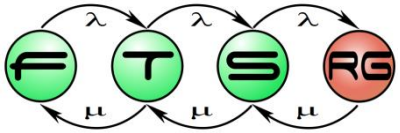
Teljesítménymodellezés

Gönczy László

gonczy@mit.bme.hu



Teljesítménymodellek

$$T_{Kiszolgálás} = T_{Várakozás} + T_{TénylegesKiszolgálás} \left(+T_{HálózatiKésleltetés}\right)$$

- Modellezés célja
  - erőforrás foglalási problémák felderítése
  - elosztott alkalmazások kommunikációs költségei
  - rendszer változásának hatásai (pl. gyorsabb szerver)
  - előrejelzés támogatása
- Ökölszabályok: teljesítménymodell elfogadható, ha
  - az erőforrások kihasználtságát 10%
  - az áteresztőképességet 10%
  - a válaszidőt 20% hibával becsli

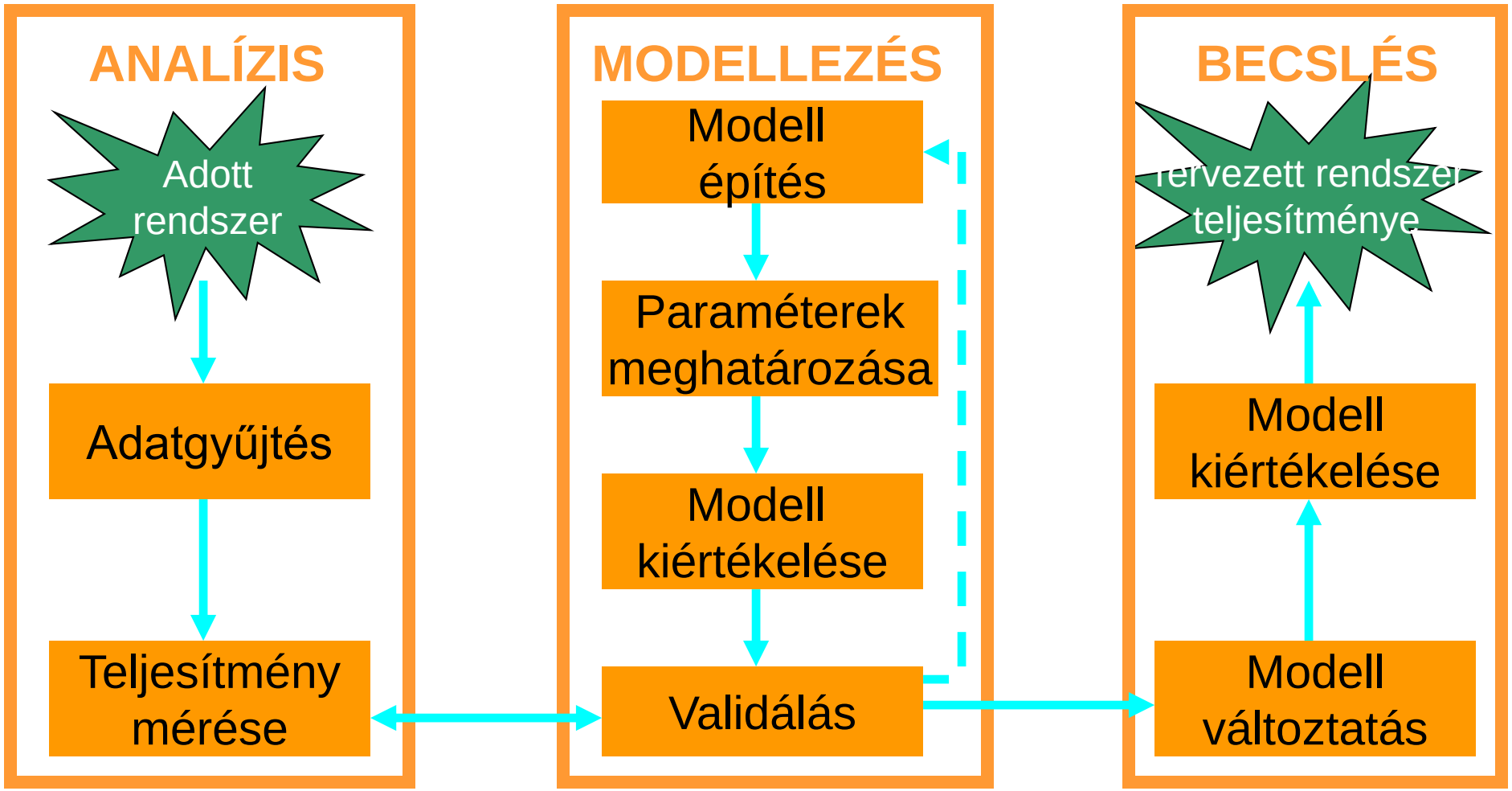
Modellek fajtái

- Analitikus modell
  - a rendszert egyenletekkel írja le, pl.

$$RT_{\min} = RTT + kérés_{\min} + Feldolgozásiidő + válasz_{\min}$$

- Szimulációs modell
  - szimulációt futtat
  - az előfordulásnak megfelelő tranzakció gyakorisággal
  - előny: általános vizsgálat ☺
  - hátrány: drága, nehéz kifejleszteni ☹

Modellezési/becslési paradigma



Honnan lesznek adataink?

