat motivația de a veni în sprijinul numărului tot mai mare de utilizatori în alegerea tipului de serviciu ce se potrivește cel mai bine cerințelor lor. Astfel, numeroasele experimente efectuate atât în sisteme Cloud publice, cât și private, rezultă în recomandări care vin în sprijinul alegerii pe care trebuie să o facă utilizatorii atunci când decid să apeleze la servicii de tip Cloud.

Pornind de la modelul arhitectural al Cloudului privat, a fost proiectată modelarea matematică a acestei infrastructuri, în vederea realizării mai multor experimente de evaluare a performanțelor. Modelarea matematică se bazează pe Teoria cozilor de așteptare și pe Teoria rețelelor cozilor de așteptare sau Teoria rețelelor deschise Jackson.

Din necesitatea de a extinde funcționalitatea modelelor matematice propuse, a fost propusă o politică de planificare bazată pe evenimente, structurată pe două niveluri: nivelul mașinilor fizice și cel al mașinilor virtuale. Sarcinile sunt atribuite mașinilor virtuale, care, la rândul lor sunt mapate unui cluster de resurse computaționale computerizate (procesoare, memorie principală, memorie secundară și lățimea de bandă a rețelei). Scopul politicii propuse este de a mapa sarcinile spre procesare, nodurilor, în funcție de puterea lor de procesare și de disponibilitate.

În virtualizare performanțele depind semnificativ de modul de planificare a sarcinilor și de volumul de lucru al sistemului. Acest argument, împreună cu dorința de a găsi soluții pentru reducerea timpului de răspuns al sistemului și a consumului de energie electrică, au reprezentat una dintre motivațiile pentru continuarea cerecetărilor în această direcție.

Pentru testele de performanță descrise, au fost folosite diverse medii de testare bazate atât pe echipamente hardware cât și pe simulare. Mediile de simulare utilizate sunt SAGEMath, CloudSim și Eucalyptus. Din punct de vedere hardware, mediul de testare este constituit din două infrastructuri Cloud: o infrastructură Cloud privată și una publică Rackspace.

Evaluarea de performanță a hipervizorilor KVM, VMware și Hyper-V în Cloud privat

Performanțele unui sistem virtualizat sunt descrise de comportamentul resurselor de calcul virtualizate. Pentru a studia acest comportament este necesară măsurarea anumitor parametrii, metrici sau indicatori. Evaluările au fost efectuate pe mașini virtuale KVM, VMware și Hyper-V în următoarea configurație: 2 core-uri (4 thread-uri), 8 GB RAM, 1 TB capacitate de stocare. Mașinile virtuale au fost create pe o mașină fizică din infrastructura Cloud privată, având următoarea configurație: CPU: 2x Intel XEON e5 2690v2 la 3 GHz, 20 core-uri (40 de thread-uri), 256 GB RAM, 18 TB capacitate de stocare. Evaluarea performanțelor s-a realizat pe sistem de operare Linux, folosind instrumentul UnixBench. UnixBench poate fi utilizat pentru evaluarea performanțelor atât atunci când se rulează o singură sarcină sau mai multe sarcini, cât și pentru prelucrare paralelă. Metricile UnixBench evaluate în experimentul curent sunt:

1. *Dhrystone:* este o metrică folosită pentru a măsura și compara performanța în condițiile aplicațiilor ce nu utilizează operații în virgulă mobilă.
2. *Whetstone:* este o metrică sintetică ce măsoară performanța în termeni de viteză și eficiență pentru operații în virgulă mobilă.
3. *Execl Throughput:* măsoară numărul de apeluri de tip ”execl” pe secundă. Rezultatele testelor conduc la concluzia că deși performanțele celor trei tipuri de mașini virtuale sunt apropiate, VMware manifestă rezultate mai bune decât Hyper-V și KVM.

