Universitatea “Politehnica” Bucuresti
Facultatea de Electronica, Telecomunicatii si Tehnologia Informatiei

Cloud Computing
Gestionarea datelor științifice

Prof. Coordonator:
Prof. Dr. Ing. Ștefan Stăncescu

Realizatori:
Ciochină Roxana Elena
An I, Master IISC

2013

1. Introducere

Popularitatea crescuta a serviciilor de Internet ca Amazon Web Services, Google App Engine si Microsoft Azure au atras atentia asupra conceptului de Cloud Computing. Cu toate ca acest termen este nou, tehnologia a aparut ca o extensie a grid-ului, virtualizarii, tehnologiei Web 2.0 si a tehnologiilor SOA (Service Oriented Arhitecture). Mai mult decat atat, interesul in Cloud Computing a aparut ca o prevalence a procesoarelor multi-core si a costurilor scazute a hardware-ului, precum si a costului din ce in ce mai mare al energiei consumate. Drept urmare, in numai cativa ani Cloud Computing-ul a devenit una dintre tehnologiile IT revolutionare, asa cum se poate vedea si in Fig. 1.

Termenul de Cloud Computing este folosit pentru a denumi Internetul. Este de obicei folosita o forma de nor (cloud) in diagramele retelelor pentru a reprezenta flexibilitatea topologiei si a abstractiza infrastructura. Aceasta tehnologie foloseste Internetul pentru a furniza servicii hardware, medii de programare si software, facand invizibila pentru useri infrastructura underlying. In ciuda popularitatii termenul este in continuare abstract, neexistand o definitie formala a Cloud Computing-ului.

Cloud Computing-ul ofera usersilor posibilitatea de a accesa diferite resurse de calcul, cum ar fi ciclurile de calcul, spatiul de stocare, medii de programare si aplicatii software (userul are nevoie doar de un browser). Cloud Computing mai ofera beneficii precum:

- Investitii mai mici: sunt oferite solutii de scalare si managementul peak-urilor la preturi mult inferioare costurilor traditionale de spatiu, timp si investitii financiare.
- Scalare: Vendorii de Cloud au centre de date ce cuprind mii se servere, oferind putere de calcul si spatiu de stocare nelimitat
- Management: Experienta userului este simplificata, nu este nevoie de configurarea sistemelor sau de backup.

Cu toate acestea, Cloud Computing ridica si multe semne de intrebare, in principal in privina securitatii, compliance si reliability. Atunci cand un un utilizator alege sa isi desfasoare activitatea in Cloud, nu are nicio certitudine ca nimeni altcineva nu ii va putea accesa datele. Daca echipamentele sunt situate intr-o alta tara pot aparea probleme legate de jurisdictie si controlul datelor. Totodata, nu este bine definit SLA-ul (Service Level Agreement) oferit de providerii de cloud.

2. Aspecte functionale ale Cloud Computing

Din punct de vedere conceptual, userii folosesc platforme computationale sau infrasctuctiri IT in Cloud si isi ruleaza in interiorul acestuia aplicatiile. Astfel, Cloud-ul ofera utilizatorilor accesul la servicii hardware, software, resurse de date ceea ce inseamna o platforma integrata de calcul sub forma unui serviciu, intr-un mod transparent:

- HaaS - Hardware as a Service
  Virtualizarea tot mai rapida a hardware-ului, automatizarile IT si posibilitatea masurarii timpului si costului serviciilor folosite au dus la o posibilitate a utilizatorilor de a „cumpara" hardware IT, sau chiar intregi cente de date sub forma „pay-as-you-go". HaaS este un concept flexibil, scalabil si usor de coordonat, capabil sa satisfaca cererile utilizatorilor. Exemple: Amazon EC2, proiectul Cloud Blue al celor de la IBM, Nimbus, Eucalyptus, Enomalism.