- Hol mérünk?
  - Felhasználó szemszögéből
  - Üzemeltető szintjén
  - Alkalmazástól függően
- Mérés/monitorozás
  - Többszintű architektúra
  - Virtualizált környezetek
  - Mintavételezés
  - Mérés többletköltsége, mérési hibák
  - Nagyszámú változó
  - Topológia változása
  - Sok mért érték kezelése (ld. virtualizált példa)

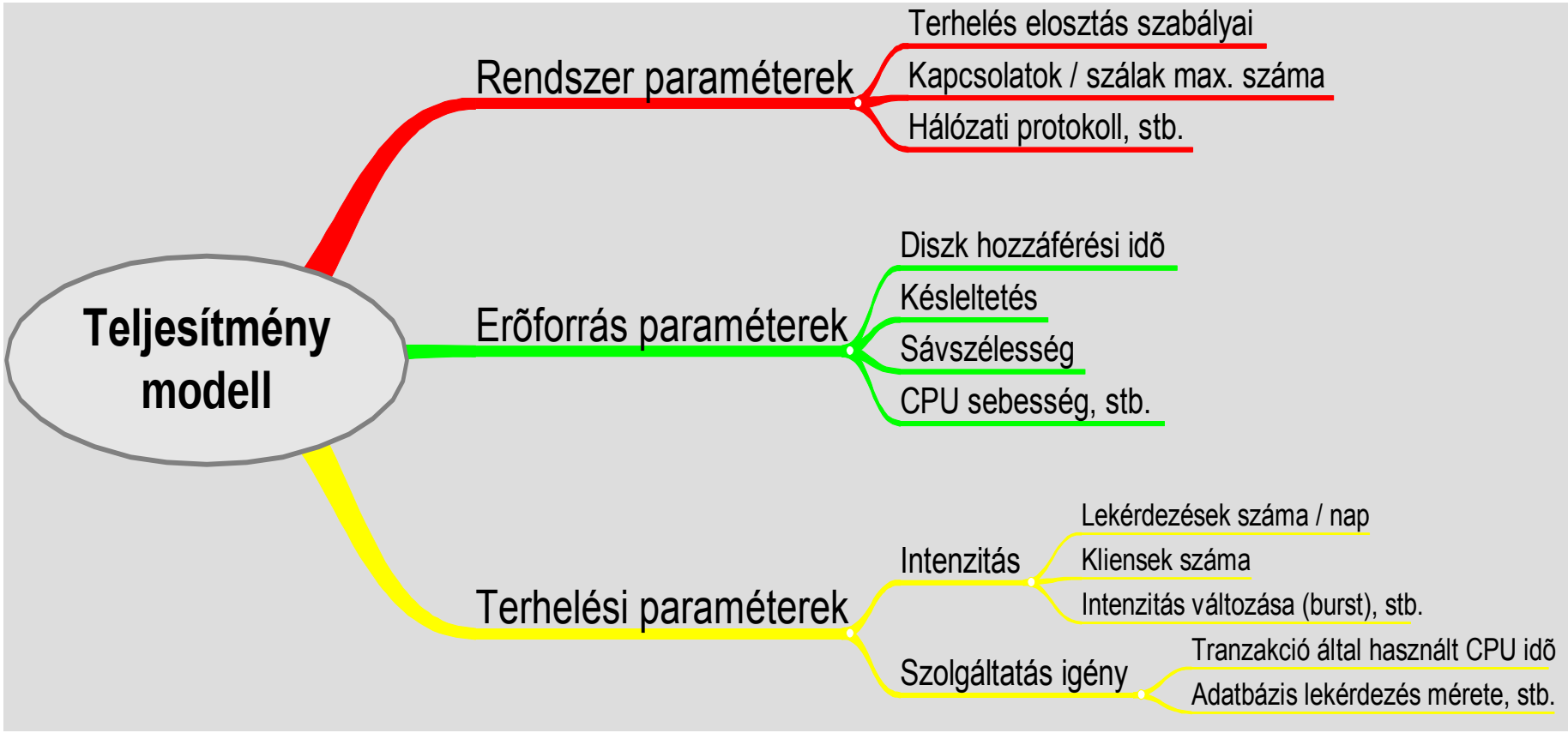
  →Valós környezetben nehéz feladat

Mérési példa

- A következő rendszert mértük 1 hónapig:
  - 2 fizikai gép (1 CPU 4 mag, 18 GB RAM, 2 db Gbit adapter, iSCSI hálózati tárhely)
  - 10 db VM (vegyesen Linux és Windows)
- 94 (fizikai) + 50 (VM) számláló rögzítése
- 20 sec felbontás

∑ 290 MB-nyi CSV fájl
∑ 89 millió mérési pont

Teljesítmény modell paraméterei



Szolgáltatás igény és idő

- Az i. erőforrásra:

D_i: egy tranzakció átlagos szolgáltatásigénye

V_i: a tranzakció átlagos erőforrás használata

S_i: egy használat átlagos erőforrás igénye

$$D_i = V_i \times S_i$$

Könnyebben meghatározható

$$S_i = \frac{D_i}{V_i}$$

Kihasználtság törvénye

- Kihasználtság törvénye (Utilization Law)

$$U_i = B_i/T = B_i/(C_0/X_i) = (B_i/C_0) \times X_i = S_i \times X_i$$

U_i: az i. erőforrás kihasználtsága

B_i: foglaltsági ideje a monitorozás alatt

T: mérési idő

C_0: tranzakciók száma

S_i: átlagos kiszolgálási idő

X_i: átlagos átbocsátás

λ_i: érkezési ráta

$$Egyensúly: \lambda_i = X_i$$

$$U_i = X_i \times S_i = \lambda_i \times S_i$$

További törvények

Forced Flow törvény:

X_i : az i. erőforrás átbocsátása

V_i : „látogatások” átlagos száma

X_0: tranzakciók átlagos száma

$$X_i = V_i \times X_0$$

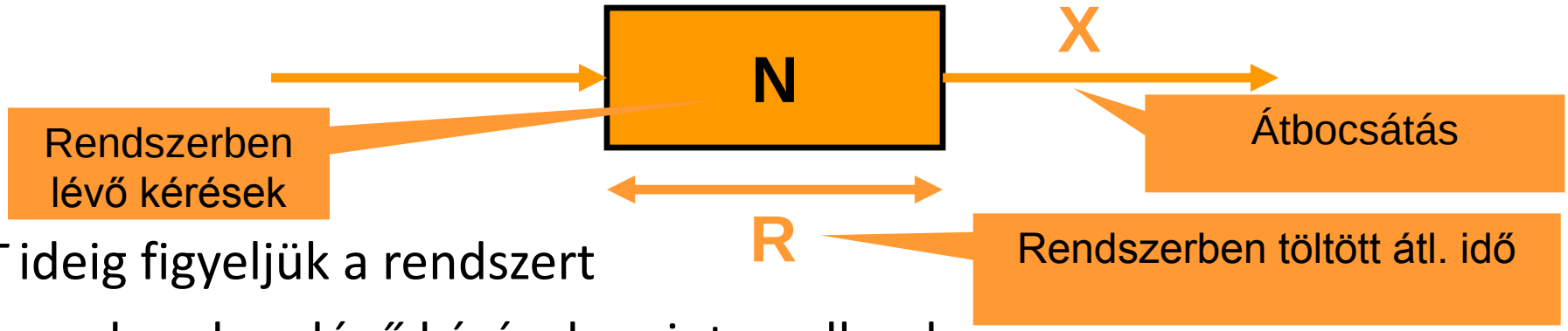
Szolgáltatás igény törvénye (Service Demand):

$$D_i = V_i \times S_i = (X_i/X_0) \times (U_i/X_i) = U_i/X_0$$

Forced Flow

Kihasználtság tv.

Little törvénye



T ideig figyeljük a rendszert

k rendszerben lévő kérések az intervallumban,

f_k az idő amíg k darab kérés van a rendszerben

$$N = \sum_k k \times f_k = \sum_k k \times \frac{r_k}{T}$$

r_k a rendszerben töltött idő összege

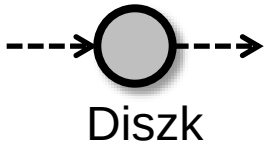
C_0: ennyi kérés hagyta el a rendszert

$$N = \frac{C_0}{T} * \frac{\sum_k k * r_k}{C_0}$$

X R

$$N = X \times R$$

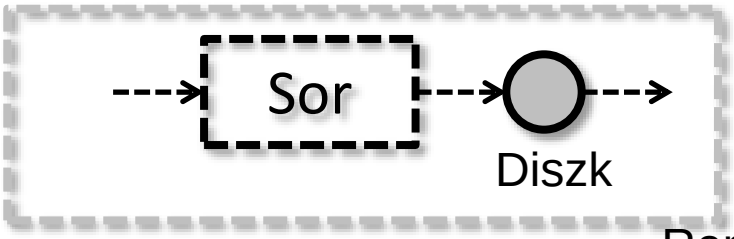
Példa



- Erőforrás: diszk
- 40 kérést szolgál ki másodpercenként (nincs átlapolódás)
- 1 kérés kiszolgálása átlagosan 0,0225 másodpercig tart
- Mekkora a kihasználtság?

$$U = X \times S = 40 \text{ kérés/mp} \times 0{,}0225 \text{ mp} = 90\%$$

Példa

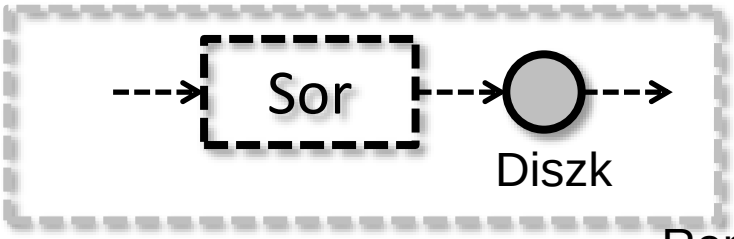


- Sorban állás is van a diszk előtt
- A diszk 40 kérést szolgál ki másodpercenként
- Kérések átlagos száma a rendszerben: 4

Átlagos rendszerben tartózkodási idő?

Átlagos sorban állási idő?

Példa



- Sorban állás is van a diszk előtt
- A diszk 40 kérést szolgál ki másodpe
- Kérések átlagos száma a rendszerbe

Sorbanállási és diszk kiszolgálási idő

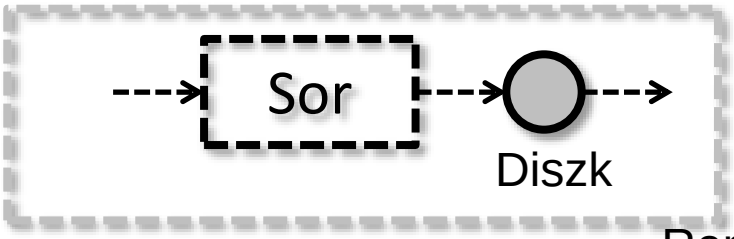
Rendszer

$N = X \times R \rightarrow R = 4$ kérés / 40 kérés/mp = 0,1 mp

Átlagos sorban állási idő?

