Dr. Toka László

BME, VIK, TMIT

HUN-REN-BME Felhőalkalmazások Kutatócsoport

Prof. Dr. Jelasity Márk

SZTE

Válaszok Prof. Dr. Jelasity Márk, az „Erőforrás-kezelés felhőrendszerekben” című MTA doktori értekezésemre adott **bírálatára**

Tisztelt Professzor Úr,

Köszönöm alapos és részletes értékelését a DSc értekezésemről. Az Ön észrevételei és javaslatai nagy segítséget nyújtanak a további fejlesztésekhez és a munka minőségének javításához.

Az Ön által feltett kérdésekre az alábbiakban válaszolok.

1.1 Igen, korrekt a modell értelmezése és a megvilágított tulajdonsága.

1.2 A kapcsolatok kialakításának költségét hivatott jelezni az alfa paraméter, ám az valóban ugyanannyi minden potenciális szolgáltató között, ebben az értelemben homogén feltevéssel élek a modellben. Bevallom ez lehet eltúlzott egyszerűsítés, ám jelentősen hozzájárul ahhoz, hogy az eredmények zárt formulában kifejezhetők legyenek.

1.3 A hyperscale felhőszolgáltatók piaci részesedésének növekedésével és a mobilhálózati szolgáltatók konszolidációjával a „peering” kapcsolatok jelentősége folyamatosan növekszik, ezért véleményem szerint a valóságban is hasonló modellek mentén születnek üzleti döntések.

1.4 A kérdés egy komoly tervezési dilemmára világít rá. Amennyiben nem feltételeznénk véletlenszerűséget a felhasználói viselkedésben, a modell gyorsan egy Bertrand játékra redukálódna. Ráadásul a jelenlegi feltevéseknél véleményem szerint rosszabb opció lenne olyan elfogyó kapacitást feltételezni, aminek tudatában van az egyszeri felhasználó. A modell megalkotásánál ezért határoztam a jelenlegi feltevések mellett, amelyeknek az egyszerűsítő jellegét természetesen elismerem. Mindazonáltal fontosnak tartom kiemelni, hogy az analitikusan levezett eredmények érdekes és intuitív konklúziókat nyújtanak a hálózati szolgáltatók szerepéről és árazási lehetőségeiről.

2.1 Az említett feltételezés hiányában jelentősen megnehezedik a javasolt algoritmus megoldásának approximációs jellemzése, illetve a viszonylag alacsony korlát bizonyítása. Érdemes lehet jövőbeni kutatásban megvizsgálni, hogy kapacitáskorlátok bevezetésével milyen nehézzé válik a probléma. Ugyanakkor megjegyezném, hogy a megengedett késleltetés szempontjából a feltételezés kifejezetten arra céloz, hogy a lehető legnehezebb feladat legyen a megbízhatóság biztosítása. A peremfelhő rendszerek elsődleges célja a gyakorlatban ugyanis az alacsony késleltetés nyújtása, és csak másodlagos jellemzőjük a kvázi kifogyhatatlan számítási kapacitás (a nagy adatközpontokkal ellentétben) hiánya.

2.2 Az offline ütemezési problémát fel lehetne írni ILP formájában, ez egy remek publikációs lehetőséget rejtő kérdés/javaslat. A 2021-ben megjelent irodalomáttekintő cikkünkben (B Sonkoly, J Czentye, M Szalay, B Németh, L Toka: Survey on placement methods in the edge and beyond, IEEE Communications Surveys & Tutorials 23 (4), 2590-2629) nem tudtunk hasonló irodalmat felsorolni, és jelenleg sem találok ilyen témájú publikációt. A gyakorlatban felmerülő problémaméretek valószínűleg lehetővé tennék pl. CPLEX solver használatát az optimális megoldás kiszámítására, hisz a késleltetéskorlátok miatt jól működik a „branch and bound" megközelítés. Megjegyezném azonban, hogy az offline ütemező módszerek jelentőségét nagyban korlátozza a felhőrendszerek használatának általában dinamikus jellege.