Testele UnixBench realizate cu scopul de a evalua operația de manipulare a proceselor au fost:

1. *Pipe Throughput*: măsoară numărul instanțelor pe secundă în care un proces poate scrie și citi 512 octeți.
2. *Creare proces:* măsoară de câte ori, în unitatea de timp, un proces poate crea și termina sub-procese.
3. *Script-uri Shell*: măsoară numărul de instanțe pe minut, în care un proces poate manipula o serie de opt copii concurente ale unui script Shell care aplică o serie de transformări asupra unui fișier de date.
4. *Supraîncărcarea la apeluri sistem*: estimează costul intrărilor și ieșirilor în și dinspre kernel-ul sistemului de operare, și anume costul realizării unui apel de sistem.

Rezutatele testelor reflectă faptul că mașina virtuală Hyper-V prezintă performanțe ceva mai slabe decât competitorii VMware și KVM. Acest lucru ce ar putea fi explicat prin rularea Hyper-V pe sistem de operare Linux (cu driverele aferente).

În urma testului Pipe Troughput, Creare Proces și Script-uri Shell, VMware se detașează semnificativ de contra-candidații săi. VMware este urmat de KVM, apoi de Hyper-V. Ultimii doi manifestă rezultate similare în cazul Creare Proces. În scenariul testului asupra comportamentului apelurilor de sistem, VMware depășește cu puțin performanțele KVM, însă Hyper-V rămâne detașat pe ultimul loc.

Testele de evaluare a memoriei au avut drept obiectiv măsurarea vitezei operațiilor de citire și de scriere efectuate asupra memoriei. De interes sunt operațiile de citire, respectiv scriere a blocurilor de date de tip întreg, și operațiile de cititre/scriere a blocurilor de date de tip virgulă mobilă. Rezultatele arată faptul că în cazul citirii din RAM a blocurilor de date de tip întreg performanțele tuturor celor trei mașini virtuale scad odată cu creșterea dimensiunii blocului de date. Comparând performanțele celor trei tipuri de mașini virtuale se observă că acestea sunt similare. Totuși, în cazul blocurilor de dimensiune mică (16, 256, 1024 octeți), se observă un același șablon de clasificare, VMware demonstrând cea mai mică viteză, urmat de KVM, apoi de Hyper-V. În cazul citirii din RAM a blocurilor de date de tip virgulă mobilă, Hyper-V se comportă, din nou, cel mai bine în cazul dimensiunilor mici ale blocurilor de date. Atunci când dimensiunea blocului de date crește mai mult de 65536 octeți, cea mai mare viteză o manifestă KVM. Rezultatele testelor arată că Hyper-V este mai performant decât celelalte două tipuri de mașini virtuale în cazul scrierii în memorie a blocurilor de date de tip întreg. În această situație, VMware manifestă vizibil cel mai slab comportament. În cazul scrierii de blocuri de date de tip virgulă mobilă evaluarea demonstrează performanțe aproximativ similare pentru cele trei tipuri de mașini virtuale, cu un mic plus pentru VMware.

În testarea performanțelor operațiilor de referință asupra sistemelor de fișiere, a fost utilizat instrumentul IOzone. Acesta generează și măsoară o gamă variată de operații executate asupra fișierelor, dintre care în secțiunea curentă sunt ilustrate rezultatele operațiilor de: scriere, supra-scriere, citire, citire aleatorie, scriere aleatorie, citire inversă (citește fișierul de la sfârșit către început) etc. În urma testului de scriere fișiere mașina virtuală KVM se comportă în general mai bine decât mașinile virtuale VMware și Hyper-V, dar la dimensiuni mai mari ale fișierelor este depășită de performanța mașinii Hyper-

V. În ansamblu, tendința este ca performanțele tuturor mașinilor virtuale să scadă odată cu creșterea dimensiunii fișierelor. În comparație cu rezultatele operației de scriere, în cazul operației de rescriere de fișiere, performanțele tuturor mașinilor virtuale cresc semnificativ. Mașina virtuală KVM este cea mai performantă, fiind întrecută de Hyper-V doar în cazul fișierelor de dimensiune mare. În cazul operației de citire, mașinile virtuale KVM și Hyper-V prezintă performanțe comparabile în timp ce mașina virtuală VMware manifestă, și de această dată, performanțe mai slabe. În cazul citirii aleatorii toate cele trei tipuri de mașini virtuale analizate au performanțe asemănătoare, cu un mic plus pentru VMware și Hyper-V. Rezultatele arată că în cazul scrierii aleatorii, Hyper-V și VMware se comportă semnificativ mai bine decât KVM.