- SaaS – Software as a Service
  In cadrul acestui concept un software sau o aplicatie sunt gazduite ca serviciu si oferite utilizatorilor prin intermediul Internetului. In acest fel nu mai este nevoie ca utilizatorul sa instaleze pe propriul calculator programele respective, sa se ocupe de mentenanta acestora, fiind eliminate majoritatea costurilor legate de acest aspect. Un exemplu in acest sens este browser-ul Google Chrome care este capabil sa ofere utilizatorului un nou desktop, prin intermediul caruia aplicatiile sunt oferite (local sau la distanta).
- DaaS – Data as a Service
  Informatia din mai multe surse si in mai multe formate poate fi accesata prin intermediull serviciilor oferite de Internet. Utilizatorii pot astfel opera date de la distanta.

  Pe baza suportului HaaS, SaaS si DaaS, Cloud Computing-ul poate furniza de asemenea si o platforma – PaaS (Platform as a Service).Utilizatorii pot astfel avea acces la configuratiile hardware, software-ul si informatiile de care au nevoie.

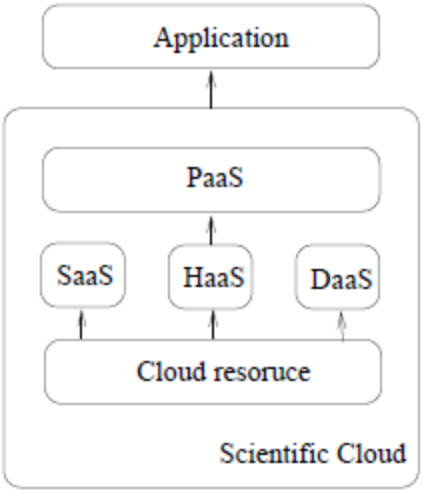

Figure 1 Platforma Cloud Computing

Cloud Computing-ul a intrat in interesul comunitatii stiintifice datorita posibilitatilor de a lucra cu volume foarte mari de date. Se realizeaza in acest fel o economie in design-ul si constructia partii hardware, grupurile de utilizatori being able sa gazduiasca, sa proceseze si sa analizeze volume din ce in ce mai mari de date din diferite surse. Vendori precum Amazon Web Services, Google App Engine, AT&T Synaptic Hosting, Rackspace, GoGrid sau AppNexus promit utilizatorilor experienta unei puteri computationale infinite, a unui spatiu de stocare care poate fi utilizat la cerere, intr-un mod "pay-only-for-what-you-use" (platesti doar pentru ceea ce folosesti).

2.1 Probleme in experimentele stiintifice

Managementul datelor stiintifice in cloud difera de abordarile anterioare prin prisma mai multor aspect. In primul rand, in cazul cloud computing-ului capabilitatile computationale si spatial de stocare sunt consolidate in centre de date de marimi inpresionante pe cand in cazul HPC (High Performance Computing) cercetatorii folosesc concomitent supercalculatoare paralele. Accesul la aceste resurse de procesare rapida este facut prin sisteme pe baza de cozi. Input-ul si output-ul computational este transferat intre nodurile de calcul si cele de stocare care sunt situate separate. In cloud, stocarea datelor si calculul se fac in acelasi loc, ceea ce suce la o schimbare de paradigm in multe dintre procesele prelucrare si analiza a datelor.

In al doilea rand, gazduirea datelor intr-o maniera centralizata poate fi un catalizator pentru sharing-ul de date dintre diferite domeni. Un exemplu de gazduire de date multidisciplinare poate si SciDB, care include date din domenii precum astronomie, biologie, meteorology, oceanografie si fizica.

Un alt aspect este sustenabilitatea acestui concept. In cloud configuratiile sunt facute in asa fel incat datele pot si reproduce in cazul unor esecuri tranzitorii sau permanente, precum si in cazul coruperii fisierelor. Conservarea datelor este un aspect critic in majoritatea domeniilor stiintifice, astfel ca managementul datelor in cloud este privit ca o solutie net superioara stocarii datelor la nivel local.

