6. gyakorlat – Teljesítménymodellezés és adatelemzés – Megoldások

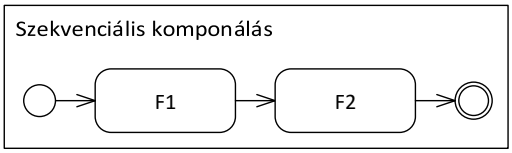
Elméleti összefoglaló

Dimenzióanalízis. A teljesítménymodellezés feladatok megoldása során érdemes a fizikából ismert dimenzióanalízist[1] elvégezni. Idézzük fel a négyzetes úttörvény: $s = v_0 t + \frac{a}{2} t^2$
Dimenziókkal: $s[\mathrm{m}] = v_0 \left[\frac{\mathrm{m}}{\mathrm{s}}\right] t[\mathrm{s}] + \frac{a}{2} \left[\frac{\mathrm{m}}{\mathrm{s}^2}\right] t^2[\mathrm{s}^2] = v_0 t[\mathrm{m}] + \frac{a}{2} t^2[\mathrm{m}]$
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: $N = X \cdot T$, $N\ [1] = X \left[\frac{1}{s}\right] \cdot T\ [s]$
Kihasználtság intuitíven és a Little-törvényből *egyetlen kizárólagos* erőforráspéldány esetén:
$U = \frac{X}{X_{\max}} = \frac{T_{\text{busy}}}{T_{\text{measured}}} = N = X \cdot T$
Átbocsátóképesség végrehajtási időből *egyetlen kizárólagos* erőforráspéldány esetén (az átbocsátóképesség az elérhető legnagyobb átbocsátás, vagyis ilyenkor a kihasználtság 100%): $X = \frac{U}{T} \Rightarrow X_{\max} = \frac{1}{T}$

1. Zárthelyi javítása

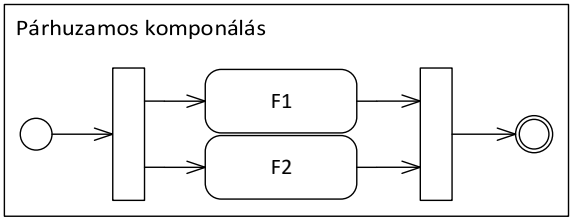
A zárthelyik megtekintése során a hallgatóknak lehetőségük van reklamálni esetleges javítási hibák miatt. Sikeres reklamáció esetén a pontszámuk módosításra kerül. Az első nagyfeladatból (F1) óránként 10 darabot képes átnézni egy javító, a második nagyfeladatból (F2) pedig 20 darabot. Mindkét feladathoz tartozik 1-1 javító, akik az adott feladatot javították. A továbbiakban készítsünk minden kérdéshez egy-egy folyamatmodellt és határozzuk meg, hogy óránként hány hallgató dolgozatát sikerül átnézni az egyes esetekben!

a) A hallgatók először az F1, majd az F2 feladatot nézetik át a javítóval.
Megoldás



Szekvenciális komponálás. A szűk keresztmetszet fogja meghatározni a teljes átbocsátó képességet, mert ott fognak feltorlódni a feladatok (hiába gyors a többi rész). Általánosan: $X^{max} = min(X_1^{max}, X_2^{max})$. Mivel $X_{F1}^{max} = \frac{10}{h}, X_{F2}^{max} = \frac{20}{h}$, F1 a szűk keresztmetszet, tehát $X^{max} = min(X_{F1}^{max}, X_{F2}^{max}) = min(\frac{10}{h}, \frac{20}{h}) = \frac{10}{h}$.

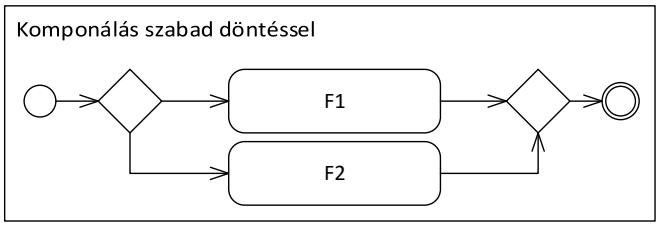
b) A leleményes hallgatók a két feladatot külön-külön már egyszerre két javítónak adják oda, mivel külön lapra voltak írva. Mit nyerünk a párhuzamosítással?
Megoldás



Párhuzamos komponálás. Mivel a feladatoknak a végén be kell várniuk egymást (szinkronizáció), ezért itt is a szűk keresztmetszet fogja meghatározni a teljes átbocsátó képességet. Általánosan: $X^{max} = min(X_1^{max}, X_2^{max})$. Tehát $X^{max} = min(X_{F1}^{max}, X_{F2}^{max}) = min(\frac{10}{h}, \frac{20}{h}) = \frac{10}{h}$.
Mit nyertünk a párhuzamosítással? Az átbocsátóképességünk változatlan maradt, de a válaszidőnk csökkent.

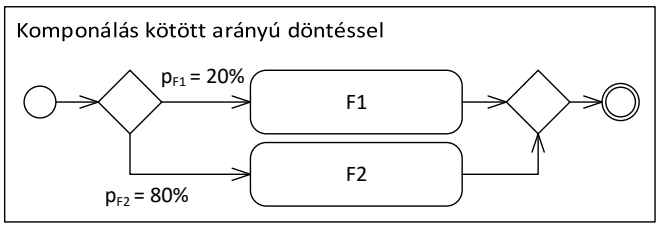
[1] Dimenzióanalízis (Wikipédia), `http://hu.wikipedia.org/wiki/Dimenzió`
[2] Ajánlott olvasmány: what if? – Droppings, `http://what-if.xkcd.com/11/`

c) A nagy tömeg miatt a hallgatók csak az egyik feladatukat nézetik át, mégpedig azt, amelyiknek a javítója éppen szabad.
Megoldás