Rezultatele experimentelor de evaluare a performanței reflectă faptul că nu există o soluție universală care să dovedească o performanță maximă în ceea ce privește toate cele trei puncte de interes. În funcție de specificul aplicațiilor utilizatorilor, este recomandat a fi utilizată o anumită soluție. În ceea ce privește performanța procesorului, soluția recomandată pentru a obține cele mai bune rezultate de performanță este VMware. KVM și Hyper-V prezintă rezultate asemănătoare în ceea ce privește performanța operațiilor asupra memoriei, iar în cazul operațiilor asupra sistemului de fișiere, recomandarea este Microsoft Hyper-V. În ceea ce privește performanța procesorului, VMware a demonstrat cele mai bune rezultate în urma testelor efectuate, atât din punctul de vedere al performanțelor procesorului, dar și în ceea ce privește manipularea proceselor. Astfel că, în cazul cerințelor de putere de calcul a procesorului, VMware reprezintă soluția optimă. KVM se situează pe locul al doilea, la o diferență nu foarte mare de liderul VMware, în cazul testului de performanță a procesorului. În schimb, Hyper-V se detașează semnificativ de cei doi, în ambele scenarii de test. Acest lucru poate fi cauzat într-o anumită măsură de driverele FreeBSD Integration Services instalate pentu a face Hyper-V funcțional. VMware prezintă, în schimb, cel mai slab comportament în ceea ce privește viteza operațiilor de scriere și citire realizate asupra memoriei. La scrierea blocurilor de date de tip virgulă mobilă, performanțele VMware sunt bune, însă în cazul blocurile de tip întreg, VMware scade seminificativ. KVM și Hyper-V sunt în permanentă competiție pentru locul întâi, în ceea ce privește operațiile asupra memoriei. Astfel, la citirea din memorie, KVM este recomandat atunci când se operează cu blocuri de dimensiuni mari, iar Hyper-V este de preferat pentru operațiile de citire ale blocurilor de dimensiuni mici și pentru operațiile de scriere ale blocurilor de date de tip întreg. În cazul blocurilor de date de tip virgulă mobilă, KVM și Hyper-V sunt la egalitate. Din punct de vedere al operațiilor asupra fișierelor, Hyper-V prezintă rezultate mai bune decât KVM și VMware. KVM este lider doar în cazul operațiilor de suprascriere, iar VMware în cazul operațiilor de citire aleatorie. La operația de scriere, KVM și Hyper-V manifestă rezultate asemănătoare.

Evaluarea performanțelor în Cloud-ul public vs. Cloud-ul privat

Din punct de vedere al performanței procesorului, în Cloud-ul privat toate cele patru mașini virtuale au rezultate similare în timp ce în cazul cloud-ului public se observă diferențe de performanțe între mșinile virtuale. Acest lucru poate fi cauzat de încărcările variabile ale mașinilor din Cloud-ul public în timp ce în Cloud-ul privat mașina fizică a fost folosită exclusiv pentru evaluarea de performanță, fiind încărcată mult sub capacitatea sa. Aceasta este și explicația pentru faptul că, per ansamblu, scorul obținut pe mașinile virtuale din Cloud-ul public sunt inferioare celui din Cloud-ul privat.

Testele de citire și de scriere de date în virgulă mobilă în memorie au fost configurate pentru citirea, respectiv scrierea de blocuri de date în virgulă mobilă cu dimensiunea de 64 KB. În cazul Cloud-ului public se observă mici diferențe de performanță între mașinile virtuale iar pe ansamblu performanțele mașinilor virtuale care rulează în Cloud-ul public sunt puțin mai scăzute decât cele ale mașinilor virtuale care rulează în Cloud-ul privat.

