6. gyakorlat – Teljesítménymodellezés – Megoldások

Dimenzióanalízis

A teljesítménymodellezés feladatok megoldása során érdemes a fizikából ismert dimenzióanalízist[1] elvégezni. Pl. a négyzetes úttörvény:

$$s = v_0 t + \frac{a}{2} t^2 \tag{1}$$

Dimenziókkal:

$$s[\mathrm{m}] = v_0 \left[\frac{\mathrm{m}}{\mathrm{s}}\right] t[\mathrm{s}] + \frac{a}{2} \left[\frac{\mathrm{m}}{\mathrm{s}^2}\right] t^2[\mathrm{s}^2] = v_0 t[\mathrm{m}] + \frac{a}{2} t^2[\mathrm{m}] \tag{2}$$

A dimenzióhasználat fő motivációja, hogy ha a dimenziók nem stimmelnek, akkor a képletet is biztosan elrontottuk valahol.[2] A dimenzióanalízis gyakran segít a megfelelő képlet kiválasztásában. Fontos, hogy a "darab", "kérés" stb. jellegű mértékegységek nem számítanak külön dimenziónak, ezért pl. a $\frac{\text{kérés}}{\mathrm{s}}$ és az $\frac{1}{\mathrm{s}}$ dimenziók megegyeznek.

Alapképletek

Little-törvény:

$$N = X \cdot T \tag{3}$$

$$N\ [1] = X \left[\frac{1}{s}\right] \cdot T\ [s] \tag{4}$$

Kihasználtság intuitíven és a Little-törvényből *egyetlen kizárólagos* erőforráspéldány esetén:

$$U = \frac{X}{X_{\max}} = \frac{T_{\text{busy}}}{T_{\text{measured}}} = N = X \cdot T \tag{5}$$

Átbocsátóképesség végrehajtási időből *egyetlen kizárólagos* erőforráspéldány esetén (az átbocsátóképesség az elérhető legnagyobb átbocsátás, vagyis ilyenkor a kihasználtság 100%):

$$X = \frac{U}{T} \Rightarrow X_{\max} = \frac{1}{T} \tag{6}$$

1. feladat

Egy diszk 50 kérést szolgál ki másodpercenként. Minden kérés kiszolgálása 0,005 másodpercet vesz igénybe. A rendszerben nincs átlapolódás.

a. Mekkora a kihasználtság?
b. Mekkora a maximálisan kiszolgálható terhelés (érkezési ráta)?

Megoldás

a. Az erőforrás kihasználtsága $U = X \cdot T$, ahol X az átlagos átbocsájtás és T az átlagos kiszolgálási idő. Tehát $U = 0,25$, így 25%-os a kihasználtság.

   A feladat józan ésszel is megoldható: a diszknek másodpercenként 50 kérés $\cdot$ 0,005 $\frac{\mathrm{s}}{\text{kérés}}$-t kell dolgoznia. Ha másodpercenként 0,25 másodpercet dolgozik, akkor 25% a kihasználtsága.

b. Ekkor a kihasználtság $U = 1$. Ekkor $X_{\max} = \frac{U}{T} = 200\ \frac{\text{kérés}}{\mathrm{s}}$. Vagyis a szabály egyetlen, átlapolódásmentes feldolgozó egységre: $X_{\max} = \frac{1}{T} = \frac{1}{0,005\ \mathrm{s}} = 200\ \frac{\text{kérés}}{\mathrm{s}}$.

[1] Dimenzióanalízis (Wikipédia)
[2] Ajánlott olvasmány: what if? – Droppings

2. feladat

Egy szerveren az alábbi teljesítményjellemzőket mértük:

Mintavétel időpontja [ms]	500	600	700	800	900
Utolsó 100 ms alatt feldolgozott kérések száma [darab]	11	12	21	18	20
Utolsó 100 ms átlagos kiszolgálási ideje [ms]	15	20	21	25	27
Utolsó 100 ms CPU kihasználtság [%]	12	13	16	17	19
Utolsó 100 ms HDD I/O kihasználtság [%]	55	63	87	61	73

a. A rendelkezésre álló adatok alapján a szerver melyik erőforrása tűnik a szűk keresztmetszetnek?
b. Ezen 5 mérés alapján milyen becslést tudunk adni az egyszerre kiszolgálás alatt lévő kérések átlagos számára?