(Teljes idő – Diszk idő) *0,1 mp – 0,0225 mp = 0,0775 mp*

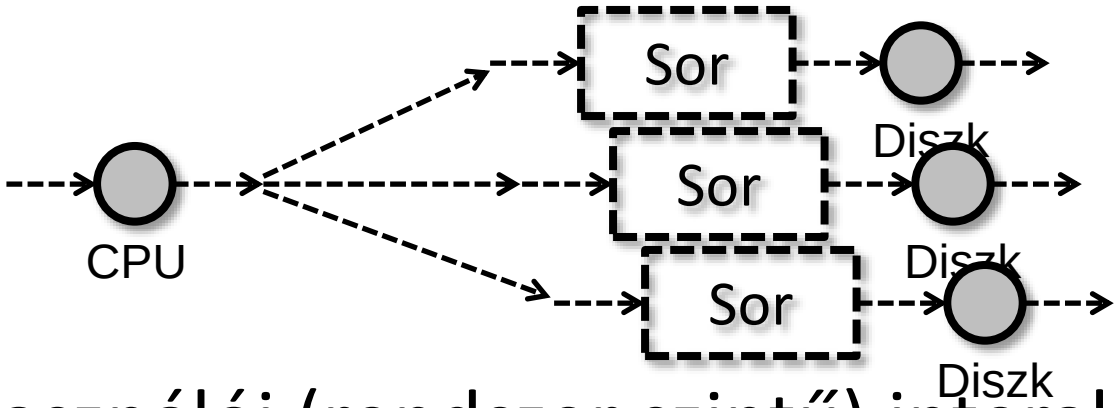
Példa



- Sorban állás is van a diszk előtt
- A diszk 40 kérést szolgál ki másodpercenként
- Kérések átlagos száma a rendszerben: 4

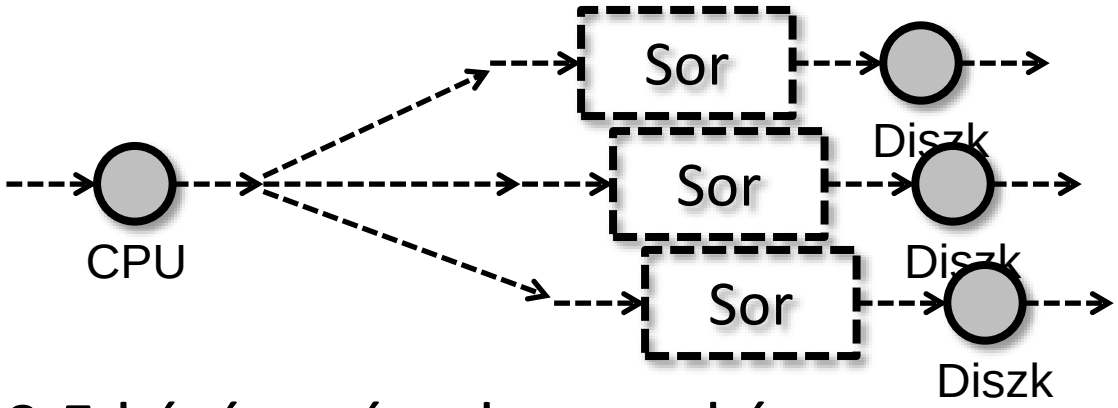
Kérések átlagos száma a sorban?
(Rendszerben lévő kérések száma – diszkben feldolgozás alatt lévő kérések száma)
4 kérés – 0,9 kérés = 3,1 kérés

Példa



- Kérés: felhasználói (rendszer szintű) interakció
- Rendszerben töltött idő
  - Nem foglalkozunk külön-külön az erőforrásokkal
  - Az az idő, ami alatt a felhasználó választ kap a rendszertől a kérésére