Gestionarea datelor in calculul stiintific include captura de date, procesarea si analiza seturilor de date. Majoritatea sunt produse de instrumente experimentale sau observationale cum ar fi: telescoape, radare Doppler, sateliti sau acceleratoare de particule – Large Hadron Collider. Colectarea acestor mari cantitati de date poate cauza ocazional probleme, in procesele initiale de acumulare, transfer sau mai apoi de stocare.

Integrarea datelor din diverse surse a fost de asemenea o provocare datorita diferentelor intre diferitele patter-uri de livrare sia diferentelor de formate. In plus, datele sunt produse si in faza de calcul sau cea de simulare. Mai mult, apar in plus fata de datele experimentale (prime, derivate sau combinate) orice rezultate sau publicatii aparute datorita acestor experimente, care sunt de asemenea colectate si gestionate ca parte integranta a datelor stiintifice.

Analiza si simularea datelor implica deseori vizualizarea lor. Datele colectate sunt in general stocate inainte de a fi accesate in procesele de analiza si vizualizare – dorindu-se in acest sens conservarea acestora pe termen lung.

Diferitele stagii de prelucrare a datelor stiintifice nu se executa numai secvential, ci si recursive si interactive cu alte etape din cadru experimentelor stiintifice. Prelucrarea acestora implica adesea schimbul dinamic intre oameni sau grupuri de oameni de stiinta.

3. Tehnologii in cadrul Cloud Computing

Exista o serie de tehnologii care alcatuiesc Cloud Computing-ul:

- **Tehnologia virtualizarii**

Se realizeaza partitionarea hardware pentru a furniza platforme computationale flexibile si scalabile. Tehnica masinilor virtuale, cum ar fi VMware sau Xen, ofera o infrastructura IT la cerere.

- **Orchestrarea fluxului de servicii si a fluxului de lucru** se realizeaza prin intermediul unui set complet de servicii sablon la cerere, care pot fi compuse din servicii efectuate in interiorul Cloud Computing. Conceptul trebuie sa fie capabil sa dirijeze in mod automat serviciile din diferite surse sau de diferite tipuri pentru a avea un flux transparent si dinamic pentru utilizatori.

- **Servicii Web si SOA (Service Oriented Architecture)**

Serviciile in Cloud Computing sunt expuse in mod normal serviciilor Web, conform standardelor industriei, cum ar fi WSDL, SOAP, UDI.

- **Web 2.0** este o tehnologie emergent descrisa de trendul inovativ de a folosi tehnologia World Wide Web si design-ul Web pentru a influenta creativitatea, sharing-ul de informatii, colaborarea si functionalitatea Web-ului.

- **Modelul de programare**

Utilzatorii introduce in cloud datele si aplicatiile; unele modele de programare sunt propuse pentru a usura adaptarea utilizatorilor la infrastructura din cloud. Pentru simplitate si acces eficient la serviciile oferite de cloud, modelul de programare al acestuia nu ar trebui sa fie prea complex sau prea inovativ.

MapReduce este un model de programare asociat cu implementarea procesarii si generarii de seturi mari de date in interiorul infrastructurii Google. Acest model implica initial aplicarea unei operatii de mapare unor seturi de date – un set de perechi cheie/valoare, si apoi aplica o operatie de reducere tuturor valorilor care au aceeasi cheie. Metoda Map-Reduce-Merge are in plus operatia “merge”.

Framework-ul Hadoop implementeaza paradigm MapReduce si furnizeaza un fisier de sistem distribuit – Hadoop Distributed File Sistem. Tehnologiile MapReduce si Hadoop au fost adoptate de proiectul de Cloud Computing realizat prin colaborarea Yahoo!, Intel si HP.