Pentru evaluarea de performanță a sistemului de fișiere, au fost efectuate teste privind citirea și scrierea fișiere cu dimensiunea de 1 MB. Atât în cazul operației de citire cât și în cazul operației de scriere fișiere Cloud-ul privat oferă performanțe mai bune decât Cloud-ul public, diferența fiind uneori semnificativă.

Chiar dacă rezultatele în valori absolute ale evaluărilor de performanțe efectuate pot fi uneori considerate ca nefind pe deplin relevante deoarece nu pot fi controlate resursele fizice din Cloud-ul public, analiza comparativă dintre cele două tipuri de Cloud-uri este pe deplin relevantă fiind foarte utilă utilizatorilor atunci când trebuie să decidă cu privire la tipul de serviciu Cloud care se recomandă pentru specificul aplicațiilor lor. Performanțele mai scăzute ale mașinilor virtuale din Cloud-ul public se explică prin partajarea resurselor fizice între mai mulți clienți, ceea ce duce la fluctuații ale încărcării procesoarelor, ale timpului de acces la memorie și ale capacității de trafic de I/O.

În toate experimentele efectuate Cloud-ul privat a dovedit performanțe superioare Cloud-ului public. Totodată, în cazul Cloud-ului public, experimentul a relevat o variație, aparent nejustificată, a performanțelor între mașini virtuale cu configurații identice. Motivul acestei instabilități a performanței este cauzat de partajarea resurselor Cloud-ului public între diverși utilizatori ceea ce duce la încărcări neuniforme ale mașinilor. Concluzia acestor evaluări de performanță este că în situația în care performanțele și securitatea reprezintă cerințele cele mai importante, se recomandă utilizarea unui Cloud privat. În cazul în care aspectele prioritare sunt prețul de cost și scalabilitatea se recomandă utilizarea unui Cloud public.

Evaluarea performanțelor modelelor matematice propuse

Modelarea matematică este, cu siguranță, o metodă utilă în îmbunătățirea performanțelor sistemelor virtualizate întrucât în urma implementării diverselor modele matematice se pot extrage concluzii interesante cu privire la comportamentul virtualizat al multor scenarii practice.

Modelele matematice propuse în capitolul 4 au avut drept obiectiv determinarea modului în care varierea ratei de sosire λ afectează timpul de răspuns al sistemului. O rată de sosire accelerată, necesită un număr mai mare de noduri de procesare pentru a optimiza timpul de răspuns. În schimb, simularea efectuată demonstrează că atunci când rata de sosire crește, utilizarea sistemului descrește asimptotic odată cu creșterea numărului nodurilor de procesare și există o limită după care, creșterea numărului de noduri de procesare nu mai determină scăderea timpului de răspuns. Acest rezultat este contrar așteptării ca mărind numărul de noduri de procesare, timpul de răspuns să scadă exponențial. Timpul de răspuns se stabilizează cu utilizarea serverelor. În cazul utilizării virtualizării, timpul de răspuns al sistemului crește ușor față de cel al modelului de bază, acest lucru fiind cauzat de procesarea cozilor de așteptare în doi pași, inițial la nivelul mașinilor fizice și apoi la nivelul mașinilor virtuale. Într-un scenariu real va trebui să se țină seama de asemenea de consumul suplimentar de resurse implicat de virtualizare.

Evaluarea performanțelor politicilor de planificare propuse

Din necesitatea de a extinde funcționalitatea modelelor matematice propuse, a fost propusă o politică de planificare bazată pe evenimente, structurată pe două niveluri. Scopul acesteia este de a mapa sarcinile spre procesare, nodurilor, în funcție de puterea lor de procesare și de disponibilitate. Politica de planificare de nivel înalt atribuie mașini virtuale, mașinilor fizice, iar cea de nivel scăzut atribuie sarcini, mașinilor virtuale. Aceste două politici au fost evaluate prin simulare folosind simulatorul CloudSim.