Megoldás

a. A HDD kihasználtsága a legnagyobb. A terhelés felskálázásával először a HDD fog telítődni.
b. Az utolsó 100 ms alatt feldolgozott kérések számából és az átlagos kiszolgálási időből adódik. Mivel az átlagos kiszolgálási idő különböző elemszámú adathalmazokból került kiszámításra, egyszerű átlagolásuk helyett a feldolgozott kérésekkel súlyozott átlagukat kell vennünk.

$$T = \frac{\sum_{i=1}^{n} k_i t_i}{\sum_{i=1}^{n} k_i} = \frac{11 \cdot 15 + 12 \cdot 20 + 21 \cdot 21 + 18 \cdot 25 + 20 \cdot 27}{11 + 12 + 21 + 18 + 20} = 22{,}39 \text{ ms} \tag{7}$$

A rendszer egyensúlyi állapotban van, ezért a c) feladatban kiszámolt átlagos átbocsátással alkalmazhatjuk a Little-törvényt:

$$N = \bar{X} \cdot T = 164 \frac{1}{\text{s}} \cdot 22{,}39 \text{ ms} = 164 \frac{1}{\text{s}} \cdot 0{,}02239 \text{ s} = 3{,}67196 \tag{8}$$

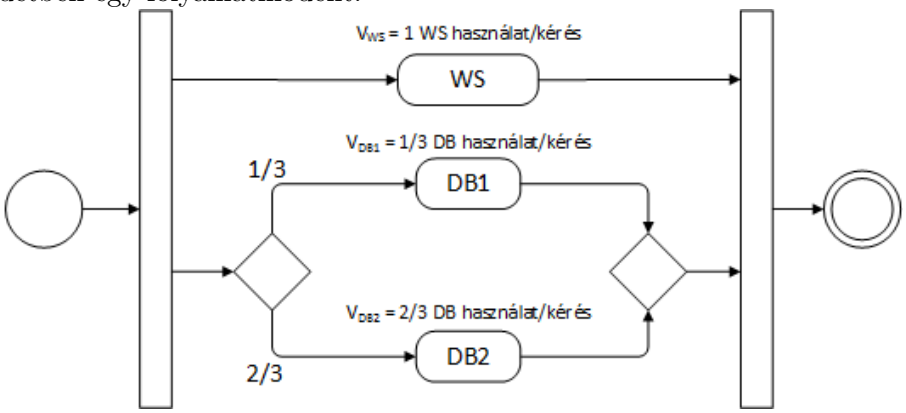
3. feladat

Adott egy webszerver (WS) és két fürtözött adatbázisszerver (DB1, DB2). A két adatbázis szerver közt súlyozott round robin terheléselosztás alapján választunk, 1:2 arányban. Minden felhasználói kérés kiszolgálása során mindkét fajta erőforrást használjuk. A csúcsidőszakban 30 percig monitorozzuk a rendszert, ezalatt 9000 kérést szolgál ki. A szerveken mért foglaltsági idők: WS – 1350 s CPU idő; DB1 – 810 s, DB2 – 1320 s diszk IO idő.

a. Mekkora az egyes szerverek jelenlegi átbocsátása?
b. Mennyi időt töltenek egy-egy kérés kiszolgálásával a szerverek?
c. Mekkora a rendszer maximális áteresztőképessége?
d. Miért nem egyféle foglaltsági időt vettünk figyelembe a két erőforrástípusnál?
e. Hol csal még így is a modell?

Megoldás

Rajzoljunk kezdetben egy folyamatmodellt!