Példa



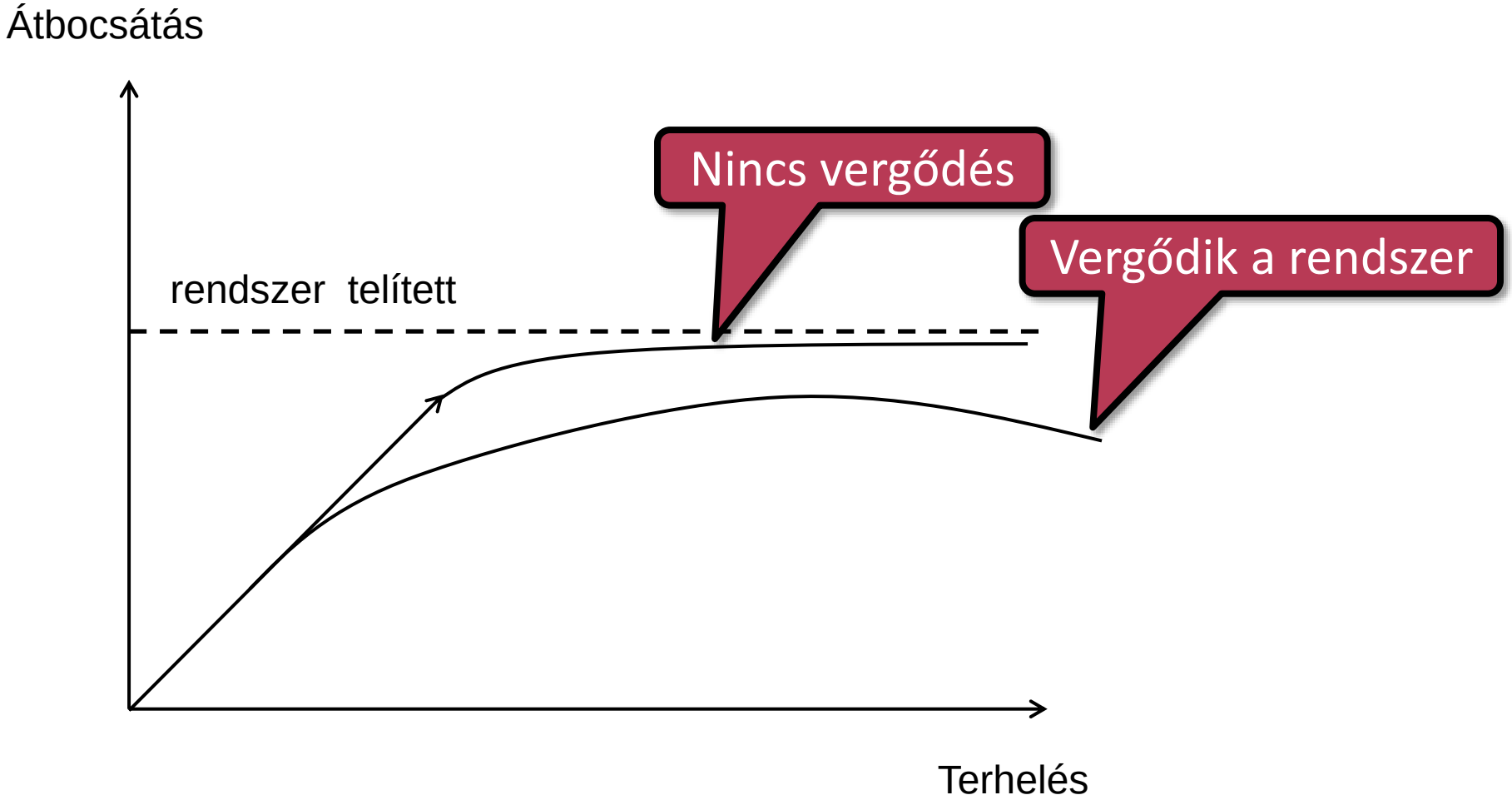
- Átbocsátás: 0,5 kérés másodpercenként
- Átlagosan 7,5 felhasználó várakozik válaszra
  - Átlagosan ennyi kérés van a rendszerben
- Átlagosan mennyi idő alatt lesz egy felhasználó kérése kiszolgálva?

