**Probleme și provocări ale tehnicilor de load balancing în Cloud computing**

Load balancing oferă posibilitatea de a distribui în mod egal volumul de lucru pe resursele disponibile. Obiectivul este de a furniza servicii continue în caz de defecțiune a oricărei componente a serviciului prin furnizarea instanțelor aplicației împreună cu utilizarea adecvată a resurselor. În plus, load balancing-ul are ca scop minimizarea timpului de răspuns pentru sarcini și îmbunătățirea utilizării resurselor, ceea ce duce la o performanța a sistemului mai bună, la un cost mai mic. De asemenea, load balancing-ul urmărește să ofere scalabilitate și flexibilitate pentru acele aplicații a căror dimensiune poate crește în viitor și necesită mai multe resurse, precum și pentru a oferi prioritate lucrărilor care necesită execuție instantanee. Alte obiective sunt reducerea consumului de energie, evitarea blocajelor, furnizarea de resurse și îndeplinirea cerințelor QoS (quality of service - calitatea serviciului) pentru îmbunătățirea load balancing-ului.

Tehnologia Cloud computing devine ținta unor cercetări mai avansate, atât în domeniul datelor, cât și al celui de calcul, vizând aspecte teoretice și practice. Cu toate acestea, cercetarea în Cloud computing se confruntă cu o mulțime de probleme, load balancing fiind una dintre provocările care necesită atenție specială. O listă cu câteva probleme de load balancing este prezentată mai jos:

1. **Noduri distribuite geografic**: În general, centrele de date din Cloud sunt distribuite geografic în scopuri de calcul. În aceste centre, nodurile distribuite spațial sunt tratate ca un singur sistem de localizare pentru executarea eficientă a solicitărilor utilizatorilor. Unele tehnici de load balancing sunt concepute pentru o zonă mai mică în care nu iau în considerare factori precum întârzierea rețelei, întârzierea comunicării, distanța dintre nodurile de calcul distribuite, distanța dintre utilizator și resurse și așa mai departe. Noduri situate la locații foarte îndepărtate reprezintă o provocare deoarece acești algoritmi nu sunt potriviți pentru acest mediu. Astfel, proiectarea algoritmilor de load balancing pentru nodurile aflate la distanță ar trebui să fie luate în considerare.
2. **Punct unic de defecțiune**: Anumiți algoritmi dinamici de load balancing sunt proiectați astfel încât deciziile pentru load balancing sunt luate de nodul central. Dacă dispozitivul central se prăbușește, atunci va afecta tot. De aceea, este nevoie de dezvoltarea unor algoritmi distribuiți în care un singur nod sa nu fie responsabil de întregul sistem de calcul.
3. **Migrarea mașinii virtuale**: Virtualizarea permite crearea mai multor mașini virtuale pe o singură mașină fizică. Aceste VM sunt independente și au configurații diferite. Dacă o mașină fizică este supraîncărcată, unele VM trebuie să fie transferate într-o alta locație, folosind o abordare de load balancing pentru migrații VM.
4. **Noduri eterogene**: În timpul cercetărilor timpurii despre load balancing în Cloud, cercetătorii au teoretizat despre nodurile omogene. În Cloud computing, cerințele utilizatorilor se modifică dinamic, ceea ce necesită executarea lor pe noduri eterogene pentru a utiliza eficient resursele și pentru a minimiza timpul de răspuns. Prin urmare, invenția unor tehnici eficiente de load balancing pentru mediul eterogen este o provocare pentru cercetători.
5. **Scalabilitatea Load-Balancer-ului**: Disponibilitatea la cerere și scalabilitatea serviciilor Cloud permit utilizatorilor să acceseze oricând serviciile pentru a reduce sau a crește rapid. Un bun load balancer ar trebui să ia în considerare schimbările rapide ale cerințelor în ceea ce privește puterea de calcul, stocarea, topologia sistemului și așa mai departe, pentru a facilita aceste schimbări în mod eficient.
6. **Complexitatea algoritmului**: În Cloud computing, algoritmii ar trebui să fie simpli și ușor de implementat. Un algoritm complex va reduce performanța și eficiența sistemului Cloud.
7. **Managementul spațiului de stocare**: Stocarea în Cloud a rezolvat problema sistemelor tradiționale, mai vechi de stocare care necesitau managementul personalului și care aveau un cost ridicat de hardware. Cloud-ul permite utilizatorilor să stocheze datele în mod eterogen, fără probleme de acces. Stocarea în Cloud crește pe zi ce trece, ceea ce necesită stocarea unei replicări a datelor pentru accesul eficient și pentru coerența datelor. Schemele complete de replicare a datelor nu sunt foarte eficiente din cauza politicii de stocare a datelor duplicate pe punctele de replicare. Replicarea parțială poate fi suficientă dar poate ridica probleme de disponibilitate a setului de date și poate crește complexitatea tehnicilor de load balancing. De aceea, trebuie dezvoltată o tehnică eficientă de load balancing care să ia în considerare și distribuția aplicației dar și a datelor aferente pe baza unui sistem de replicare.

Când vine vorba de load balancing în Cloud, încă există multe probleme și provocări care trebuie discutate și rezolvate în viitor. Din analiza literaturii de specialitate, se pot găsi câteva direcții în care Cloud-ul are nevoie de îmbunătățiri. Pentru a menține performanța în Cloud, trebuie luate în considerare câteva aspecte, precum: calitatea serviciului (QoS), acordul privind nivelul de servicii (SLA), furnizarea resurselor și load balancing-ul, printre alte aspecte. Au fost dezvoltate mai multe multe tehnici de load balancing care iau în considerare diferite metrici, cum ar fi performanța, timpul de răspuns, timpul de execuție, timpul de migrare a sarcinilor și utilizarea resurselor. Nicio tehnică nu a luat în considerare toți parametrii de echilibrare a sarcinii, pentru îmbunătățira performanței generale a centrelor de date.

**Bibliografie**

1. Ghomi EJ, Rahmani AM, Qader NN. Load-balancing algorithms in cloud computing: A survey. Journal of Network and Computer Applications. 2017 Jun 15;88:50-71.
2. Kumar P, Kumar R. Issues and challenges of load balancing techniques in cloud computing: A survey. ACM Computing Surveys (CSUR). 2019 Feb 4;51(6):1-35.
3. https://www.techtarget.com/searchcloudcomputing/definition/cloud-load-balancing