Mivel a feladatban nem volt egyéb megkötés, azt feltételeztük, hogy a kérések kiszolgálása a különböző erőforrásokon párhuzamosan történik. Ehelyett a modell lehetne szekvenciális is (az átbocsátás szempontjából nincs különbség, *de a végrehajtási időben igen*!), viszont az előbbi általánosabb, hiszen a WS használata átlapolódhat az adatbázis használatával. A valóságban persze a WS az adatbázishívás előtt és után is dolgozik, sőt, időnként még közben is. A mostani modell azt fejezi ki, hogy – pontos információ híján – ezeket a szakaszokat aggregáljuk és elfelejtjük, hogy milyen sorrendben futottak (absztrakció!).
Emlékeztető: A vizitációs számmal (többek között) a rendszer és a komponensek átbocsátása és átbocsátóképessége között tudunk váltani. Ha átbocsátással dolgozunk, akkor rendszerint a rendszer átbocsátásából számítjuk a komponensek átbocsátását – ilyenkor a vizitációs számmal szorozni kell, hiszen minden rendszerbe belépő tokent átlagosan annyiszor kell feldolgoznia a komponenseknek, mint amennyi a vizitációs szám. Ha átbocsátóképességet szeretnénk számolni, akkor rendszerint a komponensek (egyszerűen számítható) átbocsátóképességéből kiindulva határozzuk meg a rendszer átbocsátóképességét – ilyenkor a vizitációs számmal osztani kell, hiszen ha minden belépő tokent annyiszor kell feldolgoznia a rendszernek, mint amennyi a vizitációs szám, akkor annyival kevesebb token érkezhet a rendszerbe túltelítődés nélkül. Ne feledjük, hogy (többek között a szűk keresztmetszetek miatt) ebben az irányban nem elegendő a vizitációs számmal számolni, gyakran szükség van a számított értékeken végzett egyéb számításokra (pl. minimumképzésre)

a. Számoljunk először a rendszerre, aztán az erőforrásokra! A feldolgozott kérések száma $C = 9000$ (“Count”), a mérés ideje $T_m = 30$ min.

   - $X_{\text{rendszer}} = \frac{C}{T_m} = \frac{9000 \text{ kérés}}{30 \text{ min}} = \frac{9000}{1800} \frac{\text{kérés}}{\text{s}} = 5 \frac{\text{kérés}}{\text{s}}$
   - $X_{\text{WS}} = X_{\text{rendszer}} \cdot v_{\text{WS}} = 5 \frac{\text{kérés}}{\text{s}} \cdot 1 = 5 \frac{\text{kérés}}{\text{s}}$
   - $X_{\text{DB1}} = X_{\text{rendszer}} \cdot v_{\text{DB1}} = 5 \frac{\text{kérés}}{\text{s}} \cdot \frac{1}{3} = 1{,}666 \frac{\text{kérés}}{\text{s}}$
   - $X_{\text{DB2}} = X_{\text{rendszer}} \cdot v_{\text{DB2}} = 5 \frac{\text{kérés}}{\text{s}} \cdot \frac{2}{3} = 3{,}333 \frac{\text{kérés}}{\text{s}}$

b. Az egyes erőforrásokra (B a mért foglaltsági idő, “Busy time”, az egyes szerverek pedig $C \cdot v_i$ kérést dolgoznak fel):

   - $T_{\text{WS}} = \frac{B_{\text{WS}}}{C \cdot v_{\text{WS}}} = \frac{1350 \text{ s}}{9000 \text{ kérés}} = 0{,}15 \frac{\text{s}}{\text{kérés}}$
   - $T_{\text{DB1}} = \frac{B_{\text{DB1}}}{C \cdot v_{\text{DB1}}} = \frac{810 \text{ s}}{3000 \text{ kérés}} = 0{,}27 \frac{\text{s}}{\text{kérés}}$
   - $T_{\text{DB2}} = \frac{B_{\text{DB2}}}{C \cdot v_{\text{DB2}}} = \frac{1320 \text{ s}}{6000 \text{ kérés}} = 0{,}22 \frac{\text{s}}{\text{kérés}}$

c. A rendszer maximális átbocsátóképessége az a legnagyobb átbocsátás, amivel egyik komponensbe sem érkezik több kérés, mint annak átbocsátóképessége. Ennek megfelelően pl. a DB1 ágra

$$X_{\text{rendszer}} \cdot v_{\text{DB1}} \leq X_{\text{DB1}}^{\max} \Rightarrow X_{\text{rendszer}} \leq \frac{1}{v_{\text{DB1}}} \cdot X_{\text{DB1}}^{\max}. \tag{9}$$