N = X × R → R = 7,5 kérés / 0,5 kérés/mp = 15 mp

Rendszer viselkedése terhelés hatására

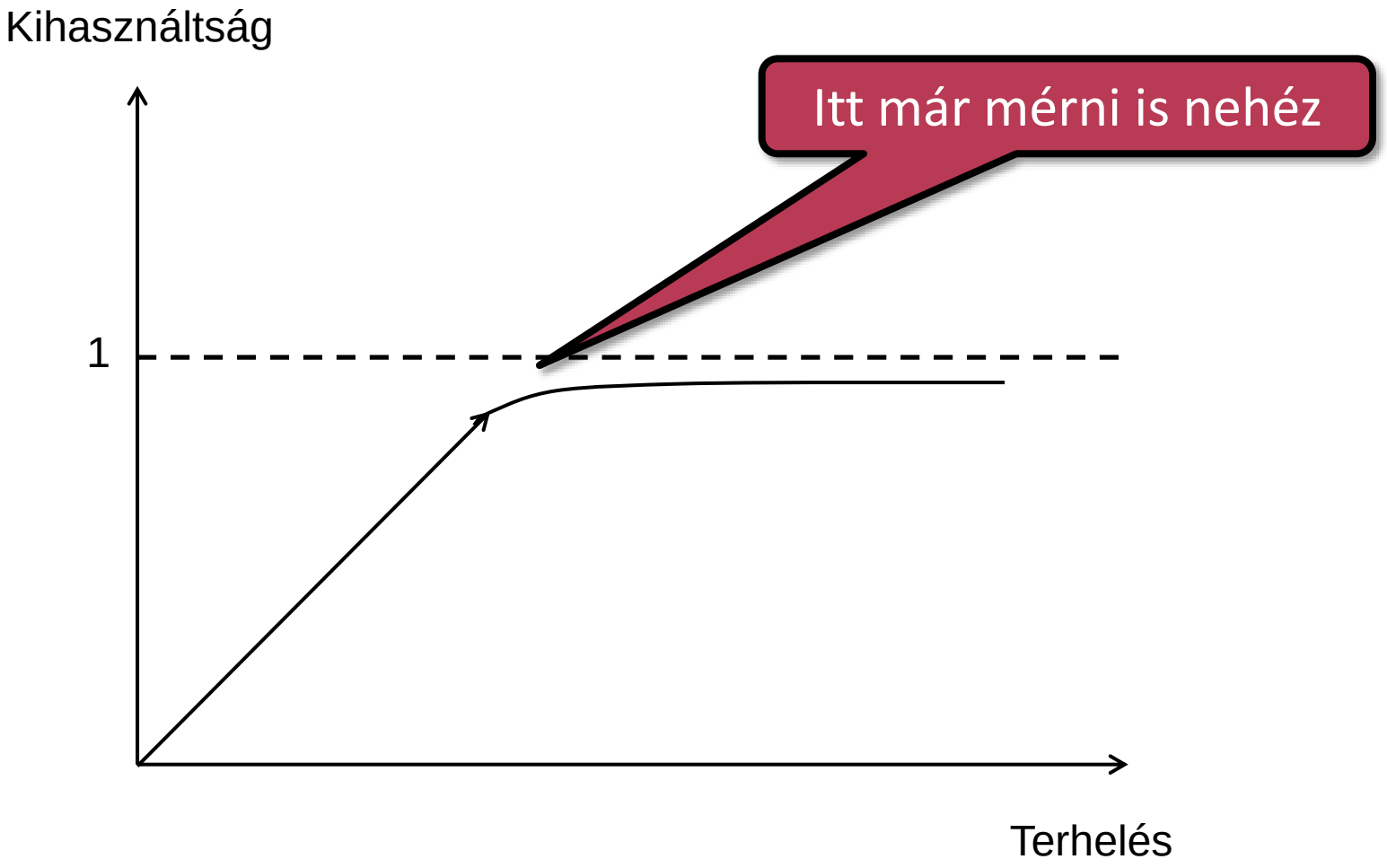
Terhelés-átbocsátás kapcsolata

- Terhelés = érkezési ráta



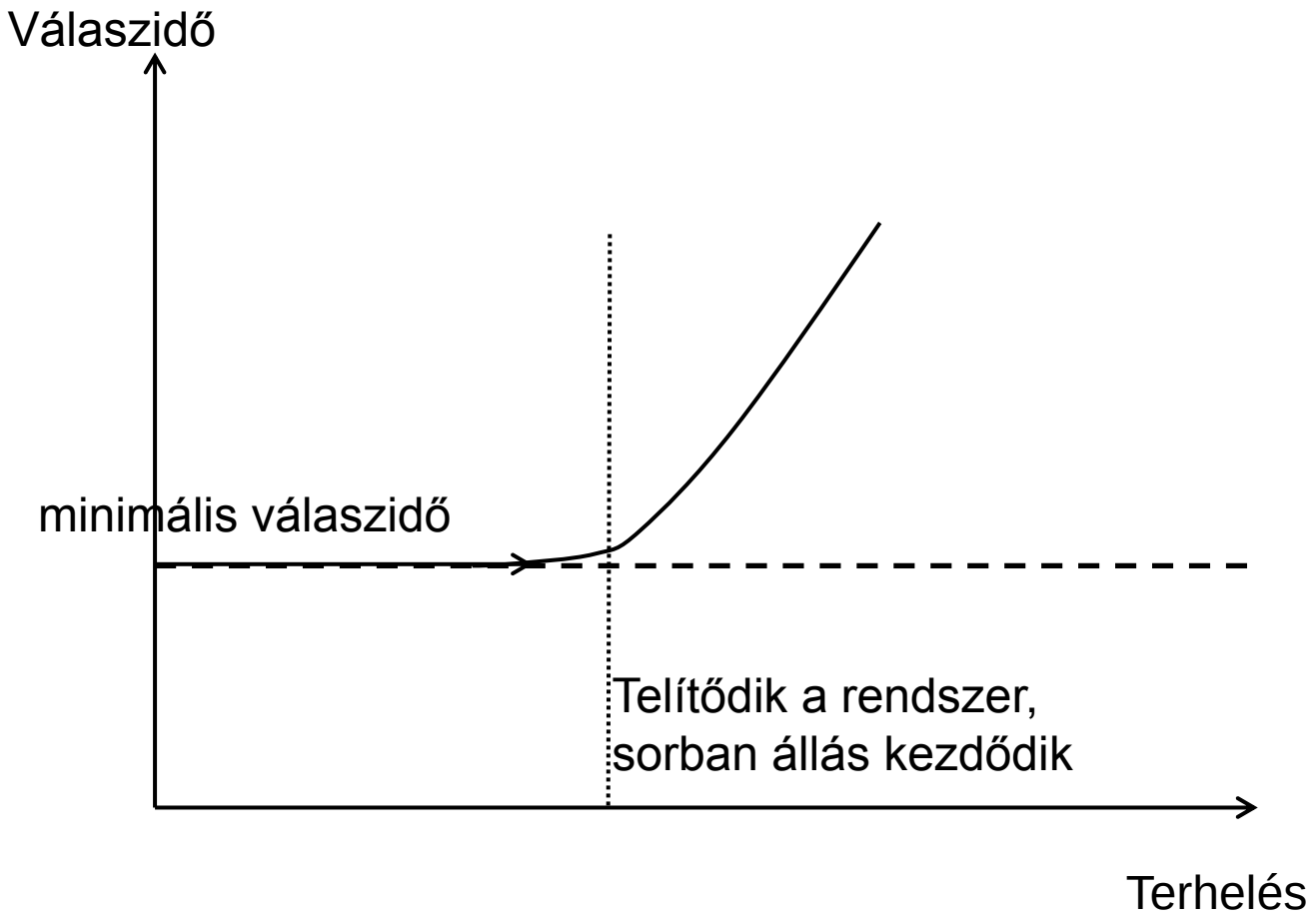
Terhelés-kihasználtság kapcsolata

- Terhelés = érkezési ráta



Terhelés-válaszidő kapcsolata

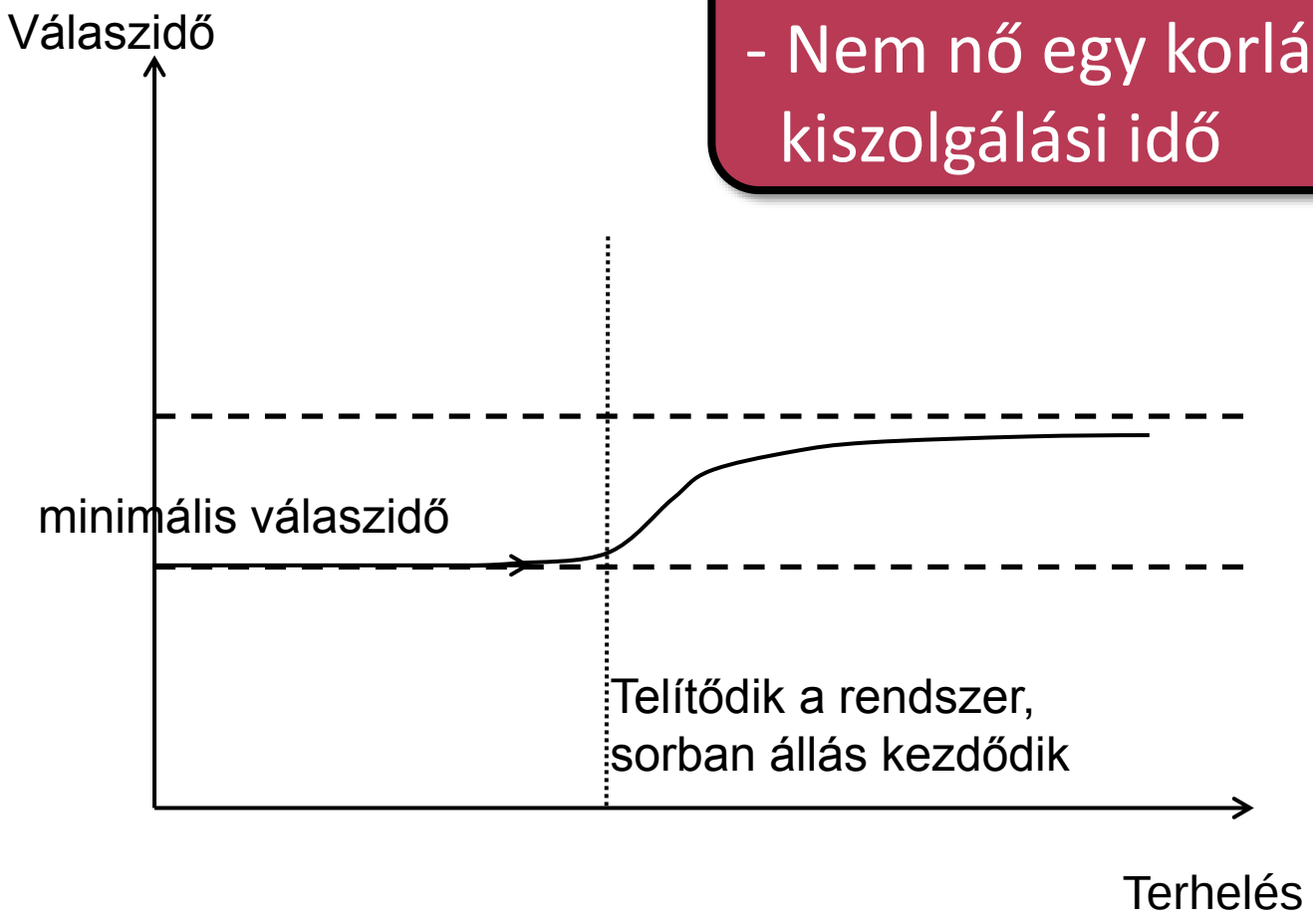
- Nem korlátozzuk a felhasználók számát



Terhelés-válaszidő kapcsolata

- Korlátozzuk a felhasználók számát (admission control)

- Elutasítunk kéréseket!
- Nem nő egy korlát fölé a kiszolgálási idő



Szűk keresztmetszet, skálázhatóság

- Szűk keresztmetszet:
  - azok az erőforrások, melyek korlátozzák a teljesítményt
  - analízis maximális terhelésnél (pesszimista érték)
- Nem mindig kell pontos szám
  - Aszimptotikus Határérték Analízis (Asymptotic Bound Analysis)
- Fogalmak
  - K: erőforrások száma
  - D_i: szolgáltatásigény az i. erőforráson
  - D_{max}: a legnagyobb szolg. igény
  - D_{min}: a legkiseb szolg. igény
  - λ_{max}: érkezési ráta

Nyílt modellek

- Nincs explicit korlát a rendszerben lévő kérésekre
- Ha az egyensúly teljesül:

$$\lambda_{max} \leq \frac{1}{D_{max}}$$

- 3 rétegű architektúra példa

Egy tipikus E-Business szolgáltatás adatai			
Réteg	**Szerverek száma**	**Szolg. átl. igénybevétele**	**Átl. szolg. idő**
Web szerver	5	1,8	110 msec
Alkalmazás szerver	3	2,5	230 msec
DB szerver	2	2,3	180 msec

- Mi a szűk keresztmetszet?
- Mennyi a max. áteresztőképesség?