Propunerea unei politici de planificare a avut drept motivație minimizarea consumului de energie. În cazul ideal, timpul de răspuns total al sarcinilor atribuite mașinilor virtuale nu ar trebui să fie afectat de planificatorul sistemului virtual. Această afirmație se bazează pe argumentul că fiecare virtuală mașină prezintă propriile cerințe în ceea ce privește puterea de calcul, și indiferent care mașină fizică este aleasă pentru găzduirea mașinii virtuale, mașina virtuală va prezenta aceste cerințe minime. Această teorie este vizată în construirea algoritmului pe care se bazează politica de planificare.

Modul în care mașinile virtuale sunt alocate gazdelor fizice determină consumul de energie al întregului sistem. Consumul de energie nu este unul liniar, iar modelarea acestuia ar conduce la o problemă analitică complexă. Caracteristicile de alimentare și de performanță ale echipamentelor de calcul de tip server au fost măsurate utilizând un cadru menit să compare puterea și performanța între diferite servere și care servește drept un set de instrumente ce vizează îmbunătățirea eficienței serverului.

Cel mai comun mod de planificare de nivel înalt este unul circular ce se bazează pe algoritmul Round-robin. Deoarece acest algoritm nu consideră minimizarea consumului de energie, o politică de planificare personalizată a fost propusă, ce are ca scop alocarea de mașini virtuale în funcție de numărul de core-uri disponibile. Principiul din spatele acestei politici este următorul: pentru fiecare mașină virtuală ce trebuie alocată, se selectează mașina fizică cu cele mai puține core-uri. În acest fel, se folosesc gazde cu o capacitate mai mică. Scăderea consumului de energie este realizată lăsând mașinile fizice mai puternice inactive, cât mai mult timp. Modul în care mașinile virtuale sunt alocate mașinilor fizice gazdă determină consumul de energie al fiecărei gazde și, în cele din urmă, eficiența energetică a întregului centru de date.

Politica propusă a fost evaluată prin simulare, în comparație cu politica Round-robin. Rezultatele demonstrează o îmbunătățire semnificativă a consumului de energie al datacenter-ului atunci când este utilizată politica de planificare propusă. Reducerea consumului, în comparație cu politica implicită de tip Round-robin, este de 7,9%. Acest lucru denotă, cum poate alocarea eficientă a mașinilor virtuale, conduce la un sistem mai eficient din punct de vedere energetic. Alocarea de mașini virtuale la gazde nu afectează timpii de răspuns ai sarcinilor, care vor fi ulterior, alocate fiecărei mașini virtuale.

Politica de mapare bloc de sarcini-mașină virtuală are un impact direct asupra timpului de răspuns al sistemul și determină timpul de finalizare al fiecărei sarcini. În această direcție a fost propusă o politică de planificare echilibrată. Rezultatele evaluării prin simulare a politicii propuse sunt comparate cu cele ale simulării în care este aplicată politica de planificare Round-robin și demonstrează că performanțele sistemului pot crește cu până la 25% dacă sunt considerați factori precum disponibilitatea și capacitatea mașinii virtuale. Această îmbunătățire se datorează nu numai considerării capacității de calcul a mașinii virtuale, dar și a stării sale curente. Recomandat este a se încărca mașinile virtuale mai puternice, evitând supraîncărcarea lor și a se lăsa cele mai puțin puternice, subîncărcate.

Evaluarea utilizării resurselor de calcul și a impactului scalabilității asupra timpului de răspuns în Cloud

Interesul pentru evaluarea utilizării resurselor de calcul (procesor, memorie RAM și lărgime de bandă) a condus la realizarea unui experiment în care s-a testat gradul de utilizare a resurselor de calcul. Inițial, comportamentul centrului de date a experimentat un vârf al utilizării procesorului la începutul simulării, care apoi s-a stabilizat în timp. Acest lucru se explică prin faptul că, în prima fază, la sosirea sarcinilor în sistem, trebuie realizată planificarea și alocarea sarcinilor la mașinile virtuale. Apoi, suma resurselor necesare pentru a menține execuția sarcinilor a scăzut până când s-a stabilizat. Memoria și utilizarea procesorului sunt strâns corelate, însă procesorul a fost utilizat la o scară mai largă. În ceea ce privește capacitatea de trafic în rețea, se poate aprecia acesta nu a fost folosită. Nu a existat comunicare între sarcinile de lucru. Aceasta este trăsătura distinctivă a aplicațiilor paralele care au fost folosite pe tot parcursul experimentului. Pentru a testa caracteristica de scalabilitate, s-au adăugat două centre cu trăsături asemănătoare primului și s-a înregistrat o scădere semnificativă a timpului total de execuție. În acest caz, CloudSim a folosit o politică de completare circulară - Round-robin și, astfel, timpul de execuție tinde să se stabilizeze.