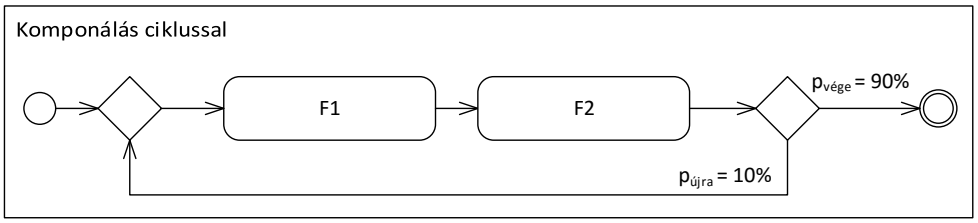
Komponálás szabad választással. Oda mennek a hallgatók ahol hely van (K db nyitott pénztár analógia). Általánosan: $X^{max} = X_1^{max} + X_2^{max}$. Tehát $X^{max} = X_{F1}^{max} + X_{F2}^{max} = \frac{10}{h} + \frac{20}{h} = \frac{30}{h}$

d) Híre ment, hogy a második feladat javítója sokkal kevésbé szigorú, így a hallgatók 80%-a inkább kivárja ennél a javítónál a sort. A maradék 20% a másik javítónál reklamál az első feladattal kapcsolatban.
Megoldás



Komponálás kötött arányú választással. Analógia: felhasználók viselkedése egy weblapon: 20% eséllyel vásárol, 80% eséllyel elvet. Általánosan: $X^{max} = min(\frac{1}{p_1} \times X_1^{max}, \frac{1}{p_2} \times X_2^{max})$, ahol p_1 és p_2 annak a valószínűsége, hogy az első, illetve a második lehetőséget választjuk ($p_1 + p_2 = 1$). Azért reciprok, mert az átlagosan F1-el töltött idő $p_1 \times T_1$, az ide eső maximális átbocsátás pedig ennek a reciproka (egy erőforráspéldány esetén). Tehát $X^{max} = min(\frac{1}{0.2} \times X_{F1}^{max}, \frac{1}{0.8} \times X_{F2}^{max}) = min(\frac{50}{h}, \frac{25}{h}) = \frac{25}{h}$.

e) A hallgatók 10%-ának a reklamáció után már csak 1-2 pont kellene a jobb jegyhez, ezért újra és újra megpróbálkoznak a reklamációval. Feltételezhetjük, hogy a hallgatók az a) részben leírt reklamációs stratégiát használják.
Megoldás



Komponálás ciklussal. Általánosan: $X^{max} = \frac{1}{\frac{1}{p_{vége}}} \times X_1^{max} = p_{vége} \times X_1^{max}$, ahol $p_{vége}$ annak a valószínűsége, hogy kilépünk a ciklusból, $\frac{1}{p_{vége}}$ pedig az iterációk várható száma (lásd később a Valószínűségszámítás tantárgyban).
Az X_1^{max} érték jelen esetben az a) részben kiszámolt érték (absztrakció), $p_{vége}$ pedig 0.9. Tehát $X^{max} = \frac{1}{\frac{1}{0.9}} \times \frac{10}{h} = 0.9 \times \frac{10}{h} = \frac{9}{h}$
Egy valós rendszerhez képest ez közelítés, mert azt feltételeztük, hogy a reklamálás itt független a reakció tartalmától.

Vizitációs szám: megmutatja, hogy a folyamat végrehajtása során átlagosan hányszor fut le az adott tevékenység/alfolyamat. Választás esetén maga a döntési valószínűség, ciklus esetén a várható iterációk száma. Átbocsátóképesség a vizitációs szám ismeretében: $X^{max} = \frac{1}{v} \times X_1^{max}$. Adott tevékenységre eső végrehajtási idő a vizitációs szám ismeretében: $T_{folyamat} = v \times T_{taszk}$.

f) Mi történne másként, ha bármelyik javító bármelyik feladatot hajlandó átnézni (de az egyes feladatok átnézése változatlan ideig tart), és így szeretnék a hallgatók az eredeti tervenek megfelelően először az F1, majd az F2 feladatot átnézetni a javítóval? **Megoldás**
Ebben az esetben a két tevékenység közös erőforrást (javító) használ, tehát nem külön-külön van felső korlátunk az átbocsátóképességükre, hanem együttesen. (Ha több erőforrás lenne, akkor az erőforrások közül kellene a szűk keresztmetszetet kikeresni, nem a tevékenységek közül.)
Az első tevékenység 6 percre, a második 3 percre foglal le egy erőforráspéldányt (javítót), tehát egy folyamatpéldánnyal (hallgatóval) összesen 9 percnyi munkája van az erőforrásnak. A két erőforráspéldány összesen óránként 120 percet tud dolgozni, tehát $\frac{120}{9} = \frac{40}{3}$ hallgató/óra a rendszer átbocsátóképessége.

2. Diszk teljesítménye

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

a) Mekkora a maximálisan kiszolgálható terhelés (érkezési ráta)?
Megoldás
Maximális terhelés mellett a kihasználtság $U = 1$. Ekkor $X_{\max} = \frac{U}{T} = 200\ \frac{\text{kérés}}{\text{s}}$. Vagyis a szabály egyetlen, átlapolódásmentes feldolgozó egységre: $X_{\max} = \frac{1}{T} = \frac{1}{0{,}005\ \text{s}} = 200\ \frac{\text{kérés}}{\text{s}}$.

b) Mekkora a kihasználtság?
Megoldás
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 $\frac{\text{s}}{\text{kérés}}$-t kell dolgoznia. Ha másodpercenként 0,25 másodpercet dolgozik, akkor 25% a kihasználtsága.