Vendors	Execution engine	Distributed data storage (unstructured)	Distributed data storage (Structured)	High-level data analysis
Google	Google Map-reduce	Google File System(GFS)	BigTable	Sawsall
Microsoft	Dryad	Azure, Cosmos	SQL Azure	DryadLINQ
Apache	Hadoop Map-reduce	Hadoop Distributed File System (HDFS)	HBase Hypertable (Zvents)	Hive, Pig Latin, and Pig
Amazon	Elastic Compute Cloud, Elastic MapReduce,	Simple Storage Service (S3), Elastic Block Storage (EBS)	Dynamo, SimpleDB, Relational Database Service (RDS)	
Facebook/ Yahoo			PNUTS, Cassandra	Pig, Hive
Other Efforts	WheelFS, Synaptic Hosting, AppNexus, GoGrid, Rackspace	Synaptic Storage		

Figure 2 Tehnologii Cloud Computing

Un sumar al tehnologiilor din cadrul Cloud Computing poate fi vizualizat in tabelul urmator:

3.1 Spatiul virtual Globus

Un spatiu virtual este o abstractizare a unui spatiu computational care este disponibil in mod dinamic utilizatorilor autorizati prin intermediul serviciilor de Grid. Aceasta abstractizare atribuie o cota a resurselor mediului de executie la implementare (spre exemplu processor si memorie) precum si aspecte ale configuratiei software a mediului (spre exemplu sistemul de operare istalat sau serviciile furnizate). Serviciul Workspace permite unui utilizator Globus sa implementeze si sa genstioneze in mod dinamic mediul de calcul. Spatiul virtual furnizeaza urmatoarele interfete de acces:

- Workspace Factory Service are o operatie denumita *create* care are 2 parametri necesari: metadata si cererea de implementare pentru metadata.
- Dupa crearea spatiulu de lucru, acesta este reprezentat ca o resursa WSRF. Spatiul de lucru poate fi inspectat si dirijat prin intermediul operatiilor *Workspace Service.*
- *Group Service* permite unui utilizator autorizat sa gestioneze un set de spatii de lucru la un moment dat.
- *Status Service* ofera interfata prin intermediul careia un client poate interoga datele colectate despre folosirea serviciului.

Pe baza serviciului Globus a fost implementat un kit denumit Nimbus cu ajutorul caruia conceptual de Cloud a patruns in spatial stiintific.

Fiind client Nimbus, utilizatorul poate: explora imaginile masinilor virtuale din interiorul Cloud-ului, adauga propriile imagini ale masinilor virtuale, implementa masini virtual si interoga statusul masinilor virtuale, si nu in ultimul rand accesarea lor.

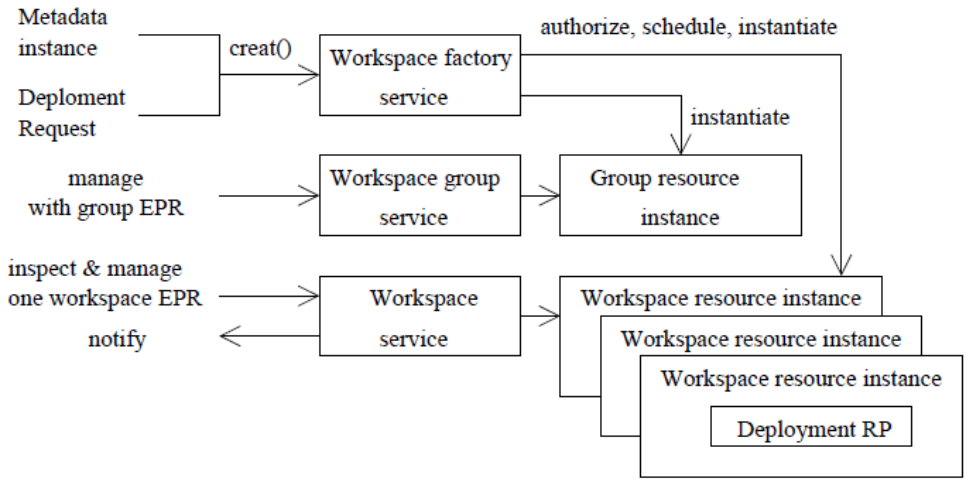

Figure 3 Spatiul Virtual Globus

3.2 Violin – Virtual Inter-networking on Overlay Infrastructure

Violin este o tehnologie virtuala de retea, care creeaza retele de IP de oridin superior si ofera careva avantaje:

- Creare la cerere de masini virtual si retele de IP virtual care le conecteaza
- O configurare a topologiei retelei virtual si a serviciilor, a serviciilor sistemului de operare si a serviciilor de aplicatii, pachete si librarii
- Obtinerea unei compatibilitati binare prin crearea mediului retelei si a timpului de lucru, sub care a fost initial dezvoltata aplicatia
- Inlaturarea impactului negativ datorat izolarii spatiului de adrese al retelei virtual

In figura urmatoare poate fi vizualizata o structura pe mai multe nivele – 2 structuri Violin urmate de o infrastructura de calcul comun. In partea inferioara, partea fizica, cu elemente de retea heterogene, se intinde pe mai multe domenii. In partea de mijloc, sistemele middleware unifica resursele de retea pentru a forma o infrastructura comuna. Partea superioara este formata din 2 structuri Violin isolate: fiecare detine propria retea, propriul sistem de operare si serviciile de aplicatii personalizate pentru aplicatia care ruleaza in interiorul acesteia.

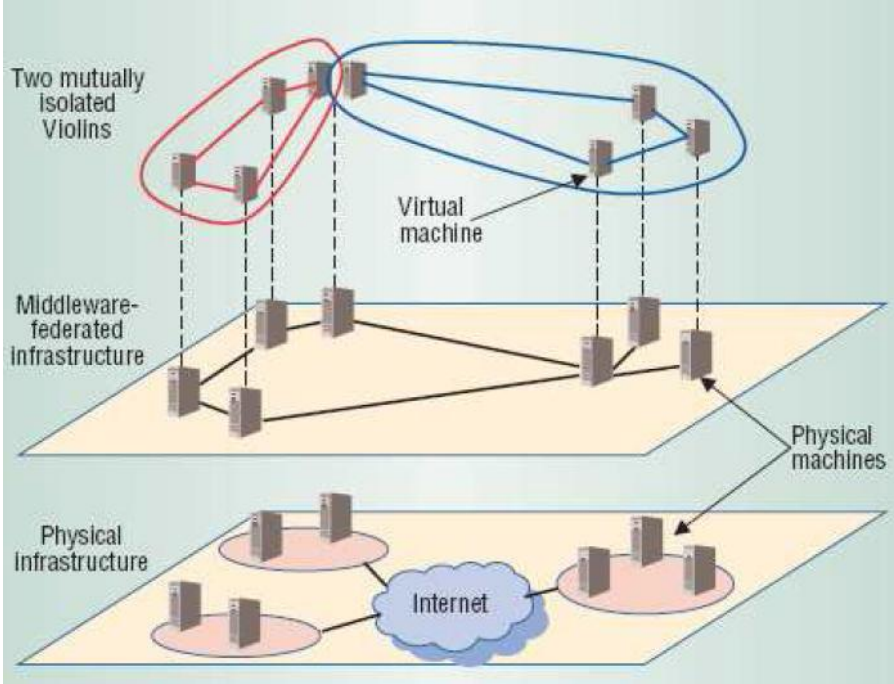

Figure 4 Structura pe mai multe nivele

3.3 Virtuoso

Acest sistem este dezvoltat de Northwestern University si urmareste crearea unei piete pentru folosirea resurselor. Furnizorii pot atfel sa isi vanda resursele consumatorilor sub forma unor masini si retele virtuale.

In sistemul Virtuoso, utilizatorul primeste o masina virtuala aflata la distant, care are configuratia dorita: tipul procesorului, marimea memoriei si resursele de stocare. Utilizatorul poate instala si configure orice software pe masina virtuala, spre exemplu sisteme de operare, compilatoare sau librarii software.