Scalabilitatea a fost, în continuare, testată adăugând inițial trei centre de date și apoi crescând numărul lor până la opt, cu trăsături similare. S-a constatat că timpul de execuție a scăzut semnificativ la adăugarea primelor centre de date după care, la adăugarea de noi centre de date, scăderile timpului de

execuție s-au redus tinzând să se stabilizeze la o anumită valoare. Prezența mai multor centre de date asigură un nivel ridicat de robustețe, în sensul că, în cazul în care unul dintre centre devine indisponibil, execuția sarcinilor sosite este asigurată, acestea fiind redirecționate către un alt centru, aplicându-se o politică euristică. Acest comportament se reflectă într-o toleranță la eroare crescută a sistemului. Totuși timpul de răspuns, este redus semnificativ până la o anumită limită, după care, tinde să se stabilizeze, iar suplimentarea numărului de centre de date nu mai conduce la o scădere a acestuia.

Direcții ulterioare de cercetare

Experimentele de evaluare a performanței demonstrează rezultate precum: reducerea costurilor hardware, reducerea consumului de energie și utilizarea eficientă a mașinilor fizice ș.a. Având în vedere utilizarea pe scară tot mai largă a serviciilor de tip Cloud bazate pe virtualizare, determinată în primul rând de considerente economice, continuarea cercetării în domeniul virtualizării în vederea îmbunătățirii performanțelor sistemelor Cloud și a reducerii costurilor pentru utilizatorii de servicii Cloud este deosebit de utilă. Rezultatele obținute, prezentate în această lucrare, dovedesc că se pot aduce multe îmbunătățiri sistemelor de calcul bazate pe virtualizare și că domeniul calculului de tip Cloud nu este doar de mare actualitate, dar și de viitor.

Problemele de cercetare de interes în domeniul optimizării performanței calculului de tip Cloud sunt multiple: disponibilitatea serviciilor, consumul de energie, furnizarea și gestionarea resurselor, gestionarea mașinilor virtuale, echilibrarea încărcăturii, securitatea etc. Acestea deschid mai multe direcții de cercetare, care introduc și numeroase provocări. Studiile asupra virtualizării și utilizării acesteia în sisteme Cloud prezentate în lucrare, pot fi extinse în diverse direcții.

Considerând rezultatele experimentelor de analiză a performanței prezentate în această lucrare, în viitor ar fi de interes dezvoltarea de noi politici de planificare luând în considerare noi criterii, dar și scenarii privind sarcini complexe ce implică sincronizarea și comunicarea între task-urile componente. De asemenea, ar fi utilă și dezvoltarea unor noi modele care să ia în considerare replicarea datelor și utilizarea serverelor multiple de stocare. De asemenea, în viitorul apropiat, se intenționează proiectarea și implementarea propunerilor din prezenta lucrare în sisteme Cloud reale prin intermediul platformei OpenStack. Acestă platformă va permite controlul unei game variate de resurse de calcul, de stocare și de comunicare fiind platforma ideală pentru infrastructuri eterogene. O altă direcție de cercetare viitoare o reprezintă scalabilitatea serviciilor Cloud în condițiile atingerii limitării impuse de resursele hardware disponibile. Într-o astfel de situație, un furnizor de servicii Cloud ar putea delega o parte din sarcinile de lucru unui alt furnizor de servicii Cloud, în mod transparent pentru utilizatori. Un astfel de scenariu necesită cercetare în domeniul managementului SLA (Service Level Agreement) și investigarea diverselor aspecte legate de SLA cum ar fi: monitorizarea QoS, violarea SLA, implicații financiare etc.