3. Szerver teljesítménye

Egy szerveren az alábbi teljesítményjellemzőket mértük (a táblázatban az első öt mérés értéke látható):

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

a) A rendelkezésre álló adatok alapján a szerver melyik erőforrása tűnik a szűk keresztmetszetnek?
Megoldás
A HDD kihasználtsága a legnagyobb. A terhelés felskálázásával először a HDD fog telítődni.

b) Az első mintavétel idején mekkora az átbocsátási ráta értéke? Az 5 mintavétel alapján mekkora az átbocsátási ráta tapasztalati átlaga és mediánja? Mi tartozik a 40%-os kvantilisbe?
Megoldás

A mintavételi időkből látszik, hogy két mintavétel között 100 ms telik el. Ebből

$$X_1 = \frac{k_1}{\Delta t} = \frac{11\ \text{kérés}}{100\ \text{ms}} = \frac{11\ \text{kérés}}{100\ \text{ms}} \left[\frac{1000\ \text{ms}}{1\ \text{s}}\right] = 110\ \frac{\text{kérés}}{\text{s}}.$$

A tapasztalati átlag kiszámítása történhet a másik négy átbocsátás kiszámításával és átlagolással, vagy a következő módon (kihasználva, hogy Δt végig 100ms):

$$\bar{k} = \frac{\sum_{i=1}^{n} k_i}{n} = \frac{11 + 12 + 21 + 18 + 20}{5} = 16{,}4$$

Ebből az átlagos átbocsátás $\bar{X} = \frac{\bar{k}}{\Delta t} = \frac{16{,}4}{0{,}1} = 164\ \frac{\text{kérés}}{\text{s}}$.
Az elemek sorba állítva $11, 12, 18, 20, 21$, ebből rögtön látszik, hogy a medián 18, tehát az átbocsátás mediánja $\frac{18}{0{,}1} = 180\ \frac{\text{kérés}}{\text{s}}$.
A p kvantilis definíció szerint az a szám, amelynél az elemek p-ed része kisebb vagy egyenlő.A p kvantilisba azok az elemek tartoznak, amelyek kisebb vagy egyenlők a p kvantilisnál. A kvantilis

speciálisabb változata a percentilis, amely egész százalékokkal dolgozik, valamint a kvartilis, amely „negyedeli” az adatot. Pl. a 35. percentilis a 35%-os kvantilisnak felel meg (a kvantilis lehetne pl. 35,7% is!), a második kvartilis pedig az 50%-os kvantilisnak.
Itt az elemek legkisebb 40%-a a 11 és a 12, ezért a 40%-os kvantilis értéke a 12 lesz, és a 11, illetve 12 elemek tartoznak bele. A kapcsolódó átbocsátási ráták 110 $\frac{\text{kérés}}{\text{s}}$ és 120 $\frac{\text{kérés}}{\text{s}}$

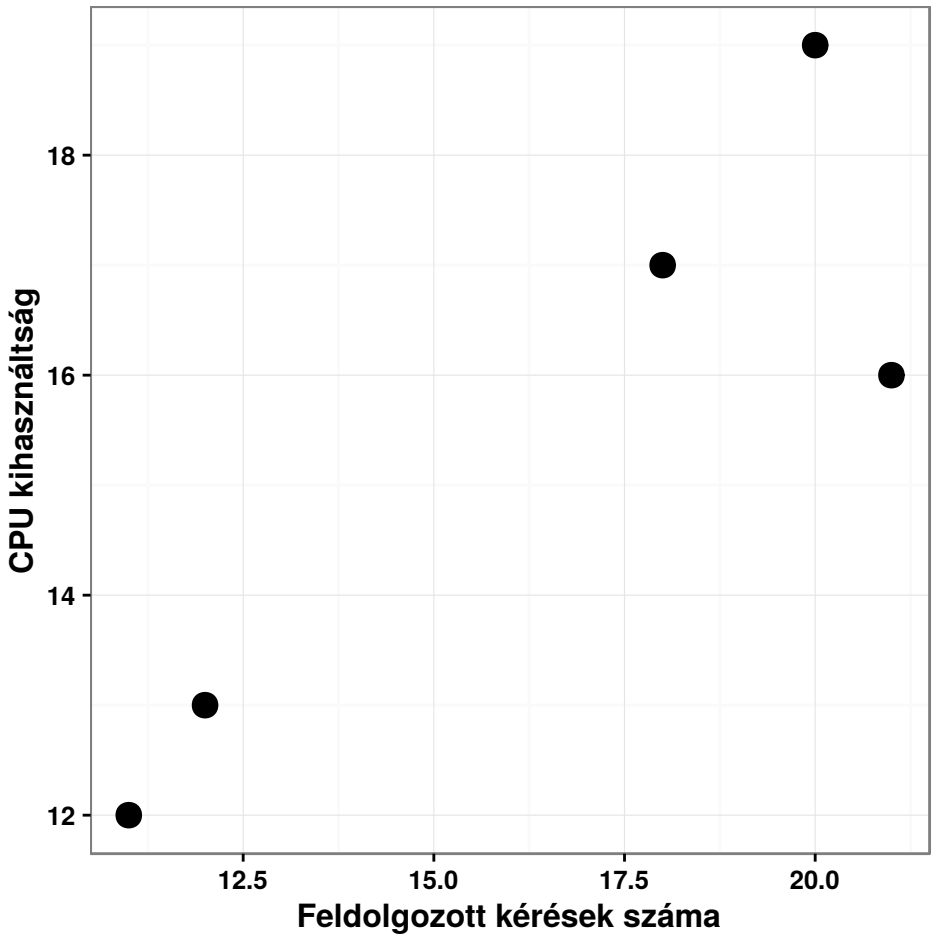
c) Ezen 5 mérés alapján milyen becslést tudunk adni az egyszerre kiszolgálás alatt lévő kérések átlagos számára?
Megoldás
Az utolsó 100 ms alatt feldolgozott kérések számából és az átlagos kiszolgálási időből adódik. Mivel az átlagos kiszolgálási idő különböző elemszámú adathalmazokból került kiszámításra, egyszerű átlagolásuk helyett a feldolgozott kérésekkel súlyozott átlagukat kell vennünk.