   Ugyanígy DB2-re és WS-re:

$$X_{\text{rendszer}} \leq \frac{1}{v_{\text{DB2}}} \cdot X_{\text{DB2}}^{\max} \tag{10}$$

$$X_{\text{rendszer}} \leq \frac{1}{v_{\text{WS}}} \cdot X_{\text{WS}}^{\max} = X_{\text{WS}}^{\max}. \tag{11}$$

   Mivel DB1 és DB2 *kötött arányú választás* (hosszú távon gyakorlatilag olyan, mintha minden “munkát” 1:2 arányban szétbontanánk és továbbküldenénk, tehát ilyen szempontból a fork-join és a szabad választás[3] közé tehető), ezért a számított értékek minimuma érkezhet meg a decision csomóponthoz túltelítés nélkül:

$$X_{\text{rendszer}} \leq \min\left(\frac{1}{v_{\text{DB1}}} \cdot X_{\text{DB1}}^{\max}, \frac{1}{v_{\text{DB2}}} \cdot X_{\text{DB2}}^{\max}\right). \tag{12}$$

[3] A szabad választású döntés akármelyik irányba továbbküldheti a kérést, tehát ha az egyik ág telítésben van, nyugodtan választhatja a másikat (a kötött arányú nem). Emiatt szabad választásnál az átbocsátóképességek összeadódnak.

A fork mindig mindkét irányba továbbküldi a kérést, és mindkét irányba a "teljes munkát" továbbítja, tehát az elágazásra számított érték és a WS-re számított érték közül a kisebb lehet a rendszer átbocsátóképessége. Ezalapján a (11) és (12) egyenlőtlenségekből a maximális átbocsátás, vagyis az átbocsátképesség képlete:

$$X^{\max}_{\text{rendszer}} = \min\left(X^{max}_{\text{WS}}, \frac{1}{v_{\text{DB1}}} X^{max}_{\text{DB1}}, \frac{1}{v_{\text{DB2}}} X^{max}_{\text{DB2}}\right) \tag{13}$$

A feladat megoldásához tehát a komponensek átbocsátóképességeit kell kiszámolnunk:

- $X^{\max}_{\text{WS}} = \frac{1}{T_{\text{WS}}} = \frac{1}{0{,}15\ \frac{\text{s}}{\text{kérés}}} = 6{,}666\ \frac{\text{kérés}}{\text{s}}$
- $X^{\max}_{\text{DB1}} = \frac{1}{T_{\text{DB1}}} = \frac{1}{0{,}27\ \frac{\text{s}}{\text{kérés}}} = 3{,}704\ \frac{\text{kérés}}{\text{s}}$
- $X^{\max}_{\text{DB2}} = \frac{1}{T_{\text{DB2}}} = \frac{1}{0{,}22\ \frac{\text{s}}{\text{kérés}}} = 4{,}545\ \frac{\text{kérés}}{\text{s}}$

A rendszer maximális átbocsátóképessége ezekből:
$X^{\max}_{\text{rendszer}} = \min\left(6{,}666\ \frac{\text{kérés}}{\text{s}}, 3 \cdot 3{,}704\ \frac{\text{kérés}}{\text{s}}, \frac{3}{2} \cdot 4{,}545\ \frac{\text{kérés}}{\text{s}}\right) = X^{\max}_{\text{WS}} = 6{,}666\ \frac{\text{kérés}}{\text{s}}$.
Érdemes megfigyelni, hogy a minimum a WS-en esett, de a DB2-höz tartozó érték ($6{,}818\ \frac{\text{kérés}}{\text{s}}$) szintén nagyon közel van. A szűk keresztmetszet tehát jelenleg a webszerver, de csak ennek a komponensnek a fejlesztésével vagy többszörözésével csak korlátozott mértékben növelhető a teljesítmény, mert nagyon hamar a DB2 válik majd szűk keresztmetszetté.

d. Azért, mert mind a DB szerver, mind a WS egy-egy kis rendszer önmagában is, és belül a diszk I/O, ill. a CPU bizonyul szűk keresztemetszetnek jelen esetben. Más rendszerben, más feladatot végrehajtva lehet, hogy az egyik erőforrás hálózati linkje, míg a másik erőforrás RAM sávszélessége fog szerepelni. Vegyük észre, hogy ez egy absztrakció, melynek célja a számítások egyszerűsítése a nem (vagy kevésbé) releváns adatok eltávolításával, ami abból indul ki, hogy az elhanyagolt adatok hatása a megtartott adatokénál jóval kisebb (itt: a webszerver memóriája vagy merevlemez sávszélessége sokkal később telítődne, mint a processzora, de ezt már el sem érjük, ha a processzor miatt vergődik a rendszer). Egyúttal emlékezzünk vissza a 2. feladat b) részére, ahol adatelemzéssel állapítottuk meg a potenciális szűk keresztmetszetet, vagyis a skálázódás és telítődés szempontjából legmeghatározóbb adatot.

e. Több egyszerűsítéssel is éltünk, pl.