Referințe

Abirami, S., & Shalini, R. (2012). Linear Scheduling Strategy for Resource Allocation in Cloud Environment. *International Journal on Cloud Computing: Services and Architecture, Vol. 2, No. 1*.

Chang, B., Tsai, H.-F., & Chen, C.-M. (2013). Evaluation of Virtual Machine Performance and Virtualized Consolidation Ratio in Cloud Computing System. *Journal of Information Hiding and Multimedia Signal Processing*.

Diogo, M. M., & Ferraz, L. G. (2012). Virtual Network Performance Evaluation for Future Internet Architectures. *Journal of Emerging Technologies in Web Intelligence, Vol. 4, No. 4*.

Falatah, M., & Batarfi, O. (2014). Cloud Scalability Considerations. *International Journal of Computer Science & Engineering Survey (IJCSES), Vol.5, No.4*.

Goswami, B., & Saha, S. (2013). Resource Allocation Modeling in Abstraction using Predator-Prey dynamics: A qualitative analysis. *International Journal of Computer Applications, Vol. 61, Nr. 6*.

Graniszewski, W., & Arciszewski, A. (2016). Performance analysis of selected hypervisors (Virtual Machine Monitors - VMMs). *International Journal of Electronics and Telecommunications*, 231-236.

Indrani, P., Sudhakar, Y., & Lizy, J. K. (2012). Performance Impact of Virtual Machine Placement in a Datacenter. *IEEE 31st International Performance Computing and Communications Conference*, (pp. 424-431). Austin, TX.

ISO-International Standard for Organizations. (2014, 10 15). Information technology - Cloud computing - Overview and vocabulary, ISO/IEC17788.

Johnsen, E., Schlatte R., R., & Tapia, S. T. (2012). Modeling Resource-Aware Virtualized Applications for the Cloud in Real-Time ABS. *ICFEM 2012. Lecture Notes in Computer Science, vol 7635.* Ber: Springer, Berlin, Heidelberg.

Kundu, S., Rangaswami, R., Dutta, K., & Zhao, M. (2010). *Application Performance Modeling in a Virtualized Environment.* Florida, USA: School of Computing & Information Sciences, College of Business Administration Florida International University.

Mancaș, C. (2015). Performance Improvement through Virtualization. *RoEduNet 13th International Conference: Networking in Education and Research.* Craiova.

Mancaș, C. (2014). Best Practices in Distributed Learning Environments. *The International Scientific Conference eLearning and Software for Education, Vol. 3* (pp. 272-279). București: "Carol I" National Defence University.

Mancaș, C., Mocanu, M., & Mancaș, D. (2013). Congestion Avoidance in Multimedia Networks. *RoEduNet 11th International Conference: Networking in Education and Research.* Sinaia, România.

Mell, P., & Grance, T. (2011). *The NIST Definition of Cloud Computing.* Gaithersburg: Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, United States Department of Commerce.

Tarasov, V., Bhanage, S., Zadok, E., & Seltzer, M. (2011). Benchmarking File System Benchmarking: It *IS* Rocket Science. HotOS.

Vinothina, V., Sridaran, R., & Ganapathi, P. (2012). A Survey on Resource Allocation Strategies in Cloud Computing. *International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 3, No. 6*.

Wang, Q., & Varela, C. A. (2011). Impact of Cloud Computing Strategies on Workloads Performance. *IEEE Computer Society*, 130-137.

Yamato, Y. (2017). Performance-aware server architecture recommendation and automatic performance verification technology on IaaS cloud. *Service Oriented Computing and Applications, Vol. 11, No. 2*, 121-135.

Zhang, X., Tune, E., Hagmann, R., Jnagal , R., Gokhale, V., & Wilkes, J. (2013). *CPI: CPU performance isolation for shared compute clusters.* ACM, Google Inc.