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

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

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

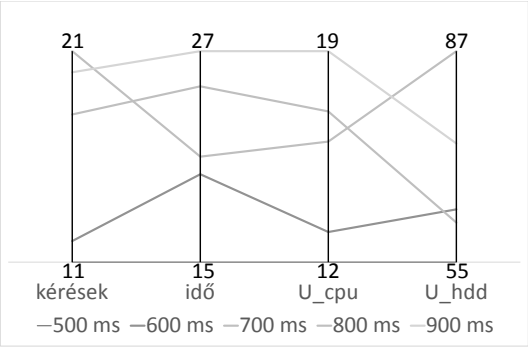
d) Ábrázoljuk a feldolgozott kérések számát és a CPU kihasználtságot pontfelhő (scatterplot) diagramon! Értelmezzük a diagramot! Vajon a mért jellemzők között sejthető ok-okozati viszony?
Megoldás



Először is: nagyon kevés adatunk van, ezért óvatosan szabad csak következtetéseket levonni.
Két klaszter (csoportosulás) látszik, tehát talán két lényegesen eltérő helyzetben figyeltük meg a rendszert. A két vizsgált változó közt úgy tűnik, nagyjából pozitív a korreláció (ezt azt jelenti, hogy ha az egyiket növeljük, akkor tipikusan a másik is nőni szokott), de nincs egyenes arányosság, sőt, nem is monoton az összefüggés (tehát valami más is befolyásolhatja az adatokat, ezért

ingadozik). Értelmezve a látottakat a CPU átlagos kihasználtsága a feldolgozott kérések számával nő. A bal alsó csoport kisebb terhelésű pillanatokat tartalmaz, míg a jobb felső nagyobbakat. Ez a megfigyelés adott esetben jó alapja lehet a terhelés vizsgálatának (pl. az egyes csoportokhoz tartozó pontok időben is közel vannak-e egymáshoz).
Ahogy az ábrán is látjuk, az átbocsátás hatással van az erőforrások kihasználtságra. A szűk keresztmetszetnek számító erőforrás (HDD) magas kihasználtsága meg is látszik a megnyúlt válaszidőkön.

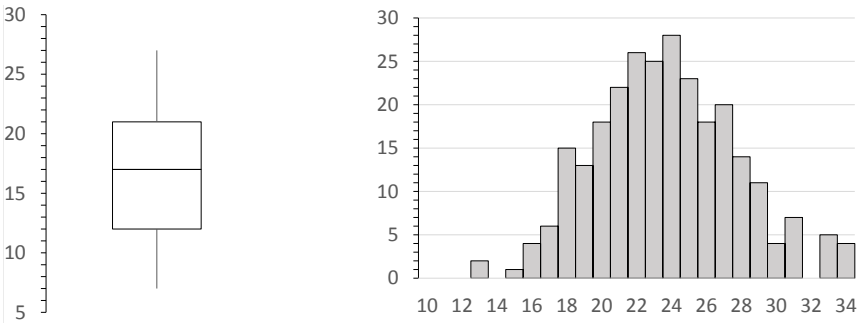
e) Ábrázoljuk az adatokat párhuzamos koordinátákon! Milyen további korreláció olvasható le a diagramról?
Megoldás
A megoldás az ábrán látható.



A diagramról leolvasható az átlagos kiszolgálási idő, és a CPU kihasználtság közti korreláció.

f) A *100ms alatt feldolgozott kérések számá*nak (több száz kísérlet során) mért értékeiből az alábbi boxplotot generáltuk. Milyen kitüntetett értékek olvashatók le a diagramról? Jól jellemzi-e az első öt mérés a teljes adathalmazt?
Megoldás
$Q1 = 12, Q2 = 18, Q3 = 21$, azaz a mért értékek 25%-a legfeljebb 12, 50%-a legfeljebb 18, 75%-a legfeljebb 20.
Az első néhány mérés nagyjából a $Q1-Q3$ intervallumba esik, látható hogy ilyen kis mintavételnél nem szóródnak úgy szét az adatpontok.

g) Az *átlagos kiszolgálási idő* (több száz kísérlet során) mért értékeiből az alábbi hisztogramot generáltuk. Mi olvasható le a diagramról?
Megoldás
A mérések nagyjából normál eloszlást követnek, melynek maximuma 24 ms körül van.

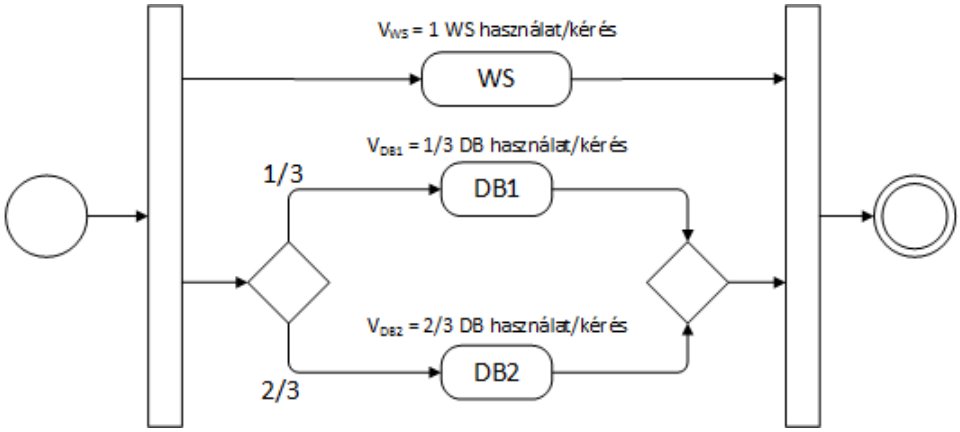


4. Kétrétegű architektúra

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) Készítsünk folyamatmodellt a kérések feldolgozásáról a szöveg alapján!