In cadrul acesteia reteaua virtual VNET uneste masinile virtuale ale consumatorilor de resurse in mos eficient in raport cu reteaua locala de resurse. VNET este o retea virtual adaptive care poate folosi interferenta traficului intre masinile virtuale, migrarea masinilor virtuale, manipularea regulilor de rutare si a topologiei in cazul supraincarcarii, rezervarea resurselor pentru optimizarea performantelor unei aplicatii distribuite sau paralele ce ruleaza in masinile virtuale ale utilizatorilor.

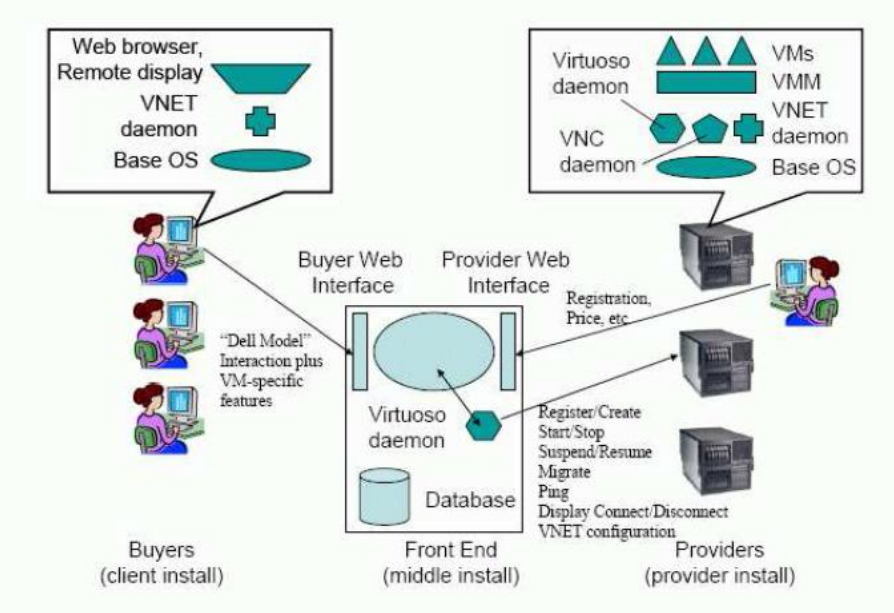

Figure 5 Sistemul Virtuoso

3.4 OpenNEbula

Este un motor de infrastructura virtual care permite implementarea dinamica si realocarea masinilor virtuale in cadrul unui cadru al resurselor fizice. Sistemul OpenNEbula extinde beneficiile virtualizarii platformelor de la o singura resursa fizica la un cadru de resurse, decupland server-ul nu numai de la infrastructura fizica ci si din locatia fizica.

OpenNEbula are urmatoarea structura: un frontend si multiple backend-uri. Frontend-ul furnizeaza accesul utilizatorilor la interfetele de acces si functiile de management. Partile de backend sunt instalate pe servere Xen (hipervizorii Xen sunt porniti si masinile virtual pot fi garantate). Comunicatiile intre frontend si partile de backend se fac prin SSH (Secure Shell). OpenNEbula ofera puncte unice de acces utilizatorilor pentru a intrebuinta masinile virtual pe infrastructuri distribuite local.