Nyílt modellek 2.

$$D_{WS} = \left(\frac{V_{WS}}{N_{WS}}\right) \times S_{WS} = \left(\frac{1.8}{5}\right) \times 0.110 = 0.0396\,\text{sec}$$

$$D_{AS} = \left(\frac{V_{AS}}{N_{AS}}\right) \times S_{AS} = 0.192\,\text{sec}$$

$$D_{DB} = \left(\frac{V_{DB}}{N_{DB}}\right) \times S_{DB} = 0.207\,\text{sec}$$

Az adatbázis szerver a szűk keresztmetszet

$$\lambda \leq \frac{1}{0.207\,\text{sec}} = 4.83\,\frac{1}{\text{sec}}$$

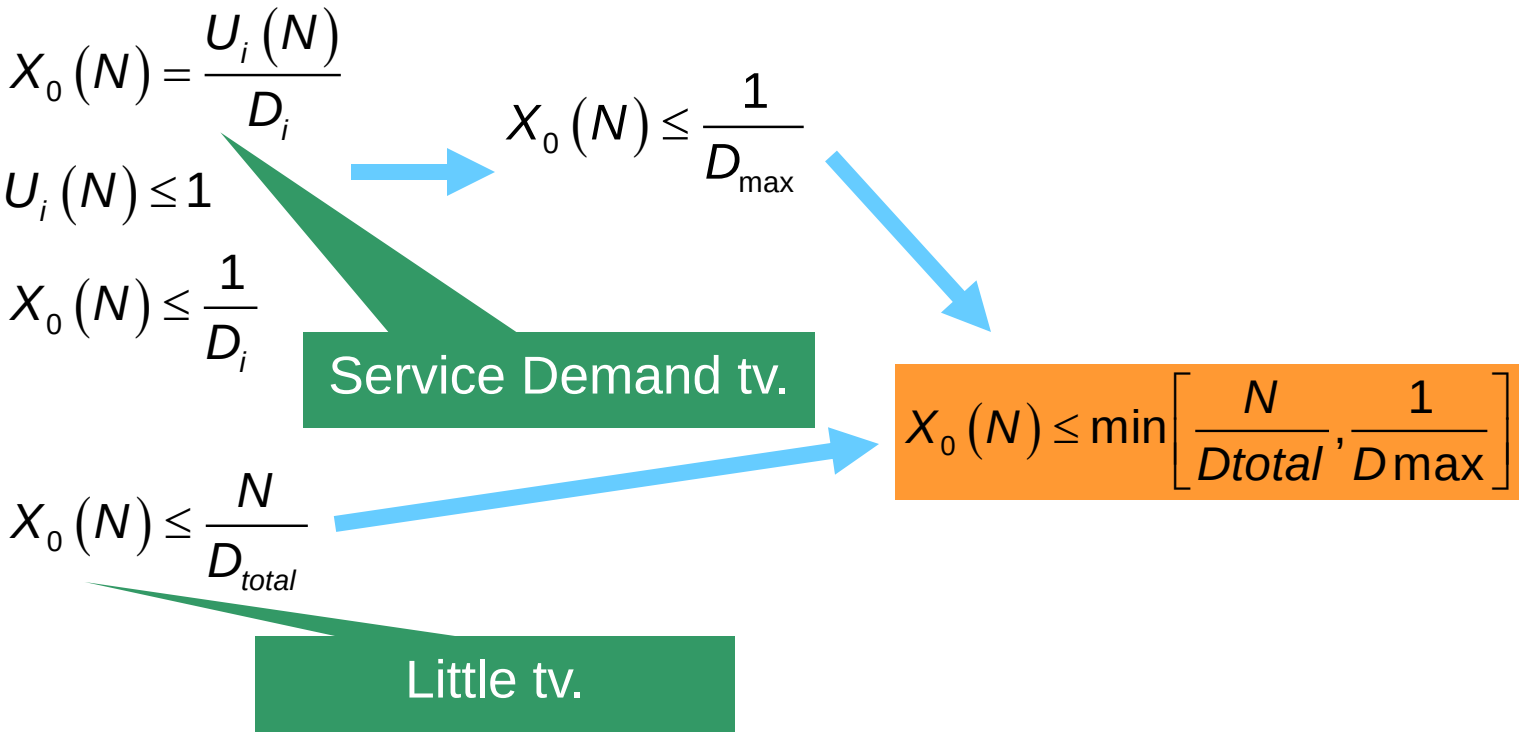
Tranzakciók max. gyakorisága

$$X_0 \leq \frac{X_{DB}}{V_{DB}} = \frac{4.83}{2.3} = 2.1\,\frac{\text{szolgáltatás}}{\text{sec}}$$

E-Business szolgáltatások max. gyakorisága (Forced Flow)

Zárt modellek

- Felső korlát a kérések számára vonatkozóan (*N*)
- Ideális eset: nincs kérés várakozási sorban



Zárt modellek 2.

- Példa: az előbbi 3 rétegű architektúra

Egyszerre max. 20 kliensnek nyújt szolgáltatást: $N = 20$

$$D_{total} = D_{WS} + D_{AS} + D_{DB} = 0.0396\,\text{sec} + 0.192\,\text{sec} + 0.207\,\text{sec} = 0.4386\,\text{sec}$$

$$X_0(20) \le \min\left[\frac{20}{0.4386\,\text{sec}}, \frac{1}{0.207\,\text{sec}}\right]$$

$$X_0(20) \le \min\left[45.59\,\frac{1}{\text{sec}}, 4.83\,\frac{1}{\text{sec}}\right]$$

$$X_0(20) \le 4.83\,\frac{1}{\text{sec}}$$