Megoldás



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 kilönbség, *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!).

b) Mekkora az egyes szerverek jelenlegi átbocsátása?

Megoldás

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)

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

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

c) Mennyi időt töltenek egy-egy kérés kiszolgálásával a szerverek?

Megoldás

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

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

d) Mekkora a rendszer maximális áteresztőképessége?

Megoldás

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

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

Ugyanígy DB2-re és WS-re:

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

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

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:

$$X_{\text{rendszer}} \leq \min\left(\frac{1}{v_{\text{DB1}}} \cdot X_{\text{DB1}}^{\max}, \frac{1}{v_{\text{DB2}}} \cdot X_{\text{DB2}}^{\max}\right). \tag{2}$$

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 az 1 és a 2 egyenlőtlenségekből a maximális átbocsátás, vagyis az átbocsátképesség képlete:

$$X_{\text{rendszer}}^{\max} = \min\left(X_{\text{WS}}^{max}, \frac{1}{v_{\text{DB1}}} X_{\text{DB1}}^{max}, \frac{1}{v_{\text{DB2}}} X_{\text{DB2}}^{max}\right).$$

A feladat megoldásához tehát a komponensek átbocsátóképességeit kell kiszámolnunk:

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

A rendszer maximális átbocsátóképessége ezekből:
$X_{\text{rendszer}}^{\max} = \min\left(6{,}666\ \frac{\text{kérés}}{\text{s}}, 3 \cdot 3{,}704\ \frac{\text{kérés}}{\text{s}}, \frac{3}{2} \cdot 4{,}545\ \frac{\text{kérés}}{\text{s}}\right) =$
$\min\left(6{,}666\ \frac{\text{kérés}}{\text{s}}, 11{,}112\ \frac{\text{kérés}}{\text{s}}, 6{,}818\ \frac{\text{kérés}}{\text{s}}\right) = X_{\text{WS}}^{\max} = 6{,}666\ \frac{\text{kérés}}{\text{s}}$.
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.

e) Miért nem egyféle foglaltsági időt vettünk figyelembe a két erőforrástípusnál?

Megoldás

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.

f) Hol csal még így is a modell?

Megoldás

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,

[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.

de rövidebb időszakokra nézve egy átlagosnál hosszabb végrehajtási idejű kérés például rövid időre telítésbe viheti a rendszert.

5. Közösségi oldal

Internetes közösségi oldalt működtetünk. Az utóbbi időben számottevően népszerűbb lett az oldal, de ezáltal a válaszidő is kellemetlenül megnőtt. 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?

Megoldás

Tehát a kiszolgáló infrastruktúránknak átlagosan 3 másodperces válaszidővel kell kiszolgálni egyszerre 1500 felhasználót. Little-törvényt alkalmazva: $N = 1500$, $T = 3\ \frac{\text{s}}{\text{kérés}}$, tehát $X = \frac{N}{T} = 500\ \frac{\text{kérés}}{\text{s}}$

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 a fenti cél eléréséhez?

Megoldás

$T_{\text{CPU}} = 20$ ms $= 0{,}02$ s, $T_{\text{DB}} = 12{,}5$ ms $= 0{,}0125$ s. A CPU-nak és adatbázisnak is legalább 500 kérést kell tudnia kiszolgálni másodpercenként, hogy a teljes rendszer is képes legyen erre (akár szekvenciális, akár párhuzamos kompozíciót alkalmazunk). Jelenleg az erőforrások egyetlen példányára: $X^{\max}_{\text{CPU}} = \frac{1}{T_{\text{CPU}}} = 50\ \frac{\text{kérés}}{\text{s}}$, $X^{\max}_{\text{DB}} = \frac{1}{T_{\text{DB}}} = 80\ \frac{\text{kérés}}{\text{s}}$. Tehát a 15 webszerver átbocsátó képessége együttesen $750\ \frac{\text{kérés}}{\text{s}}$, míg az 5 adatbázis szerveré csak $400\ \frac{\text{kérés}}{\text{s}}$. Tehát még *kell 2 db adatbázis szerver*, hogy az adatbázis réteg elérje a kívánt átbocsátó képességet.

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 15 webszerver átbocsátó képessége együttesen $X^{\max}_{\text{web}} = 750\ \frac{\text{kérés}}{\text{s}}$, a csúcsidőszakban a szükséges átbocsátás pedig $X_{\text{web}} = 500\ \frac{\text{kérés}}{\text{s}}$. A kihasználtságuk tehát $U_{\text{web}} = \frac{X_{\text{web}}}{X^{\max}_{\text{web}}} = \frac{2}{3}$. Ugyan ezzel a módszerrel: $U_{\text{DB}} = \frac{X_{\text{DB}}}{X^{\max}_{\text{DB}}} = \frac{500}{520} = 0.96$.

Ha 50%-os kihasználtságot szeretnénk, akkor $\frac{X_{\text{web}}}{U_{\text{web}}} \left(= \frac{X_{\text{DB}}}{U_{\text{DB}}}\right) = \frac{500\frac{\text{kérés}}{\text{s}}}{0.5} = 1000\frac{\text{kérés}}{\text{s}}$ átbocsátó képességgel kell rendelkeznie az infrastruktúrának csúcsidőszakban. Ehhez 20 webszerver és 13 adatbázis szerver kell.

d) Tekintsünk csak 2 db webszervert és 3 db adatbázis szervert. Készítsünk állapot alapú modell(eke)t, amely(ek) az infrastruktúra erőforrásait modellezi(k) az elérhetőségeik (szabad/foglalt) szerint. Milyen tervezői döntésekkel szembesülünk? Mik az egyes lehetőségek előnyei és hátrányai?