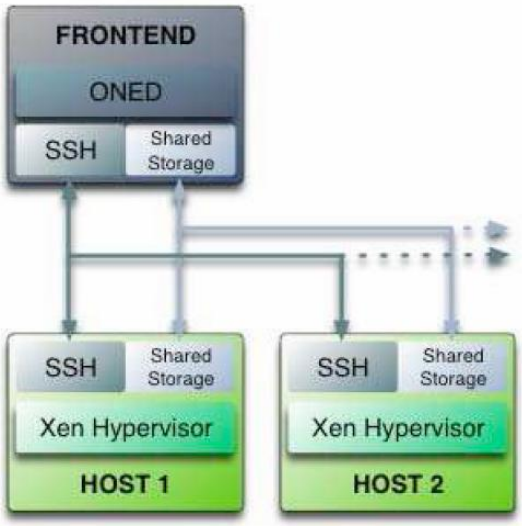

Figure 6 Sistemul Open NEbula

4. Studiu comparativ: Cloud Computing vs. Grid Computing

Grid Computing-ul este un concept oricentat pe calculul distribuit la nivel inalt, si urmareste folosirea in comun a resurselor de calcul pentru executarea la distanta si rezolvarea problemelor la scara larga. Conceptul evidentiaza partea de resurese facand eforturi impresionante pentru a construi un sistem distribuit complet si independent. Cloud Computing-ul furnizeaza servicii si functionalitati utilizatorului pentru ca acesta sa isi configureze dupa preferinte mediul de calcul. Cloud Computing-ul este orientat pe industrie si foloseste un model orientat pe aplicatii.

Din punct de vedere al infrastructurii Grid Computing -ul se prezinta ca un sistem descentralizat, care se intinde pe suprafete geografice distruibuite, lipsite de un control central. In mod normal contine resurse heterogene, cum ar fi configuratii hardware si software sau acces la interfete. Cloud Computing -ul opereaza ca un server central cu un singur punct de acces. Infrastructura Cloud se poate intinde pe mai multe centre de calcul, spre exemplu Amazon sau Google, si contin in genral resurse omogene, operate la nivel central.

Grid Computing -ul doreste sa ofere acces dependent, consistent si ieftin la calabilitati computationale de nivel inalt. Totusi, utilizatorii neexperimentati intampina dificultati in adaptarea propriilor aplicatii la Grid Computing. Mai mult decat atat, este dificila garantarea performatelor in cazul calculului in Grid. Cloud Computing -ul pe de alta parte, ofera medii de calcul customizabile, scalabile, ce garanteaza QoS-ul pentru utilizatori, cu un acces usor si extins.

Grid Computing -ul sta la baza multor povesti de success din multe domenii. Un faimos exemplu recent este proiectul LCG, de procesare a datelor generate de LHC-ul (Large Hadron Collider) de la CERN.

Se poate aforma faptul ca Grid Computing -ul a stabilit o infrastructura bine organizata si o experienta a aplicatiei. Cloud Computing -ul depaseste aceste performante prin oferirea unor medii de calcul cu caracteristici diferite, garantarea QoS si configurarea conform specificatiilor utilizatorului.

5. Capabilitatile Cloud Computing-ului

Nevoile curente si viitoare in gestionarea si procesarea datelor stintifice sunt partial satisfacute datorita unui numar de insuficente in Cloud Computing. In unele cazuri sistemele deja existente vin cu solutii ad hoc, pe cand in altele inca se mai fac cercetari pentru inlaturarea defectelor.

5.1. Functionalitatea tehnologiilor

Nepotrivirea intre modelul de programare si capabilitatile bazelor de date este un aspect demn de luat in considerare. Tehnologiile din cloud sunt incapabile deseori sa furnizeze functionalitatile cerute de aplicatiile stiintifice. Indexarea schemelor pe baza unor tipuri adecvate de date pentru obiectele de date stiintifice intampina de asemenea probleme in cloud.

5.2 Toleranta la erori

Mediile de calcul de tip cloud sunt in general construite folosind hardware ieftin, ducand la aparitia erorilor. Probabilitatea de eroare in cazul unui task de analiza a datelor care se intinde pe o durata mare de timp este foarte mare. Spre exemplu Google reporteaza o medie de 1.2 erori la un task de analiza. Detectia rapida a erorilor si schemele de recuperare ofera pentru cercetatori un mediu de analiza a datelor mult mai sigur.