- lineáris skálázódást feltételeztünk, holott a valós rendszerek ennél általában rosszabbul skálázódnak (ráadásul telítődés közelében hajlamosak leromlani),
- nem vettük figyelembe a valódi rendszerben előforduló összes erőforrást (lásd előző feladat),
- feltétleztük, hogy a kéréseket statikus módon elosztva tökéletes terheléselosztást kapunk, holott ez általában nem igaz: az átlagos értékek hosszú távon a számított módon alakulnak, de rövidebb időszakokra nézve egy átlagosnál hosszabb végrehajtási idejű kérés például rövid időre telítésbe viheti a rendszert.

4. feladat

Egy sziget lakói minden reggel munkába menet átkelnek a szigetet ölelő tavon. Észak felé híd vezet, dél felé autóskomp. Az irányonként egysávos híd 200 m hosszú, és 60 km/h sebességgel szabad rajta haladni, a követési távolság (hátsó lámpától hátsó lámpáig 30 m) betartása mellett. A négy komphajó egyenként 15 percenként teszi meg a sziget-szárazföld-sziget kört, és így óránként négyen együtt legfeljebb 800 autót tudnak átvinni a szárazföldre.

a. Mekkora a híd átbocsátóképessége (észak felé)?
b. Hány autó fér el egy kompban?

c. A reggeli csúcsforgalomban mekkora a szigetet elhagyó két útvonal együttes átbocsátóképessége?
d. Ha délben a szárazföldi főutat baleset miatt lezárták, és a szigeten keresztül (a hídon, majd a kompon átkelve) terelik a forgalmat, mekkora a terelőútvonal átbocsátóképessége?
e. Valamelyik reggel 7:00 és 8:30 között 900 autó hagyta el a szigetet komppal. Mennyi volt ebben az időszakban a kompok átbocsátása és kihasználtsága?
f. A fenti mérésben átlagosan hány autó állt sorba egyszerre a parton, ha az autók jól időzítve, átlagosan fél perccel a beszállásuk előtt érkeztek kompkikötőhöz?

Megoldás

a. Little törvényében az átbocsátás szerepel, nem az átbocsátóképesség – de abban a speciális esetben, amikor pont telítve van a rendszer, a kettő megegyezik:

- $N = X \cdot T \to X = \frac{N}{T}$;
- $N = \frac{200\ \text{m}}{30\ \text{m/kocsi}} = \frac{20}{3}$ kocsi;
- $T = \frac{200\ \text{m}}{60\ \text{km/h}} = \frac{0{,}2\ \text{km}}{60\ \text{km/h}} = \frac{0{,}2}{60}$ h; tehát
- $X = \frac{20/3}{0{,}2/60} = 2000\ \frac{\text{kocsi}}{\text{h}} = X^{max}$.

b. Az előzőhöz hasonlóan Little törvényéből az átbocsátóképesség:

- $N = X \cdot T; X = 800\ \frac{\text{kocsi}}{\text{h}}$;
- $T = 15\ \text{min} = 0{,}25$ h; ekkor $N = 200$, tehát egyszerre 200 autó utazik. Mivel 4 hajó van, ezért egy hajóra 50 kocsi fér fel.