Megoldás

Lehetőségek:

- Az erőforrásokat *típusonként összevonva* modellezzük aszerint, hogy mennyi foglalt belőlük. Tehát lesz egy 0–1–2 állapotláncunk a webszerverekre, valamint egy 0–1–2–3 állapotláncunk az adatbázis szerverekre. Az erőforráskészlet teljes modellje ezek aszinkron szorzata lesz.
  A megoldás *előnye*, hogy egyszerű. Ha például szeretnénk erőforrás foglalást is modellezni, akkor az könnyen megvalósítható kooperáló állapotgépekkel: ha az erőforrás állapotgépe nem az utolsó állapotban van, akkor sikerül a foglalás, és ezzel szinkronban az erőforrás állapotgépe is lép egyet „jobbra” (már eggyel kevesebb erőforrás szabad). Az erőforrás felszabadítás hasonlóan történik.
  A megoldás *hátránya*, hogy nem szolgáltat arról információt, hogy melyik erőforrás példány mikor szabad vagy foglalt, így nem tudunk például pontos kihasználtságot mondani az egyes példányok esetén, csak egy átlagos értéket, ami az összes szervert jellemzi.
- Minden erőforrás *példányt külön modellezünk* egy szabad-foglalt állapotpárral (vagy akár még részletesebben). Tehát annyi állapotgép régiónk lesz, ahány erőforrás példányunk van. Az erőforráskészlet teljes modellje ezek aszinkron szorzata lesz.

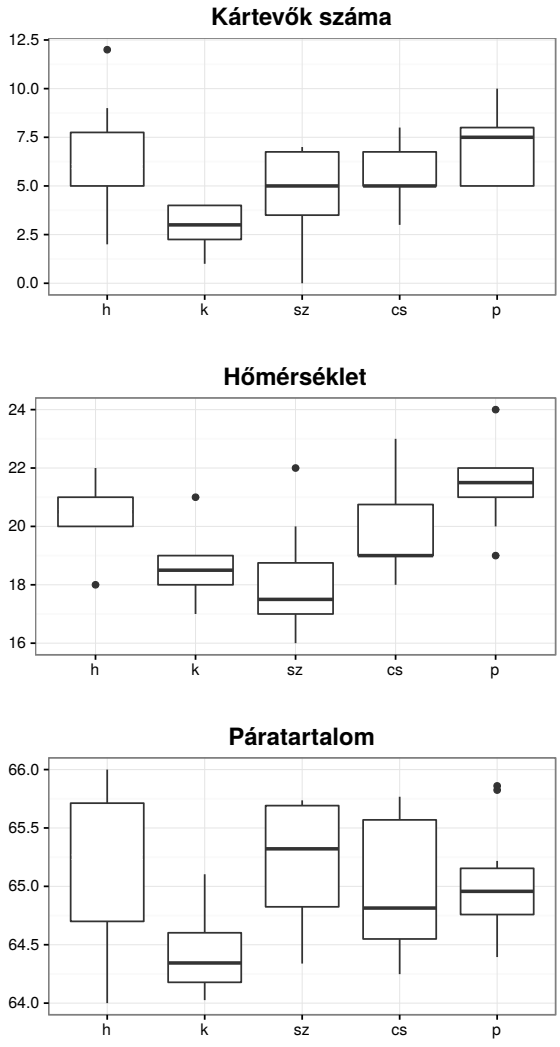
A megoldás *előnye*, hogy konkrét erőforrás példányokra is tudunk például kihasználtságot számolni. Vagy ami még érdekesebb: tudunk erőforrásonként meghibásodást és javítást is modellezni és ennek fényében megnézni az egyes metrikák változását. A meghibásodási és javítási ráták különbözhetnek is az egyes példányok esetén, így lehetőség nyílik heterogén erőforrás kollekció (vagy alkatrész elöregedés) modellezésére is.
A megoldás hátránya, hogy mostantól a fogyasztók felé is több erőforrás példány látszik, ami például megnehezíti a foglalás modellezését. Egy erőforrás foglalásához meg kell keresni egy szabad erőforrást, majd a végén pontosan azt kell felszabadítani. Ez a szituáció még tovább bonyolódik, ha egy művelethez több erőforrásra is szükség van (holtpont, éheztetés). Ebben az esetben célszerű (és szokás is) bevezetni egy erőforrás menedzser komponenst, amely elrejti ezt a folyamatot a fogyasztóktól.

Kiegészítő feladatok

6. Szenzorhálózat (korábbi zh feladat) – adatelemzés

Adott egy mezőgazdasági szenzorhálózat, amellyel a szabadföldes, üvegházi, ill. fóliasátras területeink állapotát követjük nyomon a mért értékek (hőmérséklet, páratartalom, fényerősség, szélsebesség, detektált kártevők stb.) alapján.

Dátum	Hőm. [°C]	Pára. [%]	Kártevők [db]
2015. 05. 04. 08:00	18	66,00	3
2015. 05. 04. 09:00	20	65,75	6
2015. 05. 04. 10:00	20	65,75	8
2015. 05. 04. 11:00	20	65,50	9
2015. 05. 04. 12:00	20	65,50	5
2015. 05. 04. 13:00	21	65,00	12
2015. 05. 04. 14:00	21	64,70	5
2015. 05. 04. 15:00	21	64,70	6
2015. 05. 04. 16:00	21	64,60	7
2015. 05. 04. 17:00	22	64,00	2



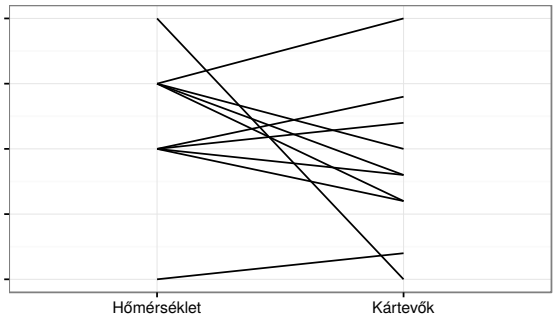
a) Sajnos a május 4. hétfői középértékek (medián) lemaradtak az ábráról, rajzoljuk őket be a táblázatban található adatok alapján!
b) Értelmezze a diagramokat: mely változó(k) első kvartilisei mutat(nak) szigorúan monoton változást az idő folyamán?
c) (Kiegészítő feladat.) Szeretnénk párhuzamos koordináta diagramon összevetni a hétfői hőmérsékleti értékeket a detektált kártevők számával.