5.3 Formatul datelor stiintifice si unelte de analiza

O sarcina critica pentru aplicatiile stiintifice este descoperirea unui subsir pentru un modul de calcul. Sirurile de date stintifice sunt deseori stocate ca fisiere in formate stiintifice precum HDF (Hierarchical Data Format), NetCDF (Network Common Data Form), FITS (Flexible Image Transport Sistem) care furnizeaza anumite attribute fisierelor, usurand descoperirea de subsiruri relevante.

Aplicatiile stintifice codeaza in fisierul de sistem attribute cheie, care reprezinta nu numai calea fisierului in fisierul de sistem ci si elemente de functionalitatea filtrarii pentru asistarea procesului de descoperire a datelor. Totusi, pe masura ce fisierul de sistem aduna milioane de fisiere de dimensiuni impresionabile, datele stiintifice au nevoie de mai multa informative pentru a fi descries. Astfel a aparut conceptul de metadata – informative despre date, stocata in general separat, si care se leaga logic sau fizic de fisierul de date.

5.4 Datele in timp real

Una dintre caracteristicile distrinctive ale gestionarii datelor stiintifice este diversitatea surselor. Senzorii moderni si echipamentul experimental digitizat introduce datele direct in sistem. Aceste date in timp real sunt procesate diferit atunci cand ajung in sistem fata de procesarea conventionala a datelor, cum ar fi paginile web sau informatiile personale. Accesul eficient la datele in timp real ar putea permite exploatarea datelor in timp real si eventual imbunatatirea performantelor de calcul.

5.5 Securitate

Un aspect important al lucrului in Cloud Computing cu date stiintifice il reprezinta securitatea accesului utilizatorilor pentru fiecare din actiunile de descoperire, browsing sau calcul. Astfel se face criptarea datelor inainte de upload-are. Pentru a evita accesul neautorizat la aceste date orice aplicatie care ruleaza deja in cloud nu ar trebui sa fie capabila sa faca direct decriptarea datelor. Totusi, decriptarea seturilor de date, mutarea acestora in si din spatiul de stocare din cloud este o sarcina care consuma intensiv latimea de banda. De aceea s-a sugerat faptul ca un sistem de analiza a datelor care poate opera direct pe date criptate ar putea imbunatati semnificativ performanta.

6. Concluzii

Cloud computing-ul ofera avantaje evidente, cum ar fi colocatia datelor cu zona de calcul si economia de scala in gazduirea serviciilor. In prezent aceste platforme sunt folosite cu predilectie in implementarea motoarelor de cautare sau a gazduirii elastic a site-urilor web comerciale. Rolul lor in calculul stiintific este in continua schimbare si evolueaza constant. In unele scenarii de analiza stiintinfica, datele trebuie sa fie stocate cat mai aproape de experiment. In altele, este preferata o banda cat mai larga.

Tendinta de a muta datele stiintifice in cloud a devenit din ce in ce mai evident. Este de asteptat ca aceasta tendinta sa continue si sa se accelereze pe vitor. Pe masura ce tot mai multe sisteme isi desfasoara activitatea in cloud, problemele enuntate in capitolul anterior devin din ce in ce mai importante, devenind o arie de cercetare prospera.

Bibliografie

[1] I. Foster, C. Kesselman. “The Grid: blueprint for a new computing infrastructure”. Morgan Kaufmann

[2] K. Keahey, I. Foster, T. Freeman, and X. Zhang. “Virtual workspaces: achieving quality of service and quality of life in the Grid. *Scientific Programming”*

[3] Monaco, Ania “A vew inside the cloud”

[4]Baburajan Rajani. “The rising cloud storage market opportunity strengthens vendors”, Infotech August 2011

[5] Amy Schurr, “Keep an eye on the Cloud Computing”, Network World

[6] Cryptoclarity.com “Encrypted Storage and Key management for the cloud”

[7] Nimbus Project, http://workspace.globus.org/clouds/nimbus.html/

[8] OpenNEbula Project, http://www.opennebula.org/