c. Az együttes átbocsátóképesség a két átbocsátóképesség összege. A hídon egy irányba óránként 2000 kocsi haladhat át, tehát 2000 $\frac{\text{kocsi}}{\text{h}}$ a híd átbocsátóképessége. A kompok óránként 800 autót visznek át, tehát az átbocsátóképesség 2800 $\frac{\text{kocsi}}{\text{h}}$ egy irányba.
d. A terelőút átbocsátóképessége (soros kompozíció): $X = \min(X_{\text{híd}}, X_{\text{komp}}) = 800\ \frac{\text{kocsi}}{\text{h}}$.
e. Átbocsátás: $X = \frac{K}{T} = \frac{900}{1{,}5} = 600\ \frac{\text{kocsi}}{\text{h}}$.Kihasználtság: $U = \frac{X}{X^{\max}} = \frac{600\ \frac{\text{kocsi}}{\text{h}}}{800\ \frac{\text{kocsi}}{\text{h}}} = 0{,}75 = 75\%$.
f. Komphoz sorbanállásra Little-törvény: $N = X \cdot T = 0{,}5\ \text{min} \cdot 600\ \frac{\text{autó}}{h} = 5$ autó.

5. feladat

Vállalatunk nyilvános szakmai tudástára egymásra is hivatkozó szócikkeket kínál a cég termékeit világszerte használó ügyfeleknek. Egyetlen szócikk lekérésének kiszolgálásához a szervert átlagosan 60 ms-ig veszi igénybe. A szócikk megtekintése után az olvasó csak az esetek 30%-ában hagyja el az oldalt, többnyire ugyanis egy újabb szócikkre mutató hivatkozásra kattint.

a. Egy olvasó összes tudásszomjának kielégítéséhez átlagosan mekkora szerveridő szükséges?
b. Tekintsük úgy, hogy az egyes kérések a szerveren nem párhuzamosíthatóak. Óránként hány egyedi látogatót képes kiszolgálni a szerver?

Megoldás

a. Egy szócikk lekérésének kiszolgálása átlagosan 60 ms, egy felhasználó pedig átlagosan $v = \frac{1}{0{,}3}$ szócikket tekint meg,[4] tehát $T = 60\ \frac{\text{ms}}{\text{szócikk}} \cdot \frac{1}{0{,}3}\ \frac{\text{szócikk}}{\text{felhasználó}} = 200\ \frac{\text{ms}}{\text{felhasználó}}$. A v most is a vizitációs szám.

b. Maximális eset, amikor a kihasználtság 100%, azaz $U = 1$. Ekkor $U = X \cdot T \to X = \frac{U}{T} = \frac{1}{0{,}2} = 5\ \frac{\text{látogató}}{\text{s}}$. Óránként 3600 s $\cdot\ 5\ \frac{\text{kéréslátogató}}{\text{s}} = 18000$ látogató.

[4] Geometriai eloszlás várható értéke (Wikipédia)

6. feladat

Egy adatbázis szervert 15 percig monitorozunk. Ez alatt az idő alatt a szerver processzora 12 percig volt foglalt. Azt figyeltük meg, hogy minden tranzakció általában kétszer használta a processzort, és átlagosan 1 ms ideig használatonként (és ezalatt teljesen lefoglalja a CPU-t, nincs párhuzamosítás). Mekkora a rendszer átbocsátása és áteresztőképessége?

Megoldás

Kis segítség: a rendszer és a komponensek átbocsátóképessége közötti viszonyt írja le:

$$X_k = C_k/T = C_k/C_0 \cdot C_0/T = V_k \cdot X_0 \tag{14}$$

$$V_k = C_k/C_0 \tag{15}$$

A visit ratio (használati arány), azt mondja meg, hogy a rendszer szintű átbocsátás hogy aránylik a komponens átbocsátásához.
Tehát: $T = 15$ perc; foglaltsági idő, busy time: $B_{\text{CPU}} = 12$ perc; Visit Ratio: $V_{\text{CPU}} = 2$ CPU használat/tranzakció; $S_{\text{CPU}} = 1$ ms/CPU használat
A processzor kihasználtsága: $U_{\text{CPU}} = B_{\text{CPU}}/T = 0{,}8$
Forced Flow törvény: $X_{\text{CPU}} = V_{\text{CPU}} X_0$
Mi a rendszer átbocsátását keressük, tehát átrendezve $X_0 = X_{\text{CPU}}/V_{\text{CPU}}$, majd pedig a kihasználtság törvényét felhasználva

$$X_0 = X_{\text{CPU}}/V_{\text{CPU}} = (U_{\text{CPU}}/S_{\text{CPU}})/V_{\text{CPU}} = 0{,}8/\frac{0{,}001 \text{ s/CPU használat}}{2 \text{ (CPU használat/tranzakció)}} = 400 \text{ tranzakció/s} \tag{16}$$

A rendszer áteresztőképessége: 500 tranzakció/s, hiszen 1 tranzakció 2 ms-ig foglalja a CPU-t.