2.3 Legjobb tudomásom szerint más kutatók még nem dolgoztak peremfelhőrendszerek megbízhatóságát javító módszereken, ezért nem találtam algoritmust, trace-t és szimulátort amivel összehasonlító elemzést lehetett volna végezni.

2.4 Természetesen csak az adott értelmezési tartományon belül látható a kvadratikus függvénnyel közelített összefüggés, szélesebb tartományon még a legjobban megtervezett és implementált felhő-natív alkalmazások is szublineáris teljesítményt mutatnak a méretezés függvényében. A kérdéses esetben az alacsony pod számoknál meglepően jó, míg 10 pod esetén kifejezetten kiváló teljesítményt lehetett mérni, ez okozza a meglepő ábrát.

2.5 Így igaz, a 3.12 ábrán a trade-off = 100 eset mutatja, hogy ha a felhő erőforrások relatíve drágák az SLA megsértések által okozott költségekhez képest, akkor a HPA+ kissé a HPA fölött van az összköltséget tekintve. A többi arányszámot nézve látszódik a jelentős költségcsökkenés, amelyet a SLA megsértések szignifikáns csökkentésével ér el a javasolt módszer. A HPA paraméterezéséhez képest a HPA+ az előrejelző gépi tanulási modellekre építve éri el ezt a képességet, azaz proaktív, amíg a HPA reaktív működési elvű.

2.6 A mikroszolgáltatásokhoz javasolt analitikai modell, ami a nagy léptékű alkalmazásfutási időszakok alatti erőforrás-többlet és az alkalmazás több skálázási egységbe rendezéséből fakadó késleltetési többlet közötti döntési helyzetet írja le, nem veszi figyelembe a skálázási egységek kommunikációs igényét. Ez valóban jelentős egyszerűsítés és komolyan befolyásolhatja a modell gyakorlati felhasználhatóságát. A kapcsolódó cikkemben a disszertációban megfogalmazottnál hosszabban tárgyalom ezt a hiányosságot, ahol is azt a feltételezést hozom mentségként, hogy várhatóan a hasonló léptékben felskálázott modulok között történik a megnövekedett mennyiségű kommunikáció, amelyek amúgy is ugyanabba a skálázási egységbe tartoznak, tehát hozzáadott késleltetés nélkül kommunikálnak.

3.1 Véleményem szerint a javasolt elosztott módszer jól szolgálja a fair erőforráskiosztás célját: a szolgáltató minden felhasználónak kiosztja a zsetonokat, és utána engedi, hogy a szükségeiknek megfelelően foglalják le vele kellő időben és kellő mértékben a korlátos erőforrást. Minden központilag történő ütemezési döntésnél vagy az igények őszinte bevallásán alapulhat a kiosztás, vagy pedig egy olyan rendszerszintű cél elérésére törekvő algoritmussal, amely nem feltétlenül veszi figyelembe a pillanatnyi felhasználói igények fontossága közötti különbségeket.

3.2 Vannak olyan esetek, amelyekben ez a feltevés kifejezetten megalapozott: pl. a javasolt módszert squash pályákon felszerelt kamerákra alkalmazva viszonylag könnyen előrejelezhető, hogy a labda mikor és a fal melyik részén fog pattanni. Hasonló helyzetet tudok elképzelni közúti forgalomfigyelés és biztonsági kamerák esetén. Természetesen ha a kamerával megfigyelt esemény nagy rendezetlenségű, akkor jogosak a bírálói fenntartások.

A kapott visszajelzések alapján fogom tovább folytatni kutatásaimat, figyelembe véve az Ön javaslatait és észrevételeit. Köszönöm az értékes inputot és a részletes értékelést.

Üdvözlettel,

Toka László

Budapest, 2024. 05. 13.