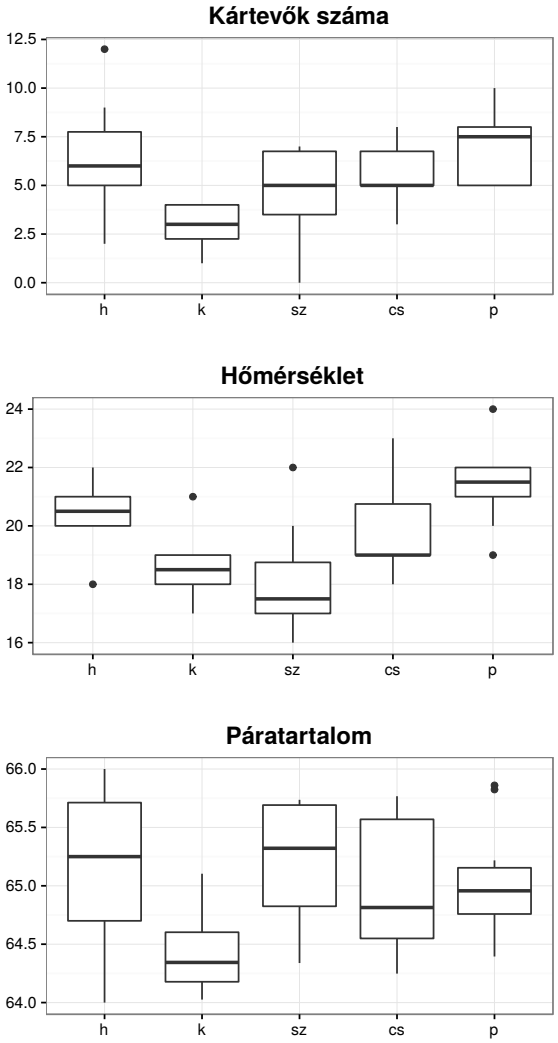
Megoldás

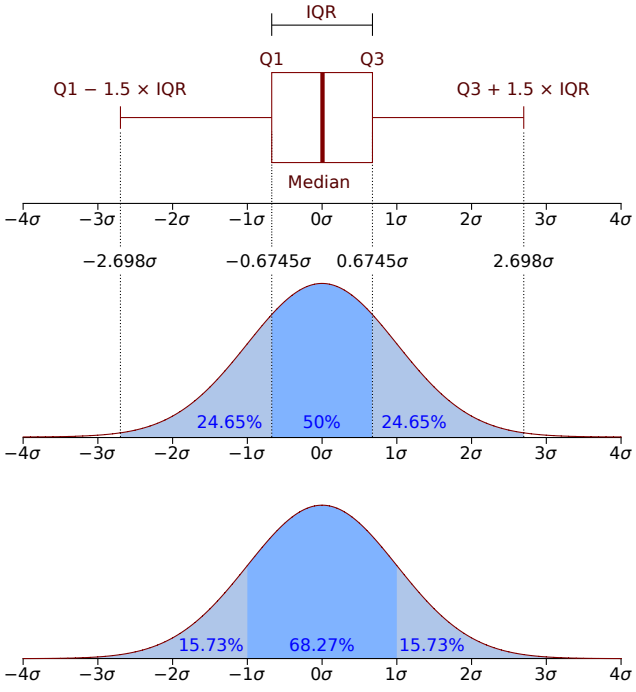
a) Rajzoljuk be a medián értékeket. Mivel páros számú értékünk van, ezért a középső kettő átlaga lesz a medián. Az első két oszlop rendezett, ezért pont a középső két érték átlaga: $\frac{20+21}{2} = 20{,}5$, ill. $\frac{65{,}5+65}{2} = 65{,}25$.
A harmadik oszlop sorbarendezve $2, 3, 5, 5, 6, 6, 7, 8, 9, 12$, így a medián $\frac{6+6}{2} = 6$.

b) Egyik sem, hiszen a „dobozok” alja nem mutat sehol szigorúan monoton változást.
A boxplot főbb jellemzőit az 1. ábra mutatja be. A $\pm 1.5 \times$ IQR-en kívül eső értékeket ponttal jelöljük.
Érdekességként megjegyezzük, hogy a 1.5 konstans használata egy statisztikai konvenció, amely analóg a a normális eloszlású adathalmazok $\pm 3\sigma$ elvével.

c) Az értékeket az alábbi párhuzamos koordináta diagramon ábrázoljuk.