7. feladat

Legalább hány aktív szálat kell engedélyeznünk egy webszerveren alkalmazásunknak, ha az egyenletes terhelés melletti teljesítményét nem szeretnénk visszafogni? Szálkorlát nélküli mérésekkel megállapítottuk, hogy egy kérés átlagosan 120 ezredmásodpercet tölt a rendszerben, és a szerver másodpercenként 50 felhasználót szolgál ki.

Megoldás

Little törvényét használva $N = X \cdot R = 50$ felhasználó/s $\cdot$ 0,120 s/felhasználó $= 6$, tehát átlagosan 6 kérés van a rendszerben, tehát 6 szálat kell indítanunk, hogy ne fogjuk vissza a teljesítményt. (Burst esetén nyilván több szálra van szükség.)

8. feladat

Internetes közösségi oldalt működtetünk. Az utóbbi időben számottevően megnőtt a népszerűsége, de ezáltal a válaszidő is kellemetlenül megnyúlt. Az üzleti cél, hogy csúcsidőszakban egyszerre 1500 felhasználót átlagosan négy másodperces válaszidővel szolgáljon ki a honlap.

a. Minimálisan mekkorára kell tervezni a kiszolgáló infrastruktúra átbocsátóképességét, ha az azon kívüli késleltetés (hálózati forgalom, HTML megjelenítés a kliensoldalon) egy másodpercnek becsülhető?
b. Az újratervezett weboldalon a mérések szerint egyetlen kérés kiszolgálása átlagosan 20 ms CPU-időt igényel a webszerveren, és 12,5 ms erejéig foglal le egy adatbázisszervert. Jelenleg 15 webszerver fogadja a kéréseket és az adatbázis 5 kiszolgálóra van replikálva. Lineáris skálázhatóságot feltételezve, milyen számítógépből és mennyit kell még legalább venni, hogy a fenti cél teljesülhessen?

c. A kibővített rendszerben mekkora lesz az egyes szervertípusok kihasználtsági aránya? Ha az a cél, hogy még a csúcsidőszakban is legfeljebb 50%-os legyen a kihasználtság, meddig kellene még bővíteni a rendszert?

Megoldás

a. Használjuk Little törvényét! $N = X \cdot R$, ahol $N = 1500$ felhasználó, $R = 3$ s (4 s teljes válaszidő – 1 s általános késleltetés), $X = N/R = 1500$ kérés/3 s $= 500$ kérés/s

b. Az átbocsátóképességek:

- **WS:** Egy szerver átbocsátóképessége: $U = X \cdot S; X = U/S = 1/0{,}02 = 50$ kérés/s (libasorszabállyal), tehát összesen 10 webszerverre van szükség, hogy másodpercenként 500 kérést kiszolgáljunk. Mivel 15 van, ez elég.
- **DB:** Egy adatbázis átbocsátóképessége: $U = X \cdot S; X = U/S = 1/0{,}0125 = 80$, tehát összesen $500/80 = 6{,}25$, azaz legalább 7 adatbázisra van szükség, hogy kiszolgáljunk másodpercenként 500 kérést. Azaz még 2 darab adatbázist kell vennünk az 5 mellé.

c. A kihasználtságok:

- **WS:** $U_{\mathrm{WS}}(= X \cdot S) = 10$ webszerver (már kiszámoltuk). $10/15 = 0{,}666 = 66{,}6\%$
- **DB:** $U_{\mathrm{DB}}(= X \cdot S) = 6{,}25$ db/(7 db) $= 0{,}892 = 89{,}2\%$ (itt a "db" az adatbázis szerverek száma).

  Meddig kellene bővíteni a rendszert?
- **WS:** W $= 10/0{,}5 = 20$ (már kiszámoltuk). Tehát összesen 20 webszerverre van szükségünk.
- **DB:** DB $= 6{,}25/0{,}5 = 12{,}5$. Tehát összesen 13 adatbázisra van szükség.