1. ábra. A boxplot főbb jellemzői

7. Szenzorhálózat (korábbi zh feladat) – teljesítményelemzés (∗)

(A 6. feladathoz kapcsolódó teljesítményelemzési feladatok.) A különböző típusú szenzorok a helyüktől számított 100 méteres körzetben lévő területekről szolgáltatnak adatokat. A szenzorok mérési eredményeiket időbélyeggel ellátva, rádiós kommunikációs hálózaton továbbítják a központnak. A központi számítógép processzora feldolgozza a kéréseket, majd archiválási célzattal kiírja őket egy tárolóegységre. A gazdaságunk összesen 4500 szenzort telepített, amelyek percenként egy-egy mérési eredményről adnak jelentést. A rendszer sikerrel kiszolgálja a terhelést. A rádiós kommunikációs hálózat 100 mérési eredményt képes másodpercenként továbbítani. A központi számítógép CPU idejének 75%-a tétlenül múlik. A tárolóegységet 8 ms-ig foglalja le minden egyes kérés kiírása.

a) Másodpercenként hány mérési adat a rendszer jelenlegi átbocsátása?
Megoldás
$X = 4500 \text{ szenzor} \cdot 1 \text{ adat}/60 \text{ s} = 75 \text{ adat/s}$

b) Mekkora a hálózat, CPU, ill. tároló átbocsátása, átbocsátóképessége és kihasználtsága?
Megoldás
$X_{\text{hálózat}} = X_{CPU} = X_{\text{tároló}} = X = 75$ adat/s, mert minden vizitációs szám 1.

$$X_{\max}^{\text{hálózat}} = 100 \text{ adat/s} \rightarrow U_{\text{hálózat}} = X_{\text{hálózat}}/X_{\max}^{\text{hálózat}} = 75/100 = 75\% = 0{,}75$$

$$U_{\text{CPU}} = 1 - 0{,}75 = 0{,}25 = 25\% \rightarrow X_{\max}^{\text{CPU}} = X_{\text{CPU}}/U_{\text{CPU}} = 75/0{,}25 \text{ adat/s} = 300 \text{ adat/s}$$

$T_{\text{tároló}} = 0{,}008$ s és nincs átlapolódás:

$$X_{\max}^{\text{tároló}} = 1/T_{\text{tároló}} = 125 \text{ adat/s} \rightarrow U_{\text{tároló}} = X_{\text{tároló}}/X_{\max}^{\text{tároló}} = 75/125 = 60\% = 0{,}6$$

c) A mérési pontosság javításához hány szenzort helyezhetünk még üzembe ugyanezen a területen az infrastruktúra fejlesztése nélkül? Feltételezzünk lineáris skálázódást!
Megoldás
Mivel minden mérés feldolgozásához igénybe vesszük mindhárom erőforrást:

$$X_{\max} = \min(X_{\max}^{\text{hálózat}}, X_{\max}^{\text{CPU}}, X_{\max}^{\text{tároló}}) = X_{\max}^{\text{hálózat}} = 100 \text{ adat/s}$$

Tehát 4/3 arányú felskálázás lehetséges, még 1500 szenzor üzembe helyezhető.

d) A rádióhálózat ügyes kódolással biztosítja, hogy egyszerre több szenzor is sugározhasson mérési eredményeket. Átlagosan hány szenzor rádiója sugároz egyszerre (vagyis hányszoros az átlapolódás) a hálózaton jelenleg, ill. a hálózat maximális terheltsége esetén, ha egy mérési eredmény sugárzása 40 ms-ig tart?
Megoldás
Alkalmazzuk Little törvényét a levegőben épp sugárzás alatt álló üzenetekre:

$$T_{\text{hálózat}} = 0{,}040 \text{ s}$$

$$X_{\text{hálózat}} = 75 \text{ adat/s} \rightarrow N_{\text{hálózat}} = X_{\text{hálózat}} \cdot T_{\text{hálózat}} = 75 \text{ adat/s} \cdot 0{,}04 \text{ s} = 3 \text{ adat egyszerre most}$$

$$X_{\max}^{\text{hálózat}} = 100 \text{ adat/s} \rightarrow N_{\text{hálózat}} = 100 \text{ adat/s} \cdot 0{,}04 \text{ s} = 4 \text{ adat maximálisan}$$

8. Sziget közlekedési hálózata (∗ korábbi zárthelyi 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)?
Megoldás
Little törvényében az átbocsátás szerepel, nem az átbocsátóképesség – de abban a speciális esetben, amikor pont telítve van a rendszer, a kettő megegyezik:
- $N = X \cdot T \rightarrow X = \frac{N}{T}$;

- $N = \frac{200\text{ m}}{30\text{ m/kocsi}} = \frac{20}{3}$ 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) Hány autó fér el egy kompban?
Megoldás
Az előzőhöz hasonlóan Little törvényéből az átbocsátóképesség:
- $N = X \cdot T; X = 800\ \frac{\text{kocsi}}{\text{h}}$;
- $T = 15\text{ min} = 0,25\text{ h}$;

ekkor $N = 200$, tehát egyszerre 200 autó utazik. Mivel 4 hajó van, ezért egy hajóra 50 kocsi fér fel.

c) A reggeli csúcsforgalomban mekkora a szigetet elhagyó két útvonal együttes átbocsátóképessége?
Megoldás
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) 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?
Megoldás
A terelőút átbocsátóképessége (soros kompozíció): $X = \min(X_{\text{híd}}, X_{\text{komp}}) = 800\ \frac{\text{kocsi}}{\text{h}}$.

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?
Megoldás
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) 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
Komphoz sorbanállásra Little-törvény: $N = X \cdot T = 0,5\text{ min} \cdot 600\ \frac{\text{autó}}{h} = 5$ autó.

9. Tudásbázis (∗)

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?
Megoldás
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) Tekintsük úgy, hogy az egyes kérések a szerveren nem párhuzamosíthatóak. Óránként hány egyedi látogatót képes kiszolgálni a szerver?
Megoldás
Maximális eset, amikor a kihasználtság 100%, azaz $U = 1$. Ekkor $U = X \cdot T \rightarrow X = \frac{U}{T} = \frac{1}{0{,}2} = 5\ \frac{\text{látogató}}{\text{s}}$. Óránként $3600\text{ s} \cdot 5\ \frac{\text{látogató}}{\text{s}} = 18000$ látogató.

[4] Geometriai eloszlás várható értéke (Wikipédia) `http://hu.wikipedia.org/wiki/Geometriai_eloszlás`